REVIEWER #1

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

General comments:

Despite the increasing significance of BGC models, the model validation is limited to the comparison with satellite estimates of surface properties, the climatological data, and/or sparse in-situ observations. In recent years, the fast growing BGC-Argo network provides opportunities to evaluate BGC models in an unprecedented spatial and temporal resolutions. Since there is a large number of floats at the global scale, it becomes difficult to evaluate the global model through the point-to-point comparison which has been used in the regional model. This study suggests some BGC-Argo-based metrics to evaluate a global model and provides some diagnostic plots to display these metrics. This manuscript is well structured and easy to follow. I would suggest to publish after minor revision.

REPLY: Thanks for the positive assessment of our work.

Specific comments:

P5 Line 18-23: The BGC-Argo-based POC concentrations were obtained from the filtered bbp signals and therefore should be the small, slow-sinking POC (i.e., 0.2-20μm) (Dall'Olmo and Mork, 2014; Lacour et al., 2019). In the following model description (P6 Line 15-16), the authors mentioned that their POC model had two size classes. Which modelled POC class was compared with the BGC-Argo based one?

REPLY: Following the approach of Gali et al. (2021) and based on the reviewer's suggestion, in the new version of the manuscript we compare the two sizes classes of phytoplankton, the small detrital particles and microzooplankton modelled by PISCES to match the small and slow-sinking POC observed by the BGC-Argo floats.

Based on Roesler et al. (2017), the BGC-Argo based chlorophyll were suggested to be divided by a factor of 2 due to the systematic error in fluorometers. It seems that the authors did not apply this correction to the chlorophyll. If not, this can partially explain the model underestimation of surface chlorophyll in the high-chlorophyll regions (please see the Figure 4). The authors should include some descriptions on how they process the BGC-Argo based chlorophyll.

REPLY: The gain adjustment of 0.5 is already implemented in the “adjusted“ chlorophyll data (Bittig et al., 2019). We have not applied any processing to the BGC-Argo data apart from those already applied at the Data Assembly Center levels as described in the given references. We have clarified this point in the revised version of the manuscript.

P5 Line 25-30: I am concerned with the comparisons between the global model and the estimates from CANYON-B neural network which is also a model. Although it has been validated with some independent observations (e.g. the GO-SHIP cruise data and BGC-Argo floats), differences between the global model and the CANYON-B neural network may come from the CANYON-B neural network’s deviation from the observations.

REPLY: We agree with the reviewer that we do not provide justification for mixing together BGC-Argo data with CANYON-B estimates. We added the following paragraph in the Data section to justify our choice.

“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.”

P8 Line 1-2: Since the POC concentrations vary a lot (~ 2 orders of magnitude) within the mesopelagic zone, the averaged POC concentrations will be skewed to the upper layers right below the mixed layer. In addition, the reference is not appropriate here since the upper bound of mesopelagic zone was defined as the base of productive layer (the maximum of mixed layer and the euphotic zone) in Dall’Olmo and Mork (2014).

REPLY: We agree with the reviewer. We used a log10-transformation to represent the data to account for the skewness in this layer. In the revised version of the manuscript, we have removed the Dall'Olmo and Mork (2014) reference.

P8 Line 13: What is the definition of H? I guess it is the mixed layer depth (MLD).

REPLY: It is an omission on our part. H is the mixed layer depth. We have replaced H by MLD.

P13 Line 26-29: I don't agree with this sentence "However, this seems to have a limited effect on the export of POC …". First, the conclusion here is anti-intuitive because the authors mixed up the POC concentration with the POC export flux. In the north Atlantic, the POC concentration is largely determined by the small, slow-sinking POC. Although the relative contributions of small, slow-sinking POC has been recently addressed, the POC export flux is dominated by the gravitational sinking flux of large, fast-sinking POC, which is estimated by multiplying the concentration and sinking velocity. Therefore, the large differences in the POC export flux can be hide by the similar POC concentrations. Second, the lower sPOC but the similar levels of POCmeso (Figure 6) can be a result of the suboptimal parameter values, e.g. the underestimated remineralization rate. However, this cannot deny the sensitivity of POCmeso to the primary productivity. This is very likely that the POCmeso will vary a lot if the authors change the modeled primary productivity.

REPLY: We agree with the reviewer that this paragraph bears lots of assumptions. We have revised this paragraph and we have removed the conclusions about the POC export flux. As pointed out by the reviewer, we do not have sufficient information to assess the skill of the model in simulating the export of POC from sPOC and POCmeso.

References

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.

Bittig, H. C., Maurer, T. L., Plant, J. N., Wong, A. P., Schmechtig, C., Claustre, H., Trull, T. W., Udaya Bhaskar, T. V. S., Boss, E., and Dall'Olmo, G.: A BGC-Argo guide: Planning, deployment, data handling and usage, Front. Mar. Sci., 6, 502, 2019.

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.

Galí, M., Falls, M., Claustre, H., Aumont, O., and Bernardello, R.: Bridging the gaps between particulate backscattering measurements and modeled particulate organic carbon in the ocean, Biogeochemistry: Open Ocean, https://doi.org/10.5194/bg-2021-201, 2021.

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.

REVIEWER #2

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

General Comments:

This manuscript is indeed a valid compendium of diagnostics for assessing global ocean ecosystem models, which has been prepared with the aim to demonstrate the use of the multi-disciplinary dataset made available by the BGC-Argo array. The authors should thus be praised for their intention to bring together the community and follow the steps taken by Russel et al. (2018). However, that paper had different entry points, since it was specifically dedicated to a poorly sampled oceanic region and offered a multi-model analysis. This manuscript is well written and constructed, but only conveys a demonstrative message. I am thus not fully convinced by the scope of this present version of the manuscript, as well as by its effective novelty, since it does not add further knowledge to the existing literature [...]
Hence, I have carefully thought about how to write this review, and realised that the most relevant point of clarity would be to illustrate some cases of how readers could approach it. From a point of view of someone approaching modelling validation as a student or early career researcher, this manuscript offers a limited perspective, and one would gain more theoretical and methodological background in the 2009 JMS special issue (Lynch et al., 2009, and all the other papers in the issue), if not from earlier papers in the ecological modelling literature (Oreskes et a', 1994; Rykiel, 1996). If a reader is interested in the validation of the global version of PISCES, this manuscript is insufficient, because it provides a series of figures with few comments and discussions. It is surely of interest to the PISCES developers who are knowledgeable of the model details and possible deficiencies, but then an internal report would suffice. Finally, for experienced global ocean modellers, this manuscript is an illustration of the minimum

set of assessments (which I prefer to the term “validation”) that serious modellers have been doing in the last ten years when evaluating their model results. In terms of “metrics”, it gives indications to compare the model output against the state variables that can be measured by the array of floats and to add derived state variables from applications of artificial intelligence. Ultimately, the assessment is based on visual comparisons of coarsely gridded spatial maps and time series, or through the use of basic univariate scores (bias and RMSD) and cumulative diagrams that combine the same skill scores (e.g. the Taylor diagram, which also includes linear correlation).

REPLY: Thanks for the careful assessment of our work. The goal of this paper is to demonstrate the use of BGC-Argo floats for the evaluation of BGC models at the global scale, through a concise evaluation of the CMEMS global BGC forecasting system. Our hope is that the methodology employed in this study can be useful and informative for other research teams interested in model assessment with BGC-Argo floats. In particular, the main points we want to highlight are: 1) how do we handle BGC-Argo data (e.g., quality control and flags) for model assessment purposes, and 2) to propose BGC-Argo metrics, which we believe are useful to assess the accuracy of model states. We have intentionally chosen simple metrics, a minimum set of assessments and basic quantitative techniques (visual inspection, bias and RMSE) to focus the message of the study on the 2 points listed above and not on the evaluation of the model simulation. Therefore, this study is not designed as a review of biogeochemical modelling validation and it does not represent a thorough assessment of PISCES either.

We agree with the reviewer that the main message conveyed by the manuscript is not clear enough and that it can be confusing for the reader. Based on the reviewer’s comments, we have modified the manuscript so that the main message of the study appears more clearly to the reader.

First, we changed the title to “*Using BGC-Argo floats for the assessment of marine biogeochemical models: a case study with CMEMS global forecasting system.* ”

In the abstract, P1, L-28, we changed to “*Here, we demonstrate the use of the global array of BGC-Argo floats for the assessment of biogeochemical models through a concise evaluation of the CMEMS global forecasting system. We first detail the handling of the BGC-Argo data set for model assessment purposes, then we present 22 assessment metrics to quantify the consistency of BGC model simulations with respect to BGC-Argo data. The metrics evaluate either the model state accuracy or the skill of the model in capturing emergent properties, such as the Deep Chlorophyll Maximums (DCMs) or Oxygen Minimum Zones (OMZs). These metrics are associated with the air-*

sea CO_2 flux, the biological carbon pump, and the oceanic pH and oxygen levels. Moreover, we suggest four diagnostic plots for displaying such metrics."

In the introduction, the paragraph starting P. 4, L2, changed to " *We aim to demonstrate the use of the BGC-Argo global array for the assessment of BGC models at the global scale. To that end, we performed a concise evaluation of the CMEMS global BGC forecasting system using the global fleet of BGC-Argo floats. 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.* "

The BGC-Argo data are certainly invaluable, and this is the reason why the community has strived to develop the technology and the financial support to deploy them. The authors did not however succeed in showing their enhanced value for model assessment, beyond the obvious consideration that this increases the number of data, which would be much more evident if this same assessment was done by comparing datasets with and without the contribution of the BGC-Argo.

REPLY: The reviewer brings up an interesting point. It is true that BGC-Argo dramatically increases the availability of data collected by traditional oceanographic cruises. It would indeed be informative to repeat the same assessment by comparing datasets with and without the contribution of the BGC-Argo, such as for example the World Ocean Atlas. While we are very interested in this question, we do not think it belongs to this paper, whose main focus is to show the use of BGC-Argo floats for model assessment rather than showing the impact of increasing the number of observations on skill scores.

In summary I have found two major issues with this manuscript that the authors have not considered to a satisfactory extent:
The loose definition of metrics and the absence of uncertainties' treatment. The authors use the term metrics in a rather ambiguous way. They also do not differentiate between measured data and artificially generated data. This implies that the evaluation process does not necessarily lead to an improvement of the model(s).

REPLY: We agree with the reviewer that our definition of metrics was somewhat ambiguous. In the introduction, we have changed our definition of metrics based on the recent review of Hipsey et al. (2020):

"*In this study, the BGC-Argo dataset is used in conjunction with the model evaluation framework developed by Hipsey et al. (2020). In particular, they propose three levels of* *assessment metrics to evaluate the skill of a model simulation: state variables validation (e.g., Chla, nitrate, oxygen), mass fluxes and process rates validation (e.g., primary production, division rates), and emergent properties validation (e.g., deep chlorophyll maximum, oxygen minimum zones). In this study we present 22 metrics for the assessment of a model simulation with BGC-Argo data. Most of them evaluate the model state accuracy through the comparison of simulated state variables with BGC-Argo observations in the mixed layer or at fixed depth. In addition, some of the metrics assess the skill of the model in capturing emergent properties. These metrics are associated with the air-sea CO_2 flux, the biological carbon pump, the oceanic pH, and oxygen levels and Oxygen Minimum Zones (OMZs). Further, our validation framework could, in principle, include the second level of assessment metrics (i.e., flux and process). Indeed recent works demonstrated the feasibility of calculation at basin scale, from BGC-Argo observations, of mass fluxes and process rates, such as primary production, phytoplankton division and accumulation rates (Yang et al., 2021; Mignot et al., 2018), net community production (Plant et al., 2016), and carbon export (Dall'Olmo et al., 2016). However, it would be arduous to achieve such estimations on the global BGC-Argo dataset as it requires ad hoc calibration that cannot be easily defined. As a consequence, the evaluation of simulated process rates with BGC-Argo data is not addressed in this study*."

Concerning the reviewer's second comment, as we explain above, the object of the paper is not a thorough analysis of the model performance. Nevertheless, the proposed concise evaluation of the model (e.g., maps of RMSD) can be further exploited (e.g., by analyzing the spatial and temporal distribution of the RMSD maps or multivariate relationships of the errors) to investigate the model uncertainty sources.

Finally, we agree with the reviewer that we do not provide justification for mixing together measured data with artificially-generated data. We have added a paragraph in the Data section that justifies our choice.

"*Finally, we complemented the existing BGC-Argo dataset with pseudo-observations of NO_3, PO_4 , Si, and DIC concentrations as well as pH and pCO_2 using the CANYON-B neural network (Bittig et al., 2018). CANYON-B estimates vertical profiles of nutrients as well as carbonate system variables from concomitant measurements of floats pressure, temperature, salinity and O_2 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 (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."

The unconvincing enhancement of the effective role of BGC-Argo data in model assessment. Basically, the question I have is: why BGC-Argo are good enough and should be used separately and not as part of a global compilation of data such as the World Ocean Atlas? (which incidentally includes or will include the BGC-ARgo data). Since BGC-Argos are ultimately increasing the availability of data that are usually collected by means of traditional oceanographic cruises, what is indeed their value in model validation?

REPLY: We thank the reviewer for bringing this to our attention. We have carefully examined the documentation that deals with the BGC-Argo data processing in the WOD (https://www.ncei.noaa.gov/sites/default/files/2020-04/wod_intro_0.pdf) but we have not found sufficient information concerning the data mode used in the WOD. As we detail in the manuscript, the "Delayed-mode" represents the highest quality of data but for some variables, only a limited fraction of data is accessible in Delayed-Mode. Consequently, for each variable, we selected the highest quality of data (i.e., "Adjusted" or "Delayed-mode") that did not compromise too much the number of available observations. We are not sure whether such data selection is possible with the WOD, so we prefer to use the BGC-Argo data directly downloaded from Argo Coriolis Global Data Assembly Centre and not as part of a global compilation of data.

Furthermore, one of the issues of large databases such as WOD, is the interoperability of the data that compose them, and which, ultimately, affect their overall accuracy. Using the BGC-Argo dataset separately is a way to ensure consistent accuracy. The GLODAP V2 data set (on which CANYON B is developed) is an illustration of an interoperable homogenous data set (with very strict data QC procedure) used for model assessment and not used as part of a global compilation of data.

Finally, concerning the last of the above reviewer's questions, the BGC-Argo floats provide observations at high vertical and temporal resolutions and for long periods of time providing nearly continuous time series of the vertical distribution of a number of biogeochemical variables. This is not possible with discrete vertical samplings provided by cruise cast *in situ* measurements.

We have commented on the two last points discussed above in the 5th paragraph of revised Introduction.

For clarity, I would like to elaborate more on the first concept above, while the second point is mostly derived from the specific comments detailed in the next section. Russel et al (2018) also use the concept of metrics in a wider sense, although they define metrics as “any quantity or quantifiable pattern that summarizes a particular process or the response in a model to known forcings”. The strength of the ACC transport at Drake Passage or the latitude of the maximum zonal mean winds over the Southern Ocean are “metrics” in this context. They are combinations of state variables, or values of state variables at specific locations.
In this context, all the surface state variables listed in Table 2, are indeed components of the biological carbon pump, but they are not metrics. They are simply state variables. Only when considered together to evidence emergent patterns they may give indications of proper process functionality (e.g. the ratio of particulate organic carbon to total chlorophyll, de Mora et al, 2016). I agree that the DCM and the “nutricline” (which would deserve a more appropriate definition, see specific points below) are “metrics”, as well as the depth of the hypoxic layer. Mixing together indicators of processes with state variables is confusing, unless a rigorous link between a single state variable and the process is established.

REPLY: As we explained above, we have changed our definition of metrics, and in the new version of the manuscript we use the framework proposed by Hipsey et al. (2020). They propose three levels of assessment metrics to evaluate the skill of a model simulation: state variables validation (e.g., chlorophyll, nitrate, oxygen), mass fluxes and process rates validation (e.g., primary production, division rates), and emergent properties validation (e.g., deep chlorophyll maximum, or oxygen minimum zones). We have inserted a new column in Table 2 to inform about the level each proposed metric is referring to. In Section 3, we have made a rigorous link between the state variable and the associate process in the section that defines the assessment metrics.

This manuscript increases the risk of misinterpretation by mixing together “metrics” and skill scores. Neither Russel et al (2018) and this manuscript expand on the concept of metrics performance and objective assessment (performance indicators, skill scores, cost functions, are all synonyms that depend on the specific discipline), which was instead done by Allen et al. (2007), Friedrichs et al. (2009), Vichi and Masina (2009) and others in the JMS special issue. For ease of simplicity, I will use the term skill score, which is the one used in the more mature field of weather forecasting. State variables can be assessed using univariate skill scores, and this is a necessary exercise for any modeller to ensure the model has some grip with reality. Figure 3 and the other density plots in the Appendix give a visual indication of the skill score, but they do not quantify it (e.g. Smith and Rose, 1995; Rose and Smith, 1998). I also have another question linked to my Point 2 (and further detailed in the specific comments): why should this exercise be done only with the BGC-Argo and not also including the other existing data? Since BGC-Argo are evaluated against cruise cast benchmarks, then those data are usually considered always superior, and should be used. Again, the real value of the BGC-Argo would have been shown if the score had been substantially modified with the inclusion of the Argo data.

REPLY: In the revised manuscript, we have inserted Table 3 that quantifies the skill scores for each metrics as done in Vichi and Masina (2009) or Doney et al. (2009). As we explained above, we believe it is more reasonable to use the BGC-Argo data as a separate dataset rather than as part of a global compilation of data.

Specific comments:

P2L1 - Earlier work has specifically addressed the impact of assimilation on the carbonate system (Visinelli et al., 2017)

REPLY: Thanks for suggesting this reference. This study showed that the assimilation of physical data improves the simulation of alkalinity, DIC and pCO_2. However, a number of recent studies have shown that the assimilation of physical observations tends to degrade the simulation of BGC state (Fennel et al., 2019; Park et al., 2018; Gasparin et al., 2021).

P2L26-29 - This sentence is mixing together sensor accuracy, which has been assessed by Johnson et al and Mignot et al, in two specific regions of the world ocean) and temporal/vertical resolutions, which have not been assessed as far as I am aware. This is misleading. 10 days may not be sufficient for all variables, as well as the vertical binning that is done. The comparisons have assessed the equivalence between rosette casts and the floats, but they say nothing about the temporal and vertical resolution. For certain processes, such as carbon exchange and phytoplankton biomass through chlorophyll and backscattering proxies, a resolution of 10 days would lead to sampling aliases either of the mean or of the variability (Monteiro et al., 2015, Little et al., 2018). These are examples from the Southern Ocean, where there is the highest density of buoys.

REPLY: We have revised the sentence removing the part about the temporal and vertical resolutions.

P2L32-34 - The authors should be more specific. Other datasets, such as for instance remote sensing, are less limited in terms of temporal and spatial resolutions. This is connected to the concerns expressed in Point 1 above.

REPLY: We have revised the sentence, being more specific about the BGC-Argo resolutions (P3L33-P4L1).

P4L3-5 This sentence seems to imply that one can only perform point-by-point comparisons when there are few floats, which is odd. Again linked to my main Point 1 above. The authors should explain why given the current computing capability, they only suggest to perform diagnostics for few selected tracks and not for the overall dataset (Section 5.d).

REPLY: We have changed this paragraph based on point 1 and point 2 (see above), consequently this sentence was removed in the revised version of the manuscript.

P4L12-16 The connection between the variables and the ocean health/ecosystem functioning is not made explicit in the text. Taking as an example the ocean health index (http://www.oceanhealthindex.org/), establishing ocean health is obtained as a multivariate analysis of several data layers, forming a selected set of drivers and their associated thresholds. The authors should be more explicit about their intent here.

REPLY: Since we have changed our definition of metrics, we no longer refer to ocean health and ecosystem functioning in the revised version of the manuscript.

P5L12-13 This is not an objective criterion. What is an acceptable level of compromise?

REPLY: We have added an objective criterion in the revised manuscript:

" However, for some variables, only a limited fraction of data is accessible in "Delayed-Mode". Consequently, for each variable, we selected the highest data modes, where at least 80 % of the data are available (see Table 1). Note that this criterion does not apply to O_2, where only delayed mode data were selected in order to generate the pseudo-observations from CANYON-B neural network (see after). "

P5L22 There are many other relationships, and they have been shown to give different results (e.g. Thomalla et a., 2017l). The authors should explain why they are recommending this one.

REPLY: In the revised version of the manuscript, we now use a POC vs b_{bp} relationship developed for the global ocean (https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-MOB-QUID-015-010.pdf) based on a global database of in situ POC and satellite b_{bp} (Evers-King et al., 2017). This relationship, developed for global application, has been shown to outperform regional relationships, such as Cetinic et al. (2012), at global scales (P6L9-18).

P6L12-15 It appears that this method of linear resampling would artificially increase the number of data, and hence bias the statistical results, especially in conditions where there are not enough data.

REPLY: We thank the reviewer for raising this interesting issue. In the revised manuscript, we have commented on the possible bias introduced by the linear resampling method on our statistical results (P7L20-22).

P7L10-12 The authors do not discuss what would happen if the MLD is different between the observations and the model.

REPLY: The dynamical component, used in this study, has been extensively validated (Lellouche et al., 2013, 2018), and demonstrate to correctly represent variables that are constrained by observations (e. g., temperature and salinity), including Argo profiles. We verified that the MLD, which is calculated on a density criterion basis, is indeed correctly represented in the model. The global bias between the model and the BGC-Argo observations is 0.3 m. In the revised manuscript, we added a sentence that specifies that we verified the model skill in simulating the MLD (P8L22).

P7L29-30 Related to my point 1 above. The relationship between the state variables and the ecosystem functions is not made explicit. The term "useful" should be motivated.

REPLY: This section has been revised making the relationship between the state variables and ecosystem function more explicit. In addition, we have added new metrics for the mesopelagic layer as explained below.

"The biological carbon pump is the transformation of nutrients and dissolved inorganic carbon into organic carbon in the upper part of the ocean through phytoplankton photosynthesis and the subsequent transfer of this organic material into the deep ocean. The functioning of this pump relies on key pools of nutrients and carbon as well as a number of processes that control mass fluxes between the pools.

The first level of assessment of a biological carbon pump simulated by a model consists in evaluating the different pools (or state variables) of the pump (Hipsey et al. 2020). In particular, the comparison of simulated surface nutrients (NO_3, PO_4, and Si), DIC, Chla and POC with BGC-Argo observations gives an indirect evaluation of the model capability to capture key processes of the biological carbon pump in the ocean upper layer, such as primary production, respiration, and grazing. A second-level, , assessment would be to directly compare these key processes with measured mass fluxes, but this assessment level is not addressed in this study. The surface nutrients , DIC, Chla and POC (hereinafter denoted sNO_3, sPO_4, sSi, sDIC, sChl and sPOC) are calculated as the average concentrations in the mixed layer.

Similarly, the assessment of the mesopelagic nutrients, DIC and POC concentration (hereinafter indicated with the subscript $_{meso}$) provides an indirect evaluation of the key mesopelagic layer processes , such as export production, respiration, etc. The mesopelagic concentrations correspond to the depth-averaged concentrations between the base of the mixed layer down to 1000 m. "

P8L7-8 Same as above, the value of DCM as an indicator should be contextualized. Why are BGC-Argo data providing a better estimate of this metric than other data?

REPLY: In the revised manuscript the use of the DCM as an indicator is better contextualized: "*In stratified systems, a Chla maximum (hereinafter denoted Deep Chlorophyll Maximum, DCM) is formed at the base of the euphotic layer (Barbieux et al., 2019; Cullen, 2015; Letelier et al., 2004; Mignot et al., 2011, 2014). It has been suggested that the DCM plays an important role in the synthesis of organic carbon by phytoplankton (Macías et al., 2014). DCMs are therefore important features to be assessed in BGC models with respect to processes involved in the biological carbon pump processes such as the primary production, however the DCM layer generally escapes detection by remote sensing. Furthermore, DCMs are also an emergent feature that develops in response to complex physical and biogeochemical interactions (Cullen, 2015). Thus, their evaluation provides critical information regarding the accuracy of the model in capturing complex patterns of key ecosystem processes.*"

As we explain above, the BGC-Argo data provide consistent profiles at high vertical and temporal resolution allowing to derive time-series of DCM depths. In comparison, discrete vertical samplings provided by cruise cast *in situ* measurements have a vertical resolution much lower (10 samples taken over a 100 m layer), without repetitive sampling.

P8L13 Please explain what H is.

REPLY: It is an omission on our part. H is the mixed layer depth. We have replaced H by MLD.

P8L14-16 This may be confusing for some readers, since it's not technically a gradient. The cited paper uses and justifies this definition. I'd suggest the authors to be more precise and give their definition and how this is an effective metric of the carbon pump. Also, there is a difference in sampling between argo and the layers of discrete models. How is this taken into account?

REPLY: We have provided a more precise definition of the nitracline depth in the revised manuscript, and we described how this is an effective metric of the carbon pump:

"*The vertical supply of NO_3 to the surface layers is a critical process of the biological carbon pump as NO_3 is often depleted in the surface layers and is a limiting factor for phytoplankton growth in most oceanic regions. The NO_3 vertical supply depends, among other factors, on the vertical gradient of NO_3 (the nitracline), and, in particular, on its depth (the nitracline depth)* (Cermeno et al., 2008; Omand and Mahadevan, 2015)*. Therefore, the comparison of the simulated nitracline depth with BGC-Argo observations allows for an indirect assessment of the model quality in reproducing vertical fluxes of NO_3. Following previous studies* (Cermeno et al., 2008; Lavigne et al., 2013; Richardson and Bendtsen, 2019)*, the depth of the nitracline corresponds to the first depth where NO_3 is detected. The detection threshold was set to 1 µmol/kg, which corresponds to an upper estimate of BGC-Argo NO_3 data accuracy* (Johnson et al., 2017; Mignot et al., 2019). "

Finally, there is indeed a difference in sampling between the BGC-Argo and the layers of discrete models. This is clearly visible in the scatterplot for the nitracline, the DCM and the OMZ depths.

P8l28-30 At P4L11 it is reported "depth of the OMZ". This the depth of the oxygen minimum. It should be explained how and why this is a good indicator, and why the BGCArgo data are superior in its identification.

REPLY: In the revised version of the manuscript, we explain why the depth of the oxygen minimum is a good indicator. "*Oxygens levels in the global and coastal waters have declined over the whole water column over the past decades (Schmidtko et al., 2017) and OMZs are expanding (Stramma et al., 2008). Assessing how models* *correctly represent ocean oxygen levels as well as the OMZs is therefore critical to monitor their changes over time. Similarly to DCMs, the assessment of OMZs is also informative on how the model simulates emergent dynamics as OMZs originate from complex physical and biogeochemical interactions (Paulmier and Ruiz-Pino, 2009).* ""

As we detailed in a previous reply, BGC-Argo floats are particularly fit in the identification of vertical characteristics of BGC variables.

P9L26 This statement about non-linearity is odd in the context of model goodness-of-fit (Smith and Rose, 1995; Pineiro et al, 2008; Vichi and Masina, 2009). If it's non-linear, then the assessment is failed.

REPLY: We have removed the sentence.

P10-8-12 The choice of the binning interval should be discussed. What is the advantage of losing the variability measured by the floats? Why not using the standard deviation as an indicator of the model skill to reproduce the proper scales? These are enhanced features that only the BGC-Argo data would allow to compute.

REPLY: We discuss the choice of the binning interval in the revised version of the manuscript. "...*To do so, the metrics from 2009 to 2017 are averaged in 4°x4° bins, excluding those with less than 4 points. The 4° distance is an upper estimate of the autocorrelation length scales for O_2, nutrients, and pCO_2 (comprised between 300 and 400 km) between 20° and 40° of latitude in both hemispheres* (Biogeochemical-Argo Planning Group, 2016)***.***"

Moreover, in section 4.c we have commented about using standard deviation maps as an indicator of the model skill in properly reproducing variability scales. However, we won't show standard deviation in the manuscript as we prefer to not overload Figure 4 and the associated supplementary figures with additional panels.

P10L22-24 Allen et al (2007) warned against the visual comparison of time series. This sentence is generic and should be explained in the context of the augmented data provided by the BGC-Argo.

REPLY: We agree with the reviewer that visual inspection relies on the subjective appreciation of the evaluator. Consequently, we have added normalized skill scores to Figures 5 and 6. Moreover, we have added the following sentence at the end of the section 4c. "*In addition to the time series of metrics, we also display normalized skill* *scores such as percent BIAS and RMSD as a function of season in order to combine quantitative metrics with visual comparison*."

P11L11-14 The results are not presented according to the concept of the biological carbon pump "metric". It is evident that the nutrients are correlated while all carbon flux variables are not performing. Which ultimately questions the use of surface nutrients as indicators of carbon cycling.

REPLY: The fact that nutrients are well represented in the model suggests that the model captures the combination of process rates that drive nutrients dynamics. Some of these process rates drive both the nutrients and carbon dynamics, but there are also rates that are specific to each state variable. This probably explains why the carbon variables are not performing while the nutrients are well simulated. However, it must be recognized that without a direct assessment of the individual rates, we cannot verify this hypothesis. We have clarified this point in the revised version of the manuscript in Section 5a (P13L27-32-P14L1-14).

P11L31 I cannot see the data "around" the line. I rather see an overestimation. (it is either Cape Verde or Cap Vert)

REPLY: We have improved the clarity of the figure in the revised version of the manuscript.

P12-L2-17 Linked to Point 2 above. The authors seem to imply that BGC-Argo data are more suitable than ocean colour for model assessment. I acknowledge that this is not explicitly written, but there is no clear rationale. This kind of map would certainly be superior in terms of spatial and temporal resolution when using that product as Benchmark.

REPLY: Indeed, such a map would be superior in term of spatial and temporal resolution. However, the BGC-Argo data allows to assess the skill of the model in estimating Chla concentration, when ocean color data are not available, i.e., during cloudy days and during winters at high latitudes.

P12-section-d This is the section that mostly led to the inclusion of Point 2 above. The shown time series is 2 years long, which is an invaluable source of data from a region that has been influential in shaping our understanding of the spring bloom. I am missing the point why the authors are writing the term spring bloom in quotes. The advantage of time series from floats that remained in a given province of the global ocean is of huge potential in model validation. The offered description is quite generic, which could have been done even using monthly climatological time series obtained from the WOA, or from the existing long-term observational ocean sites (BATS, PAPA, HOT). The BGC-Argo floats are an unprecedented source of multiple opportunities to do validation in several regions of the world ocean (with some limitations), but this present form of the manuscript does not offer any specific recommendation of what numerical modellers should do to unleash this potential. I would be very interested in seeing an exploitation of the multivariate nature of BGC-Argo, while I only see multi-panel plots.

REPLY: Based on this comment, we have revised this section. We have removed the unnecessary description of the spring bloom, while highlighting the invaluable opportunities of such time series for the assessment of models by showing other time series in an oligotrophic region where in situ data are scarce. Concerning the evaluation of the multivariate nature of BGC-Argo, we agree that it is an interesting point to pursue. We are very interested in applying the multivariate approach proposed by Allen et al. (2007) to the BGC-Argo data set. However, we prefer to focus this manuscript on the presentation of the metrics and to exploit the multivariate approach in another study.

P13L4-5 The authors should do more than simply say “correctly represented”. This is a subjective statement, which is based on a visual comparison, exactly what the community challenged in the last 10-15 years. The advantage is that now we can use a frequency of 10 days, when initially phenology analysis was based on monthly data. Again, the authors are missing an opportunity to demonstrate the intrinsic value of this new data set.

REPLY: As suggested by the reviewer, we have included normalized skill scores to avoid relying only on subjective visual inspection. We agree that the frequency of 10 days is a significant progress over previous data sets. However, as explained in the conclusion, we do not address phenology metrics in this study because the number of observations per month and per bins is still too low to perform a global analysis.

P13-L13-20 This is a more detailed analysis of this specific model, which indeed brings in some of the advantages of a multivariate data set. However, there is a combination of measured and derived variables, which are treated as if they were equivalent. Quite a few questions come to mind: Is there a possibility that there is artificial correlation in the derivation of the phosphate and silicate concentration? What is the error associated with the CANYON-B method? Which is the effective (measured) variable mostly responsible for the response of the other estimated nutrients? The reduced consumption occurs during the spring period, and is continued during summertime. Hence, there is a factor at play during the late spring period, which is less likely to be reduced uptake from smaller phytoplankton during summer as suggested. It may thus be a delayed onset of the phytoplankton succession, or maybe a faster remineralization occurring in the upper layers, which retain more inorganic nutrients closer to the surface. This may indeed be beyond the scope of the manuscript, but it has been the authors' decision to propose some mechanistic explanations of this discrepancy. Showing a complete example of how the use of multivariate data allows modellers to investigate model deficiencies would offer guidelines to other modellers.

REPLY: As explained above, we have included a paragraph in the Data section that discusses the error associated with the CANYON-B method. Concerning the second comment, we have removed the mechanistic explanation of this discrepancy. As suggested by the reviewer, we agree that this is beyond the scope of the manuscript.

P13-L22-23 This sentence bears lots of assumptions. This is really where BGC-Argo can make a difference. The related uncertainties should however be highlighted, together with recommendations to other modellers on how to best approach the assessment of the carbon cycle metrics.

REPLY: Please, see next REPLY.

P13L26-29 This argument is flawed. If the occurrence of the peak is matched in the mesopelagic layer rather than at the surface, it is a clear indication of vertical mismatches in the export. I would thus argue that POC concentration is a proper metric for the export component of the carbon cycle. I would again encourage the authors to replace the use of subjective terms such as "consistent" with objective indicators (see Allen et al., 2007). For instance the comparison of the skill score computed in two consecutive years would give indication if there is some variability or if the model tends to repeat the same pattern.

REPLY: We have removed the conclusions about the oceanic carbon cycle and POC export flux, and we have removed the time series of SDIC, sPOC and POC_{meso} in Figure 5. As pointed out by the reviewers #1 and #2, this paragraph bears a lot of assumption and we don't have sufficient information to assess the model skill in simulating the process rates that drive sPOC, and POC_{meso}.

P14L16-19 I would recommend more clarity on this statement. Are these sensors not available on the global ocean floats? It is not clear why this example is presented for Mediterranean floats, and not introduced earlier as one major advantage of the BGC-Argo floats.

REPLY: We have clarified this statement adding that the sensors are available on the global ocean. However, the global model used in the study does not resolve the spectral and directional properties of the underwater light field. That's why we didn't use the global model but a model of the Mediterranean Sea equipped with a multispectral light module, as clarified in the new manuscript version.

P14L26-28 This sentence is similar to the statements done in the earlier sections. This is not technically a perspective statement.

REPLY: We agree with the reviewer that this section does not provide a perspective statement, thus we have added a perspective statement at the end of this section(P17L1-2).

P15L1-6 The question is whether these data should be used "on their own" or in conjunction with the other existing datasets. The authors should clearly explain in the conclusion why this dataset should be exploited as a separate unit.

REPLY: As explained above, in the introduction we have added motivation about using this dataset as a separate unit.

P15L32-P16L3 I would thus recommend the authors to thoroughly address the issue of how the uncertainties should be treated. This is particularly important in the case of mixing measured and derived variables. If BGC-Argo are capable, within their limits, to reduce uncertainties in model assessment exercise, this should be adequately argumented. The fact that there are more data available is undoubtedly of relevance, but I wonder if it does help to reduce uncertainties in model states.

REPLY: As explained previously, we have added a paragraph in the Data Section that provides justification for mixing together measured data with artificially-generated data. We also removed the paragraph about fluorescence quenching as it can be misleading for the reader. As discussed above, we have verified that the RMS difference between model and BGC-Argo Chla is always lower than the BGC-Argo Chla RMS error, so that the comparison of simulated Chla with the BGC-Argo Chla leads to an evaluation of the skill of the model in simulating Chla concentrations.

P16L15-18 Please highlight in which part of the results this is shown.

REPLY: We have highlighted in which part of the results this is shown (P18L21).

P17L2 Please add in the caption the meaning of the codes (or a link to where they are explained more in detail). Also, in the heading of the 3rd column, correct Date with Data. Figure 2 Taylor diagrams are based on geometric properties of the circle. Hence they should be presented using equal axes.

REPLY: In the revised manuscript, we have added the meaning of the codes, changed Date with Data and presented the Taylor diagram using equal axes.

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.

Barbieux, M., Uitz, J., Gentili, B., Pasqueron de Fommervault, O., Mignot, A., Poteau, A., Schmechtig, C., Taillandier, V., Leymarie, E., Penkerc'h, C., D'Ortenzio, F., Claustre, H., and Bricaud, A.: Bio-optical characterization of subsurface chlorophyll maxima in the Mediterranean Sea from a Biogeochemical-Argo float database, Biogeosciences, 16, 1321–1342, https://doi.org/10.5194/bg-16-1321-2019, 2019.

Biogeochemical-Argo Planning Group: The scientific rationale, design and implementation plan for a Biogeochemical-Argo float array, https://doi.org/10.13155/46601, 2016.

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.

Cermeno, P., Dutkiewicz, S., Harris, R. P., Follows, M., Schofield, O., and Falkowski, P. G.: The role of nutricline depth in regulating the ocean carbon cycle, Proc. Natl. Acad. Sci., 105, 20344–20349, https://doi.org/10.1073/pnas.0811302106, 2008.

Cetinic, I., Perry, M. J., Briggs, N. T., Kallin, E., D'Asaro, E. A., and Lee, C. M.: Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment, J. Geophys. Res.-Oceans, 117, https://doi.org/10.1029/2011JC007771, 2012.

Cullen, J. J.: Subsurface Chlorophyll Maximum Layers: Enduring Enigma or Mystery Solved?, Annu. Rev. Mar. Sci., 7, 207–239, https://doi.org/10.1146/annurev-marine-010213-135111, 2015.

Dall'Olmo, G., Dingle, J., Polimene, L., Brewin, R. J. W., and Claustre, H.: Substantial energy input to the mesopelagic ecosystem from the seasonal mixed-layer pump, Nat. Geosci., 9, 820–823, https://doi.org/10.1038/ngeo2818, 2016.

Doney, S. C., Lima, I., Moore, J. K., Lindsay, K., Behrenfeld, M. J., Westberry, T. K., Mahowald, N., Glover, D. M., and Takahashi, T.: Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data, J. Mar. Syst., 76, 95–112, https://doi.org/10.1016/j.jmarsys.2008.05.015, 2009.

Evers-King, H., Martinez-Vicente, V., Brewin, R. J. W., Dall'Olmo, G., Hickman, A. E., Jackson, T., Kostadinov, T. S., Krasemann, H., Loisel, H., Röttgers, R., Roy, S., Stramski, D., Thomalla, S., Platt, T., and Sathyendranath, S.: Validation and Intercomparison of Ocean Color Algorithms for Estimating Particulate Organic Carbon in the Oceans, Front. Mar. Sci., 4, 251, https://doi.org/10.3389/fmars.2017.00251, 2017.

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.
Lavigne, H., D'Ortenzio, F., Migon, C., Claustre, H., Testor, P., d'Alcalà, M. R., Lavezza, R., Houpert, L., and Prieur, L.: Enhancing the comprehension of mixed layer depth control on the Mediterranean phytoplankton phenology: Mediterranean Phytoplankton Phenology, J. Geophys. Res. Oceans, 118, 3416–3430, https://doi.org/10.1002/jgrc.20251, 2013.
Lellouche, Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C.-E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., and Le Traon, P.-Y.: Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1⁄12° high-resolution system, Ocean Sci., 14, 1093–1126, https://doi.org/10.5194/os-14-1093-2018, 2018.
Lellouche, J.-M., Le Galloudec, O., Drévillon, M., Régnier, C., Greiner, E., Garric, G., Ferry, N., Desportes, C., Testut, C.-E., Bricaud, C., Bourdallé-Badie, R., Tranchant, B., Benkiran, M., Drillet, Y., Daudin, A., and De Nicola, C.: Evaluation of global monitoring and forecasting systems at Mercator Océan, Ocean Sci., 9, 57–81, https://doi.org/10.5194/os-9-57-2013, 2013.
Letelier, R. M., Karl, D. M., Abbott, M. R., and Bidigare, R. R.: Light driven seasonal patterns of chlorophyll and nitrate in the lower euphotic zone of the North Pacific Subtropical Gyre, Limnol. Oceanogr., 49, 508–519, 2004.
Macías, D., Stips, A., and Garcia-Gorriz, E.: The relevance of deep chlorophyll maximum in the open Mediterranean Sea evaluated through 3D hydrodynamic-biogeochemical coupled simulations, Ecol. Model., 281, 26–37, 2014.
Mignot, Claustre, H., Uitz, J., Poteau, A., D'Ortenzio, F., and Xing, X.: Understanding the seasonal dynamics of phytoplankton biomass and the deep chlorophyll maximum in oligotrophic environments: A Bio-Argo float investigation, Glob. Biogeochem. Cycles, 28, 856–876, https://doi.org/10.1002/2013GB004781, 2014.
Mignot, Ferrari, R., and Claustre, H.: Floats with bio-optical sensors reveal what processes trigger the North Atlantic bloom, Nat. Commun., 9, https://doi.org/10.1038/s41467-017-02143-6, 2018.
Mignot, A., Claustre, H., D'Ortenzio, F., Xing, X., Poteau, A., and Ras, J.: From the shape of the vertical profile of in vivo fluorescence to Chlorophyll-a concentration, Biogeosciences, 8, 2391–2406, https://doi.org/10.5194/bg-8-2391-2011, 2011.
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.
Omand, M. M. and Mahadevan, A.: The shape of the oceanic nitracline, Biogeosciences, 12, 3273–3287, https://doi.org/10.5194/bg-12-3273-2015, 2015.
Paulmier, A. and Ruiz-Pino, D.: Oxygen minimum zones (OMZs) in the modern ocean, Prog. Oceanogr., 80, 113–128, 2009.
Plant, J. N., Johnson, K. S., Sakamoto, C. M., Jannasch, H. W., Coletti, L. J., Riser, S. C., and Swift, D. D.: Net community production at Ocean Station Papa observed with nitrate and oxygen sensors on profiling floats, Glob. Biogeochem. Cycles, 30, 859–879, https://doi.org/10.1002/2015GB005349, 2016.
Richardson, K. and Bendtsen, J.: Vertical distribution of phytoplankton and primary production in relation to nutricline depth in the open ocean, Mar. Ecol. Prog. Ser., 620, 33–46, https://doi.org/10.3354/meps12960, 2019.
Schmidtko, S., Stramma, L., and Visbeck, M.: Decline in global oceanic oxygen content during the past five decades, Nature, 542, 335–339, https://doi.org/10.1038/nature21399, 2017.
Stramma, L., Johnson, G. C., Sprintall, J., and Mohrholz, V.: Expanding Oxygen-Minimum

Zones in the Tropical Oceans, Science, 320, 655–658, https://doi.org/10.1126/science.1153847, 2008.
Vichi, M. and Masina, S.: Skill assessment of the PELAGOS global ocean biogeochemistry model over the period 1980–2000, Biogeosciences, 6, 2333–2353, https://doi.org/10.5194/bg-6-2333-2009, 2009.
Yang, B., Fox, J., Behrenfeld, M. J., Boss, E. S., Haëntjens, N., Halsey, K. H., Emerson, S. R., and Doney, S. C.: In Situ Estimates of Net Primary Production in the Western North Atlantic With Argo Profiling Floats, J. Geophys. Res. Biogeosciences, 126, https://doi.org/10.1029/2020JG006116, 2021.