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.

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.

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.

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.

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.

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.

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.

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).

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?

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

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?

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.

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

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

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.

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.

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?

L191: are these ratios fixed or time-dependent?

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, …'

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

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

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

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

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?

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

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

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?