Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-306-RC1, 2022

Comment on essd-2022-306

Anonymous Referee #1

Referee comment on "Water masses distribution in the Canadian Arctic Archipelago: Implementation of the Optimal MultiParameter analysis (OMP)" by Alessandra D'Angelo et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-306-RC1, 2022

Review Angelo et al, submitted to ESSD

https://doi.org/10.5194/essd-2022-306

This manuscript describes a data set of 53 CTD stations from the Canadian Arctic Archipelago (CAA) region. The data consists of SBE911 C(S)TP and nutrients and dO18 data analysed from rosette sampler bottle samples. The observational datasets have been published in different repositories while the whole collection of data can be found at https://arcticdata.io/catalog/view/doi:10.18739/A2W66995R

Besides the observational data, water mass mixing analysis are applied to the data and the results are also part of the whole data collection and archived via the at https://arcticdata.io/ handle.

I my view this data set is not ready for a publication and that has to do with various aspects in the data formatting data, data archiving, but also in the way the water mass mixing analysis have been applied and results are archived.

My first major concerns are on observations metadata. I could not find any information about international data and metadata standards used e.g. which variable names are used (e.g. CF)? which instruments been used (e.g. in SeaVOX standard names)? Along those line – I would recommend you to store the data in a machine readable (and thus FAIR) format such as netCDF.

Second, no information is given about the data quality assurance (QA) that has been designed for the data. As a consequence of no data QA, the data quality control (QC) is not outlined and no references to the applied Standard Operation Procedures (SOP)s are given. And, most importantly, no information about the expected uncertainties of the data sets is provided.

These are shortcomings to the manuscript but also this information is to be added to the data sets in the respective repositories (e.g. via data fields and the metadata).

In addition, I see a problem the way the different datasets, stored in other repositories, are "merged" at http://arcticdata.io . There is not information in the metadata at http://arcticdata.io about where the "underlying repositories" are located – and hence I do not understand how it is ensured that updates/revisions of data sets will propagate into all data bases? For example, it is said that the C(S)TP data is found at CCHDO (where exactly?) but how is it ensured that any correction to the C(S)TP is, after an update to CCHDO, communicated to https://arcticdata.io? What would be your strategy to notify the data repositories about any alternation/correction to the underlying data sets? And what is true for the arcticdata.io also is true for the dO18 and salinity stored at Pangae.de (https://doi.org/10.1594/PANGAEA.937543)?

Making all such information transparent is key when it comes to AIR data – certainly a standard that is motivated if not required by ESSD.

My second major concern is related to the two water-mass mixing analysis, that were applied and the results are added to the data set under discussion. As I see it, two mixing analysis techniques have been used: (1) one is a derivative of a mixing triangle (TRI) that aims on reconstructing dO18 data, and (2) the other is a technique that is called here an "Optimum Multiparameter" analysis (MP; but I will detail below why it is not) and that aims in estimating fractions/percentage of source water types.

First a general comment on water mass mixing analysis - in all cases the user of such analytical tools must be aware that a mathematical technique is used for an oceanic application – and the mathematical tool will always estimate a solution of a given problem – regardless if it makes sense or not. For a water mass mixing analysis being a meaningful application two factors matter most: 1) the choice of boundary conditions (here, the source water types) and 2) the formulation of the mixing problem as a system of equations that take into account all factors affecting the mixing - from the origin of the water masses to the region the analysis is applied to. Obviously, all source waters expected to contribute to a regional water mass composition need to be considered. So, if multiannual variability in source water is high (e.g. the various blends of Labrador Sea Water, LSW) and the interest is decomposing an interior field that has ventilation times >1 year the interior field is potentially be impacted by various different types of LSW – and thus various different source water types for LSW must be defined and considered (even so oceanographers use only one single “nick name” for the water mass: which is “LSW”).

As said, here two mixing techniques are applied (TR and MP) – but without ensuring the necessary link between them. From an oceanographic point of view they must be linked because both assume that certain data points can be decomposed in underlying source water masses. What I mean is that if you consider the source waters types used in the TRI (which are the triangles corner points) are meaningful (which I hope you do) they also must be used as source water types for the MP. Why should it makes sense to re-create the d18O field from various source waters but then use different source waters for the MP? What you could do as an alternative is to define the d18O for the few (4) source waters that are considered in the MP, and reconstruct all d18O from that – this would at least bring both methods into the same reference frame.

Another problem in the application here is including the surface waters in the analysis. Typically, these waters are excluded from mixing analysis because local air/sea/ice fluxes constantly change over time and this is what creates upper ocean (mixed layer) properties. In a water mass mixing context: the varying air/sea/ice properties would need to be represented by many, many source water types… and thus many, many parameters are required to ensure an overdetermined system to be solved. This has fact has not been addressed in the manuscript.

The last comment is on the MP. It is said that the authors used the Optimum Multiparameter analysis. Please first of all note that the Optimum Multiparameter analysis goes back to Tomczak and Large 1989 (https://doi.org/10.1029/JC094iC11p16141) but only was used by Tomzack and Liefrink (2004). For details on the method I recommend you to see the Tomczak and Large 1989 and where you find that the OMP analysis is based on an overdetermined system because it does NOT apply a simple least square (as written in the manuscript) but a non-negative least square (based on an algorithm by Lawson & Hanson 1974). The nonlinear least square ensure all water mass fractions are >=0 (which is oceanographically meaningful) but this is on the costs of the degrees of freedom and in fact one more parameter then the number of unknown source water type fractions is required for determining a solution of the system of equations. In fact there is an earlier version of the technique – the Multiparameter mixing analysis (Tomzak 1981) and that does use a simple linear regression, maybe you mean that?

In addition: as for the data, it also is important to determine uncertainties in the OMP analysis fields – e.g. by applying Monte Carlo simulations with randomly disturbed source water types and also observational data to quantify the robustness of the fits. These uncertainties should be added to the analysis results and part of the data set you aim to publish here.

Specific comment:

- The arcti.io data base XML says salinity is in PSS 78 scale, while in the text you refer to using the TEOS-10 framework – why is that?
- Was there an oxygen sensor on the CTD? Has that been calibrated?
- Moreover, the data is further analysed with a water mass mixing analysis.