Review of the manuscript "A synthetic database generated by radiative transfer simulations in support of studies in ocean optics and optical remote sensing of the global ocean", by Loisel et al.

General comments

This manuscript describes the construction of a synthetic dataset for optical studies in the ocean, using Hydrolight. This topic is very familiar to me right now as I am pursuing a similar goal, so it was an easy read. Authors must make sure that it is accessible to a broader audience though.

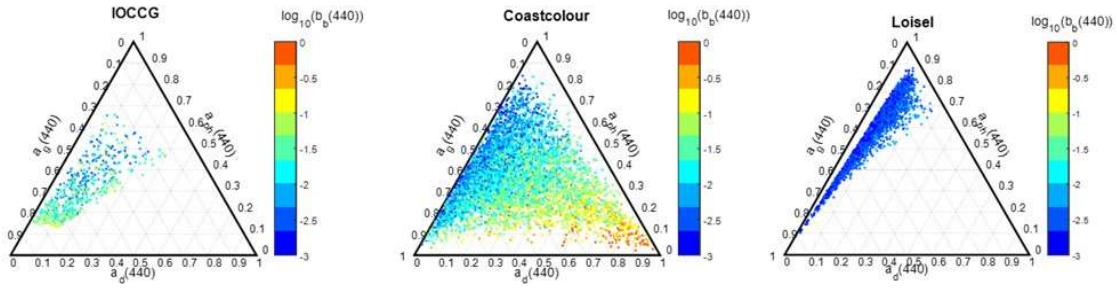
It proceeds the usual way, as in the old IOCCG dataset from Lee in 2003: first it assembles a set of phytoplankton absorption spectra, then the rest of IOPs are built with relationships that include some randomness. Finally, a single wind speed (5 m/s) and three sun angles (0°, 30° and 60°) are set, as well as various combinations of inelastic scattering on and off. I downloaded and saw the dataset as part of the review.

Things I liked:

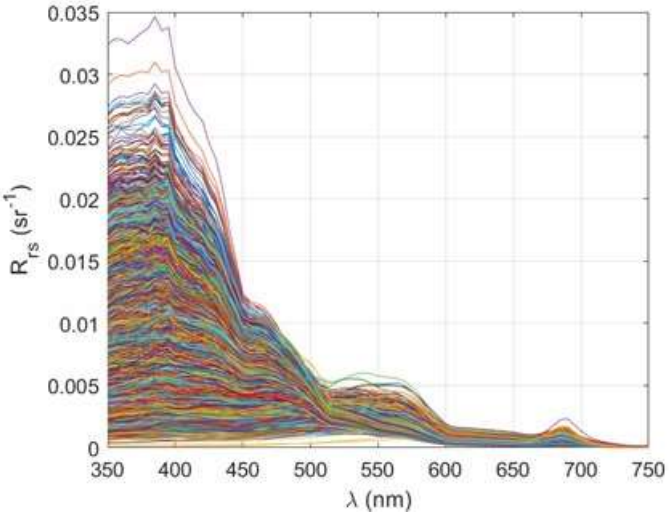
- The randomness in the bio-optical relationships, that will reproduce the spread in the relationships that is observed in nature.
- The Petzold phase function is abandoned and the much more realistic Fournier-Forand is considered for non-algal particles. Maybe a remark by the authors would be better.
- The 50 nm gap left for Raman scattering. In fact, I checked with my own simulations that the spectral memory of Raman scattering is about 50 nm, so it makes sense. A comment by the authors would be appreciated.
- The organization in netcdf files is quite handy compared to the Hydrolight text files.

Now I have a list of things I liked less:

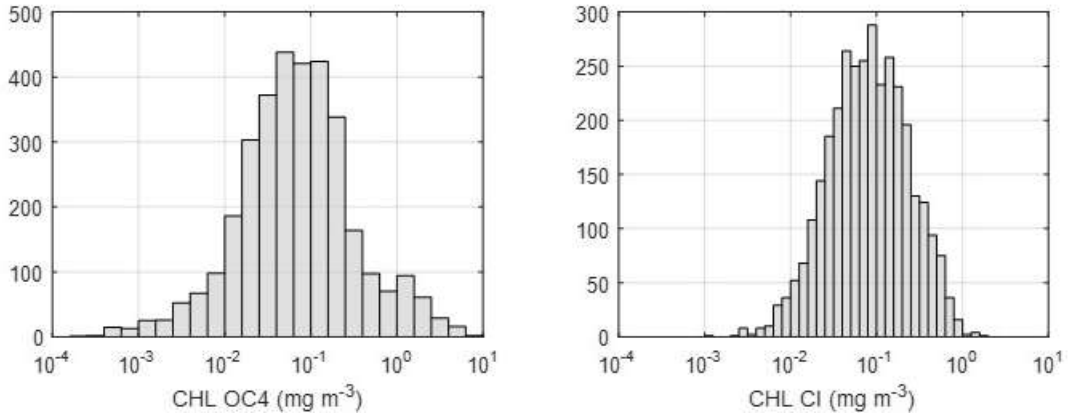
I have made a ternary plot of the absorption budget and I have compared it with the IOCCG (Lee) and the Coastcolour (Nechad) datasets. What I see here is a disproportionately low amount of non-algal particles, even compared to the IOCCG dataset, which was developed for ocean applications. I am not saying that IOCCG is right and this one is wrong, but authors should verify that such absorption budget is what it is actually found in the global oceans. Compared to other datasets, b_b appears lower too.



I have also plotted the remote-sensing reflectances (no inelastic scattering, sun at 30°):

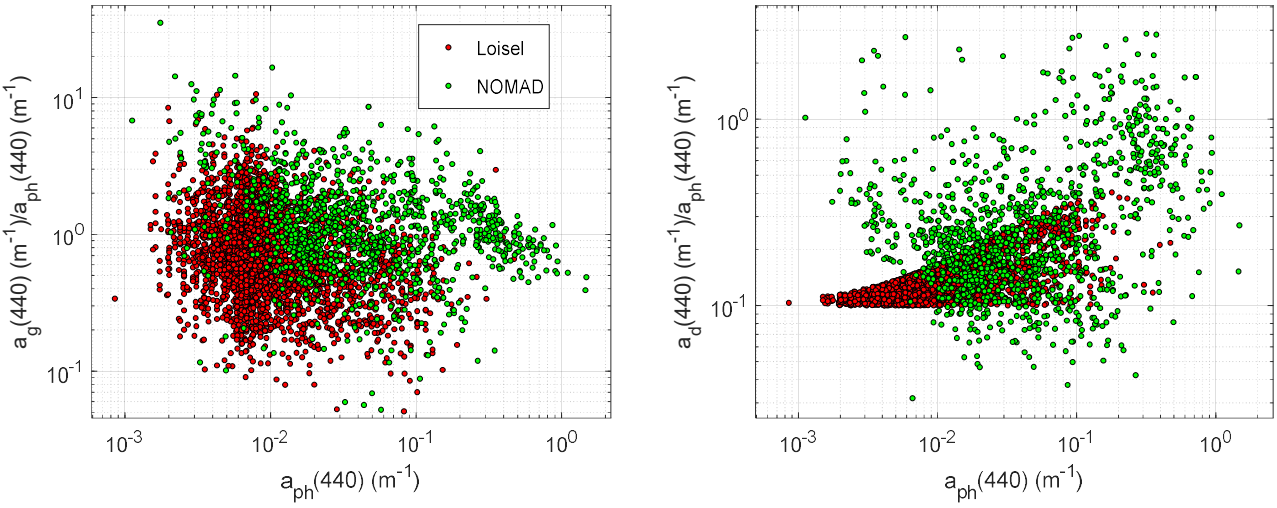


Some R_{rs} look crazy for me. I have never seen anything that high in the blue, even for the most oligotrophic waters. To verify, I have calculated the maximum band ratio (MBR) and I have applied the OC4 to it, according to O'Reilly and Werdell (2019). I have also calculated the chlorophyll index (CI), by Hu et al. (2012), for the most oligotrophic waters and I have applied his algorithm too. I get two chlorophyll histograms for the whole dataset:



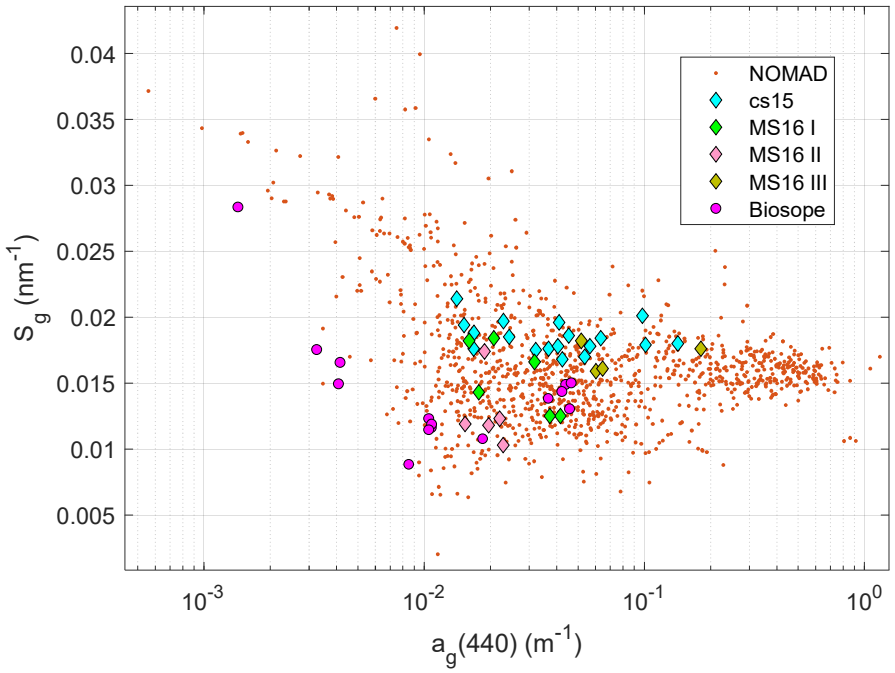
Considering that the lowest CHL measured in Valente et al. (2019), cited in O'Reilly and Werdell (2019), was 0.012 mg m^{-3}, that leaves us a very high amount of simulations whose CHL is unlikely low, whether we use OC4 or CI (Hu) to compare with. I also checked with Morel "clearest" waters and these values are definitely off. I therefore encourage redefinition of the dataset. I do not have an explanation for this artifact considering that the authors have reproduced the histograms seen by satellite data. I can hypothetise (1) the retrievals were biased the $a_{ph}(440)$ is actually higher or (2) the bio-optical relationships affect the CHL algorithm and need redefinition.

Related to this, there are datasets that may help in getting bio-optical relationships that are realistic. For example, I compared some absorption ratios to NOMAD:



I think I see that for the same $a_{ph}(440)$, there is a general lower value for $a_g(440)$ compared to NOMAD. Regarding $a_d(440)$, I see a lack of spread.

This is not the only example of what the authors can do. For example, I have plotted the CDOM slope S_g as a function of $a_g(440)$ for the NOMAD and Biosope datasets, as well as for three cruises in very clears waters of our group:



One can see some tendency to spread, especially to high S_g, when $a_g(440)$ is small, and a tendency to concentrate around Sg≈0.016 nm^{-1} for high $a_g(440)$. But the authors use a uniform distribution between 0.01 nm^{-1} and 0.02 nm^{-1}. This could therefore be improved.

I could revise the rest of IOPs and bio-optical relationships but I believe that at this point the authors got my point.

Specific comments

Abstract: it lacks a motivation on why another dataset is needed

Lines 51-52: “Recent technological developments and broader accessibility of optical in situ instrumentation” I believe this is unfortunately not the case. Seabird (old Wetlabs and Satlantic) has discontinued many in situ optical instrument, HobiLabs has closed and is not selling instruments anymore. All we have is Sequoia and Seabird in a situation of monopoly with little or no incentive to innovate and imposing high prices in already old design instruments, with a general lack of market competition.

Lines 60-63: the most important motivation for a synthetic dataset is that we will never have complete optical datasets across the widest dynamic range, and with declared and low uncertainties.

Lines 118:120: this is unclear to me.

Line 145: I would avoid the word "specific" as it is usually referred to the absorption divided by the concentration.

Lines 152-153: I think all IOPs matter equally, not only a_{ph}.

Line 160: “the measured values of $a_{ph}(\lambda)$ were used in the calculations of these IOPs”. Alright, but Lee did the same 20 years ago, so it is not a big novelty. I would not emphasize.

Lines 238-241: this comment is totally right. In fact, it is a pity that in 2023 there are still new datasets that are degrading spectral resolution to only few bands. Not to mention the aggregation of a_g and a_d in Valente, which makes us still rely on NOMAD when we want them separately.

On the reconstruction of hyperspectral a_{ph} from multispectral: I believe that a decently sized of hyperspectral a_{ph} can be compiled without the need to worry about this.

Line 276: When extrapolating a_{ph} to the UV, how is exactly the UV part “glued” to the rest?

Line 311: probably instead of “shifted”, I would say “biased”.

Lines 345-346: I think it is stated that the Mediterranean Sea is ultraoligotrophic, when it is actually not, not even the eastern basin (maybe this place in Summer, yes).

Line 460: “m^2/(mg Chla)”. Mass is mass, so please delete the “Chla”. Yes, it is common to write it like that among some biologists, but it does not make sense metrologically.

Lines 460-461: it is much more accurate to use a red wavelength of a_{ph} rather than a blue one to estimate CHL.

Lines 536-540: I wonder what are the reason to not consider the pure water measurements by Mason and Fry in 2016.

Lines 551-553: I wonder whether saving the whole profile is very useful, considering that Hydrolight already calculates for you the “K’s”, “z’s” and these depth-related quantities.

Figure 7 is not an efficient way to show the differences. Of course, everything increases with $a_{ph}(440)$ to a first order, but we want to know the differences among datasets. I prefer if the ratios are represented e.g., $a_g(440)/a_{ph}(440)$ as a function of aph(443), etc.

Line 602: “The scatter plots show a significant degree of overlap” Very roughly, but see my comment above.

Lines 688-689: there is no complementarity of this dataset and Nechad’s as both have different assumptions regarding the bio-optical modelling, so they are not consistent with each other.

The plots in Fig. 10 are not telling anything new as we know what happens with E_d profiles for different water types.