Dear Editor,

Thank you for reviewing our manuscript. We have revised the manuscript according to the Reviewer's comments. Below, you will find our point-by-point responses detailing the main changes made in the manuscript according to the Reviewer's comments
The Reviewer's comments are in blue, our responses follow each comment in black and the new rewritten version in *black italic*. Additionally, in our response, we provide the line of insertion referring to the track-changes version of the manuscript.

We would like to sincerely thank all reviewers for dedicating the time and effort necessary to review the manuscript, and provide feedback in such a constructive and useful way.

Best regards,
On behalf of all the authors,
Carolina Amadio

General comments on the new version in response to comments of Review #1, #2 and #3.

We conducted an overall review of the manuscript, focusing on improving the English language and ensuring the avoidance of speculative conclusions (as suggested by the Reviewer 2). All figures have been redone enhancing their readiness (by refining elements such as color palette, font size, etc.).

#Reviewer 1

In the Introduction we have markedly revised the text and the flow to enhance its readability. We've further added information about the availability of the BGC-Argo data [Track-changes version at P2 Lines 30-37].

Line 58: May mention here or later that these examples are time series of chlorophyll, while the use of reconstructed nitrate is novel.
Thank you. We've added more information by specifying the variable(s) involved in each cited work focused on reconstructing BGC-datasets from OC [Track-changes version at P3 Lines 83-88] and from BGC-Argo [Track-changes version at P4 Lines 89-94].

Line 72: May be worth stating what the first release included for a full account of the developments
Thank you, we have corrected as follows:
"*Starting from the first release that included OC data assimilation in the open ocean (Teruzzi et al., 2014), the assimilation has progressively developed to handle coastal OC observations (Teruzzi et al., 2018), chlorophyll and nitrate profiles from BGC Argo (Cossarini et al. 2019 and Teruzzi et al. 2021 respectively).*" [Track-changes version at P4 Lines 102-104].

Line 81: Not clear at this point what "sequential modular approach" means

As described in Buizza et al. (2022) the combination of DA and NN can be addressed by ‘fusing’ the DA and NN modules together or keeping them as independent entities. In this work, we have chosen “*allowing the flexibility of choice between different modules depending on the needs of the overall system.*” (i.e., modular approach). To make this choice clearer in the manuscript, we have corrected the paragraph as follows:
“*In this paper, the OSE experiment, which combines data assimilation and neural network in a modular approach, aims to quantify how the Argo and BGC-Argo network can be exploited. The sequential use of the NN and DA schemes provides flexibility in using one module independently of the other, depending on the needs of the overall system (Buizza et al., 2022). The DA module used in this work is the 3DVarBio data assimilation scheme described in Teruzzi et al. 2021 and updated to assimilate oxygen BGC-Argo profiles. The NN module is the NN-MLP described in Pietropolli et al., 2023 for the Mediterranean Sea (hereafter NN-MLP-MED)*” [Track-changes version at P4 Lines 112-118].

Line 87-100: The paragraph about the MedSea oceanography feels out of place and may be covered at the beginning of the first results section.
We gently disagree. We believe that a paragraph about Mediterranean Sea oceanography in the Introduction can help the reader to understand the context (semi-enclosed regional sea but with marked internal variabilities) where our combined approach is tested.

Line 114 & Figure 1: the flow of information between 3DVarBio and OGSTM-BFM is implied to be one-way, but presumably it’s two-way, with OGSTM-BFM fields also an input to 3DVarBio? Also, “3DVarBio" is used in the text, but “3D-VarBio” in the figure (and on line 116) - these should be consistent.
Thank you for noticing the typo. We have modified the arrow in Figure 1 and corrected the “*3D-VarBio”* notation.

Line 150: “preserve optimal values” - a better wording would be “preserve existing values” or “preserve background values”, there’s no guarantee they’re optimal.
Thank you, we added the following clarification:
Specifically, for the assimilation of chlorophyll, the VB operator includes a balance scheme that maintains the ratio among the phytoplankton groups and preserves the physiological status of the phytoplankton cells (i.e., preserves the internal ratios between the chlorophyll, carbon and nutrients as described in Teruzzi et al. 2014). [Track-changes version at P7 Lines 186-189].

Line 152: “spurious assimilation” - please be more specific. “spurious correlations”?
Thank you, we have changed the text to be clearer. The localization is used to limit the impact of observations on physically distant state variables to reduce spurious error correlations.
“including a localization function to avoid unrealistic corrections due to possible spurious error covariances in the deepest part of the water column” [Track-changes version at P7 Lines 191-192].

Line 154: “it barely affects other variables” - is it known how model-dependent this finding is? Since the models used here and in Skákala et al. (2022) are very similar, this is a reasonable approach to take here, but it could be worth clarifying that this lack of effect on other variables is in the model, not necessarily the real world.

Thank you for the comment. In the BFM model (which is similar to the ERSEM model used in Skákala et al., 2021), few formulations depend on oxygen concentration, for example nitrification and an oxygen regulating factor for the switch between aerobic and anaerobic conditions for bacterioplankton (Vichi et al., 2004 and 2007, Di Biagio et al., 2022). In general, for simulated values in the water column, the effect of oxygen on these dynamics is very small. On the other hand, oxygen is changed by several and contrasting processes (such as Primary production and respiration) which makes multivariate covariance including oxygen not reliable.
We have proposed the following clarification:
"VB included only a new direct relation for oxygen (i.e., oxygen assimilation updates only the oxygen itself), given that it has been shown that it barely affects other variables (Skakala et al., 2021). In the BFM model equations, few formulations depend on oxygen concentration (e.g., nitrification). Indeed, when the euphotic zone of the open ocean is well oxygenated, oxygen dynamics has a limited impact on the biogeochemical cycles." [Track-changes version at P7 Lines 193-196].

Line 161: "we decided to not use different values of error for the two nitrate subsets in order to show the highest potential impact of the OSE." A caveat needs adding either here or in the discussion that as a result of this decision, the assimilation may be non-optimal in terms of fitting the true state (as opposed to just fitting the observations). The same could be said about the lack of accounting for representation error.
Thank you. The nitrate error used in this work is as in Mignot et al., 2019 for the floats (observation error). In the Discussion, we motivated our choice of using a uniform observation error among the two nitrate subsets.
Based on the Reviewer's comment, we have discussed the representative error and possible over-fitting towards the observations. In particular: i) the nitrate error used in this work is an evolution of the one used in Teruzzi et al. (2021) with the addition of a larger error at depth to avoid inconsistency between the deeper part of the assimilated layer (0-600 m) and the lower one; and ii) the representation error was not added in this work, since the results of the previous works demonstrated a good balance between assimilation impacts and over-fitting towards the observations (instead that true state). [Track-changes version at P28 Lines 594-605].

Line 163-4: Is there a reference for the oxygen observation error values used? If not, please state how these values were chosen.
We added the reference on the dataset used to define the oxygen observation error: *Feudale et al., (2022).*

Section 2.3: While it is fine to refer the reader to Pietropolli et al. (2023) for details, it would be helpful to have a slightly longer and clearer description of the NN-MLP-MED methodology in this section.
Line 174: "The error of reconstructed nitrate, obtained by using the EMODnet as validation dataset, was 0.5 mmol m−3". As this figure contrasts with the uncertainty value of 0.87 mmol m−3 given in the previous section, a little more context would be useful. For instance, introduce the EMODnet dataset (that hasn't been done yet), state that the NN-MLP-MED method was trained on 80% of the EMODnet data, then had an RMSE of 0.5 mmol m−3 when tested against the remaining 20% of EMODnet data, and an RMSE of 0.87 mmol m−3 when the methodology is applied to BGC-Argo data that is not in EMODnet (if I have

interpreted Pietropolli et al. (2023) correctly).

According to the suggestions of the Reviewers we propose a more detailed version of the paragraph. On the other hand, we did? not specify the percentage of data used for training, since we adopted a standard approach (80% for training and 20% for testing).

Line 180: It would be useful to put the information about added reconstructed profiles into context. As a suggestion, that could be in the form of stating for each aggregated region or the sub-regions how much reconstructed data is added. Having this information about added data per region may be useful in later sections e.g. when looking at RMSE changes between the DA runs, to enable linking the change in coverage to a change in RMSE (or highlighting where this does not link for any reason).

We have added a new figure (Figure 3 Track-changes version at P10) to fill the gaps of information on the distribution of reconstructed nitrate and nitrate data.

Line 184: "Adjusted and delayed mode data were selected for oxygen and chlorophyll, while exclusively DM data were considered for nitrate." - A sentence or two explaining the reasons for these choices would be useful.

Thank you for the comment. The information we provided is not clear enough. We propose the following changes taking into account all the reviewers' comments:

"*We collected both AM and DM data for oxygen and chlorophyll. For nitrate we selected DM data, while AM data were incorporated after undergoing correction via Canyon-b NN method or using the World Ocean Atlas (WOA18) collection (Garcia et al., 2019) as explained in Johnson et al. (2021). For the three variables we use data flagged as good, probably good, changed and interpolated values (flags 1, 2, 5 and 8)".* [Track-changes version at P9 Lines:246-250].

Line 196: I may not understand the approach, but what happens if a float lives less than a year, which is when the largest drift occurs (Line 193)? Will the drift correction be applied to t0, or not because this is for operational purposes?

Thank you for the comment. If a float lives less than 1 year the trend analysis is not performed (the profile is never corrected).

"[..] Conversely, if the available float time series is less than one year, the profiles are not corrected because the float lifetime is considered too short to account for in situ sensor drift." [Track-changes version at P11 Lines:272-273].

Line 197: Please give more details about the splitting into "inliers and outliers".

Thank you for the comment. Both the RANSAC and Theil Sen methods automatically split the data into inliers and outliers. We can define some parameters of the methods (e.g., the "number of iterations", "min_sample") but not choose the inliers and outliers. Detailed information can be found at https://scikit-learn.org/stable/modules/linear_model.html#ransac-regression. We have proposed to following correction:

"Used for linear and non-linear regression problems, the RANSAC and Theil-Sen methods automatically partition the oxygen dataset into inliers and outliers. In order to avoid possible biases (Dang et al. 2008 and Fischler and Bolles 1981), these methods calculate the drift based on the data subset identified as inliers." [Track-changes version at P11 Lines:274-276].

Line 201: If drift is expected to linearly increase with depth, why use the average drift between 600 m and 800 m, rather than just the drift at 600 m? This may be reasonable (we're not experts on oxygen sensor drift), but it's not clear from the explanation.

Thank you for the comment. We chose two methods and two depths to obtain a solid basis in evaluating if a float has an oxygen sensor drift. The calculation at two depths avoids the possible fake detection of drifts because of changes in the water masses. Thus, given that the calculations have been done at the two depths we took the arbitrary, but we think more precautionary, choice to use their average. In fact, when the drift is detected (i.e., the four values have the same sign and are higher than 1 mmol/m^3 in absolute term), the two values are generally quite close. The new version of the text is deeply changed [Track-changes version at P11 Lines:265-284].

Line 216-222: This paragraph needs to be clearer, especially around the oxygen saturation procedure.

Thank you for the comment. the paragraph was deeply revised:

"At surface, the oxygen profile exclusion is evaluated by calculating the difference between the uppermost oxygen measurement and the oxygen saturation (derived from temperature and salinity data from the Argo dataset as in Garcia et al. 2019). Profiles are
excluded when this difference reaches the threshold of 10 mmol m−3. At 600 meters, the difference between oxygen and a climatological reference oxygen at depth is calculated. Profiles are excluded when the difference reaches the threshold of 2 times the standard deviation of the same reference dataset. As reference dataset, we chose the EMODnet2018_int data collection that integrates the in situ aggregated EMODnet data (Buga et al., 2018)) and the datasets listed in Lazzari et al. (2016) and Cossarini et al. (2015b). The EMODnet2018_int dataset is available for 16 sub-basins in the Mediterranean Sea (Figure 2)" [Track-changes version at P11 Lines:303-311].

Line 223-227: How were these thresholds arrived at?

According to the comments from the other reviewer we have clarified the choice of the threshold and corrected the paragraph. Additionally, we changed "misfit" with "innovation". [Track-changes version at P11 Lines:311-321].

Line 243: "After removing of drift, the deep oxygen concentrations results to be closer to the EMODnet climatological data, allowing to include a higher number of profiles" - does this mean that in the absence of the drift correction the profiles would be expected to fail QC checks and be excluded, rather than the uncorrected profiles being assimilated?

Thank you for the comment. We have not statistically analyzed it, but in most cases the correction makes our data at 600m closer to the EMODnet climatology. As a consequence, we logically have inferred that it is more likely that a profile is excluded when not corrected (drift correction).

"The removal of drift brings the oxygen concentration at 600 m closer to the EMODnet climatological data (as exemplified in Figure 4, green star). This leads us to infer that our drift correction enables the inclusion of more profiles in the assimilated oxygen datasets" [Track-changes version at P13 Lines:341-344].

Line 246-247: "While for the satellite comparison the model daily averages are considered, the model first guess (i.e. the model state before the assimilation) is used for metrics based on BGC-Argo." - This is reasonable given that BGC-Argo is assimilated and ocean colour not, but a clearer reasoning for the decision should be given. Furthermore, is the first guess instantaneous (at midnight? at the observation time?) or an average? Also, it states here that for the satellite comparison the model is a daily average, but two paragraphs later that the observations are a weekly average?

Thank you for the observation. The weekly was a typo, we actually used the daily L3 map of satellite chlorophyll from Copernicus. They are given as daily maps thus the comparison uses the model as daily output. The first guess for BGC-Argo comparison is instantaneous at 1pm (i.e. right before the assimilation). Here the new version:

"*Skill performances of the simulations listed in Table 1 are evaluated by comparing model results with satellite Copernicus300 OC product (i.e., OCEANCOLOUR_MED_BGC_L3_MY_009_143 from marine.copernicus.eu, last visited in July 2023) of chlorophyll and BGC-Argo profiles (Argo, 2022). The satellite comparison used daily model output. The model first guess (i.e., the model state at 1pm before the assimilation) is instead used for the metrics based on BGC-Argo profiles. While the use of the first guess is a common practice in DA applications (Hollingsworth et al., 1986), it is worth to remind that this comparison should be considered as a semi-independent validation, given that two consecutive profiles of the same BGC-Argo float can share a certain degree of correlation*" [Track-changes version at P13 Lines 346-354]

Line 248: RMSE has its place, including here, but could usefully be supplemented by other validation statistics. Furthermore, RMSE is only optimal for Gaussian variables, is this the case for the variables considered? If not, then more robust statistics may be preferable.

Following the comments of all reviewers, we clarified the RMSE results from Figure 5 (OLD VERSION, Figure 6 Track-changes version). We have decided to use RMSE to be consistent with product uncertainties (NN-MLP-MED or previous work such as Teruzzi et al., 2021).

Line 250: "the aggregated combination" was not mentioned before. Could be done with the description of Figure 2.

Thank you for highlighting the issue.

Following all the Reviewers comment we provided information about the names of the 16 sub-basins in the Mediterranean Sea, classifying them into eastern or western sub-basins. [Track-changes version at P9 Lines 255-262]. Additionally, we've introduced a new figure and modified Figure 4 (OLD VERSION) by inserting a vertical line to divide west and east sub-basins [Track-changes version at: P10 Figure 3 and Figure 5 P15]. Furthermore, we've described the 6 aggregated basins used in the Results Section [Track-changes version at P13 Lines 357-360].

To enhance clarity, the 16 sub-basins are defined using lowercase letters, while the 6 aggregated basins are defined by uppercase letters.

Line 281: "Assimilating oxygen profiles enable reducing the model-BGC floats RMSE" - is it possible to know how much this is due to the oxygen assimilation, and how much to the chlorophyll and nitrate assimilation? The lack of impact of reconstructed nitrate is an indicator here, but some further comment would be useful.

Thank you for highlighting this point. The reduction in oxygen RMSE is to be ascribed to the oxygen assimilation. The lack of impact of reconstructed nitrate is mainly due to the fact that reconstructed nitrates come from the oxygen dataset. It means that everytime we assimilate reconstructed nitrate we also assimilate oxygen (that directly updates and affects oxygen dynamics). The oxygen dynamics is not directly affected by chlorophyll and nitrate increments by the assimilation. [Track-changes version at P16Lines 410-412].

Section 3.3.1 may benefit from rewriting for clarity. It is difficult to pick out the key message. As a suggestion (definitely not a requirement) you may test describing the BGC differences one region at a time instead of structuring the paragraph by variable. Possibly that improves the understanding.

Thank you. We have revised Section 3.3.1. (From P18 Track-changes version) and modified the Figure 7-8-9-10.

Figure 10: Experiment names on the y axis differ from the main text. Write "npp" instead of "ppn". I think it would help to include the basin boundaries for orientation. An idea to better visualize the results may be to plot the difference in the subplots for the DA experiments compared to HIND instead of absolute values but that's not a necessary change.

We corrected the typos in Figure 10 (OLD VERSION,Figure 11 new version).

Line 340: If I understand correctly, the results thus show that nitrate suggests reduced NPP and chlorophyll enhanced NPP? Does that point to a bias in the model or representation of a specific component? (e.g. PFTs) If that's the case, it may be worth noting in the discussion.

Thank you for the comment. Following the comment of the other Reviewers, we decided to not infer any conclusions on the effect of chlorophyll assimilation on the primary production. We removed the speculative conclusion on chlorophyll assimilation at L 339-341 (OLD VERSION)

Line 345: When introducing the impact indicator, please add information about how that differs from other statistical metrics such as RMSE or a simple comparison between fields at the end of the simulation. What is the advantage of using this metric?

Thank you for the comment. We choose the same metrics used in similar biogeochemical assimilation experiments (Teruzzi et al., 2021). We acknowledge that other metrics can be used and that each of them can have strengths and disadvantages.

For example the use of RMSE metric considers only the location of the observation that can be unevenly distributed, thus providing misleading interpretation about the assimilation impact. The comparison between fields at the end of the simulation might limit the significance of the comparison given that many biogeochemical variables, such as chlorophyll, assume low values in December that are not representative of other biogeochemical conditions occurring during the year.

Thus, we think that the metric introduced in Teruzzi et al., 2021, by looking at the 95th of the distribution of the differences, can highlight the largest relative impact in each point of the domain and in different seasons considering the peculiarity of biogeochemical variables.

Line 360: Where does this threshold come from?

Thank you for the question. The threshold is the mean value of the impact indicator.

Line 368: Do you mean "initial conditions" as in using the analysis to initialise a forecast? If so, that may need clarification because it may be confused with general initial conditions for

ocean simulations. For initial conditions in a general sense the QC'd oxygen profiles may not qualify.
Thank you for the comment. We meant the initial condition for simulations. Indeed, once oxygen profiles from BGC-Argo are quality checked (official ADTM QC plus the additional QC proposed in the present work), they can represent a qualified dataset for computing ICs. As shown in Figure 2, more than 2000 profiles are available in the Mediterranean Sea for the period 2017-2018. However, some areas of the Mediterranean Sea are still undersampled by the BGC-Argo, thus it will require the integration with other in situ datasets.

Line 377: "threshold on 1mmol/m3" – can you add a value for decadal variability in the sentences before, which puts this threshold into context to illustrate it is indeed a justified choice please.
We added a value of decadal variability in the Discussion section. [Track-changes version at P21 Lines 454].

Line 399: "more than 30 profiles" - what was that before? How much larger is the data availability?
Following the other Reviewer comments, our OSE experiment shows that the basin coverage rate of nitrate can potentially be as high as the BGC-Argo equipped with an oxygen sensor. We explained better this concept as follows:
" *Through the integration of NN and DA, the count of nitrate profiles ingested can potentially be as high as the BGC Argo equipped with an oxygen sensor (i.e., more than double of the nitrate profiles), which corresponds to a density of 1 profile in each 2.5deg x 2.5deg box every 10 days for the 2017-2018 period.*" [Track-changes version at P27 Lines:573-577].

Line 401: "can effectively be constrained" is that referring to previous papers such as observing system simulation experiments? If this is meant as a conclusion from your results, this statement may need more explanation.
Thank you for the comment. The statement "can effectively be constrained" refers to our results. Indeed, by increasing the density of available observations, it was possible to achieve the seasonal temporal scale and the sub-basins spatial scale for nitrate dynamics.
We explained better this concept as follows:
"Apart from an increase in the numbers of floats, a further increase of the area impacted from a float assimilation can be achieved by redefining horizontal covariance errors in the data assimilation scheme. Indeed, benefits of non-uniform correlation radius in the horizontal scale have been previously investigated (Cossarini et al., 2019) and additional improvements could be provided by a 3D varying correlation radius (Storto et al., 2014)".
[Track-changes version at P27 Lines:580-583].

Line 436: Since ocean colour is not assimilated in this study the statement "should be used in conjunction with…" should have a reference to literature
As discussed in the previous comment, we revised this part of the discussion. [Track-changes version at P27 Lines:562-572].

#Reviewer 2

Specific comments:

P6.l146: "VH is built using a Gaussian filter whose correlation radius modulates the smoothing intensity"

> What is the size of correlation radius in average? This information is important to understand how far BGC Argo profile assimilation leave impact in the analysis.

Thank you for the suggestion. We've added the information on the correlation radius. "*The correlation radius ranges between 12-20 km".* [Track-changes version at P7 Lines:184-185].

P7.l184: "Adjusted and delayed mode data were selected for oxygen and chlorophyll".

> Can you describe which QC flag was used for selecting "good" data both for oxygen and chlorophyll?

We've revised the text to include essential details about our criteria for selecting BGC profiles based on their data mode and data flag.

"*We collected both AM and DM data for oxygen and chlorophyll. For nitrate we selected DM data, while AM data were incorporated after undergoing correction via Canyon-b NN method or using the World Ocean Atlas (WOA18) collection (Garcia et al., 2019) as explained in Johnson et al. (2021). For the three variables we use data flagged as good, probably good, changed and interpolated values (flags 1, 2, 5 and 8)".* [Track-changes version at P9 Lines:246-250].

P8.l199: **"**when all four drift estimates agree in sign"

> Not clear what do you refer here by "four drift estimates". In P7.l196, it is mentioned that drift is evaluated at two different depths and what are the rest of two estimates? Or the number of four has nothing to do with that?

We chose two methods (RANSAC and Theil Sen) and two depths (600 and 800 m).

"In our approach, the presence of a drift is established when all four drift estimates (RANSAC at 600 and 800 m, Theil-Sen at 600 and 800 m) agree in sign and their average value (D_avg) exceeds 1 mmol m−3 y" . [Track-changes version at P11 Lines:278-280].

P10.l246-247: "While for the satellite comparison the model daily averages .. the model first guess is used for metrics based on BGC-Argo"

> I believe the choice of these different RMSE metrices between satellite OC data and Argo profiles are based on the experiment settings of Argo profiles being assimilated while satellite OC data are not assimilated. By choosing the first guess state to be compared with not yet assimilated Argo profiles, you can use the Argo profiles as independent data. If it were the case, better to describe so here.

P10.l251: "Satellite L3 products from Copernicus Marine Service catalogue .."

> Usage of this data set requires proper citation. Plus, this sentence is floating without clear connection in 3.2. Does it mean satellite OC RMSE metric is based on this weekly averaged data? If so, does weekly cycle coincide with an analysis cycle? Please make its significance clear.

Thank you for your comments. Considering the suggestions from other Reviewers regarding this point, we've extensively revised this section, merging all comments and deeply revising this part from the old version of the manuscript. Here the new version:

"*Skill performances of the simulations listed in Table 1 are evaluated by comparing model results with satellite Copernicus300 OC product (i.e.,* *OCEANCOLOUR_MED_BGC_L3_MY_009_143 from marine.copernicus.eu, last visited in July 2023) of chlorophyll and BGC-Argo profiles (Argo, 2022). The satellite comparison used daily model output. The model first guess (i.e., the model state at 1pm before the assimilation) is instead used for the metrics based on BGC-Argo profiles. While the use of the first guess is a common practice in DA applications (Hollingsworth et al., 1986), it is worth to remind that this comparison should be considered as a semi-independent validation, given that two consecutive profiles of the same BGC-Argo float can share a certain degree of correlation*" [Track-changes version at P13 Lines 346-354]

P11.l274-l275: "Here, improvements related to chlorophyll assimilation can be observed in nwm, ion and lev in winter and at depth in tyr, ion and lev in summer (Figure 5 middle panel)"
P11.l278: "the direct chlorophyll assimilation is more effective than .."

> Since there is no experiment with only assimilating chlorophyll in this study, it is not easy to point out degree of "improvements related to chlorophyll assimilation" and if direct chlorophyll assimilation is more effective than the dynamical model adjustment after nitrate assimilation. You need to provide extra analysis to support these statements.

We do agree with the Reviewer that our text was in part speculative. The objective of the present work is to assess the impact of the addition of extra nitrate profiles from NN (DAnn). Thus, our present results show that there is no further improvement of RMSE of chlorophyll after DAnn with respect to the improvement shown by the DAfl run. The discussion about the relative impacts of nitrate vs chlorophyll profile assimilation have been already assessed in Cossarini et al., 2019 and Teruzzi et al., 2021. We removed the text of all the speculative conclusions.

P13. "3.3.1 Impacts on biogeochemical vertical dynamics".

> There is no description on how figures 6, 7, 8 and 9 are plotted. Are they sub-basin averaged value? "the basin wide averages of DAnn display .. (Figure 6)" at P13.l310 infers these figures are basin-average, but it is never be stated clearly.

P13.l297: "Nitracline depth"

> There is no definition of nitracline depth. Please be specific.

Section 3.3.1 has been completely rewritten, including definitions for phosphocline and nitracline. In the Hovmoeller figures, we plotted averaged values over two selected sub-basins (NMW and Ion2) and the entire Mediterranean Sea. Additionally, we've added a brief explanation for Figure 7-10. [Track-changes version at P13-14].

P13.l296-l297: "decreases by 8% and 11% in DAfl and DAnn runs, respectively"

> Contrary to the clear difference in impact of nitrate assimilation in DAfl and Dann at nwm in Figure 7, RMSE profiles in Figure 5 (especially Summer Nitrate at Nwm) does not show such difference in the two DA experiments. Can you explain why?

Figure 7 (OLD VERSION) is plotted using the model average daily outputs computed over the whole area of the sub-basin, while Figure 5 (OLD VERSION) reports the RMSE statistics computed on model background values of the grid points corresponding to the locations of BGC-Argo nitrate profiles.

"This representation offers additional details on the vertical impact of the reconstructed nitrate profile assimilation with respect to the validation of Figure 6 that considers only model points corresponding to the location of BGC-Argo profiles. Nwm and ion2 represent distinct trophic conditions in the Mediterranean Sea and are also characterized by high number of

assimilated reconstructed nitrate profiles (Figure 3)." [Track-changes version at P18 Lines 419-423].

P13.l299: "eventually reach a stationary phase"
> What does it mean by a stationary phase and how do you measure it?

The term stationary was misleading. We refer to the fact that during the second year, the rate of accumulation of DA impact decreases or is quite null. [Track-changes version at P18 Lines 437-439].

P13.l306: "As a consequence of both the direct assimilation of chlorophyll profiles and the dynamical model adjustment after nitrate assimilation"
> Again, how can you argue a consequence of dynamical model adjustment only from nitrate assimilation with this OSE settings? For example, why would phytoplankton biomass change as a consequence of direct assimilation of chlorophyll not affect chlorophyll concentration in the DCM as a consequence of its dynamical model adjustment? If extra material not provided, this statement is speculative.

Thank you for the comment. The second Hovmoeller of Figure 6 (OLD VERSION) shows the difference between DAfl and HIND runs. This difference is due to the direct impact of chlorophyll assimilation and indirect effects (e.g., model dynamical adjustments) after nitrate assimilation. It is correct that we can not distinguish between the two runs (DAfl and DAnn), however the original sentence was meant to highlight the difference between DAfl and HIND. "Considering chlorophyll (Figure 9), the main difference between DAfl and HIND is a slight reduction of the DCM chlorophyll concentration (e.g., variation smaller than 5% with respect to HIND simulation) and a correction of the timing of the surface winter blooms (second row in Figure 9)". [Track-changes version at P18 Lines 449-451].

P13.l313: "oxygen profiles assimilation (DAfl, second row in Figure 9) provides positive or negative corrections"
> As is described by authors in the subsequent sentences, changes in phytoplankton biomass also change oxygen through primary production and remineralization process as dynamical model adjustment. Thus, assimilation of chlorophyll and nitrate both have a potential to alter oxygen. How can you judge what can be seen in Figure 9 is sole consequence of oxygen assimilation? This sentence contradicts with statements following about impact if reconstructed nitrate profile assimilation in oxygen.

We do agree with the reviewer. By comparing DAfl and HIND we can only provide an assessment of the overall impact on oxygen of the multivariate (nitrate, chlorophyll and oxygen). We've made the correction as follows:
"Corrections on oxygen dynamics after the multivariate assimilation (DAfl, second row in Figure 9) are either positive or negative depending on the area and the period of the year. In particular, corrections are mostly positive in ion2, while the NWM sub-basin shows negative corrections in the subsurface layer and positive ones in the upper layer of the second year.
[Track-changes version at P19 Lines *456-460*].

P13.l316-l318: "The only noticeable difference .. > This is one of the most important findings in this study as an impact of reconstructed Nitrate profile assimilation, but difference between DAfl and DAnn in figure 9 (summer period in NWM) can not be found in RMSE profiles in figure 5 (summer Oxygen in Nwm) and we can not judge if this difference in DAnn against DAfl is improvement or not. Can you explain why?

Thank you for the comment. Figure 9 (OLD VERSION) is plotted using the model's average daily outputs computed over the whole area of a given sub-basin, while Figure 5 (OLD VERSION) reports the RMSE statistics computed on model background values of the grid points corresponding to the locations of BGC-Argo oxygen profiles. Given that oxygen profiles are assimilated in DAfl, the positive message is that the assimilation of extra nitrate profiles does not degrade the quality of DAfl (with oxygen assimilation). It is reasonable to expect that DAnn can perform at maximum as good as DAfl with respect to oxygen. On the other hand, Figure 9 (OLD VERSION) shows an additional detail: nitrate extra profile assimilation can provide changes to the model dynamics that is wider than oxygen one. This change can impact oxygen in areas distant from the location of oxygen BGC-Argo profiles. The text is changed following all reviewers' comments as for the comment above (see above comment labeled as P13.l313 comment).

P16 entire section of 3.3.2

> Since difference between DAfl and DAnn is almost impossible to see in Figure 10, readers can not confirm what is described in this subsection. Please reevaluate how to present different impact of DA settings in NPP.

Following the suggestions of all Reviewers we corrected the readability of Figure 10 (OLD VERSION). We modified and added a contour line in black in the third row to highlight the areas mainly impacted by the different assimilative setup where the difference in Net Primary Production (NPP) between Dann and DAfl exceeds 15 mgC m−2 d−1. [Track-changes version at P23 Figure 11]

P16.l339-l341: “In fact … after chlorophyll assimilation”

How can you measure that weak negative correction of macronutrients is the main cause of reduced NPP outweighing the effect due to change in phytoplankton biomass after chlorophyll assimilation? As far as I read, there is no concrete material supporting this statement is provided. Unless extra material provided, this statement is speculative.

Thanks for the suggestion. We've removed the speculative conclusion regarding chlorophyll assimilation

P17. 3.3.3

> In figure 11, figure title indicates Nitrate Iij(t) is evaluated over 0-600m depth range rather than 0-300m depth specified in equation (2). If it were the case, please specify so. If not, please fix the figure titles in figure 11.

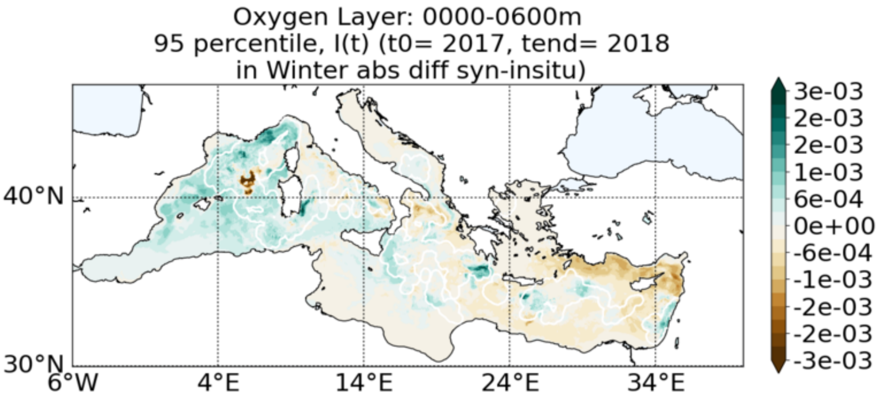
Thank you for identifying the typo. We've utilized 0-300m and 0-600m layers for chlorophyll and nitrate, respectively. Additionally, we've replaced the subscript "300" with the more general term "maxdepth" in Equation 2.

P17.l365: “since the same QC oxygen dataset was assimilated in DAfl and DAnn”

> But the authors just described in P13.l316-l318 that impact of the reconstructed Nitrate is noticeable in oxygen at least at NWM where density of the reconstructed Nitrate is large. Then it does not make sense that you do not see difference in the two DA experiments. Why do you not see the difference in the Iij 95th percentiles maps for oxygen?

Thank you for your comment. We observe low differences (10^{-3}) mainly in the NWM, between the two DA experiments (DAfl and DAnn) maps. Since the “*oxygen assimilation updates only the oxygen itself*” (*L 154 OLD VERSION*) these differences are due to the reconstructed profiles of nitrate that cause a decrease of productivity, a loss of oxygen

production and a loss of remineralization.
Moreover, in our DA scheme, the oxygen dynamics is not directly affected by chlorophyll and nitrate increments by the assimilation. This explains the relatively minimal differences observed between the oxygen impact maps of DAfl and DAnn. In the updated manuscript, we've expanded on this topic. For sake of clarity, we also provide hereafter the differences between the Impact indicator for DAfl and DAnn of oxygen:



PP19.l396-l.397: "In this work, important impacts are also observed in summer for all variables, as a consequence of the increased number of assimilated profiles."
> It is not clear what does it mean by "a consequence of the increased number of assimilated profiles". Increased number of nitrate profile from DAfl to DAnn? Or about something else? As far as I understand, main reason why we see impact of DA in summer in DA experiments in this study compared to Teruzzi et al. (2021) is because satellite OC can not see DCM while Argo float profiles see the signal by multiple sensors. In that sense, you could see the impact of Argo profile assimilation no matter how small or large number of profiles is. Please reevaluate this statement.
Thank you for your comments. By comparing the impact indicator maps for DAfl and DAnn, we illustrate the potential additional benefit of incorporating extra nitrate profiles in a multiplatform data assimilation simulation. We've clarified this concept as follows:
"*Previous findings (Teruzzi et al., 2021) have primarily demonstrated the efficiency of ocean colour assimilation in constraining chlorophyll dynamics especially during winter and the advantages of assimilating BGC-Argo profiles in summer. Our work highlights the larger and more extensive benefits of profile assimilation during summer due to the incorporation of reconstructed nitrate profiles.*" [Track-changes version at P23 Lines:568-572]

PP19.l399-l400: "Indeed .. box every 10 days"
> I do not understand which "results" in this study support this statement. Basin coverage rate of BGC-Argo floats equipped with oxygen sensors is simply determined by deployment plan. Or do you like to say that the new O2 QC module prove enough number of O2 profile survives to be ingested to nn module? I read 3.1, but could not get such information. Please be clearer about meaning of this statement.
Our aim here is to highlight that the OSE experiment shows that the basin coverage rate of nitrate can potentially be as high as the BGC-Argo equipped with an oxygen sensor. Considering also comments of the other Reviewer, we revised this concept:
" *Through the integration of NN and DA, the count of nitrate profiles ingested can potentially be as high as the BGC Argo equipped with an oxygen sensor (i.e., more than double of the*

nitrate profiles), which corresponds to a density of 1 profile in each 2.5deg x 2.5deg box every 10 days for the 2017-2018 period." [Track-changes version at P27 Lines:573-577]

PP19.l401-l406: "while, up to … by a 3D varying correlation radius (Storto et al., 2014)"> This discussion on improvement in meso-scale dynamics look out of topic and I can not see the reason why it is needed to be discussed here. Especially confusing knowing that 2.5 degree by 2.5 degree horizontal resolution in BGC profiles potentially could be achieved by nn with oxygen profile is far below meso-scale resolving resolution of o (50km).
Thank you for the comment. By redefining horizontal covariance error we can only increase the spatial area in which each float has an impact. We've rephrased this concept, considering all the review comments:
"Apart from an increase in the numbers of floats, a further increase of the area impacted from a float assimilation can be achieved by redefining horizontal covariance errors in the data assimilation scheme. Indeed, benefits of non-uniform correlation radius in the horizontal scale have been previously investigated (Cossarini et al., 2019) and additional improvements could be provided by a 3D varying correlation radius (Storto et al., 2014)". [Track-changes version at P27 Lines:580-583].

PP19.l423-l.429: "Indeed … Li et al.(2021)»
MLP base Sauzède et al. (2017) overcame of this issue by adding pressure as input variables in MLP. Why do you believe choosing other NN approach such as 1D CNN is important before using pressure or depth information in MLP-NN-MED?
As shown in Pietropolli et al., 2023 (GMDhttps://doi.org/10.5194/egusphere-2023-1876), MLP does not explicitly consider that close points in a profile share information (the back propagation during the training treats two close points in a profile as not-correlated values of the target variable). As a result, a profile reconstructed with MLP and T, S, O2 and pressure input from BGC-Argo can show discontinuities that need to be filtered with additional steps in the procedure (see line L176-179 OLD VERSION). This potential pitfall is overcome by 1D convolutional NN, which learns explicitly the shape of the vertical profiles during the training, thus exploiting the fact that each point of a profile shares information with its neighbors.
In Pietropolli et al., 2023 there is also a comparison between vertical profiles predicted through MLP-NN-MED and PPCon, which is the proposed 1D CNN approach. Results demonstrate that changing the architecture leads to more smooth profils, which better approximate the original sampled vertical profiles.

Technical corrections:
P7.l182: 2.4 BGC-Argo data and post-deployment oxygen quality control
> I assume subsection 2.4 is about QC-O2 module, but the module name is never referred to in this section but found in the next section, 2.5. Please make it clear that this is about QC-O2.
We corrected the title of the Section as follows: " *2.4 BGC-Argo data and the post-deployment QC O2 module"* [Track-changes version at P9].

P10.l248: "is evaluated in winter (from February to April, FMA) and summer (from June to August, JJA)"
> Since your experiment period is two years from Jan 2017 to Dec 2018, do you use both 2017 and 2018 results for this evaluation?
Yes, we used both 2017 and 2018 results. Following all the reviewers' comments, we revised

this paragraph and added this information as well. [Track-changes version at P13 Lines 354-355].

P10.l255. “the eastern sub-basins”

> Please define which sub-basins (lev1, lev2,…etc) are included in the definition of the eastern sub-basins.

P8.Figure 2 caption: “lev=lev1+lev2+lev3+lev4; ion=ion1+ion2+ion3; tyr=tyr1+tyr2; adr=adr1+adr2; swm=swm1+swm2”
P10.l257: “alb, swm and nwm”
P11.l263: ” Alboran, South West Mediterranean, North West Mediterranean, Tyrrhenian, Ionian and Levantine Seas”
P11.l271: ”is observed in nwm and tyr (winter) and in ion (summer).”
P11.l275: “in nwm, ion and lev in winter and at depth in tyr, ion and lev in summer”

> Association of long and short names of each sub-basin such as Alboran (alb), South West Mediterranean (swm) etc. is never clearly defined in this article. Please do in section 2.3 or add extra table to do so.

We provided information about the names of the 16 sub-basins in the Mediterranean Sea, classifying them into eastern or western sub-basins. [Track-changes version at P9 Lines 255-262]. Additionally, we've introduced a new figure and modified Figure 4 (OLD VERSION) by inserting a vertical line to divide west and east sub-basins [Track-changes version at: P10 Figure 3 and Figure 5 P15]. Furthermore, we've described the 6 aggregated basins used in the Results Section [Track-changes version at P13 Lines 357-360].
To enhance clarity, the 16 sub-basins are defined using lowercase letters, while the 6 aggregated basins are defined by uppercase letters.

P19.l367-l.389: Five paragraphs about oxygen QC.

> This information do not fit to “Discussion”, but rather should be integrated to 2.4.

We have reduced the discussion about QC O2.

#Reviewer 3

Section 2.3 I agree with Reviewer 1 that details about the neural net approach are missing. Especially the sentence "incorporating nonlinear
functions, adjusting neuron count, and optimizing the training algorithm" needs to be expanded, since we could wrongly understand that the Fourier et al. approach does not incorporate nonlinear functions (while in reality, they use the nonlinear sigmoid function).
Thank you for the comment. We've revised the entire section [Track-changes version at P8 Line 210].

L246 "the model first guess" does it correspond to the background?
- About the assimilation: how frequent is the assimilation update? Is it 10 days?
Yes, the first guess is the background. It is the state of the system before the assimilation. Given that BGC-Argo floats have a profiling (or measurement) frequency of nearly 5 days, the first guess corresponds to the 5-day predictions in the local areas around the location of a given profile. Considering also other comments, the text at old L246 is changed as follows:
“*The satellite comparison used daily model output. The model first guess (i.e., the model state at 1pm before the assimilation) is instead used for the metrics based on BGC-Argo profiles*” [Track-changes version at P13 Lines 349-350].

Furthemore, the information on the BGC-Argo floats profiling frequency:
“ [..] *a generally 5-day temporal sampling frequency. Higher sampling frequencies (< 5 days) are registered for the 20% of profiles*.” [Track-changes version at P9 Lines 264-265]

About the validation: Can you comment a bit on the choice of using the RMSE between BGC-Argo profile and model first guess as a validation. Since a previous measurement of a BGC-Argo profile was already assimilated, can a new measurement be considered independent? It could be interesting to have a quick discussion about the lagrangian autocorrelation…

Thanks for the comment. Due to the lack of independent in situ data, our validation has used the common practice of comparing the first guess with assimilated observations (Hollingsworth, et al., 1986). We introduced the information in the new version as follows:
“*Skill performances of the simulations listed in Table 1 are evaluated by comparing model results with satellite Copernicus300 OC product (i.e., OCEANCOLOUR_MED_BGC_L3_MY_009_143 from marine.copernicus.eu, last visited in July 2023) of chlorophyll and BGC-Argo profiles (Argo, 2022). The satellite comparison used daily model output. The model first guess (i.e., the model state at 1pm before the assimilation) is instead used for the metrics based on BGC-Argo profiles. While the use of the first guess is a common practice in DA applications (Hollingsworth et al., 1986), it is worth to remind that this comparison should be considered as a semi-independent validation, given that two consecutive profiles of the same BGC-Argo float can share a certain degree of correlation*” [Track-changes version at P13 Lines 346-354]