**Responses to Reviewer 1**

Below the complete reviewer comments are shown in black font along with detailed responses to each comment in blue font.

Review:
This is a very interesting and valuable contribution to the present scientific literature addressing the impact of (DEnKF) assimilation of physical and biogeochemical data, including Argo floats, on the simulated state of the ocean. More specifically, the paper addresses an important question of the impact of (mostly) surface data assimilation on sub-surface physics and biogeochemistry, focusing on the Gulf of Mexico. The simulated physical and biogeochemical tracers are validated with an independent BGC-Argo data-set. The paper is well written and I have mostly only minor comments to address, which can be found below. One remark: please note that in the downloaded .pdf version of the manuscript the first digits of the three-digit line numbers are running off the left side of the page, so beneath the section 1 I avoid referring to the line numbers and refer to the paragraph and section number instead.
Response: We thank the reviewer for the constructive comments and suggestions which will be very helpful as we revise the manuscript.

Section 1:
The introductory section is nicely written and the only comments that I have are regarding the use of references:
- in particular on line 52 when the paper talks about assimilation of optical properties, the included references are very far from exhaustive. On top of my head I could add to the list the following papers: 1. Jones et al (2016), Biogeosciences, https://doi.org/10.5194/bg-13-6441-2016, 2. Gregg and Rousseaux (2017), Frontiers in Marine Science, https://doi.org/10.3389/fmars.2017.00060, 3. Skakala et al (2020), Journal of Geophysical Research: Oceans, https://doi.org/10.1029/2020JC016122. I completely understand that sometimes there simply are too many references and you cite only the first ``pioneering'' papers, but in such case it would be good to put "e.g." in front of the citations, or add after the citations "with few more recent follow-ups", or something similar. Otherwise this might look like the cited papers are all that has been published on the topic and it might lead to the other papers being omitted when someone new-to-the-area uses your references as part of literature review on the topic. Similarly to this, I would make sure you somehow emphasize that the citations aren't exhaustive also in the other cases, e.g. the phytoplankton size-class chlorophyll has been assimilated also in Ciavatta et al 2019, Journal of Geophysical Research: Oceans, https://doi.org/10.1029/2019JC015128 and for total surface chlorophyll there's many references left out.
Response: Thank you very much for pointing out this. We will add these references and revise the corresponding text in our revised manuscript as suggested.

- more importantly the text on the lines 65-70 suggests that the subsurface validation was done only in the 2 cited studies and only on climatological, or basin-wide scale. Firstly, I believe it is appropriate to include here the Cossarini et al 2019 reference that is cited elsewhere in the paper (which looks at vertical profiles along BGC-Argo locations), but secondly there is a new study of Skakala et al (2021), Journal of Geophysical Research: Oceans, https://doi.org/10.1029/2020JC016649 that validates both free runs and surface data assimilative runs along glider trajectories (in the North-West European Shelf) to determine the impact of surface data assimilation (and some other DA as well) on the simulated sub-surface tracers. This study, although it uses a very different, 3DVar, system, is of direct relevance here, since it addresses similar questions to this submitted manuscript, and, as mentioned before, it does model validation along a specific glider 3D transect (which is of course spatially limiting), rather than on basin-wide scales. Btw. please note that there is another study that assimilated BGC-Argo oxygen data that perhaps deserves to be cited in this paper: Verdy and Mazloff, Journal of Geophysical Research: Oceans, https://doi.org/10.1002/2016JC012650.
Response: We will include these references in our revised manuscript and rephrase this paragraph into:

*"The insufficient availability of subsurface and interior ocean biogeochemical observations is not only reflected in the immaturity of biogeochemical data assimilation but also its skill assessment. When compared with the surface, the subsurface has received less attention in skill assessments of biogeochemical data assimilation systems. Although there already have been studies which compared their vertical structures with in-situ observations and/or climatological datasets (e.g. Fontana et al., 2013; Ford and Barciela, 2017; Mattern et al., 2017; Ourmières et al., 2009; Teruzzi et al., 2014), these validations are often limited to low spatio-temporal resolution. The recent growth of autonomous observation systems, esp. BGC-Argo floats and gliders, make it possible to evaluate biogeochemical data assimilation systems below the surface in high resolution (e.g. Cossarini et al., 2019; Salon et al., 2019; Skákala et al., 2021; Verdy and Mazloff, 2017)."*

Section 2:
I find it would be perhaps beneficial for the reader to provide slightly more information on the system set-up and maybe a bit rearrange the text. In particular:
- could you please mark the exact model domain? The Fig.1 shows the Gulf of Mexico region with its bathymetry, but I guess what's in Fig.1 is not precisely the spatial model domain (?) Or is it the red rectangle in Fig.1, which however isn't explained (?) Maybe you can mask out the ``irrelevant'' regions?
Response: The model domain is represented by the red rectangle. We will make it clearer in the main text and figure caption of our revised manuscript.

- a very minor (bottom of first paragraph in the section 2.1): could you please provide a rough km scale for the 1/8 degree resolution. Of course this is slightly variable depending on the latitude, but is it something like ~12km?
Response: The longitudinal distance is about 12 km and the latitudinal distance is about 14 km. We will add this into our revised manuscript as suggested.

- again extremely minor: just beneath Eq.2 "measurement operator" should be "observation operator"?
Response: We will revise it as suggested.

- I would perhaps suggest to put the first two paragraphs in section 2.4 behind the section 2.1, or maybe behind the section 2.2, so you say clearly how the ensemble is generated, how many members are run (and so on), around the same time when you talk about ensemble DA. I got firstly confused how little information is provided about the ensemble generation until I realized it's been put behind the "Observations" section.
Response: Thank you very much for this suggestion. We will move these two paragraphs to behind the section 2.2 in our revised manuscript.

- regarding the ensemble generation: how was the parameter sensitivity determined, in order to perturb the ``right'' parameters to account for the model uncertainty? Physical model parameters weren't perturbed (e.g. vertical diffusion parameter values)? If so, why? I know 20 member ensemble is itself limiting, but perhaps some more reasoning on how the ensemble was generated would be desirable.
Response:
The biological parameters were selected by sensitivity tests. Specifically, we ran a 1D version of this model multiple times by incrementally perturbing one parameter at a time and setting other parameters unchanged (hereafter as **Test case**). Each parameter ($p$) was perturbed by multiplying factors ranging from 0.25 to 1.75 with an increment of 0.25. Then parameters were sorted based on their sensitivity which was quantified by the normalized absolute differences from the unperturbed run (hereafter as **Base case**):

$$Q(y,p) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\frac{|y_{Base} - y_{Test}|}{y_{Base}}$$

where $m$ is the number of parameter increments (here 7) and $n$ is the number of base-test pairs including all 1D model grid cells throughout the whole simulation period. Based on the sensitivity tests, the most

sensitive parameters were selected and perturbed in our data assimilation experiments. The details about the 1D model can be found in the section 3.3 of Wang et al. (2020). In our revised manuscript, we will add explanation about the choice of sensitive biological parameters.

The physical model parameters (e.g. vertical diffusion coefficient) weren't perturbed in this study. To create an ensemble with sufficient spread, we perturbed the model initial conditions, the open boundary conditions, and the wind forcing. Details in these perturbations can be seen in the section 2.4 of our original manuscript.

It is worth noting that this data assimilation framework and configurations (e.g. observational errors, ensemble perturbations, ensemble members, and etc.) are obtained from Yu et al. (2019) which performed twin experiments in the same region (the Gulf of Mexico). In this study, we extend the work of Yu et al. (2019) to jointly assimilate the physical and biological observations.

- it seems that the observation uncertainties don't come with the observation product (?), but based on what you estimate the 35% observational uncertainty for total chlorophyll, or the physical observation errors? This might be of importance, since you are aiming at estimating model variances quite accurately with the ensemble, so the true state estimate can be quite sensitive to the observation error estimates.
Response: An observational error of 35% (or 30%) is a common practice to specify the error of satellite surface chlorophyll (e.g. Fontana et al., 2013; Ford, 2021; Ford and Barciela, 2017; Hu et al., 2012; Mattern et al., 2017; Santana-Falcón et al., 2020; Song et al., 2016; Yu et al., 2018). The observational errors of SSH and SST are also based on references (Song et al., 2016; Yu et al., 2018, 2019).

- btw do you assimilate chlorophyll, or log-chlorophyll? If the former, how do you treat the problem of non-Gaussianity?
Response: We assimilate the actual chlorophyll concentrations in this study. Based on our prior tests, assimilating log-chlorophyll can yield more improvements of the surface chlorophyll in coastal regions but degrade it in the deep ocean (with depth>1,000m). In addition, assimilating log-chlorophyll results in less improvements of subsurface biological properties, including the chlorophyll, phytoplankton, and POC. As this study focused on the subsurface biological properties in the deep ocean of the Gulf of Mexico, we decided to assimilate the actual chlorophyll. Also, there already have been successful examples which assimilated the actual chlorophyll (e.g. Hu et al., 2012; Yu et al., 2018). Nevertheless, we acknowledge that assimilating the actual chlorophyll is suboptimal because of its non-Gaussian nature and will state this more clearly in the revised manuscript.

–what exactly is the 7 day window assimilation at the end of third paragraph of section 2.3, and how exactly does it help with data sparsity, should be explained more in detail.. E.g. is this that you assimilate more data that have been otherwise merged in the OC CCI products?
Response: In this study, we applied update every week. Since the spatial coverage of surface chlorophyll is limited (with average of ~ 9.5%), we not only assimilate the daily records of surface chlorophyll at the date of update (e.g. 7 Jan 2015, the first update date in our data assimilative experiments), but also assimilate all daily records within 7 days before (e.g. from 1 Jan 2015 to 7 Jan 2015). Literally, the abundance of surface chlorophyll data that we assimilated will increase by approximately 7 times. We will explain this in more detail in our revised manuscript.

- I am not particularly great expert on EnKF, but maybe you could say for a general readership a bit more about how exactly are the correlation length-scales (especially the vertical) calculated from the ensemble? There is something on the spurious correlations and localisation, but perhaps you can say a bit more? Also I understand correctly that SSH, T & S are updated without updating also the horizontal velocities? Could that be also discussed slightly more?
Response:
The data assimilation framework and configurations (e.g. localization radius, observational errors, ensemble perturbations, ensemble members, and etc.) used in this study are the same as in Yu et al. (2019)

which has successfully performed twin experiments in the same region (the Gulf of Mexico). The localization radius was tuned by initial sensitivity tests in Yu et al. (2019). Specifically, the different values of localization radius (i.e. 50, 75, and 100km) were tested and the optimal value (50km) was selected. The vertical correlation length-scale is not used in this study.

In our data assimilation experiments, the 3D distributions of temperature and salinity were updated. The SSH and horizontal velocities were adjusted by dynamics in the following forecast steps. In the deep Gulf of Mexico, the circulation features (e.g. the Loop Current and its associated mesoscale eddies) are primarily geotropically balanced. Therefore, an update of temperature and salinity can improve the 3D circulation features in large scales effectively. This can be also supported by the twin experiments in Yu et al. (2019). As this study mainly focuses on the biological aspect, discussing the update of horizontal velocities may be a little distracting, but we will refer to the explanations in Yu et al. (2019) more clearly in the revised manuscript.

- the skill metrics at the end of section 2: why don't you also include the overall bias? I find that calculating bias always carries worthy information (at least worthy enough to be mentioned somewhere in the paper).
Response: We agree that the bias can carry supplementary information to the RMSE. As also suggested by the review 3, we will report the bias and correlation coefficient in our revised manuscript.

Section 3
Nice, just some relatively minor clarifications would help:
- section 3.2: I find it surprising that addition of Argo (physical) data to the assimilation (Daargo) has such impact on surface chl (Fig.6), especially if it didn't change much T & S (Fig.4)? Could you comment a bit more?
Response: The degradation of surface chlorophyll in DAargo run is caused by the multivariate update between the Argo T&S profiles and chlorophyll, rather than the dynamical adjustment. Figure r1 shows the forecast and analysis of surface chlorophyll at the first update date (7-Jan-2015). The white circles in Figure r1c represent the assimilated Argo T&S profiles. We can see that assimilating the Argo T&S profiles can yield extremely high surface chlorophyll around some profiles (e.g. the profile located at about 96.5°W, 24°N). This is due to the spurious correlations between the T&S profiles and surface chlorophyll. This issue has been also reported in Goodliff et al. (2019) which assimilates SST to update both physical and biological variables. In their study, the multivariate update can yield unrealistic chlorophyll concentrations no matter when the chlorophyll has been log-transformed. Muting the multivariate update on phytoplankton, zooplankton, and detritus can make chlorophyll concentrations more realistic.
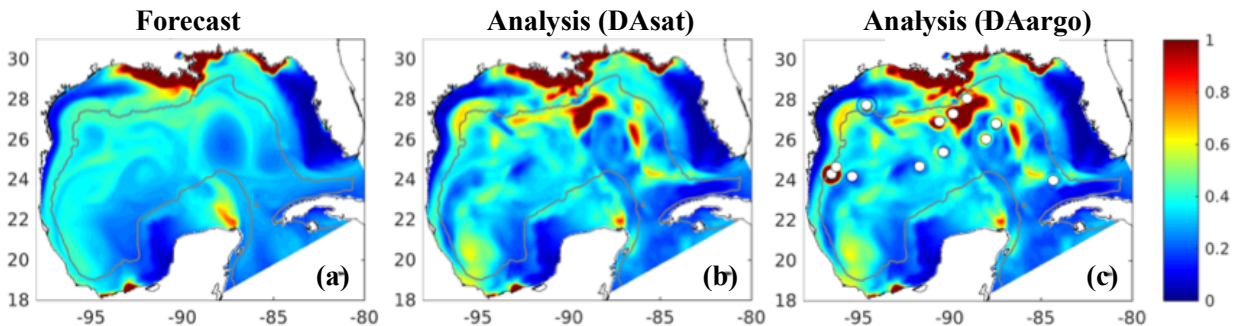


Figure r1. The forecast and analysis of surface chlorophyll at the first data assimilation cycle (7-Jan-2015). The white circles in panel c represent the assimilated Argo T&S profiles.

Section 3.2 paragraph 2: where are the nitrate data coming from? Maybe I'm getting confused, but nitrate data weren't mentioned between the BGC Argo data in section 2.
Response: The NO3 is not measured by the BGC-Argo floats. It is estimated based on its empirical relationship with temperature (Figure S1 of our original manuscript), which is derived from the

climatological data in the Gulf of Mexico from the World Ocean Atlas (WOA). The description of estimated NO3 is in the end of the section 2.4 in our original manuscript. To make it clearer, we will move it to the end of section 2.3 (observations) in our revised manuscript.

specific Figure-related comments:
Fig.3: although I understand what the Figure is trying to demonstrate, it might be better to show the model –observation differences for the different runs, or at least show them for the free run next to the existing panels?
Response: Figure 3 in our original manuscript aimed to show reductions in RMSE and therefore the improvement obtained by data assimilation. We agree with the reviewer and we will add panels to show RMSE of the free run as a benchmark in our revised manuscript as suggested.

Fig.4: its interesting how little difference there is in subsurface T between sat DA and argo DA! Can you discuss how much of this is due to the assimilation length-scales and how much due to the model dynamical adjustment? Btw what was the relationship between the spatial positions of Argo floats and BOEM floats? This hasn't been shown..
Response:
    Figure r2 shows positions of Argo profiles (gray dots) at each data assimilation cycle (e.g. 7 Jan 2015, the first update date in our data assimilative experiments) and BOEM profiles (colored squares) before the next one (e.g. from 7 Jan 2015 to 14 Jan 2015). The solid black circles represent areas within one localization radius (50km) from each Argo profile. Colors of squares represent the days of each BOEM profile after each data assimilation cycle. As shown in Figure r2, most of BOEM profiles are outside of one localization radius from the Argo profiles and therefore are barely updated by assimilating the Argo profiles. Figure r3 shows the root-mean-square-difference (RMSD) of temperature from each BOEM profile between two data assimilative runs ($RMSD = \sqrt{\frac{1}{n}\sum(DA_{Sat} - DA_{Argo})^2}$). The x-axis represents days of each BOEM profile after each data assimilation cycle and the y-axis represents distance to the nearest Argo profile. In general, the RMSD between two data assimilative runs decreases with the distance but shows no significant decreasing trends with the days after update. This means that the differences induced by assimilating Argo profiles can be well sustained locally by model dynamical adjustments. The overall similarities between two data assimilative runs in Figure 4 can be explained to some extent by the large distances between BOEM and Argo profiles. However, it doesn't mean that increasing the localization radius can improve the data assimilation performance. We note that the current localization radius is determined by initial tests in Yu et al. (2019)
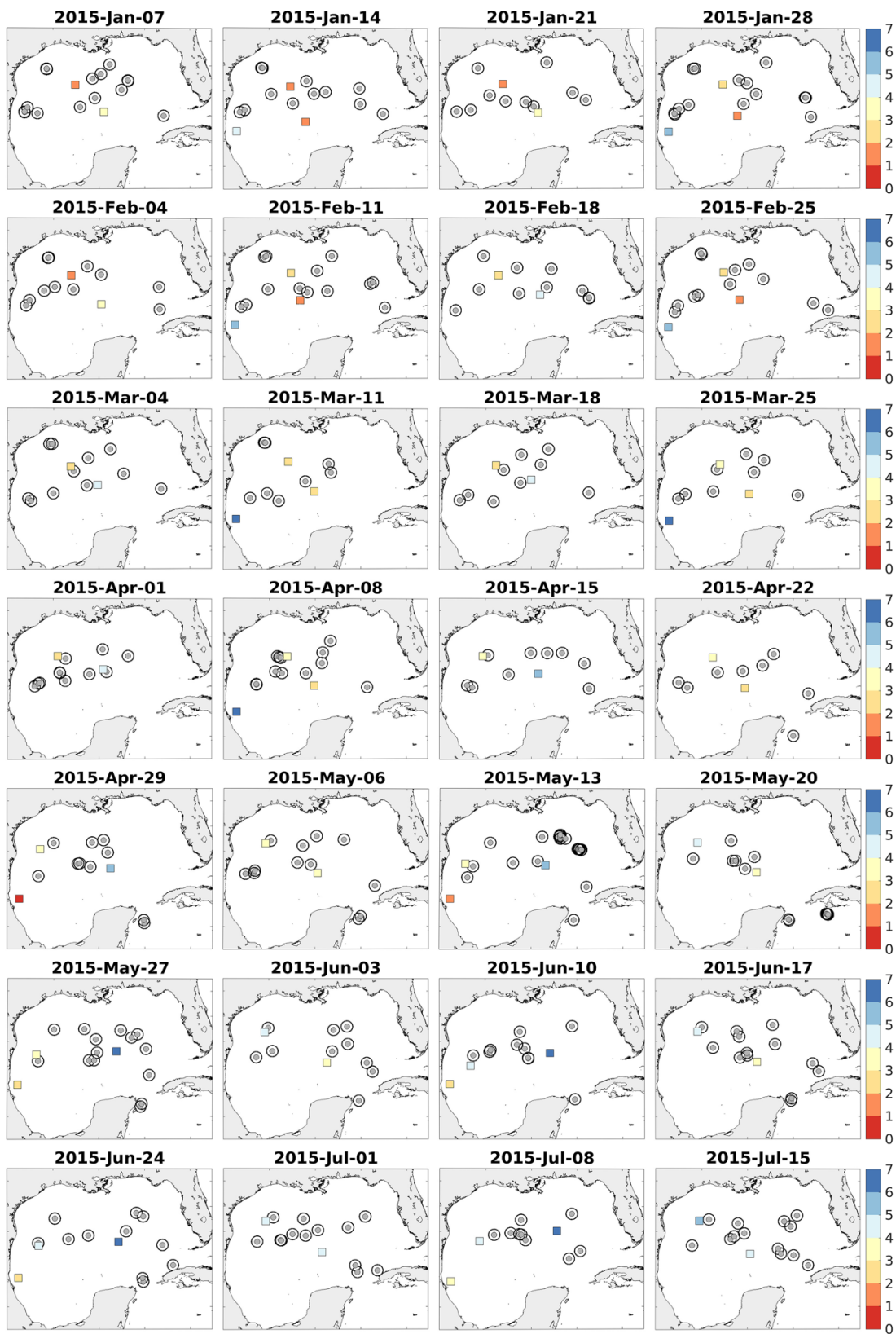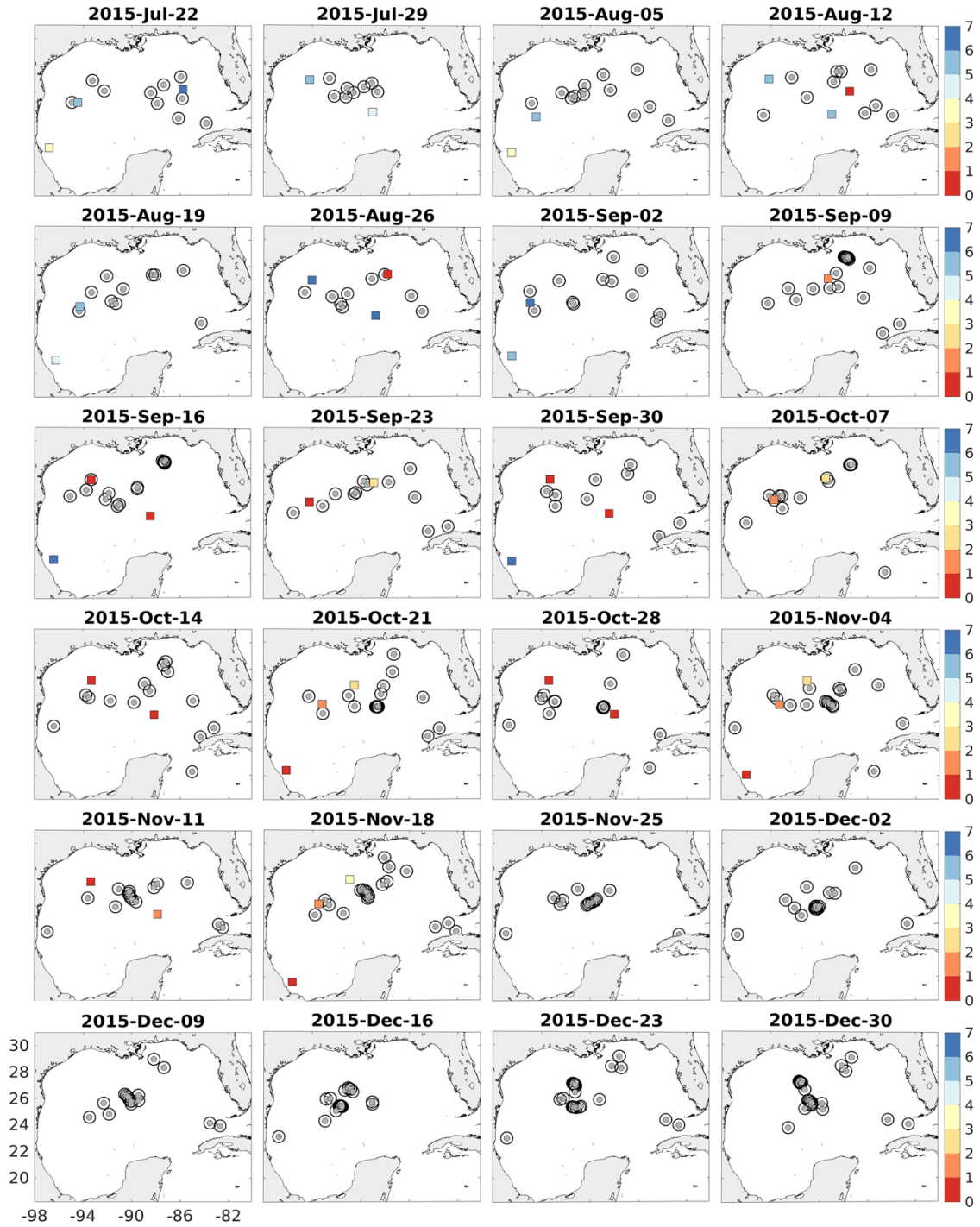
(continued)

Figure r2 Positions of Argo profiles (gray dots) at each data assimilation cycle and BOEM profiles (colored square) before the next one. Solid black circles represent areas within one localization radius from each Argo profile. Colors of squares represent the days after each data assimilation cycle.
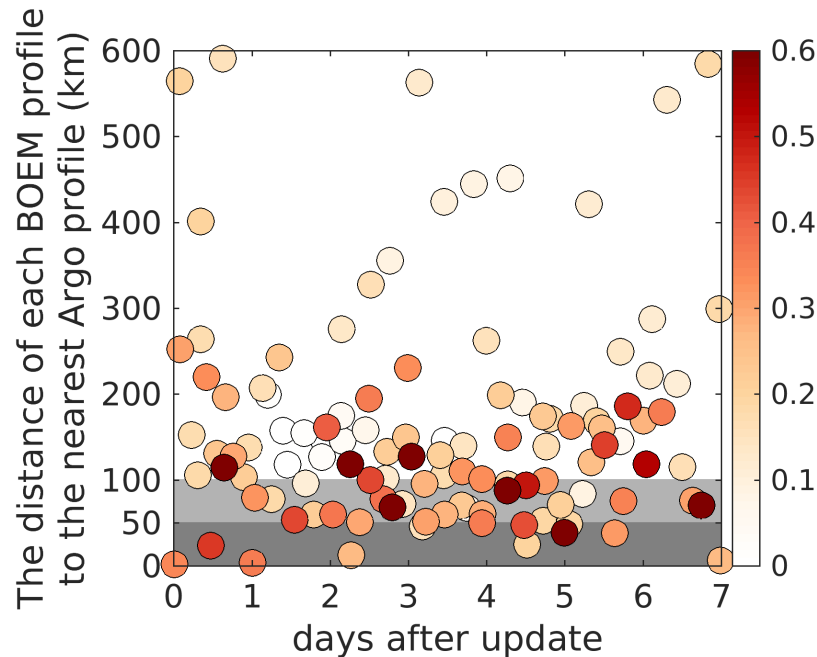
Figure r3 The root-mean-square-difference (RMSD) of temperature from each BOEM profile between two data assimilative runs, DAsat and DAargo (indicated by the color). The x-axis represents days of each BOEM profile after each data assimilation cycle and the y-axis represents distance to the nearest Argo profile

Fig.7: perhaps in this case there is no need to reproduce exactly Fig.4, since the nitrate concentrations are very similar between the three simulations? Maybe it's better to have the first row and then differences between the DA and the free run, since this would show more clearly the changes introduced by DA? Also why there aren't the observed nitrate concentrations similarly to T in Fig.4? It's interesting that nitrate and surface chlorophyll aren't correlated as much as nitrate with temperature (Fig.10), I understand that's because of non-linearity, but in my region and model of experience this non-linearity still produces overall strong correlations, just highly variable in time (including changing signature)..

Response:

The NO3 distributions are quite different before and after data assimilation. They look similar in Figure 7 of our original manuscript possibly because the NO3 are plotted in log-scales. As suggested, we replotted Figure 7 (referred as Figure r4 here) to show NO3 distributions in the free run and increment of NO3 due to data assimilation in DAsat and DAargo runs. Red colors represent increases while blue colors represent decreases of NO3 by data assimilation. As shown in Figure r4, the NO3 distributions are significantly modified and improved by data assimilation. For example, the anticyclonic eddies which are not reproduced by the free run will depress the nitraclines and decrease the NO3 concentrations (e.g. during June of the float 286, during July and October of the float 287, and during August to October of the float 289) in the two data assimilative runs. In contrast, the data assimilation will increase NO3 concentrations when the SSH is overestimated in the free run and corrected in the two data assimilative runs (e.g. during April of Float 285, during February and April of Float 287, during April to July of Float 289).

We didn't show NO3 observations because NO3 is not measured by the BGC-Argo floats in the Gulf of Mexico. In this study, we estimated NO3 based on its climatological relationship with temperature (Figure S1 of our original manuscript which is replot as Figure r5 here) and compared it with the model results. The estimated NO3 distributions following the BGC-Argo floats are shown in Figure r4. However, the estimated NO3 tend to be overestimated in the high temperature regions (Figure r5), which typically

occurs within the euphotic layer. Therefore we used the unbiased root-mean-square-error to quantify the model-data misfit of NO3.

For the relatively weak correlations between surface NO3 and chlorophyll, the deep ocean of the Gulf of Mexico (depth>1,000m) is an oligotrophic region. Apart from NO3, ammonia (NH4) can be also used to support a large fraction of primary production. In addition, the photo-acclimation is accounted for in our biological model. The chlorophyll to phytoplankton carbon ratio can vary a lot due to environmental factors, i.e. light, temperature, and nutrient concentrations (Geider et al., 1997). All these features will increase nonlinearities of the biological model and make the correlations between surface NO3 and chlorophyll weak.
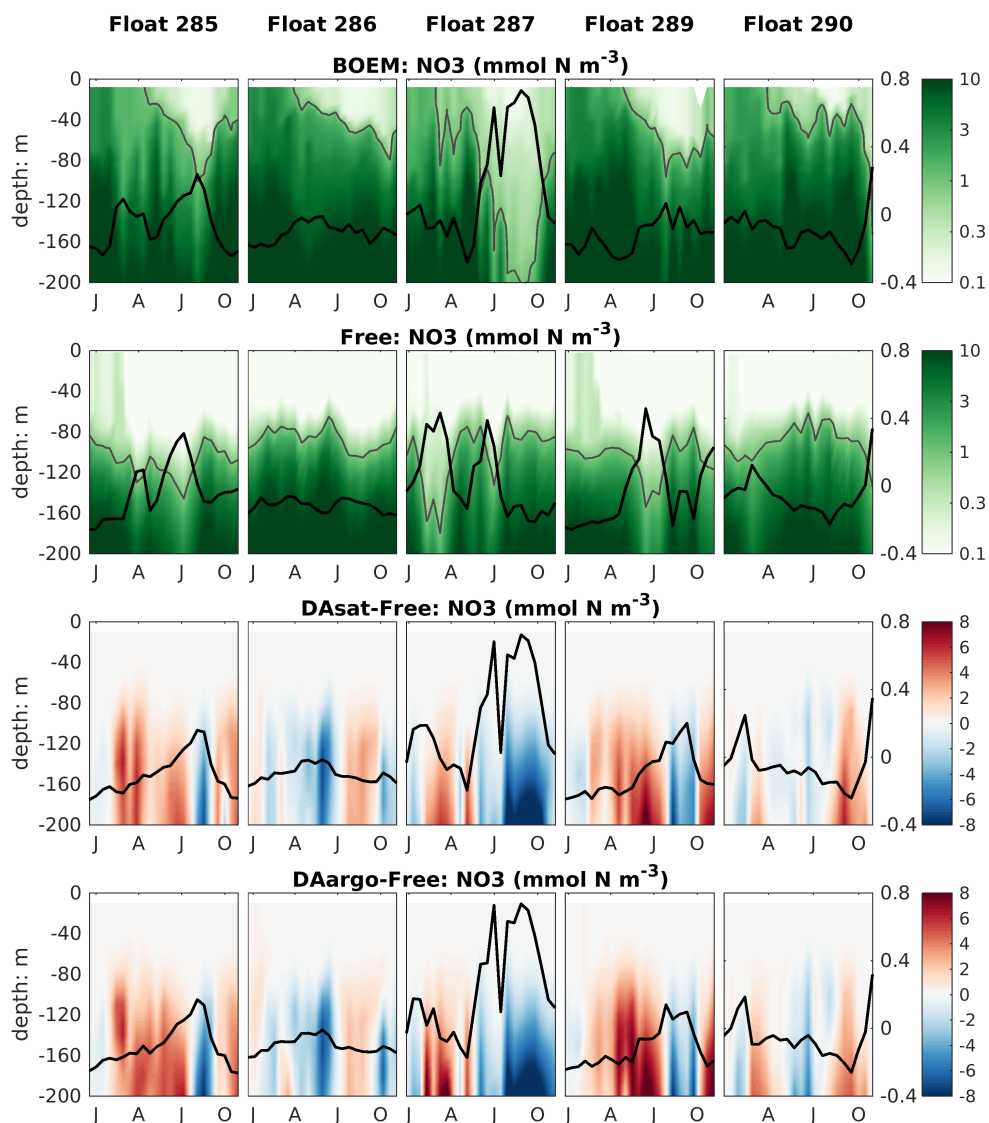


Figure r4 Vertical distributions of NO3 estimated based on its climatological relationship with temperature and modelled by the Free run, and increment of NO3 due to data assimilation in the DAsat and DAargo run. Black lines represent SSH.
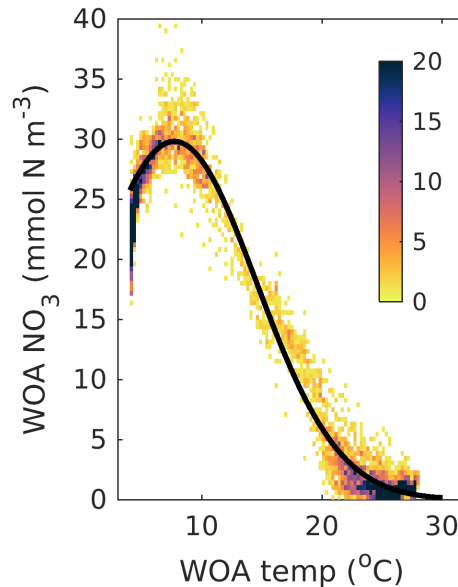
Figure r5 Empirical relationship of temperature-NO3 derived from the World Ocean Atlas in the Gulf of Mexico. Colors indicate the number of observations within each bin.

**Reference**:

Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A. and D'Ortenzio, F.: Towards operational 3D-Var assimilation of chlorophyll Biogeochemical-Argo float data into a biogeochemical model of the Mediterranean Sea, Ocean Modelling, 133, 112–128, doi:https://doi.org/10.1016/j.ocemod.2018.11.005, 2019.

Fontana, C., Brasseur, P. and Brankart, J. M.: Toward a multivariate reanalysis of the North Atlantic Ocean biogeochemistry during 1998 – 2006 based on the assimilation of SeaWiFS chlorophyll data Toward a multivariate reanalysis of the North Atlantic Ocean biogeochemistry during 1998 – 2006 based on, Ocean Science, 9, 37–56, doi:10.5194/os-9-37-2013, 2013.

Ford, D.: Assimilating synthetic Biogeochemical-Argo and ocean colour observations into a global ocean model to inform observing system design, Biogeosciences, 18(2), 509–534, doi:10.5194/bg-18-509-2021, 2021.

Ford, D. and Barciela, R.: Global marine biogeochemical reanalyses assimilating two different sets of merged ocean colour products, Remote Sensing of Environment, 203, 40–54, doi:https://doi.org/10.1016/j.rse.2017.03.040, 2017.

Geider, R. J., MacIntyre, H. L. and Kana, T. M.: Dynamic model of phytoplankton growth and acclimation : responses of the balanced growth rate and the chlorophyll a : carbon ratio to light , nutrient-limitation and temperature, Marine ecology progress series, 148, 187–200, 1997.

Goodliff, M., Bruening, T., Schwichtenberg, F., Li, X., Lindenthal, A., Lorkowski, I. and Nerger, L.: Temperature assimilation into a coastal ocean-biogeochemical model: assessment of weakly and strongly coupled data assimilation, Ocean Dynamics, 69(10), 1217–1237, doi:10.1007/s10236-019-01299-7, 2019.

Hu, J., Fennel, K., Mattern, J. P. and Wilkin, J.: Data assimilation with a local Ensemble Kalman Filter applied to a three-dimensional biological model of the Middle Atlantic Bight, Journal of Marine Systems, 94, 145–156, doi:10.1016/j.jmarsys.2011.11.016, 2012.

Mattern, J. P., Song, H., Edwards, C. A., Moore, A. M. and Fiechter, J.: Data assimilation of physical and chlorophyll a observations in the California Current System using two biogeochemical models, Ocean

Modelling, 109, 55–71, doi:https://doi.org/10.1016/j.ocemod.2016.12.002, 2017.

Ourmières, Y., Brasseur, P., Lévy, M., Brankart, J.-M. and Verron, J.: On the key role of nutrient data to constrain a coupled physical–biogeochemical assimilative model of the North Atlantic Ocean, Journal of Marine Systems, 75(1), 100–115, doi:https://doi.org/10.1016/j.jmarsys.2008.08.003, 2009.

Salon, S., Cossarini, G., Bolzon, G., Feudale, L., Lazzari, P., Teruzzi, A., Solidoro, C. and Crise, A.: Novel metrics based on Biogeochemical Argo data to improve the model uncertainty evaluation of the CMEMS Mediterranean marine ecosystem forecasts, Ocean Science, 15(4), 997–1022, doi:10.5194/os-15-997-2019, 2019.

Santana-Falcón, Y., Brasseur, P., Brankart, J. M. and Garnier, F.: Assimilation of chlorophyll data into a stochastic ensemble simulation for the North Atlantic Ocean, Ocean Science, 16(5), 1297–1315, doi:10.5194/os-16-1297-2020, 2020.

Skákala, J., Ford, D., Bruggeman, J., Hull, T., Kaiser, J., King, R. R., Loveday, B., Palmer, M. R., Smyth, T., Williams, C. A. J. and Ciavatta, S.: Towards a Multi-Platform Assimilative System for North Sea Biogeochemistry, Journal of Geophysical Research: Oceans, 126(4), e2020JC016649, doi:https://doi.org/10.1029/2020JC016649, 2021.

Song, H., Edwards, C. A., Moore, A. M. and Fiechter, J.: Data assimilation in a coupled physical-biogeochemical model of the California current system using an incremental lognormal 4-dimensional variational approach: Part 3—Assimilation in a realistic context using satellite and in situ observations, Ocean Modelling, 106, 159–172, doi:https://doi.org/10.1016/j.ocemod.2016.06.005, 2016.

Teruzzi, A., Dobricic, S., Solidoro, C. and Cossarini, G.: A 3-D variational assimilation scheme in coupled transport-biogeochemical models: Forecast of Mediterranean biogeochemical properties, Journal of Geophysical Research: Oceans, 119(1), 200–217, doi:10.1002/2013JC009277, 2014.

Verdy, A. and Mazloff, M. R.: A data assimilating model for estimating Southern Ocean biogeochemistry, Journal of Geophysical Research: Oceans, 122(9), 6968–6988, doi:https://doi.org/10.1002/2016JC012650, 2017.

Wang, B., Fennel, K., Yu, L. and Gordon, C.: Assessing the value of biogeochemical Argo profiles versus ocean color observations for biogeochemical model optimization in the Gulf of Mexico, Biogeosciences, 17(15), 4059–4074, doi:10.5194/bg-17-4059-2020, 2020.

Yu, L., Fennel, K., Bertino, L., Gharamti, M. El and Thompson, K. R.: Insights on multivariate updates of physical and biogeochemical ocean variables using an Ensemble Kalman Filter and an idealized model of upwelling, Ocean Modelling, 126, 13–28, doi:https://doi.org/10.1016/j.ocemod.2018.04.005, 2018.

Yu, L., Fennel, K., Wang, B., Laurent, A., Thompson, K. R. and Shay, L. K.: Evaluation of nonidentical versus identical twin approaches for observation impact assessments: an ensemble-Kalman-filter-based ocean assimilation application for the Gulf of Mexico, Ocean Science, 15(6), 1801–1814, doi:10.5194/os-15-1801-2019, 2019.