# HYBRID MACHINE LEARNING DATA ASSIMILATION FOR MARINE BIOGEOCHEMISTRY

I. HIGGS, R. BANNISTER, J. SKÁKALA, A. CARRASSI, S. CIAVATTA

In this paper, the authors present multivariate data assimilation methods based on the introduction of machine learning (ML) approaches with application to 1D ocean biogeochemical models. The aim is to avoid the use of an ensemble of simulations in the forecast step of ensemble Kalman filters (EnKF) that can have a large computational cost while applying time-dependent multivariate analysis scheme when updating the unobserved variables form the observations. They suggest to first update sequentially the chemical components of the surface phytoplankton functional types (PFT) from surface observations of chlorophyll concentrations based on the BLUE while maintaining constant the background ratio between the components and the chlorophyll. Then, a fully connected neural network is used to update the surface unobserved variables either to emulate their forecast error correlation with the surface chlorophyll in the BLUE (ML-OI scheme) or to emulate directly the analysis increment of an EnKF (ML-EtE scheme). Numerical experiments are done with a 1D marine biogeochemical model (GOTM-FABM-ERSEM) at two locations in the English Channel with different biogeochemical dynamics. At this stage, only synthetic observations are assimilated in order to assess the performances of the different approaches.

The manuscript is globally well written and the topic is relevant. Building data assimilation systems for coupled ocean-biogeochemical models is challenging. However, this work has some weaknesses:

- I feel that many choices of the authors - both in the definition of the algorithms and the configuration of the numerical experiments - coud be better justified. Why building the update of the chemical component of the PFTs in that way? Why using the surface update for the variables in the mixed layer and not updating them in a similar way as done at the surface? Why not updating the variables under the mixed layer? Why comparing the ML schemes to an univariate scheme and not to an EnKF? See specific comments for more details. I think that this paper would benefit from additional justifications.
- To my point, the discussion on the results of the ML-OI and ML-EtE schemes are too optimistic, especially regarding the framework of updating all the unobserved variables (see specific comments). For instance, the 7-day forecast root mean square errors (RMSE) in the nutrients and some detritus are larger with the ML approaches compared to an univariate scheme updating only the PFTs. This may highlight assimilation biases due to inconsistent updates. From the results shown in this work, I am not fully convinced that they "have successfully shown that ML methods can make improvements to the DA schemes" as claimed by the authors. I think the discussion of the results should be qualified and better higlight the remaining issues of their approaches.
- The extension of these approaches to 3D models does not seem to be straigthforward. It is mainly due to generalization issues well known in ML and observed in this work - the NN trained with data from the L4 station perform poorly at the CWEC location - that could require to sample the English Channel with a large number of 1D vertical ML-based DA systems. Furthermore, the training of the NN is based on existing ensemble of simulations (forecast correlations) or

advanced multivariate DA systems (analysis increments). The data they produce are uncertain (large number of unknown parameters) and potentially not relevant notably in case of changes in the system (observation network, increases in the temperature that would result in the occurence of out-of-distribution data during the inference) leading to poor performances of the approaches. In that case, it may require to produce a new data set and retrain the NN which may be costly and time consuming.

To conclude, I think that some improvements are required before publication. I have listed some comments that the authors may consider.

**Specific comments.**

- p.2, fourth paragraph "we investigate the capability of ML to learn the hidden, non-linear, and complex relations between BGC variables and to use the learned functions within a DA scheme or fully substitute it": In this work, the data assimilation algorithms used are derived from an ensemble Kalman filter (EnKF), which means that a linear update is applied to the variables. Common ocean biogeochemical models are non-linear and subject to constraints such as the positivness of the variables that make the use of an EnKF challenging. So, one might wonder what are the benefits that can be expected to learn complex non-linear relations when using linear data assimilation methods. The authors may comment this point and hightlight the constraints associated with operational oceanography. Furthermore, the authors use ML algorithms to learn the forecast error correlation (ML-OI) or the analysis increment of an EnKF (ML-EtE). So it is not clear that the model did actualy learn the "hidden, non-linear, and complex relations between BGC variables". The authors should consider to rewritte the sentence to better highlight their contribution.
- 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.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 occurence of assimilation biases in the water column. In this study, only surface variables are updated from the observations based on a BLUE-like equation. The same increment is then used to update the corresponding pelagic variables in the mixed layer while no updates are applied below this layer. It means that the impact of the observations on the variables below the mixed layer is only due to the adjustement of the model to the update in the mixed layer. Because it can be weak and compensate by the relaxation towards nutrients, it may prevent the occurence of assimilation biases such as trends in the nutrient concentration in the deep layers. I think that the authors should add a comment on it to highlight the limits of this work for more realistic applications.
- p.5, first paragraph "The resulting variation in wind strength [...] prevents ensemble collaps induced by the previously mentionned nutrient relaxation or [..] are absent in a 1D set-up.": I presume that this strategy is effective in preventing an ensemble collapse because only the surface variables are updated. In this case, it might be sufficent to maintain the spread of the ensemble at the surface, independently of the deeper layers. However, it may not be effective in a setting where the variables are update throughout the water column during the analysis. I think that a comment should be added to highlight that point.
- p. 6 Eq.(4) "$\mathbf{H} = [1, 0_1, ..., 0_N]$ [...] with the first element being surface total chlorophyll": The total surface chlorophyll is not a state variable and rather "an

observation" that can be computed from the phytoplankton function types (PFT, state variables). This is quite unusual reagarding the mathematical framework of data assimilation. So, I think that the authors should better motivate this choice, notably in relation to the following comment.

- p.6, third paragraph "The PFTs are excluded from the state [...] where $\chi$ 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 $\frac{x^a_\chi}{x^a_c} = \frac{x^f_\chi}{x^f_c}$ and $\frac{x^a_\zeta}{x^a_\chi} = \frac{x^f_\zeta}{x^f_\chi}$. 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.
- 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. 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).
- p.7, Table 2 "Eq.(4) with (5)": Eq.(4) refers to the observation operator $\mathbf{H}$ and (5) to the PFT update. Does it mean that $\Delta x_i$ correspond to PFT only?
- 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?
- 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.
- 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".
- p.8, second paragraph "where $P^f_{i,c}$ is the background error [...] where $COR_{i,c}$ 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.

- p.8, fourth paragraph "is scaled according its climatological" → is scaled according to its climatological (?)
- p.8, fourth paragraph "The resulting surface increments of the unobserved variables are then propagated to the other levels in the mixed layers as described previously": I wonder why the authors did not apply the same ML strategy to learn the forecast error corrrelation $Corr_{i,c}$ between the surface chlorophyll and the unobserved variable in the mixed layer. Could the authors motivate their choice?
- p.8, last paragraph "For the first application of ML-OI in Sect 4.2, the target [...] between total chlorophyl and nitrate. In the later application in Sect. 4.3 onwards this is extended from just nitrate to a wider set of variables.": Could the authors motivate their choice to start with the correlation between total chlorophyll and nitrate only and not directly presenting the results of the more general case?
- p.9, second paragraph "In ML-EtE, we assume that the essential properties of each statistical object [..] that are combined in a single value": It could also be mentionned that directly predicting the analysis increment (a vector) allows to tackle problems in large dimensions that would not be possible when predicting a error covariance matrix required in the assimilation step.
- 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.
- 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?
- 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.
- p.9, seventh paragraph "Each ML approach is tested in the following scenarios [...] results of (2a)": The auhtors may add a comment to explain the motivation of this splitting.
- p.10, fourth paragraph "MGBC": Could the authors explain this acronym (first occurence)?
- 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?
- p.11, first paragraph "In this section, we explore the performance of ML-OI and ML-EtE in updating only nitrate as an unobserved variable [...] a limiting nutrient at the L4 location (see Fig.A2 in the Appendix) [...] has a clear, explainable relationship with total chlorophyll": Following previous comments, it is not clear to me what are the motivations of updating only the nitrate variable and what could be learned that would meaningful for the multivariate case updating all the nutrients. Fig.A2 highlights correlations between surface total chlorophyll and silicate (or phosphate) as strong as those between surface total chlorophyll and nitrate. So, one may anticipate similar magnitudes of update for the other nutrients. In that case, it is not clear what would be the evolution of the system updating all the nutrients after several assimilation cycles based on the evolution of the system updating only the nitrate variable. For instance, one notes significant differences in the forecast and analysis RMSE in the nitrate between the versions of the ML-EtE algorithm that update only the nitrate or all the nutrients (see Fig.7). For this case, updating all the nutrients damages the forecast and

analysis RMSE in nitrate while improves the RMSE in ammonium and silicate. This result is interesting and may highlight assimilation biases induced by the different update strategies. Thus, it is worth to show results for the strategy updating nitrate only. However, I think that a focus on this strategy would require to show additional results: for instance the RMSE in the other nutrients at the surface and in the mixed layer depth. Then, the introduction of this subsection may be motivated as alredy done (nitrate is a limiting nutrient) while questioning the relevance of this approach and the need for specific results.

- p.11, first paragraph "Offline referes here to a setup in which the ML-OI analysis is not then used as initial condition for the next cycle": For confirmation, the total chlorophyll and the PFTs are not updated as well, are they?
- 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.12, second paragraph "We also see that during the light-limited regime [...] at this time of the year": It could be highlighted that the correlations tend to be too weak for both methods compared to the reference. It can be due to the fact that both methods are not able to reproduce the interannual variablity observed for the correlation in the reference. On the contrary, both systems tend to reproduce the same weak negative correlations every year while larger negative or weak positive correlations (winter 2018) occur in the reference depending of the year. Even if the chlorophyll uptades are expected to be weak, it suggests that both systems can not reproduce the interactions between chlorophyll and nitrate during this period of the year.
- 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?
- 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.14, Fig.5: It may be recalled in the legend that the increment is computed at the surface.
- 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.15, second paragraph "We can also see that each approaches make little to no adjustment during the ligth-limited regime": The ML-OI algorithm leads to a large negative increment at the end of 2023 while the concentration of nitrate in the forecast is significantly weaker than the one in the truth. Could it be related to my previous comment on the ML-OI algorithm being not able to capture positive correlation between the nitrate and total chlorophyll in the ligth-limited period when they occur? It may be interesting to see the evolution of the increment in total chlorophyll as well.
- p.15 Fig.6: I presume that the RMSE is computed in the surface layer. Is it correct? This information may be missing.
- p.15., last paragraph "where errors are normalised against the error in the RUS scheme": I wonder why the authors do not normalise against the error of the EnKF, which is the algorithm with the best RMSE in nitrate (see Fig.4). Maybe the best would be to plot the error without normalisation and add the reult of the RUS and EnKF as benchmarks.
- p.17, Fig.7: I presume that the RMSE is computed in the surface layer. Is it correct? This information may be missing.
- p.17, first paragraph "Before discusing the extended schemes, we can see from Fig. 7 the dynamical impact of the updates of the ML-EtE (N03) have on other (i.e. non-updated) marine BGC variables [...] particularly microzooplankton.": It could be added that it also slightly degrades the ammonium and silicate concentrations that are not updated during the analysis.
- 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.
- 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.
- p.18, second paragraph "The ML-OI (ALL) scheme [...] and ML-OI (ALL) schemes": I do not agree with the statement that the ML-OI (ALL) is effective. The forecast RMSE are even larger than those of the ML-EtE (ALL) for most of all the unobserved variables. For several variables, the analysis is not able to reduce the RMSE below the value of the RUS model. To my point, it rather hightlights assimilation biases due to the inconsistency of the updates that accumulate over time. I think that the authors shoud comment on it.
- 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.
- 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.
- 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.

- p.19 first paragraph "(Sect. ??)"
- 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?
- p.21 "This is likely because the correlations predicted by the ML-OI scheme represent a more locationn-agnistic relationship in the marine BGC variables": this statement could be qualified by noting that the updates of the ammonium, bacteria and labile DOM lead to an increase in RMSE.
- 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.
- 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.
- 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.
- p.23, third paragraph "With machine learning (ML)", we achieve significant improvements over conventional approaches that rely on climatological statistics or omit updates altogether. Our analysis of ML-predicted ntirate updates illustrates [...] for improving BGC DA.": The ML approaches result in an improvement compared to the use of climatological statistics in the nitrate-only update framework. However, the impacts of using climatological statistics is not assessed in the general case. Because the ML scheme can degrade the forecast RMSE in unobserved variables compared to the RUS or the nitrate-only ML schemes, we may wonder if they can be useful in this more complex setting. The sentence may be qualified in that sense.