Response to review by referee #4, Dr. Nancy Williams

We thank Dr. Williams for the helpful comments and suggestions, each one is adressed below (comment in black, response in red).

General Comments:

This is an update to the GLODAPv2.2019 by adding 106 new cruises from 2004-2019, expanding the coverage of GLODAP to 946 cruises over 47 years, 1972–2019. Most of the new cruises are from the western North Pacific and the Davis Strait, with a few from the Atlantic, South Indian, and U.S. West coast. The methods for primary and secondary quality control (QC) are essentially the same as in the earlier version. However, there has been no full consistency analysis of the entire data product as was done with the original GLODAPv2 product. A full consistency analysis will be performed in the future for the next GLODAP update (will be termed "GLODAPv3") which is set to occur after the completion of the third GO-SHIP survey around year 2023. The researchers have also fixed some minor errors in the GLODAPv2.2019 dataset.

Throughout the manuscripts the researchers discuss alternate ways of adjusting the dataset, and tend to take a conservative approach, saving any major changes for the next full GLODAP update, i.e., GLODAPv3. As such, this update could be considered by some to be incremental, but it should be noted that incremental and timely updates to GLODAP are critical to advancing ocean observing. GLODAP, and other such data products that have come before it, forms the backbone for studying largescale changes in water column properties and has also become increasingly important as autonomous platforms and sensors rapidly begin to fill the world's oceans. Many autonomous biogeochemical sensors are prone to drift and rely on GLODAP data and methods such as linearly interpolated regressions (LIRs; Carter et al. (2016, 2018) or machine-learning methods such as CANYON/CONTENT (Bittig et al., 2018, Sauzède et al. 2017) for ongoing quality control after deployment. GLODAP also serves as a benchmark for background concentrations in ocean and earth system models.

Where available, the researchers have also added isotopic data for _13C, _18O, and **D**14C which are not quality controlled/adjusted in the same way as the core GLODAP variables but can provide context for the other data.

They have also added discrete fCO2 values which will be useful in addressing inconsistencies in the carbonate system variables. Importantly, fCO2 has not been subjected to any secondary QC. There has also been more extensive use of CANYON-B and CONTENT predictions to evaluate offsets in nutrients and CO2 data.

One important change that has been made to this version is that there is no internal consistency evaluation of seawater CO2 chemistry variables to evaluate pH. This leads to an inconsistency between the pH data for cruises added in this version, and pH data in previous versions of GLODAP. My understanding is that this will likely manifest as a bias, and not a random uncertainty. This potential bias is indeed encompassed by the stated consistency of "0.01 to 0.02 pH units," but will be critically important for those using this dataset and should be explained more clearly earlier in the manuscript, and perhaps even in the abstract. I also do not think that the consistency for pH should be stated as a range. Yes, it varies by region but unless each region/cruise/data point has its own uncertainty estimate, the overall consistency should be stated as ± 0.02 pH units. If it is the case that there is only

one region where the consistency is ± 0.02 pH units, and the rest of the ocean is closer to ± 0.01, then that region should be explicitly defined.
Indeed, no internal consistency evaluation was conducted for pH for the data added in this version. No pH data were adjusted either. If adjustments had been made, they would adjust the data from the new cruises, to the pH values of cruises already part of GLODAP (which are used as reference) and evaluated in the earlier efforts. As such, this would not have led to inconsistencies between the pH data for cruises added in this version, and pH data in previous versions of GLODAP.

Regarding stating the consistency for pH as a range. We agree that this was somewhat murky in the submitted manuscript, and we now provide clearer reasoning and identify regions of high vs low uncertainty.
Changes made: The final paragraph of section 3.2.4, where these issues were discussed, have been substantially expanded, to: "In contrast to past GLODAP pH QC, evaluation of the internal consistency of CO_2 system variables was not used for the secondary quality control of the pH data of the 106 new cruises; only crossover analysis was used as supplemented by CONTENT and CANYON-B (Sect. 3.2.5). Recent literature has demonstrated that internal consistency evaluation procedures are subject to errors owing to incomplete understanding of the thermodynamic constants, major ion concentrations, measurement biases, and potential contribution of organic compounds or other unknown protolytes to alkalinity (Takeshita et al., 2020), which lead to pH dependent offsets in calculated pH (Álvarez et al., 2020; Carter et al., 2018): these may be interpreted as biases and generate false corrections. The offsets are particularly strong at pH levels below 7.7, when calculated and measured pH are different by on average between 0.01 and 0.02 units. For the North Pacific this is a problem as pH values below 7.7 can occur at the depths interrogated during the QC (>1500 dbar for this region, Olsen et al., 2016). Since any corrections, which may thus be an artifact, are applied to the full profiles, we assign an uncertainty of 0.02 to the North Pacific pH data in the merged product files. Elsewhere, the uncertainties that have arisen are smaller, since deep pH is typically larger than 7.7 (Lauvset et al., 2020), and at such levels the difference between calculated and measured pH is less than 0.01 on average (Álvarez et al., 2020; Carter et al., 2018). Outside the North Pacific, we believe, therefore that the pH data are consistent to 0.01. Avoiding interconsistency considerations for these intermediate products helps to reduce the problem, but since the reference data set (also as used for the generation of the CONTENT and CANYON-B algorithms) has these issues, a full re-evaluation, envisioned for GLODAPv3, is needed to address the problem satisfactorily. "

The original and adjusted data, a detailed adjustment table, and a "known issues" document are available online at the links provided in several formats, and as both global and regional subsets. The "known issues document" is updated regularly and users are encouraged to consult that document when using the data products and identify new issues when they find them.

I was also expecting to hear if/when the next GLODAP gridded product will be produced. Will it always only come with "major" GLODAP updates or are there any plans to do incremental updates?
There are no plans for making incremental updates to the GLODAP gridded product. The changes would likely be rather small anyhow, as the main source of uncertainty in the gridded product is lack of observations in certain regions. The data added in GLODAPv2.2019 and GLODAPv2.2020 are mostly repeat observations, extending the coverage in time and not in space. We cannot commit, now, to making new climatologies for the next full update. While we hope it will be possible, it will depend on the funding situation. Therefore, we

simply add a statement that the intermediate products are not accompanied by a gridded product update.

- Changes made: The sentence "Additionally, the GLODAP mapped climatologies (Lauvset et al., 2016) are not updated for these intermediate products." has been included in the second final paragraph of the introduction.

Specific comments:

Line 249: An adjustment of -3 μmol/kg is made for a cruise which has a mean offset of 3.68 μmol/kg. Are adjustments always whole numbers? If so, do you always round down?

Adjustments are typically round numbers relative to the precision of the variable considered. There are no particular rules about rounding down or up; we look for example, on whether there is a difference in the offset in recent vs older crossovers. We also consider additional evidence from the other methods. Here, we settled for -3 μmol/kg, as the CANYON-B and CONTENT analyses suggested a bias of 3.4 and 2.7 μmol kg^{-1}, respectively. This also helps to make the adjustment as small as meaningfully possible, in case there actually is an increasing trend in TCO_2 from uptake of anthropogenic carbon.

Changes made: The sentence in question has been revised to : "In this case -3 μmol kg^{-1} was applied: this is somewhat less than indicated by the crossover analysis, but a smaller adjustment is supported by the CANYON-B and CONTENT results (Sect. 3.2.3). Adjustments are typically round numbers relative to the precision of the variable being considered (e.g., -3 not -3.4 for TCO_2 and 0.005 not 0.0047 for pH) to avoid the communicating that the ideal adjustments are known to high precision."

Line 251: Because they are an exception, provide more detail about how these eight Japanese Sea cruises were adjusted.

Changes made: The following paragraph has been added in section 4.2: "For the Sea of Japan cruises, (where two existed in GLODAPv2.2019 and six were added in this version - Sect. 3.2.2), the crossover results showed biased TCO_2 data for one of the older cruises (49HS20081021, which is now adjusted up by 6 μmol kg^{-1}), and biased TAlk data for two of the presently added cruises (49UF20111004 and 49UF20121024, adjusted up by 5 and 6 μmol kg^{-1}, respectively)."

Line 319-320: Needs editing for clarity.

This has now been edited for clarity, and we have included an example as well, following a suggestion by reviewer 2.

- Changes made: The text has been revised to: "Another advantage of CANYON-B and CONTENT is that these procedures provide estimates at the level of individual data points, e.g., individual pH values are determined for every sampling location and depth were T, S, and O_2 data are available. Cases of strong differences between measured and estimated values are always examined. This has helped to identify primary QC issues (outliers) for some variables and cruises, for example a case of an inverted pH profile at cruise 32PO20130829, which has been amended."

Lines 280-282: While it is stated that TAlk estimated from 67 times salinity is sufficient for such pH conversions, it would be useful to explicitly state the amount of uncertainty introduced to pH by such a TAlk approximation.

Yes, we agree.

- Changes made: The following text has been added in Sect. 3.2.4: "The uncertainties introduced with this approximation are negligible (order 10^{-7} pH units) for the scale conversions and order 10^{-3} pH units for the temperature and pressure conversion (evaluated by repeating conversions with 2 times the standard deviation of the ratio,

i.e., 67 ± 4.1). This is sufficiently accurate relative to other sources of uncertainty, which are discussed below."

Lines 427-429: Why was this decision made to replace measured values with calculated values?

This decision was made when GLODAPv2 was prepared. Often, for such cruises where the number of measured data points for a CO_2 chemistry variable is much less than the number that can be calculated, the accuracy of the measured data cannot be confidently established – there are too few data for good crossover analyses – and it makes most sense replacing these with values calculated from the two other better QC'd variables. Evaluating the approprioate action on a per cruise basis is time consuming, so we made the decision to draw the line at less than 1/3 (of the combined number of calculated and measured values)

- Changes made: We have simplified the sentences a bit, and added the reason for replacing measured values "For calculations involving TCO_2, TAlk, and pH, if less than a third of the total number of values, measured and calculated combined, for a specific cruise were measured, then all these were replaced by calculated values. The reason for this, is that secondary QC of the few measured values was often not possible in such cases, for example due to a limited number of deep data avaliable"

Lines 537-541 and 558-559: It is acknowledged twice in the summary that the surface data are both seasonally biased and not examined for consistency in GLODAP. This is an important caveat and should be stated in the introduction.

- Changes made: We have added the following sentence to the introduction (in former line # 98: "The secondary quality controlled focused on deep data, where natural variability is mimimal"

Figures 3, 5, 8, 10: Include a legend for the colors

Figure 3 and 5 are produced by the crossover and CANYON-B/CONTENT software. It is not possible to add legends at this stage. The meaning of the colors are now explained in the caption.

- Changes made: Legends have been added to Fig. 8 and Fig 10.