Earth Syst. Sci. Data Discuss., referee comment RC4
https://doi.org/10.5194/essd-2020-308-RC4, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on essd-2020-308

Anonymous Referee #4

Referee comment on "Climate-Biogeochemistry Interactions in the Tropical Ocean: Data collection and legacy" by Gerd Krahmann et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-308-RC4, 2021

General comments:

This manuscript provides an "inventory" (line 27) to the data legacy of a major research program („SFB 754"). In fact, it states (line 58) that "The aim of this article is to describe and list the published observational data sets collected by the SFB 754 for easy access and find-ability." Does it really describe a "recombination of existing (historical) datasets"? What, beyond providing a compact listing and cross-reference, is the added value of this inventory?

One could easily justify it as a highly valuable companion report, supporting the overall final report of SFB 754. It ties together and identifies some 1,000 datasets from 34 cruises performed over more than 10 years, from a wide range of (oceanographic) disciplines (lines 51–53) and links them with their supporting documentation (e.g., cruise reports, best practice documents, etc.). The manuscript provides some interesting indications about the program's data policies and practices (including the existence of data processing pipelines from raw to published data, lines 96-97, which appear, however, only accessible to SFB consortium members, lines 93-94).

According to the manuscript (line 92), 421 publications were already generated from that research. Due to their intimate knowledge of instrumentation and methods, the authors and reviewers of those publications were certainly convinced that the underlying data were indeed fit for their 421 purposes.

But as ESSD is about supporting the reuse of data in other contexts than that of their creators' research, the most important criterion on the datasets and/or the ESSD articles is to provide clear (and, as far as possible, easily interpreted) information which helps to evaluate or assess the datasets' fitness for the purposes of a third party (aka, its quality). (ESSD review criteria ask „Are error estimates and sources of errors given (and discussed

in the article)")

Does this manuscript support these aims and does it meet this criteria of ESSD? I think not, as none of the subsections on specific parameters does provide direct information on quality:
Each of the subsections, on individual observed parameters, describes the measurement methods explicitly – some very briefly (4.1.2, 4.2.4), some in much more detail (most subsections of 4.3 and 4.5). Many, but not all subsections (on physical oceanography) then refer to individual sections of the "GO-SHIP best practices". Many subsections refer to individual research articles (as well). While we can only hope that the GO-SHIP website will survive until it will be needed by a user of these datasets, at least some of the articles referenced are behind paywalls: E.g., 4.1.3, line 176, Hahn et al. €37.40; 4.1.6, line 221, Foltz et al. $8-$49. As the Hahn article, in particular, is referred to in the context of measurement errors (line 179), this defies the dedication of ESSD to provide such information in Open Access (above the inconvenience of scrutinizing yet another article for the information which should have been provided explicitly here or in PANGAEA tables).

ESSD review criteria further ask reviewers to: "Consider article and data set: are there any inconsistencies within these, implausible assertions or data,…" Specific comments, below, show that one will encounter – right from the beginning - problems doing this kind of consideration for this manuscript and its 1,000 (heterogeneous) datasets, and it appears completely out of the capacity of one or even 4 reviewers to comply with this, even on a random but representative sample from the 1,000. This would put any such manuscript out of the clearly delimited aims of ESSD – except it would provide an algorithmic way and an execution environment to perform these tasks (however that might work – but it appears to this reviewer as one possible benchmark of FAIRness: That data and metadata are machine-readable and interoperable).

As a last general comment, ESSD review criteria ask: "Is there any potential of the data being useful in the future?". The thoroughness of the description of methods (and, presumably, their consistent execution), and the broad interdisciplinary application within the SFB research program hint at further applicability. Authors mention this in Conclusions, but do not provide examples beyond the (immediate) thematic realm of SFB 754. (Would it be possible and useful to compare these with similar (climate-BGC) data from non-tropical areas? Is the collection as a whole significant to track effects of climate change, and if so: when might – selective - repetition be advisable, thus setting a lower time limit of necessary preservation of this data collection and its interoperability?)

In summary, I do not see how this "review article" provides for the "recombination" of existing data, nor do I think that it meets the general aims of ESSD to help assess the usefulness of data for readers' purposes. It is up to editors, if they wish to establish a new manuscript type or extend the scope of review articles to include the description of data inventories. In the absence of such decision, I do not suggest to accept this manuscript for ESSD, useful as it may be in another context.

Specific comment #1:

While the manuscript may be the best an inventory can do to describe the maze of individual datasets and their listings, ordered by parameter observed, cruise where such observation was performed, and of pertinent references to ancillary information (such as cruise reports, best practices), it makes it truly hard for someone, say, interested in ocean currents, 25 years from now:

- Section 4.1.2 goes so far as to name the instruments used, but why does it not tell us that ADCP-specific datasets at PANGAEA actually do carry a "+/-"-column in their data table (it does for cruises AT08-04, https://doi.org/10.1594/PANGAEA.811565, and M80/1, https://doi.org/10.1594/PANGAEA.811718, – but also for all others?). (The ADCP example demonstrates how important this information is, as in the above cases the median of the *relative* error is 60 and 100%, resp., and just 20 and 10% of all values, respectively, claim less than 30% relative error. A "naïve" future user - e.g., from another discipline, or from a time when there might be a much more accurate instrument - would not suspect as much.)
- Instead, section 4.1.2 sends us on a journey through references:
  - First, at line 157, to Krahmann and Mertens 2021b, which is a PANGAEA reference, https://doi.org/10.1594/PANGAEA.926065, listing, however, all datasets from all "CTD data and additional sensors used on the CTDO system". Actually, clicking on list item 13 (on the Atalante cruise), one is directed to a dataset which does not include any LADCP data!
  - Second, also line 157, it directs us to table 2 (of this manuscript), which indeed has a row 4.1.2 with a PANGAEA reference to all SFB LADCP data. Why don't authors provide this reference here, directly?
- Fun fact: PANGAEA names "[+/-]" as the unit of the absolute error, which of course should be [cm/s].

Specific comment #2:

Immediately on going into the data reference, I encountered another irritant: While the main chapter 4 directs us, in its first paragraph (line 122), to "cruise reports where additional information about the data collected and methods used can be found", there appears to be an inconsistency about error margins between the first of these reports (Atalante cruise AT08-04) and the associated first CTDO dataset: On the end of p.10 the cruise report, http://dx.doi.org/10.3289/ifm-geomar_rep_19_2008, appears to claim an error ("rms difference") of 0.052 ml/l in the oxygen data, while at PANGAEA, the CTDO data for this cruise, https://doi.org/10.5194/bg-10-5079-2013, appears to claim a "nc_uncertainty_o = 1.000000" in the metadata "comment" item, which one might, tentatively, amend with the unit for oxygen from the "parameter(s)" table, namely µmol/kg.

As these claims appear to be inconsistent (https://ocean.ices.dk/tools/unitconversion.aspx), one is forced to think: Should I take the PANGAEA uncertainty value as the most authoritative, since the somewhat substantial description of data processing and calibration in this article seems to refer to it – and

disregard the (preliminary?) estimates from the cruise report? Should I do this in general, that is: for all cruises and parameters? Will I then find nc_uncertainty_xy values or error columns in all datasets?