Response to Referee #2

The manuscript by Ford presents an OSSE experiment to investigate a number of assimilation strategies for ocean colour (OC) and biogeochemical Argo (BGC-Argo) observations using an already published DA method. The simulations are performed using a global model and sets of synthetic observations that resemble the current L3 chlorophyll OC and two potential arrays of BGC-Argo based on the current Argo network.

Thank you for your review and constructive comments. I will answer these in turn below.

The manuscript is well written and the performed modelling experiments allow novel and useful insights on the integration of BGC-Argo and OC data into global model assimilation. However, as presented, results seem rather superficial. The work would be the basis for a very valuable paper, but the results need to be explored further before I can recommend publication. My main concern is that results present a single month of simulation (i.e., the last month of a 2-year simulation: 1-year spinup and 1-year assimilation) and a single global statistics (e.g., Figs. 3, 4 and 10). Conclusion/discussion on assimilation strategies and impact of the observing systems are possibly misled by the limited results. Results of the whole year of the assimilation runs should be presented and the MEAred maps and profiles enriched with spatial and temporal statistics to provide quantitative insights on the impact of BGC-Argo in different seasons and regions of the global ocean. In particular, which are the areas and seasons that could benefit most from BGC-Argo assimilation and the integration of the two observing systems? I feel that the manuscript misses the objective to provide useful indications to design future observing system strategy, as it is proposed in the title.

Thank you for your comments. I agree that more detailed assessment could be presented, and that greater analysis of regional and seasonal results would aid with the aim of providing recommendations for future observing system strategy. I will add these results to the revised manuscript as suggested.

A second issue concerns the comparison between the PHY (phytoplankton update) and NIT (nitrogen balancing scheme) assimilation schemes. While the novelty of the OSSE experiment is related to the integration of OC and BGC Argo, the lack of the NIT implementation for BGC-Argo assimilation is a significant limitation of the manuscript that should be discussed. L425-427 are misleading. In fact, the choice of the PHY update scheme is explained at L229-230 (i.e., apparently, the NIT method has not been implemented for the joint BGC-Argo and OC assimilation). The joint OC and BGC-Argo assimilation strategy should be clearly described at L347-348. Then, it is not clear the objective of the first set of experiments (OC assimilation, which conclusions are mainly already published) if its results are not used for the second set of experiments. Even if the NIT method has not been implemented, the manuscript can be completed by a more unbiased discussion on pro and cons of the two methods and by presenting the work to be done and the the benefits to have the NIT method working with the BGC-Argo assimilation. In fact, some conclusions seem misleading. While the nitrogen balancing scheme is reported as the method with more potentiality (L406-407 in results and L510-L515 in discussion), results on BGC-Argo assimilation runs do not support this conclusion. I suggest to clarify better the proposed assimilation strategy and to detail better what would be needed to have the NIT method working with BGC-Argo assimilation. The conclusion that "only minimally altered for use with MEDUSA, so more specific tuning may help (L516-517)" seems misleading.

Referee #1 also questioned the value of including the comparison of different ocean colour assimilation strategies in this manuscript. Based on the feedback of both referees, I propose to remove this section of the paper. This should result in a more focussed manuscript and avoid the confusion my description of the different strategies seems to have caused. It will also allow more space to further explore the results of the BGC-Argo experiments. I will leave discussion of the multivariate update strategies to the final Discussion section, when detailing future work.

Minor points:
Line 173: why should the two methods provide similar increments? One is uniform with depth from surface to MLD depth, while the second method is not limited to the MLD and vertical increments are mediated by covariance.

In the 3D method the vertical correlation length-scales are defined to allow surface information to be spread to the base of the mixed layer but not below it. I will rephrase my description of the assimilation method to make this clearer.

Line 335: can the author provide some more details on how the observation and background errors have been matched? and which is the value of inflation?

The background error standard deviations estimated using the Canadian Quick method were output on the model grid, and the global mean value calculated. For each variable, the estimated standard deviations were multiplied by a constant so that the global mean value now matched the constant observation error standard deviations of 0.638 mmol m^{-3} for NO_3, 2.767 mmol m^{-3} for O_2, and 0.006 for pH. This meant that the global mean of each field matched the observation error standard deviations, while maintaining the spatial variation of the original estimates. I will add this information to the revised manuscript.

Line 336: this sentence is not clear: the observation error is set from the real global BGC-float array, so what will change when the system is functioning with real BGC-floats? More generally, the discussion missed to tackle this topic: how much are the results affected by the selected observation and background errors? would it be a different impact of the two observing systems using different observation errors?

The observation error will remain the same, but the background error will change, which I will clarify in the revised manuscript. In part, this will be due to the different model parameterisations used in the OSSE framework compared with the standard model setup. Furthermore, the background error should reflect the error in the assimilative system, which will be dependent on the number and locations of BGC-Argo floats in the real ocean. It is likely that a different specification of the observation and background errors would give different quantitative results, but show a similar qualitative impact. I will add discussion of this to the revised manuscript.

Table 3 can be enriched to improve the identification of the runs at a glance. I would suggest to add 2 new columns for the assimilated and updated variables, and to split the note column in two new ones: background error (i.e., vertical propagations: 2D and 3D) and type of increment.

Removing the comparison between ocean colour assimilation strategies means much of this information is now redundant, and I will simplify the table accordingly to aid identification of the runs.

L354-355: explain how MAEosse and MAEcontrol are computed for pixels in the maps of Figures 5-9 and for points of the profiles in Figures 3, 4 and 10. Which are the distribution compared?

For the maps, the MAE is calculated independently for each model grid cell by calculating the absolute difference between the model run and the nature run on each day of the given time period (the 31 days of December in this case), and then calculating the median of those 31 values. For the profiles, at each model depth level the absolute difference between the model run and the nature run on each day of the given time period is calculated, giving a set of values comprising of 31 days x 1442 longitudes x 1207 latitudes (with land points then excluded). The median of this set of values, weighted by the area of each grid cell, is then calculated to give the global MAE value for that depth level. I will add details of the calculation to the revised manuscript.

L377: the absolute differences in Atlantic and Indian Oceans appear significant (i.e., 1 order magnitude). Can the author provide more details about which modifications of the FREE run (w.r.t. NATURE run) should have served to increase nutrient concentration? And which modifications compensate it (L382)? I agree that achieving a global appropriate level of error with a complex model (with uniform parameterization) can not be managed, however the author should provide some details on which modifications didn't work as expected. This can be helpful in understanding the effectiveness of the data assimilation in those areas.

The nutrient differences in the Atlantic and Indian Oceans are of an order of magnitude, but with low absolute values, typically increasing from O(0.01) to O(0.1) mmol N m^{-3}. It is difficult to pinpoint the exact cause of any given change, as several parameters have been altered, and these will have complex interactions depending on the underlying concentrations of different variables. One potentially significant change though is that the nutrient uptake half-saturation concentration for phytoplankton was greatly increased for nitrogen, and decreased for iron. In areas which are nitrogen-limited, phytoplankton will therefore be less efficient at taking up nutrients. Furthermore, a decrease in zooplankton grazing half-saturation concentration means zooplankton become more efficient at grazing low phytoplankton populations. In NATURE, the Atlantic and Indian Oceans are the areas with the lowest DIN and phytoplankton concentrations. A first-order explanation may therefore be that the increased nitrogen uptake half-saturation concentration means phytoplankton take up less DIN, resulting in higher DIN concentrations. This then allows greater phytoplankton growth, as more DIN is available, though it is used less efficiently. This is then balanced by an increase in grazing, resulting in slightly elevated DIN and zooplankton concentrations, but largely unchanged phytoplankton concentrations. In other areas, which aren't so nitrogen-limited, the balance of processes is different, leading to different changes. I will add discussion of this to the revised manuscript.

L385: the absolute difference of NO3 and pCO2 between FREE and NATURE seems much lower than that between FREE and real observation. Can the author discuss which are the implications of the low difference for the OSSE assimilation results? I would argue that the effectiveness of the assimilation might be limited in some areas by the low differences between the synthetic observations generated from NATURE and the FREE run. Further, I would argue that MAEred might not be a good metric because of the MAEcontrol at the denominator in the areas where NATURE and FREE are so close.

From previous studies, the conclusion has been that an insufficient level of error would lead to "an overestimation of impact when sparse data are assimilated and an underestimation when dense (e.g., satellite) data are assimilated" (Halliwell et al., 2014, https://doi.org/10.1175/JTECH-D-13-00011.1). This suggests that this study may overestimate the impact of BGC-Argo observations in some regions. I mentioned this in the original manuscript, but agree it's a point which deserves further discussion, which I will add to the revised text. I will also add some assessment of the absolute as well as percentage reduction in MAE, to avoid the issue of having $MAE_{control}$ in the denominator. This gives generally similar conclusions, but I agree is a useful additional way of looking at the results.

L395: Since large areas of the global ocean are characterized by a DCM below 60m depth, the degradation of MAE below 60m would deserve a more detailed comment.

Given that phytoplankton biomass is not degraded, I would speculate that depth variations in carbon-to-chlorophyll ratios are not being correctly characterised in these runs. However, as I plan to remove the comparison of ocean colour assimilation strategies, and the remaining OC_3D_PHY run does not show this degradation, I propose to remove this discussion from the revised manuscript.

L405: why should the similar behaviours of OC_2D and OC_3D demonstrate that the use of NEMOVAR to create 3D increments for the combined OC and BGC-Argo float assimilation is fit-for-purpose?

This comment was simply intended to imply that because OC_3D gives similar behaviour to the proven strategy of OC_2D, assimilating ocean colour in this manner (which is a prerequisite for combining OC and BGC-Argo chl-a) should also give acceptable results. As I propose to remove the comparison between ocean colour assimilation strategies, I will also remove this comment in the revised manuscript.

Figure 4: I would suggest to reduce y-axis to 0-250m depth for the chlorophyll, phytoplankton and zooplankton plots to increase their readability. Are the high positive values below 250 in chla and phytoplankton plots due to the very low values of those variables below the euphotic zone?

I will alter the figure as suggested. The high percentage values are indeed due to the low absolute values, and I will add comment on this in the text.

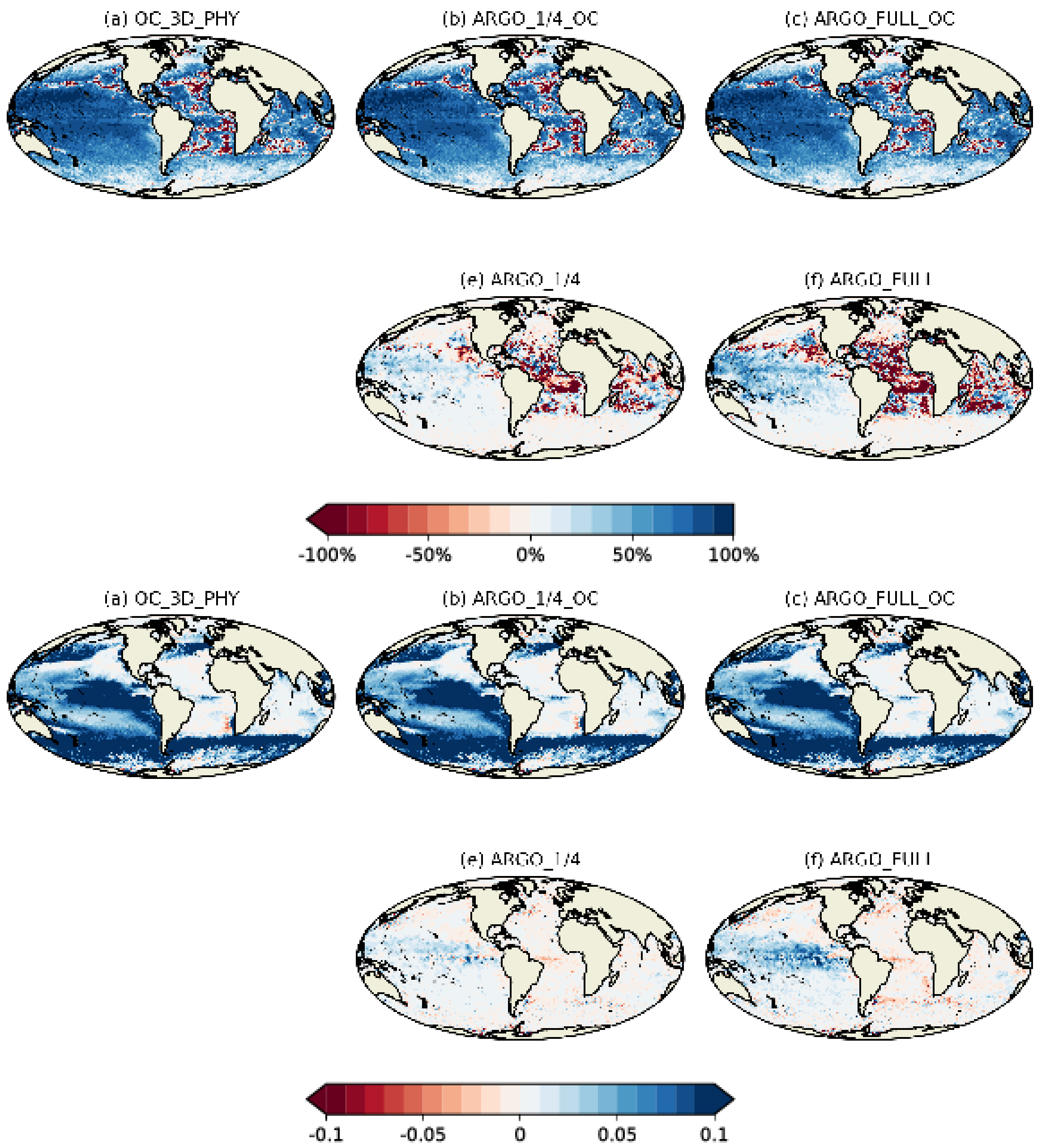
L450-451 and L477-478: Can the degradation of Alkalinity be due to an improper working of the smallest combined change of DIC and Alk with pH assimilation? A more detailed analysis is expected to show the pro and cons of the method when pH is used instead of pCO2.

Testing of the scheme shows the calculation is being performed correctly, as also indicated by the overall improvement in pH. The cause is likely to be that making the smallest combined change to DIC and alkalinity is not necessarily the approach that minimises errors in both DIC and alkalinity. In some circumstances it might be more appropriate to e.g. make a smaller or no change to alkalinity, and a larger change to DIC. Or even to make a change of the opposite sign to alkalinity, and an even larger change to DIC to compensate. Unfortunately, without concurrent observations of DIC or alkalinity, this information is not known at the time of assimilation. This is equally the case whether pCO_2 or pH is being assimilated. An assumption therefore needs to be made, and during the development of the original pCO_2 assimilation scheme it was decided that the safest assumption would be to make the smallest combined change in DIC and alkalinity – an assumption adopted for pH in this study. In light of these results it may be worth revisiting this assumption, but to do so effectively would involve a great deal of experimentation which is best left for a future study. I will add discussion of these issues to the revised manuscript though.

L454-L455: why are the negative values of MAEred in the Atlantic and Indian Oceans (Fig. 5) related to the compensating errors introduced in FREE? Since the FREE and NATURE differences are very low in those areas (Fig. 2), the impact of the assimilating synthetic observations (generated from NATURE) should be negligible. I

wonder whether the MAEred is not a good metric because of the MAEcontrol at the denominator for those areas.

Below is the original version of Fig. 5, showing the percentage reduction in MAE, and an alternative version showing the absolute reduction in MAE. It is true that the response in the Atlantic and Indian Oceans is minimal in absolute terms, and I will rephrase the text accordingly. I will also present results showing both absolute and percentage reduction in MAE in the revised manuscript, as these do show complementary information.



Figures 8 and 10 seem redundant and not necessary. For example, the MAEred over OC_3D_PHY of ARGO_FULL_OC (Fig. 8a and Fig 10a) provides the same information (except for the normalization of denominator) of the difference between MAEred over FREE of ARGO_FULL_OC and MAEred over FREE of OC_3D_PHY (Fig. 4a and Fig 5a and c). I suggest that the relative impact of adding BGC-Argo can be shown by a new table of the MAEred over FREE numeric values. The table can report values for

selected regions and different seasons/months providing indications of which areas of the global ocean and periods of the year can benefit most by the BGC-Argo assimilation.

As suggested, I will expand the range of assessment performed, including looking at seasonal and regional statistics, and will modify the presentation of the results accordingly.

L505-508: which parameter settings between NATURE and FREE caused the degradation of the other variables? Can additional details be added?

The interaction between different parameter changes is complex, and varies depending on the underlying concentrations of each of the variables. The biggest impact on zooplankton though is likely to have come from alterations to the grazing half-saturation concentration, which was changed from 0.8 mmol N m^{-3} in NATURE to 0.36 mmol N m^{-3} in FREE for microzooplankton, and from 0.3 mmol N m^{-3} in NATURE to 0.135 mmol N m^{-3} in FREE for mesozooplankton. Other significant changes to the ecosystem dynamics are likely to have come from changing the nutrient uptake half-saturation concentrations for phytoplankton. For nitrogen, this was changed from 0.5 mmol N m^{-3} in NATURE to 2.13 mmol N m^{-3} in FREE for non-diatoms, and from 0.75 mmol N m^{-3} in NATURE to 3.195 mmol N m^{-3} in FREE for diatoms. For iron, this was changed from 0.00033 mmol Fe m^{-3} in NATURE to 0.00011 mmol Fe m^{-3} in FREE for non-diatoms, and from 0.00067 mmol Fe m^{-3} in NATURE to 0.00022 mmol Fe m^{-3} in FREE for diatoms. I will add more details to the revised manuscript.

L510-511 please explain: this seems not supported by results or references.

Assimilation schemes which use ensembles to generate cross-correlations are reliant on the model relationships between variables being correct, as it is these model relationships which the cross-correlations are based on. If the response of zooplankton to an increase in phytoplankton in the model ensemble differs from that in the real ocean, then the cross-correlations used in the assimilation will lead to a zooplankton response which follows the (incorrect) model rather than the real ocean, in exactly the same way as seen in this study. I will clarify this in the revised manuscript.

L522: BGC-Argo assimilation has a small and positive impact on O2 as shown in Figure 10f (BGC-Argo assimilation). The degradation of O2 at surface seems due to OC assimilation (see figure 4f).

Agreed. This was clumsy wording on my part, and I will clarify this in the revised manuscript.

L531-L535: DA method improvements are important, but the paper does not really tackle this aspect; thus, those lines may fit better in the introduction and not as the last conclusion.

I believe that the Discussion section is the best place for this discussion, as it details future work and recommendations arising from the results presented. I acknowledge that the original manuscript could have done this more effectively, and propose to expand the discussion around assimilation improvements, and relate it better to the results, in the revised manuscript. I also take the point that something else, such as recommendations on observing system strategies, would fit better as the last conclusion.