Anonymous Review 1

Review of Subhas et al. "A tracer study for the development of in-water monitoring, reporting, and verification (MRV) of ship-based ocean alkalinity enhancement"

This is a thoughtful and well-conducted study which attempts to take the challenge of quantifying the effects of marine carbon dioxide removal (mCDR) from the realm of theory, where it has mostly been confined, out into the real ocean. The authors convincingly use a tracer approach to simulate the addition of alkalinity to Atlantic waters and monitor the dispersal of the tracer over multiple days.

We thank the Reviewer for their constructive thoughts and comments on our manuscript. We provide detailed responses below.

MAJOR COMMENTS

-The paper proposes an MRV framework, but it's a little unclear what the framework is, exactly. Is it Equations 1a-1d? Is it the conceptual diagrams in Figure 1? Is it all the text in Lines 90-125? Is it all of these combined? I may have missed it, but after reading the manuscript I would struggle to describe the framework to someone else in a sentence or two. It would be good to clarify the basic points of the framework, maybe in some bullet points or a callout box.

Thank you for this point, we will clarify this in the appropriate section of the manuscript, and may shift our language to discussing it as our "OAE analytical framework" rather than as an actual MRV approach. We appreciate the suggestion of a callout box, and will consider this, or a succinct set of bullets in the framework section, to make it clear what we see as the important steps. To us, this focuses on the following statement made in the manuscript:

> "We explicitly distinguish between OAE-driven signals and CDR-driven signals."

The analytical framework then follows as three main steps. The first two are solely related to OAE, while the last one quantifies the resulting CDR:

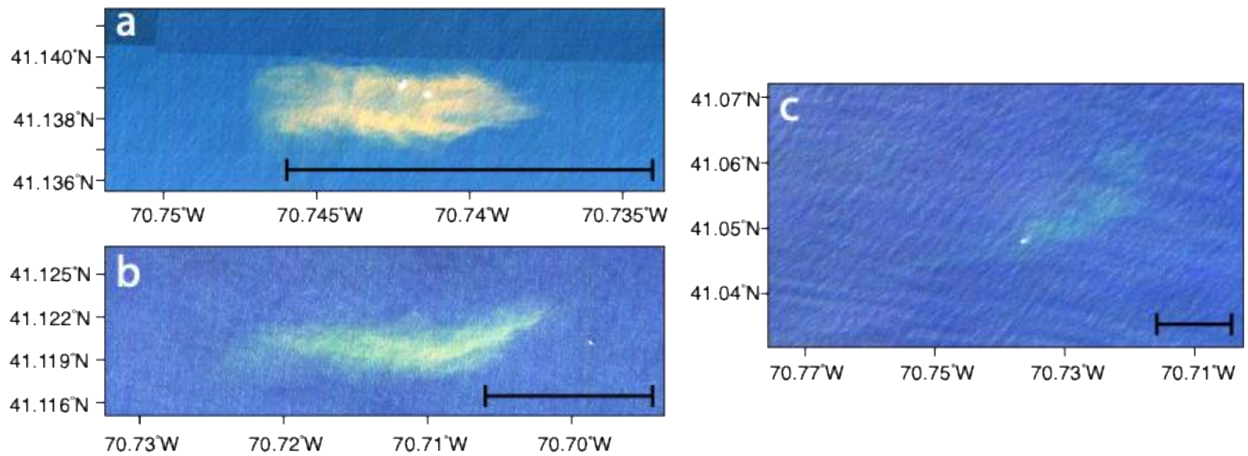
- Net Alkalinity transfer from alkaline feedstock into seawater via dissolution, or (in the case of addition of dissolved alkalinity) confirmation of no loss due to precipitation(not addressed here as everything is completely dissolved);
- Tracking of dissolved alkalinity and its dispersion (and for solid feedstocks, particle transport, dissolution, and settling);

- Calculation of CDR due to the above processes, via direct measurement, models, and/or a combination of both.

-One major question I am left with regarding the dispersal of the tracer is the path of the patch and its north-south extent. Figure 5 shows that the team followed the tracer in generally east-west transects, which slowly moved from north to south over the experiment as the tracer was advected. But did the tracer move north to south as a fairly coherent patch, with the total surface area not changing much, or did it 'smear', so the tracer patch grew in total surface area over time? Essentially, it seems to me that the patch was well-sampled in one direction (east-west), but lightly-sampled in the north-south direction due to understandable time restrictions. Figure 3b shows a lot of heterogeneity in the patch concentrations: how can the authors know that they were always sampling 'peak' patch RT concentrations in later Figures (i.e. Figures 6, 10). This seems to have large implications for the estimate of CDR efficiency starting at Line 621, since the efficiency calculation is predicated on all the added TA remaining in the patch (i.e. no precipitation). Perhaps this is mostly unknowable, but Figure 3a shows that some drone imagery was taken, and perhaps maybe satellite imagery could be available as well, which might inform this question of the patch extent.

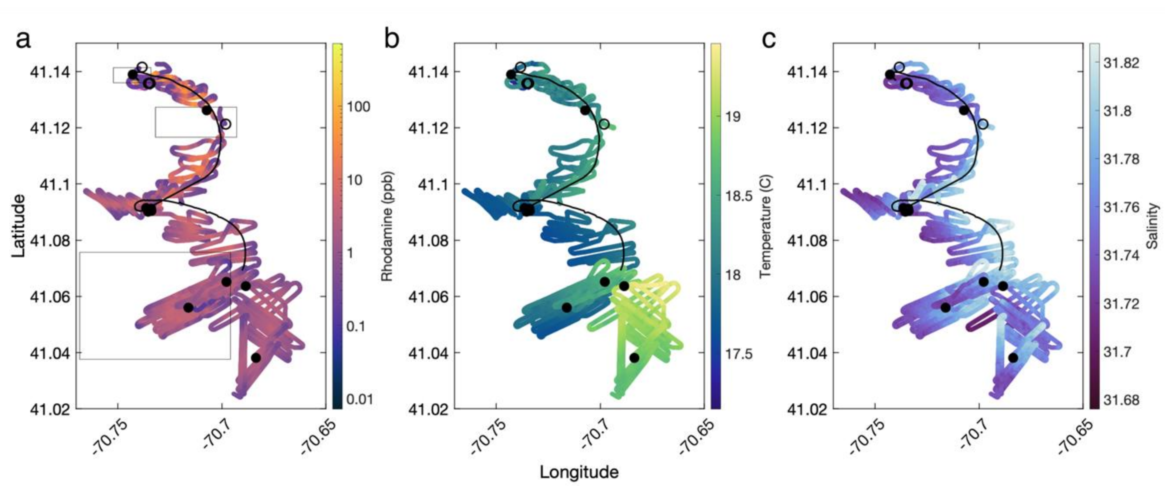
We now include satellite imagery (see below) to confirm that the patch started to smear east-west, and that we sampled through the longest part of it. While we can never be absolutely sure that we sampled the peak, the constancy and decay of the peak RT signal in the underway system suggests that we were not randomly sampling the patch but instead systematically passing through the peak concentration every time.

In all three images, the patch can be clearly visualized, and demonstrates that the patch did stay coherent and move as a unit, rather than getting dispersed irregularly. Over time it stretches mostly east-west, with some southwest-northeast trending observed at 24 hours. This matches well with the long axis of sampling demonstrated in fig. 5. We plan on adding context to Figs. 5 and 6 to indicate where and when these images were taken with respect to the spatial ship track and underway timeseries data streams.

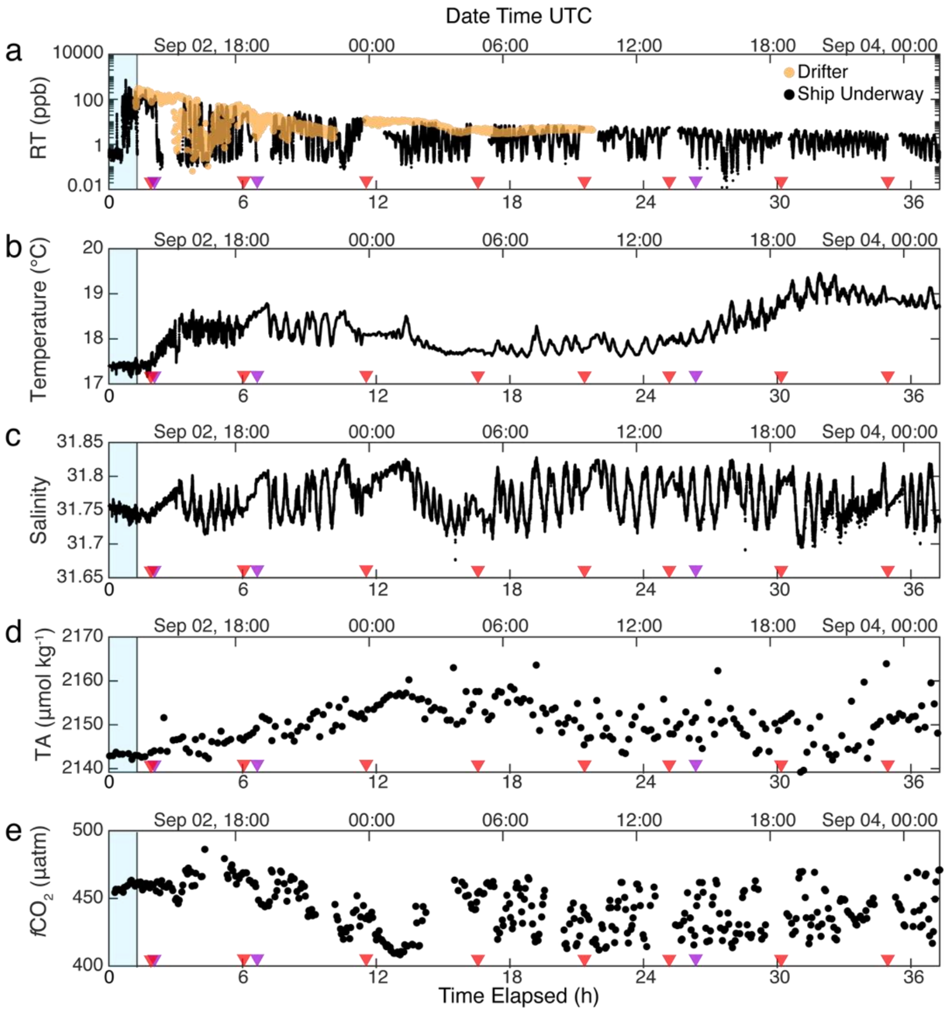


Proposed figure addition: Satellite imagery collected via Planet Labs at 3 time points: a) approximately 1 hour after dispersal; b) approximately 6 hours after dispersal; and c) approximately 24 hours after dispersal. Images shown are true color, scaled to enhance image brightness. Scale bars in each image are 1km.

Methods addition: High resolution satellite imagery was collected during the cruise from Planet Labs via two methods: 1) Ultra-high resolution (0.5m per pixel) multiband imagery (red, green, blue, near-infrared, and panchromatic) was collected via the Planet SkySat constellation through tasked image collection and 2) high resolution (3.0m per pixel) multispectral imagery (8-band) through the PlanetScope near-daily revisit product via the Dove/SuperDove constellation. Three images were collected during the cruise, on September 2, 2023 at 14:14:25 UTC, September 2, 2023 at 18:58:16 UTC, and September 3, 2023 at 14:43:07 UTC. The first two images were collected via SkySat tasking, and the third via PlanetScope. Level 0 images were internally processed by Planet's algorithms for orthorectification and atmospheric correction to produce Level 3 surface reflectance data. Orthorectification was verified using shipboard location data, which required small corrections for images two and three.



Updated underway data map showing satellite image capture windows in panel a.



Updated underway timeseries figure showing satellite image collection times in purple triangles along the x-axes.

MINOR COMMENTS

-L51: “Research” Fixed, thank you.

-L562-63: I don’t think semicolons are the right punctuation here. Revised accordingly.

-L74: I was not familiar with the concept of ship wake dilution models. A sentence quickly explaining what these are and what they are used for might be helpful. Thank you, we will be sure to add some context here.

-L84: fCO2 has not yet been defined Thank you, we will add a definition of fCO2.

-Figure 1 and caption: most of the paper uses fCO_2, not pCO_2. The difference is minor, but consistency would be good. Thank you for the catch, we have now changed to fCO_2 throughout.

-Figure 1: these plots show what would happen to the carbonate system in the surface ocean with infinite dilution, I think. My understanding is that for a fixed volume of seawater, OAE eventually results in a DIC that returns to near-baseline levels, but at a slightly higher concentration, and that pH also remains just slightly elevated above baseline levels. It's a little hard to tell from Fig 1, but I think DIC and pH return to baseline conditions. This is a subtle difference that might be worth mentioning. Thank you, yes we assume infinite dilution, i.e. the case where the intervention is relatively small relative to the volume of the body of water. Interventions would have to become significantly larger (gigatonne-scale), or the body of water significantly more restricted, for there to be a steady-state mean shift in TA, DIC, or other carbonate system parameters. We now make this clear.

-L109-110: Is it worth keeping the $\Delta TA(CDR)$ term in Equation 1a if it is always zero? We prefer to keep it in for completeness, as it is unclear for now how feedbacks associated with CDR on the TA signal will be made and assigned.

-L141-142: does storing CO_2 in biomass require a sustained, measurable fCO_2 gradient? This is an interesting point. Largely, the answer is no, because some groups/companies are considering terrestrial biomass waste storage in the deep ocean, which would not generate a CO_2 gradient in seawater during the biomass formation. But if the biomass is grown in seawater, measuring the decrease in the fCO_2 in surface water as a result of that photosynthetic activity would help with its MRV.

-L153: "(90-foot)" Thank you, changed.

-L163: strange citation format with all four authors listed Thank you, fixed.

-L190: define "UHMW" Thank you, we just decided to remove that acronym for clarity as it did not add much.

-L193: refer reader to Figure 2. We think the best figure to reference here would be Fig. 3, which shows the image of the patch. We do so, but note that this does now reference figures out of order in the text.

-L200: 38 knots is very fast for the *Connecticut*, unless it is a cigarette boat Thank you for the catch! It should be 3.8 knots, changed.

-Equations 2 and 3: I'm curious if these are constructed to apply in straight-line travel conditions. How does the spiral dispersal pattern used in this experiment work in these models? It seemed like the swaths of RT sometimes dispersed into each other, which I imagine would have an effect. This is an interesting point. These equations were designed to operate on a ship traveling in a straight line, discharging waste or ballast behind it. At least with the first pass over the patch, the underway system intake is always upstream of the RT discharge, so we would only "see" the original material on the first pass-through. On later passes, though, there would be some cumulative effect and the dilution would actually be less than observed. We now make a note of this, and indicate that this approach provides an upper bound on dilution, which we may not have reached by dispersing in this way.

-L217, it might be useful here to describe what the "in-patch"/"out of patch" thresholds were. It's later mentioned on L328. Thank you, added here for clarity.

L240-241: include metric units here (inches and feet used) Thank you, added.

-L239: the TA was distributed at the surface, but the underway intake was at 1.5. Could this have affected anything in terms of peak RT (and thus estimated TA) concentration, patch edge detection, etc? We don't think so, as the dye appeared relatively well-mixed in the upper portion of the ocean, as shown by Fig. 7 and S1. CTD sensor readings less than 1-2m using the CTD rosette are very difficult to interpret, regardless, because of ship heave.

-L245: It seems the total time from the ship's hull to the fluorometer was about 30 seconds (mostly delayed in the debubbler). Does that lag time seem about right? Yes, that is about what we calculated as well, and we now add this to the methods for clarity.

-L260: "analyzed" Thank you, fixed.

_L293: "what does "subsequently routinely" mean? Were all samples not poisoned? Edited for clarity. All samples were poisoned.

-L294: the 13-C DIC data are only presented in the supplementary, and not discussed anywhere that I can see. These data and associated methods probably don't need to be included. If they are needed for some reason, how was the 13-C isotope ratio calibration done (Dickson CRM is not certified for 13-C). We provide these data because we collected them, and think it is important to include here for completeness. But agreed that we did not use them in the paper. We provide some details on our calibration procedure for d13C in the methods now.

-L317: were the four drifters all deployed in the same place, at the same time? Yes, they were deployed essentially all at once, at the same location. We were interested to see how far they would drift from each other, which the traces demonstrate is minimal. We add some clarification and some discussion on this now.

-L339: what concentration of liquid NaOH is assumed? We assume the addition of 50% by weight NaOH (19.1 molar).

-L345-349: this section shifts into present tense. Thank you, edited for consistency.

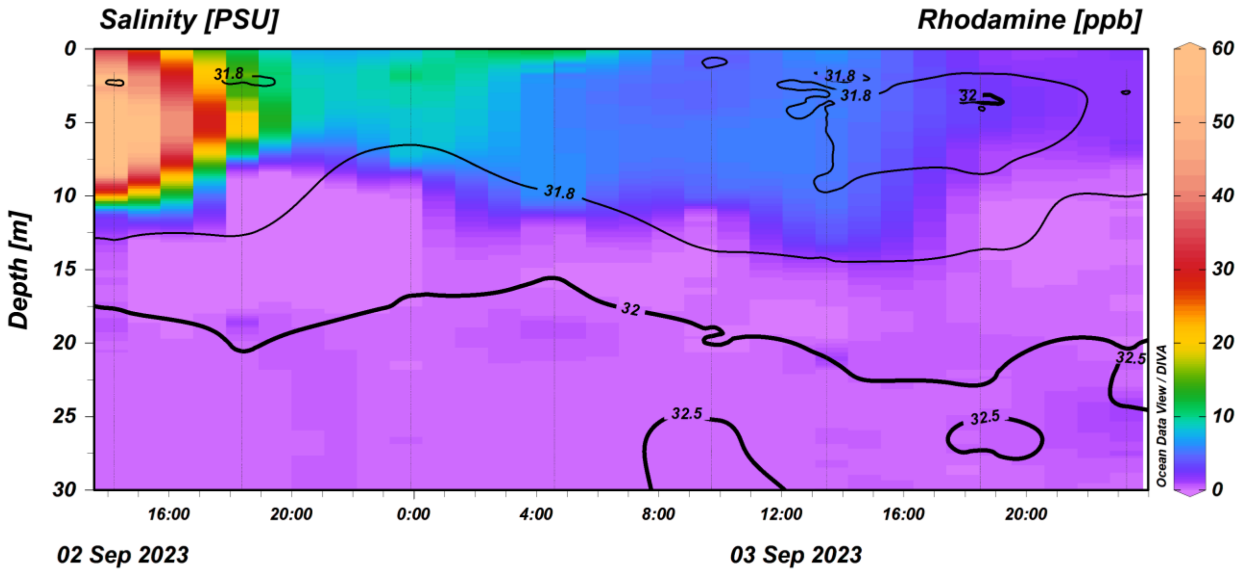
-L423-424: this gets back to the convening swaths of RT dispersal. It would be good to discuss the implications of this some more, in terms of OAE. Thank you, we add a sentence of two of discussion here on this point as well. This may influence the small scale "patchiness" of the patch, and will not be captured in larger-scale models, which will assume a uniformly distributed patch of OAE. Something similar could potentially happen with coastal outfalls, given tides, currents, and wind that could "pile" up water on top of itself. Therefore, it is likely critical to investigate these very near-field processes in some more detail, with both models and observations.

-L439: the decrease in RT concentration is hard to see with the color scale used in Fig. 5a. Also can mention here the decrease is seen in Fig. 6a as well. Thank you, added this for ease of reading/comprehension.

-L460: It's hard to parse the three exceptions noted here. Perhaps number or identify them individually somehow. Thank you, we will do this.

-Figure 6: Seelmann et al. (2019 p. 526) note the occurrence of outliers in underway TA data from the Contros HydroFIA instrument. Just going on looks, the high TA points in Figure 6d of this manuscript seem to match those criteria. The Seelmann paper cites a method to filter out these outliers as well. Thank you for this – we did filter outliers, but did it by eye. The Seelmann et al. (2019) approach was for a single seawater sample measured over time, and therefore the Grubbs outlier test they used was sufficient. However, because we sample spatial and temporal gradients, this univariate approach is not appropriate. We have tried a few different methods of removing outliers mostly focusing on calculating a running mean of 5-10 samples (corresponding to 40 minutes to 80 minutes in window width) and finding outliers either through a standard deviation cutoff or a z-test ($p<0.05$) cutoff. We found similar results with these statistical methods and our by-eye method, all resulting in about 2-3% of the data points being removed. We will discuss the approach briefly in the paper, and can include a figure in the supplemental identifying the outliers removed.

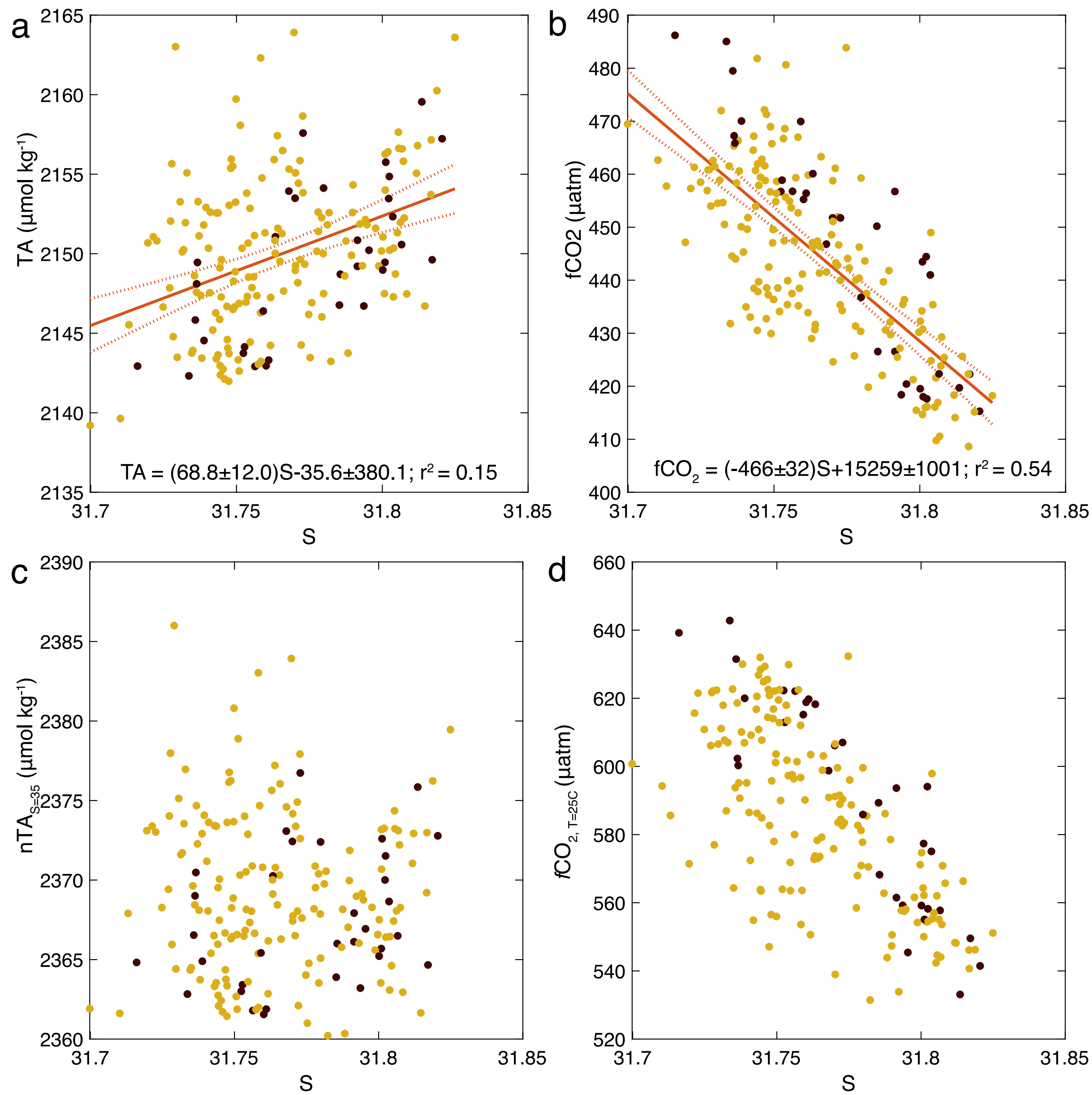
-Figure 7: the salinity 31.8 contours are very faint, and will be hard to see in the final paper Thank you, we made them stronger in this latest version:



-L501: Is "interpolation" the right word here? I usually think of interpolation as filling in gaps in lower-frequency data to match higher-frequency measurements. As in, the TA data in this experiment could be linearly interpolated to match the higher-frequency T, S etc. Were the higher-frequency data subsampled, or averaged somehow, to match the 10-minute TA data? Thanks for this – yes, the appropriate word may be "downsampled", as we downsampled the T,S, and fCO2 data, matching the closest timestamped measurement to the TA measurement timestamp.

-L508: describe the statistical test used to compare the fCO2 data categories. we used a two-way t-test and now provide p values for the comparison. For fCO2, the categories are the same ($p=0.4$), and similarly for TA ($p=0.15$).

-L515:516: describe the statistical tests and coefficients which indicate correlation between fCO2 and TA. We used a linear model ("fitlm" in MATLAB) and now provide those as supplemental figures, along with the goodness of fits (r-squared). Another reviewer also wanted to see salinity-and temperature normalized TA and fCO2, which we also include in this supplemental figure as panels c and d.



-L530: should be Fig 3c. Thank you, edited.

-L555 same question regarding interpolation Changed to clarify and be more specific.

-Figure 9: it is really hard to tell the two boxes/sets of horizontal lines apart. The open box and shaded box are easy to mix up. One option is to use dashed horizontal lines for the mean and +/- standard deviation of TA_bl, and solid lines for mean +/- standard deviation for TA_bl,interp, or some other similar scheme. The shading doesn't really add anything for me. Thank you, we have edited this figure to add colors for clarity.

-L578-587: I thought this section was really well done. Thank you.

-L595, fCO2 returns to near-baseline values Edited

-Figure 10: do all the points in this Figu correspond to points in Figures 8 or 9? Are they all matched up to the underway TA measurements? These are now the peak values in each hourly bin, so are a subsample of the values plotted in Figures 8 and 9. Figure 8 shows all TA and fCO2 data, downsampled to match the TA sampling resolution. Figures 9 and 10 only show the peak hourly data. we clarify this now in the text.

-L606: should be uatm, not ppm. Fixed, thank you.

-Figure 10 caption, Line 2: should reference Figure 8 I think, not Figure 7. Also, do you mean "RT concentration" instead of "RT signal"? Fixed this, thank you

-L612: what is a "small but mean" offset? We mean that there is a significant offset between the mean values, edited for clarity.

-L623: should reference Figure 10h. Line 593 days the TA enhancement was >20 umol/kg) Thank you, added.

-L624: should reference Figure 10e Added

-L627-630: Isn't this a big deal? How do the authors suggest the spatial heterogeneity of the patch should be addressed? This is a good question. This spatial heterogeneity is what makes MRV so difficult using measurements, and yet these measurements are critical to provide some level of validation to models being used to calculate CDR. One approach would be to try to capture the tracer distribution with imagery, as we have done with satellites. Another method would be to apply statistical and/or gas exchange models to the RT (and OAE) distribution over time, but one would somehow need to know the representative area associated with each concentration measurement. This could be accomplished using drifting assets equipped with sensors, sampling a larger area of the plume continuously over time.

-L627: this focused on the peak RT concentration measured in the patch. Could it not have been higher in another spot? Wasn't the patch quite patchy (i.e. Figure 3b)? This is true, but we do think we were sampling the peak concentration consistently, as described above.

-L634: how likely is it for the water mass to remain at the surface? The mass in this experiment advected quite a bit, and presumably mixed as well. This is a good point. It really depends on the conditions, and whether there are other water masses in the area that could interact with the patch. For instance, in coastal areas with significant freshwater input, seawater will readily subduct under freshwater lenses. So, this study is likely more appropriate for open-ocean scenarios where the water column is better stratified and there is less horizontal heterogeneity. It also clearly mixed, as the concentration decreased steadily over time. However, interestingly, it did stay as a coherent “unit”, meaning that these features can be tracked over time with precision.

-L643-644: this returns to my question of patch behavior. Did the whole mass travel south, or did it spread out south with some concentration remaining at the original northerly dispersal location? The patch did move as a coherent unit, and while it did, it spread and diluted with surrounding water. We hope this is made clear with the satellite images.

-L678: “may advect may disperse” Fixed, thank you.

-L705-710: I thought this section was well done. We appreciate your feedback!

-Final thought: I suspect that some will read this work, see the 8% CDR efficiency number, and be disheartened for the prospect of OAE. Perhaps the authors can speculate about this some more. How much longer, or how much more broadly spatially, would observations need to be carried out to observe a higher efficiency? What might be the highest detectable efficiency, given uncertainties in the baseline estimates and analytical instrumentation? More broadly, what even represents a successful CDR? One that is measureable? One that is profitable? Presumably the CDR effects would have continued after the monitoring efforts in this experiment were done, but how might someone realistically do MRV for these ongoing CO2 removals? These are interesting, at times philosophical, but highly relevant questions for in-water measurements for the purpose of MRV of OAE. The 8% number might not sound like a lot, but many papers are suggesting that near-complete CDR via OAE takes years to complete. Thus, we feel that constraining the initial 8% of the process via direct measurements is actually a strong argument for pursuing in-water research on OAE and its CDR. While we do not feel we can extrapolate our signals to beyond our dataset, obviously the best scenario would be to follow the patch and measure it until it completely returns to below detection limits, but even in ideal scenarios for small-scale experiments, this timescale is likely on the order of days. This would allow a more complete budgeting and assess more complete efficiency directly from measurements. We do acknowledge that at some point, models must be used to estimate CDR, because it will become unmeasurable, but measurements that can help ground -truth and validate these models are essential at this stage. We add some discussion on this topic now.