Ocean Sci. Discuss., referee comment RC1
https://doi.org/10.5194/os-2021-22-RC1, 2021

Comment on os-2021-22

Anonymous Referee #1

Referee comment on "Filtering method based on cluster analysis to avoid salinity drifts and recover Argo data in less time" by Emmanuel Romero et al., Ocean Sci. Discuss., https://doi.org/10.5194/os-2021-22-RC1, 2021

General Comments

I find that the aim of this study is interesting and the tool the authors provide is useful. However, it presents one main problem that is the applicability of the tool in any polygon of any area chosen by the user. I don't think that the tool can be used globally and so, its limits need to be specified by the authors.

The cluster analysis is carried out within a region that can be defined by the user. So, if this region is large, how can we suppose that the profiles should be similar to each other within the region even below 1500 m? How did the authors chose this criterion? The choice of 1500 m need to be justified, why not deeper? It could be spatial variability that is not easy to distinguish from the salinity drift. I believe that the authors need to prove that below 1500 salinity does not vary.

Also, a more detailed explanation of this cluster analysis is needed. The centroid of the groups is considered in salinity? Or in both temperature and salinity ? The cluster analysis bases on iterations to approximate the centroids in data space (what is data space? T-S space?) to their closest centroid. An schematic of the functioning of the algorithm would be very useful for the readers.

I suggest to add more information about the processes of this real time quality control as performed by Argo. In this way, the readers can realize in what the proposed technique differs from the one already existent that discards the profiles with salinity drift. Note that the Argo Quality Control is a very detailed process, so it is not easy to justify and propose an alternative method that inspires the confidence of the users. The description and justification of the proposed alternative must be very detailed. I understand that the authors are not saying that their method is an alternative better than Argo, rather, they propose a solution to have available more data in short time only in the case that they were discarded at first instance by Argo due to a salinity drift. Still, this point should be very clear to the readers.

Figures seem inadequate to me, the first 3 figures could be easily summarized in a single figure, and they are not well referenced in the text. Moreover, some of the figures are explained in a too exhaustive way in my opinion (i.e. figure 3) while others that contain more substantial information are too briefly described (fig. 7)

The second filter needs further explanation.

How do the authors deal with the data that on top of showing salinity drift, show any other problems? Do they discard them or not?

Specific Comments

Abstract

“In the study area selected as an example, it was possible to recover around 80 % in the case of the first filter and 30 % in the case of the second of the total real time quality control data that are usually discarded due to problems such as salinity drifts”

This sentence is not clear, (a) what is the first and second filter? (b) the Argo quality control is not only based on salinity drift, so can you explain the method a bit more?

Introduction:

Line 20:-22: It would be useful to say here what percentage of total data are flagged good in average instead of saying that in places with low concentration of profiles the good quality data are (not ‘is’ as it’s written in the text) scarce.

Data collection and methods:

Line 47-49: It seems to me that, if you set the range of latitudes and longitudes as described you'll end up with a regular polygon, what happens if the selected polygon by the user is irregular?

If that's not the case please explain this paragraph better

Technical corrections

47: change "them" by "the data" for example. The reader won't know to what "them " makes reference.

52: The characteristics of the region are relevant for the study presented here? If so please say in which way.

65: What are the corrections of the data according to TEOS-10?. Actually I do not understand well this whole paragraph. The authors used monthly means of T and S from WOA and then the PIP algorithm was used to separate these data in those that are inside and outside of the polygon, but what happened next? In what consist the comparison between these WOA data and those from DMQC from Argo? Please explain

60-65: Also, This paragraph describes the statistical mean of each quarter degree (¼) in WOA. It seems that this is shown in Figure 3, why it is not referenced here?

69: What does "around 30% of the data are part of the RTQC" mean? That are these data flagged good by the RTQC?

71: "one being the best and the fourth being the worst. Tests were performed by graphing the TS diagrams using these flags" please review the writing.

71: These tests are not related with Figure 5?, if so, please reference it.

80: The centroids are computed in the T-S space?

88: "it calculates the mid-ranges of each quality control"… The mid ranges of what variable? Please describe what mid-ranges is.

In line 85 The authors say that Algorithm 1 is used to automate the process of the enumeration of the groups to be searched for the k centroids. However then, in line 93 the authors claim that "The procedure described above is the first filter of the RTQC data". I did not understand well how the explained algorithm is a filter and how it relates with the stated from line 85. I would suggests a better explanation of this process since it seems to be key for the paper.

93-94: The second filter need further explanation (already said in general comments)

100: “In Figure 2, the blue line delimits the EEZ of Mexico and the yellow box delimits the TPCM.” it is the opposite way

115: In my opinion the authors describe too in detail the processes of selecting the data that are inside the polygon. It is not a complicated task to accomplish and I don’t think that it deserves a whole figure with two panels to show the same thing.

120: the DMQC in Argo and the same DMQC in WOA18? Please specify, since the WOA data also have quality flags

Figure 4 needs a legend indicating what the two colors are (and same for Fig 5 and 7)

121-122: That seems to be true, but in the figure we cannot see where the 1500 m depth limit is

125: So...data that have been qualified as good by Argo in their RTQC present salinity drift? Why were they qualified as good then? Probably the Argo system knows that they can be corrected? That's why I recommend to include in this paper some more information of the more relevant choices of the Argo QC.

128: And here the authors say that these data with drift are labeled as erroneous...I don't understand this contradiction with previous lines. Is it a mistake or am I confused?

134: The authors need to show that this is true and where it is true. This is a major shortcoming of the study

144: one of the blue profiles shows evident salinity drift, why is that? Argo error? Authors mistake in plotting the profiles with different colors?

145: discarded by who? By Argo RTQC or but the authors of this study? This figure needs more explanation

146: "both groups contain data in DMQC "? Is this true? Or mainly both groups contain data that match those of DMQC? I'm not sure I'm understanding. And under which criterion is the matching defined? Also, shouldn't panel c plots be in blue color for consistency?

151-153: The results of the second filter seem quite good and promising. However, I insist that this second filter is not explained enough. Please provide a more detailed explanation in the methods section.

Table 1: What does “meas” mean?

164: “the researcher may simply not use the data from those months”. I strongly advise to delete this sentence for two reasons: (I) the problem is probably not with these months but with the data and if we change the region, the wrong data would be in different months. (ii) it can be not easy for researchers to go and look for the data that are wrong and decide if the are wrong enough to discard them. The I’d propose to the researchers to only use the data from the second filter (and using data from the first one would be on their on risk).

171: I would start a new subsection here, something like: “Web application”

171-173: This figure and result is very similar to those in fig 3 and which I have already advised to reduce. Now I insist that the authors could joint together Fig 9 and 3 and summarize the description. Choosing data that belong to a given polygon (even if it has a complicated shape) is a very simple task in my opinion, and it doesn’t deserve that much of attention. The most interesting subject of the paper is the filtering procedure that could gain more attention and more space for its description on the paper.

189-191: The authors talk a lot about the example in the ETP off Mexico, but, at which degree is their method applicable to larger or different polygons?

216-219: “The current platforms already provide graphics and data from the profilers, as well as

filters to display or download the data, however, the geographical filter they use is by maximum and minimum coordinates, so it is only possible to filter by polygons in rectangle or square shape without rotation”

I see now the interest on showing that with the tool provided by the authors users can choose irregular polygons. This advantage in comparison with other platforms is great, and it should be mentioned earlier in the text. However, I still thing that it can be said in one or two sentences and that too much detail on this is included in the text before (in the discussion is fine).

219: define JCOMMOPS (and change analyzes for analyses)

230: This sentence is not a conclusion of this study, it should be removed. This is something between Argo and WOA.

233: 80% regarding what? Earlier in the text, the authors said that the data recovered in comparison with the DMQC of Argo were 30% and 10% respectively for the first and second filter. I recommend to define the criterion for the recovering percentage, either regarding the total amount of data or regarding the data that are discarded by the Argo DMQC.

Technical comments:

- data is plural, please correct the concordance with the verbal tenses throughout the manuscript

- 113: New sentence after “worked correctly”

113-115: “in addition to establishing the range of maximums and minimums of the latitude and longitude of the polygon to discard the profiles measured outside it, allowed the PIP algorithm to filter only the profiles made near or inside the polygon”. This sentence is oddly written and seems kind of obvious.

- 194-195: “since these processes are automatic and search for data that is impossible or outside the global and regional ranges” Please rewrite