

Thank you, once again, for taking the time and care to provide valuable feedback and contributions to this manuscript. Please see our responses to the comments below, along with the changes to be implemented in the new version. Copy of comments from both the editor (EC) and reviewer (RC) are given below, followed by the author comments (AC). We have also made some minor grammatical changes to improve readability, which are visible in the “diff” file.

Editor Comments

EC	Many thanks for submitting a revised manuscript, which has now been seen by one of the referees again. I would like you to address the referees comments in a revised manuscript and improve the readability of the text, following the referees suggestions.
EC01	Additionally, In response to the referees comment on DA results and synthetic vs real observations, I would like you to extend the discussion on lines 841-843 (line numbers referring to the revised manuscript) further explaining the potential differences that may arise from using real world observations in a next step.
AC	Added some additional discussion on this: “Given the dominant role of the total chlorophyll analysis increment in ML-EtE (ALL), it is important to note that the influence of observations could differ when using real observations rather than synthetic ones, owing to potential differences in error characteristics, representativeness, or systematic biases. We would still expect the analysis increment of the observed variable to exhibit comparable importance if trained on increments derived from real observations, as demonstrated in this experiment. However, because of its strong influence, the overall structure and nature of the resulting updates may change, reflecting the impact of more complex and realistic observational error structures.”
EC02	Additionally, I would like you reconsider the colour use of Figure 6, as the parallel use of red and green colours might challenge visually impaired or colourblind readers. If you would like to keep the colours, an alternative option would be to use different symbols per line.
AC	In response, we have changed the red in several plots to a “magenta/pink” that should be more easily interpreted for colourblind readers.
EC	Overall, I believe these changes are of minor nature.

Reviewer Comments

General

RC	At this point I have mainly one major comment that is regarding the accessibility/readability of the manuscript text. The manuscript is difficult to follow in places and at times I had to reread sentences several times to understand their intended meaning.
RC01	Some sentences are convoluted and difficult to follow, for example, "Also, this indicates that the improvement in correlation estimation in the offline experiment of Fig. 3 translates (at least on average) into the online testing period, where the updates of a given DA cycle feed into subsequent cycles." (l. 658) Here, it would be helpful to put the main message of the sentence front and center.
AC	Updated as suggested: "ML-EtE shows similar improvements during the online testing period as those observed in the offline experiment (Fig.3), indicating that the benefits persist even when the updates from each DA cycle feed into subsequent ones. Moreover, ML-EtE exhibits a lower standard deviation than ML-OI, indicating reduced sensitivity."
RC02	Other sentences seem to just restate what was described before: "This exceeds a value of 1 as we are measuring expected error of ensemble members, not error to the ensemble mean, as described in Sect. 3.4.1. ... 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." (l. 640).
AC	Re-worded for clarity: "Figure 4 displays the performance of the EnKF, the RUS scheme, the climatological statistics scheme (CliC), and the ML schemes. Each panel shows the mean expected error of ensemble members for ensemble sizes ranging from 4 to 96. In the left panel for total chlorophyll, the relative RMSE is calculated as a ratio of the observation error. The EnKF achieves near-optimal performance for ensemble sizes greater than 16, after which the mean expected error reaches a plateau. The relative analysis error of total chlorophyll exceeds a value of 1 because it is based on the expected error of ensemble members, not the ensemble mean (see Sect. 3.4.1). This reflects the additional stochastic error introduced by the EnKF's generation of perturbed observations."

RC03	There are also instances where it is still not entirely clear to me what is meant: "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." (l. 578) Here, I am not sure what exactly is meant by "specific results", and the sentence again seems a bit convoluted.
AC	Revised for clarify: "However, as will become clear in Sect. 4.3, this strategy for assigning importance to variables in the DA scheme does not necessarily reflect their true dynamical role in the model, highlighting the need to evaluate how different assimilation strategies and variable-update choices affect performance."
RC--	I have highlighted other instances below; overall I would encourage all coauthors to go through the text again with fresh eyes and rephrase sentences that may be difficult to understand for readers, including those without extensive knowledge of BGC DA.
AC	See specific comments below.

Specific comments

RC04	L 99: "non-linear/flow-dependent relations": This statement seems to imply that flow-dependent is equivalent to non-linear. Because this term is used throughout the manuscript, it would be useful to define and explain it briefly.
AC	Changed to clearly separate/explain these terms, so there is no implication that these terms are equivalent: "In this work, we investigate whether ML can capture the complex, non-linear relationships among biogeochemical (BGC) variables, and how these relationships depend on the evolving state of the system (i.e., are flow-dependent). Here, flow-dependent indicates that the statistical relationships or errors between variables vary with the current flow or dynamical state, rather than being constant or solely climatological. The relationships learned by the ML model are then used within a DA scheme, or as a full substitute for it."

RC05	L 101: "Thus, we are not attempting to emulate or improve BGC models via ML, ...": But of course the goal is to improve the estimates of BGC models -- perhaps not directly, but indirectly through improvements to the DA system. I would suggest adding a "directly" to this sentence, as in "Thus, we are not attempting to emulate or improve BGC models directly via ML, ...".
AC	Agreed. Added the word "directly": "Thus, we are not attempting to directly emulate or improve BGC models via ML, but instead use ML to improve DA,..."
RC06	L 103: "problem of propagating information from observed to unobserved variables in single-model deterministic runs": As opposed to probabilistic multi-model ensembles using different types of BGC models? I think the sudden emphasis on single-model is confusing -- also does not the model itself propagate information from observed to unobserved variables? I would suggest rephrasing the description of this problem in the context of DA.
AC	Added clarification: "Thus, we are not attempting to directly emulate or improve BGC models via ML, but instead use ML to improve DA, and specifically to cope with the challenging problem of propagating information from observed to unobserved variables in single-model deterministic runs that, by definition, do not have an ensemble to estimate the necessary statistics."
RC07	L 118: "BGM" -> "BGC"
AC	No more instances of BGM.
RC08	L 118: "end-to-end learning": Please introduce or define this term.
AC	Rewritten this paragraph to give a clearer description of this: "In a second configuration, instead of merging DA and ML, DA is used to generate a training dataset, from which ML learns the full DA step in an "end-to-end" emulation (Barth et al., 2020; Fablet et al., 2021). In this context, "end-to-end" refers to a ML model trained to map inputs (such as forecast states and observational information) directly to analysis increments, bypassing traditional, hand-crafted intermediate steps. As such, the ML model learns to predict the analysis increments for unobserved variables, given the surface background state and the increment for the observed variable. Recent work by

	Bocquet et al. (2024) has demonstrated that the entire EnKF analysis step can be learned in this data-driven, end-to-end manner.”
RC09	L 123: "the existing univariate scheme": add "DA" to make it clear to the reader what scheme this statement is referring to. Here, too, would be a good place to mention that the scheme is 1D only.
AC	Changed according to suggestion: “(a) Can we make improvements to the existing univariate DA scheme by updating a limited set of additional variables in 1D water column model, with an ML model to estimate correlations or analysis increments?”
RC10	L 220: "Offline refers here to a setup in which the ML-OI analysis is not then used as the initial condition for the next forecast": This is the first mention of "ML-OI" and it is not explained or defined. Either introduce the term or use "DA" instead of "ML-OI" in this sentence.
AC	Agreed. Changed to “DA”.
RC11	L 251: It should be plural: "state vectors".
AC	Amended.
RC12	L 254: "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.": Mention here somewhere that the mixed layer consists of more than one GOTM grid cell; otherwise, it appears like a handwavy argument that 0D is really 1D because it represents a column of water.
AC	We have updated the text to more accurately reflect the structure of the model: “While a 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. The mixed layer consists of a number of vertical GOTM grid cells, which can vary throughout the year based on the physical drivers determining mixed layer depth, such as temperature and salinity. This approach assumes that surface conditions are representative of the mixed layer as a whole (which is broadly true in this model configuration, where vertical

	<p>mixing is strong enough to ensure that surface and mixed layer conditions remain well-coupled). As a result, the increments derived at the surface also provide a reasonable correction at depths/other grid cells within the mixed layer.”</p>
RC13	<p>L 267: "We assimilate only total chlorophyll, y, at the surface.": Why introduce the symbol y here? It is not used in the following sentences and makes it seem like y is used to from hereon to denote "chlorophyll" when it is really meant to represent a chlorophyll observation. I would suggest introducing y when it is first used, which may be in Eq. 6.</p>
AC	<p>It is first defined in Eqs. 1 & 2, where we introduce the standard EnKF equations. But we can make the important clarification here that y is observations of total chlorophyll in our context:</p> <p>“We assimilate only observations of total chlorophyll, y,”</p>
RC14	<p>L 244: "x^b is the background state (which in this work is equivalent to a seven-day forecast state)": The phrase "seven-day forecast state" suggests a vector containing entries for 7 days of model states. Is that the case?</p>
AC	<p>The model state is at an instant in time, so rephrased to make this clearer:</p> <p>“(which in this work is the state of the system after a seven-day forecast period)”</p>
RC15	<p>L 247: What does it mean that P^b is "appropriately flow-dependent"? The authors just mentioned that P^b is of special interest in this study, but then this sentence does not provide a good explanation why it is of special interest. I would suggest including more information.</p>
AC	<p>Added some more information to make this clearer:</p> <p>“the matrix \mathbf{P}^b is of special interest to this work. Ideally this matrix should be appropriately flow-dependent, but in practice it is often not, such as in many operational schemes.</p> <p>For instance, in marine BGC, we would expect the onset of a phytoplankton bloom to trigger changes in nutrient concentrations, which in turn affect the correlations between the two. This response cannot be captured by climatological statistics alone where a fixed climatological estimate is used.</p>

	The purpose of this work is to introduce such flow-dependency to $\mathbf{P}^{\mathbf{b}}$, or to the analysis increments $\Delta \mathbf{x}$, with ML techniques.”
RC16	L 248: Here would be a good place to mention that the DA schemes described later use different ways to update other variables, but that all schemes use the same technique to update PFT chlorophyll which is described in the following.
AC	Added this clarification: “While each of the DA schemes described later will use different methods to update unobserved variables, all schemes will update total chlorophyll, the observed variable, and it's associated PFTs in the same way. (using the RUS scheme described in Sect.2.4.1).”
RC17	Eq 4. and 5.: The mathematical notation here is not ideal as χ and ζ are each meant to represent 4 different symbols, one for each PFT. It is not clear, for example, that the $x^{\mathbf{b}}_{\chi} / x^{\mathbf{b}}_{\zeta}$ ratio is based on the same PFT.
AC	We have added subscripts to clarify the notation – making it much clearer.
RC18	L 253: What does the "potentially" imply in this statement?
AC	Assuming this is referring to the use of the word potentially on L325: In the nitrate only experiments, N = 1. In the extended framework, N = 30. The use of the term potentially is confusing here, so we have removed it.
RC19	L 323: "long training EnKF run": Does this mean it was trained or run for a long time? It is unclear to the reader why "training" is used here.
AC	Rephrased to make this clearer: “Climatological variances, which are used to estimate $P_{\{\text{chl}\},\{\text{chl}\}}^{\mathbf{b}}$, are calculated from the same period of the EnKF run (Sect. 2.4.2) used to train the ML models.”

RC20	L 368: "ANN" has not yet been defined.
AC	We now define ANN (Artificial Neural Network).
RC21	L 400: "while making a bold assumption": This phrasing is not very scientific and suggests that the assumption is likely false.
AC	Agreed. Removed "bold".
RC22	L 405: This text is difficult to follow and I do not understand the jump to variances here. The previous sentences described the way correlations are computed and normalized and now there is a jump to variances and a statement about variances and how they can be estimated more reliably than correlations. How does this help when estimating the correlations?
AC	Added transition sentence to make less of a jump: "Apart from correlation, the other aspect of covariance to determine are the variances. Since each variance is a single-variable statistic that can be estimated more reliably than correlations..."
RC23	L 548: Is the light limitation due to the deep mixing or low seasonal light level? If it is the seasonal light level, mention this point first.
AC	Reworded to mention this point first: "The light-limited regime approximately spans the period from October to the start of the next spring bloom. Here, there is little-to-no phytoplankton growth due to the reduced light-levels at this time of year, meaning nutrients can be mixed throughout the water column without being used by the phytoplankton. During this period, phytoplankton concentrations are very low and mostly decoupled from nutrient dynamics.
RC24	L 615: "Though, it is clear that the correlations of the ensemble still vary more strongly than either other method." It is not the methods that vary but their correlation estimates: "Though, it is clear that the correlations of the ensemble still vary more strongly than the climatological and ML-based correlation estimates."
AC	Changed as above.

RC25	L 619: Contrary to the statement, based on Figure 3, it looks like both the climatological and ML estimates can capture the "moderately negative correlation" state. What am I getting wrong?
AC	In Fig. 3, we see that both methods do not appear to capture the weakly positive correlations, and both methods under-estimate the moderately negative concentrations. Modified the text to reflect this: "Finally, we also see that during the light-limited regime, the system can exist in either a ``weakly positive or no correlation" state, or a ``moderately negative correlation" state. However, both the climatological and ML estimates fail to capture these possible states (both predicting a weakly negative correlation over the period)."
RC26	L 622: "... 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.": This last sentence is confusing: the "This means" links the lack of great improvement to the small DA updates in the previous sentence. However, the second part of the sentence beginning with "as there is weak relationship" links the lack of great improvement to the weak relationship between chlorophyll and nitrate increments. What is the main reason? Please rephrase.
AC	Both contribute, so we have modified to reflect this. "Furthermore, as the concentrations of total chlorophyll are very near zero at this time of year (both in the ensemble and observations) and there is little spread in the ensemble - the resulting DA updates are 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 and the updates are small."
RC27	L 730: "It also slightly degrades the ammonium and silicate concentrations": What is meant by degrading concentrations?
AC	Changed to be more precise:

	“It also slightly worsens the forecast error for ammonium and silicate concentrations that are not updated during the analysis.”
RC28	L 941: I can see why one would choose synthetic observations in some initial DA experiments but I would not consider it a "necessary" first step. In fact, as I have mentioned in my last comments, synthetic observations may be too idealised to provide guidance for DA experiments with real observations. I would suggest removing "necessary" here.
AC	Agreed. Done.