Check for updates

LIMNOLOGY AND OCEANOGRAPHY
Letters

ASLO
Open Access

Limnology and Oceanography Letters 2023
© 2023 The Authors. *Limnology and Oceanography Letters* published by Wiley Periodicals LLC on behalf of Association for the Sciences of Limnology and Oceanography.
doi: 10.1002/lol2.10319

LETTER

How many independent quantities can be extracted from ocean color?

B. B. Cael [1]* *Kelsey Bisson* [2] *Emmanuel Boss* [3] *Zachary K. Erickson* [4]

[1]National Oceanography Centre, Southampton, UK; [2]Oregon State University, Corvallis, Oregon; [3]University of Maine, Orono, Maine; [4]NOAA Pacific Marine Environmental Laboratory, Seattle, Washington

Scientific Significance Statement

The reflectance of sunlight from the ocean can be observed from satellites and is used to derive many biologically relevant parameters, such as the concentration of chlorophyll in the upper ocean. Reflectances are currently observed at about 10 different wavelengths, but this will soon be expanded to hundreds with the upcoming launch of a new ocean color satellite, PACE, in early 2024. Many new algorithms are being proposed to make use of the wealth of ocean color data which will be provided. However, there are strong correlations between reflectances at different wavelengths; these correlations mean there will be far fewer products that can be independently derived than there will be reflectance wavelengths observed. Here, we use ship-based measurements similar to what will be provided from PACE to suggest that, on a global scale, only a few independent variables can be calculated from hundreds of reflectance wavelengths. Current and past satellites provide a similar amount of independent data to what is projected from PACE. We then show that, on a global scale, a set of six derived parameters only contains one independent piece of information, suggesting that more information exists in ocean color data than is being currently used.

Abstract

Products derived from remote sensing reflectances ($R_{rs}(\lambda)$), for example, chlorophyll, phytoplankton carbon, euphotic depth, or particle size, are widely used in oceanography. Problematically, $R_{rs}(\lambda)$ may have fewer degrees of freedom (DoF) than measured wavebands or derived products. Here, we show that a global sea surface hyperspectral $R_{rs}(\lambda)$ dataset has DoF = 4. MODIS-like multispectral equivalent in situ data also have DoF = 4, while their SeaWiFS equivalent has DoF = 3. Both multispectral-equivalent datasets predict individual hyperspectral wavelengths' $R_{rs}(\lambda)$ within nominal uncertainties. Remotely sensed climatological multispectral $R_{rs}(\lambda)$ have DoF = 2, as information is lost by atmospheric correction, shifting to larger spatiotemporal scales, and/or more open-ocean measurements, but suites of $R_{rs}(\lambda)$-derived products have DoF = 1. These results suggest that remote sensing products based on existing satellites' $R_{rs}(\lambda)$ are not

*Correspondence: cael@noc.ac.uk

Associate editor: Raphael Kudela

Author Contributions Statement: BBC lead and all other authors assisted with all aspects of this study.

Data Availability Statement: Code is available at https://github.com/bbcael/eifoc. Remote sensing data were downloaded from https://oceancolor.gsfc.nasa.gov/ and http://sites.science.oregonstate.edu/ocean.productivity/index.php. In situ data were downloaded from https://seabass.gsfc.nasa.gov/.

[Correction added on Mar 22, 2023, after first online publication: "In situ data were downloaded from https://seabass.gsfc.nasa.gov/." is included as the last sentence in Data Availability Statement section]

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

independent and should not be treated as such, that existing $R_{rs}(\lambda)$ measurements hold unutilized information, and that future multi- or especially hyper-spectral algorithms must rigorously consider correlations between $R_{rs}(\lambda)$ wavebands.

Ocean color satellites have revolutionized the study of ocean ecology and biogeochemistry in recent decades by providing a near-continuous global picture of surface ocean properties (Hovis et al. 1980; O'Reilly et al. 1998). Satellites measure the spectral radiance emanating from the ocean and atmosphere. Remote sensing reflectance ($R_{rs}(\lambda)$) is obtained following the removal of the contribution of atmospheric and surface effects and normalization to downwelling solar irradiance. Algorithms have been developed to estimate numerous biogeochemcally relevant surface variables from $R_{rs}(\lambda)$, such as chlorophyll concentration (Chl, [$\mu g\,L^{-1}$]) (O'Reilly et al. 1998; Hu et al. 2012), the spectral slope of the particle size distribution (ξ) (Kostadinov et al. 2009), the concentrations of phytoplankton and particulate organic and inorganic carbon (C_{phyto}, POC, and PIC, [$\mu g\,L^{-1}$]) (Graff et al. 2015; Evers-King et al. 2017; Mitchell et al. 2017), euphotic layer depth (Z_{eu} [m]) (Lee et al. 2007), and, using additional input variables, net primary production (NPP, [$mg\,m^{-2}\,d^{-1}$]) (Behrenfeld and Falkowski 1997; Westberry et al. 2008; Silsbe et al. 2016). Such products are used in a wide variety of applications, such as validation of complex ocean ecosystem and biogeochemistry models (Dutkiewicz et al. 2020; Cael et al. 2021) or as inputs for simpler models that predict other variables such as vertical particulate organic carbon fluxes from ocean color (Siegel et al. 2014; Cael et al. 2017; DeVries and Weber 2017; Bisson et al. 2020; Nowicki et al. 2022).

Existing $R_{rs}(\lambda)$ data are multispectral, measured at several wavebands. Derived products generally rely only on a subset of these wavebands and are commonly expressed as functions of band ratios between just two wavelengths (Hu et al. 2012). Some algorithms attempt to simultaneously estimate multiple products to match the full $R_{rs}(\lambda)$ spectrum, for example, the generalized inherent optical properties approach (Werdell et al. 2013). However, the most widely used products, such as for Chl and POC, treat all outputs as independent quantities and are fully empirical.

Correlations between $R_{rs}(\lambda)$ at different wavebands are strong (Huot and Antoine 2016) presenting multiple potential issues for both users and developers of derived products. If multiple products are used simultaneously and treated as independent when they are in fact not, this can lead to overconfidence in model skill or miscalculation of uncertainties. Adding different (yet correlated) satellite products to a model can result in model output redundancy (Bisson et al. 2020). These issues will be exacerbated by the hyperspectral resolution of the next generation of ocean color satellites, namely the Plankton, Clouds, Aerosols, and Ecosystems (PACE) satellite scheduled to launch January 2024 (Werdell et al. 2019). In addition to the common suite of multispectral products, PACE also plans to enable characterizations of phytoplankton communities, for example, Chase et al. (2017), substantially increasing the number of products available from $R_{rs}(\lambda)$.

The strong correlations among $R_{rs}(\lambda)$ wavelengths can be framed in terms of the degrees of freedom (DoF) of $R_{rs}(\lambda)$ measurements and suites of derived products. DoF represents the effective number of dimensions of a dataset after accounting for correlations and uncertainties between variables and is in essence the number of independent variables in that dataset. It has been shown that the DoF of globally distributed near-surface measured hyperspectral absorption spectra is ~ 5 (Cael et al. 2020). This could be considered a possible upper limit for the DoF of satellite-measured $R_{rs}(\lambda)$ given higher uncertainties on satellite measurements—particularly associated with atmospheric correction (Cael et al. 2020; Bisson et al. 2021). The DoF of PACE's hyperspectral measurements might then be expected to be much lower than the number of wavelengths for which it will measure $R_{rs}(\lambda)$, which will appreciably affect how hyperspectral satellite $R_{rs}(\lambda)$ products should be constructed. For both existing and future satellite $R_{rs}(\lambda)$, understanding the DoF of $R_{rs}(\lambda)$ measurements and derived products is crucial for appropriate usage and optimal construction of such products.

Here, we investigate the DoF of $R_{rs}(\lambda)$. We show that a global sea surface hyperspectral $R_{rs}(\lambda)$ database has four DoF. Coarsening hyperspectral $R_{rs}(\lambda)$ to their moderate resolution imaging spectrometer (MODIS) equivalent retains four DoF, though the sea-viewing wide field of view sensor (SeaWiFS) equivalent only has three DoF. Both multispectral equivalents predict individual hyperspectral $R_{rs}(\lambda)$ wavelengths within nominal uncertainties for satellite sensors. For climatological $R_{rs}(\lambda)$ and derived products, both MODIS-Aqua and SeaWiFS $R_{rs}(\lambda)$ have two DoF, suggesting $R_{rs}(\lambda)$ complexity is lost either through atmospheric correction, relatively more inclusion of open-ocean data, or averaging over larger scales in space and time. Suites of derived products only retain one DoF. Therefore derived products should not be treated as independent by users. These findings have substantial implications for the construction and use of multispectrally and hyperspectrally derived ocean color products.

Sea surface R_{rs}: Hyperspectral vs. multispectral

We first analyze a global sea surface hyperspectral $R_{rs}(\lambda)$ dataset to determine its DoF and how the DoF depends on spectral resolution (Chase et al. 2017; Kramer et al. 2022). The dataset includes $R_{rs}(\lambda)$ data at 191 locations at an effective 3.35 nm resolution (Chase et al. 2017) from 400 to 800 nm, linearly interpolated to 1 nm (Fig. 1). We trimmed spectra to

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

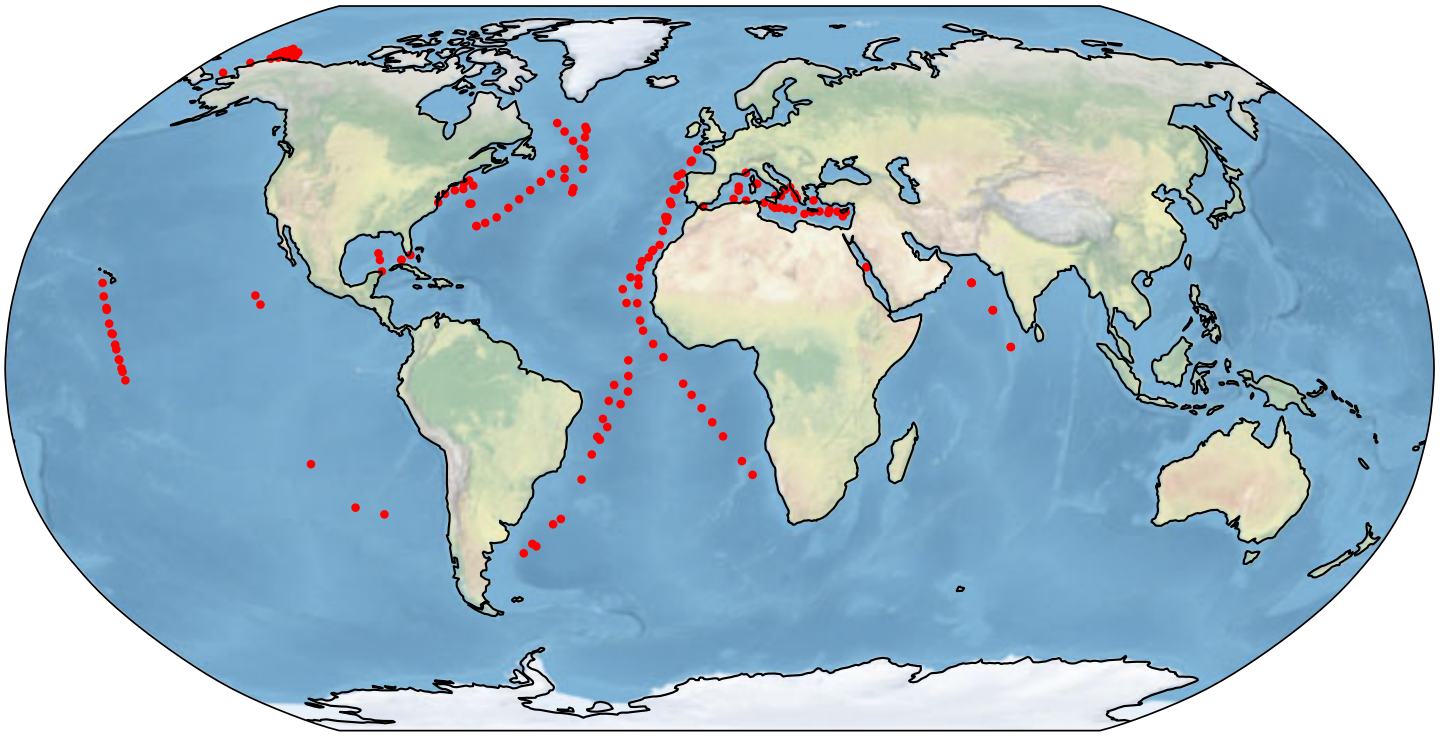

Fig. 1. Locations of the 191 stations considered in this study (red dots).

700 nm due to the large fraction of missing values and zeros >700 nm; our conclusions are not affected by using a lower or higher maximum wavelength. The dataset includes measurements taken from 2004 to 2018 evenly distributed across months of the year, and from all major ocean basins ranging in latitude from 41°S to 74°N. We also compare these data to their MODIS-Aqua and SeaWiFS multispectral equivalents by convoluting the hyperspectral $R_{rs}(\lambda)$ with the MODIS-Aqua and SeaWiFS spectral response functions (available at https://oceancolor.gsfc.nasa.gov/docs/rsr/HMODISA_RSRs.txt and https://oceancolor.gsfc.nasa.gov/docs/rsr/SeaWiFS_RSRs.txt) to generate 10-waveband and 6-waveband datasets which correspond to what each instrument would have measured from the same optical input that the radiometer received when generating the hyperspectral $R_{rs}(\lambda)$ data.

We then apply principal component analysis (PCA) (Wold et al. 1987) to these 301-, 10-, and 6-dimensional $R_{rs}(\lambda)$ datasets. PCA is a widely used method to reduce the dimensionality of datasets by identifying orthogonal vectors that explain the most variance in the data. PCA is linear in nature, which may result in an overestimation of effective dimensions by poorly approximating nonlinear relationships between variables (e.g., a PCA performed on the pair (x, y) where $y = x^2$ will yield two DoF). Nonlinear generalizations do exist (Weinberger et al. 2004; Scholz et al. 2008), though these are less widely applied due to their additional complexity and computational requirements that make interpretation challenging. One may therefore consider the DoF we report to be upper bounds. We perform a PCA on each $R_{rs}(\lambda)$ dataset, after standardizing each waveband. We use the broken-stick rule to choose the DoF, which states that the DoF is equal to the

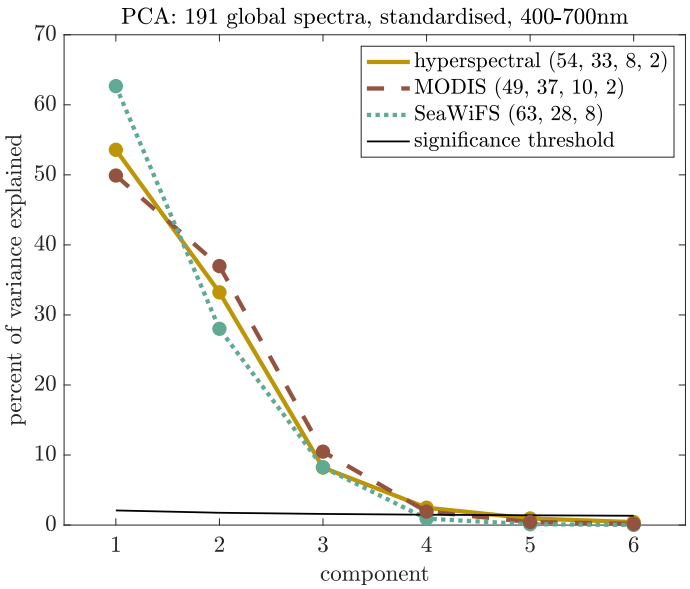

Fig. 2. Scree plot of percent variance explained vs. component for hyperspectral $R_{rs}(\lambda)$ dataset and MODIS-Aqua and SeaWiFS equivalents calculated from their spectral response functions. Black line indicates broken-stick significance threshold for hyperspectral data; numbers in legend give percent variance explained for each mode above this threshold in each case.

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

number of components that explain more variance than would be expected by randomly distributed data; this method was shown to be more consistent than a suite of others in a comparison (Jackson 1993; note that this threshold $b(i) = \frac{1}{n}\sum_{k=i\ldots n} 1/k$ is a function of the dimensionality of the dataset, so the significance cutoffs are different in Figs. 2, 4). These results can be shown visually as a "scree" plot, which plots the percentage of variance explained by each component and for randomly distributed data; the DoF is the number of components with a higher percentage of variance explained than would be expected for randomly distributed data. Our figures also visibly demonstrate that one would get the same results from using the scree plot rule, which states that the DoF is equal to the number of components not sitting on the straight line made by the higher-order components, and was found to consistently capture the correct DoF plus one when the first point on this straight line was included (Jackson 1993).

PCA reveals the hyperspectral $R_{rs}(\lambda)$ dataset has four DoF (Fig. 2); the first four components explain 54%, 33%, 8%, and 2%, totalling 97%, of the variance. The first four MODIS-Aqua equivalent $R_{rs}(\lambda)$ principal components have very similar percentages of variance explained: 49%, 37%, 10%, and 2%, totalling 99% of the total variance. In contrast, the first three SeaWiFS equivalent $R_{rs}(\lambda)$ principal components explain 63%, 28%, and 8%, totalling 99%, of the variance. This suggests that the hyperspectral $R_{rs}(\lambda)$ have four DoF, or four independent variables within the data, and that these four variables are effectively captured when reducing spectral resolution to the 10 MODIS-Aqua wavebands, but not to the 6 SeaWiFS wavebands.

Note that this difference of 1 DoF between the SeaWiFS- and MODIS-equivalent data is because MODIS includes a band centered at 645 nm which captures spectral variation in the range of $\sim$ 610–650 nm, a spectral region which is not covered by any SeaWiFS waveband. When this 645 nm-centered waveband is excluded from the MODIS analysis, we find three DoF for the remaining nine MODIS wavebands (this is not true of other wavebands, e.g., the fluorescence waveband at 673–683 nm). Irrespective of the exact number of DoF, Fig. 2 demonstrates that the bulk of the variance in these data, whether hyperspectral or multispectral, collapses along a few modes of variation, with the first two modes containing almost all ($\sim$ 90%) of the explanatory power. This suggests that only two independent quantities can be estimated accurately with these data. Even if one or two additional quantities can be estimated independently (i.e., DoF = 3 or 4), regardless of the exact number of quantities, these will necessarily be estimated with low signal-to-noise ratios.

The ability of coarsened, MODIS-equivalent data to obtain the same number of DoF as the hyperspectral dataset is further supported by predictions of hyperspectral $R_{rs}(\lambda)$ from multispectral equivalents. To illustrate this, for each

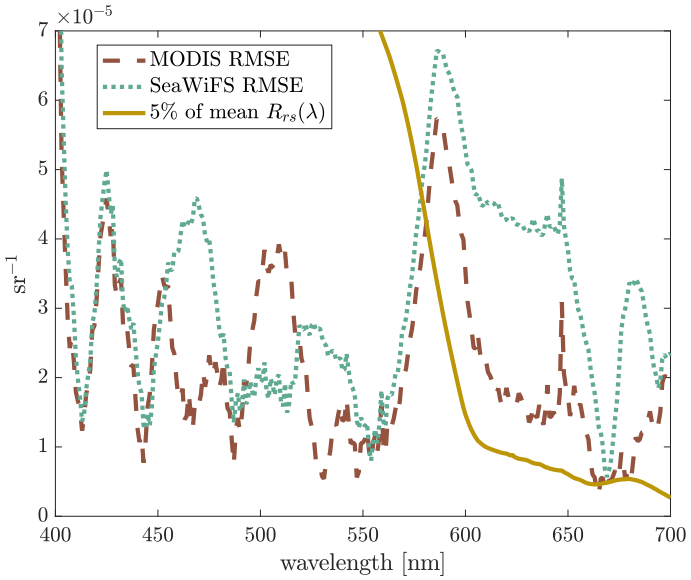

Fig. 3. Root-mean-square-error of multivariate linear regressions of each hyperspectral wavelength vs. the MODIS-Aqua and SeaWiFS equivalent $R_{rs}(\lambda)$. Solid line is 5% of the mean of each wavelength's hyperspectral $R_{rs}(\lambda)$.

hyperspectral wavelength we perform a multivariate linear regression of $R_{rs}(\lambda)$ at that wavelength regressed against $R_{rs}(\lambda)$ at each waveband of both the MODIS-Aqua and SeaWiFS equivalent $R_{rs}(\lambda)$. For all wavelengths < 578nm in the SeaWiFS case and 582 nm in the MODIS-Aqua case, the root-mean-square-error (RMSE) is smaller than 5% of the mean $R_{rs}(\lambda)$ at that wavelength, where 5% is a nominal relative uncertainty for satellite $R_{rs}(\lambda)$ (Fig. 3). Even for wavelengths greater than this, the RMSE is still very small in absolute terms, $< 0.00007\,\mathrm{sr}^{-1}$, far smaller than the nominal $0.0003\,\mathrm{sr}^{-1}$ absolute error for 1 km-by-1 km pixels for PACE (Gordon and Wang 1994). This underscores the extent to which different wavelengths' $R_{rs}(\lambda)$ are correlated and explains the ability of MODIS-Aqua equivalent multispectral $R_{rs}(\lambda)$ to preserve the dimensionality of hyperspectral $R_{rs}(\lambda)$. The fact that SeaWIFS-like $R_{rs}(\lambda)$ can accurately predict hyperspectral $R_{rs}(\lambda)$ to within PACE uncertainties but has fewer DoF than the in situ hyperspectral dataset is a reflection of the lower uncertainty on the in situ dataset than the expected PACE $R_{rs}(\lambda)$, and suggests that PACE $R_{rs}(\lambda)$ may have fewer DoF than the in situ hyperspectral dataset.

Climatologies: R_{rs} vs. products

The analysis above is based on instantaneous, sea surface $R_{rs}(\lambda)$ values. The power of satellite $R_{rs}(\lambda)$ and derived products, however, lies in their near-continuous global spatial coverage, and many users are primarily interested in climatological data, that is, the the coarsest spatial and

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

temporal scales. In this section we therefore analyze climatological $R_{rs}(\lambda)$ and derived products, again via PCA to determine DoF.

We generated a $1^\circ \times 1^\circ$ monthly climatology for SeaWiFS $R_{rs}(\lambda)$, excluding 2009–2010 due to known instrument issues (Siegel et al. 2014), using data downloaded from https://oceancolor.gsfc.nasa.gov/. Note that as we are using climatological data, this is a global scale analysis. We did the same for MODIS-Aqua, spanning July 2002–June 2022. We generated analogous climatologies, from each satellite over the same periods and spatiotemporal resolutions, for Chl, C_{phyto}, POC, PIC, Z_{eu}, ξ, the fraction of biovolume in the microplankton size class f_{micro} calculated from ξ as described in (Kostadinov et al. 2009), the particulate backscatter to chlorophyll ratio b_{bp} : Chl, and NPP as estimated by the CAFE (Silsbe et al. 2016) and CbPMv2 (Westberry et al. 2008) models. Chl, POC, and PIC were downloaded from https://oceancolor.gsfc.nasa.gov/, as was b_{bp} to calculate C_{phyto} according to Graff et al. (2015) and b_{bp} : Chl and the diffuse attenuation coefficent at 490 nm to calculate Z_{eu} according to Lee et al. (2007); SeaWiFS ξ and f_{micro} were derived as in Kostadinov et al. (2009); and NPP products were downloaded from http://sites.science.oregonstate.edu/ocean.productivity/index.php. In total we then have climatologies for MODIS-Aqua, SeaWiFS $R_{rs}(\lambda)$, and 10 derived products. We consider the six products Chl, C_{phyto}, POC, PIC, ξ, and Z_{eu}, to be core products and f_{micro}, b_{bp} : Chl, CAFE NPP, and CbPMv2 NPP to be ancillary products as these are either derived from the core products or rely on ancillary data other than $R_{rs}(\lambda)$.

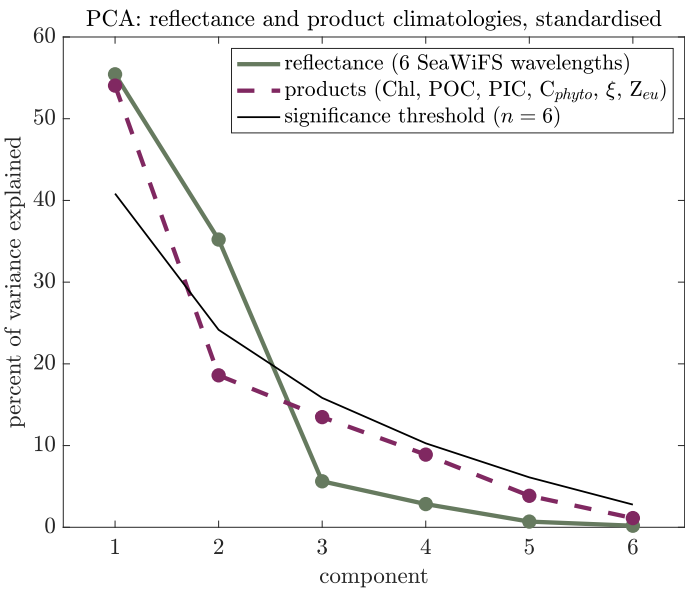

Fig. 4. Scree plot of percent variance explained vs. component for climatologies of SeaWiFS $R_{rs}(\lambda)$ and of six SeaWiFS $R_{rs}(\lambda)$-derived products. Black line indicates broken-stick significance threshold for six-dimensional data.

We note that a PCA on the MODIS-Aqua climatologies of $R_{rs}(\lambda)$ and products other than ξ and f_{micro} yields the same results as those for SeaWiFS below, so we focus here only on the SeaWiFS climatologies because ξ and f_{micro} are not readily available for MODIS-Aqua. We find two DoF for climatological SeaWiFS $R_{rs}(\lambda)$, but only one for the products (Fig. 4). This result is not sensitive to which combination of products is used; for example, including all the ancillary products as well still results in one DoF for the products. This result is also not sensitive to log-transformations of the variables that are log-normally (e.g., Chl, POC, PIC, C_{phyto}; Campbell 1995) or log-skew-normally (e.g., NPP; Cael 2021; Cael et al. 2018) distributed, or removal of outliers, zeros, or negative values.

That $R_{rs}(\lambda)$ have more DoF for the data in the previous section than for satellite-derived climatologies suggests that some reduction of complexity of the data occurs via some combination of increased sensor noise relative to ship-based data, atmospheric correction, or averaging over large space and time scales (Scott and Werdell 2019). Two DoF remain in satellite climatological $R_{rs}(\lambda)$ for both SeaWiFS and MODIS-Aqua, indicating the possibility of generating two independent products from these data. The suite of products tested above, however, has one DoF. This is likely due to derived products' appreciable uncertainties and/or strong correlations with chlorophyll. POC, ξ, and Z_{eu}, for instance, have Spearman rank correlations (across all months and 1° grid cells) of >0.9 with Chl. C_{phyto}'s rank correlation with Chl is still fairly high, at 0.61, and is lower largely due to small fluctuations when both are small; a simple spline fit of $\log(C_{phyto})$ against $\log(\text{Chl})$ yields an r^2 of 0.7.

The exception is PIC, which has a rank correlation with Chl of 0.11. For typical $R_{rs}(\lambda)$ values, however, PIC is highly uncertain—that is, PIC estimates are very sensitive to small variations in $R_{rs}(\lambda)$—as substantiated by the following analysis. We performed a simple sensitivity analysis with the standard two-band PIC algorithm used by NASA for all but the most optically bright waters (see https://oceancolor.gsfc.nasa.gov/atbd/pic/). We calculated PIC for the climatological median $R_{rs}(\lambda)$ at 443 and 555 nm and for 5% variations, converting to normalized water-leaving radiance by multiplying by the global mean extraterrestrial solar irradiance. We then perturbed these $R_{rs}(\lambda)$ values with Gaussian noise at the 5% level, corresponding to the nominal uncertainty in $R_{rs}(\lambda)$. This noise at 443 nm results in 68% noise in PIC. By contrast, POC only varies 5% with these 5% variations in $R_{rs}(\lambda)$ at either wavelength. This indicates that in the bulk of cases, satellite-derived PIC is highly uncertain, on the order of 70% (and note the PIC uncertainty will be magnified more when considering documented uncertainties for $R_{rs}(\lambda)$ of 15–40% in some regions; Bisson et al. 2021). In contrast, for relatively bright waters, the same exercise resulted in PIC variations of $<10\%$, indicating that this algorithm performs well in instances when PIC values are high. Nonetheless, the high

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sensitivity to typical uncertainty in $R_{rs}(\lambda)$ for median waters explains why we find one DoF for the products even though PIC and Chl are not strongly correlated: derived PIC is noisy most of the time.

These results have two key implications. One is that there is additional information in climatological $R_{rs}(\lambda)$ that is not included in current derived products, because climatological $R_{rs}(\lambda)$ has more DoF than a suite of climatological products. The other implication is that these products are not at all independent, because a suite of them only has one DoF. For instance, a numerical ecosystem model that reproduces the satellite-derived climatology of chlorophyll and of the particle size distribution's spectral slope should not be considered to be capturing two independent properties of the Earth system. When using satellite products as inputs to other models, these products and their propagated uncertainties must be treated simultaneously rather than independently.

The results presented here are appropriate for global and hence primarily open ocean analyses, composed primarily of Case 1 waters where optical variability is dominated by chlorophyll (Morel and Prieur 1977). It is therefore arguably unsurprising that the suite of $R_{rs}(\lambda)$-derived products produced one DoF. Coastal and inland waters' optical variability is influenced by other constituents, such as colored dissolved organic material (CDOM), inorganic particles, and other pigments (Brown et al. 2008; Nelson and Siegel 2013). Analyses focused on these waters is likely to reveal a higher number of DoF from both $R_{rs}(\lambda)$ and derived products. However, we note that the in situ dataset used here (Fig. 1) represents waters with $R_{rs}(\lambda)$ variability similar to that of the ocean as a whole, which can be seen by comparing the variation in $R_{rs}(\lambda)$ at each MODIS-Aqua wavelength from global satellite data with the same satellite data subsampled to the locations with in situ measurements (or the closest non-cloudy location). Subsampled satellite measurements have similar, and slightly lower, $R_{rs}(\lambda)$ in bluer wavelengths, indicating that the in situ dataset is oriented more toward optically complex coastal waters with substantial CDOM. This suggests that part of the explanation for the drop in DoF in satellite-derived climatologies comes from the fact that the in situ dataset sampled, as a whole, more optically complex waters. For future work, it would be valuable to perform a similar analysis to what we have done here for coastal waters, regional or shelf seas, or other more optically complex environments.

We find that both $R_{rs}(\lambda)$ and variables derived from $R_{rs}(\lambda)$ are highly inter-correlated, reducing the number of DoF associated with each, with a greater reduction in DoF in the derived products. This becomes a problem when products are derived using empirical relationships with $R_{rs}(\lambda)$, and especially when the same wavelengths are used for the products that are assumed to be independent of each other; for example, over much of the ocean PIC, POC, and chlorophyll all are functions only of $R_{rs}(\lambda)$ at two wavelengths, at (or near, depending on the sensor) 443 and 555 nm. Certain combinations of PIC, POC, and chlorophyll, which may occur in the surface ocean, are therefore impossible to find using these algorithms. This is distinct from algorithms, typically called "quasi-analytical" or "semi-empirical", that use known or assumed spectral shapes for absorption and scattering properties of optical constituents that can be related to the same derived products, such as PIC, POC, and chlorophyll (Werdell et al. 2013). These approaches may result in similar correlations and DoF between derived products, but do not inherently have the same problems as empirical approaches. We note that PACE will have, in addition to hyperspectral visible bands, UV bands from 350 nm as well as spectral polarized bands. These measurements are expected to both improve the atmospheric correction (hence reduce the $R_{rs}(\lambda)$ uncertainties) as well as provide their own ocean signals, both of which may increase the DoF compared to those found here. UV data in particular is potentially rich with information about phytoplankton physiology and community structure, and independent of variability at other wavelengths, though there will be challenges associated with simultaneously using it for atmospheric correction and extracting biogeochemical information, and not enough data exist at present for reliable statistical analysis. In addition, it has been shown that adding other environmental variables such as SST can add useful information to inversions of phytoplankton groups, for example, Chase et al. (2022) and thus another approach to increase DoF for inversions by adding relevant and independent information (e.g., mixed-layer depth and nutrients from BGC-Argo assimilating models). It may also be fruitful to include spatiotemporal information to specify, for example, where blooms of a particular plankton type are expected.

Conclusion

The results presented here highlight the high degree of codependence between remote sensing reflectances at different wavelengths and of the products derived from these reflectances. For users of products based on existing reflectances, this primarily means factoring in the relationships between products when using more than one simultaneously. For the algorithms that generate these products from existing reflectances, these results indicate a potential to improve the suite of available products to be more accurate and precise, and to account for the relationships between products and $R_{rs}(\lambda)$ wavebands. One way to do this, consistent with the findings above, would be to derive a single product such as chlorophyll as a function of all reflectance wavebands, derive an anomaly from chlorophyll-based expectations of a secondary product, then specify all other products explicitly as a function of these two, along the lines of Alvain et al. (2005).

These findings are most relevant for algorithms that will generate products from hyperspectral reflectances in the future. The small number of DoF in hyperspectral

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reflectances indicates that only a few quantities can be estimated independently, and that different wavelengths' reflectances as measured from space will be strongly correlated. Complex algorithms that utilize the full spectrum of reflectance will need to factor in these correlations in order to generate reliable products. Crucially, if more than a few products are generated from hyperspectral reflectances, as is likely the case, such algorithms will also need to output the covariance information encoding the uncertainty in each product and the relationships between them. The fact that hyperspectral reflectances can be predicted within nominal uncertainties by their multispectral equivalents suggests that hyperspectral resolution can play a role in improving ocean color products, but that it will be challenging to provide a substantially finer-grained picture of surface ocean ecosystems and biogeochemical cycles. In particular, these results present a fundamental challenge to (or at least ceiling on the ecological resolution of) algorithms that attempt to extract the abundance of different phytoplankton functional types from remote sensing reflectance. Here by relying on PCA we have focused on broad, first-order variations, but where such resolution may be most useful and generate novel insights is in investigating outliers and rare events, such as blooms or binning data over coherent features like eddies, where e.g., monospecific signatures may be resolved with spectral precision.

References

Alvain, S., C. Moulin, Y. Dandonneau, and F.-M. Breon. 2005. Remote sensing of phytoplankton groups in case 1 waters from global seawifs imagery. Deep-Sea Res. I: Oceanogr. Res. Pap. **52**: 1989–2004. doi:10.1016/j.dsr.2005.06.015

Behrenfeld, M. J., and P. G. Falkowski. 1997. Photosynthetic rates derived from satellite-based chlorophyll concentration. Limnol. Oceanogr. **42**: 1–20. doi:10.4319/lo.1997.42.1.0001

Bisson, K., D. A. Siegel, and T. DeVries. 2020. Diagnosing mechanisms of ocean carbon export in a satellite-based food web model. Front. Mar. Sci. **7**: 505. doi:10.3389/fmars.2020.00505

Bisson, K., E. Boss, P. J. Werdell, A. Ibrahim, R. Frouin, and M. Behrenfeld. 2021. Seasonal bias in global ocean color observations. Appl. Opt. **60**: 6978–6988. doi:10.1364/AO.426137

Brown, C. A., Y. Huot, P. J. Werdell, B. Gentili, and H. Claustre. 2008. The origin and global distribution of second order variability in satellite ocean color and its potential applications to algorithm development. Remote Sens. Environ. **112**: 4186–4203. doi:10.1016/j.rse.2008.06.008

Cael, B. 2021. Variability-based constraint on ocean primary production models. Limnol. Oceanogr.: Lett. **6**: 262–269. doi:10.1002/lol2.10196

Cael, B., K. Bisson, and M. J. Follows. 2017. How have recent temperature changes affected the efficiency of ocean biological carbon export? Limnol. Oceanogr.: Lett. **2**: 113–118. doi:10.1002/lol2.10042

Cael, B., K. Bisson, and C. L. Follett. 2018. Can rates of ocean primary production and biological carbon export be related through their probability distributions? Glob. Biogeochem. Cycles **32**: 954–970. doi:10.1029/2017GB005797

Cael, B., A. Chase, and E. Boss. 2020. Information content of absorption spectra and implications for ocean color inversion. Appl. Opt. **59**: 3971–3984. doi:10.1364/AO.389189

Cael, B., S. Dutkiewicz, and S. Henson. 2021. Abrupt shifts in 21st-century plankton communities. Sci. Adv. **7**: eabf8593. doi:10.1126/sciadv.abf8593

Campbell, J. W. 1995. The lognormal distribution as a model for bio-optical variability in the sea. J. Geophys. Res.: Oceans **100**: 13237–13254. doi:10.1029/95JC00458

Chase, A. P., E. Boss, I. Cetinić, and W. Slade. 2017. Estimation of phytoplankton accessory pigments from hyperspectral reflectance spectra: Toward a global algorithm. J. Geophys. Res.: Oceans **122**: 9725–9743. doi:10.1002/2017JC012859

Chase, A. P., E. S. Boss, N. Haëntjens, E. Culhane, C. Roesler, and L. Karp-Boss. 2022. Plankton imagery data in form satellite-based estimates of diatom carbon. Geophys. Res. Lett. **49**: e2022GL098076. doi:10.1029/2022GL098076

DeVries, T., and T. Weber. 2017. The export and fate of organic matter in the ocean: New constraints from combining satellite and oceanographic tracer observations. Glob. Biogeochem. Cycles **31**: 535–555. doi:10.1002/2016GB005551

Dutkiewicz, S., P. Cermeno, O. Jahn, M. J. Follows, A. E. Hickman, D. A. Taniguchi, and B. A. Ward. 2020. Dimensions of marine phytoplankton diversity. Biogeosciences **17**: 609–634. doi:10.5194/bg-17-609-2020

Evers-King, H., and others. 2017. Validation and intercomparison of ocean color algorithms for estimating particulate organic carbon in the oceans. Front. Mar. Sci. **4**: 251. doi:10.3389/fmars.2017.00251

Gordon, H. R., and M. Wang. 1994. Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with seawifs: A preliminary algorithm. Appl. Opt. **33**: 443–452. doi:10.1364/AO.33.000443

Graff, J. R., T. K. Westberry, A. J. Milligan, M. B. Brown, G. Dall'Olmo, V. van Dongen-Vogels, K. M. Reifel, and M. J. Behrenfeld. 2015. Analytical phytoplankton carbon measurements spanning diverse ecosystems. Deep-Sea Res. I: Oceanogr. Res. Pap. **102**: 16–25. doi:10.1016/j.dsr.2015.04.006

Hovis, W. A., and others. 1980. Nimbus-7 coastal zone color scanner: System description and initial imagery. Science **210**: 60–63. doi:10.1126/science.210.4465.60

Hu, C., Z. Lee, and B. Franz. 2012. Chlorophyll aalgorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. J. Geophys. Res.: Oceans **117**: C01011. doi:10.1029/2011JC007395

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Huot, Y., and D. Antoine. 2016. Remote sensing reflectance anomalies in the ocean. Remote Sens. Environ. **184**: 101–111. doi:10.1016/j.rse.2016.06.002

Jackson, D. A. 1993. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. Ecology **74**: 2204–2214. doi:10.2307/1939574

Kostadinov, T., D. Siegel, and S. Maritorena. 2009. Retrieval of the particle size distribution from satellite ocean color observations. J. Geophys. Res.: Oceans **114**:C09015. doi:10.1029/2009JC005303

Kramer, S. J., D. A. Siegel, S. Maritorena, and D. Catlett. 2022. Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. Remote Sens. Environ. **270**: 112879. doi:10.1016/j.rse.2021.112879

Lee, Z., A. Weidemann, J. Kindle, R. Arnone, K. L. Carder, and C. Davis. 2007. Euphotic zone depth: Its derivation and implication to ocean-color remote sensing. J. Geophys. Res.: Oceans **112**:C03009. doi:10.1029/2006JC003802

Mitchell, C., C. Hu, B. Bowler, D. Drapeau, and W. Balch. 2017. Estimating particulate inorganic carbon concentrations of the global ocean from ocean color measurements using a reflectance difference approach. J. Geophys. Res.: Oceans **122**: 8707–8720. doi:10.1002/2017JC013146

Morel, A., and L. Prieur. 1977. Analysis of variations in ocean color. Limnol. Oceanogr. **22**: 709–722. doi:10.4319/lo.1977.22.4.0709

Nelson, N. B., and D. A. Siegel. 2013. The global distribution and dynamics of chromophoric dissolved organic matter. Ann. Rev. Mar. Sci. **5**: 447–476. doi:10.1146/annurev-marine-120710-100751

Nowicki, M., T. DeVries, and D. A. Siegel. 2022. Quantifying the carbon export and sequestration pathways of the ocean's biological carbon pump. Glob. Biogeochem. Cycles **36**: e2021GB007083. doi:10.1029/2021GB007083

O'Reilly, J. E., S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain. 1998. Ocean color chlorophyll algorithms for seawifs. J. Geophys. Res.: Oceans **103**: 24937–24953. doi:10.1029/98JC02160

Scholz, M., M. Fraunholz, and J. Selbig. 2008. Nonlinear principal component analysis: Neural network models and applications, p. 44–67. *In* A. N. Gorban, B. Kégl, D. C. Wunsch, and A. Y. Zinovyev [eds.], *Principal manifolds for data visualization and dimension reduction*. Springer. doi:10.1007/978-3-540-73750-6_2

Scott, J. P., and P. J. Werdell. 2019. Comparing level-2 and level-3 satellite ocean color retrieval validation methodologies. Opt. Express **27**: 30140–30157. doi:10.1364/OE.27.030140

Siegel, D., K. Buesseler, S. C. Doney, S. Sailley, M. J. Behrenfeld, and P. Boyd. 2014. Global assessment of ocean carbon export by combining satellite observations and food-web models. Glob. Biogeochem. Cycles **28**: 181–196. doi:10.1002/2013GB004743

Silsbe, G. M., M. J. Behrenfeld, K. H. Halsey, A. J. Milligan, and T. K. Westberry. 2016. The CAFE model: A net production model for global ocean phytoplankton. Glob. Biogeochem. Cycles **30**: 1756–1777. doi:10.1002/2016GB005521

Weinberger, K. Q., F. Sha, and L. K. Saul. 2004. Learning a kernel matrix for nonlinear dimensionality reduction. Proceedings of the 21st International Conference on Machine Learning.

Werdell, P. J., and others. 2013. Generalized ocean color inversion model for retrieving marine inherent optical properties. Appl. Opt. **52**: 2019–2037. doi:10.1364/AO.52.002019

Werdell, P. J., and others. 2019. The plankton, aerosol, cloud, ocean ecosystem mission: Status, science, advances. Bull. Am. Meteorol. Soc. **100**: 1775–1794. doi:10.1175/BAMS-D-18-0056.1

Westberry, T., M. Behrenfeld, D. Siegel, and E. Boss. 2008. Carbon-based primary productivity modeling with vertically resolved photoacclimation. Glob. Biogeochem. Cycles **22**:GB2024. doi:10.1029/2007GB003078

Wold, S., K. Esbensen, and P. Geladi. 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. **2**: 37–52. doi:10.1016/0169-7439(87)80084-9

Acknowledgments

It is a pleasure to thank the many scientists whose collective work has generated the data on which this work relies, and the editors and reviewers for constructive critique. Cael acknowledges support from the National Environmental Research Council through Enhancing Climate Observations, Models, and Data, and the European Union under H2020 grant agreement No. 820989 (project COMFORT) and Horizon Europe grant agreement 10109915 (project BIOcean5D). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union, nor the granting authority, nor the European Commission nor their executive agency can be held responsible for any use that may be made of the information the work contains. Bisson acknowledges support from NASA grant 80NSSC18K0957. Boss acknowledges support form NASA grant 80NSSC20M0203. As stated in the acknowledgments, I (B. B. Cael) also need to acknowledge support from Horizon Europe (grant number 10109915).The authors have no competing interests to declare. This is PMEL contribution number 5445.

Submitted 15 December 2022
Revised 20 February 2023
Accepted 27 February 2023

23782242, 0, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lol2.10319 by Azadeh brown - University Of Southampton , Wiley Online Library on [15/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License