Referee #2

This manuscript presents the results of an intercomparison between several pH sensors and laboratory measurements that were deployed in a challenging (in terms of biofouling) field environment. The experimental approach was very thorough and it seems likely that the dataset is excellent for doing the presented analysis. Overall it will be a useful contribution to the field. The concept of using a high-accuracy, low-resolution sensor to calibrate a low-accuracy, high-resolution one is interesting and does need more work in the context of specific sensor setups but it is not novel. There are a couple of limitations with the analysis. Primarily more evidence is needed to support the proposed approach, e.g. comparisons with other possible approaches and improvements to mitigate identified limitations, if the authors wish to present it as a guideline for the community to follow. My major points are in titled sections below, followed by minor comments and then technical corrections.

Instability of a 2-point regression

1. The issue of instability in a linear regression with 2 points (lines 331-333) is the major problem with the approach presented; it is mentioned briefly but not convincingly dealt with. Were this a study focused on reporting a particular observational dataset to interpret in some environmental context, it would probably be sufficient, because the uncertainties for the method used have been calculated and reported. But given the aim of this manuscript to provide guidelines for the research community on how to do this correction, it becomes essential here to do the extra work to see if the approach can be adapted to eliminate this issue, or at least to demonstrate that adaptations don't add any value. The authors have already collected the data needed to test these things. For example, linear fits could be made over 3+ consecutive LOC points to reduce the sensitivity to individual points. Doing linear fits also means there are sharp transitions between gradients as points are crossed; how does doing some smoothing fit (e.g. PCHIP, moving average) between the subsampled LOC points affect the quality of the corrected data?
   I recognise this requires some more work, but a paper proposing community guidelines should have done the due diligence to show that the guidelines are actually the best way of doing something. (While fully accepting that "the best way" will be a balance between complexity of the approach and accuracy of the results.) Alternatively, the relevant parts should be rephrased to indicate that this is a manuscript proposing and evaluating one potential way to do something, not claiming that this is a guideline that others should follow. Taking the latter choice would also reduce the impact of this study.

We thank the Referee for this comment and the opportunity to clarify our data fitting methodology. The linear fit used in our approach is not intended to perform a statistically rigorous regression in the traditional sense but rather functions as an enhanced offset correction. Specifically, instead of applying a simple constant offset (i.e., a correction only along the y-axis in pH space), the two-point line fit provides additional information about the local gradient in the LOC data within a user-defined correction window. This allows the correction to account for gradual drift in the high frequency optode measurements while maintaining the short-term variability captured by the optode.

Although LOC measurements were collected every 1–2 hours during this study, we demonstrate in Table 3 that a substantially reduced LOC sampling frequency (e.g., once every 24 hours) is sufficient to generate correction metrics that bring the drifting/offset high-frequency optode data *into close agreement with an independent pH sensor (SeapHOx). Increasing the LOC sampling frequency would provide additional fitting points, but this comes at the expense of increased reagent consumption, power demand, and waste generation. Importantly, our analysis indicates that these additional measurements yield only marginal improvements in performance.*

Nonetheless, to evaluate the Referee's suggestion more fully, we tested a range of alternative fitting approaches, including multi-point linear fits, moving-average smoothing of LOC data, offset-only corrections, and higher-order interpolation methods (e.g., PCHIP). Performance was evaluated using the mean bias, standard deviation and root mean square error (RMSE) of the differences between corrected optode measurements and the discrete co-samples measured in the laboratory. While some alternative approaches slightly reduced bias or produced comparable precision, none consistently improved both metrics simultaneously. In particular, more complex interpolation methods tended to increase systematic bias or introduce signs of overfitting. Across the tested methods, the two-point line fitting approach described in the manuscript provided the most balanced performance, maintaining low bias while also minimising the spread of residual differences relative to the reference co-sample dataset. For this reason, and given its conceptual simplicity and low data requirements, we consider it the most appropriate method for the correction approach presented here.

To highlight this, we have now expanded ESI section 3 to also report the results of different fitting methods. The changes are highlighted in blue to the ESI on ***pages S7-S8, lines 78-109*** *and in the main manuscript (using text below) on* ***page 11, lines 292-295****.*

"We evaluated several alternative correction methods e.g., offset correction, moving average smoothing, multi-point line fitting, etc. that can be seen in ESI 3. While these methods produced comparable results in some cases, the method described here provides the most favourable balance between correction performance, simplicity of implementation and a reduced LOC sampling frequency."

Terminology: accuracy

2. Throughout, especially e.g. Table 3, Fig. 7 and associated discussion: the mean offset is referred to as "accuracy" which is not always helpful terminology. This leads to e.g. describing an apparent accuracy "minimum" at some intermediate correction interval (line 326). An alternative, and to me more convincing, interpretation of Table 3 is that there is some constant offset (-0.018) e.g. due to an offset between LOC and co-samples that cannot be improved upon by increasing resolution, but because this is negative and the initial offset is positive (+0.111), you necessarily have to pass through zero to get from one to the other. But it doesn't mean that the results are "more accurate" at that intermediate point. Indeed if the initial offset happened to be more negative than this final constant value (or the constant value positive) then we the apparent "accuracy minimum" would probably not appear. In this case, the accuracy minimum is a fluke and not a reproducible feature that would necessarily be found in other datasets that had a different offset between the LOC and co-samples.

We thank the Referee for this viewpoint. We agree that considering the mean offset alone can lead to an apparent accuracy minimum when a positively biased dataset (e.g., optode) is corrected using a second dataset that itself exhibits a negative bias (e.g., LOC) relative to validation co-samples. In this situation the mean difference must pass through zero as the correction interval *changes (i.e., as the performance improves in our case), and we agree that this crossing point does not by itself imply that the measurements are most accurate at the zero point.*

For this reason, in the present work we do not interpret the minimum in the mean difference alone as indicating maximum accuracy. Instead, the sensor performance is evaluated using the combined statistics of the mean difference, the standard deviation of the residuals, and the RMSE all relative to the discrete co-sample measurements. The RMSE metric incorporates both systematic bias and random variability and therefore provides a more representative measure of total error. When these metrics are considered together (Table 3), the intermediate correction intervals correspond to reduced overall error rather than simply reflecting the zero-crossing of the mean offset.

However, to remove any potential ambiguity we have amended the below text in the Results & Discussion on ***pages 13, lines 339-345*** *to clarify and define explicitly how accuracy is quantified in this study:*

"Specifically, sensor performance in the present work is evaluated using three statistical metrics relative to the independent discrete co-sample reference measurements, i.e., the mean error ($\bar{x}$ ΔpH), the standard deviation of the error (1σ ±ΔpH), and the root mean square error (RMSE ΔpH). The mean error (i.e., $\bar{x}$) represents systematic bias (offset) relative to the reference measurements, while the standard deviation (1σ) reflects the spread of residual differences and therefore the measurement precision. The RMSE incorporates both the systematic bias and the random variability and therefore provides a combined estimate of total measurement error."

Lastly, we have also removed nearly all inferences to "accuracy" throughout the manuscript and instead have referred to the above metrics as performance (or difference) relative to discrete validation co-samples.

3. Not helping my interpretation of the above is that I found it unclear exactly how this "accuracy" error was calculated. I'm assuming it's corrected optode vs lab co-samples in the comment above. Please clarify or make more obviously explicit in the relevant parts of the discussion.

We have reviewed the manuscript and ensured that all sensor performances are reported as being relative to discrete co-sample reference data. Please refer to Author Response in above point (Referee #2, point 2) for further details.

4. Finally the points stated to represent these local minima in the text (2 days for x and 1 day for 1-sigma) are not the lowest local minima in the table (1 week for x and 2 hours for 1-sigma), so I don't follow why they were selected to be highlighted.

We thank the Referee for highlighting this point. The values referenced in the manuscript text and reported in the table were intended to illustrate local minima in the trend rather than the absolute minima. When the relationship is examined across the full range of sampling intervals (e.g., from high to low or vice versa), local minima are observed at approximately the 2-day and 1-day points for the average and standard deviation, respectively. Therefore, these points were highlighted in the text. We agree that the table also contains lower absolute minima, however, we have illustrated that their presence is likely an artefact of the fitting process itself. We also highlight that this is also influenced by the bias of the LOC sensor itself, which is driving the correction and is itself at a slightly negative bias relative to the discrete co-sample dataset. However, to clarify this we have added the below text on ***page 14, lines 368-372****:*

"Furthermore, the decreasing trend in $\bar{x}$ with higher correction frequencies will pass through zero difference offset (relative to discrete co-samples) as the optode data are being corrected by the LOC dataset which is itself at a negative offset relative to the reference dataset. Therefore, $\bar{x} = 0$ does not necessarily mean that at that point it is the most accurate point, as the sensor performance is evaluated across three collective metrics (i.e., $\bar{x}$, 1σ, and RMSE) relative to validation samples."

*Readers are also referred to **page 14, lines 360-368** which discusses this further.*

I think a big step towards a solution here would be to be more specific about what is meant in each statement and avoid using the somewhat ambiguous term "accuracy" when a more specifically meaningful alternative word is available.

We thank the Referee for this helpful suggestion. We agree that the term "accuracy" can be ambiguous if not explicitly defined. In the revised manuscript we have further clarified our terminology (see Author reply to Referee #2, point 2) and continue to report the specific statistical metrics used to evaluate performance. We believe this now provides a clearer and more quantitative description of sensor performance.

What is accurate?

5. The manuscript refers often to producing measurements that are "accurate" but does not define what this means – what constitutes "accurate" and whether something is accurate enough? It depends on the research question being asked of the data. Please could this be addressed briefly where relevant (e.g., Introduction and Conclusions, maybe relevant parts of R&D). Often this is done with reference to the GOA-ON "weather" and "climate" uncertainty targets (Newton et al., 2015), although other approaches are possible.

*We appreciate the Referee for pointing this out as it helps bring context to the larger aim of sensor development. We have added the following sentences to the Introduction (**page 2, lines 43-46**) and Conclusion (**page 19, lines 496-498**) sections, respectively to comment on this point:*

"Current observational targets for monitoring ocean acidification have been proposed by the Global Ocean Acidification Observing Network (GOA-ON) framework, which defines approximate uncertainty thresholds of 0.02 for weather quality data and 0.003 for climate quality data (Newton et al., 2015). These benchmarks provide a useful reference point for evaluating sensor performance."

and

"The corrected optode measurements achieved RMSE values on the order of ~0.02 relative to discrete co-samples, thus meeting the GOA-ON weather-quality observational target, which is sufficient for resolving short-term variability in coastal carbonate chemistry while minimising reagent consumption and operational complexity."

Manufacturer claims

6. Manufacturer accuracy claims are presented in the Introduction and Methods. They are sometimes alluded to in the R&D but it might be useful to have a short paragraph or section that directly addresses if these accuracy claims could indeed be achieved by the various sensors in the tests here.

*We have added the below text on **page 17, lines 441-444** to mention this point:*

"It is worth noting that the SeapHOx pH sensor performance in the field from our work was well within the manufacturer reported limits of ± 0.050. Furthermore, the field measured performance of the LOC pH sensor was close to achieving its manufacturer reported performance of <0.009. Therefore, this 6-month shallow water field study has demonstrated that it is feasible to achieve these manufacturer-reported performances."

Minor comments

7. 158 PyroScience have several different sensor caps available, with different pK values and returning results on different pH scales; please specify what was used.

*This has now been updated with the following text **page 6, line 171** to clarify this:*

"…a fresh optode pH cap (PHCAP-PK8T-SUB, PyroScience GmbH) was soaked…"

8. 159 The PyroScience optode software also has the option to add a third calibration point of a buffer (e.g., tris) within the measuring range to improve accuracy. Could the authors comment on if and how excluding this step may have affected their results and conclusions?

This is a fair point raised by the Referee. The PyroScience optical pH sensors (e.g., AquapHOx-L-pH as used in the present study) are typically calibrated using a 1- or 2-point calibration in (non-seawater) acidic and basic pH buffers as recommended and supplied from the manufacturer. This is commonly pH 2 to represent the fully protonated form and pH 11 for the fully deprotonated form. The sensor response from these two buffers represents the plateau regions in a response curve, which is a sigmoidal shape, and intentionally occurs outside of the operable range of the sensor which is typically pH 7-9. An optional third calibration point in a known seawater sample near the pH of the anticipated deployment pH may be applied to adjust the absolute offset within the measurement range. Recent work from Wirth et al. (2024) has shown that including this third point can further reduce offsets and improve absolute accuracy in laboratory settings. However, the two-point buffer calibration defines the sensor response function, and omission of the third point largely affects the absolute bias rather than precision of the measurements. Furthermore, our findings do confirm what Wirth et al. (2024) have proposed in their Comments and Recommendations section: 'It may be possible to correct for drift with multiple validation samples taken throughout the deployment. Validation samples every 3–4 weeks and at the end of the deployment are recommended to maintain the weather objective quality of 0.02 pH.'

*As such we have added the below text on **page 10, lines 263-266** to clarify this point:*

"Recent work has shown that using a three-point buffer calibration (in lieu of a two-point buffer calibration) can be beneficial in reducing signal offsets and improving absolute accuracy for optode-based pH sensors in laboratory settings, but for field deployments discrete co-samples are required to maintain this level of sensor performance (Wirth et al., 2024)."

9. 167 Was it really possible to always get the sample from the harbour, into the lab, in an optical cell, equilibrated to 20 °C, injected with mCP and measured in under 5 minutes? Impressive if so, but the relevant time to report would be the actual moment of measurement, not just the moment that the sample handling in the lab began – please check & confirm.

The lab spectrophotometer and setup were always prepped prior to discrete co-sampling, and the laboratory is directly accessible to the quayside at NOC. Effectively all we had to do was inject *the sample and allow it to equilibrate. However, as the temperature equilibration employed was indeed 5 minutes, we have increased the total time to measurement indicated to be ≤10 minutes. The measurement jacketed glass cell is connected to a recirculation water bath which affords relatively fast temperature equilibration of the sample. Furthermore, the relevant time to report is not the point of measurement in the lab but rather the point of sample collection in the harbour. Each discrete co-sample represents a water mass collected at a specific point in time, which allows us to have a meaningful comparison to sensor measurements made at that same water mass and point in time. As such, we have also updated the text* ***page 6, lines 181-183*** *to clarify which time is used for the co-sample:*

"...the time between co-sampling and lab measurements was ≤10 minutes (i.e., the harbour was directly accessible from the NOC), and the time of the discrete sample was noted upon collection (i.e., the collection time was used for comparison to sensor data)."

10. 202 Does a "battery failure" refer to the battery running out of charge, or something else more dramatic? Please clarify.

We have amended the text on ***page 7, line 218-219*** *to clarify this point:*

"...the optode sensor's battery was never depleted even during periods of..."

Technical corrections

pH is dimensionless; please remove references to "pH units" throughout.

We have changed the wording throughout the manuscript and ESI to not report any values as "pH units".

42 Grammar: change "and until recently, was" to e.g. ("which until recently was").

Done.

43 If pH is calculated from DIC, TA or fCO_2 then it is not a "measurement of pH", please rephrase.

This wording has been rephrased as suggested.

55 Provide a location for the NOC.

Done.

58 The ocean goes deeper than 6000 m, please rephrase "full ocean depth".

This wording has been rephrased as suggested.

67 Not clear specifically what "This" refers to.

We have clarified this as "This technology has..."

77 Grammar: either "version" => "versions" and "are a cylinder" => "are cylinders", or "are" => "is". Also probably "sensors" => "sensor".

This wording has been rephrased as suggested.

108 Presumably "The NOC" should be "The harbour", or make it clear that the harbour is at the NOC in the previous sentence.

This wording has been rephrased as suggested.

155 “calibration-less” is a bit awkward; “calibration-free”?

This wording has been rephrased as suggested.

Section 3.2 Several aspects of results that should be in past tense are written in the present. Also applies to other parts of the Results & Discussion. Some of these I have noted as technical corrections here but my list will be incomplete so please check through carefully.

We have reviewed the Results & Discussion section to ensure that past tense was used.

290 Brackets around the two *b* terms at the end of Eq. (5) are unnecessary.

Done.

Table 3 Should *n* be the same in every row? I would have guessed it is how many LOC points were used in the calibration, which would be smaller for the longer correction intervals. If not, then please rewrite the caption to make it clearer what *n* means. If it is supposed to be the same then it doesn’t need to be a column in the table. Also, please mention in the caption which sensor data are being shown and what they are being compared to.

*This has now been updated as suggested. The caption for Table 3 (**page 13, line 355**) has been updated as shown below to clarify this as well:*

“Performance of the optode sensor, relative to discrete lab-validated co-samples, is reported as the mean sensor error ($\bar{x}$ ΔpH), standard deviation of the error (1σ ±ΔpH), and the root mean square error (RMSE ΔpH). The number of samples (n) for determining these metrics were n = 44 for all correction intervals.”

343 “30/80/2023” => “30/08/2023”.

Done.

359 “are reporting” => “were reporting” or “reported”.

This wording has been rephrased as suggested.

376 “are tracking” => “were tracking”.

This wording has been rephrased as suggested.