Anonymous review 3

I found this paper to be extremely well written, very nicely presenting the results of the field work tracking rhodamine tracer as a preliminary step to tracking a future OAE experiment via the dispersal of a NaOH solution. I commend the authors on stepping through every aspect of the experimental design, deployment, interpretation, and implications for the future OAE experiment in a highly readable and comprehensive way.

We thank the reviewer for their kind words.

I have only one somewhat major suggestion, which doesn't require new analysis but might trigger some careful thinking about reframing the motivation and recommendation of the paper. I think my suggestion is related to Reviewer #1's question about "What is your MRV framework?" I see the work presented in the paper as foundational research that can inform an MRV framework, but not itself MRV nor a framework. The term MRV signals a set of processes used in a voluntary or regulated carbon marketplace, whereby carbon removal is monitored and reported to a third party who verifies that it met a given set of predefined standards. The language in the paper should make this distinction.

We thank the reviewer (and Reviewer 2) for pointing out this key distinction, and it is one that we have now worked into the manuscript throughout – specifically in the introduction and in the recommendations at the end -- to distinguish this research effort from what industrial/practical MRV could look like. In the end, we decided to reframe our approach as an OAE analytical framework, rather than as a specific MRV methodology, which we agree would be difficult to implement across the board.

Further, I suggest the authors to reflect on how much of this kind of extraordinary research effort (deep domain expertise, research vessel and state-of-the-art laboratory infrastructure, a valuable equipment pool, and the funding to support the team) they would recommended for future MRV. At a minimum, the target audience for the work should be made explicit (my guess: researchers trying to design a comprehensive field experiment with relevance for tracking surface anomalies influencing air-sea gas exchange). It would be slightly more ambitious, but commensurately more valuable, for the authors to reflect on a trajectory for mCDR researchers and practitioners that uses the knowledge gained from this research to ultimately create a more parsimonious MRV framework for operational mCDR (or at least provide recommendations on how to bridge that gap).

Thank you – we have now added some content to the paper to address this important point. While we agree that this level of effort might not be conducted by every group attempting MRV in the future, we do feel that it is important for in-depth, in-water research to accompany early field projects to continue developing and refining MRV for OAE. These

measurements will continue grounding models with critical observations, and may provide additional insights into the OAE and CDR processes. Right now, it is unclear what "operational" MRV should look like, and studies like this one can help guide the field towards a meaningful, trusted, and scalable MRV framework.

Minor comments:

The first line of abstract is deeply misleading: The solution to a 10^10th ton emissions problem is not meaningfully "supplemented" by processes that now collectively sum to 10^4th tons of removal. Please remove this sentence. Removed.

Line 30: Wording is a little confusing. I think you mean that "Over 36 hours, the ensuing anomalous carbon dioxide uptake by the ocean was driven by the enhanced air-sea gradient in fCO2. The calculated CDR signal was detectable as a 3 uatm surface ocean fCO2 increase, a pH decrease of 0.003 units, and…". If this is an accurate interpretation of what is written, it would be more understandable to the reader to spell it all out. Thank you, we edited this for clarity.

51: typo: Researc Fixed.

Figure 1 schematic is very helpful and well done. Thank you, we appreciate the feedback.

Line 133 and throughout - specify air-sea gradient in fCO2 (or name a variable ΔfCO_2) for clarity thank you, done.

357 - Why not make the gas exchange coefficient wind- or wind- and wave- dependent, as we know it is sensitive to wind speed and wave height. It would seem hard to justify using a constant, given the variability in Figure 2. We now calculate an hourly gas exchange coefficient from the Block Island wind data shown in Figure 2, and use that in our calculations. We note that wind speed was a bit faster on average than typical conditions, meaning that this should only increase the CDR potential of our hypothetical experiment. In addition, we note that small-scale (e.g. hourly) variability in windspeed may not make a huge difference in the overall CDR, because of the CO2 gas exchange timescales being so long – similar to the reason why small-scale baseline variability also does not make a huge difference, as already described in the manuscript.

Figure 2: Specify that bathymetry is in m. Zooming closer to the ship track would be helpful, as would adding a circle of the size of the initial dispersal spiral. We include another panel of this figure that shows a zoomed in view of the patch and the dispersal location itself.

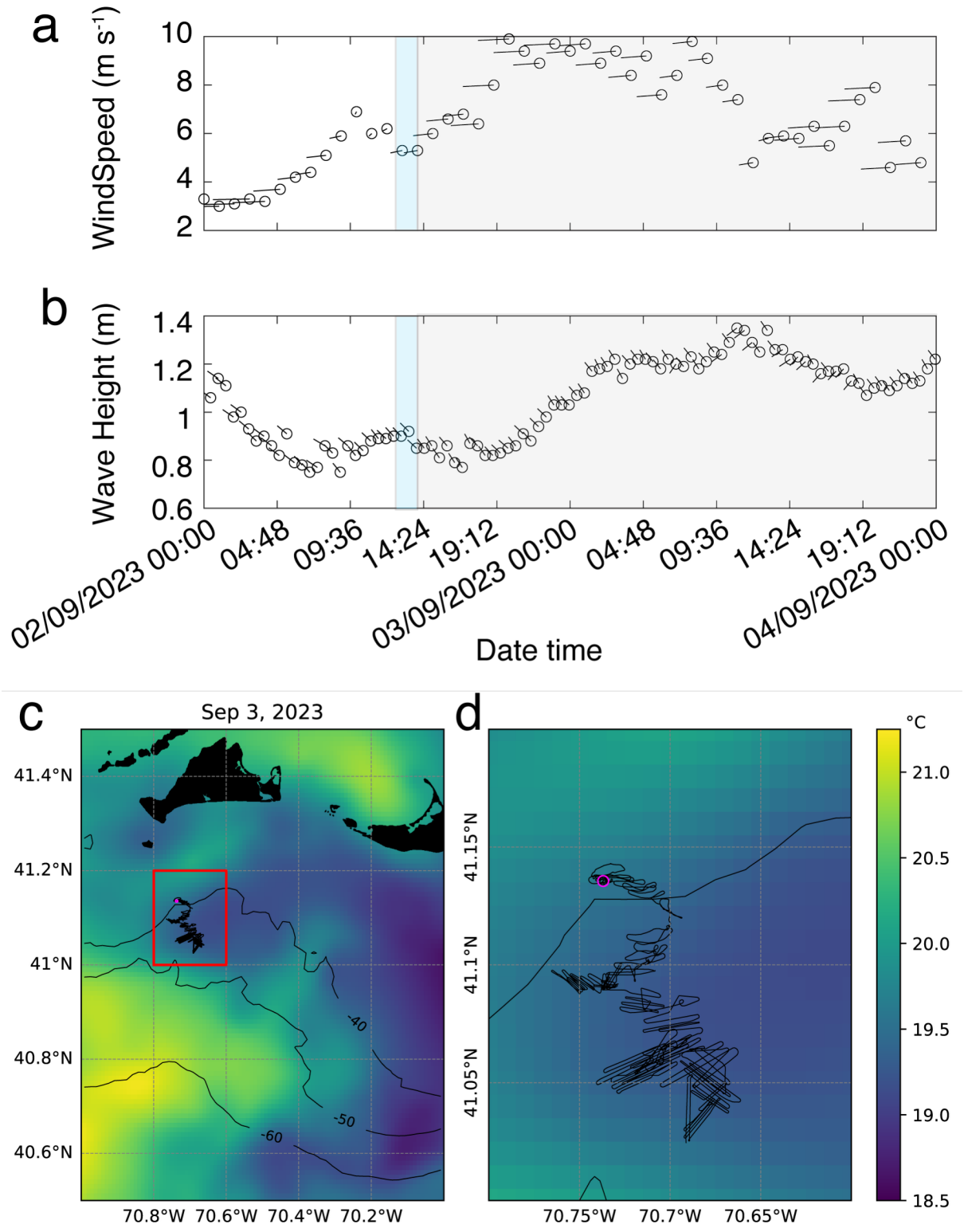
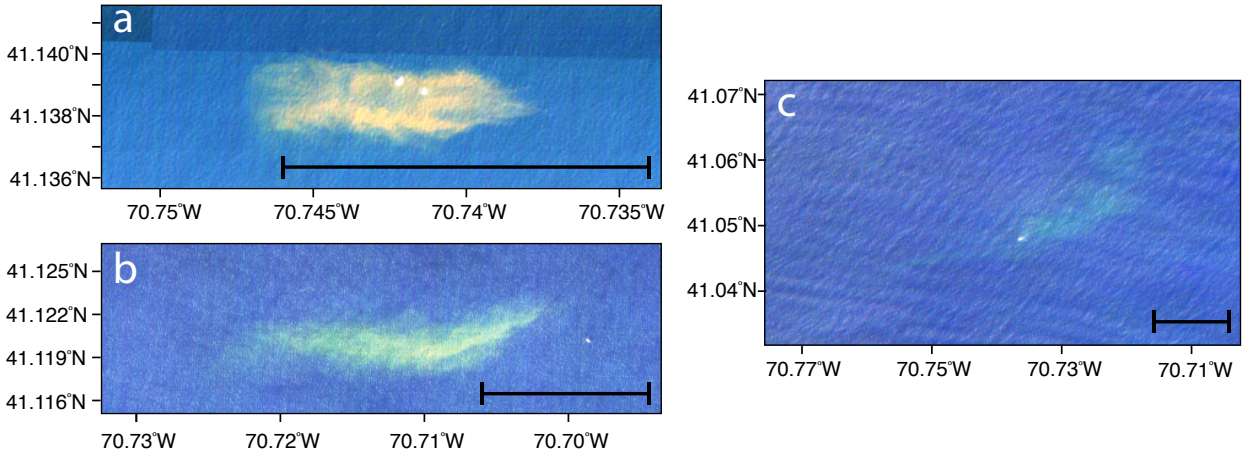


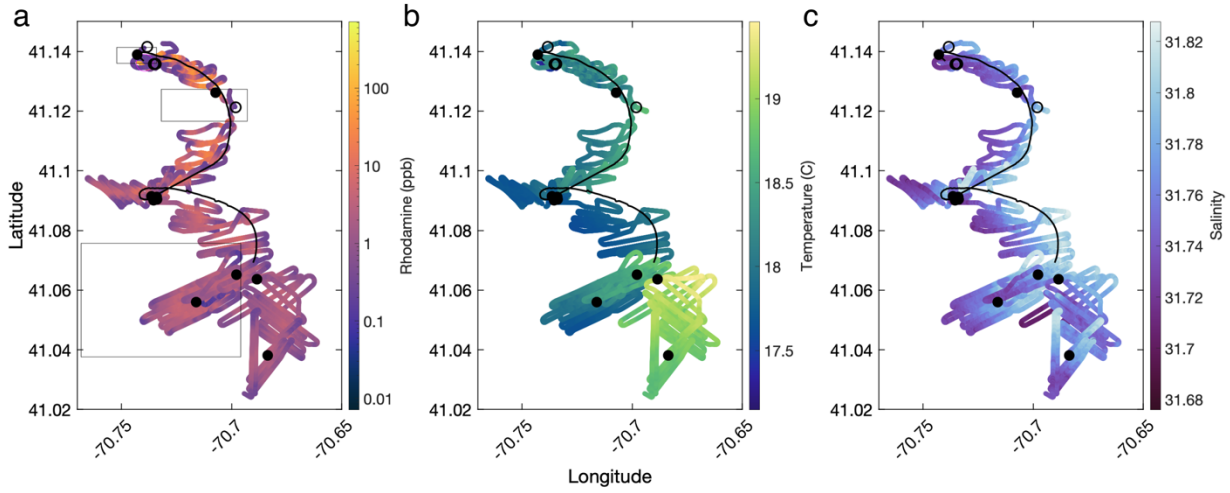
Figure 3: If you had a drone photo of the patch taken at a later time than Figure 3a that showed the stretching and spreading of the dye, I think it would be super helpful for the reader. We now include satellite imagery for 3 timepoints which clearly shows the patch

and its spreading (and stretching), and include some discussion on this point and how it maps on to our monitoring track in Fig. 5.

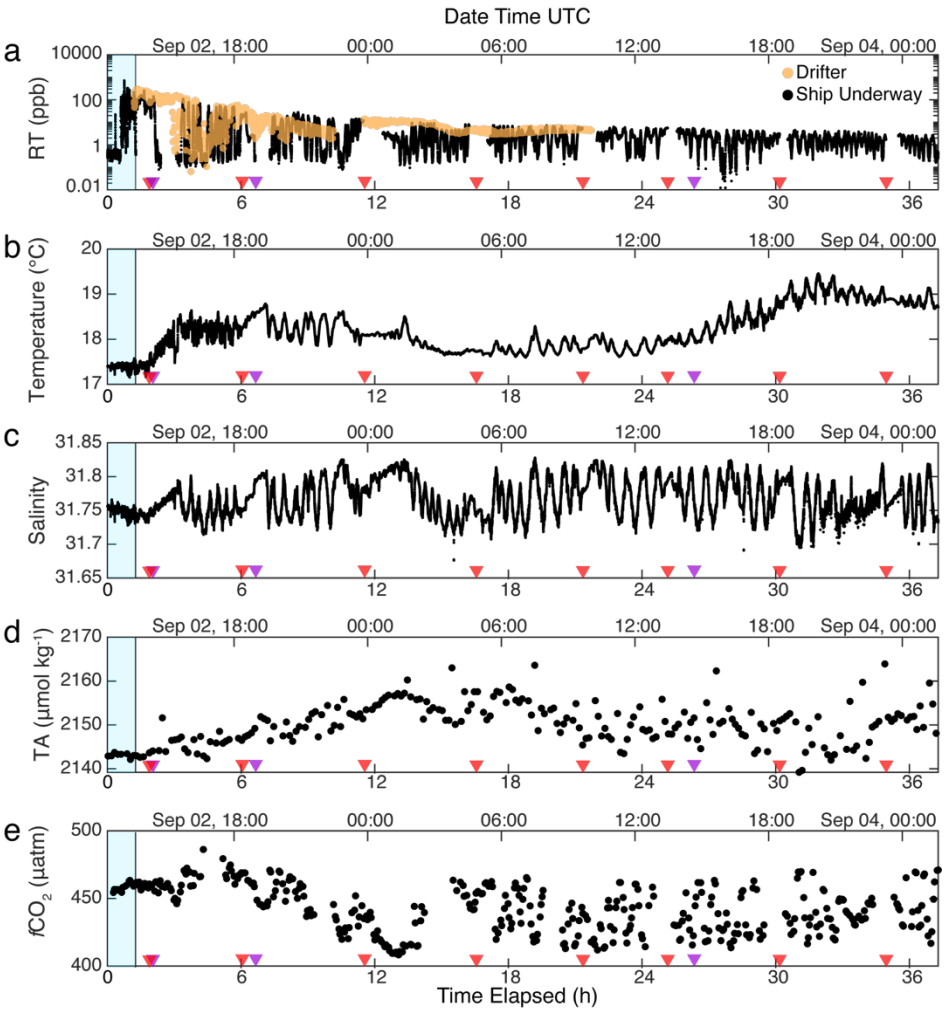


Proposed figure addition: Satellite imagery collected via Planet Labs at 3 time points: a) approximately 1 hour after dispersal; b) approximately 6 hours after dispersal; and c) approximately 24 hours after dispersal.

In all three images, the patch can be clearly visualized, and over time 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 have added context to Figs. 5 and 6 to indicate where and when these images were taken with respect to the ship track and underway data streams.

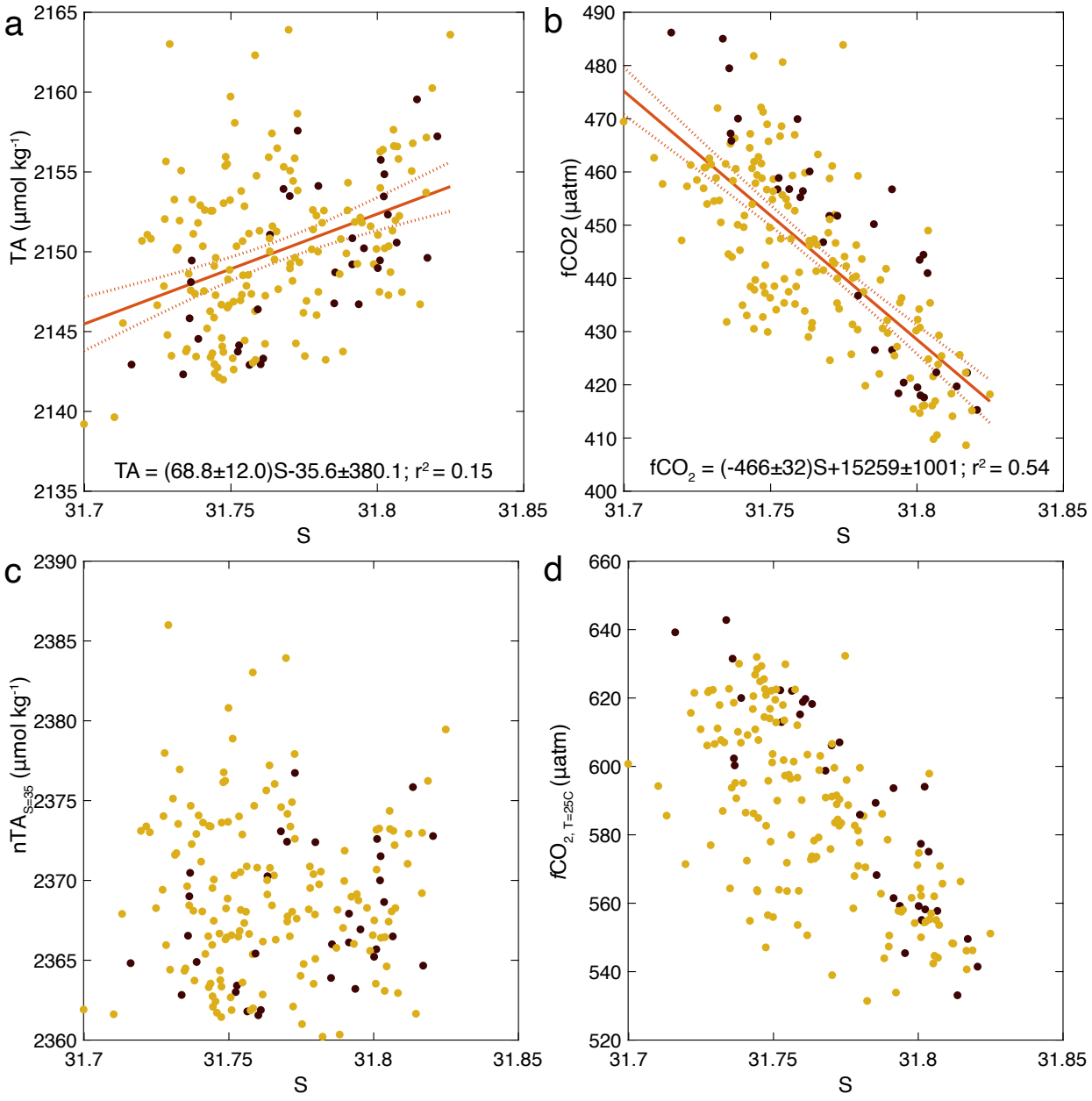


Updated underway data map showing satellite image capture windows in panel a.



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

Figure 6: consider adding salinity-normalized TA to panel d, and temperature and salinity-normalized fCO2 to panel e. The variability in T and S will be reflected in the carbon variables in a way that would be helpful to separate. We feel that salinity-normalizing fCO2 is overly complicated, but agree that temperature-normalizing fCO2 is interesting to look at, as is salinity normalized TA. We have done this now, and as expected, nTA shows no significant variability with salinity (implying that salinity drives the TA gradient between the water masses). Temp-normalized fCO2, however, shows an even larger gradient as a function of salinity – upwards of 100 uatm-- suggesting that thermodynamic effects cannot be driving these changes. Instead, it is likely the biogeochemical signature of these two water masses. We include these panels in the supplemental, rather than in the main text.



569 - DIC should also be sensitive to biology, no? This is true, but likely not on the 36-hour timescale shown here, given the magnitude of any biological changes and the size of the DIC reservoir (for the same reason that OAE-based CO2 uptake takes a long time to happen, and to see the DIC increase).

Figure 9 - I suggest using some colors! There are several solid black lines, so that it's hard to track which one the caption refers to. Thank you, we have now added colors to this plot to make it clearer.

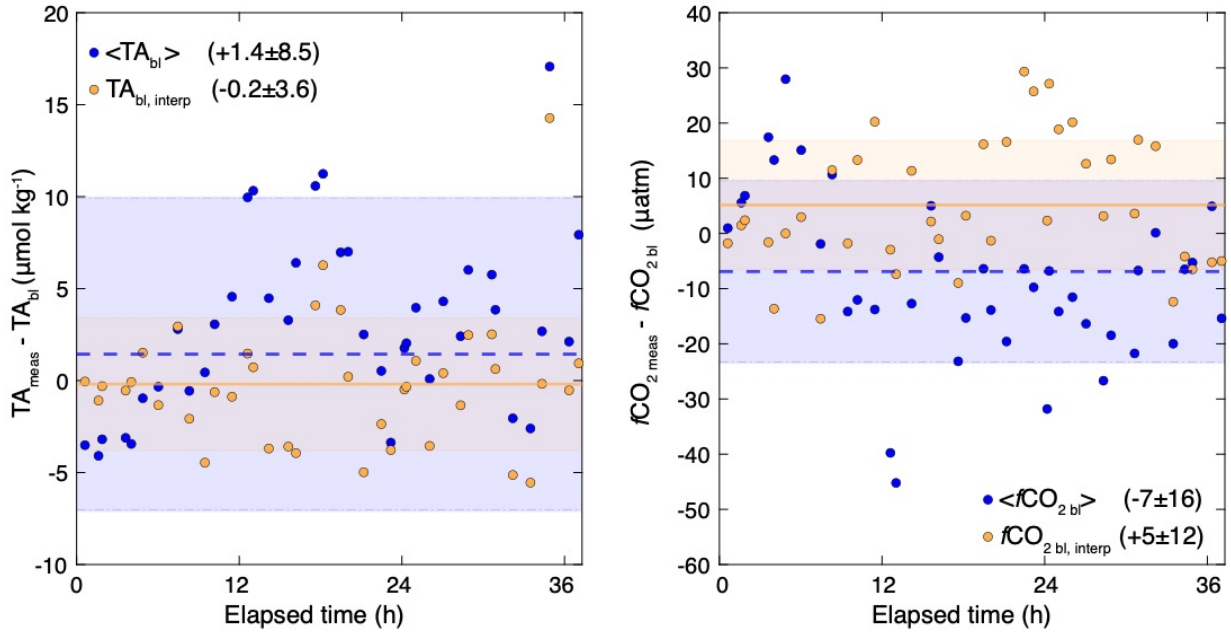


Figure 10 - I'm confused why there seems to be only 1 dot per hour, when the sampling frequency was 10-minute. We discussed in the methods and above why we chose to use the peak value every hour – it makes for a more interpretable and cleaner result for this thought experiment. During real OAE deployments, data should obviously be collected and analyzed at as high of a frequency as possible, and per another reviewer's suggestion, we add some discussion about spatial heterogeneity and how to handle/measure it.

606 - yellow minus blue values — add note here (and/or in the caption) that the points are barely visible because of overlap in Figures 10 f and g Thank you, made this clarification.

628 - 638 "MRV approaches that can accurately capture the entire patch budget … " This is related to my "major suggestion" above: This statement makes it sound like MRV would be done in a way that mimics this experiment (with RT or similar tracer and a research vessel tracking it for days). But doing such hugely expensive field campaigns is a research-level activity that is likely beyond the scope of scalable MRV. In fact, I'd be shocked if 1 ton of CO2 wasn't emitted by the MRV activities in their totality (4 days on the R/V Connecticut + chaser boat + all the instruments, supplies, and shipping), cancelling most/all of the mCDR here. Thus, I see this as a research activity, rather than an MRV approach, and I am very interested in the authors' thoughts on the difference. We thank the reviewer for this important clarification, and will make sure to clearly separate research activities (such as this one) from "practical" MRV in the revised manuscript.