Response to Referee #1

General comments:
This study tackles an important question regarding the potential benefit of assimilating BGC-Argo profiles in improving the global ocean biogeochemical reanalyses. The manuscript consists of 3 sections: 1) establish an OSSEs framework; 2) assess different strategies of updating BGC model state variables when assimilating ocean color data; 3) evaluate the benefit of assimilating different numbers of BGC-Argo profiles in addition to ocean color data over the assimilation of profiles or ocean color data alone. Overall, I think this work is well-conceived and will make an important contribution to the state-of-art ocean biogeochemical data assimilation combining the routinely available ocean color data and the emerging BGC-Argo observations.

Thank you for your positive assessment and constructive comments. I will address each of these in turn below.

I have two major comments. First, I feel that the second section of the manuscript on different DA strategies, at its present form, does not add much value to the story and the selection of best DA strategy involving a nitrogen balancing scheme before thorough tuning doesn't seem fair. Second, the first section on OSSEs requires a bit more analysis to prove its credibility. I'll provide more detailed explanations below. Aside from that, I have some minor comments, mostly technical, for the author to consider.

Thank you for these comments. On reflection, I agree that the section on ocean colour assimilation strategies does not add much value to the core aims of the manuscript, and is not fully fledged. I therefore propose to remove this section, just presenting the results of the OC_3D_PHY run alongside the BGC-Argo assimilation. I will potentially develop the ocean colour assimilation work further in a future publication.

I also agree that presenting further analysis of the OSSE framework would help demonstrate its credibility, and propose to include the types of assessment requested below.

Upon appropriately addressing these comments, I'll recommend publication of the manuscript in Biogeosciences.

1. I question on the value of including section 2 on comparing different update strategies when assimilating surface chl data for following reasons:

1) While I acknowledge the efforts and time needed for comparing 6 different DA strategies, I feel that the present comparison is not sufficient for fairly selecting the best DA strategy. I would argue that the more sophisticated nitrogen balancing scheme failing to outperform other strategies is largely because the parameter values used in the scheme are directly adopted from Hemmings et al. (2008) without proper tuning. These parameters reflect the BGC model's inherent relationships between chl and other model state variables. Since the model used here (MEDUSA) has quite different structure from that of Hemmings (HadoCC), a careful calibration of the parameters in the N balancing scheme is needed before its usage. That maybe contribute to a separate manuscript focusing on the benefit of multivariate BGC update over single-variable update.

I agree with this assessment. Recalibrating the parameterisations of the Hemmings et al. (2008) nitrogen balancing scheme for specific use with MEDUSA would be a considerable amount of work, but necessary to achieve the best results. As suggested, I will therefore remove this section from the current manuscript, and potentially develop the work further as part of a separate paper.

2) At its present form, I didn't see strong connection between section two and three in the manuscript. To me, the most significant findings are from section three and this section stands out even if section two is completely removed. This is because the comparison in section two didn't suggest a clear winner and the ultimate decision of using the DA strategy of an intermediate complexity rather than the most sophisticated one (the N balancing scheme) for section three further reduce the value of including the entire comparison in section two.

I agree that the most significant findings, in relation to the core aims of the study, are in the final section, and that this section deserves the most attention.

3) If section two was removed, the author can have more space to elaborate and focus on the impacts of assimilating BGC-Argo profiles on different variables and suggesting directions for future work to improve. Currently, I feel the discussion on this part is relatively short compared to the emphasis it receives in the title, abstract and Introduction.

I agree. I therefore propose to remove the ocean colour section, and use the extra space to further develop the assessment and discussion surrounding BGC-Argo assimilation, as suggested.

2. I think section one on establishing the OSSEs framework is key to the credibility of assessment on assimilation impact. Presently the only analysis provided to show the credibility of OSSEs is a comparison of the errors between FREE and OBS and between FREE and NATURE for surface chl, NO3 and pCO2 in Figure 2. According to the criteria of designing rigorous ocean OSSE system detailed in Halliwell et al. (2014), I would request the author to comment and/or provide some information on following aspects:

1) Can the NATURE run reasonably capture the key features measured by the observing systems (in this case the surface chla, and the BGC profiles)? The author refers the performance of NATURE run to references given in Section 2 which is not very clear to me which one exactly has the same configuration and time period as the one here. A brief summary and/or some figures on the performance of the NATURE run will help.

The references in Section 2 detail assessment of the model components used, but not of the specific NATURE run which is newly presented in this manuscript. I will be clearer about this in the text. I will also present some figures and assessment in the revised manuscript demonstrating the performance of the NATURE run as requested. I can confirm that it is able to reasonably capture key features, but agree that this isn't currently shown in the manuscript and should be.

2) Figure 2 only presents the surface comparisons. Since the assessment involves the vertical profiles, can the author also comment on whether the errors between FREE and NATURE are comparable to those between FREE and OBS in terms of the vertical distribution pattern of observable BGC variables.

I will present further assessment in the revised manuscript comparing the errors between FREE and NATURE and FREE and OBS in terms of vertical distribution for biogeochemical variables for which there are appropriate observation products for the comparison.

3) How about the error growth rate? One important criterion of credible OSSE evaluation is that the differences between the FREE and the NATURE ("truth") grow at the same rate as errors that develop between the state-of-the-art ocean models and the true ocean (Halliwell et al. 2014).

I will also present further assessment of the error growth rate, I agree this is important to demonstrate.

Halliwell, G. R., Srinivasan, A., Kourafalou, V., Yang, H., Willey, D., Le Hénaff, M., and Atlas, R.: Rigorous evaluation of a fraternal twin ocean OSSE system for the open Gulf of Mexico, J. Atmos. Ocean Tech., 31, 105–130, https://doi.org/10.1175/JTECH-D-13-00011.1, 2014.

Specific comments:

L21: '… half the planet's primary production.' Reference?

I will add a reference to Field et al. (1998, http://doi.org/10.1126/science.281.5374.237).

L75-77: Could you briefly add the outcome of assimilating these BGC-Argo observations in these two studies?

I will add details of these. Verdy and Mazloff (2017) produced a five-year state estimate of the Southern Ocean using an adjoint method, and were able to capture over 60% of the variance in oxygen profiles at 200 m and 1000 m depth. Cossarini et al. (2019) assimilated chl-a into a model of the Mediterranean Sea, and found this was successful in adjusting the shape of chlorophyll profiles, and that with the present number of BGC-Argo floats they could constrain phytoplankton dynamics in up to 10% of the Mediterranean Sea.

L87-88: 'Two groups perform … and the Met Office … presented here.' this information may be meaningful for the groups involved but doesn't seem informative for general readers. Do the two groups aim at different perspectives of the BGC assessment? What are they then?

I will rephrase this section to be less focussed on the details of the project, and instead give a brief summary of the results of Germineaud et al. (2019), who presented a probabilistic evaluation at a single assimilation time step, finding that chl-a from BGC-Argo floats added value at surface locations where ocean colour was unavailable, and at depth.

L116-118: Isn't that oxygen and dissolved inorganic N are also simulated? Or are they implicitly included in the 'coupled carbon cycle'? It's not clear to me what 'coupled' means here.

This sentence was poorly phrased and incomplete. Oxygen and DIN are indeed also simulated, as are iron, silicate, and fast- and slow-sinking detritus. I will rephrase the sentence to give a more complete description of the model variables. The word "coupled" is unnecessary in this context, and I will remove it. The model DIC and alkalinity are the state variables representing the carbon cycle.

L154: This is acceptable, but could you comment if the physics of the NATURE run without DA is reliable to conduct the OSSEs?

The NATURE run is able to capture key features of the physics, I will present assessment of this alongside the similar assessment of biogeochemical variables discussed above.

L156: Just curious if there is any particular reason for using log10 instead of log-normal transformation.

In practice, it should make no difference to the assimilation whether log-normal or $\log_{10}$ is used. The shape of the distribution is the same (the ratio of log(x) to $\log_{10}$(x) is identical for all values of x), except that $\log_{10}$ gives a smaller variance. It is the shape of the distribution that matters for the assimilation, so as long as the same transformation is applied consistently to both model and observations, it should not matter whether log-normal or $\log_{10}$ is used. In the literature, some studies use $\log_{10}$ (e.g. Gregg, 2008, https://doi.org/10.1016/j.jmarsys.2006.02.015), while others use log-normal (e.g. Ciavatta et al., 2011, https://doi.org/10.1029/2011JC007219). The decision to use $\log_{10}$ here is a historical one, following Ford et al. (2012, https://doi.org/10.5194/os-8-751-2012), but as stated the choice should make no difference.

L160: How large is the correlation length-scale? Water et al. (2015) is on physical DA. Same length scale used for the BGC assimilation here? I'm thinking that BGC fields are more dynamic and thus have a smaller correlation length-scale.

The correlation length-scale is the same as in Waters et al. (2015, https://doi.org/10.1002/qj.2388), and varies with the Rossby radius, from a value of 25 km at low latitudes to 150 km at the Equator (see Fig. 3 of Waters et al., 2015, https://doi.org/10.1002/qj.2388). For a 1/4° resolution ocean model, which is limited in its resolution of mesoscale features, using the same correlation length-scale for BGC as for physics is probably appropriate for an initial implementation. It is true though that the appropriate correlation length-scale(s) to use for assimilating BGC-Argo is an open question, and this should be addressed in future development of the assimilation. I will therefore clarify in the text the value of the length-scale used, and add some discussion of these issues to the final Discussion section.

L165-170: Can the surface information help constrain the BGC fields below the mixed layer? '… below the mixed layer the vertical length-scale increases with the model's vertical grid resolution.' this is confusing to me.

I will rephrase and expand the description of the vertical correlation length-scales to try to be clearer. The length-scales are designed to limit the spreading of information across the base of the mixed layer. For an example see Fig. 4 of Waters et al. (2015, https://doi.org/10.1002/qj.2388) and the surrounding description in their Section 3.6.

For surface observations the length-scale is set equal to the mixed layer depth, meaning that information from the surface observations is spread to the base of the mixed layer, but has limited impact on BGC fields below it. This is a deliberate decision based on the lack of correlation between water mass properties in and below the mixed layer. This is likely to be the case as much for BGC as for physics (see e.g. Fontana et al., 2013, https://doi.org/10.5194/os-9-37-2013).

The vertical correlation length-scale is set to a minimum value at the mixed layer depth, and then increases with depth (see Fig. 4 of Waters et al., 2015, https://doi.org/10.1002/qj.2388). This increase is proportional to the increase in vertical model grid spacing that occurs with depth. I will rephrase the description in the text to make this clearer.

L172: 'The increments … from the two methods should be similar, though not identical.' Why? Isn't that the two methods have different treatments below the mixed layer?

As I propose to remove the section comparing ocean colour assimilation strategies, I will also remove this sentence.

L191: are these ratios fixed or time-dependent?

Again, as I propose to remove the section comparing ocean colour assimilation strategies, I will also remove this section. But to answer the question, the ratios are time-dependent, based on the background ratios in each assimilation cycle.

L216-217: this sentence should be reworded, something like: 'The approach taken to the assimilation of partial pressure of CO2 (pCO2) into HadOCC (While et al., 2012) is therefore adopted here with pH. In HadOCC, pCO2 is a function of temperature, …'

I will reword the sentence as suggested.

L261: Fujii et al. 2019 suggested the assimilative model to be configured either in reduced resolution or sufficiently different physical parameterizations.

I will clarify this in the text.

L272: 'year 5000', is it true or typo?

This is true. Spinning up UKESM1 for CMIP6 was a massive endeavour, as documented by Yool et al. (2020, https://doi.org/10.1029/2019MS001933).

L311: 30% is fine for estuarine and coastal waters, but would it be too large for chl-a profiles in open ocean?

30% is a commonly used value in open ocean chlorophyll assimilation studies (e.g. Pradhan et al., 2020, https://doi.org/10.1029/2019MS001933), and especially for daily products I would expect this to be appropriate. For instance, Krasemann et al. (2017, https://esa-oceancolour-cci.org/sites/esa-oceancolour-cci.org/alfresco.php?file=6d534e45-fbfd-4cc5-8125-d84f0b3abea6&name=OC-CCI-PVIR-v3.20170303.pdf) found an RMSD of 0.31 for ocean colour matchups of $\log_{10}$(chl-a) against in situ observations. Maritorena et al. (2010, https://doi.org/10.1016/j.rse.2010.04.002, Fig. 10) estimated errors to be in excess of 30% across much of the ocean for daily products, though lower for monthly composites.

L327: for these variables, are the error standard deviations fixed or monthly varying as well?

They are fixed, I will make this clearer in the text.

L357 & Table 1: would it be clearer to reserve the term 'control run' for the definition in Eq 2 only and call the 'non-assimilative run' the 'free run' throughout the text?

It would, I agree. I will change that.

L391: What's the DA impact for depths below 250 m?

As suggested above, I will remove this section. But the impact reduces quickly below 250 m, as can be seen in Fig. 4.

Figure 7 does not include O2 or pH while Figure 8 does. What's the rationale of presenting different set of variables here?

Fig. 5-7 were chosen to match the variables shown in Fig. 2, with extra variables just presented in Fig. 8 to limit the number of figures. I agree that it would be better to expand the range of information presented in this section, as suggested in detail by Referee #2, and so will present extra assessment accordingly.

L522: Any comment on why O2 is not improved by BGC-Argo data? And why "in situ technologies such as gliders" can play a role?

I think the likely reason that surface O_2 and chl-a are not really improved by BGC-Argo data, whereas surface NO_3 and pH are, is due to the relative importance of top-down versus bottom-up control for these variables, and the density of data required for the assimilation to have a major impact. In the case of NO_3, and DIC which helps control pH, concentrations typically increase with depth, and the supply of NO_3 and DIC from below the mixed layer is a major contribution to surface values. Therefore, changes at depth due to the assimilation will alter surface values through indirect processes. O_2 and chl-a typically decrease in concentration with depth, and dynamics within the mixed layer are much more important in setting surface values. For O_2, major drivers are temperature and ocean-atmosphere exchange. For chl-a a major driver is light availability. It seems that the BGC-Argo data is too sparse, even in ARGO_FULL, to have a widespread impact in these circumstances. More dedicated observations within the mixed layer may be likely to have more of an impact on surface values. For chl-a, this can be provided by ocean colour. For O_2, an obvious candidate is gliders. I will add a fuller description of these points in the revised text.