



review of Jiang et al - essd-2020-402

Anonymous Referee #2

Referee comment on "Coastal Ocean Data Analysis Product in North America (CODAP-NA) - An internally consistent data product for discrete inorganic carbon, oxygen, and nutrients on the North American ocean margins" by Li-Qing Jiang et al., Earth Syst. Sci. Data Discuss., <https://doi.org/10.5194/essd-2020-402-RC2>, 2021

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.

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.

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.

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?

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.

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

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.

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.

L146, 147 These abbreviations are not used below.

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.

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.

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.

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.

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.

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.

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.

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?

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.

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.

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.

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

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.

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.