

Ocean Sci. Discuss., referee comment RC1 https://doi.org/10.5194/os-2021-22-RC1, 2021 © Author(s) 2021. This work is distributed under the Creative Commons Attribution 4.0 License.

## 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

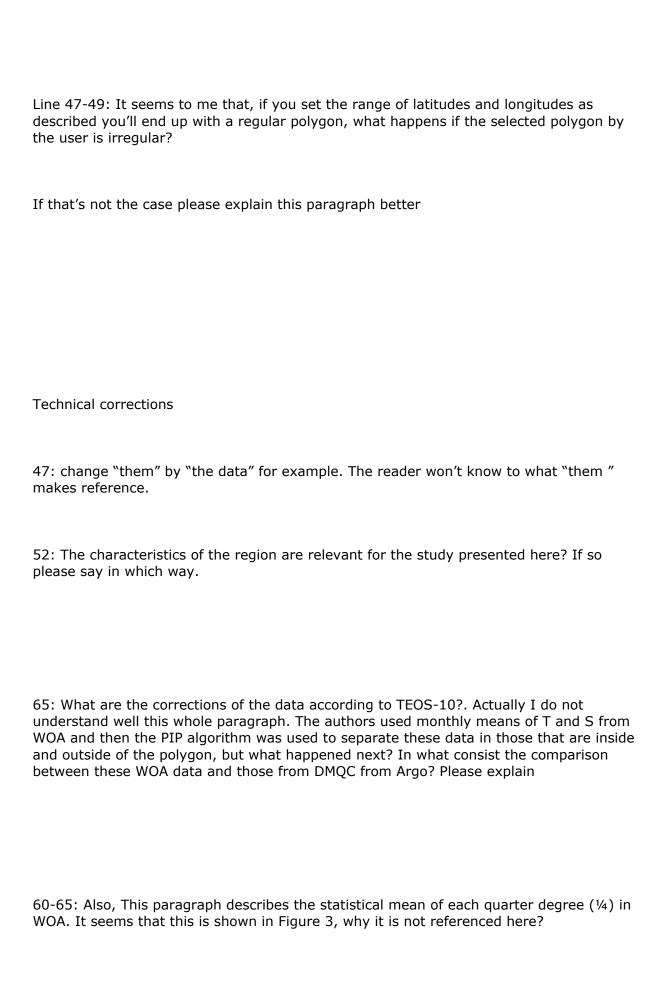
## **General Comments**

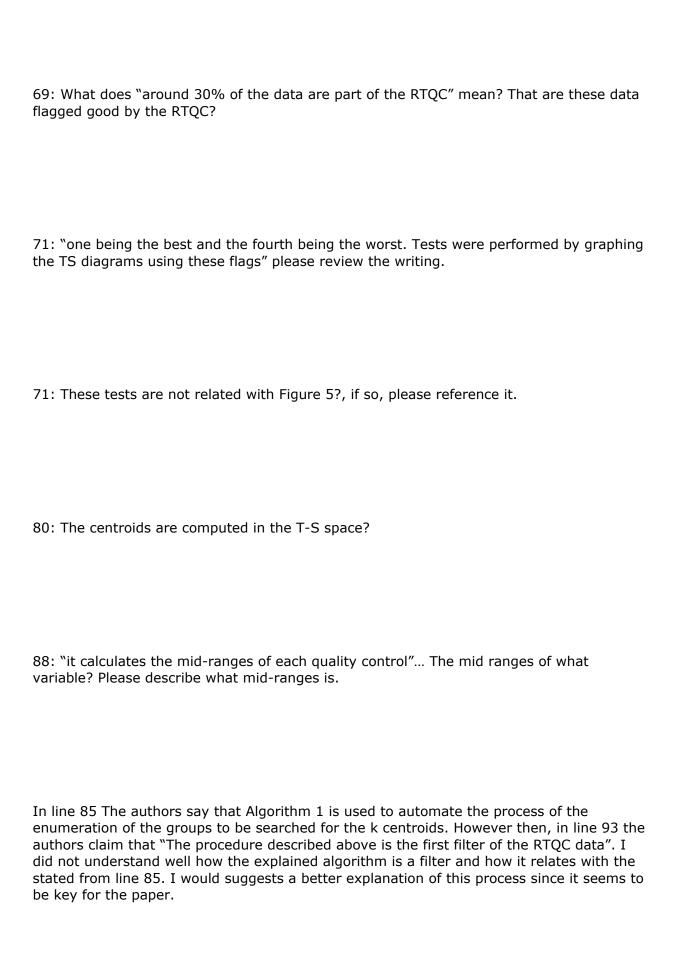
I find that the aim of this study is interesting and the tool the authors provide is useful. However, it presents one main problem that is the applicability of the tool in any polygon of any area chosen by the user. I don't think that the tool can be used globally and so, its limits need to be specified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

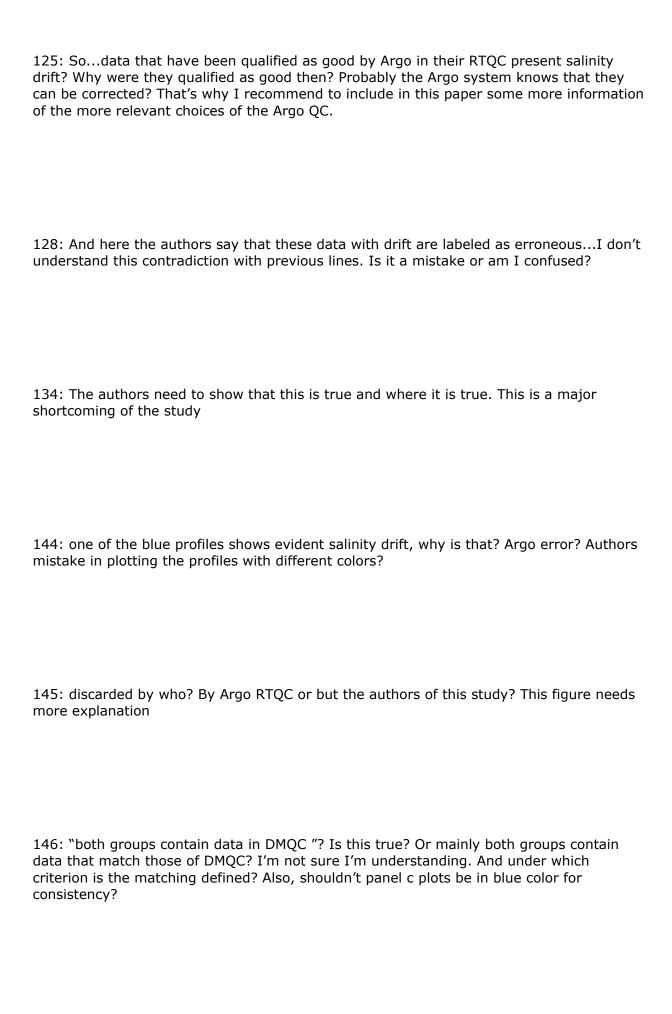
Also, a more detailed explanation of this cluster analysis is needed. The centroid of the groups is considered in salinity? Or in both temperature and salinity? The cluster analysis bases on iterations to approximate the centroids in data space (what is data space? T-S space?) to their closest centroid.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

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:





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





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
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