Technical note: Enhancement of float-pH data quality control methods: A study case in the subpolar northwest Atlantic Ocean

In this response, the "original manuscript" refers to the first submitted manuscript that has been evaluated by the reviewer and the "revised manuscript" refers to the manuscript that has been modified according to the reviewer´s comments. Comments from the reviewer are pasted below in black font; our point-by-point responses immediately follow in blue font. Blue italic sentences are those that have been modified / added in the revised manuscript. When indicated, line numbers refer to the new version of the manuscript (Latex version). We have also submitted a Word document with the "track changes" function activated, which should help the reviewer in figuring out the changes.

Responses to REVIEWER 1 - Brendan CARTER

Wimart-Rousseau et al. have put together an interesting new data set and float-cruise comparison. They also review float adjustment algorithms, adjustment reference depths, and choices regarding how and when adjustments are updated. Their findings show significant offsets between their floats and discrete samples collected from a nearby cruise. They arrive at a plausible set of conclusions and suggest several useful measures. The resulting paper could be a useful contribution to the literature, but I believe it should nevertheless be returned to the authors for revision for the following reasons:

We thank the reviewer for his constructive and helpful comments about our paper, his positive opinion concerning the interest of this manuscript for the community and the time he spent reading and reviewing our manuscript. We most appreciate his relevant comment about the recent global pH algorithm update which has been added in the revised manuscript. Comments and suggestions made by the referee on text and/or figure/table edits are discussed below and addressed in the revised version of our manuscript.

1. This is too long to be a technical note, which, to my read on the journal policies, is limited to "a few pages." This is 21 pages (before references) in the review format. Also, too much of the discussion is qualitative rather than quantitative and did not "feel" technical. This paper needs to be re-worked as a full length paper or shortened and focused.

Following the referee's recommendations and comments listed hereafter in this review, we modified several paragraphs of the manuscript by removing sentences (e.g. Section 3.3) and clarifying some others. Figure 6 has been reduced and Figure 8 removed to simplify the reading. See also our point-by-point responses indicating the changes done. Overall, we accepted all the suggestions proposed by the referee. We believe that these modifications improve the readability of the manuscript.

This paper aims to describe and discuss the main limitations and uncertainties associated with the current float-pH data correction procedure as well as to propose a way forward to enhance the float-pH quality control process. According to the Biogeosciences guideline, we believe that our manuscript is relevant for publication as a technical note as it presents "novel aspects of experimental and theoretical

methods and techniques which are relevant for scientific investigations within the journal scope". For this type of manuscript, no clear indications about the length are given on the webpage, except a "few pages". In its revised form, the manuscript is now 20 pages. As a comparison, a technical note of 18 pages (before references) has been published in 2021 (Canning et al., 2021[1]). We thus believe that the revised manuscript could be published as a technical note too.

2. The discussion about adjustment update methodologies was ultimately unconvincing. I feel it could be reduced to a quick comment that it would be useful to have a community consensus regarding how this is done, which I don't feel needs much justification. Alternatively, the authors could rework and/or expand upon their rationale for their preferred method in a full paper.

In the original manuscript, a discussion about the impact of the correction method used to correct float-pH data was presented in Section 3.1.3. and Figure 5 aimed to illustrate our presentation. The purpose of this section was to discuss the noticeable step-like changes observed with the current correction procedure (i.e., the SAGE method) and to find the best way to represent the smooth sensor drift over time, as observed when looking at the pH time-series recorded at the parking depth (previous Figure A1 in the Supplementary Material). Indeed, in comparison with the pattern of the cycle-by-cycle correction, the high pH changes of ca. 0.01 pH units observed between linear drift phases with the SAGE method appear to be unrealistic. In our view, the sensor rather shows undulations in response with smooth and less smooth phases. This statement is somehow confirmed by the pH sensor behavior when the float drifts at its parking depth. In consequence, we believe that an adaptation of the current correction procedure could be done to better maximize the smoothness of the corrections and to avoid introducing artificial jumps. This presentation could be thus useful for the community and discussion concerning the current procedure could arise from it.

However, we agree that explanations were missing in the original manuscript and that the original idea to put Figure A1 in the Appendix was not relevant as it is critical for our argument. Following the recommendation of the reviewer, Section 3.1.3 has been modified, explanations have been added, and Figure 5 re-drawn: now 4 panels representing differences between raw and corrected float-pH data following the SAGE method (panel A), the GEOMAR methods (panel B), pH data measured at the parking depth (panel C) and pH data measured at the parking depth minus reference (CANYON-B pH data, panel D) are presented. Figure A1 has been modified and is now included in the revised manuscript.

3. The quantitative aspects seemed, in places, potentially incorrect. Other float-to-pCO2 comparisons have not shown as large of offsets at the surface as are found in this study (see discussion in https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022AV000722, also confirmed by some unpublished community studies, and compare to the pCO2 offset implied by the pH offset observed herein of ~50 uatm). Worryingly, the correlation shown herein of the upper-ocean discrete-to-float pH offset to the temperature is comparable in magnitude to the sensitivity of pH itself to temperature, so I would urge a double check on the pH temperature adjustments made in this comparison. If they

[1] Canning, A., Fietzek, P., Rehder, G. and Körtzinger, A. (2021). Technical note: Seamless gas measurements across the land–ocean aquatic continuum – corrections and evaluation of sensor data for CO_2, CH_4 and O_2 from field deployments in contrasting environments. Biogeosciences, 18, 1351–1373. https://doi.org/10.5194/bg-18-1351-2021

were performed correctly, then the calibration of the pH probe sensitivity to temperature seems to be nearly 100% in error (which might be the author's point, though it is odd then that this is not seen elsewhere). The author's point stands that the North Atlantic is a worst case scenario in many respects is accurate, but that it is not an especially problematic location from the perspective of vertical temperature gradients. The authors' comment regarding the differences between the algorithm estimates is supported by a listed standard deviation of the algorithm differences of 0.051… for reasons given in line-by-line comments I believe this represents an order of magnitude error in the listed value or an indication that the standard deviation is not the appropriate statistic in this case. If my suspicions here are correct, and the true standard deviation is <0.01, as it appears from the histogram, then this is roughly consistent with the published algorithm uncertainties at depth when comparing two (only arguably independent) algorithm estimates. This then would give us no reason to doubt the earlier Williams et al. uncertainty propagation.

In the open ocean, we agree with the referee and the current literature that float-derived pCO_2 (pH, TA) estimates have a theoretical uncertainty of ~11 µatm. We also thank the referee for suggesting the paper written by Bushinsky and Cerovčki (2023). Notwithstanding, in the studied area, two crossovers comparisons have been performed using two independent datasets and on two different floats. Despite the limited number of floats and crossovers associated with this study providing only one showcase, and although the actual pH values may be slightly different due to the regional variability, the preliminary results point at unacceptably high and almost identical biases in surface pH values from the 2 floats which have been corrected in the exact same way. It calls for an additional independent pH reference in the surface ocean. Indeed, in both cases, pH offsets are positively correlated with temperature, being the smallest at the temperature of the reference depth. In agreement with the referee, and as it is stated in the manuscript, we argue that it points towards an imperfect representation of the temperature and/or pressure dependencies of the pH sensor (Page 16). Our findings show that corrected float-pH data may be biased by several hundredths of a pH unit near the surface in this deep convection region, suggesting that an adjustment of the reference depth might be done in such oceanic areas. Finally, it could also be pointed out that a current publication from Gattuso et al. (2023)[2] presents even higher pH uncertainties for two SeaBird pH sensors deployed in the high-Arctic fjord (Kongsfjorden, Svalbard).

Concerning the pH temperature adjustments: Discrete pH measurements used for the comparison have been converted to in situ temperature and pressure for this study. It has been done using the CO2SYS software with measured pH data and TA values as input variables. Converted discrete pH data have been re-checked for this review and no conversion mistake has been found.

Concerning the standard deviation of the algorithm differences: As stated in our answer to one specific referee's comment, the standard deviation value given in the manuscript is the right one. This value is large compared to the 5^{th}/95^{th} percentiles because of very wide tails in the distribution. Indeed, this wide tail distribution and important standard deviation value are located, for some parts, in the Black Sea causing quite a consistent (and high) difference of ca. -1.6 pH units. In our case, we agree with the referee that this noticeable standard deviation value implies/indicates that the distribution is not Gaussian and raises the question of the utility of this metric considering that percentiles give a more

[2] Gattuso, J.-P., Alliouane, S., and Fischer, P. (2023). High-frequency, year-round time series of the carbonate chemistry in a high-Arctic fjord (Svalbard). Earth System Science Data, Earth Syst. Science Data, 15, 2809–2825. https://doi.org/10.5194/essd-15-2809-2023

robust accuracy assessment metric than standard deviation. In this study, we still wanted to present this value as it is a metric commonly used but also because it reflects one of the main messages of the manuscript: differences are much larger than we would expect from a comparison.

4. A somewhat recent global pH algorithm update (ESPER-LIR and ESPER-NN) were omitted from the analysis. Presumably this is because they are not yet included in SAGE but including them should nevertheless be helpful for this discussion because the updated algorithms take a different approach to several of the algorithm limitations discussed in this paper and are slated to become incorporated into SAGE (to my understanding). For the algorithms that were used, important information was not provided (that I saw) regarding the use of an optional adjustment for ocean acidification. Also, it is arguably appropriate to keep the "spectrophotometric" pH adjustment (as the authors chose to do), but it should be pointed out that this limits the comparability of the two algorithms that were considered (unless such an adjustment was applied to CANYON-B estimates?). As a note here, we (Ocean Carbonate System Intercomparison Forum) are drafting a paper that is taking a step back from the recommendation that this adjustment should always be applied given the finding that the apparent slope is not present in a subset of cruise measurements made with purified dye. The recommendation is to instead treat uncertainty in the comparability of pH and DIC + TA as a component of uncertainty in whatever calculation is being attempted... this recommendation would support the overall thesis of this paper that we need to be cautious about the uncertainty that we claim we can achieve from float-based pCO2. A related comment that this paper uses the term "correction" when in many cases I believe "adjustment" should be used. The distinction I draw is when we believe we understand the reason for the apparent offset it is appropriate to call it a correction, and I would use adjustment in all other cases.

We agree with the referee that a comparison with the new ESPERs methods was missing in the original manuscript. Following the comments and suggestions of the reviewer, the revised manuscript now includes a presentation of pH data estimates with the ESPERs routines. Figures and tables have been subsequently adjusted. In particular, Sections 3.1.1. and 3.1.2. have been modified substantially. Precisions about the optional ocean acidification adjustment have been added too, both in the main manuscript and in figure legends.

Concerning the spectrophotometric pH adjustment: As stated by the referee and precise in the manuscript, this adjustment has been kept in this study. Thus, CANYON-B pH data have been adjusted to align estimates with spectrophotometric pH measurements made using purified dye. As the LIR-pH training dataset consists of values either measured or calculated but adjusted using the same purified-dye adjustment (adjustment 3; Carter et al., 2018), we consider that the two algorithms' results are comparable. In the revised manuscript, a sentence has been added to more clearly state that this adjustment has been used.

Finally, we have paid more attention to the nomenclature and the meaning of the word adjustment in the revised manuscript.

5. Some of the figures are missing information, and the writing is difficult to understand in a few places (but excellent in many other parts). Some of the notation is inconsistent (see line by line comments). There is a comparison to “weather” and “climate” quality

data thresholds, which I argue below is inappropriate. The discrete pH measurement uncertainty is unrealistically low.

In the original manuscript, labels for Figures 4C and 6D were incomplete and explanations about some numbers presented were missing (e.g. Figure 5). Following comments by the reviewer, several figures have been modified in the revised manuscript and explanations added in the legends. All the writing comments and modifications proposed by the referee have been adressed.

Concerning the "weather" and "climate" goals comparison: We agree with the referee that climate and weather goals are related to precision and that the presentation done in the original manuscript could be confused as this discussion occurred after the comparison between corrected float-pH data and in situ discrete measurements (Section 3.2.). On the other hand, Section 3.1. presents a detailed description of the dispersion of the corrected float-pH data in response to the reference pressure choice, the reference depth selection as well as the choice of the method used to correct float cycles. In our manuscript, we believe that, whereas Section 3.2. aims to assess the accuracy of the correction procedure and to discuss the errors, uncertainties presented in Section 3.1. are relevant and allow comparison against the GOA-ON goals. In the revised manuscript, Section 3.3 has been shortened and clarified.

Concerning the pH uncertainty: During the MSMS94 cruise, samples were poisoned onboard following the current standard procedure (SOP) and analyzed at GEOMAR. pH measurements were tested regularly against CRM reference samples to check the accuracy of our measurements. While CRMs are certified only for AT and DIC, pH measurements were also performed for each bag by Dickson's lab and made available for us. The resulting uncertainty in pH measurements for discrete water samples was ± 0.002 pH units. In the studied area, as all the best practice recommendations have been followed, we have no reasons to doubt the resulting pH uncertainty.

Ultimately I was left uncertain of what to make of the results. This is perhaps a useful outcome for a paper that is arguing that we need to be less optimistic about the quality of the data generated by certain approaches, but I feel like the authors should take advantage of a revision to address the issues noted above and below, should double check for potential errors in the presentation and the analysis (particularly the pH- temperature and pressure conversions), should more rigorously compare how the algorithm estimate variability compares to the expected variability in a statistical sense, and should shorten the paper and distill the main ideas (particularly if it is to be kept as a technical note). That said, I tend to agree with all of the conclusions made by the authors, and their presentation did seem to support the conclusions. The subject matter is important and the statements in the conclusions are worth making to the community, so I hope the authors resubmit this paper.

We thank the referee for his insightful and stimulating recommendations regarding our paper. In addressing his concerns, we think that the revised manuscript has been improved substantially. Overall, we have accepted all the suggestions proposed by the referee, significantly modifying the main text and tables, and adding new information in the discussion section.

Line-by-line comments

1: "this" refers to a subject that has not yet been defined

The sentence has been modified as follows: "*Since a pH sensor has become available that is suitable for demanding autonomous measurement platforms, [...]*" (L.1).

6: "decipher punctual events" is awkward phrasing. Suggest "Measure the impacts of short-term events"

The sentence has been modified as suggested by the referee. (L.6).

8: This is a matter of personal preference so please feel free to ignore this comment, but I feel this sentence is written backwards. It is shorter and easier to read when written as, e.g., "Quality control is needed to correct sensor offsets or drifts." A recommendation a past advisor gave me is to establish the subject and verb of a sentence early in the sentence so the readers know what the sentence is about. The subject and verb are among the last words in this sentence, and this is a common element of many of the most difficult sentences in this paper.

We thank the referee for this comment and his suggestion. The sentence has been clarified accordingly: "*In consequence, a consistent and rigorous quality-control procedure has been established to correct sensor offsets or drifts as the interpretation of changes depends on accurate data*." (L.7).

9: Again, this feels backwards

This sentence has been modified as suggested by the referee in his comment #8 (L.7).

12: LIRPH should be LIPHR or LIR-pH

The acronym LIRPH has been replaced in the manuscript as well as on all the figures by LIR-pH.

32: This sentence is correct, but it is probably better to say "ocean acidity will increase" because "alkalinity" will not decrease by this mechanism.

This sentence has been modified as suggested: "*Depending on emission scenarios, ocean acidity will increase with a projected pH decline ranging from 0.16 to 0.44 pH units by 2100 [...].*" (L.32)

49: The majority of what we know about surface ocean pCO2 arguably comes from ships of opportunity, which deserve greater mention in this discussion.

The description of the SOOP network previously done in Section 2.2. of the original manuscript has been shortened and explanations about the SOOP program and its interest added in the revised manuscript.

L.49: "*Since the 1990s, the Ship Of Opportunity Program (SOOP; Goni et al., 2010) aims to obtain data from autonomous instrumentation installed on volunteer merchant ships regularly crossing certain areas. This network contributes to building sustained carbon observing datasets and complements the limited capacity of classical observational strategies as the standard-SOOP framework features, at least, routine pCO_2 observations (e.g. Lüger et al., 2004). In the Atlantic Ocean, parts of the SOOP network are operated in the European Research Infrastructure 'Integrated Carbon Observation System' (ICOS) and the 'Surface Ocean CO_2 Reference Observing Network' (SOCONET).*"

58: Maurer et al. show large pH sensor adjustments, suggesting the pre-deployment calibration is not a major factor since it is seldom used… except perhaps for characterizing the dependence on, e.g., pressure.

We agree with the referee that the pre-deployment calibration is not a major factor with regard to the recent literature results. Nevertheless, these sentences aimed to present the regular and general procedure to follow to obtain reliable and consistent data. Indeed, each sensor system should be processed following a calibration scheme ensuring and demonstrating unequivocally accurate pH determinations.

71: Missing ESPER-NN and ESPER-LIR (which are updates to LIR).

This precision has been added in the revised manuscript: "*Recently, two Empirical Seawater Property Estimation Routines (ESPER; Carter et al., 2021) have been included in SAGE as reference methods. The ESPER-NN method generates estimates from neural networks while the ESPER-LIR routine is based on locally interpolated regressions.*" (L. 77)

98: the "claimed" pH accuracy…

The sentence has been modified as suggested. (L. 107)

104: worth pointing out that this is a single point calibration that doesn't reflect the conditions often found at the reference adjustment depth for pH.

We agree with the reviewer that the uncertainty given was the most optimistic one. We toned down this assessment in the revised manuscript.

L. 113: "*A stringent referencing and adjustment process for the oxygen can yield accuracies around 1.5 µmol kg^{-1} (Bittig et al., 2018a), although depending on the details of the optode calibration, handling, and usage scenario, the accuracy of O_2 measurements can vary considerably.*"

120: the GLODAP data are not directly used in the float pH correction procedure in most instances.

This sentence has been shortened: "*In situ pH data measured from water samples are generally considered as reference data for float-based observations and are useful tools to independently estimate pH data accuracy and, if needed, apply additional adjustments.*" (L. 132).

135: this is not a CRM for pH, and there is no certified value

We agree that this is not certified reference material (CRM) for pH. However, our logic is that the CRMs from the Dickson lab are known to be stable for carbon parameters (and thus for pH). We believe that the pH value from the Dickson lab is accurately determined even though it is not a real certification. The uncertainty in the reported pH value for the Dickson CRM will be an order of magnitude less than the difference we observed between the pH data from the SOOP line and the floats. Nevertheless, by moving the sentence line 152 in the original manuscript right after this statement in the revised manuscript (L. 152), we believe that it clarifies the situation. The certified value (7.8417 ± 0.0014 at 25°C) has been added too (L. 153).

137: on what basis is it assigned this very low pH uncertainty given that the best methods are typically assigned an uncertainty of 0.01-0.007? The uncertainties in the calculations from DIC and TA are also quite large, and the uncertainties from the conversion to in situ conditions are thought to be large and poorly known. This claim might be justifiable if it is expressed in terms of reproducibility, but uncertainty has a more expansive definition than reproducibility.

While CRMs are certified only for AT and DIC, pH measurements were also performed for each bag by Dickson's lab and made available for us. The comparison done against these certified materials leads us to conclude that the resulting reproducibility was ± 0.002 pH units. However, we agree that the term uncertainty might be misleading. The sentence is changed to reproducibility. (L.153)

152: This caveat should go earlier.

This sentence has been moved earlier in the revised manuscript and is now line 152.

200: note

The sentence has been modified.

230: and in

The sentence has been modified.

236: measure pH

The term pH has been added.

250: Figure 4 does not obviously support this statement without further explanation

We agree with the reviewer that the reference to Figure 4 wasn't sufficiently supported in the original manuscript. In the revised version of the manuscript, Figure 4 has been modified and is now better introduced and explained: "*Figure 4 exhibits spatial distributions of estimated pH data at the classical reference 1500m depth level using either LIR-pH (with the OA adjustment), CANYON-B, ESPER-LIR, or ESPER-NN and differences between the estimated datasets with uncertainty between reference algorithms in the order of 0.015 pH units in the SNWA area.*" (L.273).

256: what does the "respectively" refer to? Are these different depths? Estimate methods?

The sentence has been clarified: *"In addition, using the SOCCOM array, Maurer et al. (2021) calculated CANYON-B and LIR-pH pH estimates and observed a larger uncertainty toward the surface compared to 1500 m with mean differences (CANYON-B minus LIR-pH pH data) of -0.025 and 0.001 pH units near the surface and at the 1500 m depth level, respectively.*" (L.281).

Figure 4: You should indicate whether LIR-PH estimates are using the OA adjustment option. I believe SAGE usually omits this adjustment, which might explain why LIR-PH is high relative to CANYON-B in the North Atlantic and low in the North Pacific. That said, as you correctly point out, LIR-pH uses an globally uniform OA adjustment that varies only by density, whereas CANYON-B uses an empirical local fit (that, I argue elsewhere, may erroneously project interannual variability forward and backward in time). Carter et al. 2021

attempt to resolve these issues, and this technical note would be more useful if it also included a comparison with the ESPER routine estimates. These routines are slated to become incorporated into SAGE. (Note, you'll probably still find large differences between ESPER pH estimates and LIR-pH estimates, particularly in the North Atlantic, which appear to be attributable to the omission of depth as a predictor variable from ESPER-pH… i.e. the values from ESPER-LIR are much more similar to LIR-PH (and ESPER-NN) when depth is included as a predictor... future updates to ESPER-LIR will likely have depth as an optional predictor to minimize the discontinuity that we'll see if and when the transition from LIPHR to ESPER-LIR is implemented).

We thank the referee for this remark. We agree with the referee that the OA adjustment is omitted by default in SAGE and we have decided to keep this option off in this study with regard to the limitations associated with the OA adjustment (i.e., LIR-pH assumes fixed OA rates over time). This information has been added in the revised manuscript (L.233, Fig.3 label). Nevertheless, in Figure 4, the OA adjustment option was used in order to clarify the figure and to not overall mean bias because of a few oceanic regions. In the revised manuscript, this precision has been added both in the main text and the caption (L.274, Fig.4 label). The spatial distribution derived without the OA adjustment (Fig. 1 below) presents a higher mean bias of 0.002 pH units (against -0.001 pH units) which is caused, at least partially, by the Black Sea causing quite a consistent (and high) difference of ca. -1.6 pH units. By removing this sea during the simulation, it appears that the std decreases considerably. High differences can also be explained by the both enclosed and undersampled Mediterrean Sea and Baffin Bay.

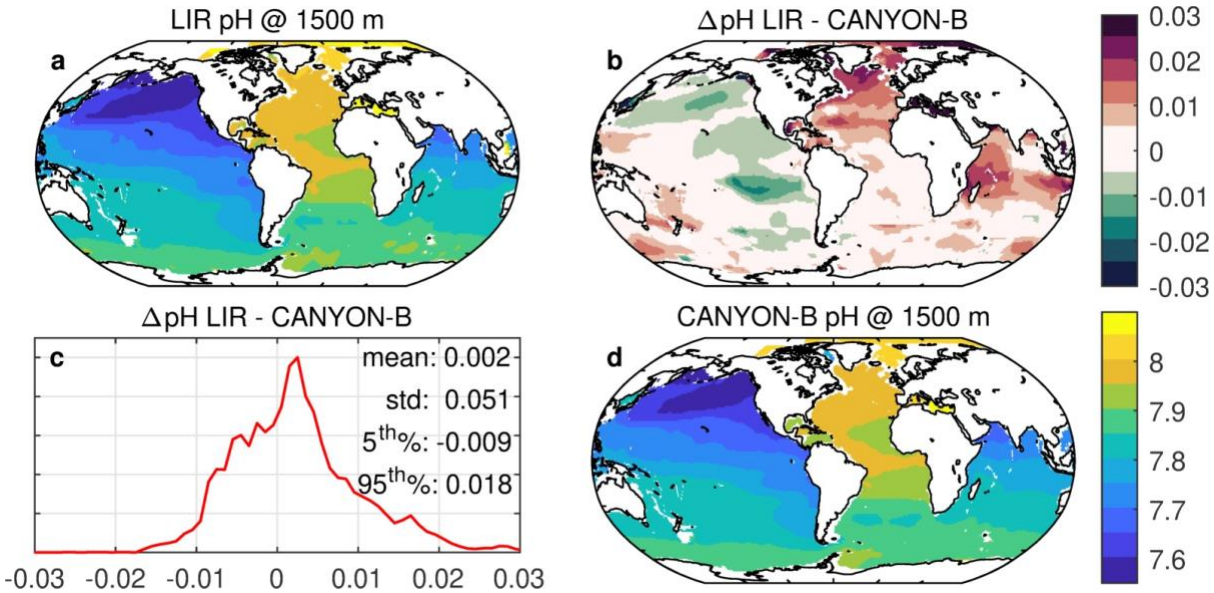

Figure 1 (not in the revised manuscript). Spatial distributions of estimated pH data at 1500 m using different reference models: LIR-pH (without the OA adjustment) (A) and CANYON-B (D). The map of the spatial difference between the two estimated pH datasets is presented in panel (B). Panel (C) shows the bias ΔpH distribution (with statistics). The upper colorbar indicates the difference between estimated pH data using the two models and the lower colorbar gives the pH values.

We also agree with the referee that a comparison with the ESPER methods was missing in the original manuscript. Following the comments and suggestions of the reviewer, Figures 3 and 4 have been updated in the revised manuscript and now include a presentation of pH data estimates with the ESPER routines.

L.284: "*The new ESPERs methods attempt to resolve the issues encountered with existing routines (especially the OA estimate) by expanding their functionality and being trained on a larger data product. In comparison with the LIR-pH estimates, large differences are observed in the SNWA region and might be attributable to the OA adjustment as well as the omission of depth as a predictor variable from ESPER-LIR (Carter et al., 2021). Updated global algorithms (i.e., ESPERs) show comparable estimates in the SNWA area with ESPER-LIR pH estimates slightly higher than pH data estimated with CANYON-B or ESPER-NN. In the dynamic and strongly human-impacted studied region, the lack of coordinate information as a predictor variable in the ESPER-LIR routine could also be argued as an explanation of the observed differences. However, according to Carter et al. (2021), regional assessment statistics obtained in the Northern Atlantic indicate almost similar biases for both the ESPERs and the CANYON-B methods, with a better RMSE statistic for CANYON-B.*"

Figure 4c: The y axis is not labeled. Also, there may be a missing 0 in the std value?: It is only possible for the STD to exceed the 5th and 95th percentiles if there are a small number of extreme outliers (that should have likely been omitted). This is among the most important numbers in this manuscript, to my thinking. If it is much smaller, then I'd ask whether the std is indeed much larger than we'd expect from a comparison of two algorithm reference adjustments?

We agree with the referee that the y-axis wasn't labeled for Figure 4c in the original manuscript. The original thinking behind this omission was related to the kind of plot presented directly: to our thinking, the relative distribution of a histogram is what is relevant to check, and the absolute number reported as frequency distribution or count is depending on the dataset resolution. In the revised manuscript, Figure 4 has been re-drawn and the y-axis label "Frequency" has been added (Page 13).

The std value given is the right one: The std is as large compared to the 5^{th}/95^{th} percentiles because of very wide tails in the distribution. In our case, this noticeable std value implies/indicates that the distribution is not Gaussian and raises the question of the utility of this metric considering that percentiles give a more robust accuracy assessment metric than std. In this study, we wanted to let this std value on Figure 4c as it is a metric commonly used but also because it reflects one of the main messages of the manuscript: differences are much larger than we would expect from a comparison. Nevertheless, as stated in the former answer to the referee, data outside the 5^{th}/95^{th} percentile, and explaining this wide tail distribution and important std value, are located, for some parts, in the Black Sea causing quite a consistent (and high) difference of ca. -1.6 pH units. To clarify this figure and not lead to misinterpretation, we have decided to remove this area in Figure 4 of the revised manuscript. The std is then reduced, even if percentiles indicate that there can be quite some deviations, especially due to the Mediterranean Sea and the Baffin Bay.

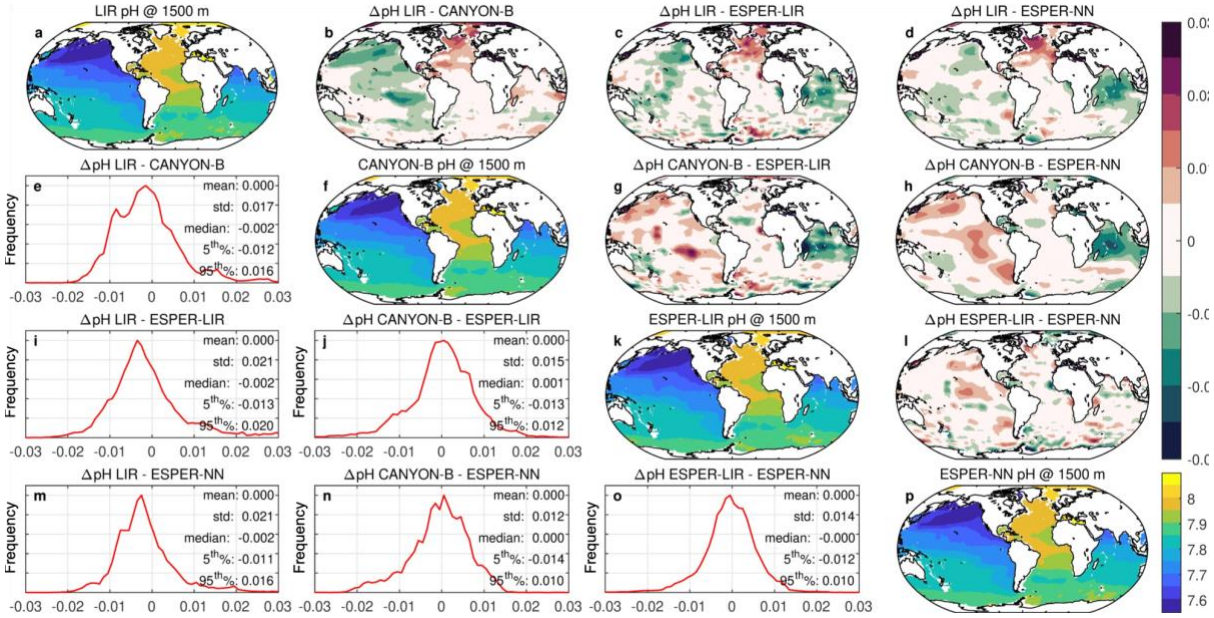

Figure 2 (Fig. 4 in the revised manuscript). Spatial distributions of estimated pH data at the classical reference depth 1500 m using different reference models: LIR-pH (with the OA adjustment) (a), CANYON-B (f), ESPER-LIR (k) and ESPER-NN (p). Maps of the spatial difference between the estimated pH datasets are presented in panel (b-d, g-h and i). Panel (e, i-j and m-o) shows the bias ΔpH distribution (with statistics). The upper colorbar indicates the difference between estimated pH data using the different models and the lower colorbar gives the pH values. For clarity, pH data estimated for the Black Sea have been removed for this simulation as they were outside the 5th/95th percentile and they caused a noticeable increase of the standard deviation (std).

261: from the way the

The sentence has been modified. (L.296)

264: is the noise uniform with depth? If not, then it should be handled by averaging the adjustment across a depth range. If it is, then should it be removed by adjusting every profile independently?

The pH sensor can generate occasional and non-uniform spikes due to electrical noise and despiking is appropriate. According to Johnson et al (2022)[3], the default Argo spike test in core variables does not work for pH because of the strong vertical gradient dependency of this test, making it regionally dependent. On the other hand, the spike test recommended for chlorophyll (Schmechtig et al., 2014)[4] is more appropriate for pH. Thus, a data point is considered a spike and marked with quality flag 4 (data bad) if the test value is > 0.04 pH. If float-pH data have a good QC, then they are adjusted and corrected following the current procedure.

266: "corrected for" should be "removed"

The sentence has been modified.

[3] Johnson, K., Maurer, T. and Plant, J. (2023). BGC-Argo quality control manual for pH, in preparation.
[4] Schmechtig, C., Claustre, H., Poteau, A. and D'Ortenzio, F. (2014) Bio-Argo quality control manual for Chlorophyll-A concentration, Version 1, December 17th 2014. IFREMER, 13pp. https://doi.org/10.13155/35385.

Figure 5: The use of 2 y axes is very confusing here and defeats the purpose of being able to compare the adjustments to one another, though the goal of increasing visibility makes sense. It would probably be better to use a single y axis, which would allow the plot to focus in on the 0.04 to 0.06 range. The greater vertical resolution should also improve visibility. If the different methods are difficult to distinguish on this unified scale, then that suggests that the methods aren't different enough to worry about on this scale. What are the numbers in the upper right? (it's not hard to guess, but it is better if it is spelled out)

We thank the referee for this comment and his suggestion. Considering this remark as well as the one about Figure A1, Figure 5 has been updated (Page 14 in the revised manuscript) and presents now 4 panels representing differences between raw and corrected float-pH data following the SAGE method (panel A), the GEOMAR methods (panel B), pH data measured at the parking depth (panel C) and pH data measured at the parking depth minus reference (CANYON-B pH data, panel D). We also agree with the reviewer that there were no explanations about the numbers in the original manuscript. In the revised manuscript, they are now explained in the figure caption.

By splitting previous Figure 5 (A and B) into two separate figures, we believe that the new organization of the figures helps the reader to identify the impact of the sensor drift correction used on the final adjustment. We also agree with the referee that the impact of the correction method on the final corrected dataset is almost non-significant, especially regarding the mean difference values. Nevertheless, as previously stated, this section aims to discuss the better representation of the sensor behavior over time and we believe that, by merging Figure A1 (in the original manuscript) to the original Figure 5 in the revised manuscript, this purpose has been clarified to the reader.

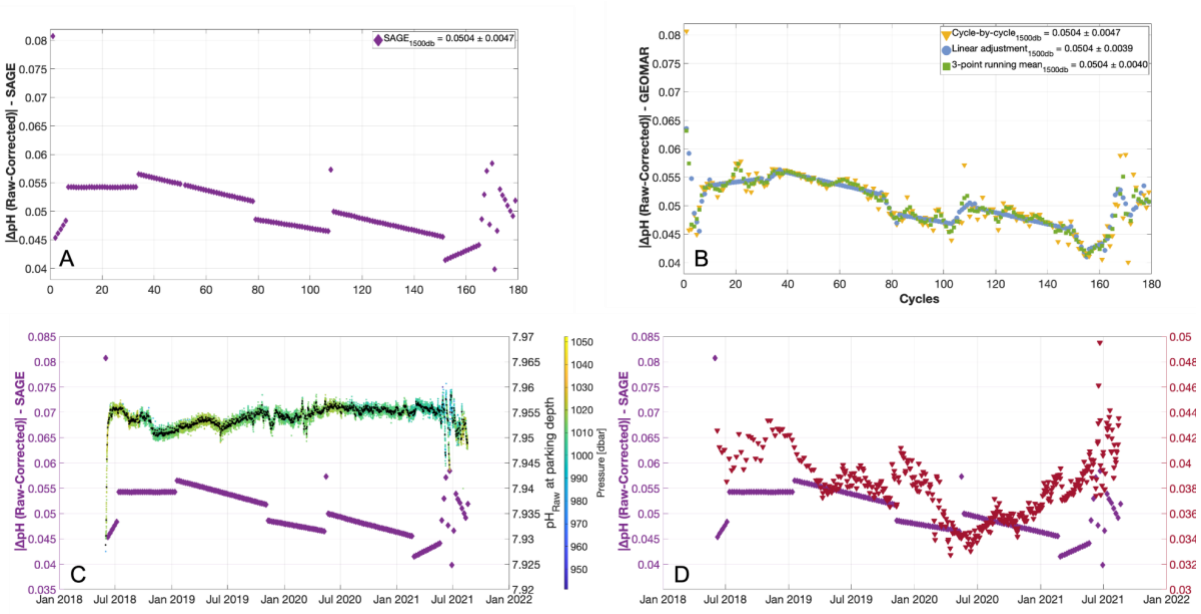

Figure 3 (Fig. 5 in the revised manuscript)**.** Mean differences between raw float-pH data minus float-pH corrected using the SAGE tool (Fig. 5A), the cycle-by-cycle GEOMAR method (yellow dots, Fig. 5B), and the linear mean regression GEOMAR method (blue dots, Fig. 5B) and the 3-point centered running mean correction method (green dots, Fig. 5B) for float WMO 3901669. In every case, CANYON-B was chosen as a reference method, and 1500 dbar was chosen as the reference depth. Mean differences between raw and corrected float-pH data with the standard deviations are shown in the legend boxes for each reference method. Figure C shows, for comparison with the SAGE correction, the uncorrected pH data measured at the parking depth (right y-axis) with black dots representing mean pH values for each day. The colorbar shows pressure. Figure D shows differences between raw float-pH data minus float-pH corrected using the SAGE tool

(purple dots, left y-axis) and differences between uncorrected mean pH data measured at the parking depth minus mean reference CANYON-B pH data calculated using measurements recorded at the parking depth (red dots, right y-axis).

278: profile-by-profile or cycle-by-cycle? Either is fine but be consistent. Also, doesn't it remove too much sensor noise, or am I misunderstanding?

We thank the referee for pointing out this mistake. The denomination has been uniformized in the revised manuscript and the term "cycle-by-cycle" is now used. Lines 315 and 322 have been modified in Section 3.1.3. Moreover, we agree with the referee that this sentence was confusing in the original manuscript, especially regarding the denomination "sensor noise" used. The line 278 (in the original manuscript) was related to short-term sensor drifts and noises along the water column, which in fact could not be only related to the correction method but also to the variability in the algorithm estimates as well as chemical reactions close to the sensor chip. With regard to these doubts, this unclear and unverifiable sentence has been modified in the revised manuscript. Indeed, the purpose of this section, generally speaking, is only to discuss the noticeable step-like changes observed with the current correction procedure and to find the best way to represent the smooth sensor drift over time.

L.315: "*The cycle-by-cycle adjustment has the disadvantage that it gives discontinuous adjustment rather than a segmented set of piecewise adjustments*".

282: "has to" is an overstatement as this is not currently done by many data centers (this is another case where the verb is among the last words in the sentence). Also, you have not completely made the case against the cycle-by-cycle approach or the SAGE approach. The basis for the SAGE approach, I believe, was based upon the first principles assumption that the reference potential jumps with discrete events and that the wiggles around the jumps reflect variability in the algorithm estimates or short term sensor noise. This hypothesis would need to be discounted to really make the case effectively.

We thank the referee for pointing out this overstatement and suggesting a reformulation of this sentence. In the revised manuscript, this sentence has been modified and tone down as: "*Therefore, the adjustment method should involve techniques such as a higher-order spline fit, a centered running mean, or a segment separation of the record into linear drift phases."* (L.318)

We agree with the reviewer that there is no clear comparison in the original manuscript. In the revised manuscript, Figure 5 has been modified and explanations regarding our interpretation of the current correction method have been added. Moreover, as stated in one former answer to the referee, the purpose of this section was to discuss the noticeable step-like changes observed with the current correction procedure (i.e., the SAGE method) and to find the best way to represent the smooth sensor drift over time, as observed when looking at the pH time-series recorded at the parking depth. Indeed, in comparison with the pattern of the cycle-by-cycle correction, the high pH changes of ca. 0.01 pH units observed between linear drift phases and leads to step-like changes with the SAGE method appear to be unrealistic. In our view, the sensor rather shows undulations in response with smooth and less smooth phases. This statement is somehow confirmed by the pH sensor behavior when the float drifts at its parking depth. In consequence, we believe that an adaptation of the current correction procedure could be done to better maximize the smoothness of the corrections and to avoid introducing artificial jumps.

L.327: *"The pH sensor behavior when the float drifts at its parking depth is in agreement with this observation (Fig. 5C). In comparison with float-pH data corrected using the SAGE method, no strong visible discontinuities in raw pH data are observed while the float drifts between its measurement's phases. In our view, the sensor rather shows undulations in response with smooth and less smooth phases over time. In order to test the impact of the reference method on the adjustment pattern, differences between uncorrected float-pH data and CANYON-B pH data derived at the parking depth are presented in Figure 5D. Once again, the pH time-series shows smoothed transitions and the general pattern does not present noteworthy jumps. Such sharp transitions can perhaps be best corrected with our modified GEOMAR segment method or alternatively with a spline fit or a 3-point centered running mean (Fig. 5B)."*

293: discontinuities are not a problem for QC. The statistics still work fine. They are perhaps a problem for studies examining biogeochemical variability over time for a specific profiling float, but these studies would also be challenged by excessively smoothed transitions if a reference potential jump occurred. I believe the authors have a strong case to make here, but this presentation leaves me more confused than convinced. I'd urge them to instead focus on the consequences for a common biogeochemical analysis that would be affected by discontinuities (leading to, e.g., discontinuities in DIC vs. time… though even then a clearly visible discontinuity in DIC might be preferable to a smooth-seeming but equally spurious excursion in DIC as the smoothed adjustment factor catches up to the true adjustment factor). These arguments lead me to the belief that the best approach would be a 1000-1500 m average adjustment applied cycle-by-cycle.

We agree with the referee that statistics are fine and almost not impacted by the discontinuities observed depending on which method is used to correct float-pH data. Nevertheless, as now more clearly stated in the revised manuscript, we believe that some corrections methods, especially the cycle-by-cycle and the SAGE linear adjustment ones, induce jumps that are not observed either on float-pH data time-series or when pH data are recorded while the float is at its parking depth. In consequence, we argue that, when a peculiar float-pH profile is used in comparison with discrete pH measurements in order to compare and examine the accuracy of the correction, such artificial variability induced by the method and not related to the sensor itself could lead to biases and possible misadjustment. In the revised manuscript, sentences have been added to clarify our point of view.

L.338: "*Indeed, and even if the impact of the adjustment method on the final corrected dataset is almost non-significant regarding the mean differences values (Fig. 5D), the possible impact of such artificial jumps induced by the method itself rather than the pH sensor could be noticeable if float-pH data related to these peculiar discontinuous cycles are compared against discrete pH measurements and then adjusted (see Section 3.2).*"

289: Figure A1 is critical for your argument. It needs to be brought into the main text if this section is retained in the final manuscript. It needs to be well-explained and your rationale for preferring the smoothed adjustments over alternatives needs to be more strongly and thoroughly defended. The rationale should go quantitatively beyond "In our view, the sensor rather shows undulations in response with smooth and less smooth phases."

We thank the referee for this relevant suggestion which helps supporting the discussion in Section 3.1.3. Figure 5 has been re-drawn in the revised manuscript and includes now the former Figure A1. We also

thank the reviewer for his suggestion to better describe the figure in the main manuscript in order to use it as an argument for our assessment. The legend of the new Figure 5 has been modified accordingly and explanations have been added in the revised manuscript.

L.327: *"The pH sensor behavior when the float drifts at its parking depth is in agreement with this observation (Fig. 5C). In comparison with float-pH data corrected using the SAGE method, no strong visible discontinuities in raw pH data are observed while the float drifts between its measurement's phases. In our view, the sensor rather shows undulations in response with smooth and less smooth phases over time."*

293: others would argue that these are correcting biases of that magnitude

In the studied area, we observed that the current float-pH correction procedure is impacted by the choice of the reference method used to correct the data (uncertainty of ca. 0.015 pH units), the choice of the reference depth (uncertainty of ca. 0.005) but also the method itself used to correct data which could lead to biases of up to 0.01 pH units. In consequence, with regard to the literature stating that for SBE pH sensors, the accuracy ranges from ± 0.05 pH units (manufacturer statement) to ± 0.005 pH units after data adjustment (Johnson et al., 2017)[5], we believe that these discontinuities have to be more constrained to decrease the adjusted error.

Figure A1: would be more useful if pHTotalInsitu were plotted as the difference from the algorithm estimate. Also, it is unclear if the parking depth pH value has any adjustments applied and, if so, on what basis. Finally, the indication of whether pH is "total scale pH" is inconsistent in these figures. Based on other conversations with people who have strong opinions about these things, my recommendation is to universally use pHT to indicate total scale pH.

We acknowledge the reviewer for suggesting a comparison between pH data measured at the parking depth and pH data estimated using an algorithm. This comparison has been added in the revised manuscript using CANYON-B as a reference method. We believe that this new figure (labeled Figure 5D) improves our presentation and highlights well that the pattern observed for corrected float-pH data (ex. Figure 5A) is not related to the reference method used to correct float-pH datasets but rather to sensor drift adjustment done by each method. In Figure 5D (in the revised manuscript), the y-axis label "pHT in situ" has been replaced by "pH raw" as these data are uncorrected. When pH data recorded at the parking depth are plotted, they are uncorrected. This precision has been added in the legend of Figure 5.

L.330: "*In order to test the impact of the reference method on the adjustment pattern, differences between uncorrected float-pH data and CANYON-B pH data derived at the parking depth are presented in Figure 5D. Once again, the pH time-series shows smoothed transitions and the general pattern does not present noteworthy jumps. Such sharp transitions can perhaps be best corrected with our modified GEOMAR segment method or alternatively with a spline fit or a 3-point centered running mean (Fig. 5B).*"

[5] Johnson, K. S., Plant, J. N., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Riser, S. C., et al. (2017). Biogeochemical sensor performance in the SOCCOM profiling float array. J. Geophys. Res. Oceans 122, 6416–6436. https://doi.org/10.1002/2017JC012838.

302: Yikes! Variants on this experiment have been performed by several members of the community, and none, to my recollection, saw consistent differences this large.

We agree with the referee that the literature showing comparisons of quality-controlled float-pH data against shipboard reference data shows differences much lower. For example, Maurer et al. (2021)[6] present a median bottle-minus float difference for pH data of 0.002 ± 0.015 pH units. In the revised study, including the ESPER methods, we found mean differences ranging between -0.0659 and -0.0150 pH units, and varying according to both the reference pressure level and the reference method used to correct float-pH data. Nevertheless, when excluding the LIR-pH method which seems to induce an over-adjustment of the dataset, and when removing the first four measurements (measured within the first 50 meters of the water column), mean differences of 0.009 and -0.0018 pH units are obtained using ESPER-LIR and ESPER-NN, respectively, and the 1950 db reference pressure. By considering the lab-to-in situ pH conversion uncertainty introduced through calibration (0.005 pH units; Williams et al., 2017[7]), lower uncertainties are even obtained. Thus, we believe that the current correction procedure is relevant at depth but that, in this area, large differences are observed near-surface and might reflect an imperfect representation of the temperature dependence. This assumption is at some points confirmed by the comparison between SOOP-based pH measurements and float-pH data pointing towards apparent biases toward the surface. Moreover, in a recent publication focusing on a high-Arctic fjord (Kongsfjorden, Svalbard), Gattuso et al. (2023)[2] present offsets between spectrophotometric reference samples and a calibrated SeaFET pH time series ranging between ± 0.02 pH units.

In the revised manuscript, some details about the uncertainties to consider (i.e., bottle pH inaccuracy and lab to in situ pH conversion uncertainty) have been added to discuss the results observed and tone down the observed differences.

L. 361: "*Moreover, the laboratory-to-in-situ temperature pH conversion uncertainty of 0.005 pH units (Williams et al., 2017), as well as the absolute uncertainty in the bottle pH measurements (here 0.002 pH units), have to be taken into account before drawing strong conclusions*."

302: There is a growing sense in the community that bottle pH samples are not well preserved even when following SOPs for DIC and AT storage. Is this possibly a discrete sample issue? Do you have some at-sea measurements to compare with?

During the MSMS94 cruise, samples for total alkalinity, dissolved inorganic carbon as well as pH measurements were taken. These samples were poisoned onboard following the current standard procedure and analyzed at GEOMAR. pH measurements were tested regularly against CRM reference samples to check the accuracy of our measurements. While CRMs are certified only for TA and DIC, pH measurements were also performed for each bag by Dickson's lab and made available for us. In consequence, and even if no at-sea measurements are available to compare with, the comparison done against these certified materials leads us to conclude that the resulting uncertainty in pH measurements for discrete samples was ± 0.002 pH units.

[6] Maurer, T. L., Plant, J. N., and Johnson, K. S. (2021) Delayed-Mode Quality Control of Oxygen, Nitrate, and pH Data on SOCCOM Biogeochemical Profiling Floats, Frontiers in Marine Sciences, 8, 683207. https://doi.org/10.3389/fmars.2021.683207, 2021.

7 Williams, N. L., Juranek, L.W., Feely, R. A., Johnson, K. S., Sarmiento, J. L., Talley, L. D., Dickson, A. G., Gray, A. R., Wanninkhof, R. Russel, J. L., Riser, S. C. and Takeshita, Y. (2017). Calculating surface ocean pCO_2 from biogeochemical Argo floats equipped with pH: An uncertainty analysis, Global Biiogeochemical Cycles, 31, 591-604, https://doi.org/10.1002/2016GB005541.

302: the two numbers in this figure only make sense after looking at figure 6, and it should be clear from the text alone

In the revised manuscript, Figure 6 is now described in the main text and the numbers, related to the observed differences, better explained.

L.347: "*Figures 6A and B present differences between discrete pH measurements and float-pH data along the water column and according to two distinct reference pressure levels. We find mean differences ranging between -0.0659 and -0.0150 pH units (Fig. 6B) between the reference pH cast and the fully corrected pH of cycle 122, with higher differences found for the "classical" reference depth of 1500 dbar, and the lowest differences reported for the two ESPER methods.*"

6D: the y axis is unlabeled and missing many negative signs (looks like the figure was just cut off on the left). Far more worrying, the delta pH is about the same magnitude as the change in pH expected from changes in temperature. It is unlikely that there is a 100% uncertainty in the calculated pH change with temperature, which leads a possible simpler explanation that the pH vs. temperature correction was performed incorrectly in this comparison. Were the discrete pH values recalculated at the in situ temperature and pressure? When you say "discrete water temperature" is this the temperature at the time of bottle triggering or the laboratory temperature at the moment of analysis?

We thank the referee for pointing out that Figure 6D was incomplete in the original manuscript. In the revised manuscript, Figure 6 (Figure 4 below) has been re-drawn (Page 16 in the revised manuscript) and now presents differences between discrete pH measurements and float-pH data along the water column and according to two distinct reference pressure levels (1500 dbar, Fig. 6A and 1950 dbar, Fig. 6B) and ΔpH (discrete pH measurements minus float-pH data corrected at the reference depth level 1950 dbar, Fig. 6C) as a function of the difference between discrete water temperature and temperature values recorded at the reference depth of 1950 dbar (i.e., 3.3733°C). Here, discrete water temperature refers to the temperature measured in situ at the time of bottle triggering (at sea). This precision has been added in the legend of Figure 6 in the revised manuscript. On every panel of the revised Figure 6, the four reference methods are also presented.

Discrete pH samples were analyzed at GEOMAR right after the cruise at standard temperature (~25°C) and atmospheric pressure and have been converted to in situ temperature and pressure for this study. The conversion was done using the CO2SYS software with measured pH data and TA values as input variables (see Table 1). A double check for potential errors has been done. Thermodynamic calculations within the carbonate system used the carbonic acid dissociation constants of Mehrbach et al. (1973) as refit by Dickson and Millero (1987), the dissociation constant for bisulfate of Perez & Fraga (1987) and Uppström (1974) for the ratio of total boron to salinity.

pH measured in the laboratory at ~ 25°C	**pH measurement - temperature in the laboratory [°C]**	**TA measured [µmol/kg]**	**Temp. *in situ* [°C]**	**Salinity**	**Pressure *in situ* [dbar]**	**pH converted to *in situ* Temp. & Pres.**
7,708	25,0173	2304,48	3,3931	34,9257	1925,8	**7,9563**
7,706	24,9913	2303,18	3,4606	34,9041	1620,4	**7,9647**
7,694	25,01433	2300,47	3,3397	34,8582	1012,5	**7,9779**
7,687	25,0256	2298,7	3,4712	34,8644	706,1	**7,9805**

7,681	25,0053	2296,33	3,4765	34,8456	504,5	**7,9816**
7,689	25,055	2296,44	3,5466	34,7806	201,6	**8,0018**
7,685	25,0313	2294,06	3,94	34,7375	100,9	**7,9947**
7,737	24,9803	2298,01	5,3817	34,6516	51,5	**8,0283**
7,848	24,9823	2295,34	8,4924	34,5503	30,6	**8,0959**
7,875	25,03267	2293,27	12,3241	34,4711	19,9	**8,0646**
7,865	24,935	2272,82	12,7456	34,3996	10	**8,0466**

Table 1 (not in the revised manuscript). Parameter values used as inputs to convert pH data from standard temperature (~25°C) and atmospheric pressure to in situ temperature and pressure using the CO2SYS software. The last column on the right side of the table presents pH data used in this study.

Example: conv=CO2SYS(7.694,2300.47,3,1,34.8582,25.0143,3.3397,0,1012.5,0,0,1,4,2);

conv(:,18)=7.9779 % 18 - pH output

with CO2SYS (pH measured value, TA value, parameter type (pH), parameter type (TA), salinity, temperature input (during the measurement), temperature output (in situ), pressure input (during the measurement), pressure output (in situ), SI concentration, PO4 concentration, selection of the pH scale, selection of the K_1K_2 constants, selection of the KSO4 constants).

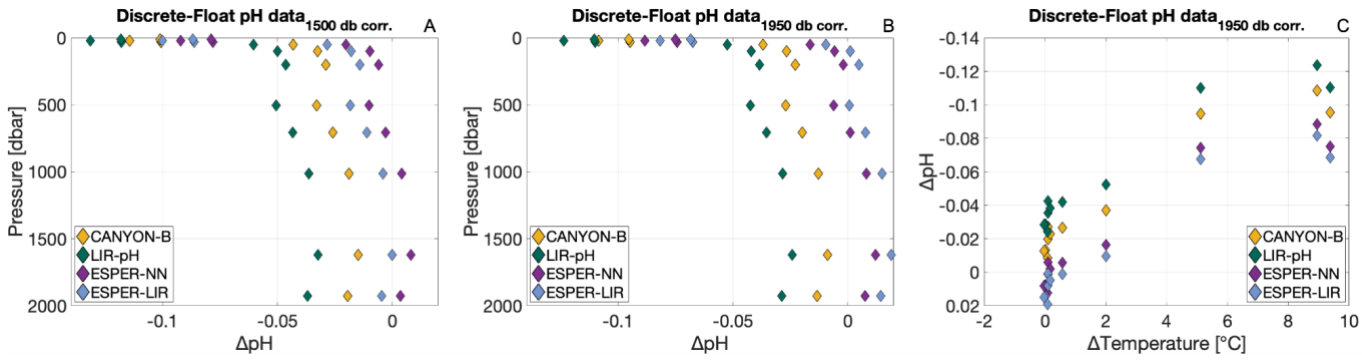

Figure 4 (Fig. 6 in the revised manuscript). (A and B) Differences between discrete and float-pH data (for the cycle 122) calculated after matching in density space to avoid biases from internal waves and corrected using corrected reference levels of 1500 dbar (Fig. 6A) and 1950 dbar (Fig. 6B). (C) ΔpH (discrete pH measurements minus float-pH data corrected at the reference depth level 1950 dbar) as a function of the difference between discrete water temperature (i.e., the temperature measured in situ at the time of bottle triggering at sea) and temperature values recorded at the reference depth of 1950 dbar. The color code refers to the reference method used to correct float-pH data: CANYON-B (yellow diamonds), LIR-pH (green diamonds), ESPER-NN (purple diamonds) or ESPER-LIR (bleu diamonds).

324: SOOP-pH is not a mature effort to my understanding. Some comments on the SOOP-pH methods and QC practices are warranted. SOOP-pCO2 comparisons, to my understanding, are showing much more modest implied float offsets

Indeed, most SOOP feature only pCO_2 measurements but other CO_2 system variables are coming along well. We have put much effort into testing, improving and assessing autonomous spectrophotometric TA measurements (CONTROS HydroFIA TA) and have reached a quite decent accuracy of about 5

µmol kg^{-1} in unattended SOOP mode (Seelmann et al., 2019[8], 2020a[9], 2020b[10]). Using a much simpler analytical setup (as it does not require sample acidification and CO_2 stripping) of this commercial spectrophotometric system for pH (CONTROS HydroFIA pH), we have gained quite some experience in SOOP-based pH measurements. Because of the relatively high stability of the pH measurement, a suite of 5-8 repeated CRM-reference measurements are performed in port before and after each 5-week autonomous roundtrip (Fig. 5 below). These pre- and post- calibration runs are rather stable for each meta-cresol purple (mCP) indicator bag. This yields a clear and consistent track of the small pH drift over consecutive roundtrips which allows us to correct the measured pH to CRM values. Given the small standard deviations of the CRM measurements we believe that the SOOP-pH is of about ± 0.003 pH units.

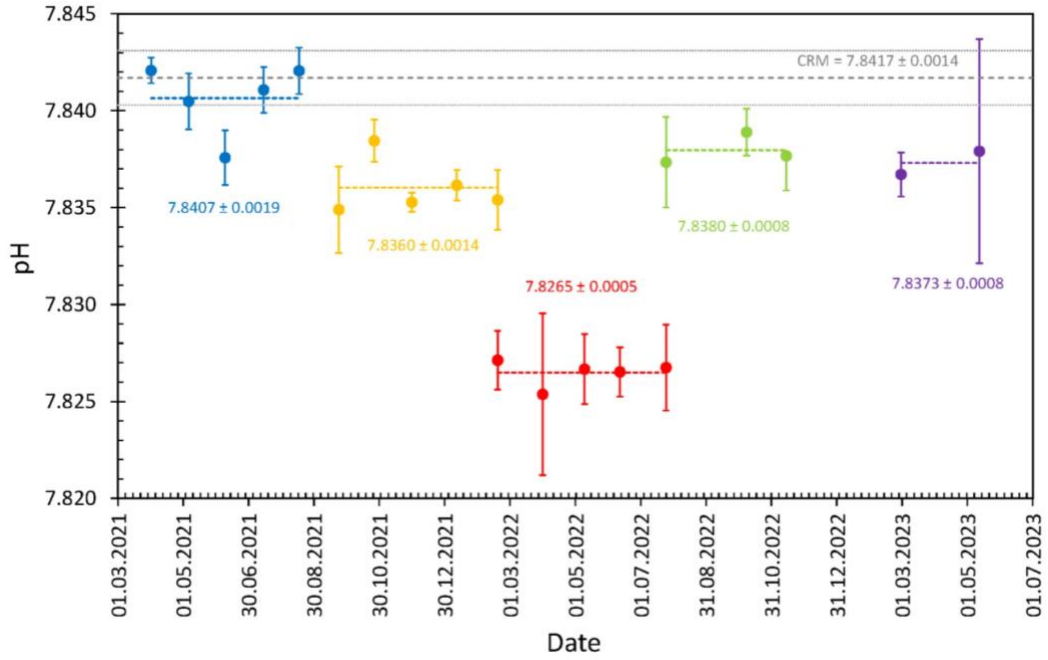

Figure 5 (not in the revised manuscript). pH measurements performed on CRM batch 190 with the CONTROS HydroFIA pH system before and after each 5-week roundtrip of the DE-SOOP *Atlantic Sail*. Adjustments of measured pH to the nominal pH value assigned to the CRM (7.8417 ± 0.0014 at 25°C) is based on the mean of all CRM measurements carried out per individual meta-cresol purple bag.

328: needs

The verb has been corrected.

Table 2: A standard deviation of 0.5 salinity units is quite large, is it not? Usually this represents an offset of hundreds of km or more in the surface of the North Atlantic except in

[8] Seelmann, K., Aßmann, S. and Körtzinger, A. (2019). Characterization of a novel autonomous analyzer for seawater total alkalinity: Results from laboratory and field tests. Limnol. Oceanogr.: Methods. https://doi.org/10.1002/lom3.10329.
[9] Seelmann, K., Steinhoff, T., Aßmann, S. and Körtzinger, A. (2020a). Enhance ocean carbon observations: Successful implementation of a novel autonomous total alkalinity analyzer on a Ship of Opportunity. Front. Mar. Sci., 7. https://doi.org/10.3389/fmars.2020.571301.
[10] Seelmann, K., Gledhill, M., Aßmann, S. and Körtzinger, A. (2020b). Impact of impurities in bromocresol green indicator dye on spectrophotometric total alkalinity measurements. Ocean Sci., 16, 535-544. https://doi.org/10.5194/os-16-535-2020.

areas of very strong salinity gradients near coasts. A mean of 0.4 is just as worrisome. Am I misreading this?

Crossovers in the surface ocean are much harder to achieve due to the typically large spatiotemporal variability there. This is particularly the case in the subpolar Northeast ocean, where our SOOP and BGC-Argo float measurements take place. The close proximity of the warm and saline waters of the North Atlantic Current and the cold and less haline waters of Arctic origin cause particularly high spatiotemporal variability. The stricter SOCAT criterion would have yielded only very few crossovers with little statistical significance. We therefore decided to enlarge the search window considerably for the sake of yielding more crossovers. Of course, these individual crossovers are per se even less statistically significant. By plotting all delta-pH vs. delta-T between the float and SOOP pH measurements (note that SOOP pH data were corrected to the float pH observations using CO2SYS and the observed SOOP-TA), we were hoping to find a reasonably linear correlation which at delta-T = 0 should yield a relatively robust estimate of the delta-pH. On average the two pH datasets differed by about 1°C for both floats (although in opposite directions). So clearly this is not a perfect match. We did the same for a correlation vs. delta-S which was slightly less well-constrained but yielded essentially the same pH offset. For both floats, we found about the same salinity offset, again in opposite directions. In a possible future approach that harnesses SOOP-pH for assessment/correction of float-pH, further thought should be put into checking and optimizing the crossovers. Still, we are convinced that the results shown despite their limitations and because of their consistency across the 2 floats and with the discrete hydrocast crossover in the Labrador Sea are clear evidence of an accuracy issue with upper ocean float pH. This we feel is important for the community and should foster similar studies.

364: Climate and weather goals are specifically formulated based on needs for characterizing (paraphrased) "changes in carbonate ion concentrations." This means they are related to precision and not accuracy. This analysis is concerned with accuracy to a much greater extent than precision, so this is not an apples to apples comparison. A better comparison would be how the offset varies over time for a given float or varies between two or more floats in the same location.

Recently, the Global Ocean Acidification Observing Network (GOA-ON) has discussed measurement quality goals that need to be met to ensure appropriate quality to address the relevant problems. Thus, GOA-ON has proposed two key goals corresponding to two levels of related uncertainties: the weather and the climate goal. We agree with the referee that climate and weather goals are related to precision, i.e., "the result of a measurement that permits a statement of the dispersion (interval) of reasonable values of the quantity measured, together with a statement of the confidence that the (true) value lies within the stated interval" (Newton et al., 2015[11]). In our manuscript, whereas Section 3.2. aims to compare float-pH data against discrete pH measurements to assess the accuracy of the correction procedure (and then discuss the errors of the corrected datasets), Section 3.1. presents the dispersion of the corrected float-pH data in response to the reference pressure choice, the reference depth selection as well as the choice of the method used to correct float cycles. We have demonstrated in the manuscript that significant differences ranging between ca. 0.003 pH units and ca. 0.04 pH units are observed between the four reference methods which can be used to correct float-pH data. In the studied area, this study also shows that differences related to the reference pressure level choice ranged between 0.0047 and 0.0141 pH units (Fig. 3B). By combining the observed possible sources of uncertainty, the

[11] Newton J.A., Feely R. A., Jewett E. B., Williamson P. and Mathis J. (2015). Global Ocean Acidification Observing Network: Requirements and Governance Plan, Second Edition, GOA-ON, http://www.goa-on.org/docs/GOA-ON.

corresponding uncertainty is either at the edge or well beyond the weather and climate goals, respectively. However, we agree with the referee that several sentences in the original manuscript were out of the scope of this discussion as they were related to accuracy and errors, which is not the purpose of these GOA-ON goals. To clarify the situation, some sentences have been removed and Section 3.3 in the revised manuscript has been shortened and modified. Figure 8 has been also removed in the revised manuscript to clarify it.

379: The TA uncertainty is also far more important if we become interested in DIC.

We agree with the referee that TA uncertainty can lead to noticeable uncertainties when this parameter is used in association to pH data to derive DIC. As stated by Millero (2007)[12], the estimated probable error is even higher when TA values are used in association with pH data to derive DIC (± 3.8 mol kg^{-1}) than fCO_2 (± 2.1 µatm). However, in the context of converting surface ocean pH measurements into pCO_2 data for the purpose to derive air-sea CO_2 fluxes and determine the ocean behavior with regard to the current atmospheric CO_2 increase, Section 3.3. focuses more on this parameter rather than on all the parameters of the marine CO_2 system. In the revised manuscript, the sentence has been slightly modified to follow the referee's comment but no in-depth description of the TA uncertainties implications is done.

L.430: "*This perhaps warrants specific tests on the accuracy of TA predictions in critical regions (or seasons) but also if this parameter aims to be used to derive other parameters of the CO_2 system, especially DIC.*"

385: The situation becomes worse still when including uncertainties in the carbonate system constants (Orr et al., 2018)… however, again, these additional uncertainties should be relatively consistent over time, and "weather" and "climate" relate to precision. Many of the concerns raised herein (and the concern over carbonate constants) fall away when you begin to examine variability in the float observations over time

To frame its weather and climate goals, GOA-ON has proposed relative uncertainties thresholds in calculated carbonate ion concentration. These thresholds have been used to back-calculate the corresponding maximum permissible uncertainties in measured input variables. For parameters of the CO_2 systems, there are uncertainty contributions from different sources: the instrumental precision, the data conversion uncertainty, and the carbonate system equilibrium constant uncertainties. In consequence, uncertainty propagation should include all those identified sources of bias as it is stated by Orr et al. (2018). In the revised manuscript, this information has been added to better depict the overall concern when deriving pCO_2 values from float-pH data and TA estimates.

L. 432: *"Finally, an additional source of uncertainty when calculating pCO_2 (pH, TA) from floats originates from uncertainties in the carbonate system equilibrium constants (Orr et al., 2018)."*

[12] Millero, F. J. (2007). The marine inorganic carbon cycle. Chem. Rev., 107 (2), 308-341. https://doi.org/101021/cr0503557.