A hyperspectral and multi-angular synthetic dataset for algorithm development in waters of varying trophic levels and optical complexity

Jaime Pitarch, Vittorio Ernesto Brando

Comments

This paper presents a new synthetic data set linking apparent and inherent optical properties based on a very substantial set of radiative transfer simulations that are intended to provide comprehensive representation of optical water types found in nature. The purpose is to support ocean colour algorithm development and there is specific effort made to cover a wide range of sun sensor geometries, high spectral resolution and other important features. I am generally supportive of the effort and believe that the ambition of the work is significant. However, there are a couple of areas where I feel there are issues that might be either addressed or at least acknowledged before publication goes forward.

We thank the appreciation by the reviewer for the importance of this dataset and the amount of work that was put into it.

Limitations of measured data sets: One of the key themes of the paper is an ambition to better replicate the true range of variability found in nature. This is particularly emphasised with respect to oligotrophic waters which are reasonably claimed to be relatively under-sampled. In several sections, the authors point to existing field data sets and attempt to replicate all of the observed variability. Whilst this appears sensible on first inspection, I believe there is an underlying issue that needs to be considered. Essentially this boils down to the quality of field data. Any measurement is going to be subject to uncertainty and in many (most?) cases this uncertainty will become more significant as signal levels become smaller. This has been explored in some papers e.g. ref 1. Examples from the current manuscript that I think need to be considered include Figures 6 and 9 which both show apparently very strong variations in spectral slopes that just happen to coincide with signa levels dropping to very low levels. Is this real variability or is it the result of poor quality fits caused by limited data quality when signals are very low? Does it make sense to reproduce this level of variability in a synthetic data set if it is actually effectively noise and therefore potentially misleading? I think this at least needs to be considered.

This study relied on bio-optical data made publicly available in open access databases or single entries related to publications. Given the large amount of data that we gathered, we decided to apply stringent quality criteria. Specifically, for CDOM, it was required to follow an excellent fit to an exponential curve, thus ensuring that there were no issues with the spectral response or with the filtration. Most importantly, CDOM data had to be measured with Ultrapath devices, and the very long optical path should provide sufficient signal to noise ratio even in oligotrophic waters. Still, we acknowledge that there is an inevitable level of uncertainty related to in situ measurements.

There are also known issues with aspects of filter pad absorption measurements (Ref 2 pathlength amplification and baseline correction - the latter can also be an issue for CDOM absorption) that are not discussed but that could lead to significant discrepancies in observed data sets. These issues are effectively being baked into the training of this synthetic data set. The description of how these data were measured is lacking detail and I think there is scope to at least mention that there may be issues of this nature.

Processing differences among in situ datasets of filter pad data are not traced in the databases. That would require to go back to the original sources. It is, however, assumed that the practitioners, based on their experience, followed best practices. Indeed, most of these data come after the funding of projects by space agencies that involve related studies, and we believe that the groups that were involved were confident enough in the quality of the data before uploading. This said, the quality criteria to select valid filter pad

data was as stringent as for CDOM, so we are highly confident also here. But to keep the readers aware, we will include part of this comment in the manuscript.

Oligotrophic under-sampling: The authors make a significant play on extending coverage of oligotrophic waters that have been historically under-sampled. Whilst this is true, it remains the case that these waters have been sampled. I am concerned that Figure 9 appears to show at least a full order of magnitude of additional CDOM (ag440) range that has never been observed, even with Ultrapath CDOM sampling. I am perfectly happy to criticise measurement quality (see above) but I am a bit concerned about the justification for effectively inventing an additional decade of variability in this parameter? It is possible that community measurements have a lower limit that inhibits resolution of lower signals, but it is also potentially true that there is a background level of dissolved organic absorption that is a natural feature. I am not convinced that this aspect of the data set is as reliable as the paper currently suggests. Again, a more careful discussion of potential merit or otherwise would be advisable I think.

There is some misunderstanding here, which we aim at clarifying here and in the manuscript. $a_g(440)$ values ranging $2 \cdot 10^{-3}$ – $1 \cdot 10^{-2}$ were observed in South Pacific gyre during Biosope cruise as reported in Figure 15 of Bricaud et al. (2010), so we are not "effectively inventing an additional decade of variability in this parameter". Figure 9 in the manuscript is used for the construction of a rematonship between parametrization of S_g from $a_g(440)$, so we can predict the former from the latter. Only the CDOM data that passed the stringent quality control mentioned above are shown in the figure. It appears that those Biosope spectra were excluded for the parametrization of S_g: the text commenting Figure 9 will be updated to clarify the issue.

For the range of low CDOM values present in the synthetic dataset itself, we are in line with empirical data and another synthetic dataset (Loisel et al. 2023). To make it totally clear in the revised version, and clarify that the synthetic data set does not introduce an oligotrophic over-sampling, the lower-left panels in Figures 10 and 11 in the manuscript will be replotted showing $a_g(440)$ in the horizontal axes, and the related text will be updated accordingly to state that the synthetic dataset covers appropriately the $a_g(440)$ ranges of in situ and other synthetic datasets. The two panels are reported here as Figure R2.1. See how Loisel's dataset, the NOMAD dataset and our dataset have few points that down to about $a_g(440) \sim 3 \cdot 10^{-4}\ m^{-1}$.

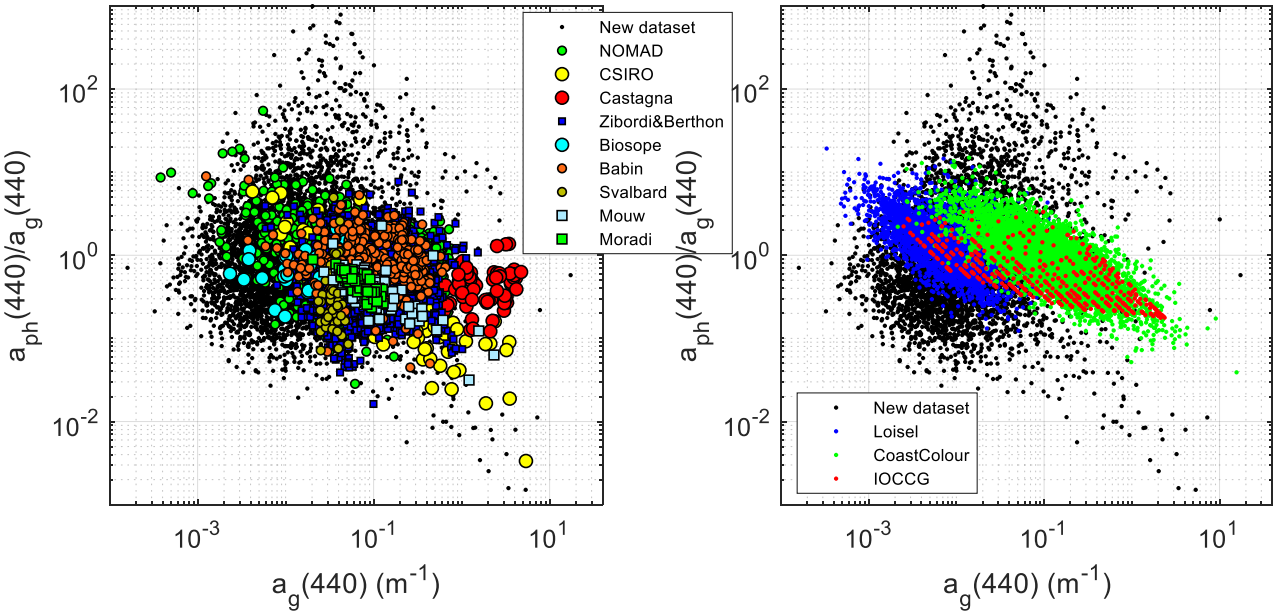

Figure R2.1 Cross-relationship comparison for $a_g(440)$ and the $a_{ph}(440)/a_g(440)$ between the synthetic dataset and (left) various in situ datasets (right) other synthetic datasets.

Parameterisation: The paper takes considerable effort to describe and justify construction of the bio-optical model and other aspects which go into parameterising the Hydrolight runs. Inevitably there are decisions that need to be made and options discarded as a result. This is fine, but in several cases here various decisions are presented as inevitable when in fact alternative option could have been chosen. I would not ask for these decisions to be revered or for models to be reworked in addition - that would be unfair. However, I think it is possible for the authors to recognise that alternatives would be available and might also be legitimate options.

We will provide more context to justify the main decisions. Still, justifying all possible options would make the manuscript much lengthier than it already is. It is usually accepted as good practice to properly validate or justify the option that was taken.

For example, they have opted to use the a version of the Hydrolight input generation where they calculate backscattering from backscattering ratios applied to scattering coefficients rather than directly inputting backscattering SIOPs.

Yes and no. Phytoplankton backscattering is calculated from backscattering rations applied to scattering coefficients as the reviewer points out, but that has been the choice for all datasets (IOCCG 2006; Loisel et al. 2023; Nechad et al. 2015). Just for the principle of parsimony, we stuck to the usual procedure here. Still, we went one step beyond them because we were able to justify this parameterization based on the independent data in Figure 4 of the manuscript.

For non-algal properties, it was actually backscattering that was fixed, after knowing absorption (Figure 7 of the manuscript). We will revise the text in the related areas to see if we can provide some additional justification of the modelling choices.

I can point to a small number of papers where there have been efforts made to directly estimate thee parameters (refs 3 and 4) and which would have provided alternative options that could be considered. Again, I would like to emphasise that I am not looking for more work to be done here, just that there is a slightly less emphatic description of what is possible and available (or not), taking into account material that is not hard to find in the literature.

Also here, we will try to add some more context in the method description, without departing too much from the main topic, which is the dataset description.

Validation: The synthetic data set produces hyperspectral remote sensing reflectance spectra that may be of great value for algorithm development. However, it is unclear how representative the simulated spectra actually are? The discussion of the outputs very rapidly branches off into cluster analysis and consideration of geometric effects, but there is no real analysis of how representative the spectra are of natural distributions. I would like to see a comparison with existing measured data sets to get a sense of where there are overlaps and divergences that may or may not be of interest when considering value as an allegedly global data set. I would emphasise that I have no trouble with the quality of the simulated reflectance spectra per se – Hydrolight will produce essentially the right reflectance spectrum for whatever conditions you tell it to work with. However, the value of this synthetic data set is very much in its ability to cover the range of naturally occurring variation and I would like to see harder evidence that it does this e.g. for turbid coastal waters as well as more open coastal and oceanic conditions.

The reviewer has a point here. In the revised version, the reviewer will find strong evidence that the generated reflectances are in line with existing measured datasets in what regards absorption, chlorophyll concentration, total suspended matter concentration, plus an addition evaluation through a spectral quality

index. All in all, the validation exercises support our dataset as representative of a wide range of natural waters. We paste them here:

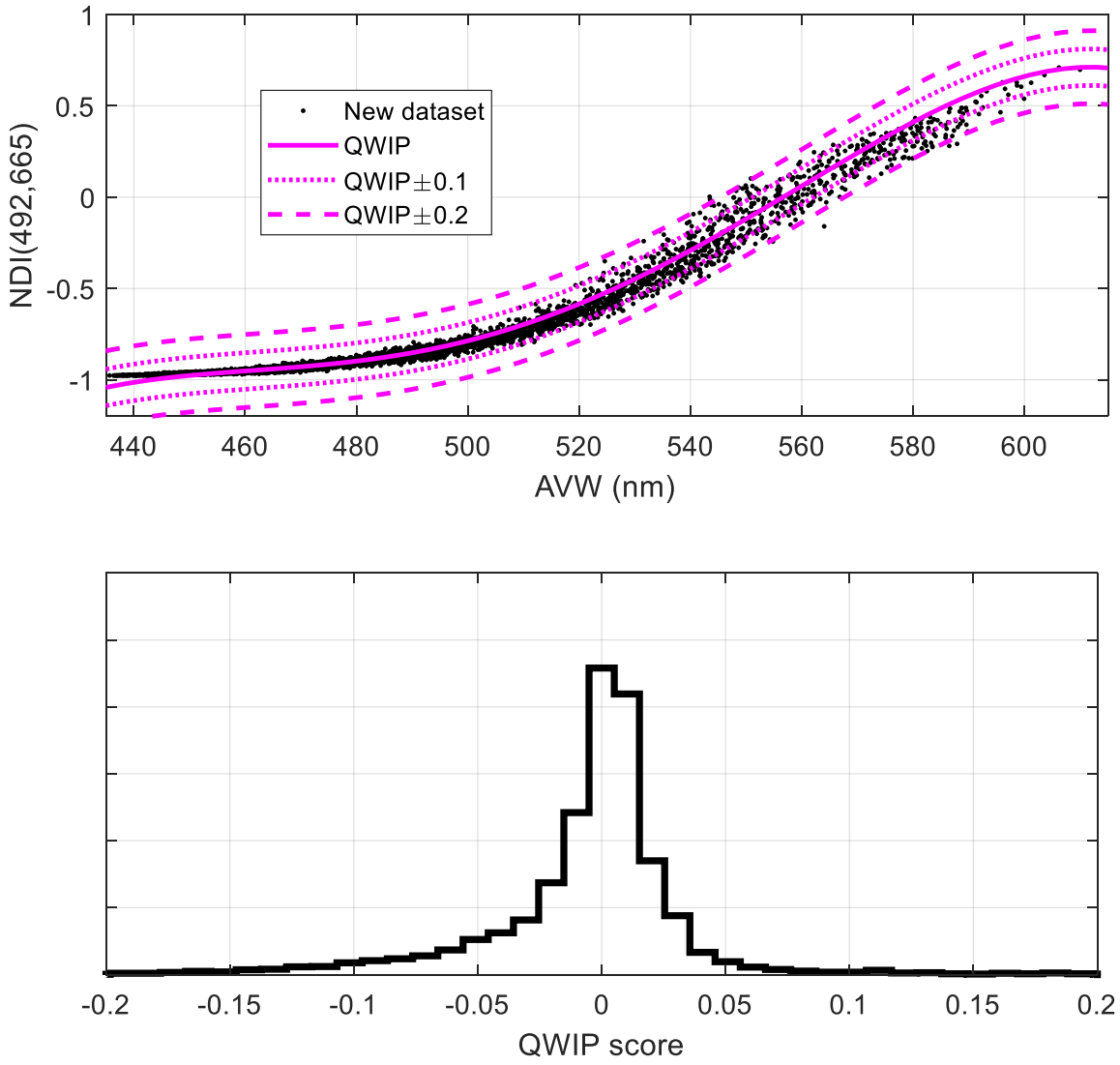

Figure R2.2 Upper plot: scatter plot between the apparent optical wavelength (Vandermeulen et al. 2020) and the NDI index: $NDI(492,665) = \frac{R_{rs}(665)-R_{rs}(492)}{R_{rs}(665)+R_{rs}(492)}$. Magenta lines: QWIP score (Dierssen et al. 2022) and error bars. Lower plot: histogram of the QWIP score, defined as the difference respect to the QWIP curve.

To generate Figure R2.2, we calculated the QWIP index by Dierssen et al. (2022) for our entire synthetic dataset. Such index aims at providing a quality estimate for a hyperspectral R_{rs}. QWIP was developed a large dataset of in situ R_{rs}, so this comparison is actually a comparison with real R_{rs} data. In Dierssen et al. (2022), it is mentioned that values within the 0.2 margins have high similarity to real spectra measured in the field, which are all 5000 but 7 spectra. Still, these 7 spectra are close to the limit, and may simply contain some bio-optical characteristics, not present in the QWIP calibration dataset. This comparison, therefore, gives confidence in the quality of our dataset.

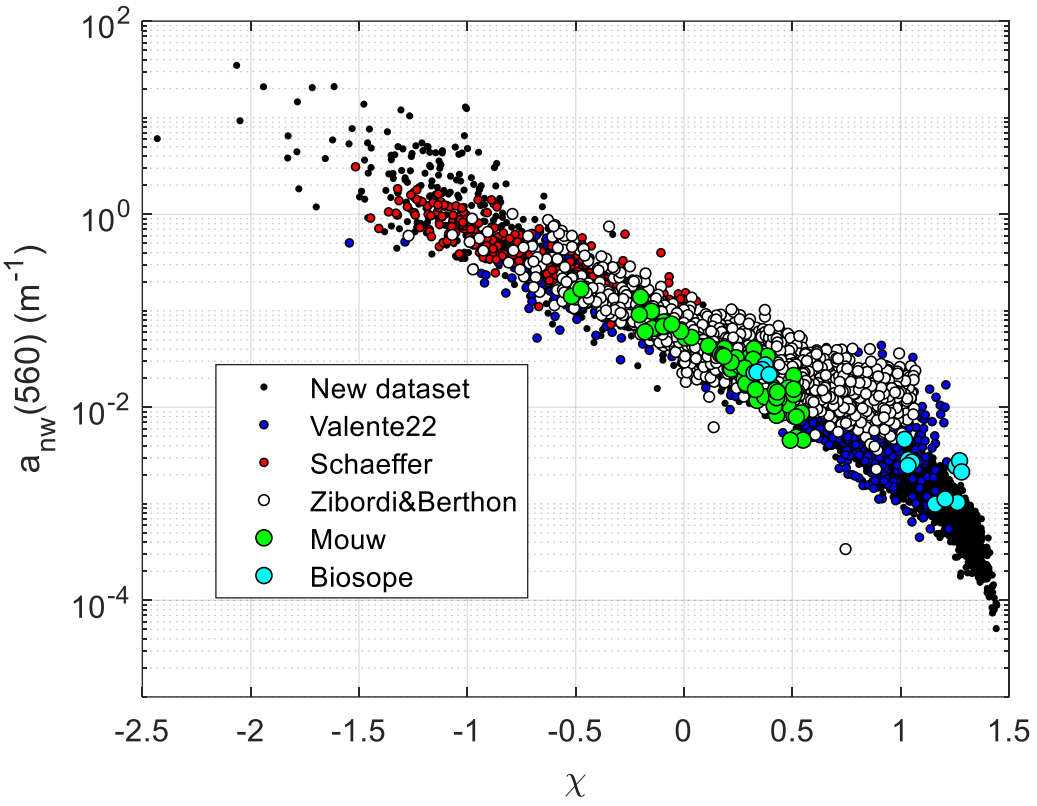

Figure R2.3 A scatter plot between the R_{rs}-generated χ index and the matched non-water absorption spectrum at 560 nm $a_{nw}(560)$. Black dots are from the synthetic dataset and coloured dots are from field data from various references (see text).

Figure R2.3 helps to assess the covariability of R_{rs} and the absorption coefficient. A one-dimensional predictor χ is derived from an R_{rs}:

$$\chi = \log_{10}\left(\frac{R_{rs}(443) + R_{rs}(490)}{R_{rs}(560) + 5\frac{R_{rs}^2(665)}{R_{rs}(490)}}\right)$$

This χ index is matched to non-water absorption spectrum at 560 nm $a_{nw}(560)$. There are several open access, freely available in situ datasets that contain both measured variables matched together, such as Valente et al. (2022), Zibordi and Berthon (2024) and the Schaeffer, Mouw and Biosope datasets (Casey et al. 2020). Figure R2.3 clearly shows the excellent average overlap between our synthetic dataset and measured data. Different bio-optical characteristics produce slight deviations from the mean curve, indicating natural variability.

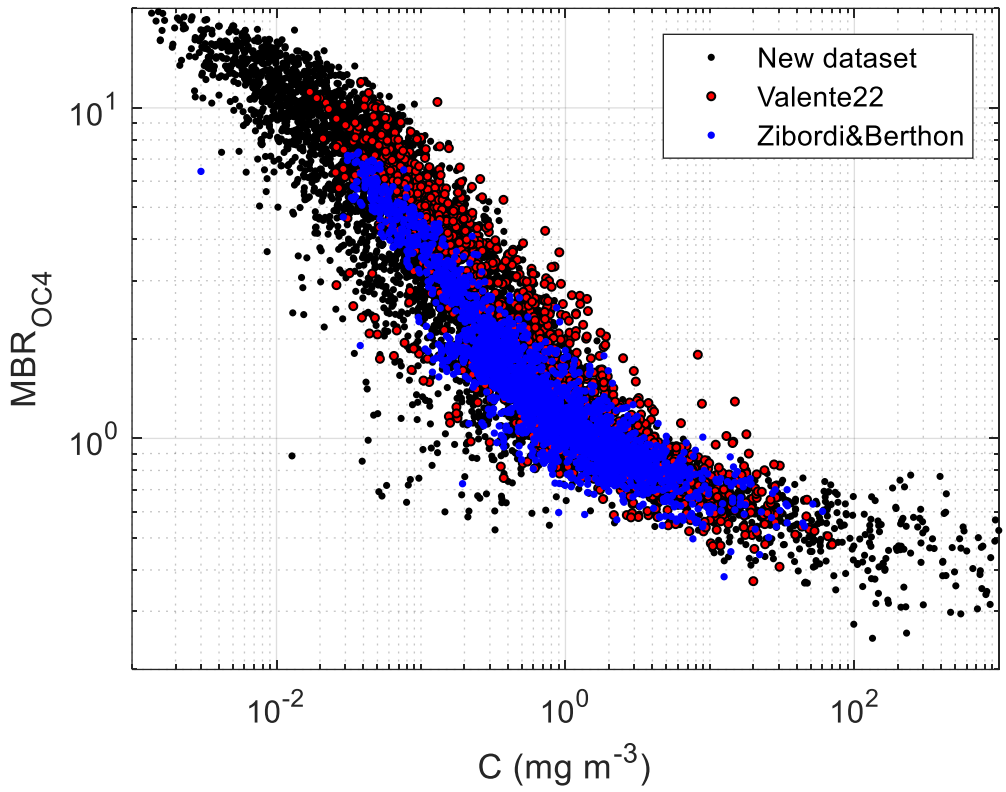

Figure R2.4 Chlorophyll concentration as a function of the maximum band ratio for OC4-type algorithms, for the synthetic dataset and for data in Valente et al. (2022) and Zibordi and Berthon (2024).

Figure R2.4 shows how a given chlorophyll concentration in the dataset relates to the generated R_{rs} through an index that is used to estimate chlorophyll in the ocean:

$$MBR_{OC4} = \frac{\max\left[R_{rs}(443), R_{rs}(490), R_{rs}(510)\right]}{R_{rs}(560)}$$

From R_{rs}, we calculate the maximum band ratio MBR_{OC4}, an index known to be a good predictor for its good correlation to chlorophyll concentration (C) in oceanic waters, but also used for studying the consistency of a given dataset in all kinds of water (Nechad et al. 2015). Here, matched MBR_{OC4} and chlorophyll concentration from two large in situ datasets are plotted (Valente et al. 2022; Zibordi and Berthon 2024), showing a good general overlap, though with some degree of differences among them, that are explainable due to different bio-optical characteristics of the seas sampled. Data from our dataset

generally agrees with the trend.

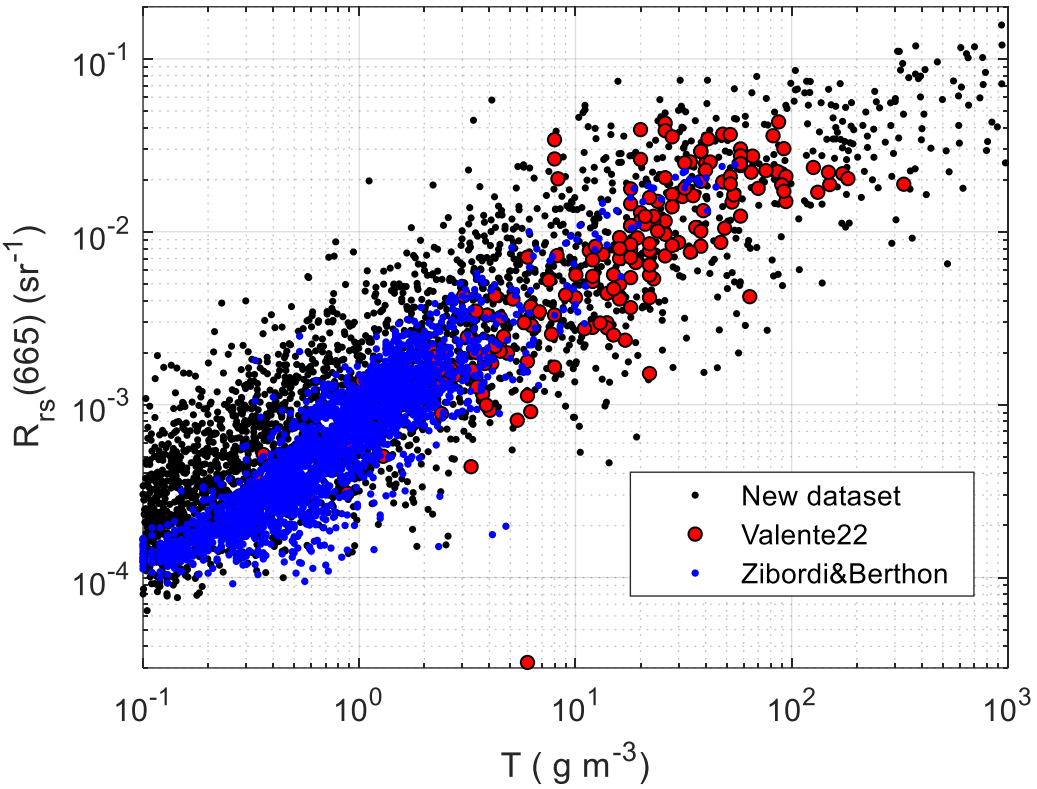

Figure R2.5 Total suspended matter concentration as a function of $R_{rs}(665)$, for the synthetic dataset and for data in Valente et al. (2022) and Zibordi and Berthon (2024).

The last comparison to real R_{rs} data involves the relationship to the total suspended matter concentration (T), relevant for coastal and inland water, which usually show higher turbidities. Our dataset does not need T for its generation, but it can be estimated as T=N+0.07C, after Brando and Dekker (2003), where N is the concentration of non-algal particles. It is known that T covaries with R_{rs} at long wavelengths, and 665 nm is commonly employed, due to the lesser disturbance by CDOM. Figure R2.5 shows that our dataset has a range of natural variability that includes that in in situ datasets (Valente et al. 2022; Zibordi and Berthon 2024), once more confirming the suitability of this new dataset for optical studies in all ranges of water.

The reader must note that for the new plots discussed above, the dot cloud amplitude in the in situ datasets is included in the synthetic dataset, meaning that the statistical treatment that was given to the inherent optical properties prior to radiative transfer simulations was such to ensure optical representativeness of many water types, as far as this plot is concerned.

Final comment: I have pointed to four references that are all from my own work. I am very uncomfortable doing this and I am NOT looking for these to be specifically referred to. They do, however, represent the basis for where my opinions have been shaped on these matters and where I believe we might have some philosophical differences that are not, however, insurmountable.

II would be more comfortable with a slightly less emphatic version of the paper that provides the reader with clear explanations of the decisions that were taken, but that notes that alternative options could have been taken in at least some cases.

The decisions that were taken were properly (although it may sound “emphatically”) justified in the paper more than in any other similar paper before, to the limit that the preprint already has 59 pages. However,

without prejudice to our responses to previous comments of this review, we will try to provide some more background and guidance to the reader on the option choice.

I genuinely think the authors need to carefully consider the rationale for reproducing all of the observed variability, including measurement uncertainties, some of which are very significant indeed.

The rationale behind this study is that not accounting for the variability would lead to a dataset that would be too self-similar, as was the case for the Coastcolour dataset. The choice incorporating the spread in crossed relationships into the bio-optical modelling goes back to the IOCCG dataset, and we applied here the same principle, with the difference that now we have much more data and the ability to constrain some crossed relationships that could not be constrained before. Unfortunately, the discussion on uncertainty is very difficult, since IOP measurements never come with an uncertainty estimate. We acknowledge that a part of the data cloud spread is due to measurement and errors and mismatches, but the validation exercises show that the reflectance shows patterns and dispersion that agrees with measurements.

Ultimately I would be unlikely to use this synthetic data set as I would struggle to accept some of the decisions that have gone into producing it, but I can imagine it being welcomed by a significant part of the community more or less as is.

We regret this reluctance as the dataset is already being used to develop the bidirectional reflectance correction algorithm to be operationally implemented in EUMETSAT's processing chain of OLCI Level 2 data. The dataset itself is valuable for many other algorithm calibration or validation purposes, but also the research leading to its generation represents a relevant contribution.

As with all of these things, the expression caveat emptor pertains. I hope that these comments will help to encourage a slightly less emphatic description of the data set and encourage potential users to be mindful of where the limitations might still be found.

We appreciate that the reviewer took the time to help improve the manuscript. The reviewer will see enhanced validation and explanations in the revised version. *In dubio pro reo*.

References

McKee, D., R. Röttgers, G. Neukermans, V. Sanjuan Calzado, C. Trees, M. Ampolo-Rella, C. Neil and A. Cunningham Impact of measurement uncertainties on determination of chlorophyll-specific absorption coefficient for marine phytoplankton. J. Geophys. Res. Oceans: 119, 9013–9025, doi:10.1002/2014JC009909, 2014.

Lefering, I., R. Röttgers, R. Weeks, D. Connor, C. Utschig, K. Heymann, and D. McKee Improved determination of particulate absorption from combined filter pad and PSICAM measurements, Opt. Express 24, 24805-24823, 2016.

Bengil, F., D. McKee, S. T. Beşiktepe, V. S. Calzado, and C. Trees A bio-optical model for integration into ecosystem models for the Ligurian Sea, Prog. Oceanography 149, 1-15, 2016.

Lo Prejato M., McKee D. (2023) Optical Constituent Concentrations and Uncertainties Obtained for Case 1 and 2 Waters From a Spectral Deconvolution Model Applied to In Situ IOPs and Radiometry. Earth and Space Science, 10 (12), art. no. e2022EA002815. DOI: 10.1029/2022EA002815

References

Brando, V. E., and A. G. Dekker. 2003. Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality. IEEE Transactions on Geoscience and Remote Sensing **41:** 1378-1387.

Bricaud, A., M. Babin, H. Claustre, J. Ras, and F. Tièche. 2010. Light absorption properties and absorption budget of Southeast Pacific waters. Journal of Geophysical Research: Oceans **115**.

Casey, K. A. and others 2020. A global compilation of in situ aquatic high spectral resolution inherent and apparent optical property data for remote sensing applications. Earth Syst. Sci. Data **12:** 1123-1139.

Dierssen, H. M., R. A. Vandermeulen, B. B. Barnes, A. Castagna, E. Knaeps, and Q. Vanhellemont. 2022. QWIP: A Quantitative Metric for Quality Control of Aquatic Reflectance Spectral Shape Using the Apparent Visible Wavelength. Frontiers in Remote Sensing **3**.

IOCCG. 2006. Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications, p. 1-122. *In* V. Stuart [ed.], Reports of the International Ocean-Colour Coordinating Group. International Ocean-Colour Coordinating Group, IOCCG.

Loisel, H., D. S. F. Jorge, R. A. Reynolds, and D. Stramski. 2023. A synthetic optical database generated by radiative transfer simulations in support of studies in ocean optics and optical remote sensing of the global ocean. Earth Syst. Sci. Data **15:** 3711-3731.

Nechad, B. and others 2015. CoastColour Round Robin data sets: a database to evaluate the performance of algorithms for the retrieval of water quality parameters in coastal waters. Earth Syst. Sci. Data **7:** 319-348.

Valente, A. and others 2022. A compilation of global bio-optical in situ data for ocean colour satellite applications – version three. Earth Syst. Sci. Data **14:** 5737-5770.

Vandermeulen, R. A., A. Mannino, S. E. Craig, and P. J. Werdell. 2020. 150 shades of green: Using the full spectrum of remote sensing reflectance to elucidate color shifts in the ocean. Remote Sensing of Environment **247:** 111900.

Zibordi, G., and J. F. Berthon. 2024. Coastal Atmosphere & Sea Time Series (CoASTS) and Bio-Optical mapping of Marine optical Properties (BiOMaP): the CoASTS-BiOMaP dataset. Earth Syst. Sci. Data Discuss. **2024:** 1-33.