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

Copy of review comments from “Reviewer 1” (RC01-RC38) are given below, followed by the author comments (AC).

Reviewer 1:

General Comments

RC01	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. 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?
AC	Thanks to the reviewer for their valuable comments. We completely agree that testing MLDA approaches using real observations is valuable. However, we have intentionally chosen a perfect model scenario for this study as a first step towards hybridising or substituting DA with ML (either by learning the full increments of the DA scheme, or learning the model correlations). Having a synthetic truth allows us to produce diagnostics that are not possible with real observations (e.g. Fig. 5, 7). Using real observations introduces 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, something that at this stage may obstacle the development of a ML-DA strategy. Note that 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. We also note that 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, and 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. Furthermore, while DA/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. 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.
RC02	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.
AC	This is an important point that is not currently well specified in the manuscript, and so 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 issues/challenges could be addressed or attempted to be addressed by transfer learning, a venue that will have only marginally touched upon here when moving from one location to the other. It may be further and more extensively studied in a follow-on investigation. 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).

RC03	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?
AC	We have inherited this aspect from the UK Met Office (UKMO; this is what our RUS scheme represents). The operational UKMO system is based on a scheme that assimilates only on the surface and propagates increments through the mixed layer. Rather than 0D, arguably our DA system is 1D (which includes the vertical spreading of the increments, and the model run between DA cycles). It can be partly expressed as a 0D problem only because all observations are at the surface. We essentially use an idealised structure function, with a vertical length-scale prescribed from the mixed layer depth, to update the 1D state of the model. The 0D/1D structure of our system allows us to focus on the multivariate aspect of this difficult problem (not the length-scales). 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).
RC04	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.
AC	This point is true in the sense that approaches in this study cannot directly correct error below the mixed layer. However: • This is true for all approaches tested in this study (which also use nutrient relaxation profiles that will control the sub-mixed layer – though we realise this is a limitation to mention as operational

	forecast systems will not necessarily use relaxation profiles), and so any error from below the mixed layer has the capacity to impact the mixed layer in every run. Moreover, updating the mixed layer only is the approach taken in many operational systems. • It is unlikely that the sub-mixed layer can be corrected from the surface observations (real or otherwise, or at least, it may not be logical to do so) as the two layers are decoupled/weakly-coupled. We agree that the manuscript could benefit from some additional discussion on this area. In fact, this will also be an excellent opportunity to discuss the multi-faceted nature of this particular problem, and how additional observational capacity (e.g. from sea-gliders) could serve to better correct the model (by observing the sub-mixed layer), rather than refinement of DA schemes and surface observations. We will also add some discussion on bias correction.
RC05	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.
AC	As the reviewer pointed out, this is beyond the scope of this study. However, we agree, and we will add some discussion on these ideas, as they would be interesting experiments for future study. Another next logical step to discuss would be to take a transfer-learning approach, 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).
RC06	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? https://pubs.aip.org/aip/cha/article/34/9/091104/3314756/Accurate-deep-learning-based-filtering-for-chaotic
AC	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.The ML-EtE model does learn 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. 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.

Specific Comments

RC08	P 1, par 1: I would suggest changing "and lead" to "which can lead to".
	Agreed.
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.
	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...).
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.
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 a little 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.
RC12	Eq. 1: There is no error here, but it looks like the Δx was meant to be included in Eq 1.
AC	Agreed, there is no error. We will rephrase slightly to make it clear that we are just defining the analysis increment.
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	We will use the notation “x^f” as it is widely used. However, we will specifically mention this alternative name when defining equation terms: “the background state (sometimes referred to as the forecast state)”
RC14	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
AC	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). We will make this clearer at this earlier point in the manuscript to improve clarity for the reader.
RC15	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?
AC	We agree that we could write the assimilation system in this way, and for updating the unobserved variables, this would be mathematically equivalent to our notation. 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): $$K_i = \frac{\sum_{j=1}^{4} P_{ij}}{\sum_{j=1}^{4}\sum_{k=1}^{4} P_{jk} + R}$$ 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. This approach would differ when updating the PFTs where, to preserve model-consistent phytoplankton physiology during assimilation, we adopt a constrained update scheme for the biogeochemical state variables, rather than allowing the ensemble Kalman filter (EnKF) to update all PFT-related state variables freely. 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. We would expect the EnKF to converge onto something similar as this is how the variables are related in the forecast, but this ensures that assimilation respects the acclimation dynamics and preserves the community structure and physiological integrity of the model. By updating only the total concentrations while maintaining forecasted stoichiometric ratios, we avoid introducing biologically inconsistent changes that may degrade model realism and performance. When considering the implementation of a MLDA hybrid system, it makes sense to rely on one of the few known relationships in the marine BGC model and only learn the ensemble updates, for use in single model runs, for the variables that have no such reliable assumptions.
RC16	Eq. 7: Mention what "R" is here.
AC	Agreed, will add text: “and R is the observation error covariance”
RC17	Table 2, line 3: I would consider a phrase like "ensemble-based" more useful than the "itself".
AC	Agreed. We will change accordingly.
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.
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.
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.
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.
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.
RC23	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. 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?
AC	We shall rephase to make this clearer. 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. 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. We will make this clear in the revision.
RC24	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".
AC	Agreed.
RC25	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.
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) …

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.
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.
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”
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.
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.”
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).
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.
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.
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...”
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. The largest improvements come during the bloom (which is a key BGC event; though this may be difficult to see with the 3-year period squashed horizontally) and nutrient-limited periods. We may split an RMSE statistics across each regime, rather than across the entire 3-year period, to better highlight the performance during each regime in this plot. Since this reviewer suggestion is the same information, but reformatted, we can redo it.
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.

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.
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)”.