Linking satellites to genes with machine learning to estimate phytoplankton community structure from space
El Hourany et al.

Review of revised manuscript

The work of El Hourany and co-authors presents a machine learning approach (specifically, Self-Organizing Maps) to estimate the relative Chla contribution and cell abundances of seven major taxonomic phytoplankton groups. The results of the trained model are applied to global satellite data, and in turn compared to both a previous SOM model developed using pigment rather than omics-based biomarker data, and a separate DPA-based approach. The study is novel in its use of phytoplankton gene information to train an ML model for assessing phytoplankton community structure from space, and the authors have clearly put thought into comparison with other approaches, and to how uncertainties also play a role in the results. After reviewing the manuscript (and in the context of previous reviewer comments and subsequent revisions), I believe the manuscript is publishable following some minor revisions and corrections. Thanks to the authors for their work on this topic.

General comments

I appreciate the background and description of functional types, the DPA and pigment-based groups in the Introduction text. However, the phrase "phytoplankton functional types" is used several times in the document, when in fact what is meant is phytoplankton taxonomic groups. Although "functional types" has been used rather loosely in the literature, the term "functional" indicates biogeochemical function (e.g. calcifiers, silicifiers), whereas phytoplankton of different sizes or even different taxonomic groups may serve the same ecosystem function. Therefore, I strongly encourage the authors to instead use the phrase 'phytoplankton taxonomic groups' when that is actually what is meant, or when referring more broadly to the variety of phytoplankton, 'phytoplankton community composition/structure'. This attention to phrasing will benefit the community of researches working on the topic of phytoplankton community composition from space in the context of interactions with any potential stakeholders and end-users.

Could you comment on the fact that the *psbO* is a proxy of individual cells, but the chain-forming phytoplankton types (e.g., *Chaetoceros*) will contain several, or many, individual cells, and will likely be found in the larger size fractions? For example, if the abundance of diatoms is high in the 20-180 fraction, but the *psbO* represents individual cells (vs. chains), the diatom contribution to Chla could be dramatically overestimated.

In my opinion the text of section 3.4 needs to be revised for clarity. Is it not fully clear what the inputs and targets of the random forest regression algorithm are. The following sentence is not clear: "In the internal node, the selected feature (i.e., pigment in this case) was used to make a decision on how to divide the dataset into separate sets with similar responses in terms of a given phytoplankton group.” Suggest revision to help the reader understand the application of the random forest regression method.

Line 351: as I’m sure the authors are aware, the CHEMTAX approach was developed decades ago to address just this. Although it does have its own caveats, it can be a useful tool to compare against, and recent work to improve it and make it more broadly applicable is worth looking into (see Hayward, Pinkerton, and Gutierrez-Rodriguez. 2023. “Phytoclass: A Pigment-Based Chemotaxonomic Method to Determine the Biomass of Phytoplankton Classes.” *Limnology and Oceanography: Methods* 21 (4): 220–41. https://doi.org/10.1002/lom3.10541.) Perhaps this additional analysis is not warranted for this study, but it is worth keeping in mind for future work related to comparison of different approaches used to estimate phytoplankton community structure.

Specific comments

While I’m not aware of the specific journal requirements, numbering the equations would make it easier for the reader to reference the equations within the text for future analyses, etc.

L79: start the sentence with “The” to avoid leading with the gene name.

Line 100: should be GlobColour (with a capital “C”), throughout.

Fig. 1 caption: please define ‘D_{RCA}’ and ‘D_{ChlF}’ for the reader here

Fig. 2 – it would be valuable to also know the absolute *psbO* values as well – for example, it is true that the Prokaryotes are over-represented in the largest size fraction, but are the absolute quantities of *psbO* very low in that size fraction? I guess more generally – what is the range in absolute quantities of the *psbO* gene across the size fractions?

L116: please define ‘CCI’

L130: sentence is awkward as written and ending in “them”; suggest revising to something like “we used two previously published algorithms:”

L149: First sentence does not add anything for the reader.

L 169: is ‘variables’ here referring to the phytoplankton groups, the satellite-derived parameters, or both?

L 174: what is meant by ‘the SOM algorithm that can deal with missing values’? can you give a sentence or two to describe mathematically what is done to account for missing values?

L 184-6: could you add reference(s) here to back up this widespread use of SOM to complete missing data?

L 195: increased from 10 to 1000 neurons at what interval?

Line 313: Curious how you decided on the threshold of 40%?

L 357: typo ‘Glocolour’ (missing ‘b’)

L 358: typo ‘Fig. 101’

L 371: typo ‘Fig. 112’

Figure 9. – suggest including a colorbar to show the number of points per pixel based on the color of the dots on the graph

Figure 12. caption – not clear what is meant by ‘original Rrs spectra’

Figure 13. Capitalize the first word of the caption. Could the x-axis of the latitude line graph be revised to label more than just the 10^0 ?

L 448: suggest revision to ‘launch of NASA’s Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission’

L 451: suggest revision to ‘the perspective of the PACE mission,’

Alison Chase