

# Overview

The paper describes GLOBESINK as a global monthly climatology derived from BGC-Argo backscattering measurements, with gridded POC, particle-size and related variables. The manuscript itself is generally thorough and well written, though the readability could be improved and some clarity could be added in some places. The impact of the dataset itself is potentially significant, and I would expect it to be widely used. Aggregated monthly climatology datasets such as this one play a significant role in increasing the reusability of data, since finding and aggregating data requires significant time and expertise. However, the impact of the dataset could be increased by improving how the netCDF files are structured, improving the metadata, and republishing to a different data centre that makes the data more easily findable and reusable. The code should also be openly available, not simply “available upon request”.

## About the review

I have focused this review on areas within my expertise, particularly data publication practices, metadata, and clarity of processing and QC. I am not an expert in POC flux estimation or flux-validation methodology.

I have categorised my comments as “recommendations” or “suggestions”. My preference is that all the “recommendations” are addressed before the manuscript is published. The “suggestions” may be my personal preferences or minor comments that the authors or editors are welcome to disagree with.

## Manuscript

The numbers used at the beginning of each bullet point refer to line numbers throughout this review.

### Recommendations

- Throughout section 2, I do not have an intuitive feel for how aggressive the QC process is. Figures 1-3 are useful in presenting examples of this, but it is difficult to gauge how representative these figures are. I assume these instances were selected to portray particular examples. This could be addressed by including statistics of what percentage of profiles or floats have been affected by each stage of the processing, or by providing further context on why certain thresholds were selected. For example, in 2.2.2., it would be interesting to quantify how much data includes values above the prescribed thresholds.
- 67 - 99: I assume that these steps were applied sequentially and in the stated order, but this is not stated anywhere. This is a thorough enough explanation of all the steps, but I think a flow chart depicting the data processing would be a useful

addition to improve readability. This should begin from the data source, include all steps, right through to the final datasets.

- Table 2: It would be good to know the total number of profiles processed for context here. I am more interested in the percentage of profiles removed than the absolute numbers.
- Figures 1-3. These are useful additions to show the effect of the automatic and QC routines. However, they should be referred to at relevant places in the text to instruct the reader when to refer to them. Furthermore, they could be annotated to highlight the changes between the different stages - perhaps using the QC flags introduced in Table 2? Whilst some changes are obvious, in other cases it takes some time to see the changes between successive processing stages. E.g. Figure 3 between raw and auto QC. Annotations would be beneficial here.
- 

## Suggestions

- A general comment. The readability would be improved by referring to the figures and tables throughout the text.
- 16: Why is “chlorophyll a” capitalised? It is not capitalised elsewhere in the manuscript.
- 21: What do you mean by the mean relative bias? Globesink - obs / obs? Or globesink / obs? Maybe my misunderstanding is due to my ignorance but this is not clear to me.
- 51: Given that the “optical backscattering spikes” particle size fractionation method is fundamental to the dataset presented, I would expect the method to be described briefly within this manuscript, rather than only referring to earlier publications.
- 63: FTP should be all capital letters.
- 79: Consider using bullet points for this?
- 92-93: Bullet points missing here? Or perhaps should be formatted as equations as done from lines 215 onwards?
- 185: “Note that the matchups do not need to be from the same year”. Okay, but are there are temporal limits implemented here? This is not clear.
- 195-196 & Figure 4: The differences in the dataset pre and post alignment are quite striking in some cases. How confident are the authors in the reliability of the aligned data? Could this confidence be quantified and stated somehow? The authors mention that only the post-alignment data are published. It will be important to include quality flags and uncertainty in the data, following best practices. I will describe how to do this in the ‘Data’ section.
- Section 2.7. I see that both interpolated and “uninterpolated” data have been published, but I don’t think this has been mentioned. It would be good to signal at what stage in the processing the “uninterpolated” data have been published. Perhaps the flow chart suggested earlier could help with this. For interpolated datasets, it is also important to include flags in the published data to describe which samples are interpolated.
- 278: Tracked changes on?

- 279, 282: Should be a space between the value and the units. This has not been done consistently throughout the manuscript, and I have not checked every instance.
- 309: A small question to the authors. Does treating 8 degree longitude and 4 degrees latitude as equidistant yield suitable results near the poles? I assume this was investigated during development, I am just curious.
- 314: Chlorophyll a is capitalised again. It is not capitalised on lines 58 or 65. I am no expert in when this should be capitalised and not, but make sure you are consistent with whatever the rules are here.
- 324: Of every year or of one particular year?
- Figures 7, 8, 9, 10, 11, 12: Which year is depicted here?
- 392: Only the year should be in brackets for this citation. /cite instead of /citep
- 482: I understand that it is difficult to commit to time intervals for these periodic updates, but some ballpark estimates could be useful. I suspect you do not plan to publish updates monthly. Annually? Every few years?

## Data

### Recommendations

- I would advise against publishing data to Zenodo. Whilst Zenodo is commonly used, it limits how discoverable data are. Zenodo follows no standards for discovery metadata (e.g. ISO 19115, GCMD DIF, DCAT, EML, schema.org), and has limited metadata requirements. Data are also not reviewed before publishing. Nor will they be curated over time in the future. It is difficult for someone to search for data that they are not already aware of. I invite you to try speculatively searching for data on Zenodo yourself. A good data centre would include the possibility to filter based on the geospatial and temporal limits of the data, some standardised keywords, etc. This is not possible with Zenodo. Metadata should also be exposed through platforms that support standard metadata harvesting mechanisms (e.g., OAI-PMH, OGC-CSW). Zenodo does not provide this either. These standards allow aggregators to provide unified access to data published in multiple data repositories, regardless of where the data are stored physically. Since it is not practical for individuals to search through all data repositories for interesting data, aggregators have a key role to play in simplifying data discovery. The truth is that the data ecosystem around this does not exist at scale yet, but is being built, and when publishing data we should consider what is best for usage in the future, not just today. Since you are based on the UK, BODC would be a good choice <https://www.bodc.ac.uk/>. BODC also supports OPeNDAP, which can be used to stream data over the internet instead of requiring direct downloads. This could be useful as the input for your data visualisation software that you linked to in a comment. If the data are updated later, the data URL could be the same (depending on how you publish future versions) meaning that users will not have to redownload, nor update their code, to access the latest versions.
- The data themselves need to be cleaned up. The metadata are not sufficient for machine readability. I will outline the changes that should be made before providing the CDL of what the data could look like

- All variables should include a standard\_name where one exists.  
<https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html>  
 Where a suitable CF standard name does not exist, new ones can be proposed - the expansion of vocabularies is a community effort! But if this is out of the scope of this work, make sure every variable has a long, descriptive long\_name as a minimum. The variable lookup table and the “source\_variable\_name” attribute should not be necessary if standard names are used. Generally speaking it is good practice to encode everything needed to understand and use the data within the file itself, rather than relying on external files. CF-NetCDF supports this.
- Similarly, the discovery metadata is far too limited. You should adhere to the Attribute Convention for Data Discovery (ACDD) documented here: [https://wiki.esipfed.org/Attribute\\_Convention\\_for\\_Data\\_Discovery\\_1-3](https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3) . These recommendations are a bit outdated and leave a lot of degrees of freedom. I like the requirements and recommendations listed here: <https://adc.met.no/submit-data-as-netcdf-cf> There are examples of how to encode each attribute at the bottom of the page.
- Time should be encoded as shown here: <https://cfconventions.org/Data/cf-conventions/cf-conventions-1.13/cf-conventions.html#time-coordinate-variables> But using the units ‘months since YYYY-MM-DD.
- The coordinate bounds for each bin can be encoded using cell bounds, see here <https://cfconventions.org/Data/cf-conventions/cf-conventions-1.13/cf-conventions.html#cell-boundaries>  
 This might also be applicable for the POC\_bbp\_ratio variables and other such variables where bounds are encoded.
- The processing flow is well described in the manuscript, but this provenance information should also be encoded within the datasets themselves. In the CF-NetCDF files, this can be encoded in the ‘history’ global attribute.
- Flags should be used to encode things like data quality. This is particularly useful for encoding which values are interpolated.
- <https://cfconventions.org/Data/cf-conventions/cf-conventions-1.13/cf-conventions.html#flags>
- Include a variable attribute coverage\_content\_type for every variable with values taken from this list: [https://wiki.esipfed.org/Attribute\\_Convention\\_for\\_Data\\_Discovery\\_1-3#Highly\\_Recommended\\_Variable\\_Attributes](https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3#Highly_Recommended_Variable_Attributes)
- Use the variable attributes valid\_min and valid\_max to encode defined limits of the variables. For example, you mention in the manuscript that values above certain limits have been discarded: <https://cfconventions.org/Data/cf-conventions/cf-conventions-1.13/cf-conventions.html#attribute-appendix>
- You should use the variable attribute ‘ancillary\_variables’ to encode how variables are related to each other. Examples here: <https://cfconventions.org/Data/cf-conventions/cf-conventions-1.13/cf-conventions.html#ancillary-variables>

[ons.html#ancillary-data](#) For example, n\_bpp is an auxiliary variable of the bbr, bbs, bbm... variables I assume.

- Is 'double' format really necessary? Usually float32 is sufficient, but depends on the precision required.
- Encoding uncertainty in CF in a standardised way is under development, likely to be included in the next version. I advise that you encode your uncertainties following the proposal linked here:  
<https://github.com/cf-convention/cf-conventions/issues/660#issue-502953237>  
8 Please download and view the cf-conventions.html file provided by davidhassel and jump to section 10. You will find examples of how to encode uncertainty here, which will be useful for your 'precision' variables.

## Code

### Recommendations

- The data should be openly available for the sake of both reproducibility and transparency. I recommend the data are uploaded to a public GitHub repository. Ideally, you should also publish a snapshot of the repository containing the version of the code used in this manuscript. You can publish this code to Zenodo. There are tutorials online on how to do this quite quickly, e.g.  
<https://www.youtube.com/watch?v=HZ6m8oxwvig>