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

Responses to REVIEWER 1 - Brendan CARTER

Dear authors,

I have gone through the revised version of your manuscript. I commend your hard work on this revised draft and your detailed responses to reviewers. You have attempted to answer all of the comments and questions and many have been addressed satisfactorily. However, I still have concerns about a few aspects:

We appreciate the time and effort that the reviewer dedicated to providing feedback on our manuscript and are grateful for the insightful comments on and valuable improvements to our paper. In the revised manuscript, we have incorporated most of the suggestions made by the reviewer and tried to address his remaining concerns in the following response. 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. Line numbers correspond to the PDF file.

1. I still have strong misgivings about the mismatch in the salinity. I'm not sure that the central comparison in this paper can be valid with that large of a discrepancy. These appear to be different water masses.

In order to yield a larger number of crossovers, a rather large search window was applied. These crossovers rarely are perfect matches in T and S. Therefore, under the assumption that differences in pH to a major extent are driven by differences in temperature, the ΔpH at $\Delta T = 0$ was calculated. By fitting a linear regression to the data (as the intercept of the regression equation), the pH offset was estimated following the assumption that this regression using crossovers achieved with a relatively wide search window yields a more robust ΔpH estimate as an average of a much smaller number of crossovers found with a smaller search window. We note that by calculating the pH offset as a function of ΔS (i.e., @ $\Delta S = 0$), the resulting ΔpH values are statistically indistinguishable from the ones based on ΔT (but have slightly larger uncertainty). This in essence means that the result does not depend significantly on whether we calculate the offset for isothermal ($\Delta T = 0$) or ioshaline ($\Delta S = 0$) conditions. The fact that the isothermal condition is not exactly isohaline (and vice versa) indicates that we do not have a perfect match in water masses. Given the slope of the ΔpH vs. ΔS regression (Table 1 below and in the revised manuscript), this mismatch seems to not introduce any discernible uncertainty in the pH offset. We therefore based the estimation on the linear regression in T space which shows a moderate sensitivity of ΔpH with ΔT. In consequence, we believe that, by increasing the research window to find crossovers only the uncertainty (±) is affected but not the mean pH offset.

Nevertheless, given the limitation of our dataset containing only 2 floats, 4 additional floats (that were not part of our pilot study) with trajectories overlapping the SOOP line transect were used to test our assumptions. The ESPER_LIR reference method was used to correct all float-pH data and only float data flagged as "good" were used for this analysis. Except for the floats WMO 1902303 and 1902304

that were corrected using a reference depth around 950 db (940-980 db) as some cycles did not go deeper than 1000 db, all the floats were corrected at 1950 db. Fig. 1A shows that differences between these floats and SOOP-based pH observations (corrected to the temperature of the respective float surface pH observations) do not show a temporal bias, indicating that the SOOP-pH dataset is not biased by a drift pattern; in agreement with Figure 2 below. Moreover, the range of apparent offsets of ± 0.03 pH observed for this 6-float dataset (Fig. 1E) is essentially independent of the search radius criterion, indicating that the large crossover search window does not introduce a bias but only adds more data points (Fig. 1B). Following the comment and suggestion of the reviewer, offsets between SOOP-pH and float-pH data as a function of temperature difference were plotted introducing an additional criterion ($\Delta S < 0.5$) for the cross selection. We note, however, that no significant difference between the calculated pH offsets based on the two different S criteria ($\Delta S < 0.5$, $\Delta S < 2.0$) is observed. As shown in Figure 1D, no slope in the ΔpH vs. ΔS regression can be reported, highlighting the lack of dependence with this parameter with data linearly spanned in comparison with the temperature-pH dependence (Fig. 1C), confirming our assumption that differences in pH to a major extent are driven by differences in temperature. The pH offset was determined at $\Delta T = 0$ °C (temperature difference between float data and SOOP data) by fitting a linear regression to the data for the float having spread ΔT values or by considering the mean pH difference when $\Delta T = 0$ (Fig. 1E). Table 1 shows the pH offsets and their uncertainties for the six floats considered and derived using either ΔT=0 or ΔS=0. While the crossovers identified for the six floats are not a perfect match, they all point toward unacceptably high and not constant biases in surface pH values that are too large to be applied as correction.

In the revised manuscript, a reduced ΔS value of 0.5 is now used to derive the pH offset at $\Delta T = 0$ and this additional comparison plot and table have been added in the Supplementary Material and discussed in the main manuscript.

"While we found no dependence between ΔpH and ΔS, an additional criterion of $\Delta S \leq 0.5$ has been applied to the crossovers selection in order to exclude major water mass discrepancies." (L. 401)

"*Calculating the apparent pH offset as a function of ΔS (Table A2) yields ΔpH values which are statistically indistinguishable from the ones based on ΔT*." (L.415)

"An extended crossover comparison with the addition of four floats (that were not part of our pilot study) yields mean pH offsets that fall in the range. ± 0.03 pH units (Fig. A2). These mean pH differences are randomly distributed in space and time, indicating an incomplete float-pH data adjustment rather than a drift in the SOOP-reference dataset." (L. 420)

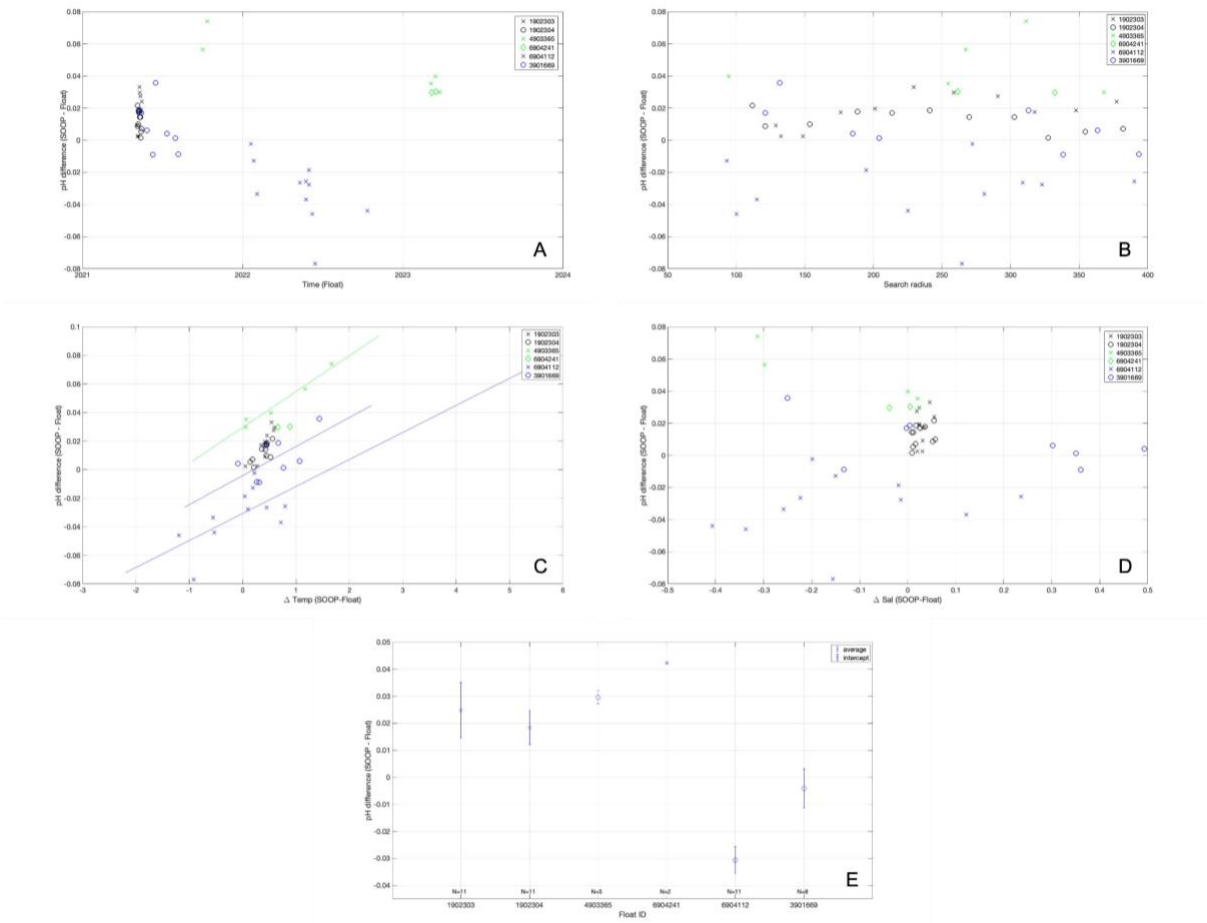

Figure 1 (Only Figs. 1A, 1B, and 1E are in the revised manuscript). Offsets between SOOP-pH and fully corrected float-pH data (y-axis) as a function of the time (Fig. 1A), the crossover criterion (Fig. 1B), the temperature difference (SOOP minus float temperature; Fig. 1C), and the salinity difference (SOOP minus float salinity; Fig. 1D). Figure 1E shows the mean offsets and their associated uncertainties for the 6 floats considered. The pH offset was determined at $\Delta T = 0$ °C (temperature difference between float data and SOOP data) by fitting a linear regression to the data for the float having spread ΔT values (dots) or by considering the mean pH difference when $\Delta T = 0$ (crosses; Figure 1E). Crossovers were calculated for $\Delta x \leq 400$ km, $\Delta t \leq 7$ d, and $\Delta S \leq 0.5$. pH values were recalculated using CO2SYS (van Heuven et al., 2011) to account for any temperature difference between matched observations. Float-pH data have been corrected with the SAGE tool using either the reference depth level 950 dbar or 1950 dbar and ESPER-LIR as reference (see Table 1 below). N stands for the number of values used to derive the statistics.

Table 1. (Table A2 in the Supplementary Material of the revised manuscript). Statistics of the crossover analysis for SOOP- and float-pH data. N stands for the number of values used to derive the statistics. Crossovers were calculated for $\Delta x \leq 400$ km, $\Delta t \leq 7$ d, and $\Delta S \leq 0.5$. pH values were recalculated using CO2SYS (van Heuven et al., 2011) to account for any temperature difference between matched observations. Float-pH data have been corrected with the SAGE tool using either the reference depth level 950 dbar or 1950 dbar and ESPER-LIR as reference.

			ΔpH at ΔT=0		**ΔpH at ΔS=0**	
Correction Depth	N	*Float WMO*	**pH offset**	**Uncertainty of the offset**	**pH offset**	**Uncertainty of the offset**
950	11	*1902303*	0.025	0.010	0.018	0.010
950	11	*1902304*	0.018	0.006	0.012	0.006
1950	5	*4903365*	0.030	0.002	0.036	0.004
1950	2	*6904241*	0.042	0.0003	0.030	0.0003
1950	11	*6904112*	-0.031	0.014	-0.029	0.008
1950	8	*3901669*	-0.004	0.007	0.013	0.006

2. I'm not certain that the point was understood about the inherent inaccuracy of pH, and why it's reassuring, but not sufficient for the authors' purposes, that measurements match those of Dickson's lab. If I recall, in this paper the authors are claiming an uncertainty in pH that is better (0.002) than the uncertainty assessed by the Dickson Lab for the measurement technique used to measure seawater reference materials (https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.4319/lom.2013.11.16) (assessed at 0.01 to 0.02 with the potential to improve to as good as 0.005). For calculations of pCO2 from pH, we care about the accuracy of the pH value and not the precision.

During the MSMS94 cruise, total alkalinity, dissolved inorganic carbon, as well as pH measurements, were achieved following the current standard procedure (SOP). During the analysis at GEOMAR, pH measurements were tested regularly against CRM reference samples. While CRMs are certified only for TA and DIC, pH measurements were also performed for each batch by Dickson's lab and made available for us. The comparison done against these certified materials leads us to conclude that the resulting reproducibility in pH measurements for discrete samples was ± 0.002 pH units. We are fully aware that the CRM is not a CRM for pH. However, we think that the Dickson lab is capable of performing accurate pH measurements (with a reported uncertainty of 0.001). With this assumption, and also because all the best practice recommendations have been followed, we are also confident that our SOOP line based pH measurements are accurate and compare well with measurements from a different laboratory (i.e. Dickson lab). Indeed, 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. 2 below). These pre- and post- calibration runs are rather stable for each individual meta-cresol purple (mCP) indicator bag (but somewhat different between individual bags). 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. In the revised manuscript, the term uncertainty has been deleted and replaced by reproducibility as suggested by the referee.

In addition, we note that the pH offsets of the 6 selected floats are in the range +0.04 to -0.03, i.e. span a range of 0.07. So even if there was a bias in the CRM pH, the range would remain exactly the same, only the absolute pH differences would shift accordingly. So, the conclusion that in the subpolar North Atlantic, the current pH cookbook procedures do not yield well enough constrained pH is still valid. Only if all 6 floats produced essentially (within error) the same offset, an accuracy issue on the CRM pH would have to be invoked.

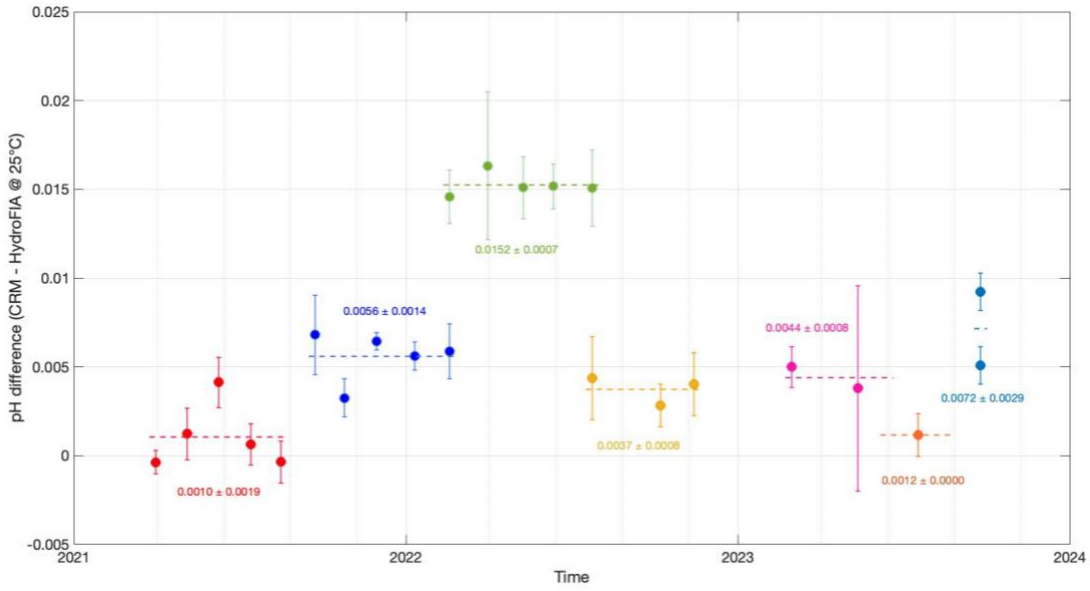

Figure 2 (not in the revised manuscript). pH measurements performed on CRM batch 190 with the CONTROS HydroFIA pH system before and after each 5-week roundtrip of the DE-SOOP Atlantic Sail. Adjustments of measured pH to the nominal pH value assigned to the CRM (7.8417 ± 0.0014 at 25°C) are based on the mean of all CRM measurements carried out per individual meta-cresol purple bag.

However, we agree with the referee that other sources of uncertainty (i.e., the lab-to-in situ pH conversion uncertainty introduced through calibration (Williams et al., 2017) and the bottle pH inaccuracy) have to be considered when comparing the discrete and float-pH datasets.
L. 375: "*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*."

Finally, we do agree with the referee that the final pH accuracy is crucial for the calculation of pCO_2 from pH. In the manuscript, after discussing the different possible sources of uncertainty, 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. The following Section 3.3 tends to link the two previous ones by presenting the impacts of the uncertainties associated with the current adjustment procedure on the final accuracy and thus the derived parameters such as the pCO_2.

3. I still don't understand why the temperature relationship seems so discrepant from other examples of similar measurements.

In the studied area, two crossover 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, in both cases, pH offsets are positively correlated with temperature, being the smallest at the temperature of the reference depth (Figure 6C in the manuscript). 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 17). One possible explanation of that imperfect temperature dependence could be linked to the TCOR ratio used in SAGE to adjust the sensor k0 (that is what is assumed to drift): it is addressed by normalizing the adjustment along the profile to the temperature at which the adjustment was derived. As the temperature gradient could be high along the

water column in this region (> 10°C), we wonder if this TCOR term may induce an under-correction when high temperatures are recorded at the surface and thus an under-representation of seasonal thermal changes. By using either the temperature at 1500 dbar or the temperature at 5 dbar, the TCOR term varies between 0.09618 and 0.9999. For a mean pH value of 7.8273 and an offset of 0.0762 (hypothetical), it represents pH values equal to 7.9006 and 7.9035, respectively, representing a difference of ca. 0.003 pH units.

Also, the pH sensor laboratory-calibration before its deployment could be another possible explanation as pressure and temperature coefficients are determined at this stage. A possible uncorrected or incomplete calibration at this stage could induce biased derived float-pH data. Finally, another possible explanation could be related to the established at-depth correction that does not seem to yield adequate pH accuracy at the surface, at least in the subpolar North Atlantic. This uncertainty may partly be incurred by the regional complication of finding a reliable at-depth reference. Following a suggestion from Reviewer #2 (added in the revised manuscript), differences between raw minus corrected float-pH data have been calculated for winter cycles during which the MLD was deeper than 1000 dbar: differences are larger when the classical reference depth is used. Conversely, lower deviations between raw and corrected pH data are measured when the deepest reference depth is used. In this area, we believe that this speaks for a deeper reference level to corrected float-pH data as late winter cycles are more prone to be perturbed at the classical reference depth and thus could not be used to correct the entire profile. An improved understanding of the temperature and pressure effects on the sensor could be a way forward to improve float-pH data adjustment.

4. Despite limiting the data used slightly, I still think that the standard deviation is presented in an unhelpful way. You have a large standard deviation relative to the interquartile ranges and even the 95th percentiles, and as pointed out that is because of a small number of extreme outliers. This is problematic because you are actually applying some of those estimation relationships in locations where those relationships are explicitly not intended to work. Thus, the resulting standard deviation is not useful and might actually confuse readers as it did me.

We agree that we could have been stricter on removing questionable regions in our comparison for the previous revision, and have done so now by removing both Baffin Bay (little effect) and the Mediterranean Sea (noticeable reduction of the standard deviation) entirely, in addition to the High Arctic and Black Sea as done previously. In the revised manuscript, Figure 4 has been modified accordingly (Page 13).

We see the point that the reviewer makes here on the mismatch between a larger-than-expected and thus confusing standard deviation compared to expectations from the 5^{th}/95^{th} percentiles, but we also do see merit in stating exactly this mismatch. Indeed, this is because of a 'smallish' number of data areas that do not conform to the expected normal distribution behavior. However, as can be seen from the color scale on the delta pH maps (and more prominently in the modified Figure 3 below where only data outside the 5^{th}/95^{th} percentiles are retained), these "outliers" largely lie in open ocean locations (such as the North Atlantic, the Indian Ocean, or the North and Tropical Pacific) where those estimation relations are indeed intended to work, but give noticeable ('extreme' in terms of desired pH accuracy) differences. We would then argue to keep the standard deviation for the above purpose.

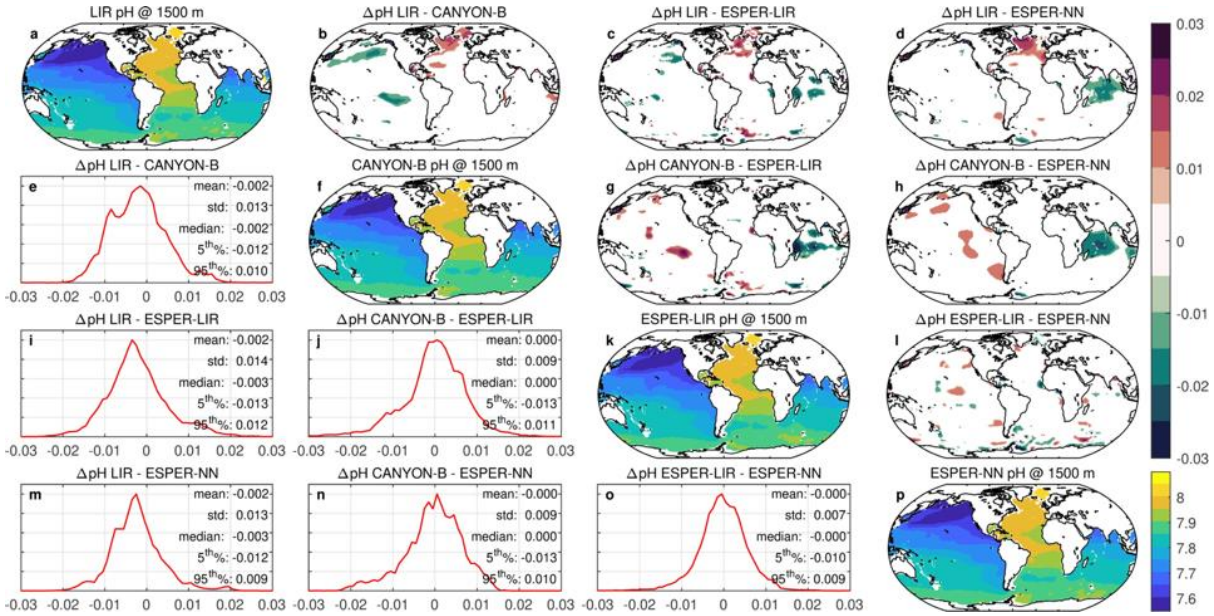

Figure 3 (Fig. 4 in the revised manuscript). Spatial distributions of estimated pH data at the classical reference depth 1500 m using different reference models: LIR-pH (with the OA adjustment) (a), CANYON-B (f), ESPER-LIR (k), and ESPER-NN (p). Maps of the spatial difference between the estimated pH datasets are presented in panels (b-d, g-h and i). Panels (e, i-j and m-o) show the bias ΔpH distribution only for data outside the 5^{th}/95^{th} percentiles (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, the Baffin Bay, the Mediterranean Sea, and the High Arctic 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).

We also would like to point out that, in response to Reviewer #2, the dataset used as the input variable is also now mentioned in the revised manuscript (World Ocean Atlas climatology). Indeed, Reviewer#2 asked a related question on the same point that has been addressed in his/her response letter.

Responses to REVIEWER 2 - Anonymous Referee

This paper uses data from two floats to point out several potential sources of bias arising from current Argo float pH calibration methods: choice of reference depth, choice of reference algorithm, fitting of the correction into straight segments with large discontinuities, and uncertainties in depth or pressure dependence of the correction. These are all important issues and so the paper makes an important contribution by demonstrating their impact on real float data in an important region for understanding ocean carbon. The paper only seeks to offer a recommendation for the treatment of breakpoints in the correction, but is still a useful contribution despite not proposing specific solutions for each issue. I have a few additional suggestions (1 new one – sorry, that I didn't notice this potential problem during my previous review! 2 other moderate ones that follow up on my previous recommendations).

We appreciate the time and effort Reviewer 2 spent on our paper. His/her thoughtful comments and suggestions have helped improve it and we would like to express our thanks and appreciation to Reviewer 2 in these over-committed nowadays. We have addressed all of the comments and included responses in italics below each reviewer's comments. 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. Line numbers correspond to the PDF file.

Moderate comments

Potential problems when using oxygen data from > 1900 dbar in calculating reference pH values: A possible issue exists in calculating the reference pH values at 1950 dbar from algorithms that use O2 data. A frequent problem exists in deep O2 data from Argo floats where the first few, deepest points of a profile are biased low, the so-called "hook". To my knowledge, the origin of this low bias is unknown, and it doesn't affect every profile. However, the authors should confirm whether this bias exists in some of their profiles and remove it before calculating reference pH. The offset shown in Figure 3 appears to be in the correct direction to be caused by 1950 dbar O2 data that is biased low, though I'm unsure at what magnitude the bias could affect the pH data given that typical hook bias is less than 10 umol-O2/kg. Also, suggest adding information about which reference estimation algorithms require O2 data. CANYON-B is mentioned for this in Line 112. However, ESPER, in particular, appears to be a family of possible algorithms depending on what input data is available, so which equation is being used should be specified.

We agree with the referee that, although Argo floats typically measure up to 2000 meters, an unlikely "hook" in the oxygen data at the deepest 50 meters trending toward low oxygen values is observed for several profiles of some Argo floats. Although the main cause is still being investigated, it has been proposed that these 'hooks' are either from optode response time or bio/particle fouling[1]. However, for the three Argo floats considered in our study (WMOs 3901668, 3901669, and 690412), such a bias has not been observed on the oxygen profiles (Figure 1). In consequence, we believe that this bias does not impact our dataset. In the revised manuscript, this statement has been added in Section 2.1.

[1]Wolf, M. K. (2017). Oxygen saturation surrounding deep-water formation events in the Labrador Sea from Argo-O2 data, (Master's thesis). Retrieved from [UVicSpace]. (https://dspace.library.uvic.ca//handle/1828/8401). Victoria, BC: University of Victoria.

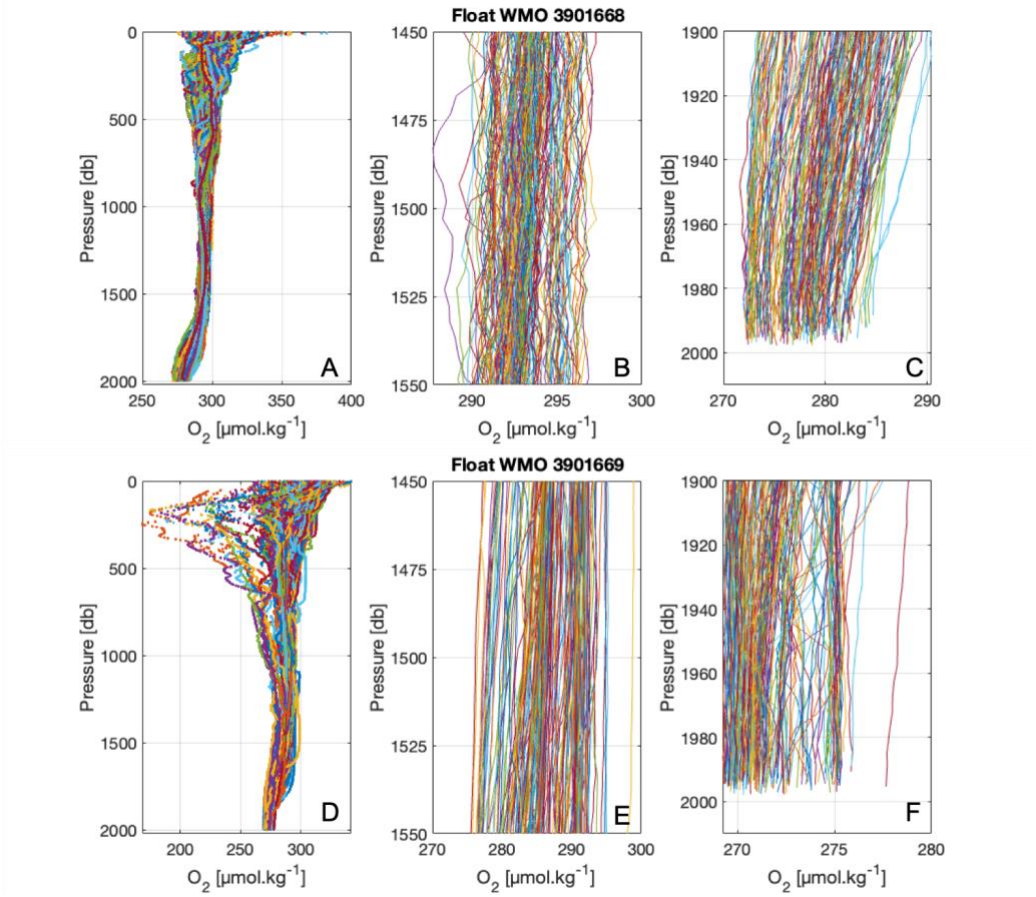

Figure 1 (not in the revised manuscript). Oxygen concentration profiles for (A & D) the entire water column, (B & E) the depth comprised between 1450 and 1550 dbar and (C & F) the bottom 1900 to 2000 db for floats 3901668 (upper panels) and 3901669 (lower panels). The low oxygen 'hook' is not visible on any profile.

In order to test the impact of an oxygen change of 10 µmol kg^{-1} on the estimated pH values, mean temperature, salinity, and oxygen values measured at 1500 db and 1950 db for the float 3901669, as an example, were used to compute CANYON-B pH values (Table 1).

Table 1 (Not of the revised manuscript). Parameter values used as inputs to determine the pH changes corresponding to an oxygen change of 10 µmol kg^{-1}. Computation has been done using the CO2SYS software.

Mean Temp. [°C]	Mean Sal.	Mean. Oxygen [µmol/kg]	Mean Depth (db]	Mean pH	Mean pH (+ 10 µmol/kg of Oxygen)	pH difference
3.3827	34.8839	277.3410	1500	7.9726	7.9784	± **0.0058**
3.3780	34.9253	271.6621	1950	7.9626	7.9678	± **0.0052**

Concerning these results, we agree with the referee that pH estimates are sensitive to other input parameters, including oxygen data. The difference between pH data estimated with oxygen values varying between ± 10 µmol kg^{-1} ranges between ±0.0052 to ±0.0058 pH units (with CANYON-B), well below the observed difference reported in Figure 3A (in the manuscript). Moreover, when the same oxygen value as the one measured at 1500 db (i.e., 277.3410 µmol kg^{-1}) is used to derive pH data using the mean temperature and salinity values recorded at 1950 db, a pH value of 7.9655 is obtained, highlighting the high sensitivity of this reference algorithm to other input parameters. In this regard, we

would like to point out that the choice of a deeper reference depth tends to use data from a stable and unperturbed reference depth. By considering a negative bias for the oxygen data measured at 1950 db, as it could happen in a "hooked" oxygen data situation, the difference between pH data corrected at 1500 db minus float-pH data corrected at 1950 + 10 µmol kg^{-1} is equal to 0.0048 pH units (i.e., 7.9726 - 7.9678), still below the difference reported on Figure 3A. In consequence, we do not believe that the difference reported between the two corrected datasets could rely on biased oxygen-deep values. Nevertheless, the oxygen-uncertainty impact on the derived pH values is discussed in the revised manuscript.

"When O_2 sensors incapable of in-air referencing are used (e.g., SBE63 optode, Sea-Bird Electronics), oxygen values typically have uncertainties up to ca. 3% (Takeshita et al., 2013), adding an additional source of uncertainty when these data are used as input parameters to derive reference-pH data." (L.116)

Finally, we thank the referee for pointing out that algorithm descriptions were incomplete in the original manuscript concerning oxygen data utilization. In this study, all the reference algorithms used employ oxygen values as ancillary data (ESPER methods and LIR regressions #7; Carter et al., 2018, 2021). In the revised manuscript, details have been added.

"*Although some Argo float profiles could be impacted by a "hook" in the oxygen data at the deepest 50 meters inducing low oxygen values (Wolf et al., 2018), a visual inspection of the oxygen profiles from these three floats has been performed and does not point toward such a bias.*" (L. 134)

"In this study, oxygen data were used as predictor variables in all reference algorithms used." (L.112)

Impact of convection depth on the 1500 dbar correction: In lines 238-246, the authors imply that deeper convection depth in this region is responsible for the offsets between the 1500 dbar reference depth and 1950 dbar reference depth corrections. I think the authors have the data to provide stronger evidence for this implication. If deep convection depths are responsible, late winter cycles where mixed layer is deep should show larger deviations or variability between the raw and reference pH for 1500 dbar. Is that the case for these floats?

In the studied region, deep convection events, water mass formation as well as decadal variability affect water masses at a depth greater than 1500 dbar. By comparing float-pH data corrected at two different depths (i.e., 1500 and 1950 db), this paper highlights that the choice of an arbitrarily chosen depth around 1500 dbar induces an uncertainty of at least 0.005 pH units. Arguing that this result highlights the implication of deep convection events on the current quality control procedure, stronger evidence arises when a comparison between raw pH data minus corrected float-pH data (at the two reference depths) is done (Table 2), as suggested by Reviewer #2. Indeed, when only winter cycles during which the MLD was deeper than 1000 dbar are considered, differences between raw minus corrected float-pH data are larger when the classical reference depth is used. Conversely, lower deviations between raw and corrected pH data are measured when the deepest reference depth is used. In this area, we believe that this speaks for a deeper reference level to corrected float-pH data as late winter cycles are more prone to be perturbed at the classical reference depth and thus could not be used to correct the entire profile. Following the referee's comment, a new Table (Table 2 below) and additional explanations have been added in the Supplementary Material and Section 3.1.1 of the revised manuscript, respectively.

"*By splitting the dataset to keep only profiles done when the MLD was deeper than 1000 dbar, the comparison between raw and corrected float-pH data using the two reference pressures reveals larger variabilities when the classical reference depth is used rather than when the deepest one is considered, highlighting the implication of deep convection events on the adjustment method (Table A1)*. " (L. 248)

Table 2 (Table A1 in the Supplementary Material of the revised manuscript). Mean absolute differences between float-pH data corrected at two distinct depths and using the four different reference methods for the floats WMO 3901668 and 3901669. SD stands for Standard Deviation. Only profiles performed when the MLD was deeper than 1000 dbar were used.

		Float WMO 3901668		Float WMO 3901669
		Winter 2019	*Winter 2020*	*Winter 2019*
		Mean MLD=1639.6 db	*Mean MLD=1712.1 db*	*Mean MLD=1240.2 db*
ESPER-NN	*Raw-1500 db*	$-0.0293 \pm 1.39\times10^{-4}$	$-0.0248 \pm 8.12\times10^{-5}$	$-0.0352 \pm 9.97\times10^{-5}$
	Raw-1950 db	$-0.0181 \pm 5.23\times10^{-4}$	$-0.0164 \pm 1.08\times10^{-4}$	$-0.0265 \pm 5.95\times10^{-5}$
ESPER-LIR	*Raw-1500 db*	$-0.0416 \pm 3.52\times10^{-5}$	$-0.0380 \pm 1.64\times10^{-4}$	$-0.0456 \pm 1.05\times10^{-4}$
	Raw-1950 db	$-0.0244 \pm 3.30\times10^{-3}$	$-0.0300 \pm 1.80\times10^{-3}$	$-0.0307 \pm 1.77\times10^{-4}$
CANYON-B	*Raw-1500 db*	$-0.0508 \pm 8.89\times10^{-5}$	$-0.0478 \pm 1.41\times10^{-4}$	$-0.0554 \pm 6.49\times10^{-5}$
	Raw-1950 db	$-0.0391 \pm 1.50\times10^{-3}$	$-0.0399 \pm 8.88\times10^{-4}$	$-0.0457 \pm 6.20\times10^{-5}$
LIR-pH	*Raw-1500 db*	$-0.0677 \pm 4.31\times10^{-5}$	$-0.0647 \pm 3.95\times10^{-5}$	$-0.0728 \pm 8.51\times10^{-5}$
	Raw-1950 db	$-0.0549 \pm 4.03\times10^{-4}$	$-0.0543 \pm 2.77\times10^{-4}$	$-0.0628 \pm 6.18\times10^{-5}$

Showing raw and reference values on the same panel: I'm glad to see the information in Figure 5 brought together, but I still think an additional panel or two would clarify the extent to which the breakpoints are caused by sudden changes in raw pH vs. potential discontinuities in the reference pH possibly from spatial patterns as the float moves to different regions. I suggest a panel showing the raw 1500 dbar pH and the reference pH from one or more of the algorithms. Two different scales (like the authors have used in panel c) would allow the signals to be at sufficient vertical resolution. The difference between the raw and corrected is interesting but doesn't provide this crucial information about how stable the reference pH itself is over the spatial movement and timing of the float. In particular, panel c suggests that the breakpoints are either not aligned with discontinuities in the raw data or that the raw data discontinuities at the parking depth are poorly related to those at 1500 dbar. Either way, this would be useful to explore further.

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 (Fig. 5C). 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 leading to step-like changes with the SAGE method appear to be unrealistic (ex. in July 2018). In our view, the sensor rather shows undulations in response with smooth and less smooth phases as it is somehow confirmed by Figure 5C as well as in Figures 5E and F showing raw float-pH data at 1500 dbar.

Indeed, 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 phases and before the application of the pH adjustment procedure. Such smooth 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). 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 doesn't present noteworthy jumps. Nevertheless, we thank the referee for suggesting modifying Figure 5 to better discuss the wiggles around the jumps and their possible link to variability in the algorithm estimates or short-term sensor changes. Indeed, high variability is observed on the reference pH time-series estimated using both CANYON-B (Fig. 5E) or ESPER-LIR (Fig. 5F), highlighting the noticeable impact of the algorithm estimate discontinuities on the final correction while raw float-pH data are not presenting sudden changes. In the revised manuscript, Figure 5 has been modified and some information added to better clarify the impact of reference methods on the breakpoint determination.

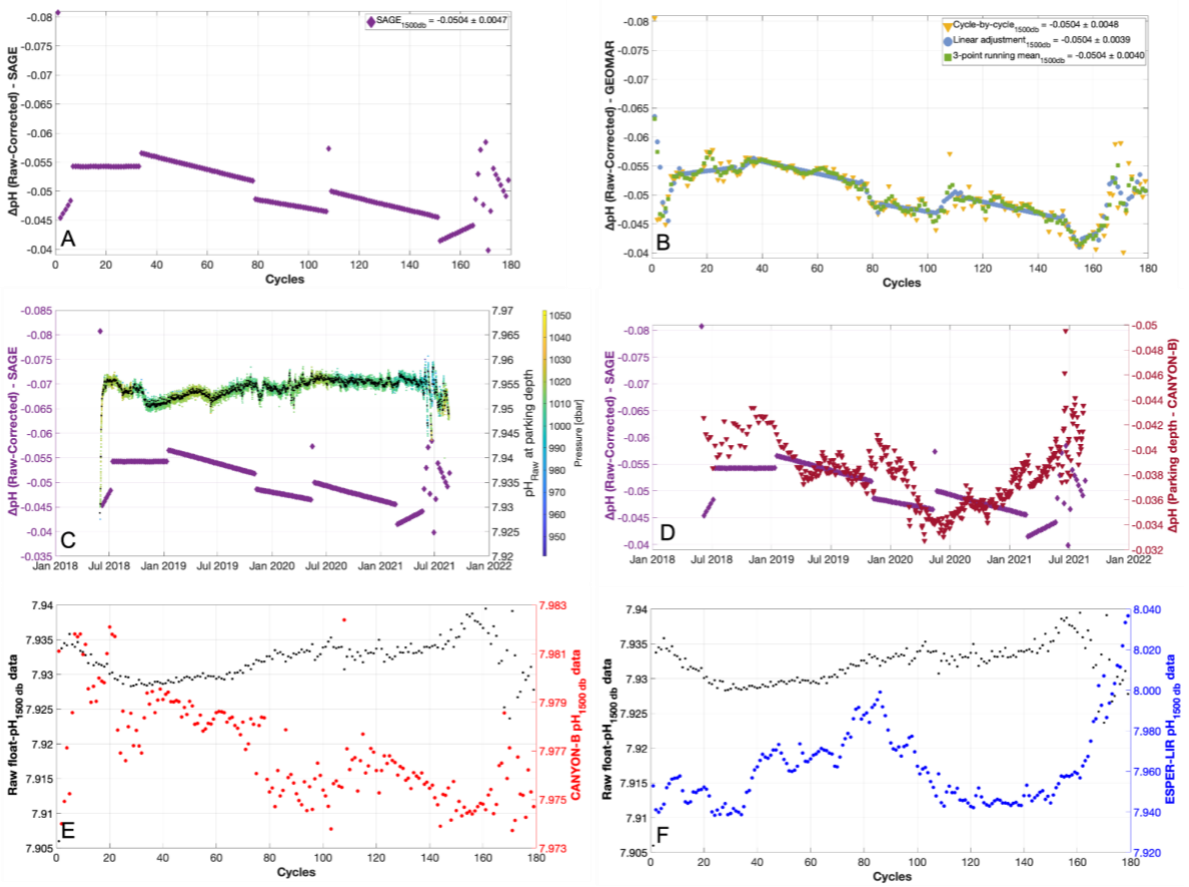

Figure 3 (Figure 5 in the revised manuscript). Differences between raw float-pH data minus float-pH corrected using the SAGE tool (Fig.5A), the cycle-by-cycle GEOMAR method (yellow dots, Fig.5B), and the linear mean regression GEOMAR method (blue dots, Fig.5B) and the 3-point centered running mean correction 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). Figures 5E and F show mean raw float-pH data measured at 1500 dbar (between 1480 and 1520 dbar) and pH data calculated by the reference methods CANYON-B (panel E) and ESPER-LIR (panel F) using as input parameters (i.e. temperature, salinity, pressure and oxygen) the values measured

by the float at 1500 dbar. For panels A to D, differences are calculated for each cycle at each depth along the entire profile and then averaged.

"*Moreover, high variability is observed on the reference pH time-series estimated using both CANYON-B (Fig. 5E) or ESPER-LIR (Fig. 5F), highlighting the noticeable impact of the reference algorithms discontinuities on the final correction while raw float-pH data are not presenting sudden changes. Indeed, the raw pH time-series shows smoothed transitions and the general pattern doesn't present noteworthy jumps.*" (L. *334*)

Minor comments

Lines 267-268: Suggest giving the mean and standard deviation of this correction at 1500 dbar for this dataset.

This information has been added in the revised manuscript as follows: "*For the two floats considered in this Section, means and standard deviations of the difference between float-pH data corrected at 1500 dbar using CANYON-B and CANYON-B adjusted are equal to 0.0055 ± 6.63×10^{-5} and 0.0055 ± 8.31×10^{-5}, respectively*." (L.278)

Line 273: Suggest briefly stating the dataset used as the input variables in the algorithms to generate Figure 4. World Ocean Atlas perhaps?

World Ocean Atlas climatology data was used to create the maps and comparisons of Figure 4. This information has been added to the revised manuscript. It was chosen to cover a reasonable data space like being encountered by profiling floats globally. While individual float profiles may be slightly more accurate examples of reality, their distribution is spotty and doesn't provide the same global coverage as climatology. Besides, differences in algorithms are not caused by climatological vs. real profile input data, but by differences in the algorithms' training data.

Here in the algorithm training data, LIR, CANYON-B, and ESPER show some differences, which could be sources for some of the more extreme differences seen in Figure 4. Both LIR and CANYON-B use the original release of GLODAPv2, however with a different treatment of anthropogenic carbon as well as on pH (from different methods, spectrophotometrically measured or calculated from other CO2 parameters). One could speculate that this is a source for noticeable differences between LIR and CANYON-B (and LIR and ESPER) in convective areas of the North Atlantic. ESPER, in contrast, uses an updated version of GLODAPv2 (GLODAPv2.2020), which includes some modifications on 90's CO_2 data in the Indian Ocean. This is a likely cause for the noticeable differences seen between LIR/CANYON-B and ESPER in the Indian Ocean.

Table 2: It's unclear to me why absolute difference is shown here. The sign of the offset seems relevant, and I suggest including it.

Table 3 has been modified in the revised manuscript and the sign of the offset has been added.

y-x		*1500 db*				*1950 db*			
		LIR-pH	**CANYON-B**	**ESPER-LIR**	**ESPER-NN**	**LIR-pH**	**CANYON-B**	**ESPER-LIR**	**ESPER-NN**
WMO 3901668	**LIR-pH**	/	0.0175 ± 0.0012	0.0264 ± 0.0026	0.0407 ± 0.0028	/	0.0155 ± 0.0011	0.0259 ± 0.0052	0.0399 ± 00027
	CANYON-B	-0.0175 ± 0.0012	/	0.0089 ± 0.0016	0.0232 ± 0.0019	-0.0155 ± 0.0011	/	0.0105 ± 0.0045	0.0245 ± 0.0024
	ESPER-LIR	-0.0264 ± 0.0026	-0.0089 ± 0.0016	/	0.0143 ± 0.0013	-0.0259 ± 0.0052	-0.0105 ± 0.0045	/	0.0140 ± 0.0053
	ESPER-NN	-0.0407 ± 0.0028	-0.0232 ± 0.00196	-0.0143 ± 0.0013	/	-0.0399 ± 0.0027	-0.0245 ± 0.0024	-0.0140 ± 0.0053	/
WMO 3901669	**LIR-pH**	/	0.0173 ± 0.0014	0.0321 ± 0.0069	0.0391 ± 0.0021	/	0.0161 ± 0.0017	0.0381 ± 0.0076	0.0356 ± 0.0018
	CANYON-B	-0.0173 ± 0.0014	/	0.0148 ± 0.0058	0.0217 ± 0.0014	-0.0161 ± 0.0017	/	0.0221 ± 0.0070	0.0196 ± 0.0014
	ESPER-LIR	-0.0321 ± 0.0069	-0.0148 ± 0.0058	/	0.0069 ± 0.0055	-0.0381 ± 0.0076	-0.0221 ± 0.0070	/	-0.0025 ± 0.0074
	ESPER-NN	-0.0391 ± 0.0021	-0.0216 ± 0.0021	-0.0069 ± 0.0055	/	-0.0356 ± 0.0018	-0.0196 ± 0.0014	0.0025 ± 0.0074	/

Table 3 (Table 2 in the revised manuscript). Mean differences (y-x) between float-pH data corrected at two distinct depths and using the four different reference methods for the floats WMO 3901668 and 3901669. SD stands for Standard Deviation.

Figure 5: The y-axis labels in this figure suggest that they show the absolute value of the raw – reference pH. Here too, I think the sign is important and the y-axis should be just the difference, not the absolute value. The caption is unclear what "mean" difference means. Given that a time series is shown, it's not the mean over the time series.

The y-axis labels have been modified in the revised Figure 5 (see Figure 3 above) and now show raw minus corrected data. In this Figure, we have plotted the mean difference per cycle, i.e. the entire raw profile minus the entire corrected profile is meant. In the revised manuscript, the caption has been clarified.

"[...] For panels A to D, differences are calculated for each cycle at each depth along the entire profile and then averaged." (Page 15)