Thank you for taking the time and care to provide valuable feedback and contributions to this manuscript. Please see our responses to the comments below, which we are ready to implement for a future revision.

Copy of review comments from “Reviewer 2” (RC39-RC90) are given below, followed by the author comments (AC).

Reviewer 2

General Comments

RC39	I feel that many choices of the authors - both in the definition of the algorithms and the configuration of the numerical experiments - coud be better justified. Why building the update of the chemical component of the PFTs in that way? Why using the surface update for the variables in the mixed layer and not updating them in a similar way as done at the surface? Why not updating the variables under the mixed layer? Why comparing the ML schemes to an univariate scheme and not to an EnKF? See specific comments for more details. I think that this paper would benefit from additional justifications.
AC	We thank the reviewer for their valuable comments and agree that we can better justify many decisions made for this work. Detailed responses are provided in the relevant “specific comments” below.
RC40	To my point, the discussion on the results of the ML-OI and ML-EtE schemes are too optimistic, especially regarding the framework of updating all the unobserved variables (see specific comments). For instance, the 7-day forecast root mean square errors (RMSE) in the nutrients and some detritus are larger with the ML approaches compared to an univariate scheme updating only the PFTs. This may highlight assimilation biases due to inconsistent updates. From the results shown in this work, I am not fully convinced that they ”have successfully shown that ML methods can make improvements to the DA schemes” as claimed by the authors. I think the discussion of the results should be qualified and better higlight the remaining issues of their approaches
AC	We agree that the claim is arguably too strong and should be more carefully qualified. The specific comments will be addressed individually, but based on the above comment, we also think it is important to modify the key claim: “... successfully shown that single model ML methods can partially or fully emulate updates from ensemble DA schemes, yielding many improvements to analysis state estimates over the operational-inspired reference univariate scheme (RUS)”.

RC41	The extension of these approaches to 3D models does not seem to be straigthforward. It is mainly due to generalization issues well known in ML and observed in this work - the NN trained with data from the L4 station perform poorly at the CWEC location - that could require to sample the English Channel with a large number of 1D vertical ML-based DA systems. Furthermore, the training of the NN is based on existing ensemble of simulations (forecast correlations) or advanced multivariate DA systems (analysis increments). The data they produce are uncertain (large number of unknown parameters) and potentially not relevant notably in case of changes in the system (observation network, increases in the temperature that would result in the occurence of out-of-distribution data during the inference) leading to poor performances of the approaches. In that case, it may require to produce a new data set and retrain the NN which may be costly and time consuming.
AC	We appreciate the reviewer's observations. We agree that generalizing ML-based 1D DA systems to 3D is challenging, particularly due to physical processes like horizontal advection that are not captured locally. As noted, models trained at one location (e.g., L4) often perform poorly elsewhere (e.g., CWEC), highlighting issues of extrapolation. Transfer learning is a potential path forward for future experiments that we will add to the discussion. Furthermore, while full 3D extension is non-trivial, a hybrid approach could help: ML models could handle local multivariate aspects at observation sites (a 0D transfor, while traditional DA methods (e.g., spatial correlations functions) manage 3D reconstruction (just as they manage the 1D reconstruction in our setup). We will also develop the discussion on clustering to identify similar regions, reducing the need for widespread retraining.

Specific Comments

RC42	p.2, fourth paragraph ”we investigate the capability of ML to learn the hidden, non-linear, and complex relations between BGC variables and to use the learned functions within a DA scheme or fully substitute it”: In this work, the data assimilation algorithms used are derived from an ensemble Kalman filter (EnKF), which means that a linear update is applied to the variables. Common ocean biogeochemical models are non-linear and subject to constraints such as the positivness of the variables that make the use of an EnKF challenging. So, one might wonder what are the benefits that can be expected to learn complex non-linear relations when using linear data assimilation methods. The authors may comment this point and hightlight the constraints associated with operational oceanography. Furthermore, the authors use ML algorithms to learn the forecast error correlation (ML-OI) or the analysis increment of an EnKF (ML-EtE). So it is not clear that the model did actualy learn the ”hidden, non-linear, and complex relations between BGC variables”. The authors should consider to rewritte the sentence to better highlight their contribution.
AC	We will rewrite this to be clearer. While the updates are linear (coming from the EnKF), the flow dependence learned from the EnKF is not. ML-OI learns a linear relationship (i.e., it predicts the correlation), but ML-EtE predicts analysis increments of observed variables from the state of the system and the analysis increment of the observed variable, which is a non-linear relationship. We will update the manuscript to make the distinctions on what is linear vs non-linear clear. We will also add some additional context on these constraints associated with BGC DA.
RC43	p.4, last paragraph ”for temperature, salinity and nutrient fields for biogeochemical relaxation profiles”: it would be worth to mention which classes of nutrients are forced with the World Ocean Atlas data.
AC	Agreed. We will specify this.
RC44	p.4, last paragraph ”A nutrient relaxation of 3 months towards [...] to prevent significant trends forming that cause the 1D model to gradually accumulate nutrients”: I wonder what are the reasons for these trends. Could the authors add a comment? Furthermore, even if the time scales of the relaxation are larger than the assimilation cycle (7 days), I wonder at which point it prevents the occurrence of assimilation biases in the water column. In this study, only surface variables are updated from the observations based on a BLUE-like equation. The same increment is then used to update the corresponding pelagic variables in the mixed layer while no updates are applied below this layer. It means that the impact of the observations on the variables below the mixed layer is only due to the adjustement of the model to the update in the mixed layer. Because it can be weak and compensate by the relaxation towards nutrients, it may prevent the occurence of assimilation biases such as trends in the nutrient concentration in the deep

	layers. I think that the authors should add a comment on it to highlight the limits of this work for more realistic applications
AC	As the model lacks lateral advection and riverine inputs, any imbalance between vertical nutrient fluxes and biological uptake can lead to systematic accumulation or depletion over time. The relaxation term acts as a simple corrective mechanism to counter this drift and keep the system within realistic bounds. As the reviewer correctly points out, the relaxation profiles could contribute to controlling the sub-mixed layer in our setup, which could help to mitigate some long-term biases in this area. As suggested, we will comment on this as a limitation for operational scale systems which do not have these relaxation profiles.
RC45	p.5, first paragraph ”The resulting variation in wind strength [...] prevents ensemble collaps induced by the previously mentionned nutrient relaxation or [..] are absent in a 1D set-up.”: I presume that this strategy is effective in preventing an ensemble collapse because only the surface variables are updated. In this case, it might be sufficent to maintain the spread of the ensemble at the surface, independently of the deeper layers. However, it may not be effective in a setting where the variables are update throughout the water column during the analysis. I think that a comment should be added to highlight that point.
AC	Agreed.
RC46	p. 6 Eq.(4) ”H = [1, 01, ..., 0N] [...] with the first element being surface total chlorophyll”: The total surface chlorophyll is not a state variable and rather ”an observation” that can be computed from the phytoplankton function types (PFT, state variables). This is quite unusual reagarding the mathematical framework of data assimilation. So, I think that the authors should better motivate this choice, notably in relation to the following comment.
AC	See response to RC15 of Reviewer 1 and response to following comment.
RC47	p.6, third paragraph ”The PFTs are excluded from the state [...] where χ stands for each of the non-chlorophyll chemical components of a PFT”: I am not sure to understand the motivation of this strategy. The authors suggest to update the total surface chlorophyll with the BLUE analysis in a first step, then to update the PFT chlorophyll based on the background ratio with the total chlorophyll, and finally to update the non-chlorophyll chemical components of the PFTs based on the background ratio between the chlorophyll and the non-chlorophyll component. I understand that it maintains the ratios chlorophyll / total chlorophyll and chlorophyll / non-chlorophyll components constant in the sense that $\frac{x_\chi^a}{x_c^a} = \frac{x_\chi^f}{x_c^f}$ and $\frac{x_\zeta^a}{x_\chi^a} = \frac{x_\zeta^f}{x_\chi^f}$. Would not it be possible to maintain this property by updating the PFT concentrations directly from the total chlorophyll concentration in the

	analysis as usually done with an EnKF? I think that the authors should motivate this strategy and explain why they can not use directly the analysis increment for the PFT.
AC	This approach follows the methodology of Teruzzi et al. (2014) and Skakala et al. (2018). In this, chlorophyll innovations are distributed proportionally across phytoplankton functional types (PFTs) and associated elemental pools (C, N, P, Si), using the internal ratios provided by the forecast. This ensures that assimilation respects the acclimation dynamics and preserves the community structure and physiological integrity of the model (which using the EnKF would not do, and in the single model runs we do not have an ensemble anyway). By updating only the total concentrations while maintaining forecasted stoichiometric ratios, we avoid introducing biologically inconsistent changes that may degrade model realism and performance. We will ensure this is clear in the revised manuscript.
RC48	p.6, fourth paragraph ”All DA methods here will only update the surface layer in the model. However the rest of the mixed layer is also updated [...] Variables below the mixed layer are not updated”: First, I think that the authors should add a comment explaining why they do not update the chemical components of the PFTs in the mixed layer according to their strategy at the surface (first the total chlorophyll, then the chemical components). Are there issues to do so? Secondly, the authors should explained why they do not update the variable below the mixed layer. I am wondering what are their motivations and what are the limits for applying this approach in realistic configurations.
AC	Propagating surface increments through the mixed layer is used operationally, and we see that the values and correlations are homogeneous within the mixed layer (and so we could use the strategy described above, but would get the same result). We do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the surface. As the reviewer has pointed out, this could have wider implications that should be considered (e.g. a system without nutrient relaxation profile might accumulate/exhaust nutrient stores in the sub-mixed layer, leading to a potential bias or source of error), and so we will add some text to better reflect the limitations and complexity that may arise in the context of an operational system where the sub-mixed layer is not controlled by nutrient relaxation profiles (as it is in our experiments). Though, we cannot necessarily expect an operational DA system with a short forecast lead-time to fix these biases occurring from processes with long time-scales, we can also include discussion on strategies that might help with this (e.g. bias correctors).

RC49	p. 6, sixth paragraph ”We call our baseline DA method the reference univariate DA scheme (RUS, Table 2, row 4)”: Could the authors motivate their choice for this baseline and not the EnKF used to produce the increments for the training of the ML-EtE algorithm? The EnKF is more advanced than the RUS algorithm, so I would compare the performance of the ML algorithms compared to a plain EnKF (multivariate analysis).
AC	The RUS scheme is much more similar to what is used operationally. While we could compare to the EnKF, it is an ensemble method, while all other methods are not. The point of interest is that the EnKF provides the flow-dependent training increments for a method that can be "learned” and more importantly applied to a single model run.
RC50	p.7, Table 2 ”Eq.(4) with (5)”: Eq.(4) refers to the observation operator H and (5) to the PFT update. Does it mean that Δxi correspond to PFT only?
AC	This is a typo, and should be clearer. We will change to read: “Eq.(8)” – and specify this is the update strategy for the unobserved variables (not PFT or chlorophyll).
RC51	p.7, first paragraph ”The EnKF updates all elements of the surface state [..] duplication of the analysis increments from the surface throughout the mixed layer”: I wonder why the authors do not apply the analysis step of the EnKF to variables in the mixed layer rather than copying the increments computed for the surface. Could the authors explain this choice?
AC	We will add the reasoning to the methodology. This is done so that there is a 1-to-1 correspondence between the strategy used to generate the training data, and the strategy applied in the single-model run schemes.
RC52	p.7, second paragraph ”The implementation of an EnKF to this problem is relatively expensive”: Considering that the problem is only 1D, I would remove this sentence.
AC	Agreed.
RC53	p.8, subsection 3.1 ”Mathematical framework for the ML-based DA schemes”: This subsection is very short. The authors may remove the structure of the subsection and start 3.1 with the subsection entitled ”Hybrid machine-learning optimal interpolation”.
AC	Agreed.
RC54	p.8, second paragraph ”where P f i,c is the background error [...] where CORi,c is their forecast error [...] respective background error”: Have ”background” and ”forecast” the same meaning? If yes, the authors may consider to use one them only.
AC	Yes, they mean the same thing. We will make the terminology consistent.
RC55	p.8, fourth paragraph ”is scaled according its climatological”→ is scaled according to its climatological (?)

AC	Agreed.
RC56	p.8, fourth paragraph ”The resulting surface increments of the unobserved variables are then propagated to the other levels in the mixed layers as described previously”: I wonder why the authors did not apply the same ML strategy to learn the forecast error corrrelation Corri,c between the surface chlorophyll and the unobserved variable in the mixed layer. Could the authors motivate their choice?
AC	The correlations and values are homogenous in the mixed layer, which allows us to make this simplification in this work (as well as in the operational system which motivates this work). We will make this clear in the revision.
RC57	p.8, last paragraph ”For the first application of ML-OI in Sect 4.2, the target [...] between total chlorophyl and nitrate. In the later application in Sect. 4.3 onwards this is extended from just nitrate to a wider set of variables.”: Could the authors motivate their choice to start with the correlation between total chlorophyll and nitrate only and not directly presenting the results of the more general case?
AC	Nitrogen makes a good first choice of unobserved variable to test the system as it is the limiting variable: *“Nitrogen is a key component of organic matter and is generally the limiting nutrient to primary production by phytoplankton in coastal marine ecosystems (Council et al., 2000).”* This makes it a primary target for incremental improvements to an operational DA scheme, without trying to update the whole state. In this, we would hope that correcting nitrate will have further benefits to other aspects of the system. However, the experiments show that this is not necessarily the case, highlighting the need for specific results of each update strategy (e.g. Fig. 7). Furthermore, updating only a single additional variable also allows us to study the increments and correlations in more detail (e.g. Fig. 3 and 5). We will make this motivation clearer in the revision.
RC58	p.9, second paragraph ”In ML-EtE, we assume that the essential properties of each statistical object [..] that are combined in a single value”: It could also be mentionned that directly predicting the analysis increment (a vector) allows to tackle problems in large dimensions that would not be possible when predicting a error covariance matrix required in the assimilation step.
AC	Very good point. Agreed. We will add: “Further to this, directly predicting the analysis increment (a vector) is better suited to problems in large dimensions (e.g., a full operational 3D system) that would not be possible if predicting the full error covariance matrix required in the assimilation step.”
RC59	p.9, second paragraph ”The resulting increments of the unobserved variables are then propagated to the other levels in the mixed layer as

	described previously": Again, why not predicting the analysis increment for the full water column? The authors should motivate this choice.
AC	See responses to RC48 and RC58.
RC60	p.9, fourth paragraph "To generate the training data for this approach, we first generate a nature run [..] analysis increment for the unobserved variables.": The authors should clarify if the reference run used to train the neural network is the same as the one used for the comparison of the performances of the methods. For a fair comparison, the network should be trained on a time window that differs for the one used for validation. Is it the case?
AC	Yes, this is the case. We will add text to the subsection titled "Skill metric and machine learning model evaluation" to explain the different train, validation, offline and online test windows.
RC61	p.9, subsections 3.4 "Machine learning architecture" and 3.5 "Purely climatological updates": These two subsections are very short (few lines). Maybe it would better to merge them in a larger subsection.
AC	We can merge 3.4 with the earlier sections on ML-OI and ML-EtE. Section 3.5 suits its own subsection as it is a different method to the others (which each get their own subsections).
RC62	p.9, seventh paragraph "Each ML approach is tested in the following scenarios [...] results of (2a)": The auhtors may add a comment to explain the motivation of this splitting.
AC	We can make this splitting clearer in the manuscript. We first tested nitrate it is the key limiting variable (as shown by the regimes, and response to RC57). After seeing that we could correct a single unobserved variable, we then aim at updating all variables in the system. We saw some variables degraded significantly, which could go on to impact other variables through the time evolution of the model. In an attempt to remedy this, we ran an additional experiment that aimed at updating all variables, except the clearly problematic ones. We can add some discussion on the limitations related to testing the different update schemes.
RC63	p.10, fourth paragraph "MGBC": Could the authors explain this acronym (first occurence)?
AC	Typo. Should read: "marine BGC"
RC64	p.10, last item "During this period, phytoplankton concentrations are very low": During this period negative concentrations may occur during the analysis step. Could the authors comment on this point and mention how they remedy to this?
AC	The model is constrained to physically non-negative values, so any negative values that arise during the data assimilation (DA) step, though extremely rare, are clipped to zero. This occurs infrequently and only in very localized

	instances. As such, this treatment has a negligible impact on the overall state and statistical distributions. Additionally, due to strong surface forcing, the model tends to quickly redistribute any localized anomalies, minimizing the persistence or propagation of these clipped values. Therefore, we are confident that this approach does not significantly influence the model performance or results.
RC65	p.11, first paragraph ”In this section, we explore the performance of ML-OI and ML-EtE in updating only nitrate as an unobserved variable [...] a limiting nutrient at the L4 location (see Fig.A2 in the Appendix) [...] has a clear, explainable relationship with total chlorophyll”: Following previous comments, it is not clear to me what are the motivations of updating only the nitrate variable and what could be learned that would meaningful for the multivariate case updating all the nutrients. Fig.A2 highlights correlations between surface total chlorophyll and silicate (or phosphate) as strong as those between surface total chlorophyll and nitrate. So, one may anticipate similar magnitudes of update for the other nutrients. In that case, it is not clear what would be the evolution of the system updating all the nutrients after several assimilation cycles based on the evolution of the system updating only the nitrate variable. For instance, one notes significant differences in the forecast and analysis RMSE in the nitrate between the versions of the ML-EtE algorithm that update only the nitrate or all the nutrients (see Fig.7). For this case, updating all the nutrients damages the forecast and analysis RMSE in nitrate while improves the RMSE in ammonium and silicate. This result is interesting and may highlight assimilation biases induced by the different update strategies. Thus, it is worth to show results for the strategy updating nitrate only. However, I think that a focus on this strategy would require to show additional results: for instance the RMSE in the other nutrients at the surface and in the mixed layer depth. Then, the introduction of this subsection may be motivated as alredy done (nitrate is a limiting nutrient) while questioning the relevance of this approach and the need for specific results.
AC	Nitrate is expected to be the limiting nutrient at L4. This might not be fully captured by correlations as the nutrients tend to be generally correlated (e.g. silicate is to a degree much less relevant as it influences only diatoms, which, although being the dominant group at L4, is only one group of phytoplankton). So, due to the importance of nitrate, this specific nutrient was selected for the simplified analysis. However, we agree with the reviewer that the impact of nitrate vs all nutrient updates is non-trivial and this should be better discussed (and the comparison is clear in Fig. 7; Also see response to RC57).
RC66	p.11, first paragraph ”Offline referes here to a setup in which the ML-OI analysis is not then used as initial condition for the next cycle”: For confirmation, the total chlorophyll and the PFTs are not updated as well, are they?

AC	Correct. This is entirely “offline”, so no updates are actually made to the system – we just see what the update “would’ve been” to make a fair comparison in like-for-like scenarios.
RC67	p12., first paragraph and legend of Fig.3 ”RMSE=0.255” and ”RMS=0.731”: Are the RMSE computed accordingly to Eq.(10) with M = 1? Are these value large or low? Even if there is an improvement in RMSE with ML-OI compared to the daily climatology, it would be nice to have a comment on the value.
AC	Yes, M=1. The RMSE of 0.731 for climatological correlation predictions is large enough to cause the CliC model to damage the analysis (Fig. 6). Since this RMSE is measuring error to the EnKF correlations (and correlations are bounded between -1 and 1), this seems large. The ~60% improvement ML-OI makes to the RMSE value is enough to make improvements in the analyses. We will also look at the RMSE within each regime, as this can be more informative on where the biggest gains are made (e.g. capturing the timing of the spring bloom). We will add further comment on this in the revised manuscript.
RC68	p.12, second paragraph ”We also see that during the light-limited regime [...] at this time of the year”: It could be highlighted that the correlations tend to be too weak for both methods compared to the reference. It can be due to the fact that both methods are not able to reproduce the interannual variablity observed for the correlation in the reference. On the contrary, both systems tend to reproduce the same weak negative correlations every year while larger negative or weak positive correlations (winter 2018) occur in the reference depending of the year. Even if the chlorophyll uptades are expected to be weak, it suggests that both systems can not reproduce the interactions between chlorophyll and nitrate during this period of the year.
AC	We will highlight this, as yes, both systems struggle to replicate the correlations during winter. Though, we should also expect that these variables are decoupled during winter (as nitrate is not the limiting factor during winter), so the correlation is less meaningful than during bloom or nitrate limited periods.
RC69	p.13, first paragraph ”The relative analysis error of total chlorophyll is normalised according to observational error. exceeds a value of 1 as we are measuring expected error [...] to the ensemble mean,”: First, it seems that a word is missing. Secondly, I am not sure to understand why the ensemble mean of the RMSE is larger than 1 while the RMSE is close to 1 for the algorithms using an univariate analysis and based on climatology or ML. Could the authors add a comment to better explain this result?
AC	Yes, we will add a comment on this. It comes from calculating the average error of ensemble members, rather computing the error of the ensemble mean, as well as the ensemble of observations used in the EnKF (which adds noise). The single model runs does not need to add noise.

RC70	p.13, last paragraph ”In contrast to this, both ML approaches result in significant improvement in performance, reducing analysis error by between 8 – 12%.”: The authors did not comment the performances of the ML approaches with respect to the EnKF. First we note that they lead to similar RMSE averaged over the experiments, which is a good result. However, the standard deviation over the experiment is way larger compared to the EnKF when the ensemble is large enough. It suggests that the performances of the ML algorithms are more sensitive to the randomness in the experimental framework (observation error) than an EnKF. It can be an issue in realistic settings where errors are badly estimated and can be large. Furthermore, it could also be noted that the ML-EtE, that learned the analysis increment of an EnKF, leads to a lower standard deviation than the ML-OI algorithm.
AC	We will expand the discussion to account for the results on the standard deviation. A few points: (1) the std in all experiments (except EnKF) is calculated on 64 cases as opposed to EnKF that uses 20 ensemble initialisations per ensemble size – the std in the EnKF is the spread in the mean of these experiments, rather than the std of the error in the ensemble members. This shows the ensemble becomes less sensitive as we increase the ensemble size, but we will think if we can present this information more carefully to be less confusing; (2) yes, we should note that the variability in the ML-EtE has a lower std than ML-OI, which is a good sign of lower sensitivity; (3) in the caption, instead of saying "... methods summarised in...." we will say the method names again and refer to the sections later.
RC71	p.14, Fig.5: It may be recalled in the legend that the increment is computed at the surface.
AC	We will recall this information as suggested.
RC72	p.15, second paragraph ”During the nitrate-limited period, we generally see comparable performance [..] the ML-EtE approach seems to more accurately capture the largest analysis increments”: Regarding the ML-EtE, this is true for the first nitrate-limited period in 2022. However, one notes too large erroneous increments both in July and August 2023, that are weaker with the other methods. I think that it should be stated in the discussion. Furthermore, have the authors an explanation for these increments?
AC	We will qualify the statement. Though it is strong to say that there are two erroneous increments in July and August 2023 by ML-EtE. As pointed out, the truth is certainly further from the background state, but the ML method still makes increments in the right direction (even if the magnitude isn't fully correct). The reason for the larger increments at this period is due to the dynamical evolution of the model (and the adjustments made in earlier cycles that

	then impact later cycles). It just so happens that CliC and ML-OI coincide at this point and ML-EtE does not.
RC73	p.15, second paragraph ”We can also see that each approaches make little to no adjustment during the ligth-limited regime”: The ML-OI algorithm leads to a large negative increment at the end of 2023 while the concentration of nitrate in the forecast is significantly weaker than the one in the truth. Could it be related to my previous comment on the ML-OI algorithm being not able to capture positive correlation between the nitrate and total chlorophyll in the ligth-limited period when they occur? It may be interesting to see the evolution of the increment in total chlorophyll as well.
AC	This is an excellent point, and makes a good link between Fig. 3 and 5. The correlation predictions of ML-OI clearly struggle to predict the positive correlations that can sometimes occur during the light-limited regime. We will add comments on the limitations of ML methods if these situations are poorly represented (or even absent) in the training data, we cannot necessarily then expect the ML models to capture these features well. However, we should also consider that we expect nitrate and chlorophyll to generally be decoupled at this point in the year, and so it is difficult to say if the correlations from the EnKF are meaningful (i.e. there may be no information in the state that is particularly informative of the correlations during winter). As it occurs during a transition between the regimes (it is close to the regime boundary defined by the seasonal behaviour of total chlorophyll and nitrate), this could also represent a period of time when the ML model is going to be particularly sensitive, as in Fig. 8, we see that ML-OI is reasonably sensitive to total chlorophyll (i.e. the observed variable). We will add discussion on these points.
RC74	p.15 Fig.6: I presume that the RMSE is computed in the surface layer. Is it correct? This information may be missing.
AC	Yes, this is computed at the surface. We will add this information.
RC75	p.15., last paragraph ”where errors are normalised against the error in the RUS scheme”: I wonder why the authors do not normalise against the error of the EnKF, which is the algorithm with the best RMSE in nitrate (see Fig.4). Maybe the best would be to plot the error without normalisation and add the reult of the RUS and EnKF as benchmarks.
AC	As the scale for the error fluctuates according to the time of year, showing the error relative to another scheme is logical in this case. As with response to RC36 of Reviewer 1, RUS acts as a comparison point for single model runs, and we feel that adding the EnKF to Fig. 5 onwards could distract from comparing the single model runs.
RC76	p.17, Fig.7: I presume that the RMSE is computed in the surface layer. Is it correct? This information may be missing.
AC	Yes, this is computed at the surface. We will add this information.

RC77	p.17, first paragraph ”Before discusing the extended schemes, we can see from Fig. 7 the dynamical impact of the updates of the ML-EtE (N03) have on other (i.e. non-updated) marine BGC variables [...] particularly microzooplankton.”: It could be added that it also slightly degrades the ammonium and silicate concentrations that are not updated during the analysis.
AC	Agreed, we will add this.
RC78	p.17, last paragraph - p.18, first paragraph ”Our next scheme, ML-EtE (ALL) [...] improving unobserved forecast and analysis RMSEs by between 10 – 50%.”: Compared to the RUS, the ML-EtE (ALL) damages the forecast RMSE in all the nutrients, almost all the detritus. Even if the analysis RMSE are better, the increase in the forecast RMSE compared to a univariate scheme is a major issue. It would rather highlight inconsistencies in the multivariate update resulting in assimilation biases.
AC	We see how this statement is overly strong and so will moderate the claim. However, we also think it is important to consider the forecast error at different lead times, rather than just the 7-day lead time which the current plot summarises. As RC81 will also point out, we provide this diagnostic in the nitrate only experiments, but not the extended set. As such, we will add additional diagnostics to support this. We will also comment on the damages at longer lead times, and how this can arise from inconsistencies in the model state as a result of the update.
RC79	p.18 first paragraph ”This also suggests they have generally weaker correlations with total chlorophyll.”: Fig.A.2 could help to strengthen this statement. However, it would not explain why the forecast RMSE are that large for the microzooplankton.
AC	We agree and will add that Fig.A.2 supports this, as well as some improved discussion. The RMSE of the zooplankton group, in our configuration, shows high sensitivity to other variables (whether we are updating only nitrate, or a wider set of variables). If the correlations between total chlorophyll and zooplankton are weak (as above), then the total chlorophyll increments do not provide much information to correct the zooplankton variable (which are already sensitive to other changes in the system).
RC80	p.18, second paragraph ”The ML-OI (ALL) scheme [...] and ML-OI (ALL) schemes”: I do not agree with the statement that the ML-OI (ALL) is effective. The forecast RMSE are even larger than those of the ML-EtE (ALL) for most of all the unobserved variables. For several variables, the analysis is not able to reduce the RMSE below the value of the RUS model. To my point, it rather hightlights assimilation biases due to the inconsistency of the updates that accumulate over time. I think that the authors shoud comment on it.

AC	Yes, we agree and will qualify this statement. ML-OI (ALL) is effective as far as mostly producing increments in the correct direction (even though the forecast itself is damaged overall). However, the sensitivity of the zooplankton variable is causing an issue here, which is evidenced by the ML-OI (Excl. Zoo) which does not update zooplankton and performs better as a result.
RC81	p.18, last paragraph ”Furthermore, even if forecasts are improved during most of the 7-day period relative to the RUS model (as can be anticipated based on Fig. 6)”: I do not agree with this statement. Fig. 6 concerns the ML-EtE (NO3) scheme that is not concerned by such large increases in the 7-day forecast RMSE in the unobserved variables shown in Fig. 7. So, it is quite speculative to suppose that the forecast RMSE remain lower than those of the RUS model during most of the days of forecast window. I think that additional diagnostics are required to support this statement.
AC	We will add some supplementary material (or appendix, whichever is more appropriate), that will show the growth/time evolution of the error, as well as some corresponding discussion. Here we can show that the benefits last for several days for multiple variables (though not all). This will further enhance the quality of the manuscript.
RC82	p.18, last paragraph ”around the 7-day lead-time it does not really outperform the RUS approach”: The authors could clearly write that the RUS model outperforms the ML-OI and ML-EtE algorithms for almost half of the unobserved variables at the end of the 7-day forecast window.
AC	Yes, we can make a much clearer statement here and will rephrase it.
RC83	p.19, Fig. 8: For a given output variable, do the feature importances sum up to a constant value? If yes, it may be useful to add this information in the legend for a better interpretation of the values. Futhermore, for the ML-OI scheme, it could be useful to recall that the output correlation are between the total chlorophyll and the considered variable.
AC	Since the input feature importances do not sum up to a constant value, and are taken as a mean absolute value, we can only consider them relative to each other. As pointed out by the reviewer in RC85, this has some potential implications for feature reduction. We agree it is useful to recall the correlation variables. We will change the text to read: “...which correspond to the correlations between total chlorophyll and each unobserved variable”...
RC84	p.19 first paragraph ”(Sect. ??)”
AC	Typo in the manuscript latex and it will be corrected to point towards what is currently “(Sect.3.6.2)”.
RC85	p.20, third paragraph ”This is to be expected, as the total chlorophyll increments contains [...] as described in Eq(1)”: It could also be recalled

	that it is an input variable for the training of the network. I wonder what are the consequences for the training of the network. Could one strongly reduce the number of input data during the training based on these shape values for reducing its computational costs without degrading too much the accuracy of the inference?
AC	Yes, this is an excellent point. We will add this to the manuscript.
RC86	p.21 ”This is likely because the correlations predicted by the ML-OI scheme represent a more locationn-agnistic relationship in the marine BGC variables”: this statement could be qualified by noting that the updates of the ammonium, bacteria and labile DOM lead to an increase in RMSE.
AC	Agreed. We will add this qualification.
RC87	p.21 ”not updating the zooplankton as in the ML-OI (Excl. Zoo) scheme (purple) produces a broadly improved result.”: I am not sure that one can conclude that excluding zooplankton results in an improvement compared to the RUS scheme. I would rather highlight that the damages due to the update are weaker compared to the ML-OI (ALL) scheme. The improvements compared to the RUS scheme seem to be marginal.
AC	Yes, this is a better way to phrase this. We will change it accordingly.
RC88	p.21 ”we see that detritus is generally improved relative to the RUS scheme.”: The improvements in detritus are marginal and of the same order of magnitude than the damages. The authors should qualify this statement.
AC	We will qualify the statement and make this clear.
RC89	p.22, second paragraph ”We have successfully shown that ML methods can make improvements to the DA schemes of marine BGC models when coupled to a 1D physical model”: I would qualify this statement which is too optimistic. The ML models used to update the unobserved variables degrade the 7-day forecast RMSE for several variables compared to the univariate scheme (RUS). This is a serious issue. While the evolution in time of the forecast RMSE is shown for the nitrate-only update, it is not the case when updating all the unobserved variables. So, one does not know how long the benefits of the analysis updates last during the forecast.
AC	We will qualify this statement as suggested. We will also provide the additional diagnostics and corresponding discussion as suggested in RC81.
RC90	p.23, third paragraph ”With machine learning (ML)”, we achieve significant improvements over conventional approaches that rely on climatological statistics or omit updates altogether. Our analysis of ML-predicted ntirate updates illustrates [...] for improving BGC DA.”: The ML approaches result in an improvement compared to the use of climatological statistics in the nitrate-only update framework. However, the impacts of using climatological statistics is not assessed in the general case. Because the ML scheme can degrade the forecast RMSE

	in unobserved variables compared to the RUS or the nitrate-only ML schemes, we may wonder if they can be useful in this more complex setting. The sentence may be qualified in that sense.
AC	Agreed, we will qualify the statement accordingly. Though, we also note that the inability for the CliC method to work when updating the nitrates only leads us to assume that it is more unlikely for it to work when updating all variables.