Responses to Reviewers' comments

Title: Coastal Ocean Data Analysis Product in North America (CODAP-NA) - An internally consistent data product for discrete inorganic carbon, oxygen, and nutrients on the U.S. North American ocean margins.

By Li-Qing Jiang et al.

Reference: essd-2020-402-manuscript-version3

We want to start by thanking the Reviewers for their very constructive comments! Please see below for our responses to the comments.

Reviewer One's comments

General Comments:

The article describes the basic elements of a data product that brings together discrete measurements of the marine carbonate system from 61 surveys conducted in different continental coastal areas of the USA covering the Atlantic, Pacific and Arctic coasts. It brings together 14 variables (temperature, salinity, dissolved oxygen, nitrate, nitrite, phosphate and silicate, DIC, alkalinity, pH and derived carbonate system parameters with a special focus on ocean acidification. The article is well written and easy to follow. It describes in detail the process of compiling the database, the variables including their definition and the flagging criteria. The process of completing the carbonate system variables by computing the complete set of variables from one or two of the measured variables is well also described.

The authors go into great length describing the importance of the internal-consistency (IC) and Quality Control (QC). They conclude that, logically, with relatively shallow sampling, there are no tools available to do solid QC, contrary to what happens in open ocean cruises. With the added difficulty of the influence of strong contributions from river waters with different geochemical characteristics of riverine water composition. The selection criterion has been based on the fact that the laboratories participating in the cruises use known quality assurance practices.

Many thanks for the great summary of this work.

After it has been reported that QC2 is not possible (Line 185 'secondary QC was not conducted for this version of the CODAP-NA and no cruise-wide offsets or multiplicative adjustments were applied') because there are not deep enough stations on most of the

cruises, it is expected to detail how to circumvent this on the basis of a good QC1. However, it does not specifically detail the modifications made to each of the 61 cruises during QC1. The steps taken are indicated, but no graphical information is given for many of them, such as those described in step four identify outliers.

Response: An Excel spreadsheet listing all of the QC related changes is now available as part of the data package. Link: https://www.ncei.noaa.gov/data/oceans/ncei/ocads/data/0219960/Table_QC_changes/.

As you can see, we made a total of over 22,601 QC related change. We've also added new Figures 8-11 to indicate the consistency of some of the parameters.

In addition, the details of QC1 are left to another article "These tools will be made available to the public soon, with a separate paper dedicated to their rationales, development details, and instructions (Jiang et al., in prep.)." In my opinion these tools should be included here because it is a key tool to validate the QC of the data set.

Responses: Unfortunately, the tools at its current stage are not yet ready to be shared. We have a two-year funded work plan to finalize these tools. We need to tap the expertise from experienced tool builders, modularize a lot of the functions, clean up the code, and make it a set of mature tools that are releasable and useful to the community. However, the important thing for readers to understand here is what precisely was done to quality control the data. We've added a new figure (Figure 2) showing the various quality control steps and the plots that were examined during the CODAP-NA QC process.

Furthermore, specifically for the IC of the carbonate system nothing is shown. Many carbonate system data have been generated, e.g. pH and fCO2, from one or two measured variables. However, I strongly recommend to show in those cruises where more than two variables have been measured the internal consistency between measured and computed values. It is needed to remember that reference materials are only available for DIC and alkalinity so for the other variables it is necessary to include another criterion on the quality of the measured pH and carbonate ion data. All this makes me express my doubts whether the current manuscript meets sufficient requirements to reach the threshold sufficient to be an article in ESSD.

Responses: We've added 3 new figures (Figures 9-11) to the paper to show the internal consistency of pH, $[CO_3^{2-}]$, and fCO$_2$.

Minor details

311 'Observation type' is not listed in the Table 2

Responses: The observation type, i.e., whether the samples are collected using Niskin bottles (Niskin) or Flow-through pumps (FT) has been added to Table 2.

315.- What changes in quality assessment has this QC brought about?

Responses: It is just a double confirmation to make sure no outliers were left out.

381 Some percentages do not really provide useful information such as CTDSAL and pH. They can even be misleading.

Responses: The percentage numbers for CTDSAL and pH have been removed.

386 'Frequency' is actually N of group of stations. This can be interpreted by the reader as a percentage of the total number of samples. Please add (N) such as 'Frequency (N)' and inform of that in the legend.

Responses: We fixed this for Figures 6 and 7.

404.- Idem for Figure 6.

Responses: Same as above, we have changed the label from "Frequency" to "Frequency (N)".

415.- Table 7. The high percentage for Nitrite and Ammonium is nonsense because these variables in so deep water usually are practically below the limit of detection.

Responses: We agree with the reviewer on that statement and just want to leave the numbers there as a reference to give the readers some idea about how much they should expect the uncertainties of these measurements.

423.- Please give the overestimation in pH or in Carbonate ion concentration to be fair that is the main AO variables. The error of fCO2 should be given in % or in logarithm as it is done for pH. In that case, it would not be such a striking value. That is why I think it is unbalanced to give only the bias value for fCO2 because it is high, and not to give it for pH and carbonate ion, which surely do not have such a high and significant apparent bias.

Responses: We did the calculations ourselves using the global surface ocean average Salinity, DIC, and TA of 34.87, 2020 umol/kg and 2306 umol/kg, respectively. The reviewer is correct. Our results show that the differences are a lot smaller than what the authors of the paper claimed. Even under a very low temperature of -4 °C, the $f\text{CO}_2$ change is only about 8uatm (6%). The pH change is 0.036 units and Carbonate ion change is 6.3 umol/kg. The original seemingly large numbers have been removed from this paper.

589: Please check the title of this article you authored.

Responses: Fixed this reference. Many thanks for catching that.

Merged data product

I have downloaded the dataset and visualized the data by performing X-Y plots to roughly inspect the flagging. I have mainly focused on visualizing the internal consistency between measured and calculated pH, and the same for O2, fCO2 and carbonate ion. The differences between measured and calculated pH showed a set of 240 data with deviations ten times (0.05) the nominal pH accuracy (0.005). This suggests to me that QC1's task has been very weak. For carbonate ion, I note that 1200 samples show values that deviate from twice the carbonate ion measurement error (about 2 micromol/kg) with only about 30 showing deviations greater than 20 micromol/kg, all of them with flag=2. Similarly, 124 samples show a deviation of more than 3% from the calculated value of fCO2. Even for Oxygen there are more than 1300 samples with differences between the oxygen measured chemically (Winkler) and measured with the CTD greater than 4 micromoles/kg, with 63 showing differences greater than 20 micromoles/kg. All of them with flag=2, which suggests to me that the QC1 performed is very, very undemanding. This is really important because as the authors suggest this dataset would be a reference product to be used for QC2 of future cruises.

Responses: As we mentioned above, we made over 16 thousand QC related changes during the CODAP-NA product development and caught a number of errors with our QC process and data ingestion during this round of revisions and we thank the reviewer for spotting these oversights. We'll note that some of the "outliers" are surface samples where the Niskin vs. CTD values are offset due to highly stratified surface conditions. In these cases, we believe most Winkler and CTD values are likely "good" data, and thus decide to keep the QC flags to be "2".

Reviewer Two's comments

General Comments:

This appears to be a useful new data set/data product, gathering regional data for the coastal areas along the North American continent. It is the kind of data that is not included in the global GLODAP data product (which contains similar kind of data) but the way of QC is very similar to that of GLODAP. It must be said that the inclusion of data seems somewhat arbitrary, where the authors (who are mainly also the data providers) do not provide clear criteria and only refer to known high quality data. It is also not clear which criteria will be valid for future additions to the data product.

Title: It contains an error. Only the U.S. North American margins are mentioned in it. However, the data product includes all coasts, those from Canada and Mexico as well. If the authors want to emphasize that this is a U.S. effort, then they must change the title.

Response: We have removed the word "U.S." from the title, and the rest of the manuscript as well. We initially used the word "U.S", because these are U.S. led research studies. That said, we agree with the Reviewer in that here the focus should be about where the sampling locations are located.

The authors explain their QC and information about the data and data product. What they did not provide is the number or percentage of bad data and which of their methods was most successful in spotting the bad data. The number of bad data or discarded data would be a measure to assess by the reader whether their initial judgement of reliable data was fine. For future QC, it is important to know which method is most successful.

Response: An Excel spreadsheet listing all of the QC related changes is now available as part of the data package. Link;
https://www.ncei.noaa.gov/data/oceans/ncei/ocads/data/0219960/Table_QC_changes/

As you can see, we made a total of over 22,601 QC related change. We also added a figure showing the QC procedures (Figure 2). The most effective approach is internal consistency checks. We have modified the text to reflect that. See Line 311.

L47 I am not sure that every reader knows what secondary QC is. Clearly, this is explained in the text but here in the abstract it is an unknown.

Response: We have changed "secondary QC" to "cruise to cruise comparisons" in the abstract.

L47-48 "We worked closely with the investigators who collected and measured these data during the QC process." The data originators are the co-authors, aren't they?

Response: Yes, most of them are either co-authors or mentioned in the Acknowledgements.

L63-64 "Despite only covering ~20% of Earth's land surface, coastal regions (from the coastline up to 200 km inland) host over 50% of the entire human population (Small and Nicholls, 2003; Hugo, 2011; Neumann et al., 2015)." This info comes out of the blue at this place. I think this info is not necessary here.

Response: The sentence has been deleted.

L78-79 "where most of the global fisheries and aquaculture industries are focused." This was already mentioned earlier. Can be deleted here.

Response: The sentence has been deleted.

Figure 1 and text in sections 1 and 2: I am surprised to read only of the US east and west coast, as the data are also from the continental shelves of Canada and Mexico. Mentioning those would be appropriate.

Response: We have removed the wording "U.S" in most of the cases. In some occasions, it was replaced with "North American".

L146 "known quality" What are the criteria for this known quality and known by whom? This does not sound objective. To change this, an explanation would be useful.

Response: We agree with the Reviewer in the lack of an objective criteria. This is very hard to quantify. Good thing is that this group of researchers know the labs who measure ocean carbon data in this region well. We explained this in the following sentence by saying these are data either collected by either AOML, PMEL or labs using their technology and quality assurance.

L146, 147 These abbreviations are not used below.

Response: The acronyms have been removed.

Table 1 Start data and End data: Is that the dates for which there are data, or the dates of the cruise? The start date is different from the date in the expocode in some cases, so I guess it is the latter. Please explain in the table title.

Response: A new sentence has been added to the Caption of Table 1. "Start date and End date refer to the dates when data were first and last collected, respectively."

Table 1 and Table 2 There is a lot of info in Table 1 that can only be understood after reading Table 2. It would be appropriate to swap these Tables. Actually the whole Sections 4 and 3 could also be swapped.

Response: The entire Section 3 has been swapped with Section 4. As a result, Table 1 has been swapped with Table 2.

L175-179 This text is written as if it is nice to have. It does not become clear whether the authors have applied this procedure in this data product.

Response: We have changed the wording slightly to make it clear that it is not optional. In our experiences, we identified a large amount of data points that need to be flagged or improved, as shown in the new Supplementary Table 1.

L184 I suggest: … where parameters were not likely to be influenced by temporal variations … (because it cannot be excluded that there is temporal variation at these greater depths). L184-185 "Due to the scarcity of cross-over stations at depths where parameters were not influenced by temporal variations (sampling depth >1500 m, Olsen et al., 2020) on coastal cruises, secondary QC was not conducted for this version of the CODAP-NA" This is not correct. Later in the manuscript such an analysis is described, even though not for all cruises. Please modify the text accordingly.

Response: The wording "likely to be" has been added to the sentence. Please check out the new Line 186.

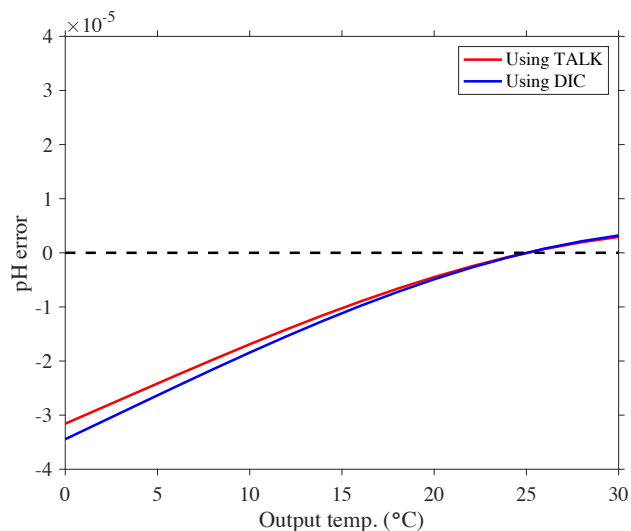
L194-196 "A new suite of QC tools was developed by this team of authors to satisfy the requirements of enhanced consistency checks. These tools will be made available to the public soon, with a separate paper dedicated to their rationales, development details, and instructions" This seems to be the other way around. For the reader and data user to judge whether the methods used are solid, useful and correct, one needs full information of those methods. In the actual case, that is not possible. I can imagine that the methods may be worth publishing in a separate paper, as they may be useful to many other data products. However, for the present manuscript and data product, information on the methods is necessary. As the paper with a description of the methods is not yet available, I suggest the following solution, as I do not want to reject the manuscript because of this. I suggest the authors add a paragraph (or two) with the most important features of these methods, in such a way that the reader may be able to assess their validity and usefulness.

Response: We agree that this is critical information for evaluating the data product. However, the important thing for readers is understanding what QC steps were performed rather than that they understand the tool that helped us perform the comparisons. We therefore have refocused the text in that section on how the comparisons were performed.

We've made that clearer by adding this sentence "Below are the major steps of the QC procedures as executed by these tools (except Step One)". Step Two to Four lays out the details of these tools. We have also added a new figure (Figure 2) to cover the main components of these tools.

We're currently funded with a two-year workplan to finalize these tools. We need to tap the expertise from experienced tool builders, modularize a lot of the functions, clean up the code, and make it a set of mature tools that are releasable and useful to the community.

L278-279 "TALK was preferentially used as the second carbon parameter. When it was not available, DIC was used." Why the preference for TALK ? Please explain.

Response: We did an analysis by adding a 5 umol/kg error to either TALK or DIC, and re-calculated output pH (see blow). In both cases, the "errors" in the converted pH are extremely small (around 0 to 4 x 10^-5), but are slightly smaller for TA conversions than for DIC conversions.

L291-292 "… as well as a measurement with one method against that with a different method (e.g., oxygen measured from Winkler vs. a sensor)." This way of working does not fit with the purpose of the measured oxygen. The purpose is namely to check and validate the oxygen sensor data, which automatically excludes its use as a quality check. One measurement is thus used for two purposes: This method of quality check is not acceptable.

Response: We agree with the Reviewer and only used the Winkler based oxygen data to check the quality of the CTDOXY sensor data. We've made it more clear in the text. See Line 314-315.

For pH, carbonate ion concentration and fCO$_2$, it is a different story. For example, we were able to use calculated pH to identify errors in measured pH. Vice versa, we were able to identify issues related to the TALK by using measured pH.

L299-300 "Consistency check-based outlier identification was the primary way of finding outliers in this study. Consistency checks were conducted for these variable pairs" For these checks the precision of the measurements is very important, as it primarily determines the possibility of comparing the data. How did the authors fit in the precision of the various variables?

Response: We estimated the measurement uncertainty based on deep level station analyses, and we also calculated the expected errors for any calculated values based on propagating uncertainties in carbonate system calculations using the CO2SYS companion errors.m program.

Unfortunately, it is challenging to apply across-the-board thresholds for their differences, as such differences vary dramatically from surface to the bottom, as shown in the new Figure 9-11.

Table 3 There are criteria for the different flags, but they seem not very stringent (as shown by the use of the word "often"). If this is the case, who did give these flags? Did single authors rate cruises or was there another way of coming to a result? Please explain.

Response: This rating was given by co-authors who are familiar with the used measurement technology and assurance.

Table 5 Although the units of all parameters are given in Table 2, I think it is a nice service to the reader to give them here again.

Response: Units have been added to Table 5.

Table 5 The minimum salinity is very low, i.e. it is almost river water. This indicates that estuarine data were included. Earlier in the text estuaries were excluded (Section 2). Please explain or correct these contentions.

Response: We did not include cruises that were collecting data exclusively from the estuaries. However, if a cruise covering the continental shelf also collected a few data points from estuaries, we did not exclude those estuarine stations. We've added a new sentence (See Line 157-159).

L382 The percentage seasonal variation for pH seems small, but pH is of course logarithmic. Would be good the mention that.

Response: We have removed the percentage numbers from pH, per request from the other Reviewer.

L449-450 and L451 These references are not necessary here. Apart from the fact that referencing in the summary and conclusion should be restricted, the method referred to here was already referenced earlier in the manuscript.

Response: These references have been removed.

L451-452 "Uncertainty analyses suggest that cross-over adjustments could be applied to future coastal data QC." This is a strange conclusion. Why would one only apply this in future cruises? If the method works well, it should also be applied in this data product. The authors did indeed do such an analysis; however, this conclusion here let suspect that its role should be more important in the analysis of the data, for example, whether corrections should be applied to certain data sets.

Response: In order to apply cruise to cruise adjustments, these below criteria must be met:

- Both cruises must have stations with sampling depth > 1500m

- These deep sampling stations must be close enough to do such analyses.

For this version of the CODAP data product, none of the identified deep sampling stations meeting the above two criteria showed any systematic offsets. That's why offsets were not applied to this study. That's also why we were able to use their differences as a proxy for estimation of uncertainties. We know this data product also contains data with less quality, e.g., those with a Cruise_flag of C and D. Unfortunately, none of them have deep sampling stations that would allow us to adjust for systematic biases.