# Comments of reviewer 1 and responses

## Major comments

**comment:** This manuscript presents a rigorous and timely assessment of how glider-based carbonate-system observations can improve coastal pH estimates through 4D-Var assimilation in ROMS–NEMUCSC. The integration of pH and alkalinity with physical and chlorophyll data, combined with a thorough evaluation of ESPER-based hybrid estimates, makes this contribution relevant for coastal carbon monitoring and DA system design.

Overall, the study is technically strong, clearly motivated, and generally well executed. The comparison between full biogeochemical DA and hybrid statistical–dynamical methods is valuable and will interest both modeling and observational communities. The manuscript is publishable after major revisions aimed at sharpening key messages and clarifying methodological choices.

We thank the reviewer for the constructive comments; please see our responses below.

**(M1.1) comment:** The manuscript is rich in experiments, but the core scientific conclusions could be distilled more explicitly. The three main findings (limited impact of physical DA on pH, strong improvement from pH+alkalinity DA, and competitive performance of hybrid ESPER approaches) should be highlighted earlier and revisited more succinctly in the Discussion.

We agree with the reviewer and in response to this comment, we modified the manuscript in several places. The three key messages identified by the reviewer are now mentioned more clearly in the abstract:

**revised text (Abstract):** In our experiments, the assimilation of physical variables and chlorophyll alone has limited impact on pH and other carbonate system estimates, while the joint assimilation including pH and alkalinity variables successfully improves these estimates. Cross-validation experiments further demonstrate that the joint assimilation typically also improves estimates near the observation network, although downstream advection of increments can occasionally degrade results. We also show that hybrid estimates that combine the output of the dynamical, physical ocean model with a statistical model produce accurate carbonate system estimates without requiring a biogeochemical model.

We further modified the first paragraph of the discussion to emphasize the first key message identified by the reviewer, which was given less weight in the previous version of the manuscript:

**revised text (Section 4, par. 1):** As a baseline we used the reference DA experiment which only assimilates observations of temperature, salinity, sea level anomaly, and chlorophyll. This reference DA experiment significantly improves the fit to the assimilated variables but has a very limited impact on model estimates of pH and oxygen. In contrast, the joint assimilation including pH, alkalinity and oxygen observations successfully improves estimates of these variables while largely maintaining the quality of physical and chlorophyll estimates. This effect is likely due to a down-weighting of the reference observations – consisting of satellite-, float-, and glider-based temperature, salinity, sea level anomaly, and chlorophyll – in the 4D-Var state estimation procedure simply due to the presence of more observations.

**(M1.2) comment:** The necessity to assimilate estimated, not measured, alkalinity (Section 2.6) is a central limitation. The discussion acknowledges this but remains somewhat cautious. The authors should explicitly quantify the sensitivity of the pH increments to TA uncertainty and clarify in which coastal regimes the ESPER TA is reliable, and where it may fail (river plumes, OM-rich waters, denitrification).

We agree with both reviewers that the need for estimated alkalinity data is a limitation of our DA setup. In response to this comment and comment M2.1, we now mention this limitation in the abstract:

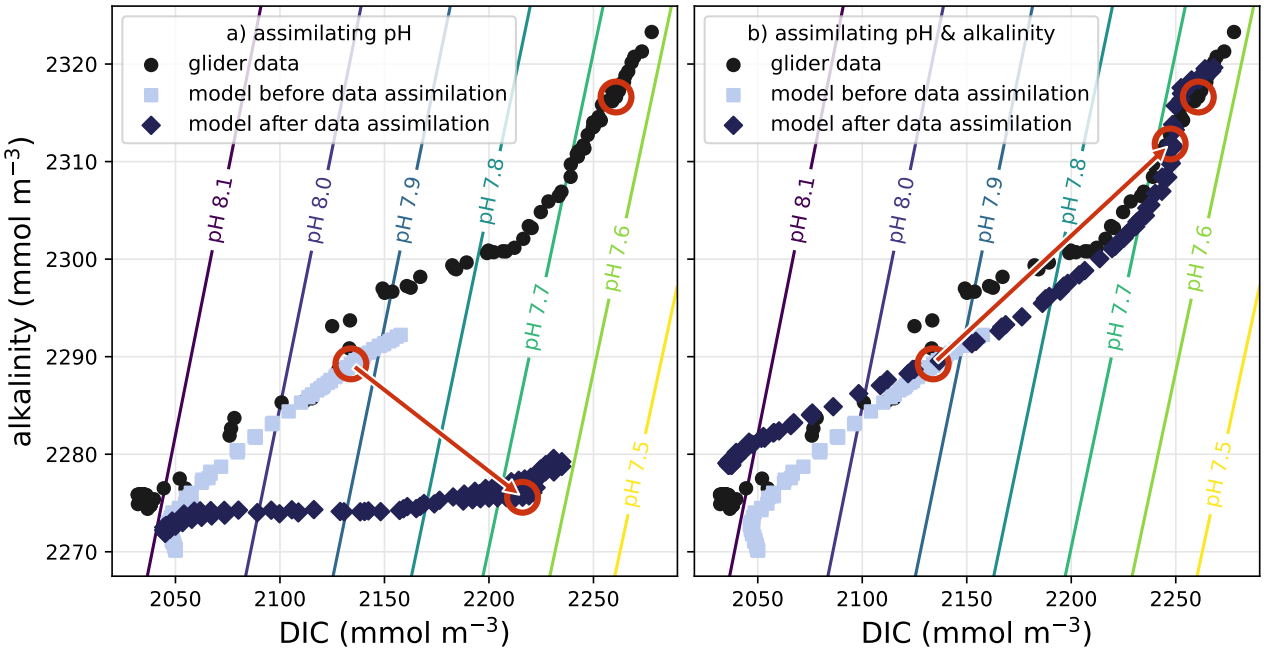


Figure R1: Replicate of Fig. 5 in Fennel et al. (2023) using alkalinity instead of DIC observations. (a) When only pH data is assimilated, the model estimates are moved closer to the observed pH values by increments in alkalinity-DIC space that degrade the model's alkalinity estimates. (b) The model state estimates improve considerably by assimilating data for alkalinity (or DIC; not shown) together with the pH observations.

---

**revised text (Abstract):** Our carbonate system data assimilation setup relies on the combined assimilation of pH and alkalinity data to obtain reliable state estimates. Because alkalinity is not yet routinely measured by gliders, we utilize statistically estimated alkalinity values and examine the limitations of this approach in our study.

---

Furthermore, we moved the paragraph in the Discussion section assessing the limitation upward so that it appears earlier, following the discussion of our main results – the key points emphasized by the reviewer such as river plumes, organic matter-rich waters and denitrification are mentioned there.

As pointed out in Section 2.6, the joint assimilation of two of the three variables pH, alkalinity, and DIC, is required to move the state estimate to the correct point in alkalinity-DIC space. When assimilating pH alone, state estimates are moved to points in alkalinity-DIC space with the correct pH value but not necessarily the right alkalinity or DIC value. When assimilating pH together with alkalinity, the sensitivity to the alkalinity value that is being assimilated depends on its location in alkalinity-DIC space (because $\partial$pH/$\partial$alkalinity is not constant) as well as the prescribed observation errors (see Section 2.5). The behavior of the DA system is nicely illustrated by Fig. 5 in Fennel et al. (2023) which is based on an earlier version of the DA system presented here. For the reviewer, we have recreated the figure here. In response to this comment, we have added a reference to Fig. 5 in Fennel et al. (2023) to Section 2.6.

---

**(M1.3) comment:** Some cross-validation experiments show deterioration of pH downstream of the lines, attributed to advection of increments. This is important for future glider network design. A brief dynamical explanation (e.g., density structure, mesoscale features along Line 67) would strengthen the argument.

---

We believe that Fig. 6 is quite illustrative of the downstream effect of the DA. While the DA increments are localized close to the glider lines (Fig. 6a), the resulting change in pH is advected downstream and after one year still remains filamentous in structure (Fig. 6b). While there is an overall increase in pH at 100 m depth in

the south of the domain, there also appear pockets of decreased pH that are being advected through the region. These pockets offer an explanation why pH estimates may deteriorate downstream. We hypothesize that after more than one year of carbonate system DA and continuous adjustments of the model state, the difference in pH will become less filamentous. We are in the process of performing new carbonate system DA experiments for the years beyond 2019, and in future work we will examine this question more thoroughly.

In response to this comment, we modified Section 3.5, added a reference to Fig. 6 and a reference to the structure of the pH change brought about by the DA:

**revised text (Section 3.5, par. 2):** This suggests that in some cases, the advection of increments may cause model-data discrepancies downstream, likely due to the filamentous structure of the change in pH after less than a year of carbonate system DA (Fig. 6b).

**(M1.4) comment:** The result that hybrid ESPER estimates outperform the full BGC model (when carbonate variables are not assimilated) is striking. The implications deserve more emphasis: under which conditions does a hybrid approach suffice operationally? Is the benefit solely from improved T–S via physical DA, or also from limitations in the NEMUCSC carbon module?

As we illustrate in the manuscript, the hybrid approach can be used operationally to obtain gapless estimates of the carbonate system variables and oxygen from a physical ocean circulation model. The answer to the question, under which condition these estimates approach suffice, in the sense that they reach a certain level of quality or perform better than a biogeochemical model, depends on the models being used and under which conditions. As pointed out in the reviewer's second comment (M1.2), under certain conditions the statistical model estimates of the carbonate system variables may degrade, and this is of course the case for the hybrid estimates as well. And while the NEMUCSC model and carbon module have undergone many improvements over the last years, we have not performed a rigorous parameter optimization experiment with the full set of carbonate system observations used in this study. Hence, it is likely that the biogeochemical estimates can be improved further – and other biogeochemical models may yield even better pH estimates. Under which circumstances a biogeochemical model or a hybrid approach will yield better estimates thus depends on many factors.

In response to this comment, we have modified the discussion with more emphasis on the point that the hybrid estimates, based solely on model temperature and salinity, often outperformed the biogeochemical model in our experiments.

**revised text (Section 4, par. 3):** Third, we evaluated whether combining output from the dynamical ROMS model with the statistical ESPER model can produce improved carbonate system state estimates without the need for a biogeochemical model. Our results show that these hybrid estimates based on ESPER using only ROMS temperature and salinity as inputs often produce better results for pH, alkalinity, DIC, and oxygen than the biogeochemical model without assimilating carbonate system variables or oxygen. Importantly, the hybrid estimates benefit from physical DA through improved temperature and salinity estimates and do not require a biogeochemical model. This result suggests that existing physical ocean models and physical DA systems can be used to obtain good carbonate system estimates without implementing complex biogeochemical models if the statistical model estimates are reliable.

**(M1.5) comment:** The study shows an expected improvement when O2 is assimilated, but the weak coupling between pH and O2 increments reflects structural constraints of the DA system. It would be beneficial to comment on whether variable-covariance specification (currently set to zero) is a limiting assumption for future biogeochemical DA.

First off, we need to point out that cross-variable covariances are always set to zero in present ROMS-based DA systems, and there is no option in ROMS to specify non-zero cross-variable covariance values. Therefore, not using cross-variable covariance terms was not a deliberate choice in our DA setup, and in response to this

comment, we have added an explanatory statement to more clearly state this fact to the reader (see revised text below).

We already mentioned the lack of cross-variable covariances in the Discussion section as a major limitation of our DA system. However, both reviewers brought up this topic in their major comments (see comment M2.4) and in response, we have added more information and mention that other commonly used DA systems, such as the broad range of ensemble-based techniques may provide better state estimates:

---

**revised text (Section 4, par. 2):** This limited impact across variables likely stems from two key factors in our DA implementation: (1) the ROMS 4d-Var DA system presently allows for univariate covariances only (i.e., with zero covariance between different variables and with no ability to specify non-zero cross-variable covariance values), and (2) our configuration, in particular the background error specification (see Section 2.5), favors increments to directly observed variables over unobserved ones. As a result, our current DA setup allows larger increments to alkalinity and oxygen directly but not to unobserved variables that have pathways with feedback to these variables, such as nitrate, a main driver of primary production that modifies both oxygen and pH. This inherent limitation of our DA system could mean that alternative techniques, such as ensemble-based DA, which automatically include cross-variable covariances, could create larger improvements in unobserved variables from the assimilation of pH data.

---

**(M1.6) comment:** The manuscript relies exclusively on ESPER for alkalinity and DIC estimation, but does not justify this choice. This is important because CANYON-B/CONTENT is widely used in the community, specifically trained for glider-type variables, and often performs better in coastal and upwelling systems due to its inclusion of oxygen and sometimes nitrate as predictors. The authors should briefly explain why ESPER was selected, and whether alternative empirical regressions (e.g., CANYON-B, LIAR, multi-sensor neural networks) were evaluated. A short comparison or rationale would strengthen confidence in the robustness of the hybrid approach. At minimum, please clarify: what variables ESPER requires in this implementation, whether CANYON-B was unsuitable due to predictor availability or training domain, whether differences between algorithms could alter the conclusions on hybrid performance.

---

We agree with the reviewer that additional justification is warranted. For the glider-based ESPER pseudo-observations, there was a slight error in the manuscript about how these were derived. For glider data, we can use T, S, O2, P, location, and date as inputs, thus allowing us to use ESPER-LIR, ESPER-NN, and CANYON-B. We assessed the performance of these three algorithms by comparing discrete measurements to algorithm-derived measurements from three cruises (West Coast Ocean Acidification Cruise 2021, and two cruises conducted by Takeshita's lab at MBARI) that span a large range of the model domain (see Fig. R2). These cruises were not included in the training datasets for the algorithms, thus, should be an independent assessment for the algorithms. In general, all three algorithms produced reliable estimates of TA, DIC, and pH across the whole model domain, although subtle differences exist between the algorithms. Given that there was no clear 'best' algorithm in this domain, we chose to average the output of the three algorithms. We corrected the error in the manuscript, and we expanded Section 2.5 with a description of TA, DIC and pH data using the three algorithms, including a validation experiment using the cruise data described above and new Fig. 3.

The reason why the ESPER algorithm was used for the hybrid approach is that it accepts inputs of T, S, P, location, and date, whereas CANYON-B requires oxygen as an additional input. Since the hybrid approach uses a physical model as inputs, we required an algorithm that only requires T and S as inputs. Therefore, CANYON-B could not be used. ESPER-LIR is an updated version of the LIR algorithms (Carter et al. 2018), and in fact, Carter et al. 2021 refers to ESPER-LIR as LIRv3. Therefore, we opted to use ESPER over LIR, since ESPER is trained on more recent data, it handles anthropogenic carbon better from first principles, and it has a DIC output. This justification has been added to the main manuscript.

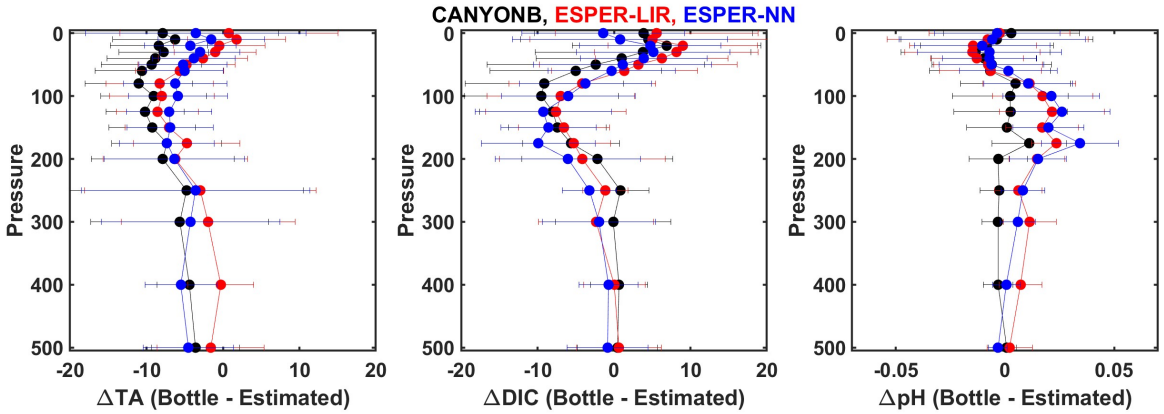


Figure R2: Profiles of measured minus estimated values for TA (left), DIC (middle), and pH (right) using CANYON-B (black), ESPER-LIR (red), and ESPER-NN (blue). The markers represent the median, and error bars represent 1 standard deviation. This figure has been added to the manuscript as Fig. 3.

## Specific comments

**(S1.1) comment:** Figures 4 and 5 are informative but visually dense; consider simplifying color scales or moving supplementary diagnostics to the Appendix.

Fig. 5 is does indeed contain a lot of information, but it is currently the only figure showing actual glider observations and model estimates. We believe it is valuable to the reader to see these estimates not just in the form of an error bar or other summary statistic. This figure uses 3 color maps to visualize 3 distinct properties: pH (observed or estimated), the difference between estimated and observed pH, and the improvement in pH estimates. We think that each has a place and helps the reader quickly identify where model-observation differences are largest and where the DA leads to the largest improvements.

Fig. 6 is a visualization of the direct impact of the DA (panel a) and its downstream effect (panel b). We use it in this document to respond to the reviewer's comment about the downstream effect of the DA (comment M1.3) and in response to that comment, we have added another reference to Fig. 6. We therefore believe it is useful to the reader and important to keep this figure in the manuscript.

**(S1.2) comment:** State the glider pH sensor accuracy explicitly when first introduced (currently only in Table 3).

We made this change and now mention the value in Section 2.5.

**(S1.3) comment:** Clarify whether ESPER was re-trained or used as published.

For generating the hybrid estimates, we use ESPER as published; this is an important point we forgot to mention in the manuscript. In response to this comment, we have added the missing information, including the version of `ESPER_LIR` in the official repository (to which we include a reference):

**revised text (Section 2.4, par. 7):** We created hybrid estimates using the published version of ESPER (ESPER_LIR V1.01 from the ESPER GitHub repository) without any additional training and based on model estimates from [...]

As clarified in response to comment M1.6, the estimated data used in the DA experiments is based on an average of three statistical model estimates, each one was used as published; see our response above.

**(S1.4) comment:** The manuscript is long; some methodological descriptions (e.g., NEMUCSC structure) could be tightened.

In preprint mode, Section 2.2 is less than a page long. In response to this comment, we shortened the section a bit, but we see no opportunity for a large reduction in text or moving a standalone section to a Supporting Information document.

# Comments of reviewer 2 and responses

## Major comments

**comment:** The manuscript is an excellent and very well-written paper. It provides a clear methodological advance for improving pH estimates and, more broadly, carbonate-system observation in oceanography by combining data assimilation with existing observing networks and hybrid approaches. The results are convincing but a few points would benefit from clarification.

We thank the reviewer for the helpful comments; please see our responses below.

**(M2.1) comment:** There is a strong dependence on estimated variables. The pH assimilation relies on a joint assimilation of pH and TA, but TA is not directly observed here: it is taken from statistical estimates (ESPER) because autonomous platforms do not routinely measure alkalinity. If TA is biased (e.g., mCDR influence, river plumes/DOM,...), the whole approach can quickly become fragile, or even unusable. This is an important limitation, because it reduces the method's operational value in complex coastal settings and outside regimes where TA is well captured by broad, global relationships. This point should be mentionned in the manuscript

Indeed, that is a point that is mentioned in the discussion but that we did not give enough emphasis. Both reviewers commented on this limitation of the state estimation setup, and we made several changes to the manuscript in response. Please see our response to comment M1.2.

**(M2.2) comment:** There is a risk of statistical "circularity". ESPER is used both to generate the pH/TA pseudo-observations and to reconstruct carbonate variables from model/DA outputs (Section 2.4, 3.3-3.6 and Figure 4). The same statistical machinery appears on both sides of the evaluation. The authors acknowledge that the good performance of the hybrid product may partly reflect this methodological proximity, rather than a truly independent skill. It would help to include a more "decoupled" validation, either with alternative algorithms (not ESPER) or, ideally, with independent carbonate-system observations.

We agree that this point required further clarification. The algorithms used for the hybrid approach are not exactly the same as the ones used for the pH/TA pseudo-observations. The hybrid approach uses ESPER-LIR with temperature and salinity inputs, whereas the pH/TA pseudo observations from the gliders are an average of three algorithms (ESPER-LIR, ESPER-NN, and CANYON-B) with temperature, salinity, and oxygen as inputs (see our response to comment M2.1). In response to comment M2.1 and this comment, we added a validation with independent carbonate-system observations to the manuscript, as suggested by the reviewer. The validation is included in Section 2.5 and includes a new figure. It shows that the pseudo-observations match the true observations well and that the addition of oxygen significantly improves the algorithm performance since it provides a constraint on biological respiration for subsurface waters.

Therefore, while ESPER is used both in the generation of the pseudo-observations and the hybrid estimates, we have now validated the pseudo-observations and provide evidence for "independent skill" based on ship-based observations. Furthermore, the inputs to ESPER are different, and we average additional algorithm outputs to generate the pseudo-observations – points that we now include in the manuscript.

**(M2.3) comment:** The sensitivity to observation/background error choices should be tested. In the section 2.5 the errors settings are tuned via FPI for T/S/SSH/Chla, but not in the same way for carbonate variables, where the system is more non-linear and where several choices remain assumption-driven (e.g., $\pm 3\sigma$-type reasoning, fixed parameters). What is missing is a real sensitivity analysis: how stable are the results if $\sigma$(pH), $\sigma$(TA), correlation length scales, or relative weights are changed within reasonable bounds? Do the main conclusions survive if carbonate uncertainties are slightly under- or over-estimated?

Setting up a fixed point iteration (FPI) requires output from an existing DA simulation and thus requires

completed 4d-Var experiments such as the ones presented in the manuscript. Furthermore, the FPI is not straightforward to implement for variables that are not directly observed, for example, pH in our application. Model pH is a diagnostic variable, mainly based on model alkalinity and DIC, so modifying $\sigma$(pH), the pH background error values, will have no effect on the estimates of the DA system. Instead, the FPI will need to adjust the background error values for DIC and alkalinity, together with the observation error values for pH and alkalinity. In response to this comment, we conducted some initial experiments, which suggest that an FPI-guided adjustment leads to a small rebalancing of the uncertainties, likely increasing the misfit for oxygen while improving the fit for pH modestly. Based on these calculations, we have no indication that any of the results we present in the manuscript would change qualitatively or strongly modify our conclusions. These results are not yet ready for publication and fall beyond the scope of the manuscript.

In terms of correlation length scales, we performed some experiments modifying the horizontal and vertical length scales for a previous version of our DA setup. We tested modifying the length scales of all variables, only the biogeochemical variables and only the carbonate system variables, and performed a grid search with length scales ranging from 10 km to 75 km. In these experiments, we found no configuration yielding a strong improvement in model estimates, and we thus stayed with the length scales employed in various previous studies.

We believe it is important to provide the reader with more detail about the FPI and in response to the reviewer's comment, we added more context about why we consider an FPI for a future study:

---

**revised text (Section 4, par. 7):** We further aim to address current methodological limitations through modification of background and observation error values in the 4D-Var system using the methodology presented in Mattern et al. (2018). For this purpose, we require statistics computed from completed 4d-Var DA experiments, and the experiments presented in this study could form the basis for improved background and observation error specifications, especially for the carbonate system variables and oxygen in our DA system. Such an approach requires additional research because model pH is a diagnostic variable, and pH observation error values will need to be balanced with alkalinity and DIC background error values.

---

**(M2.4) comment:** In the section 3.6 and discussion section, the ROMS 4D-Var assumes zero cross-variable covariances. As a result, assimilating pH mainly adjusts DIC/TA, but it does not propagate much information to other biogeochemical controls (e.g., nutrients), and therefore cannot strongly improve oxygen indirectly. The manuscript flags this as a key limitation and points toward ensemble-based approaches or multivariate formulations where cross-covariances are allowed. In its current form, the benefit remains largely "diagnostic" for the carbonate system, rather than truly improving coupled biogeochemical dynamics.

This limitation of our DA system was also picked up on by reviewer 1; please see comment M1.5 and our response there. As pointed out in our response and the discussion section in the manuscript, we believe that the lack of cross-variable covariance values is a significant limitation of our DA system. With reliable cross-variable covariance values, the pH and alkalinity DA could perhaps significantly improve model oxygen estimates as well. However, we also believe that improving model DIC and alkalinity estimates through DA represents an important improvement in biogeochemical dynamics, even when oxygen estimates are not improved simultaneously.

**(M2.5) comment:** There is a limited spatial/temporal generalization. The analysis is restricted to 2019 and a region that is heavily constrained by the observing network. This is fine for a first demonstration, but it limits how far the conclusions can be generalized. Extending the assessment to other years (including extreme or anomalous conditions) and bringing in additional datasets (e.g., BGC-Argo, high-frequency $pCO_2$ products) would be important to demonstrate interannual robustness and transferability to other coastal margins.

We agree with the reviewer; as mentioned in the discussion section, we are in the process of preparing an extension of the DA runs beyond 2019 and utilizing more data for assimilation and verification. We learned many valuable lessons from the 2019 DA experiments, identified issues in the DA setup, and found solutions to

some of them as examined in the manuscript. These lessons, issues, and solutions can be obtained from one-year experiments without the need for longer simulations, and we believe our results are of interest to readers and the ocean modeling community. Future work will build on the 2019 experiments and may answer questions we cannot yet address with our results presented here.

**(M2.6) comment:** In highly dynamic waters (Line 67), DA "pH" does not always beat the reference unless pH is directly assimilated. For the "measured" pH dataset (glider line), DA experiment 1 (CUGN pH/TA estimates) is sometimes slightly worse than the free/reference run. The clear improvements in the upper 0-150 m come mainly from the hybrid approach or from direct assimilation of the pH sensor (DA experiment 2). In the manuscript's conclusion, denser spatio-temporal observations are needed, seems justified. It means that "pH via ESPER pseudo-observations" is not a universal solution in regions dominated by strong meso/submesoscale variability. This should be mentionned in the discussion and conclusion

We agree with the reviewer's comment in pointing out the difficulties of joint physical-biogeochemical DA in highly dynamic systems. In response to this comment and comment M1.2 by reviewer 1, we now emphasize the limitations of the assimilation of estimated alkalinity in the abstract and discussion sections. We further added a mention of this limitation to the conclusions:

**revised text (Conclusions):** For carbonate system DA, our approach relies on assimilating observed pH jointly with statistically estimated alkalinity data which succeeds in our application but may lead to issues in scenarios where alkalinity estimates are unreliable.

**(M2.7) comment:** To summarize, these points should be considered:

- Add a sensitivity analysis ($\sigma$(pH), $\sigma$(TA), correlation scales, localization choices, 4D-Var window length) to show that the qualitative conclusions are robust.
- Strengthen independent validation: use a different algorithm than ESPER and/or independent carbonate data (BGC-Argo, discrete samples, $pCO_2$ products), especially outside the CUGN footprint.
- Discuss concrete options for allowing cross-covariances (EnKF/EnVar, multivariate B), since this is identified as a structural limitation.
- Better frame situations where TA is likely perturbed (mCDR, terrigenous inputs): detection criteria, potential switch to a more regional/statistically adapted framework, or assimilation of alternative variables.

We addressed these comments in our responses above.

## References

*ESPER GitHub repository* (n.d.). URL: https://github.com/BRCScienceProducts/ESPER.

Fennel, Katja et al. (2023). "Modelling considerations for research on ocean alkalinity enhancement (OAE)". In: *State of the Planet Discussions* June, pp. 1–47. DOI: 10.5194/sp-2-oae2023-9-2023.

Mattern, Jann Paul, Christopher A. Edwards, and Andrew M. Moore (2018). "Improving Variational Data Assimilation through Background and Observation Error Adjustments". In: *Monthly Weather Review* 146.2, pp. 485–501. DOI: 10.1175/MWR-D-17-0263.1.