In Casey et al., the authors investigated the performance of calibration models developed for ambient O3 and CO2 across time and space using field deployments spanning 2014 – 2017 as case studies. Specifically, they looked at the impact of post-deployment calibration vs pre- and post- calibration, and the impact of applying a calibration model developed in one location on U-Pods deployed in other locations. Calibration models investigated included linear models and artificial neural networks.

The size and scope of the study is impressive, and I believe there is a significant quantity of insightful information within this paper. However, in general, I found the narrative of the paper to be confusing (it is hard to effectively distill such a breadth of research) and the take home points could be made considerably clearer. Additionally, I think this paper would benefit with a few more analyses of general model performance implications and a closer look at the impact of relative humidity. Following these corrections to comments identified below, I believe the publication is suitable to be published in Atmospheric Measurement Techniques.

GENERAL COMMENTS

In general, I found this manuscript a little hard to read, because I felt like a cohesive narrative was missing. A lot of information is presented in a rapid-fire manner such that the results and discussion section reads more like a results section, with limited discussion. Although there is no straight forward solution to this problem, I would suggest that the authors think about the three to five key messages they wish to convey in the manuscript and that they tune and streamline the text to support this narrative. Many of the figures are needlessly complicated by an overload of case studies, unintelligible sensor signal labels, and colours. If there is any way of summarizing this data more cohesively, it would significantly improve the paper.

It would be good if the authors could elaborate on which U-Pods were where over all these campaigns. Given that temporal degradation / time was investigated in detail in this paper, some assessment of U-Pod changes over the three years of campaigns would be helpful if possible. Also, when comparing sensor performance spatially, are the U-Pods that are compared the same age?

Directly addressing the size of the training and testing windows should be included. It is hard to make generalizable conclusions from the study when there is so much variability in training and testing window size. Is there a reason why some training windows are shorter than others? This should be directly addressed in the manuscript.

I found the discussion of ANN and LM model building to be significantly under-developed, especially considering that this is a measurement techniques journal. This paper relies too heavily on the prior 2017 study, and has too much assumed knowledge that should be summarized in Section 2.4. The resulting LM and coefficients should be provided. As well as some mention of model performance metrics like MAE or r2.

It would be good to include an explicit discussion of % reduction in error by using established models vs. “best fit” models. Can we generalize? What is the quantitative impact of using your prior models vs making a new model every time. My interpretation from this paper is that we need a new model for every U-Pod for every deployment – is there any way around this? I feel there is a significantly missed opportunity to be quantitative here.

Section 3.3 could be substantially enhanced using some sort of summary figure/table (other than a target diagram) that gives percent change in bias, random error, r2, mae etc. by switching from pre/post to just post, or by switching location. Given that there are many pairs of sensors looking at impact of pre/post vs. just post or impact of location switching, you could show average % change in model fitting statistics as well as confidence intervals or standard deviations to show the spread across the case studies. This might be a helpful way of streamlining the paper.
This is mentioned in the specific comments, but I would like to see a quantitative assessment of the impact of swapping out RH data if a U-Pod failed. You could accomplish this by taking a U-Pod with valid RH data, replacing it with the Picarro or nearby station RH data, and quantitatively assessing the impact on model performance. That way, you could transition from hypotheticals about the impact of this data swapping to some actual numbers.

**SPECIFIC COMMENTS**

P1 - L13-14: Seems like an oxymoron to say “Generally” if the circumstances for best model performance are case study specific. Recommendation to remove the word “generally”.

P3 - L19-24: Please discuss why ozone is elevated near O&G production.

P3 – L27: Can you quantify “small spatial scales” in this context? Is well pad combustion and diesel traffic really contributing so much that it is universally increasing ozone? Most of the construction traffic would occur during active drilling and less so during production when well pad sites are very quiet. I think some further thinking or elaboration on this train of thought it warranted.

P4 – L1-2: What do you mean by “pooling” of compounds - I am not sure I understand this sentence.

P4 – L5-8: I think some short discussion of the operating principles of the sensors would be helpful here.

P4 – L25: Not sure what is meant by “toward” here

P5 – L27: Is this “clean air” normalization done dynamically/in real-time in parallel with the actual measurement? Or is the clean air measurement established during some calibration/maintenance? Please clarify.

P5 – L30-32: Is there expected to be spatial variability of RH? Why not just replace the RH sensors directly?

P7 – Section 3.0 first paragraph – Are there some general conclusions from the SM that you can discuss here? Some discussion of model performance is warranted vs just describing what figures are in the SM.

P8 – L17: What is eltCO2?? Can you better define all the model parameter inputs? This comes up in Figure 9 as well.

P8 – L30-33: Is this early morning under prediction really true? Bloomfield doesn’t look like it is exhibiting any diurnal variation in residual error at all… I feel like given the small number of U-Pods, it is hard to make this conclusion definitively.

P9 – L4-13: I am confused now – why did you use the model with three inputs (eltCO2, abshum, and temp) if the best performing model had more variables? I feel like the model selection discussion is substantially underdeveloped. There could be many good reasons to not choose a more complex model, but any discussion of this seems to be completely omitted, or the reasoning is too difficult to follow.

P9 – L28-30: It would be good to do a more comprehensive assessment of the impact of replacing RH sensor signal on model performance. Could you conduct a dummy experiment where you replace RH data you actually logged with that of a nearby or alternate monitor and then quantify the impact on model outcome? Given that it seems that a) RH/abshum is an important variable and b) that you had significant data loss issues, I feel a more quantitative assessment of the impact of these data substitutions is needed.

P11 – L2-13: Can you comment on the quality of the fit at Dawson vs CAMP in addition to the ideal model. The discussion is fairly qualitative. Also, the LM should be much better at extrapolating vs the ANN (which cannot extrapolate I think…not sure) – can you comment on this difference? Does LM perform better because it can extrapolate?
Section 3.2.2: If the calibration is immediately after deployment, I am not surprised that there wasn’t much of an effect. Do you anticipate there should be a significant time effect on such short time scales? What is the lifespan of the U-Pods?

I am confused by the introduction of discussion around figCxHy – should we expect this sensor to play an important role?

How long is “so long”? This is related to my comment on P11 – Section 3.2.2.

I am confused – did you switch to humidity measured by Picarro or omit humidity entirely? It is not clear to me what happened here.

I don’t really understand what is meant by “relative circumstances” – could you be more explicit about each of the case studies? Perhaps a table that outlines case study, with a one sentence description, and a column describing limitations would be more appropriate (and should be introduced at the beginning of the paper).

What is meant by “extrapolated significantly?” Can you be specific?

I find Table 1 almost impossible to follow. It is not very clear which sensor measures which pollutant, as the input codes are frequently indecipherable. I am honestly not sure what I am supposed to get out of this table.

I find it difficult to interpret this table. What do the black diamonds mean? What do you mean by “relative circumstances”??

There is way too much going on in this figure, it is almost difficult to look at. Is there a more streamlined way of presenting the findings that is less complicated? I feel there is valuable information in the Figure, but it’s hard to determine what that is, due to information overload. Ditto for Figure 9, though it isn’t as bad. Also you would do well to remind the reader what each variable represents.

TECHNICAL CORRECTIONS

P1 – L18: “in time than to…” vs. in time that to

P2 – L20: Delete word “Specifically”

P7 – L8: Rephrase “…showed successfully reduced over fitting”

Figure 1: Enhance figure caption to explicitly state that blue is training, pink is testing