#REV 1

Please note that this is a co-review by an early-career and mid-career scientist.

This paper details new processing of BGC-Argo data in the Mediterranean Sea, including oxygen sensor drift correction and the use of a neural network to reconstruct nitrate from other measured variables. An existing data assimilation scheme, previously used to assimilate chlorophyll and (non-reconstructed) nitrate from BGC-Argo, is extended to also assimilate reconstructed nitrate and measured oxygen profiles. Test runs demonstrate a positive impact on model analyses of assimilating these new variables.

The study is novel, of interest to the community, and within scope for Ocean Science. The study is generally well-conceived and well-presented, but there are aspects which should be more clearly explained.

In many places, the manuscript is hard to follow and would benefit from being made clearer. Some specific examples are given in the comments below, but are not exhaustive. As a general example, the use of passive voice, in particular in the methods section, makes it challenging in some parts to distinguish your work from previous studies. As another example, in the introduction the topics that should be introduced are introduced, but the text lacks flow and links between the topics, and so does not lead to the question you are addressing.
The manuscript would also benefit from English language copy editing, but we believe the journal offers this service as standard, so will not list such technical corrections as part of this review.

This paper details new processing of BGC-Argo data in the Mediterranean Sea, including oxygen sensor drift correction and the use of a neural network to reconstruct nitrate from other measured variables. An existing data assimilation scheme, previously used to assimilate chlorophyll and (non-reconstructed) nitrate from BGC-Argo, is extended to also assimilate reconstructed nitrate and measured oxygen profiles. Test runs demonstrate a positive impact on model analyses of assimilating these new variables.

The study is novel, of interest to the community, and within scope for Ocean Science. The study is generally well-conceived and well-presented, but there are aspects which should be more clearly explained.
We appreciate the constructive comments and suggestions from the Reviewer. We present our point-by-point responses to the Reviewer’s comments below. The Reviewer's comments are in blue, our responses follow each comment in black. In each response, we detail the changes we propose to make to the manuscript and include the proposed modified text and/or figure (in red).
For clarity, we have numbered some of the reviewers' comments so that similar ones are aggregated to provide a single response. Comments are labeled and highlighted with specific colors to distinguish reviewers (e.g. Rev. 1: **comment1a**, Rev. 2: **comment1b,** Rev. 3: → **comment1c**).

In many places, the manuscript is hard to follow and would benefit from being made clearer. Some specific examples are given in the comments below, but are not exhaustive. As a general example, the use of passive voice, in particular in the methods section, makes it challenging in some parts to distinguish your work from previous studies. As another example, in the introduction the topics that should be introduced are introduced, but the text lacks flow and links between the topics, and so does not lead to the question you are addressing.
Thank you for the comment. We will revise the manuscript having particular attention to the use of the passive voice.

comment1a
As another example, in the introduction the topics that should be introduced are introduced, but the text lacks flow and links between the topics, and so does not lead to the question you are addressing.
We will revise the introduction to fill the lack of links between paragraphs.
Some examples are in **comment2a, comment4a, comment5a, comment8a, comment10a.**

The manuscript would also benefit from English language copy editing, but we believe the journal offers this service as standard, so will not list such technical corrections as part of this review.
Thank you for the comment. We will review the manuscript paying more attention to the English language.

comment2a
The paper aims to “address availability gaps” (Line 4) but this objective is not clear throughout the paper. The introduction does not clarify how, where, or why the data gaps affect the analysis. Results of the two model runs with and without reconstructed observations clearly show differences, but these are not always linked to the change in coverage.
The 'data availability gap' will be explained more explicitly in the introduction, as proposed hereafter (in bold the new text).
In addition, we will include information on the seasonal availability of the observations that will help to comment on the results throughout the manuscript considering the BGC-Argo data availability.
L29: Among the BGC sensors, oxygen (O2) is currently the most common measured variable with approximately 270,000 profiles worldwide (**as of July 2023-07**), which is double that of suspended particles and chlorophyll and more than four times those of nitrate, downwelling irradiance, and pH (https://biogeochemical-argo.org).
Since 2019, the availability of nitrate and chlorophyll profiles has been gradually decreasing due to the high cost of the sensor (Dall’Olmo personal communication). The number of oxygen profiles instead decreased initially (2019-2022), but since 2022 is stable or slightly increasing. In the future, Argo Italy envisages mounting oxygen sensors on all Argo floats in the Mediterranean Sea (Discussion in the workshop on “Copernicus Marine requirements for the In-situ Observing System”, 14-15 September 2023).

comment3a

Also, the abstract and discussion conclude with a note about Argo data being complementary to satellite ocean colour assimilation, but this study does not show that.

The complementarity with satellite assimilation will be removed from the abstract but we will keep the topic in the discussion modifying the sentence as follows (**comment25a**):

Previous works demonstrated the complementary of vertical profiles with respect to satellite ocean colour assimilation (Cossarini et al. 2019, and Ford et al., 2021, Skákala et al., 2021 and Teruzzi et al., 2021).

Line 20: “Array for Real-time Geostrophic Oceanography” - this acronym form does not appear to be widely or currently used, suggest just using the name “Argo”.

Thanks for the comment. We will remove the acronym form.

Line 29: “approximately 270,000 profiles worldwide until now” - better to put “as of [date]” rather than “until now”.

OK, we will insert the date as follows (in bold):

Among the BGC sensors, oxygen (O2) is currently the most common measured variable, with approximately 270,000 profiles worldwide **(as of July 2023-07)**, which is double that of suspended particles and chlorophyll and more than four times those of nitrate, downwelling irradiance, and pH (https://biogeochemical-argo.org).

comment4a

Line 39: “By improving the accuracy…” is the result of the QC and would therefore fit better at the end of the sentence to increase clarity.

Line 45: “encouraged” replace with “is necessary”

Considering all the comments on the Introduction section (e.g., **comment1a**) we propose the following modified version of the paragraph that does no longer include “By improving the accuracy”. Moreover, the term “encouraged” is replaced in order to highlight the final aim of the oxygen QC method:

When the sensor drift exists, it is higher in the storage, out of the water, than during the deployment. As described in Takeshita et al. (2013) and in Maurer et al. (2021), raw oxygen data from floats can have errors of up to 20% in terms of oxygen saturation (at the surface) due to sensor drift during the storage. This drift is generally corrected by multiplying the oxygen concentrations for a gain factor term that is derived from a reference dataset (Johnson et al., 2015). Despite this correction can improve the accuracy up to 5-10%, Maurer et al. (2021) and Bushinsky et al. (2016) found a drift in about 25% (with a mean of -0.07% per year, a standard deviation of 0.65%, and a total range of 1.1 to 1.2% per year) and 70% of the analyzed floats, respectively. Given the logistical challenges in recovering deployed floats, an in situ (or during deployment) drift >1% per year can be likely observed (Bushinsky et al., 2016). Here the drift can be both positive or negative as found in Johnson and Claustre (2016) and Bittig et al. (2018b).

The development and dissemination of a post-deployment oxygen QC aims to avoid spurious results (Wang et al., 2020) and to distinguish between ocean signals or trends (e.g., deoxygenation) and potential drifts. This allows to obtain more robust datasets suitable for specific numerical modelling applications.

Line 46: “optode” was not mentioned before

We will solve the comment by introducing the meaning of optode (before line 46) as follows (in bold):

in line 36:

The implementation of O2 QC is mainly devoted to improving the long-term reliability and accuracy of autonomous measurements (Sauzède et al., 2017) **in particular with respect to the sensor drift (the optode drift).**

comment5a

Line 47: Suggestion to link the topics for better flow of the text: say that purely observation-based Argo studies are regional, and using data assimilation has the potential to create a synthesis

Thank you for the suggestion. We propose the following version to better link the paragraphs (in bold the new insertion). The concept of “using data assimilation has the potential to create a synthesis” will be introduced later (L50, see **comment4a and comment6a**).

The development and dissemination of a post-deployment oxygen QC aims to avoid spurious results (Wang et al.,2020) and to distinguish between ocean signals or trends (e.g., deoxygenation) and potential drifts. **This allows to obtain more robust datasets suitable for specific numerical modelling applications.**

Aiming at optimally combining observations and model information to obtain a closer description of reality, DA underpins decades of progress in ocean prediction (Geer, 2021).

comment6a

Line 50: “DA underpins …” - suggest rephrasing this sentence slightly to articulate more clearly the aims and principles of DA.

Thank you, we propose the following correction:

Aiming at optimally combining observations and model information to obtain a closer description of reality, DA underpins decades of progress in ocean prediction (Geer, 2021). **In on one hand, progresses began with an increase in the number of available observations over the past decade (number of measured variables and number of observations) on the other hand DA scheme were progressively updated to be able to perform multivariate and multiplatform assimilation, retrieve associated uncertainty into a prediction model, and solving problems connected to uneven distribution and scarcity of the observations (Buizza et al., 2022).**

comment7a

Line 54: “NN algorithms” - “NN” hasn’t been defined yet in the main text. Need to add at least a couple of sentences introducing neural networks at the start of this paragraph.
Line 54: “match specific DA tasks” - a better phrasing might be something like “have the potential to perform specific tasks related to observation processing and DA”. The discussion of the following studies could be made clearer too, as well as stating that it is the method of Pietropolli et al. (2023) that is used in this study.
We will add the extended name and introduce the acronym NN. We will propose the following correction (in bold) to solve both comments referring to L54 (merged above). Your suggestion “*The discussion of the following studies could be made clearer too*”, is further developed in the next comment for lines 57-66 (**comment8a**), however, we prefer to do not add information on the works cited at L. 55-56, since here the aim is to list some examples of NN applications in DA.
Finally, the suggestion “stating that it is the method of Pietropolli et al. (2023) that is used in this study” is fulfilled in a later comment (**comment9a**)
In recent years, neural network (NN) algorithms have been increasingly used to solve and analyze specific tasks related to observation processing and DA. The main strength of NN algorithms lies in their ability to approximate continuous functions (Hornik et al., 1989) in remarkably low computational times. For these reasons, DA techniques have been recently augmented with NN-based tools, e.g., for: bias correction (Kumar et al. 2015 and Zhou et al. 2021), cross calibration (Lary et al., 2018), reformulation of observation operators (Storto et al., 2021) and new product creation or dataset reconstruction (Lary et al., 2018).

comment8a
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 propose the following corrections (also on the basis of the previous **comment7a**). The new insertions are in bold:
As an example, ocean color (OC) datasets were employed to test Multi-Layer Perceptrons (MLP, namely the most common NN) by retrieving past and long-term BGC time series **of phytoplankton and chlorophyll** (Martinez et al. 2020a, Martinez et al. 2020b, Roussillon et al. 2023). Moreover, in Sauzède et al. (2016), MLP serves to infer **chlorophyll** vertical BGC distribution from OC. High performance in predicting biogeochemical states (e.g., oxygen) from physical profiling floats measurements were achieved in Stanev et al. (2022) for the Black Sea.
In Sauzède et al. (2017), an [........] input data.
A further update of the application of the MLP method in the Mediterranean Sea is provided in Pietropolli et al. (2023), by achieving a lower error in the prediction of **nutrients** through a larger training dataset, a hyperparameter refinement and a two-step quality control of the input data. **Given its potential in predicting nutrient profiles, the MLP-NN model provides a valuable dataset to be used to fill the gap in the availability of in situ observations in data assimilation.**

Line 72: May be worth stating what the first release included for a full account of the developments
Thank you, we will correct as follows:

Starting from the first release that included OC data assimilation in the open ocean (Teruzzi et al., 2014), the assimilation has been progressively developed to handle coastal OC observations (Teruzzi et al., 2018), and chlorophyll and nitrate profiles from BGC-Argo (Cossarini et al. 2019 and Teruzzi et al. 2021, respectively).

comment9a

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 propose to correct the paragraph as follows (new insertions in bold):

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 3D-VarBio 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**).

comment10a

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.

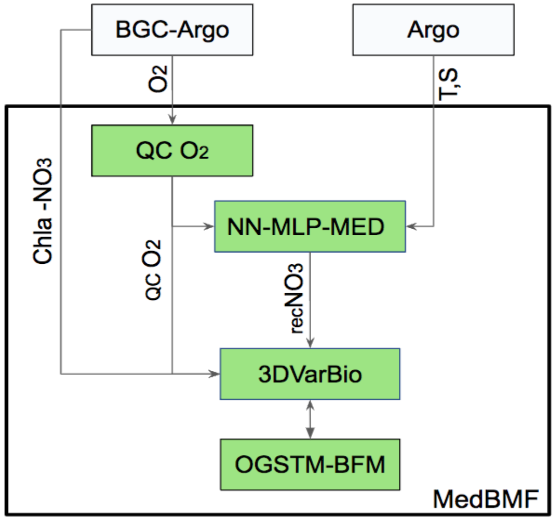
We gently disagree. We believe that a paragraph about Mediterranean Sea oceanography in the Introduction can help the reader to understand the context (semi-enclosed regional sea but with marked internal variabilities) where our combined approach is tested. However, also following a previous comment (**comment1a**) we will change the paragraph to increase its flow as follows.

Spatial and temporal impacts of the OSE are evaluated using classic and new skill performance metrics in three two-year (2017-2018) numerical experiments performed using the MedBFM coupled with the 3DVarBio: a control run (HIND) without assimilation; a multivariate run (DAfl) with assimilation of BGC-Argo chlorophyll, nitrate, and oxygen; and a multivariate run that also assimilates in situ observations and reconstructed ones (DAnn). **Given its particular characteristics and the density of BGC-Argo profiles, the Mediterranean Sea represents an ideal site for OSE experiments to evaluate the potentiality of the BGC-profiles assimilation.**

The Mediterranean Sea is an anti-estuarine…

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 propose the following clarification (in bold):

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., preserve the internal ratios between the chlorophyll, carbon and nutrients as described in Teruzzi et al., 2014**).

Line 152: “spurious assimilation” - please be more specific. “spurious correlations”?

Thank you, we will change 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.**

comment11a

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 propose the following clarification (in bold the new text):
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 oxygen 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.**

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 (L. 418-422), we discussed the choice of using a uniform observation error among the two nitrate subsets. Based on the Reviewer’s comment, we will discuss 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).

Line 163-4: Is there a reference for the oxygen observation error values used? If not, please state how these values were chosen.
We will add the reference on the oxygen observation error:
Observation error for oxygen is set to 5 mmol m^{-3} in the upper 200 meters of depth and gradually goes to 20 mmol m^{-3} in correspondence of the maximum assimilation depth. These values correspond to the uncertainty associated with the oxygen dataset described in Feudale et al., (2022).

comment12a)
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 (e.g., **comment1c**), we propose the following more detailed version of the paragraph (in bold the new text). On the other hand, we will not specify the percentage of data used for training, since we adopted a standard approach (80% for training and 20% for testing)

The NN-MLP-MED (Pietropolli et al., 2023) is the evolution of previous MLP architectures developed to predict low-sampled variables (e.g., nutrients) starting from high-sampled ones (e.g., temperature) (Sauzède et al. 2017, Bittig et al. 2018c, and Fourrier et al. 2020). **NN-MLP-MED is a deterministic Feed-Forward Neural Network based on a MLP structure. The NN-MLP-MED consists of the merging of 10 different MLP architectures, each one with the same input and output features, composed by the same number of hidden layers (i.e., 2), but composed by a different number of neurons per layer. The final prediction resulting from the NN-MLP-MED is the mean of all the predictions of these components.**

The data flow of this MLP-based approach follows the forward direction from the input to the output layers through the neurons which composed the layers.

In our OSE experiment, the trained NN-MLP-MED reconstructs nitrate profiles (output) from temperature and salinity (Argo), oxygen (BGC-Argo) and float date, latitude and longitude (inputs).

The NN-MLP-MED model presents some novel elements with respect to the mentioned methods (and in particular with respect to Canyon-Med in Fourrier et al. 2020). Firstly, the input dataset includes a larger sample size and wider coverage of the Mediterranean Sea region, **i.e., the quality controlled EMODnet2018_int data collection which integrates the in situ aggregated EMODnet data (Buga et al., 2018) and direct observations (i.e., campaigns) as in Lazzari et al. (2016) and Cossarini et al. (2015b).**

Secondly, the quality of the input dataset benefits from a two-step quality check process, removing noisy and unreliable samples. The neural network architecture was also modified to enhance prediction performance by accurately selecting a performing nonlinear function, adjusting and optimizing the amount of neurons for each layer of the MLP model, and choosing a different optimization strategy to train the algorithm. NN-MLP-MED also includes a vertical smoothing (running mean of 5-10 m window) step and a climatological adjustment at depth 600m that is derived from EMODnet (Salon et al., 2019).

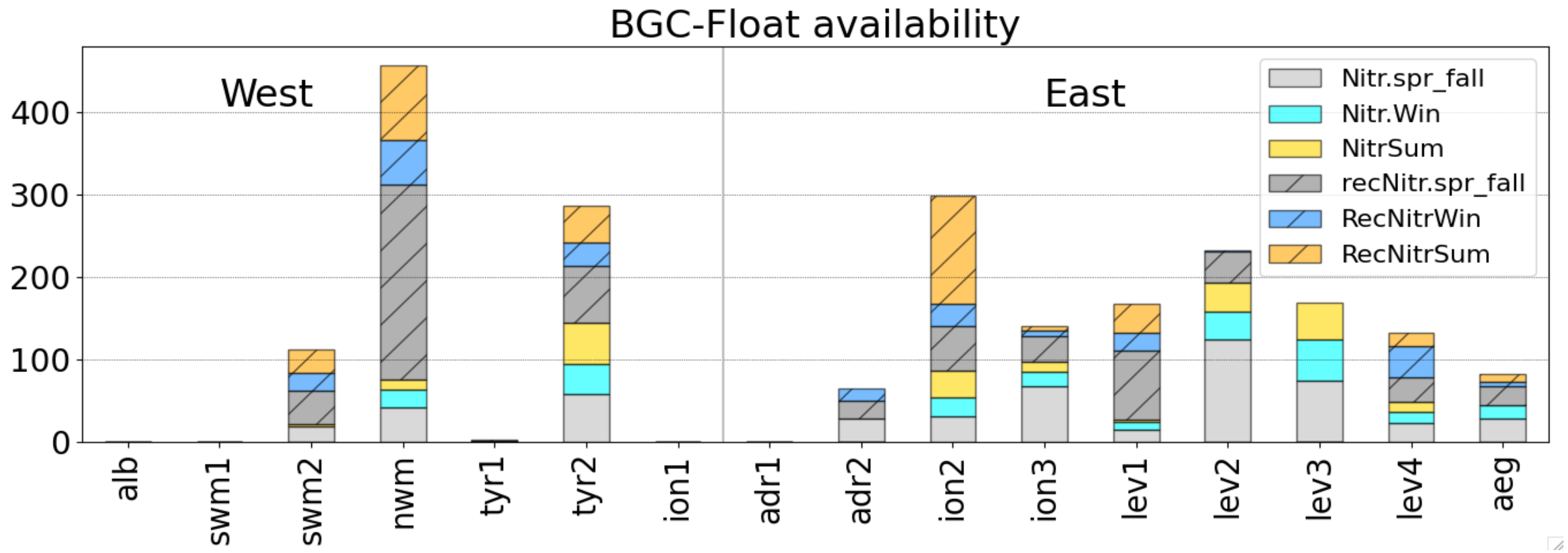
The input nitrate dataset for assimilation contains 938 BGC-Argo profiles and 2146 reconstructed nitrate profiles (Table 1). The reconstructed nitrate profiles are located 61% in the western and 39% in the eastern Mediterranean Sea, thus providing a larger and more homogeneous spatial coverage asin Figures 2.

Uncertainty of reconstructed nitrate associated to the EMODnet validation dataset is 0.5 mmol m−3, while it reaches 0.87 mmol m−3 when it predicts the BGC-Argo dataset (Pietropolli et al., 2023).

comment13a

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 propose the following figure in order to fill the gaps on the distribution of reconstructed nitrate and nitrate data (by subbasin, by season and by assimilated dataset).



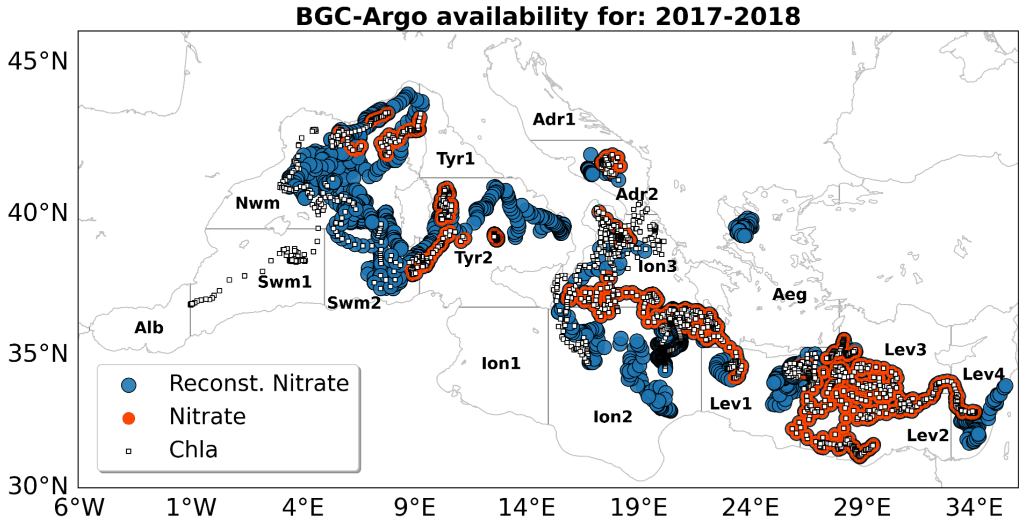
comment14a)

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. In particular, what level of drift correction for oxygen has been done in these data sets?

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 (e.g.,**comment2b**), moreover the updated Introduction will reply to the Reviewer's comment on oxygen drift correction.

We select AM and DM data of oxygen and chlorophyll, and all DM data for nitrate. **AM data of nitrate are included after being corrected using CanyonB or WOA as explained in Johnson et al. 2021. For the three variables we use measures that are flagged as good, probably good, and interpolated.**

Fig. 2: “of chlorophyll-a (red), Nitrate in situ (orange) and reconstructed Nitrate (grey)” - this may just be a matter of perception, but the colours used don’t look like red/orange/grey.
Following the comments of the other reviewers, we propose this revised version of the Figure 2:



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).
Generally, the in situ drift is found after 1 year of the deployment (Bittig at al., 2018, Maurer et al., 2021 and Thierry et al., 2021) and more rarely after months. Consequently, at T0 the drift analysis is never calculated, because we believe that the effect of calibration is still strong.
Following the comments of the other reviewers, we will add more details on sensor drift in the rewritten version of the introduction. In this way, the reader should have enough information about the method. Moreover, we propose to add the following explanation (in bold):
Tests are applied when the life of a float is longer than 1 year. **Conversely, if the available float time series is shorter than 1 year, their profiles are not corrected because the float lifetime is considered too short to account for in situ sensor drift.**

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 propose to correct the sentence as follows:

Used for linear and non-linear regression problems, the RANSAC and Theil Sen methods **automatically** split the oxygen dataset into a set of inliers and outliers. In order to avoid possible biases (Dang et al. 2008 and Fischler and Bolles 1981), the methods calculate the drift from the inliers set of data.

comment15a

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 sentences are rewritten as follows also taking into account of other comments (e.g., **comment3b** and **comment 4b**):

L. 195 Here, the optode sensor in situ drift is evaluated through non parametric methods (RANSAC and Theil Sen) at two different depths (600 and 800 meters) **to avoid possible fake drift detection because of changes in the water masses.**

[…]

L. 201 The presence of a drift is established when all four drift estimates **(RANSAC at 600 and 800 meters, Theil-Sen at 600 and 800 meters)** agree in sign and their average value is greater than 1 mmol 200 m−3 y. This threshold was chosen on the basis of results in Bittig et al. (2018b). **Then, the drift is removed from the oxygen profiles by setting the computed drift average at 600 meters and linearly interpolating toward the surface, where drift is set equal to zero. According to Thierry and Bittig, 2021, there is a lack of specific tests at depth, while several (i.e., 14) tests applied near-surface are already performed by the GDACs. The presence of near-surface tests motivates our choice to reduce the effect of our correction at the surface.**

Line 207: “Marine Copernicus Service” - “Copernicus Marine Service”

Sorry for the error, we will correct it.

Line 208: “initial conditions from EMODnet dataset (Details are provided in Salon et al. 2019).” - does this include the same spin-up procedure as in Salon et al.? That should be detailed.

Thank you for the comment. The information are added as follows:

[..] initial conditions from EMODnet dataset (details are provided in Salon et al. 2019) and a 3-year spin up using the 2017 forcings in perpetual mode.

comment16a

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

Thank you for the comment. Information has been added as follows (new text is in bold) taking into account other reviewers' comments (see comment16b):

Before integrating data in the 3D-VarBio, the same pre-assimilation assessment described in Teruzzi et al. (2021) is applied to the chlorophyll profiles. **Nitrate profiles are rejected if concentration at the surface is higher than 3 mmol m−3.**

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 and Gordon 1992). 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 (see Figure 2) in the Mediterranean Sea.

comment17a

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

According to comment from the other reviewer (comment5b) we will clarify the choice of the threshold and rewrite the paragraph as follows (in bold new text):

L223 During the data assimilation, profiles are excluded when innovation exceeds specific threshold rules. For the chlorophyll the threshold is set to 2 mg m−3, for nitrate, the two thresholds are 2 and 3 mmol m−3 for 0-50 m and 250-600 m layers, respectively (as in Teruzzi et al., 2021). Oxygen thresholds are 30 and 50 mmol m−3 in the 0-150m and 150-600m layers respectively (thresholds are roughly 3 times the standard deviation of the climatology computed on Emodnet data for the different sub-basins). Exceeding values have to be found in at least 5 vertical levels in the specific layers. Exclusions are set to avoid corrections that can trigger unstable dynamics after the assimilation (Teruzzi et al., 2021, Storto et al., 2011, Sakov et al., 2017 and Waller et al , 2018). The excluded profiles ranged from 0.1% for chlorophyll to less than 1% for nitrate.

Additionally, we will change "misfit" with "innovation", introducing the term at old L137:

which relies on the innovation (i.e., the difference between the observations y and the model background xb)

Fig. 3: What does the horizontal line at 600 m represent?

We apologize for the inaccuracy. The horizontal line refers to the standard deviation calculated on the EMODnet dataset at 600 m. We will include the information in the text.

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). We propose the following modified text (in bold):
After drift removal, the oxygen concentration at depth is closer to the EMODnet climatological data. **This allowed us to infer that it is likely that the drift correction allows more profiles to be included in the assimilated oxygen datasets (Figure 1).**

comment18a
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). Please consider that we will change old L246-247 lines also considering the comments of the other two Reviewers (**comment6b** **comment7b** **comment2c**) as follows:
Skills performance of the simulations listed in Table 1 are evaluated by comparing model results with satellite Copernicus 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. **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 floats can share a certain degree of correlation.**

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 aim to clarify RMSE results from Figure 5. 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. We will list all the aggregates in the text. We decided to aggregate the basins to make the results clearer (since 16 sub-basin * 2 seasons * 3 variables makes a high number of profiles).

Line 253 and following: The changes in RMSE should be linked to the change in coverage. From visual inspection, most regions of reduced RMSE are regions of higher pseudo-nitrate (Figure 2), but not all of them e.g. Nwm. Other regions have no (additional) float data yet show changes in the RMSE.
Line 272: "directly ascribed to the increased number…" – this is not clear to me as the Figures do not show how the reconstructed obs are distributed over seasons.
Thank you for the comment. We will add information about the coverage in a table/plot as proposed for **comment13a** and to make clearer the link between coverage and impacts of DA.

Fig. 5 and associated discussion: it is not at all clear what is displayed in the figure. Absolute values? RMS errors? Percentage RMS errors? Are the x-axis values identical for all variables or have they just been cut off for all except the bottom panel?
We will redo Figure 5 to make results clearer. Please, consider that the x-axis refers to the value of the RMS errors (model vs BGC Argo) for each variable. Moreover, in the present version of the manuscript, the x-axis cannot be seen due to a Latex-formatting error in the figure (we will increase the space between the rows to correct this error).

comment19a
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). As commented in **comment11a** oxygen dynamics is not directly affected by chlorophyll and nitrate increments by the assimilation.

comment20a
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 for the comment. We prefer to keep the description by variable, unless the reviewer strongly suggests otherwise. However, also taking into account the comments from other reviewers, we will change the order of the figures (i.e., nitrate, phosphate, chlorophyll and oxygen), and we will markedly revise the section to improve its readability. New version of the section reads as follows (as in **comment9b**):

To assess the impact of profile assimilation in changing the vertical gradients of biogeochemical variables, the Figures 6, 7 and 8 show the Hovmoller diagrams of the spatial averages for two selected sub-basins (NMW and Ion2 in map of Figures 2) and for the whole Mediterranean Sea. This representation offers additional details on the vertical impact of the reconstructed nitrate profile assimilation with respect to the validation of Figure 5 that considers only model points corresponding to the location of BGC-Argo profiles. The two sub-basins represent two different trophic conditions in the Mediterranean Sea. The North Western Mediterranean (nwm) has higher level of nutrient concentrations and more intense surface blooms in winter (Siokou-Frangou et al., 2010, and Di Biagio et al., 2022). In summer, nwm has higher chlorophyll concentration at the deep chlorophyll maximum (DCM), shallow nitracline, and shallow subsurface oxygen maximum (SOM) (first column in Figures 6, 7, 8, and 9). On the contrary, more oligotrophic conditions and deeper nitracline and DCM are found in the eastern subbasin (ion2, second column of Figures 6, 7, 8, and 9).

Considering nitrate (Figure 6), the multivariate assimilation (DAfl) corrects a general positive bias of the model in all the Mediterranean areas (blue pattern in Figure 6). The addition of reconstructed profiles makes the corrections stronger. On average the nitrate concentration below the nitracline decreases by 8% and 11% in DAfl and DAnn runs, respectively. Both the assimilation runs also show changes of the nitracline depth (i.e., depth at which the vertical gradient of nitrate is maximum) with more intense deepening in the DAnn simulation.

Differences between the assimilation and the reference run accumulate over time. The rate of this accumulation is highest during the first year while during the second year it decreases and the differences remain almost constant in sub-basins with a high number of BGC-Argo and reconstructed profiles (e.g., NWM in Fig. 6).

On the other hand, considering other areas (e.g., ion) and the whole Mediterranean Sea, which comprises some under-sampled areas (e.g., southern Ionian and southern western basin), the effect of DA corrections is propagating after the two year(third column of Figure 6).

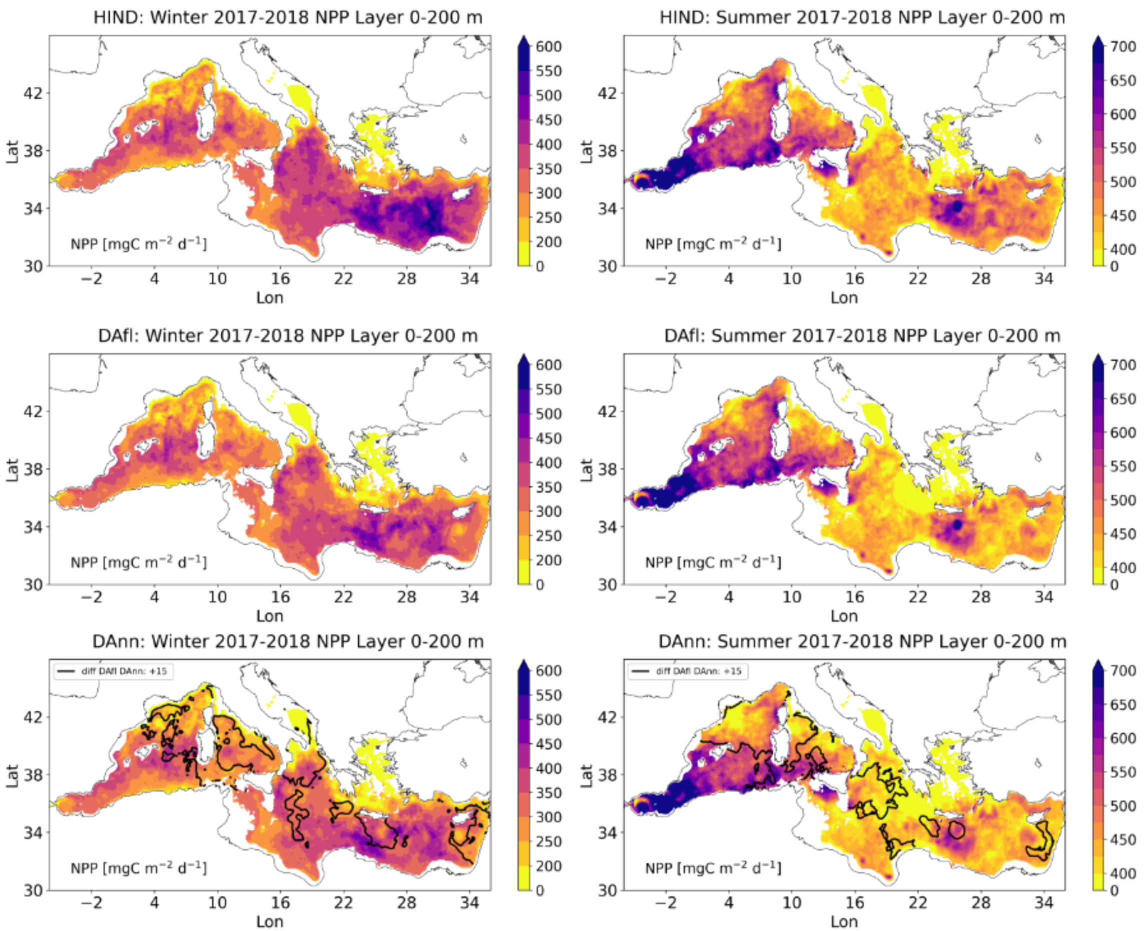
Very similar patterns are also observed in the Hovmoller diagrams of phosphate (Figure 7), which is an updated variable of the multivariate variational assimilation scheme through nitrate-phosphate covariance. In fact, the general negative corrections on phosphate fields are linked to the high positive values of the covariance matrix between nitrate and phosphate (Teruzzi et al., 2021).

Considering chlorophyll (Fig. 8), the main differences between DAfl and HIND are a slightly 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 8). Even if the chlorophyll validation (Figure 5) does not show significant differences between DAfl and DAnn, the basin wide averages of DAnn display more intense corrections with respect to DAfl in terms of DCM depth and chlorophyll intensity and overall chlorophyll concentration (Figure 8). Over the 0-200 m layer of the whole Mediterranean Sea, the chlorophyll decreases with respect to HIND are 4% and 5% for DAfl and DAnn, respectively.

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. On the Mediterranean basin-wide scale, the average correction is 0.2% for the 0-200m layer. The addition of the nitrate reconstructed profiles does not alter the correction pattern with an average correction of 0.3%. However, the largest differences between the two assimilation runs can be spotted in areas with a high density of reconstructed profiles during summer (e.g., NWM, first column in Figure 9). As observed in the nitrate and chlorophyll figures, the assimilation of reconstructed profiles causes a decrease of the summer productivity in the DCM layer. Consequently, less oxygen is produced generating the negative changes in the DCM layer in the bottom left panel of Figure 9. Because of the smaller amount of subsequent sinking organic matter, less oxygen is consumed in remineralization processes in layers below the DCM in late summer and autumn, and positive oxygen changes are generated, particularly during 2018.

Line 302: How do you distinguish if a region is still drifting in Figure 7? To me, ion2 (second column) looks as if it is drifting still, but the differences have smaller magnitude than in Med (third column)
Thank you. We will modify the comment to Figure 7 in agreement with **comment20a**

comment21a
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 will correct the typos in Figure 10. Following the suggestions of the Reviewers (see **comment10b**) and in order to better highlight the NPP differences between the three simulations, we changed the colorbar and added the15 mg/m3 contour line on the last row. The contour refers to the difference between NPP in DAfl and NPP in DAnn.



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 (**comment11b**), we decided to not infer any conclusions on the effect of chlorophyll assimilation on the primary production. We will remove the speculative conclusion on chlorophyll assimilation at L 339-341. The sentence we propose is the following:

The weaker fertilization of the surface layer in DAnn, which occurs for both macronutrients after assimilation (Figures 7 and 8), causes a reduction of the net primary production.

comment22a

Line 346: "0-300 m", Figure 11 says 0-600 m in the title

Thank you for spotting the typo. We used 0-300 and 0-600 m layers for chlorophyll and nitrate respectively. We will correct the text and the equation. We will replace subscript "300" with a more general term "maxdepth" in the equation 2.

In this work, we adopt the impact indicator Iij (t) described in Teruzzi et al. (2021). The impact indicator allows quantifying the integrated response of assimilating BGC Argo profiles with respect to the no assimilative run: [equation_corrected]. HIND is the reference, while Sim refers to one of the different DA setups (DAfl or DAnn). |Simij (t) − HINDij (t)| is the absolute difference between simulations (for each day and grid point), while the subscript maxdepth represents the integral over the 0-300 m and 0-600m for chlorophyll and nitrate respectively.

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 median value of the impact indicator (see colorbar), as we will clarify in the text:

Compared to a threshold of 0.4 (**the median of the impact indicator in the Mediterranean Sea**), the impacted areas increase from 18.2% to 29.8% in winter and from 10.8% to 14.5 in summer in the DAfl and DAnn runs

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 Fig 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 will add a decadal variability value to provide a term of comparison for the chosen threshold.

comment23a
Line 399: “more than 30 profiles” - what was that before? How much larger is the data availability?
Following the other Reviewer comments (see also **comment14b**), 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 will better explain this concept at L399-400 as follows:
Through the integration of NN and DA, the number 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.

comment24a
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 propose the following changes (in bold the new text) also considering comment from the other Reviewer (**comment15b**):
This means that seasonal sub-basins scale dynamics (e.g., bloom or stratification) can effectively be constrained, while, as stated in d’Ortenzio et al., 2021, mesoscale dynamics is still limited to be only locally studied.
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).

Line 409: The decrease in available BGC Argo observations was not mentioned before, but feels like this should be a major motivation of this work (for the introduction)
Thank you for raising the point. We solved the issues of the BGC Argo availability by providing more accurate information in the Introduction.

Line 415 & 443: “feed-forward” - this term is suddenly introduced in the discussion and conclusion when describing the method used, it should be introduced and explained in the methods section.

We have introduced this information in the Methods following the comments: **comment12a** and **comment1c**

comment25a

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, **comment3a**, we will rewrite part of the discussion by adding the following sentences:

Previous works demonstrated the complementary of vertical profiles with respect to satellite ocean colour assimilation (Cossarini et al. 2019, and Ford et al., 2021, Skákala et al., 2021 and Teruzzi et al., 2021)