Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-230-RC1, 2021

Comment on essd-2021-230

Anonymous Referee #1

Referee comment on "Collection and analysis of a global marine phytoplankton primary-production dataset" by Francesco Mattei and Michele Scardi, Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-230-RC1, 2021

The authors have compiled an impressive set of phytoplankton production data, which they have compiled based on the specific criteria of temporal-spatial information, depth-resolved ^{14}C productivity, and chlorophyll a profiles. These data are highly useful for a wide range of applications and generating a central data product is certainly a worthwhile endeavor. My main comments outlined below are not directed specifically at these primary variables of interest, but more at the numerous ancillary variables that the authors have also included. While some of these additional variables are of a more extraneous nature, the addition of others is concerning given how they may be interpreted by subsequent users.

My main concern is the addition of variables that are derived from separate data products and are not directly associated with the in-situ data they are describing. Furthermore, there isn't a clear distinction in the dataset between when the data are in-situ and when they were missing values that the authors have filled in with a data product. The most pronounced example of this is SST, which contains both in-situ observations associated with the ^{14}C incubation experiments and values derived from remote sensing products or even a climatology for the pre-satellite era. This is not the same as the in-situ temperature observations; the derived products have associated errors and are spatially and temporally coarser which will introduce substantially more uncertainty when utilizing the productivity estimates. This is especially concerning because any users that download this data may not realize that the SST data is a blend of in-situ and remotely sensed observations, without fully reading through the associated manuscript. My suggestion to the authors is to either flag the data that have been filled in with a separate product, or to not fill-in this data at all. I would lean more towards the latter, as many data users will already have a specific data product in mind for a variable like SST, that may work better for their region or application of interest. This also allows the user to have a clear understanding of the uncertainty surrounding that SST estimate too.

PAR is another example of the previously mentioned issue, which contains a blend of in-situ and satellite-derived values. Additionally, the authors actually discarded nearly 800 profiles due to the lack of an in-situ or remotely sensed PAR estimate. I don't follow the reasoning for why these data were discarded considering that PAR was not one of the four criteria that the authors established for generating the dataset.

Apart from these concerns with SST and PAR, there are some other variables that seem more extraneous and in my opinion would be better left for the user to define, if necessary. For example, the Northern hemisphere seasonal classification is not applicable for the Arctic, where a significant portion of the data are located. This variable is probably best left for the user to define based on their specific application. The Jenks (1967) data classification schemes are useful for generating the manuscript figures that illustrate data distribution, though I'm not sure how useful these variables will be for other users. As the authors describe in the text, the classifications are generated specifically for the entire dataset, meaning that they will change if a user selects a subset of the data.

In the introduction, I suggest that the authors provide a little more background on the distinction between chlorophyll a and primary productivity. The authors allude to important differences between the two variables but do not go into much detail. This is especially relevant in order to drive the motivation behind this dataset, since productivity data are generally more scarce than Chlorophyll.

Technical comments:

Line 26: Change to "… of global productivity"

Line 60: Here and in the supplement, please be more specific than just "The National Oceanic and Atmospheric Administration" regarding where you pulled the data from. NOAA is a large entity and it's not clear where these data are virtually located.

Line 74: What is "CZCS"?

I would think the "Conclusions" section should come before the "Data Availability" section at the end of the manuscript, unless this specific format required by the journal.

Figure 2 caption: Should the seasonal definitions be winter (January to March), spring (April to June)…?