

Reviewer 1

In the following, RC = Reviewer's Comment, AC = Updated Authors' Comment, CH = Change

General Comments

RC01	<p>Machine learning (ML) techniques will likely gain more and more traction in marine biogeochemical data assimilation applications. The authors present an interesting study with compelling results, but I worry that the choice of using synthetic data limits strongly how much we can learn about how to implement ML-assisted DA in less idealized situations. The allure of synthetic data is apparent: there is a true model state, we know all about this truth and can thus examine DA improvement even regarding unobserved variables. However, in this study, the synthetic observations were generated from a nature run that uses the same biogeochemical and physical model as the DA systems. Thus, the setup is highly idealized, and the ML techniques can learn correlations between variables that directly relate to the dynamics that generated the observations. Consequently, the ML-based DA strongly benefits from accurately improving (some) unobserved variables. But is this still an advantage if the true unobserved variables do not follow model dynamics? I would suggest a follow-up experiment in which the DA techniques are confronted with "real" data.</p> <p>Various data are collected for the L4 station, satellite data could be used. Do the improvements in chlorophyll forecast skill seen for the synthetic data translate to improvements for real data compared to the more standard DA approaches?</p>
AC	<p>Thanks to the reviewer for their valuable comments.</p> <p>We agree that testing MLDA approaches using real observations is valuable, but we have intentionally chosen a perfect model scenario for this study.</p> <ul style="list-style-type: none">• Having a synthetic truth allows diagnostics that are not possible with real observations (e.g. Fig. 5, 7).• Real observations introduce model error into the problem and would make it difficult or impossible to disentangle model errors from the errors introduced by exploring ML. Results and their interpretation would then also be more model-dependent, which at this stage may impede the development of a ML-DA strategy. Note: even using real observations, the ML aspects would still be based on characteristics of the model used in the assimilation, so that aspect of our current work is not a surrogate for a real system.

	<ul style="list-style-type: none"> • If ensemble covariances (derived from the model) do not reflect reality well, then both ensemble DA and MLDA (as its emulator) will struggle. However, this is a broader challenge of DA system design and ensemble generation. While important, is not the focus of our current study. Our aim here is not to improve the fidelity of the model or the ensemble representation per se, but to examine the potential of ML to approximate the DA correction process. • While ML can be used to improve aspects of the model performance, there are generally trade-offs, and the unique dynamics and characteristics of a marine BGC model (unique compared to physical modelling such as atmosphere and ocean), mean that there is much to learn from applying ML techniques even in a synthetic scenario. <p>Applying these methodologies to real data would make an interesting direction for future work, and so we will include recommendation of this in the conclusions.</p>
CH01	<p>Conclusions, par1: We have made it clear in the conclusions that using “real” observational data would make for an interesting, and necessary direction for future work.</p> <p>“In this study, we used a synthetic “perfect model” setup as a necessary first step to explore ML-assisted DA under controlled conditions, but a natural direction for future work is to apply these approaches to real observations to assess their practical value.”</p>
RC02	<p>The authors single out zooplankton as unobserved variables that "do not update well" (Sec 5, par 3). However, it could be that this issue with zooplankton updates is due to the way the biogeochemical model is parameterized. Depending on the way zooplankton grazing, phytoplankton mortality and other processes are parameterized in the model, phytoplankton/PFTs and zooplankton can be more strongly or weakly linked. This effect could mean that the result that zooplankton, specifically is difficult to estimate, does not generalize to other model configurations. Yet, I agree with the authors that there can always be some variables that are difficult to estimate, perhaps even more so in a DA setup without synthetic observations.</p>
AC	<p>We shall add some qualification to make this distinction clear. Since, each component in this model is potentially parameterisation dependent, it is possible that the ability to update some variables well (or not) can be altered by parameterisations. In practice, new parameterisations would require new emulators, or an emulator that is trained on many different parameterisations to accommodate the range of possible dynamics. The latter issue could be addressed by</p>

	<p>“transfer learning”, an avenue that we have only marginally touched upon in this work when “transferring” the DAML from one location to the other. Transfer learning from one model configuration (i.e. parametrization) to another may be further and more extensively studied in a follow-on investigation.</p> <p>Using a controlled ground truth that is similar to the ensemble is a reasonable first step, and we should generally expect variables that are less closely linked, or less sensitive, to the observed quantity to be harder to update well (we will modify the conclusion to reflect this broader point, which in our case is the zooplankton).</p>
CH02	<p>Conclusions, point (b): We have added text to highlight this limitation, and potential avenues forward to handle it in future study: “However, this limitation may reflect the particular parameterisations of zooplankton–phytoplankton interactions, grazing, and mortality in the underlying BGC model, and may not generalise to other model setups. More broadly, we expect that variables less directly linked to or less sensitive to the observed quantity will be more difficult to update well. Since parameterisation choices can alter these relationships, new parameterisations would likely require retraining emulators, or alternatively, more flexible ML strategies such as transfer learning. Exploring such approaches in more diverse configurations remains an important avenue for follow-on investigations.”</p>
RC03	<p>When I first read the title, I was expecting an approach to better inform 3D biogeochemical models with observations. The abstract then mentions that 1D models are used -- which seems like a good choice given the complexity of 3D models. However, when I read that the DA was basically performed in 0D, I felt that an opportunity was missed to examine if ML approaches can yield improved spatial increments. Variational DA requires the specification of length scales, ensemble-based approaches typically rely on localization to limit the DA update spatially, but ML approaches could potentially learn how to best spread the increment spatially. But by eliminating all spatial dimensions from the DA and simply applying any DA update throughout and only in the mixed layer, this important DA aspect is not examined in this study at all. Why are only 0D increments used in this study, how much additional effort would be required to create a full 1D update?</p>
AC	<p>We have inherited this aspect from the UK Met Office (UKMO; this is what our “RUS” scheme represents). The UKMO system assimilates only on the surface and propagates increments through the mixed layer.</p> <p>Rather than 0D, arguably our DA system is 1D (which includes the vertical spreading of the increments, and the model run between DA cycles). We use an idealised structure function, with a vertical</p>

	<p>length-scale prescribed from the mixed layer depth, to update the 1D state of the model. This structure of our system allows us to focus on the multivariate aspect of this difficult problem.</p> <p>This is a reasonable choice for correcting the leading order of error as the mixed layer and sub-mixed layer in the vertical structure are generally decoupled (especially on shorter timescales).</p>
CH03	<p>We have addressed this on page 6 – restructuring large amounts of the text to improve the clarity. Specifically, we clarify the above point, with additional justification:</p> <p>“While the state vector only considers surface values (0D), the entire DA process itself is 1D, using an idealised structure function to spread increments uniformly throughout the mixed layer. This approach assumes that surface conditions are representative of the mixed layer as a whole (which is broadly true in this model configuration), so that increments derived at the surface also provide a reasonable correction at depth within the mixed layer. We do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the surface. Extending increments deeper would risk introducing spurious vertical structures, distorting stratification, or interfering with biogeochemical processes that are driven by different controls (e.g., remineralisation). By restricting updates to the mixed layer, the assimilation scheme ensures consistency with the available observations and avoids imposing unsupported corrections in the sub-mixed layer. Strategies for updating below the mixed layer may therefore require additional observation types (e.g., from floats or sea-gliders) or bias-correction approaches, rather than relying solely on surface observations combined with DA.”</p>
RC04	<p>Depending on the setup (and more obviously in a DA configuration with real observations), the surface-only DA approach could lead to worse performance of the DA systems. The study does not provide many specifics about the ensemble generation or the nature run, but if the model state below the mixed layer is sufficiently different in the data-generating nature run and the simulation used in the DA, then there is an error source that the DA system (with or without ML) cannot adequately correct. An example: Higher nitrate (or higher DOM, perhaps simply more C or N) below the mixed layer in the DA compared to the nature run could result in continuous overestimation and subsequent downward corrections of nitrate, chlorophyll and zooplankton in the mixed layer without addressing the underlying issue of too much nutrient input. This type of scenario could also manifest itself in zooplankton drifting away from the truth despite a series of beneficial corrective DA updates.</p>
AC	<p>This point is true as the surface observations cannot directly correct error below the mixed layer. However:</p>

	<ul style="list-style-type: none"> • This is true for all approaches tested in this study, and so any error from below the mixed layer has the capacity to impact the mixed layer in every run. • This is a reflection of many real systems. • It is unlikely that the sub-mixed layer can be corrected from the surface observations (real or otherwise) as the two layers are decoupled/weakly-coupled. <p>We agree that the manuscript could benefit from additional discussion on this. This will be an excellent opportunity to discuss the multi-faceted nature of this problem, and how additional observational capacity (e.g. from sea-gliders) could better correct the model by observing the sub-mixed layer.</p>
CH04	<p>Page 6: “We do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the surface. Extending increments deeper would risk introducing spurious vertical structures, distorting stratification, or interfering with biogeochemical processes that are driven by different controls (e.g., remineralisation). By restricting updates to the mixed layer, the assimilation scheme ensures consistency with the available observations and avoids imposing unsupported corrections in the sub-mixed layer. Strategies for updating below the mixed layer may therefore require additional observation types (e.g., from floats or sea-gliders) or bias-correction approaches, rather than relying solely on surface observations combined with DA.”</p>
RC05	<p>The authors trained their ML models in one location and show that this approach yields comparatively bad results when transferring the models to a new location. Here it would have been interesting to train the models on data from 2 or more (sufficiently different) locations to examine if these more generalized models would (1) perform similarly well on one of their training locations as the specialized models and (2) if the generalized training would make the models more transferable to new locations. This is likely beyond the scope of this study, but perhaps the authors could discuss this point.</p>
AC	<p>As the reviewer pointed out, this is beyond the scope of this study. However, we agree, and we will add some discussion for future study.</p> <p>Another next logical step to discuss would be to take a transfer-learning approach that we mentioned above, using the original trained model as pre-conditioning for training on the data of a new location (with the hope that the pre-conditioned model would require less new data than an unconditioned one).</p>
CH05	<p>We have added an additional paragraph on page 29 to address this: “An additional avenue worth exploring is whether training ML models on data from multiple, sufficiently different locations could improve</p>

	<p>their generalisability. Such an approach would allow for testing whether a more generalised model can (i) perform as well as a specialised model at its training locations, and (ii) transfer more successfully to new, unseen locations. While this is beyond the scope of the present study, it represents an interesting line of future work, particularly when combined with transfer learning strategies. For instance, a model trained at one location could be used as pre-conditioning for training at a new site, with the expectation that the pre-conditioned model would require less additional data to adapt effectively (e.g., Hu et al., 2016). Together, these directions highlight the importance of balancing specialisation and generalisation when scaling ML-assisted DA from idealized 1D configurations to realistic 3D systems.”</p> <p><i>Hu, Q., Zhang, R., and Zhou, Y. (2016). Transfer learning for short-term wind speed prediction with deep neural networks. Renewable Energy, 85:83–95.</i></p>
RC06	<p>As mentioned in the manuscript, the DA update in Eq. 1 represents the best linear unbiased estimator (BLUE) and the authors kept the ML-based DA approaches in the linear framework by estimating elements of Eq. 1. But one could imagine going beyond that and estimate nonlinear relationships between model state, observations, and increment, for example using neural networks. In how far do the authors think the DA framework presented here could be improved further using other ML techniques?</p> <p>https://pubs.aip.org/aip/cha/article/34/9/091104/3314756/Accurate-deep-learning-based-filtering-for-chaotic</p>
AC	<p>This is an interesting issue. The ML-OI model estimates a correlation, which is used to perform a ‘linear’ update to the system state. Even though the assimilation is based on the Kalman filter, the relationship between the state and the correlations is non-linear, and this is the relationship that the ML model must learn.</p> <p>Likewise in the work pointed to by the reviewer, our ML-EtE model does learn (in an end-to-end fashion) to predict an increment for an unobserved variable, based on the increment for the observed variable (and so includes the observational information) and the state. Again, this is potentially non-linear.</p> <p>The increments still ultimately come from the “analysis-background” of an EnKF, which is providing a “linear” update to the system. To truly go beyond this limitation, we would need to either train on increments to the truth (e.g. truth-background), or use a non-linear DA system, such as a particle filter. This also warrants some discussion in the conclusions of the paper.</p>
CH06	We add additional discussion on this idea.

	<p>Section 3.1, par 1: “(and while the EnKF update is linear, the relationship between state and correlation is likely non-linear)”</p> <p>Section 3.2, last par : “[...] while Bocquet et al. (2024) pursued a similar end-to-end replacement of the analysis step [...] Here, however, our primary goal is to examine the feasibility of ML-EtE and its ability to learn the EnKF updates effectively. Using this approach, the increments still ultimately come from the “analysis–background” of an EnKF, which provides a linear update to the system. Going beyond this limitation would require either training on increments relative to the true state (e.g., truth–background) or employing a non-linear DA system, such as a particle filter, to capture more complex, non-linear corrections.”</p>

Specific Comments

RC08	P 1, par 1: I would suggest changing "and lead" to "which can lead to".
	Agreed.
CH08	P1, par1: Changed as above.
RC09	P 2, par 1: "size-class chlorophyll [...] and other types of in situ data": This makes it sound like size-class chlorophyll is an in situ data product, when it typically is based on satellite estimates.
AC	Agreed. We will rephrase to: Other products are assimilated in reanalyses or research and development (R&D) versions of the operational systems, such as optical variables (refs...) and size-class chlorophyll (refs ...), as well as types of in situ data, such as chlorophyll, oxygen and nutrients from gliders (refs...).
CH09	P2, par1. Changed according to above.
RC10	P 2, par 2: "The multivariate updates can happen in the DA step, through ensemble-informed background covariances (as in the EnKF), or, through balancing schemes, such as the scheme of Hemmings et al. (2008) based on nitrogen mass conservation ...": In variational DA, multivariate updates also happen through the tangent-linear and adjoint models. However, this type of update has issues as well and can lead to unrealistic updates because there are typically no prescribed covariance terms between different variables. For example, in Mattern et al. (2017; DOI: 10.1016/j.ocemod.2016.12.002), the authors had to reduce the updates to unobserved nitrate in order to avoid unrealistic nitrate accumulation at the ocean surface.
AC	This is a good point, and we will add this point and reference to the manuscript.
CH10	P2, par 2 – we have added this point:

	The multivariate updates can happen in the DA step, through ensemble-informed background covariances (as in the EnKF), through balancing schemes, such as the scheme of Hemmings et al. (2008) based on nitrogen mass conservation applied to Nutrient-Phytoplankton Zooplankton-Detritus models (Hemmings et al., 2008; Ford et al., 2012), or through the tangent-linear and adjoint models Mattern et al. (2017).
RC11	P 3, Sec 2.1: "It provides a sufficient balance between realism and computational cost": How "sufficient" a 1D model is, may be very dependent on the application. I would suggest motivating it a bit better based on the goals of this study.
AC	Agreed. This current statement is too vague. We will better motivate this – realism refers to the use of real forcing data and relaxation profiles, as well as the full complexity of ERSEM (which can even run in 0D if the physical model was 0D). The computational benefits of running the 1D model vs the 3D model allow for a wider range of test scenarios (e.g. testing different update strategies) which can inform the strategies to test in a 3D system, which is much more expensive.
CH11	P3, Sec 2.1 – updated the text to reflect the above: “It provides a balance between realism and computational cost by using real atmospheric forcing data, relaxation profiles, and coupling at full BGC complexity, while sacrificing the explicit representation of 3D processes. This makes the system ideal for this work, where we are primarily interested in the error relationships between different biogeochemical quantities (e.g., chlorophyll to nitrate), rather than the spatial error characteristics.”
RC12	Eq. 1: There is no error here, but it looks like the Δx was meant to be included in Eq 1.
AC	Agreed. We will rephrase slightly to make it clear that we are just defining the analysis increment.
CH12	We have removed what was previously Eq (2), and added a notation brace to highlight the analysis increment (Δx) in Eq (1). This makes it much simpler to read.
RC13	P 5, Sec 2.4: " x^f is the background state": This notation is used in many studies but maybe for some who do not know it, either motivate the superscript "f" by mentioning "forecast" or change it to "b" to match "background".
AC	Agreed with above. Further adjustments to improve the clarity have also been made (see CH13 below).
CH13	We have made the decision to change equations to use " x^b ". A footnote is also added to the introduction: In data assimilation, the terms “forecast” and “background” are often used interchangeably. Strictly, the background refers to the forecast state used as the prior in the assimilation step.

	<p>We also specify at the point where the equations are introduced: “x^b is the background state (which in this work is equivalent to a 7-day forecast state)”</p> <p>Making this distinction is important, as we also analyse different forecast lead times (which are not background states, as they are not used in an assimilation step).</p>
RC14	<p>P 6, par 1: "In this work, the state of the system, both x^f and x^a, comprises the surface values of most pelagic variables in ERSEM.": Based on the abstract and previous text, I was expecting a 1D DA system, and not 0D. Reading a bit further, I see that this is not simply a 0D update, but that the full vector is updated but only within the mixed layer</p>
AC	<p>Yes, we can simplify the DA system based on the behaviour of the mixed layer. So, while the ML model predicts a 0D increment for each unobserved variable, we take the mixed layer depth as a length-scale and use an idealised structure function to propagate the increment through the vertical layer. So, the 1D model is updated beyond the surface (as also discussed in response to RC03).</p> <p>We will make this clearer at this earlier point in the manuscript to improve clarity for the reader.</p>
CH14	<p>We now clarify this in section 2.4:</p> <p>“While the state vector only considers surface values (0D), the entire DA process itself is 1D, using an idealised structure function to spread increments uniformly throughout the mixed layer. This approach assumes that surface conditions are representative of the mixed layer as a whole (which is broadly true in this model configuration), so that increments derived at the surface also provide a reasonable correction at depth within the mixed layer. We do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the surface.”</p>
RC15	<p>Eq. 4: Why is the summing of the 4 chlorophyll variables to total chlorophyll not included in H? That is, why doesn't H have the form [1, 1, 1, 1, 0, 0, ...] or similar, where the four non-zero entries are associated with the chlorophyll variables?</p>
AC	<p>We could indeed write the system in this way, and for updating the unobserved variables, and would be mathematically equivalent. However, on reflection, in our case it would provide a less concise formulation to describe the relationship between the observed quantity and the updated quantities (Eqs 7 and 8 become more cumbersome, and make it less clear as to what the ML model is estimating/predicting):</p> $K_i = \frac{\sum_{j=1}^4 P_{ij}}{\sum_{j=1}^4 \sum_{k=1}^4 P_{jk} + R}$

	While this is equivalent, it obscures the fact that we are interested in the relationship between total chlorophyll and the observed variables. Especially for those not familiar with DA, as pointed out in RC13.
CH15	Additional justification is added in section 2.4, on page 6/7: “In principle, one could keep only the PFT chlorophyll concentrations in the state and represent total chlorophyll via the observation operator. However, this would require explicitly summing across PFTs at every assimilation step, making the DA equations more cumbersome. Instead, we treat surface total chlorophyll directly as part of the state vector. This simplifies the presentation of the analysis equations, since the observation operator then reduces to a simple selection operator:...”
RC16	Eq. 7: Mention what "R" is here.
AC	Agreed, will add text: “and R is the observation error covariance”
CH16	Changed as above.
RC17	Table 2, line 3: I would consider a phrase like "ensemble-based" more useful than the "itself".
AC	Agreed. We will change accordingly.
CH17	Changed as above.
RC18	Eq. 9: This may be a Latex issue, but the "COR" looks very much like "CO*R" with the R from the previous equation. Why not use a lower-case rho, which is often used to denote correlations?
AC	Agreed. COR will be changed to lower-case rho.
CH18	Changed as above.
RC19	Eq. 9: Are these properties obtained from the long simulation or prescribed some other way? Reading on, I see that $COR_{i,c}$ is being estimated using the ML techniques. This could be made explicit here.
AC	This is outlined in Table 2, as the source for each term depends on the run/scheme. We will redefine the relevant sources of information at Eq. 9.
CH19	This is now outlined early in Section 2.3: Climatological correlations and variances are calculated using a free-run ensemble over the training period. The CWEC location uses a run spanning 2000 - 2010 to generate climatological statistics, and the online test is performed for (2022-2023). And reiterated at the relevant point when describing the ML-OI method: Together with climatologically estimated values of σ_i and σ_c (estimated using a free-run ensemble over the training period as described in Sec. 2.4)

RC20	Sec 3.2, par 1: "to predict the state-dependent correlations between observed and unobserved quantities in Eq. (9)." I would suggest adding the symbol (currently COR _{i,c}) so the reader knows immediately what is being estimated. Furthermore, I would suggest using "estimate" rather than "predict", as this is not done in a forecast sense.
AC	Agreed. We will add the symbol. "Estimate" also makes sense in this case, as "predict" is more typical of pure ML nomenclature.
CH20	Changed as above. We have also replaced other instances of "predict" with "estimate" where appropriate.
RC21	Sec 3.2, par 1: "is scaled according its climatological maximum": Does this mean that COR _{i,c} is estimated with data from all locations, and it is then rescaled for each location? A better description and some more information would be useful here. Apparently COR _{i,c} is location-specific, is some time-dependence assumed, if so, what kind? How many parameters need to be estimated here? Stating these basic facts early on helps the reader understand the main idea before going into implementation details.
AC	Yes, we will add more information to clarify. The ML-OI model is only trained on L4 data (as are all models in this work). The number of targets to be estimated is equal to the number of unobserved variables.
CH21	Additional information added to Sec 3.1 to clarify the above: "For each variable input into the ML-OI model to estimate the correlation, the background state is additionally divided by its climatological maximum at the corresponding location (before the regular standardisation procedure is applied to the data). This normalization accounts for differences in the amplitude of seasonal variability while making a bold assumption that the underlying correlative relationships between variables remain consistent across locations. Since correlations are dimensionless, this scaling does not affect their interpretation, and the subsequent analysis increments (which are constructed by combining the estimated correlations with the location-specific climatological variances) remain physically consistent." We also state earlier in Sec 3: "In each setup, the number of outputs to be estimated by the ML model corresponds to the number of unobserved variables. In the case of ML-OI (see Sect. 3.2), the standard deviations must also be estimated from climatology."
RC22	Sec 3.2, par 1: What motivates the use of the OI name here?
AC	Optimal interpolation is the standard name for a DA update scheme that is structured in this manner. Since we use ML to predict the correlation only, this 'structure' is essentially the same. This differentiates itself from the full "end-to-end", which emulates the entire increment rather than one term in an equation that calculates the increment.
CH22	No change required.

RC23	<p>P 9, par 2 "In ML-EtE, we assume that the essential properties of each statistical object that creates an analysis increment, such as covariances and observation uncertainty, can be more effectively captured by directly predicting the analysis increment rather than predicting every component individually and allowing errors to compound across multiple independent predictions that are then combined into a single value.": This sentence is too long and difficult to understand.</p> <p>Also, "each statistical object that creates an analysis increment, such as covariances and observation uncertainty" sounds like the covariances create an analysis increment and also the observation uncertainty creates an analysis increment. Please rephrase. But furthermore, if I understand this sentence correctly, it seems to argue for directly estimating the Kalman gain matrix rather than its components?</p>
AC	<p>We shall rephrase to make this clearer.</p> <p>Yes, the idea of end-to-end (EtE) emulation is that the error that is inherent to any neural network will be smaller if it is used to predict a single value, rather than the product of multiple predictions which will compound error.</p> <p>The EtE scheme does not quite estimate the Kalman gain (KG) matrix. Instead of transforming the innovation to assimilation increments (as a KG matrix would do), the EtE is designed to transform the analysis increment of total chlorophyll to analysis increments of other variables.</p> <p>We will make this clear in the revision.</p>
CH23	<p>Sec 3.2, par 2:</p> <p>"In ML-EtE, the analysis increment is predicted directly. By contrast, ML-OI predicts correlations, which are then combined with climatological standard deviations to form a Kalman gain, and only then applied to the innovation to yield the increment. The ML-OI scheme introduces potential sources of error - both from the uncertainty of data-driven correlation estimates and from the reliance on climatological statistics.</p> <p>Directly estimating the analysis increment (a vector) is also naturally more scalable for high-dimensional applications (e.g., operational 3D systems), where manipulating the full error covariance matrix - or even reduced or reformulated versions - becomes computationally costly."</p>
RC24	<p>Sec 3.4, par 1: "(1) a set-up where we choose to update only nitrate": I presume chlorophyll is updated as well? Maybe rephrase to "a set-up where the only unobserved variable that is updated is nitrate".</p>
AC	<p>Agreed.</p>
CH24	<p>Changed as above.</p>
RC25	<p>Sec 3.6.1: What are "cycles" here, and is the trajectory more than just a snapshot? Based on Eq. 10, one could assume a cycle is a time step.</p>

AC	Agreed this may not be clear. We will add clarifying text: For a system that runs for τ cycles (where a cycle represents a complete 7-day forecast and analysis) ...
CH25	Updated text to read: “For a system that runs for τ cycles (where a cycle represents a complete 7-day forecast and analysis)”
RC26	Sec 3.6.1: "The expected RMSE": Why not call it the "ensemble average RMSE"?
AC	Thank you for the suggestion but we think that may cause confusion with the RMSE of the ensemble average, which is a different quantity." We will think about a better name, but we hope we can leave it as it is if we cannot think of one.
CH26	No change required (we could not think of a better name).
RC27	Fig. 2: I think it is counter-intuitive to have nitrate shown in green and chlorophyll in black. I would suggest switching the colors. Also counter-intuitive, but maybe to a lesser extent: the light-limited time-period is shown in the brightest color. Finally, why show an unspecified "arbitrary" year instead of the climatology in the top panel, when the correlation is based on climatological values?
AC	We will switch the colours of nitrate and chlorophyll. We have also received comments from the Associate Editor on the suitability of colours for colour-blindness and so will review each plot accordingly. We agree that swapping the shading of the regimes would be a nice aesthetic change as well. We will also show the ensemble correlation for this particular year, to highlight the differences that can occur between the climatological values and the ensemble derived ones. This will link into Fig 3, which then shows the difference between the ensemble derived correlations, the climatological ones and the ML predicted ones.
CH27	We have changed the colours, and added the ensemble correlation example, as above.
RC28	P 12, par 1: Make it explicit that these RMSE values are for the correlation estimates.
AC	We will rephrase to: “outperforms the climatological correlation estimates”
CH28	This has been made clearer, with new text added to the caption: “The root mean square errors between the estimated and target correlations are given at the top of the figure, calculated separately over each regime window and for both estimation methods (with corresponding colour).”

RC29	P 12, par 3: Explain better what the terms offline and online are referring to here. I would suggest introducing the offline term at the very start of the section when the experiment is introduced.
AC	Agreed. We will move the text to make the difference between offline and online clearer.
CH29	We now state earlier in Sec. 2.3 that: “Offline refers here to a setup in which the ML-OI analysis is not then used as initial condition for the next forecast, and so it does not impact successive DA cycles. Conversely, online refers to a setup in which updates to the system can have dynamical impact on later DA cycles as the model integrates forward in time.”
RC30	P 12, par 3: "so that any update to the system can have dynamical impact on later DA cycles as the model": This sentence ends abruptly without a period.
AC	A formatting quirk caused this sentence to spill over to the next page, so it is understandable that this was easily missed. The full sentence reads: “This is done in an “online” setting, so that any update to the system can have dynamical impact on later DA cycles as the model integrates forward in time.”
CH30	No change required.
RC31	Fig. 4: The "relative ratio" label is not helpful, "relative to observational error" would be better. The title says "Nitrates"; there is a syntax error in the unit label of the second panel -- which also appears in Fig. 5.
AC	Agreed, the figure will be recreated with the suggested label; with a title that reads “Nitrate”; and the syntax error will be fixed (assuming the reviewer is referring to the ^ symbol).
CH31	Changed as above. See Figs 4 and 5.
RC32	Fig. 4: Why is the performance of the EnKF seemingly much worse than the RUS scheme for observed chlorophyll? Mention this in the text.
AC	We will mention this. The EnKF generates an ensemble of observations (with noise based on the uncertainty). This is going to introduce additional error relative to the single-model runs, which use the exact value of the observation. This means that when we calculate the error of each ensemble member, rather than the error of the ensemble mean, we get this difference in error. The key here is that the skill improves as the ensemble size increases. We will mention this in the text.
CH32	P16: “Since the EnKF generates an ensemble of observations (with noise based on the uncertainty), an additional source of error is introduced relative to the single-model runs which are not stochastic. This means that when we calculate the error of each ensemble member, rather than the error of the ensemble mean, we get this difference in error.

	We also see, in the right panel, that the error decreases with ensemble size for the unobserved nitrate, indicating that the system converges towards more correct nitrate updates at larger ensemble sizes.”
RC33	P 13, par 1: "as a percentage of the observation error": Are these truly percentage values?
AC	You are correct to point this out. These are ratios, not percentages. We shall amend the text.
CH33	Corrected as above.
RC34	P 13, par 1: "error. exceeds": There is an issue with the sentence.
AC	The first word of this sentence is missing. It shall be corrected to: “...error. This exceeds...”
CH34	Corrected as above.
RC35	Fig. 5: I find it difficult to see improvement in the plots, but perhaps I haven't stared at them long enough. I would think it would be more informative to show "forecast-truth" and "analysis-truth": it would clearly show when improvement occurs ("analysis-truth" closer to zero) and analysis-forecast is simply the space between the curves (which could be shaded in the figure).
AC	This plot shows increments (i.e., “analysis-background” and “truth-background”), as we are primarily interested in the increments, and if the increments from the DA scheme are of the correct size and direction. Since this reviewer suggestion is the same information, but reformatted, we can redo it.
CH35	We have redone the plot as suggested, showing the forecast-truth and analysis-truth. We have also shown the analysis RMSE for each seasonal regime, to highlight where the errors are the lowest. The text has been updated to reflect these changes to the plot.
RC36	Fig. 5 and others: Why aren't EnKF results shown for comparison?
AC	EnKF is primarily used for training and an initial benchmark for the nitrate only schemes. We feel that adding the EnKF to Fig 5 onwards could distract from comparing the single model runs.
CH36	No changes made.
RC37	P 17, par 1: "so values shown in Fig. 7 are RMSEs for 7-day forecasts relative to the RMSE of the RUS method": But Fig. 7 only shows a one value for each RMSE value, are these averages across several 7-day forecasts? Please explain better what is shown.
AC	Yes, the dot represents the average error at 7-day forecasts during the online testing period. We will improve the text to explain this better.
CH37	Texted added to clarify the text: “(averaged across the 104 forecast-analysis cycles of the entire test period)”

RC38	P 19: "(Sect. ??)"
AC	Typo in the manuscript latex and it will be corrected to point at what is currently "(Sect.3.6.2)".
CH38	Now correctly refers to the section (now 3.4.2).

Reviewer 2

In the following, RC = Reviewer's Comment, AC = Updated Authors' Comment, CH = Change

General Comments

RC39	<p>I feel that many choices of the authors - both in the definition of the algorithms and the configuration of the numerical experiments - could be better justified.</p> <p>Why building the update of the chemical component of the PFTs in that way?</p> <p>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?</p> <p>Why not updating the variables under the mixed layer?</p> <p>Why comparing the ML schemes to an univariate scheme and not to an EnKF?</p> <p>See specific comments for more details. I think that this paper would benefit from additional justifications.</p>
AC	<p>We thank the reviewer for their valuable comments and shall provide the detailed responses in the relevant "specific comments".</p>
CH39	<p>See responses and corresponding changes below.</p>
RC40	<p>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.</p> <p>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 highlight the remaining issues of their approaches</p>
AC	<p>We agree that the claim is arguably too strong and should be more carefully qualified.</p> <p>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 successfully partially or fully emulate updates from ensemble DA schemes, yielding many improvements to analysis state estimates over the operational-inspired reference univariate scheme (RUS)".</p>

CH40	<p>We have substantially increased the discussion, providing additional qualification to the statements specified in specific comments. The conclusions have also been modified to more accurately reflect these changes.</p> <p>Furthermore, additional diagnostics have been provided (e.g, Fig.8 in the new version) to provide more evidence for the relevant improvements.</p> <p>The key claim has also been modified to read: “We have shown that ML methods can make improvements to the DA schemes of marine BGC models when coupled to a 1D physical model - particularly in the shorter lead times of the training location.”</p>
RC41	<p>The extension of these approaches to 3D models does not seem to be straightforward. 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.</p> <p>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 occurrence 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.</p>
AC	<p>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 transformation, 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.</p>
CH41	<p>Multiple modifications or additions have been made throughout the paper to clarify that this problem is not solved during this work. Some of these are already mentioned in Sec 4.5: “However, state-of-the-art ensemble marine BGC systems are still limited in scale and may not (yet) accurately represent the statistics needed for multivariate DA (Sk ´akala et al., 2024). Also, this approach would need to</p>

	<p>be repeated if/when the observation network changes, which is likely given new observation missions and strategies (Telszewski et al., 2018)”</p> <p>We have also added additional discussion points on this – see P29: “Furthermore, ML models could handle local multivariate aspects at observation sites (a 0D transformation), while traditional DA methods (such as spatial correlations functions) manage 3D reconstruction (just as they manage the 1D reconstruction in our setup).</p> <p>[...]</p> <p>An additional avenue worth exploring is whether training ML models on data from multiple, sufficiently different locations could improve their generalisability. Such an approach would allow for testing whether a more generalised model can (i) perform as well as a specialised model at its training locations, and (ii) transfer more successfully to new, unseen locations. While this is beyond the scope of the present study, it represents an interesting line of future work, particularly when combined with transfer learning strategies. For instance, a model trained at one location could be used as pre-conditioning for training at a new site, with the expectation that the pre-conditioned model would require less additional data to adapt effectively (e.g., Hu et al., 2016). Together, these directions highlight the importance of balancing specialisation and generalisation when scaling ML-assisted DA from idealized 1D configurations to realistic 3D systems.”</p> <p><i>Hu, Q., Zhang, R., and Zhou, Y. (2016). Transfer learning for short-term wind speed prediction with deep neural networks. Renewable Energy, 85:83–95.</i></p>

Specific Comments

RC42	<p>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 positiveness 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 highlight 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 actually learn the “hidden, non-linear, and complex relations between BGC variables”. The authors should consider to rewrite the sentence to better highlight their contribution.</p>
------	---

AC	<p>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 the relationship between the system state and the correlation is non-linear. 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 - again, the relationship between the system state and this increment is non-linear. 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.</p>
CH42	<p>Sec 3.1 has been updated to clarify what is non-linear (i.e., flow dependence) and what is linear (i.e., the enkf updated). We also removed the word “hidden”:</p> <p>“Then, an ANN estimates the state-dependent correlations $\rho_{i,c}$ between observed and unobserved quantities in Eq.(8) as a function of the background state (and while the EnKF update is linear, the relationship between state and correlation is likely non-linear).”</p> <p>We also make clarification in sec 3.2: “Using this approach, the increments still ultimately come from the “analysis–background” of an EnKF, which provides a linear update to the system (even if the relationship between the state at the analysis increments is non-linear). Going beyond this limitation would require either training on increments relative to the true state (e.g., truth–background) or employing a non-linear DA system, such as a particle filter, to capture more complex, non-linear corrections.”</p>
RC43	<p>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.</p>
AC	<p>Agreed. We will specify this.</p>
CH43	<p>Specified as suggested: “(nitrate, phosphate and silicate)”</p>
RC44	<p>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</p>

	<p>variables below the mixed layer is only due to the adjustment 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 occurrence 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</p>
AC	<p>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.</p> <p>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.</p>
CH44	<p>We have added text, in Sec 2.3, to better highlight the limitation that using the relaxation profiles can cause:</p> <p>“However, the relaxation profiles could contribute to controlling the sub-mixed layer in our setup (which is not updated during assimilation), which could help to mitigate some long-term biases in these areas. This is a potential limitation for operational scale systems which do not have or use these relaxation profiles.”</p>
RC45	<p>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.</p> <p>I think that a comment should be added to highlight that point.</p>
AC	Agreed.
CH45	<p>See CH44, and we also make the following adjustment to another part of the section:</p> <p>“...prevents ensemble collapse (at least within the mixed layer) induced by the previously mentioned nutrient relaxation, or lack of representation of error growth processes like horizontal advection which are absent in a 1D set-up.”</p>
RC46	<p>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</p>

	types (PFT, state variables). This is quite unusual regarding 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 (reviewer 1).
CH46	Additional justification is added in section 2.4, on page 6/7: “In principle, one could keep only the PFT chlorophyll concentrations in the state and represent total chlorophyll via the observation operator. However, this would require explicitly summing across PFTs at every assimilation step, making the DA equations more cumbersome. Instead, we treat surface total chlorophyll directly as part of the state vector. This simplifies the presentation of the analysis equations, since the observation operator then reduces to a simple selection operator:...”
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.
CH47	Added text on P7, following eq.5: “This constrained redistribution scheme (Teruzzi et al., 2014; Skakala et al., 2018) ensures that phytoplankton updates preserve forecast stoichiometric ratios and remain physiologically consistent, rather than

	<p>allowing the EnKF to update PFTs freely. While an unconstrained EnKF would eventually converge towards similar balances, this explicit approach guarantees that assimilation respects acclimation dynamics and maintains the community structure of the model. Importantly, the same ratio-based balancing scheme can also be applied outside an ensemble framework in single-model runs, where it provides a consistent way of updating PFTs from a total chlorophyll correction.”</p>
RC48	<p>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.</p>
AC	<p>Propagating surface increments through the mixed layer is used operationally, and we see that the behaviour of this model (or at least this location’s configuration) means that the values are homogeneous within the mixed layer (and so we could use the strategy described above, but would get the same result).</p> <p>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).</p>
CH48	<p>Added text to page 6: “This approach assumes that surface conditions are representative of the mixed layer as a whole (which is broadly true in this model configuration), so that increments derived at the surface also provide a reasonable correction at depth within the mixed layer. We do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the surface. Extending increments deeper would risk introducing spurious vertical structures, distorting stratification, or interfering with biogeochemical processes that are driven by different controls (e.g., remineralisation). By restricting updates to the mixed layer, the assimilation scheme ensures</p>

	consistency with the available observations and avoids imposing unsupported corrections in the sub-mixed layer. Strategies for updating below the mixed layer may therefore require additional observation types (e.g., from floats or sea-gliders) or bias-correction approaches, rather than relying solely on surface observations combined with DA.”
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.
CH49	No change made.
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 Δx_i correspond to PFT only?
AC	This should read: Eq.(8)
CH50	Updated to refer to what is now Eq.(7). The caption is also updated to specify the index i: “Index chl refers to total chlorophyll, while index i refers to an unobserved variable (e.g., nitrate).”
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	This is done so that there is a 1-to-1 correspondence between the strategy used to generate the training data, and that which will be applied in the single-model run schemes.
CH51	We have added text to Sec. 2.4.2 to specify the reasoning for this: “This is done so that there is a one-to-one correspondence between the strategy used to generate the training data, and the strategy applied in the single-model 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.
CH52	Removed as suggested.

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.
CH53	We have removed the subsection title as suggested.
RC54	p.8, second paragraph "where $P_{f,i,c}$ is the background error [...] where $CO_{Ri,c}$ 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	We will make the terminology consistent.
CH54	<p>While in most places, they are interchangeable, we have added a footnote to clarify some technical differences: In data assimilation, the terms "forecast" and "background" are often used interchangeably. Strictly, the background refers to the forecast state used as the prior in the assimilation step.</p> <p>We also specify at the point where the equations are introduced: "x^a is the background state (which in this work is equivalent to a 7-day forecast state)"</p> <p>Making this distinction is important, as we also analyse different forecast lead times (which are not background states, as they are not used in an assimilation step).</p>
RC55	p.8, fourth paragraph "is scaled according its climatological" → is scaled according to its climatological (?)
AC	Agreed.
CH55	<p>We have rewritten this for clarity: "For each variable input into the ML-OI model to estimate the correlation, the background state is additionally divided by its climatological maximum at the corresponding location (before the regular standardisation procedure is applied to the data). This normalization accounts for differences in the amplitude of seasonal variability while making a bold assumption that the underlying relative relationships between variables remain consistent across locations. Since correlations are dimensionless, this scaling does not affect their interpretation, and the subsequent analysis increments (which are constructed by combining the estimated correlations with the location-specific climatological variances) remain physically consistent. Since variances are single-variable statistics that can be estimated more reliably from long climatological records, we assume they are sufficiently robust to provide a stable basis for use in the Kalman gain. In contrast, correlations describe joint variability and therefore require the more</p>

	sophisticated, data-driven approach described above, and cannot be approximated in the same way.”
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 correlation 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 the operational system which motivates this work).
CH56	See CH48.
RC57	p.8, last paragraph ”For the first application of ML-OI in Sect 4.2, the target [...] between total chlorophyll 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	<p>Nitrogen makes a good first choice of unobserved variable to test the system as it is the limiting variable: <i>“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).”</i></p> <p>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).</p> <p>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).</p> <p>We will make this motivation clearer in the revision.</p>
CH57	<p>This has been expanded and made clearer (this also links in with suggestions of further comments):</p> <p>“We choose nitrate for these initial experiments because it is a limiting nutrient at the L4 location (see Fig.A.3 in the Appendix and Smyth et al, 2010), and therefore has a clear, explainable relationship with total chlorophyll as discussed in Sect.4.1 (see also Fig.2). Since nitrate is the key driver limiting primary production among nutrients, addressing it through DA could have a significant knock-on effect on the whole model state (through dynamical evolution from the corrected state).</p> <p>Moreover, this also provides us a more understandable proof-of-concept with reduced complexity to analyse initially, before we later extend the updates to more than 30 additional pelagic variables (see Table 1) in a higher complexity scenario.</p>

	However, as will become clear in Sec.4.3, this strategy for assigning importance or priority to variables in the DA scheme does not necessarily correspond to the dynamical importance of a variable in the model, highlighting the need for specific results.”
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 mentioned 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.”
CH58	Added text to Sec 3.2: “Directly estimating the analysis increment (a vector) is also naturally more scalable for high-dimensional applications (e.g., operational 3D systems), where manipulating the full error covariance matrix - or even reduced or reformulated versions - becomes computationally costly.”
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 RC48, RC58.
CH59	Added text to page 6: “This approach assumes that surface conditions are representative of the mixed layer as a whole (which is broadly true in this model configuration), so that increments derived at the surface also provide a reasonable correction at depth within the mixed layer. We do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the surface. Extending increments deeper would risk introducing spurious vertical structures, distorting stratification, or interfering with biogeochemical processes that are driven by different controls (e.g., remineralisation). By restricting updates to the mixed layer, the assimilation scheme ensures consistency with the available observations and avoids imposing unsupported corrections in the sub-mixed layer. Strategies for updating below the mixed layer may therefore require additional observation types (e.g., from floats or sea-gliders) or bias-correction approaches, rather than relying solely on surface observations combined with DA.”
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	The periods for the train and test are different. 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.
CH60	<p>This has now all been clarified in Sec 2.3:</p> <p>"For the purposes of training ML models, and generating climatological statistics, time periods for the L4 location are partitioned as follows: training data (2000-2014), validation (2015-2017), offline test (2018-2020), online test (2022-2023). Offline refers here to a setup in which the ML-OI analysis is not then used as initial condition for the next forecast, and so it does not impact successive DA cycles. Conversely, online refers to a setup in which updates to the system can have dynamical impact on later DA cycles as the model integrates forward in time. Climatological correlations and variances are calculated using a free-run ensemble over the training period. Because the forecasts extended 7 days into the future, the number of available samples for any specific calendar day was limited, even though forecasts were issued throughout the full training period. To address this, the climatological statistics were computed as daily values, defined by averaging the statistic for a given day across all years in the training set. To increase the sample size and smooth out variability, a ± 30-day window around each calendar day was applied.</p> <p>The CWEC location uses a run spanning 2000 - 2010 to generate climatological statistics, and the online test is performed for (2022-2023). No validation or offline test period is required for CWEC in this work, as we are primarily interested in a "naive transferability" of the model trained and evaluated at L4."</p>
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).
CH61	This has been merged as described above.
RC62	p.9, seventh paragraph "Each ML approach is tested in the following scenarios [...] results of (2a)": The authors may add a comment to explain the motivation of this splitting.
AC	<p>We can make this splitting clearer in the manuscript.</p> <p>We first tested nitrate it is the key limiting variable (as shown by the regimes, 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</p>

	additional experiment that aimed at updating all variables, except the clearly problematic ones.
CH62	We now clarify in the preamble of Sec 3: The progression from (1) to (2a) allows us to move from a controlled test of a single key limiting variable to a comprehensive update of the full system, while (2b) represents a refinement step that is only possible after evaluating the performance of (2a). In this way, the experimental design not only tests the limits of updating all variables, but also demonstrates how excluding problematic variables can improve robustness without discarding the broader benefits of multivariate updates.
RC63	p.10, fourth paragraph "MGBC": Could the authors explain this acronym (first occurrence)?
AC	Typo. Should read: "marine BGC"
CH63	Fixed as above.
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.
CH64	Added a footnote to clarify this point in Sec 4.1: "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 be meaningful for the multivariate case updating all the nutrients. Fig.A2 highlights correlations

	<p>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).</p> <p>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 already done (nitrate is a limiting nutrient) while questioning the relevance of this approach and the need for specific results.</p>
AC	<p>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.</p> <p>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).</p> <p>Also see response to RC57.</p>
CH65	<p>As with CH57:</p> <p>“We choose nitrate for these initial experiments because it is a limiting nutrient at the L4 location (see Fig.A.3 in the Appendix and Smyth et al, 2010), and therefore has a clear, explainable relationship with total chlorophyll as discussed in Sect.4.1 (see also Fig.2). Since nitrate is the key driver limiting primary production among nutrients, addressing it through DA could have a significant knock-on effect on the whole model state (through dynamical evolution from the corrected state).</p> <p>Moreover, this also provides us a more understandable proof-of-concept with reduced complexity to analyse initially, before we later extend the updates to more than 30 additional pelagic variables (see Table 1) in a higher complexity scenario.</p> <p>However, as will become clear in Sec.4.3, this strategy for assigning importance or priority to variables in the DA scheme does not necessarily correspond to the dynamical importance of a variable in the model, highlighting the need for specific results.”</p>

	<p>Additional discussion of Fig.7 has also been added, comparing the difference between nitrate only, and updating all:</p> <p>“Figure. A.2 shows that correlations between surface total chlorophyll and silicate (or phosphate) are as strong as those with nitrate. This suggests that updates of similar magnitude might be expected for the other nutrients as well. However, it is not straightforward to infer how the system would evolve when all nutrients are updated simultaneously, based solely on the behaviour observed when only nitrate is updated. For example, Figure 7 reveals clear differences in the forecast and analysis RMSE for nitrate between the two ML-EtE configurations: one updating only nitrate and the other updating all nutrients. In the case for ML-EtE (ALL), updating all nutrients degrades the nitrate RMSE in both forecast and provides negligible impact on the analysis, while improving the analysis RMSEs for phosphate, ammonium and silicate. This outcome is notable, as it may point to assimilation biases introduced by the choice of update strategy.”</p>
RC66	<p>p.11, first paragraph “Offline refers 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?</p>
AC	<p>Correct. This is entirely “offline”, so no updates are actually made to the system – we just see what the update “would’ve been”.</p>
CH66	<p>No change required.</p>
RC67	<p>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.</p>
AC	<p>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.</p>
CH67	<p>Discussion of the plot has been modified. Instead of a comparison of RMSEs across the entire offline test period, we instead split the RMSE values across the different regimes so that the improvements at different times in the seasonal cycle can be compared. Also importantly, we make the clarification for interpreting these RMSE values: “It is important to note, however, that the RMSE here only reflects the capability of a method to estimate the ensemble-derived chlorophyll-nitrate correlations. Such improvements do not necessarily translate directly to</p>

	performance in an online, cycled data assimilation system (shown later) where past estimates can influence future states through the model's dynamical evolution.”
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 variability 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.
CH68	Discussion of this added, P16, to specify the link (or lack of) between the correlations at this time of year: “However, both the climatological and ML estimate fail to capture these possible states. During this time, total chlorophyll and nitrate are generally decoupled and there is no clear link between the state of the system and the correlations estimated. Furthermore, as the ensemble concentrations in total chlorophyll are very near zero, the DA updates are also small and have very little impact. This means that any improvement or degradation in correlation estimates at this time of year are less likely to result in any great improvement to the system, as there is weak relationship between chlorophyll and nitrate DA increments. However, updates at the start of this period can be important, as the resulting store of nitrate in the upper water column could have dynamical impact in later DA cycles when light is no longer limiting and the next bloom period starts.”
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 (i) calculating the average error of ensemble members, rather than the error of the ensemble mean, and (ii) the adding of observation noise to the ensemble of observations used in the stochastic EnKF. The single model runs do not need to add noise.
CH69	<p>Added some additional explanation for this difference: “Since the EnKF generates an ensemble of observations (with noise based on the uncertainty), an additional source of error is introduced relative to the single-model runs which are not stochastic. This means that when we calculate the error of each ensemble member, rather than the error of the ensemble mean, we get this difference in error. We also see, in the right panel, that the error decreases with ensemble size for the unobserved nitrate, indicating that the system converges towards more correct nitrate updates at larger ensemble sizes.”</p>
RC70	<p>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.</p>
AC	<p>We will expand the 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.</p>
CH70	<p>This plot has been revised, to better reflect the relevant statistics. In the previous version, the standard deviation of the EnKF was showing the standard deviation in the mean error of ensemble members in each EnKF run. This has been updated to show the standard deviation of error in</p>

	<p>ensemble members aggregated across all EnKF runs. This is a more accurate comparison between the single run models and the EnKF, when considering the spread in error.</p> <p>We also make note of the reduced spread in ML-EtE errors: “The standard deviation of ML-EtE is also lower than the spread for ML-OI, which is a good sign of lower sensitivity.”</p>
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.
CH71	We have put this information on the Y label of Fig.5.
RC72	<p>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?</p>
AC	<p>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).</p> <p>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.</p>
CH72	<p>We have updated this plot to show the difference to truth of both forecast and analysis, rather than the increments in the previous version.</p> <p>We have also included analysis RMSE values for each regime period, to give more detailed analysis of where each method is better or worse, compared to others.</p> <p>This has led to the revised discussion of Fig.5, on page 19 of the diff.</p>
RC73	<p>p.15, second paragraph ”We can also see that each approaches make little to no adjustment during the lighth-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.</p> <p>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 lighth-limited period when they occur? It may be interesting to see the evolution of the increment in total chlorophyll as well.</p>

AC	<p>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 temperature and total chlorophyll. We will add discussion on these points.</p>
CH73	See CH68.
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.
CH74	Included this information in the caption for Fig.6.
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 result of the RUS and EnKF as benchmarks.
AC	<p>As the scale for the error fluctuates according to the time of year, showing the error relative to another scheme (especially a representation of the current operational scheme) is logical in this case.</p> <p>As with response to RC36, 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.</p>
CH75	No change required.
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.
CH76	Included this information in the caption for Fig.7.
RC77	p.17, first paragraph "Before discussing 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.
CH77	Added text to as suggested: “It also slightly degrades the ammonium and silicate concentrations that are not updated during the analysis.”
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 also points out, we originally provided 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.
CH78	We have added a qualification to the paragraph to not overstate the scheme, but also point towards some additional analysis based on forecast at different lead times (new Fig. 8): “We emphasise that while many of the 7-day forecast RMSEs are similar or even degraded compared with RUS, many of the benefits from an improved analysis persist over a significant portion of the forecast window (as is shown and discussed later).”
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. 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).
CH79	We have added additional discussion on this point: “Figure A.2 shows that correlations between surface total chlorophyll and silicate (or phosphate) are as strong as those with nitrate. This suggests that updates of similar magnitude might be expected for the other nutrients as well. However, it is not straightforward to infer how the system would evolve when all nutrients are updated simultaneously, based solely on the

	behaviour observed when only nitrate is updated. For example, Figure 7 reveals clear differences in the forecast and analysis RMSE for nitrate between the two ML-EtE configurations: one updating only nitrate and the other updating all nutrients. In this case, updating all nutrients degrades the nitrate RMSE in both forecast and analysis, while improving the RMSE for ammonium and silicate. This outcome is notable, as it may point to assimilation biases introduced by the choice of update strategy.”
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 highlights assimilation biases due to the inconsistency of the updates that accumulate over time. I think that the authors should 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.
CH80	We have qualified the statement: “This method is also shown to be somewhat effective, generally providing similar behaviour to the ML-EtE (ALL) scheme, or at least reducing the RMSE from forecast to analysis (even if it is still worse than the RUS in some cases). It does suffer from the same difficulty in estimating zooplankton updates, to an even greater degree, which causes some further imbalance in the system.”
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.
CH81	We have added Fig.8, which shows the time evolution of the forecast error for a representative set of pelagic variables (page 23 of diff).

	This clearly shows how some of the benefits from the analysis persist over the shorter lead times.
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.
CH82	We have updated the summary paragraph for this: "In summary the experiments from Figs. 7 and 8 show that the impact of DA analysis updates on the model forecast is not straightforward due to the non-linear and complex nature of the BGC model. Although in general it is true that increasing the number of updated variables benefits the forecasts, especially at shorter lead times (e.g., less than 5 days lead time), this is definitely not true for every variable (which can often show similar or worse forecast RMSE at 7-day lead time than RUS) and highlights the need for specific results on each update strategy."
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 meaning of the correlation variables. We will change the text to read: "...which correspond to the correlations between total chlorophyll and each unobserved variable"...
CH83	Changed according to above.
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)".
CH84	Changed according to above (sec 3.4.2 in the new version).
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.
CH85	We added this qualification:

	“ Such a Shapley analysis also has the potential benefit in reducing the number of features needed for training future models (though this was not the aim of our experiments).”
RC86	p.21 ”This is likely because the correlations predicted by the ML-OI scheme represent a more locationn-agnostic 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.
CH86	We have added the qualification: “This is likely because the correlations estimated by the ML-OI scheme represent a more location-agnostic relationship in the marine BGC variables, which can be used in combination with the climatological variances of CWEC to produce more location-appropriate increments (though this is not true for ammonium, bacteria and labile DOM, which produce a worse analysis than forecast).”
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.
CH87	We have qualified the statement to reflect this: “Furthermore, this scheme still struggles to estimate zooplankton correlations, and so not updating the zooplankton as in the ML-OI (Excl. Zoo) scheme (purple) reduces the damages compared to ML-OI (ALL) - with any improvements being marginal at best.”
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.
CH88	“Again, we have qualified the statement: In both ML-OI (ALL) and ML-OI (Excl. Zoo), we see that the analysis for detritus is generally improved relative to the RUS scheme (though these improvements are marginal, and of similar magnitude to the worsened forecasts).”
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.
CH89	We have qualified the statement to be less optimistic, but it is also now supported by the new Fig.8: “We have shown that ML methods can make improvements to the DA schemes of marine BGC models when coupled to a 1D physical model - particularly in the shorter lead times of the training location.”
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 nitrate 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.
CH90	We have qualified the statements, to make it clear what is being referred to: “With machine learning (ML), we achieve improvements over climatological statistics in the nitrate-only update framework. Our analysis of ML-estimated nitrate updates illustrates that the ML methods behave in a largely coherent and meaningful manner. While ML can degrade forecast skill in some unobserved variables compared to RUS or nitrate-only ML schemes, they nonetheless show promise, with their usefulness in more complex assimilation settings requiring further assessment.”