applied optics

Information content of absorption spectra and implications for ocean color inversion

B. B. CAEL,[1,*] ALISON CHASE,[2] AND EMMANUEL BOSS[2]

[1]National Oceanography Centre, European Way, Southampton SO14 3ZH, UK
[2]School of Marine Sciences, University of Maine, 5706 Aubert Hall, Orono, Maine 04473, USA
*Corresponding author: cael@noc.ac.uk

Received 28 January 2020; revised 24 March 2020; accepted 24 March 2020; posted 25 March 2020 (Doc. ID 389189); published 27 April 2020

The increasing use of hyperspectral optical data in oceanography, both *in situ* and via remote sensing, holds the potential to significantly advance characterization of marine ecology and biogeochemistry because, in principle, hyperspectral data can provide much more detailed inferences of ecosystem properties via inversion. Effective inferences, however, require careful consideration of the close similarity of different signals of interest, and how these interplay with measurement error and uncertainty to reduce the degrees of freedom (DoF) of hyperspectral measurements. Here we discuss complementary approaches to quantify the DoF in hyperspectral measurements in the case of *in situ* particulate absorption measurements, though these approaches can also be used on other such data, e.g., ocean color remote sensing. Analyses suggest intermediate (∼5) DoF for our dataset of global hyperspectral particulate absorption spectra from the *Tara Oceans* expedition, meaning that these data can yield coarse community structure information. Empirically, chlorophyll is an effective first-order predictor of absorption spectra, meaning that error characteristics and the mathematics of inversion need to be carefully considered for hyperspectral data to provide information beyond that which chlorophyll provides. We also discuss other useful analytical tools that can be applied to this problem and place our results in the context of hyperspectral remote sensing.

https://doi.org/10.1364/AO.389189

1. INTRODUCTION

In many instances, in both science and life, light provides a wealth of information about our environment. The study of ocean ecology and associated elemental cycles is no exception; optical instruments produce detailed information about many different aspects of ocean ecosystems, and remote sensing has been invaluable in the synoptic study of marine ecology and biogeochemistry for decades [1–4]. Optical information is most useful not for its own sake, but rather because of what it reveals about the material that the light interacts with. *Inversion*—the process of inferring causes from effects—is therefore crucial to optical oceanography; here the effects are the properties of observable light, and the causes are the biological and chemical components of the water that determine these properties. Frequently, we are interested in microphysical properties of particles: their bulk size distribution, shape, and composition, as well as associated pigments, such as chlorophyll a, that absorb light. This fundamental importance of inverse problems is of course not at all restricted to optical oceanography, and extends to arguably all of science. However, inversion problems are challenging, because causes often produce only slightly differing effects, which can greatly amplify always-present error, noise, or variability. Additionally, ambiguity in the result of a given inversion arises from the fact that different combinations of materials can produce a similar optical signature [5]. Great care must therefore be taken to construct an inversion that provides a satisfactory and meaningful answer, though this can be done in many ways.

To invert in the inevitable presence of error is notoriously difficult, and a rigorous approach needs to be taken. In optical oceanography, the most widely used inversions to date [6,7] have been relatively simple compared to those in other fields, using a handful of inputs to estimate one or a small number of variables (e.g., using several wavelengths of remote sensing reflectance (R_{rs}) to estimate chlorophyll a concentration; a notable exception is the estimation of the particle size distribution by laser diffraction [8]).

Current ocean color remote sensing products, such as particulate organic carbon and diffuse attenuation at 490 nm, both of which are concentration dependent, strongly covary with chlorophyll a. Even mean particle size, a property that is not concentration dependent, has been found to covary with chlorophyll [9]. On the other hand, different phytoplankton taxonomic groups (e.g., diatoms, dinoflagellates, prymnesiophytes, cyanobacteria) have different assemblages of accessory pigments in addition to chlorophyll a, which in turn results in differing spectral signatures. Therefore, in principle, different phytoplankton groups could be distinguished with hyperspectral data (i.e., data with spectral resolution of 10 nm or less), provided that information on the differential absorption and scattering properties of phytoplankton groups is extractable from spectral measurements.

The upcoming NASA Plankton, Aerosol, Cloud, and Ocean Ecosystems (PACE) satellite will be the first to host a space-born hyperspectral radiometer designed for global open ocean applications when it launches in 2022 [10,11].The PACE mission is eagerly anticipated by the ocean optics community because the hyperspectral information provided by the satellite's Ocean Color Instrument (OCI) is hypothesized to contain more information on surface ocean ecology and biogeochemistry than is presently available from multispectral satellite instruments. The possibility of obtaining information on community structure from space is extremely appealing, as it permits a much more detailed understanding of communities' ecological dynamics and biogeochemical function and has been investigated previously [12–18]. Despite the appeal, inverting for community structure is a fundamentally challenging problem because the spectral signatures of the different phytoplankton groups of interest may not always be spectrally distinct and thus the problem may be ill-posed.

The number of degrees of freedom (DoF) in a hyperspectral measurement is determined by the error characteristics of the measurement and the similarity of the spectral shapes being inverted for [19]. For instance, in a hypothetical limit case where all of the wavebands were perfectly correlated among themselves such that there was no variation in spectral shape in the whole ocean (but only of intensity), the DoF of any measurement would be one, regardless of the spectral resolution. Using such data to infer more than one variable would then be fraught, and would not be able to provide independent estimates of the quantities being inverted for. The same is true if there is variation in spectral shape but not enough relative to the error of the measurement to be significant. Such errors are especially important when inverting for quantities with relatively similar spectral signatures, as the covariance of these spectral signatures can significantly amplify errors.

Past approaches to address this issue include a study to determine the minimum number of wavelengths needed to capture the information in R_{rs}. Lee *et al.* [20] determined empirically that 15 bands should suffice to capture the variability in R_{rs}. Vandermeulen *et al.* [21], on the other hand, based on derivative analysis determined that to optimally resolve spectral variability hyperspectral absorption and reflectance data with 5–7 nm resolutions are optimal. While relevant, the information content (with respect to the concentration of various substances within the water) was not addressed directly in the above studies. The latter is the focus of our paper.

Here we investigate the DoF of hyperspectral ocean color signals with an eye towards the construction of inversions that are well-posed and meaningful, i.e., that do not attempt to invert for more quantities than there are DoF in the signal and therefore can provide independent estimates of each of the

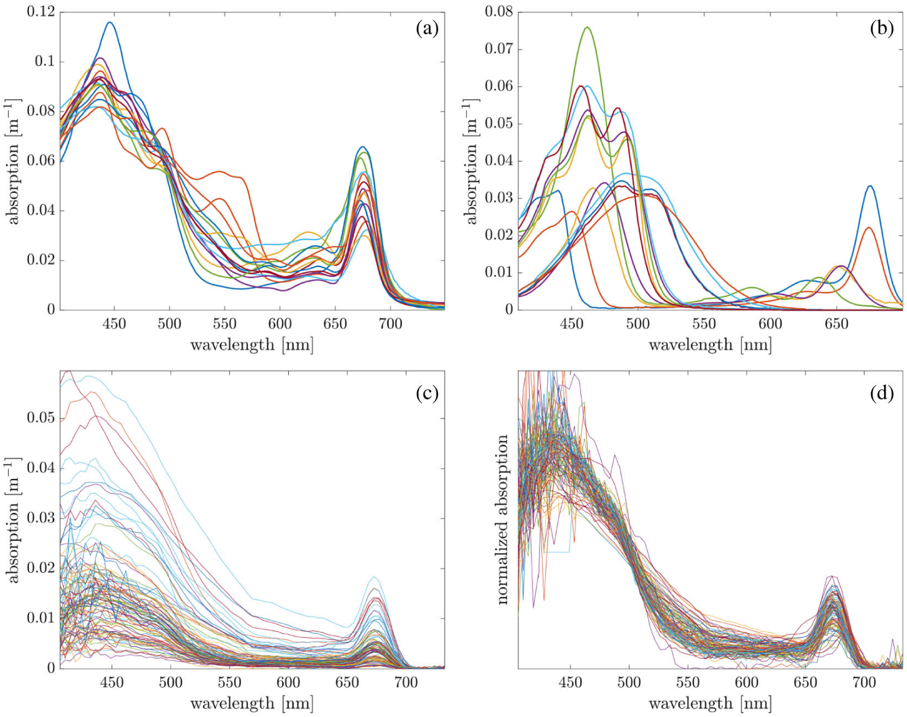

Fig. 1. (a) Example absorption spectra for phytoplankton from different groups; data from Ref. [27]. (b) Example absorption spectra for different pigments; data from Ref. [9]. (c) 100 randomly chosen absorption spectra from the data described in Section 3. (d) Same as Fig. 1(c) but with spectra normalized such that they integrate to one.

Table 1. Various Error Sources for Absorption Spectra and R_{rs} and Their Likely Characteristics[a]

Spectral Absorption Uncertainties (collected with an AC-S)	**Magnitude of Uncertainty**	**Spectrally Correlated?**
Calibration	0.01 m^{-1}	Depends on quality of calibration water
Detector sensitivity	1/SNR*path-length^{-1}	No
Scattering correction	10's of % in the blue, could have significant offset in the red in the presence of inorganic particles [28]	Yes
Binning	Quantified using standard deviation or percentiles; typically <0.004 m^{-1} decreasing from blue to red [24]	No
Spectral Remotely Sensed R_{rs} Uncertainties	**Magnitude of Uncertainty**	**Spectrally Correlated?**
Vicarious calibration	Max (5%, 0.001 Sr^{-1}) [29]	Maybe
Atmospheric correction	Several % [30]	Yes
Glint/whitecaps correction	<10% [31]	Yes
Detector sensitivity	1/SNR	No

[a] SNR denotes signal-to-noise ratio. Path-length designates the distance between source and receiver (25 cm for the AC-S whose data we use here).

quantities being inverted for. We employ two simple and complementary analyses—information content analysis (ICA) and principal component analysis (PCA)—to address this question as applied to hyperspectral particulate absorption ($a_p(\lambda)$). There is both a great deal of global $a_p(\lambda)$ data available, as well as a substantive body of work decomposing the $a_p(\lambda)$ spectra into that of non-algal particles (NAP; $a_{\mathrm{NAP}}(\lambda)$) [22] and absorption by phytoplankton pigments ($a_\phi(\lambda)$) and further from $a_\phi(\lambda)$ to different sizes of plankton or to different pigments [23–25]. The particulate absorption, $a_p(\lambda) = a_{\mathrm{NAP}}(\lambda) + a_\phi(\lambda)$, together with the absorption due to water ($a_w(\lambda)$) and of colored dissolved organic matter (CDOM, or gelbstoff, i.e., "yellow stuff," $a_g(\lambda)$) comprise the total absorption coefficient, $a(\lambda)$, a major determinant of R_{rs}, which is often approximated as a polynomial in $u = b_b(\lambda)/(a(\lambda) + b_b(\lambda))$, where $b_b(\lambda)$ is backscattering [26].

Our analysis is conducted with both an "output-based" approach (Section 2)—can we extract N pieces of information from a signal given the overlap in the underlying spectra of interest (i.e., spectral signatures of signal-causing substances), and the error associated in that signal—and an "input-based" approach (Section 3)—how many DoF are in the data themselves, or into what dimensional subspace do the data collapse when neglecting variation below measurement error? Several related issues are likely to contribute to the limited information available in hyperspectral data: 1) the spectral signatures of the desired constituents are largely similar [Figs. 1(a) and 1(b)]; 2) measurements include errors (Table 1); and 3) measured shapes of $a_p(\lambda)$ spectra tend to be quite similar [Figs. 1(c) and 1(d)]. In short, we are looking for small differences in noisy measurements to parse between covarying pieces of information. We present our approaches and findings in sections defined by analysis type; Section 2 addresses the ICA, Section 3 addresses the PCA, and Section 4 reports on other types of analyses and conclusions.

2. INFORMATION CONTENT ANALYSIS

The first approach we consider derives DoF from the inversion procedure and measurement error characteristics. ICA allows us to ask: given that error characteristics of our signal, and that the signals we are interested in inverting for (and that contribute to our measurement) are not independent, is our inversion well-posed? That is, are we trying to extract as much (or less) information from our measurement as is actually possible? Ultimately, it is the covariance of different signals of interest, in combination with the error characteristics, that determine the DoF (and thus the well-posedness) of the inversion. If $\mathcal{C}$ is the covariance matrix of the spectra being inverted for, i.e.,

$$\mathcal{C}_{jl} = \int_{\min(\lambda)}^{\max(\lambda)} \hat{k}_j(\lambda)\hat{k}_l(\lambda)\mathrm{d}\lambda, \tag{1}$$

where the k are the "kernels," i.e., spectra being inverted for (such as a given known absorption spectrum, in this case), $j = l = 1 \ldots m$ are indices of the m different spectra (kernels) being inverted for, the $\hat{}$ notation indicates normalization with a square norm, i.e., that the square of $\hat{k}$ integrates to one over $[\min(\lambda), \max(\lambda)]$, and λ corresponds to wavelength [nm], then it is the eigenvalues of $\mathcal{C}$—for which we will use the notation ν—that determine the DoF of the inversion. For an inversion of a measurement with relative error ε into the spectra $k_1(\lambda) \ldots k_m(\lambda)$ to have at least i DoF, the condition

$$\varepsilon \ll \sqrt{\nu_i} \tag{2}$$

must be met, where ν_i is the ith eigenvalue of $\mathcal{C}$ (see [19] for a more extensive description of the above). This is due to the fact that via inversion, errors are magnified by a scaling factor of ν^{-2}, so a very small eigenvalue can produce a very large error. This condition can be adapted to account for absolute error, spectrally varying error, and correlation of errors at different wavelengths by adjusting Eq. (1) accordingly [19].

The k's of interest might be pigments or phytoplankton groups when inverting $a_\phi(\lambda)$, but will also include NAP when inverting $a_p(\lambda)$, and when inverting the total absorption will also include $a_w(\lambda)$ and $a_g(\lambda)$. To illustrate how ICA can be used to identify DoF, here we consider all of these in turn. Figure 2(a) shows seven pigment absorption spectra from [32], and the square roots of the eigenvalues of the associated covariance matrix (access code available from Ref. [33]). As $\sqrt{\nu_7} = 0.11$,

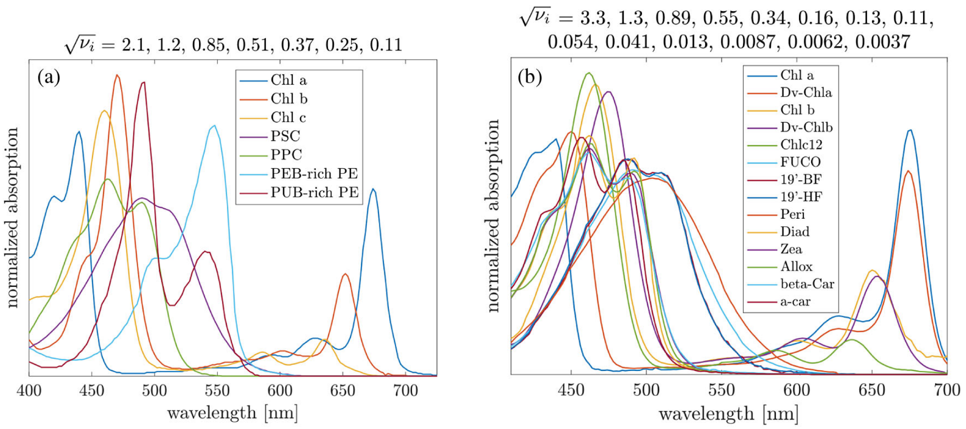

Fig. 2. (a) Square-normalized pigment absorption spectra from Ref. [32]. The list of values $\sqrt{\nu_i}$ is the square root of the eigenvalues of these normalized spectra's covariance matrix [see Eq. (2)], i.e., the ith value is the estimated error tolerance for i degrees of freedom in an inversion; in all subfigures for Figs. 2–5, the number of eigenvalues corresponds to the number of spectra plotted. (b) As Fig. 2(a) for the pigment absorption spectra from Ref. [9].

this indicates that for this inversion to be well-posed, a measurement error of $\varepsilon \ll 11\%$ is required. This result is perhaps somewhat relieving, as 11% error is not a particularly strict requirement for a spectrophotometric measurement, and is perhaps unsurprising, as pigments tend to absorb in distinct and narrow wavebands. Figure 2(b) shows, however, that there is a limit to this; inverting for more than 10 of the pigment spectra reported by Ref. [9] requires a measurement error of $\varepsilon \ll 1.3\%$, a far more stringent requirement. Note that ICA yields a discontinuous measure of DoF, so for example in Fig. 2(a), an ε of 24% is equivalent in terms of DoF to an ε of 12%, though the former would still obtain results with twice the signal-to-noise ratio.

Absorption by a particular phytoplankton species or functional group is the net result of absorption by various pigments and the modification of their absorption properties by the package effect [34]; as many phytoplankton types within broad taxonomic groups share pigments, their absorption spectra can be expected to show a much higher degree of covariance. This is readily seen in Fig. 3(a), which applies the same ICA analysis as above to absorption spectra used for different phytoplankton functional types (PFTs) in Ref. [35]. These are spectra used in a virtual simulation where measurement error is not a relevant concept but illustrates the additional challenge associated with inverting for phytoplankton groups rather than pigments and is based on measured representative absorption spectra. After the first several $\sqrt{\nu_i}$, values become very small; to invert for all eight PFTs here requires $\varepsilon \ll 1.5\%$, an order of magnitude smaller than for the comparable DoF for pigments. Nonetheless, a reduced subset can be identified that can yield a well-posed inversion for reasonable measurement error; including only the model *Synechococcus*, high-light *Prochlorococcus*, small eukaryote, and *Trichodesmium* (i.e., by excluding organisms that are not expected to contribute significantly to overall biomass in the surface oligotrophic ocean), rather than all of the spectra from Fig. 3(a), results in $\sqrt{\nu_4} = 0.082$.

The issue is further compounded by the fact that absorption spectra vary between different organisms within functional groups and even for a single species depending on environmental conditions affecting its physiology (e.g., light and nutrients). This uncertainty can be included in Eqs. (1) and (2), similar to uncertainty in the measurement itself, but will vary with each individual inversion if different k_j have different uncertainties. As one is trying to determine the contributions c_j to a measured spectrum $a \pm \varepsilon = \sum_j c_j k_j$, if the spectra have associated uncertainties ϵ_j, the sum $\sum_j c_j \epsilon_j$ contributes to the inversion uncertainty ε. Thus, when the k_j have uncertainties, ε can be replaced by $\varepsilon + \sum_j c_j \epsilon_j$. Uncertainty in any $k(\lambda)$ will therefore always result in decreases in $\sqrt{\nu_i}$, thus reducing the DoF of any inversion for phytoplankton groups or species. Furthermore, converting from an absorption as yielded by inversion to a concentration requires a conversion coefficient, which itself can vary (for example due to photo-acclimation) and brings with it appreciable uncertainty.

It is worth emphasizing that this difficulty is more the product of measurement error than of spectral resolution. Figure 3(b) shows the same spectra as Fig. 3(a), but coarsened to a 25 nm resolution rather than 5 nm (n.b., this choice is both because it is informative in terms of the spectral contrast of PACE versus multispectral ocean color satellites, and because the numerical model in [35] uses this 25 nm resolution; there is undoubtedly value in comparing different hyperspectral resolutions, but this is a more nuanced question that deserves careful consideration of its own; also see [21]). Despite this five-fold decrease, the $\sqrt{\nu_i}$ values are virtually unchanged, given that these are not exact thresholds but rather estimations for the number of DoF given by the error amplitude where results are obtained with a signal-to-noise ratio of one [19]. This strongly suggests that as long as spectral bands are placed so as to resolve the major spectral features in question, spectral resolution is not the limiting factor in extracting information from absorption spectra—as long as the spectral resolution is fine enough to capture major features [21]. While a five-fold increase in spectral resolution does not appreciably affect the DoF between Figs. 3(a) and 3(b), a five-fold decrease in ε could increase the DoF by two (i.e., $\sqrt{\nu_i}/\sqrt{\nu_{i+2}} < 5$ for any $n \leq 6$). This is in large part because the variation in absorption spectra is largely controlled by lower-frequency variability than requires hyperspectral data to resolve.

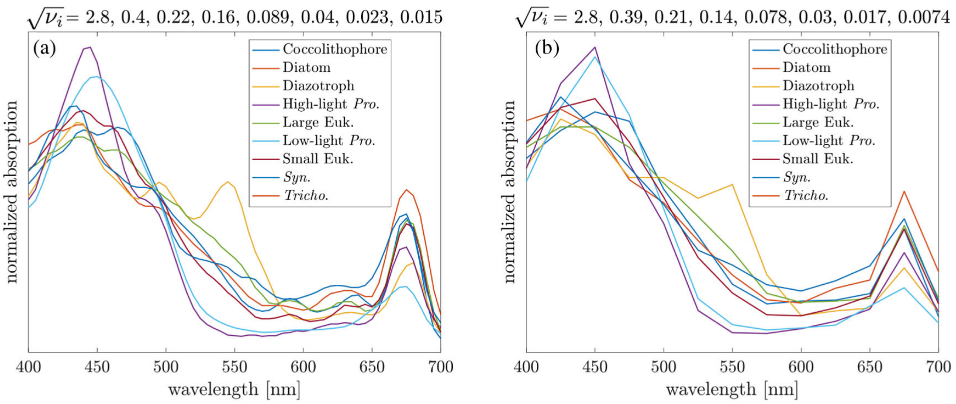

Fig. 3. (a) As Fig. 2(a) for the PFT absorption spectra from Ref. [34]. (b) Same as Fig. 3(a) at 25 nm resolution.

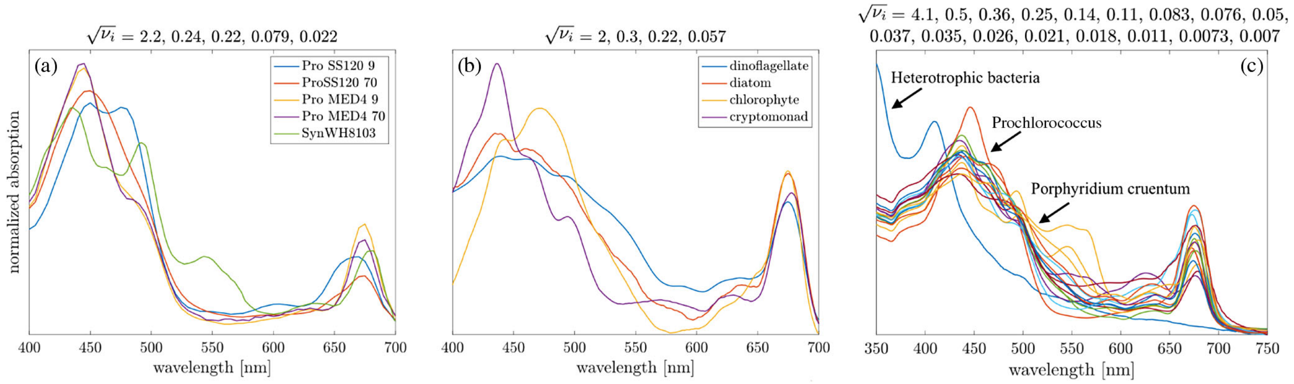

Fig. 4. (a) As Fig. 2(a) for the phytoplankton absorption spectra from [36]. (b) As Fig. 2(a) for the phytoplankton absorption spectra from Ref. [37] (c) As Fig. 2(a) for the phytoplankton (plus one heterotrophic bacterium) absorption spectra from Ref. [27]; these include heterotrophic bacteria, *Prochlorococcus*, *Synechococcus*, phycocyanin-rich picophytoplankton, *Pavlova pinguis*, *Thalassiosira pseudonana*, *Pavlova lutheri*, *Isochrysis galbana*, *Emiliania huxleyi*, *Porphyridium cruentum*, *Chroomonas fragariodes*, *Prymnesium parvum*, *Dunaliella bioculata*, *Dunaliella tertiolecta*, *Chaetoceros curviestum*, *Hymenomonas elongata*, and *Prorocentrum micans*.

The relative difficulty of identifying a well-posed inversion for phytoplankton groups as compared to pigments also holds for specific organisms. Figure 4(a) shows the results of the same ICA for two strains of *Prochlorococcus*, each grown at two different light levels, and a strain of *Synechococcus* that can be taken as a representative example of inverting for five different populations in an oligotrophic gyre (n.b., *Prochlorococcus* has more than this many ecotypes, with varying photophysiology [36]). Here even for five types, an error of $\varepsilon \ll 2.2\%$ is necessary. However, a reduced subset results in a well-posed problem; inverting for only three types (neglecting the *Prochlorococcus* strains grown in low light conditions) yields a well-posed inversion for $\varepsilon \ll 19\%$ (n.b., eliminating spectra in this fashion is not in general a sensible way to arrive at a good inversion; one should of course not expect satisfactory results if one's inversion does not account for the constituents that determine the shape of the spectra being inverted). Thus, a relatively coarse inversion attempting to estimate just a few species/types can be well-posed and provide meaningful information as long as the spectra being inverted for are defined appropriately. This is not only the case for oligotrophic regions; Fig. 4(b) shows a similar covariance structure for four co-occurring coastal species [37]. It is worth noting, however, that even in oligotrophic gyres, a suite of organisms is often present at high enough abundances to contribute appreciably to the particulate absorption, including, e.g., various diazotrophs, coccolithophores, and diatoms. Note that here we emphasize the finite information content of a spectra, not the correct choice of groups to invert to (an important topic all by itself). That is, inverting a spectrum from the Arctic, we may find it projects onto a *Prochlorococcus* strain even though none grows there.

Given that well-posedness for reasonable error values appears to require inversions including only a few spectra, it then is essential to incorporate additional information to reduce the complexity of any inversion to just a few representative spectra. That is, one cannot apply the same inversion across a gradient of different ecosystems and hope to derive accurate community structure information. One must instead restrict each inversion based on what is expected to dominate a particular measurement. Fig. 4(c), which shows the ICA performed on an assemblage of 17 different species that span a range of environmental niches, evinces the hopelessness of such an approach.

In many cases, one is interested in inverting a signal that contains more than just $a_\phi(\lambda)$, and potentially contributions from $a_{\text{NAP}}(\lambda)$, $a_g(\lambda)$, and $a_w(\lambda)$. Even though these latter three have spectral signatures very different from $a_\phi(\lambda)$, they still affect the covariance structure captured by $\mathcal{C}$ and therefore will affect the DoF. This is especially the case if $a_{\text{NAP}}(\lambda)$ and $a_g(\lambda)$

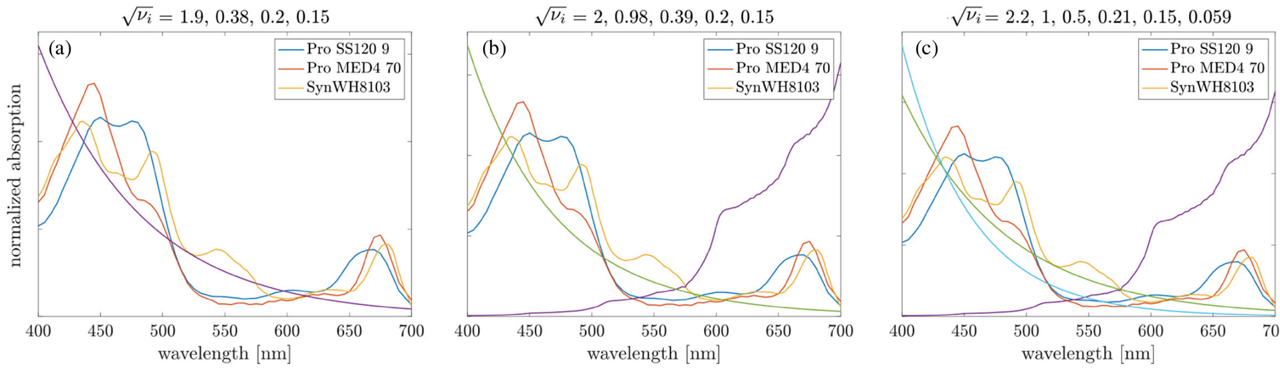

Fig. 5. (a) As in Fig. 4(a) but for only two *Prochlorococcus* strains, one *Synechococcus* strain, and a typical a_{NAP} absorption spectrum (purple line) [39], to represent overall particulate absorption. (b) As Fig. 5(a) but including the H_2O absorption data from Ref. [40] (purple line) and a stretched exponential a_{dg} as in Ref. [38] (green line), to represent overall *in situ* absorption without parsing between CDOM and NAP contributions. (c) As Fig. 5(b) but with separate exponential absorption spectra for CDOM (blue line) and NAP (green line) [39].

are considered separately, as these have relatively similar spectral shapes, and so one might expect to expend a DoF distinguishing between the two. Figure 5(a) shows that, at least for this example, inverting $a_p(\lambda)$ versus $a_\phi(\lambda)$ adds little difficulty, as the characteristically exponential NAP spectral shape is sufficiently different from that of phytoplankton (or pigment, for that matter) absorption. However, attempting to parse between $a_{\mathrm{NAP}}(\lambda)$ and $a_g(\lambda)$ is rather different, as seen in Fig. 5(c). This inversion uses fixed spectral slopes CDOM and NAP and does not consider the variability in spectral slope for both CDOM and NAP, which will significantly lower $\sqrt{\nu_5}$ (see above). If, however, one uses a combined $k(\lambda)$ for $a_{\mathrm{dg}} = a_{\mathrm{NAP}} + a_g$, such as a stretched exponential function [38], one recovers the same $\sqrt{\nu_i}$ as for the a_p case [Fig. 5(b)]. That is, one DoF is indeed lost attempting to distinguish between a_{NAP} and a_g. This exercise is the same for the spectra in Fig. 4(b). However, it is worth noting that PACE will measure in the UV, which likely will improve its ability to distinguish between NAP and CDOM contributions to absorption.

In summary, ICA of these spectra demonstrates several important points regarding their inversion:

- In all cases investigated above, the number of DoF was less than the number of spectra being inverted for, given likely error magnitudes—though we were able to choose subsets of spectra where this was not the case. To increase the available DOFs, one would likely have to augment the spectra with additional and independent sources of information [e.g., sea surface temperature (SST)].
- Additional information must be leveraged to constrain inversions to a relatively small number of spectra. Inversions can then be tailored to specific regions or questions (see Section 4.B).
- The difference between hyperspectral and multispectral resolution may be less important than measurement error in terms of impact on DoF (at least for the absorption spectra analyzed herein). This suggests that carefully binning of hyperspectral data to reduce uncertainties/minimize measurement error may be more useful for providing meaningful inversions than direct inversion of hyperspectral data (see Section 4.A), and is likely due to the large bandwidth of features (pigment absorption bands) in natural spectra.
- Knowledge of error characteristics is essential to ensure well-posed inversions, as correlations, spectral variation, and type of error all affect the DoF thresholds. When inverting for phytoplankton groups or species, this also critically includes uncertainty or variability in the spectral shape of what is being inverted for.
- Inversions for pigments tend to have higher DoF, not only because pigment spectra are far less variable/uncertain than those for species or groups, but also because they covary less.
- When inverting a signal that includes contributions by both $a_{\mathrm{NAP}}(\lambda)$ and $a_g(\lambda)$, because their spectral shapes are relatively similar and also uncertain/variable, it is useful to use a single spectrum for $a_{\mathrm{dg}}(\lambda)$ so as not to expend a DoF parsing the relative contribution between these.

3. PRINCIPAL COMPONENT ANALYSIS AND PIGMENT DECOMPOSITION

The above analysis derives DoF from the inversion procedure and measurement error characteristics, i.e., does not consider characteristics of the data beyond what their error needs to be to resolve more information. The other limit case we consider is the opposite approach, i.e., how to derive DoF from only data with no explicit knowledge of their error characteristics. PCA provides a readily available means to do so. PCA is used widely across a range of scientific disciplines, and in this case identifies orthogonal spectral shapes, or modes, that account for the highest fractions of the variance in the data. PCA is described exhaustively elsewhere, e.g., [41], but in short works by sequentially identifying the vectors along which the data have the most variance, and can be thought of as fitting an N-dimensional ellipsoid to the data. The first mode of the PCA $\vec{p}_1$ is then defined as

$$\vec{p}_1 = \arg \max_{\|\vec{p}\|=1} \|\boldsymbol{X}\vec{p}\|^2, \tag{3}$$

where $\boldsymbol{X}$ is the matrix of data, and the remaining $\vec{p}$ are defined iteratively by subtracting the previous modes from $\boldsymbol{X}$. As the spectral shapes that are identified as accounting for the most

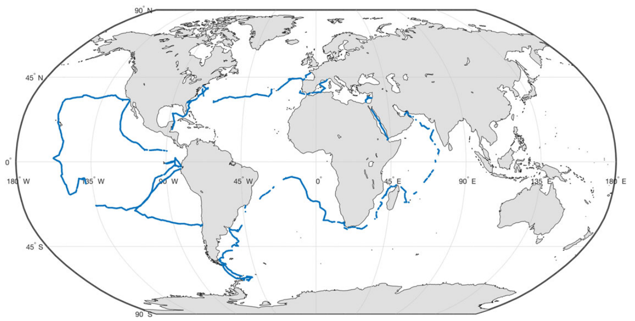

Fig. 6. Map of locations at which absorption spectra and temperature and salinity data were collected during the *Tara Oceans* expedition, 2009–2012.

variance are wholly empirical, they require subsequent interpretation. In the case of $a_p(\lambda)$, one approach for this may be by pigment decomposition, as shown below.

To illustrate how PCA can be used to identify DoF, we analyze the *Tara Oceans* hyperspectral particulate absorption data obtained by subtracting spectra of adjacent measurements of filtered water from those of non-filtered water (Slade et al., 2010). This dataset comprises 303,022 1-min-binned spectra, that have been acquired using a WETLabs AC-S deployed underway over very diverse environments [42, Fig. 6]. The spectra in this dataset have been "unsmoothed" to account for filter factors applied automatically by the AC-S instrument [24] and are available in NASA's SeaBASS repository [43]. The spectra were collected from 2009–2012 across the globe and by multiple personnel and several different instruments, thus removing potential specific user and instrument biases. Associated uncertainties are described in Table 1. We analyze these spectra three ways: (1) "raw," i.e., considering both their amplitude and shape, (2) normalized, i.e., considering only their spectral shape, and (3) subtracting the spectral shape associated with chlorophyll alone and thereby considering only the residual difference between the measurement and the chlorophyll-based prediction. In all cases, we perform a weighted PCA, where individual wavelengths are weighted inversely to the variance in absorption of that wavelength, such that all wavelengths contribute equally to determining the resulting spectra [n.b., in all cases, unweighted PCA yielded steeper dropoffs in fraction of variance (FVA) accounted for with mode number]. Note that standard PCA assumes a uniform uncertainty across all measurements, but that factoring different wavelengths' uncertainties, or those of individual samples, or even that of individual sample–wavelength pairs (i.e., individual measurements), is possible, though how best to do this is an active area of research [41,–48]. Also note that when applying PCA to normalized spectra as we have done here, one must calculate weights for the PCA *after* normalizing, and when measurement uncertainties are incorporated into said PCA, the uncertainty will also have to be rescaled by the normalizing factor.

Figure 7(a) shows the spectral shapes and the associated FVAs accounted for, resulting from a weighted PCA of the total dataset. The first mode explains almost all (>94%) of the variance in the data, consistent with Fig. 1, showing that spectral shapes tend to be quite similar and therefore that variation in these data is driven largely by amplitude. The next several modes appear mostly to be combinations of a NAP-absorption-like spectrum and a modulation of the Chl-peak in the absorption spectrum, indicating that the remaining variability is likely due mostly to changes in the ratio of Chl and NAP concentrations or slight variations in packaging, NAP spectral slope, and accessory pigments.

DoF can be assessed from PCA in various ways, but arguably the best method in terms of balancing simplicity of calculation with accurate evaluation of dimensionality is the "broken stick" method [49], which compares modes' FVAs with a random division of variance into N parts. In other words, a dataset has d DoF if the dth mode of the PCA explains more variance than would be expected if the variance was uniformly distributed, given by $(1/N)\sum_{i=d}^{N}(1/i)$ (n.b., other methods such as the Kaiser–Guttman criterion yielded ± 1 DoF for the data we considered). This method indicates only one DoF in the raw spectra, as the second mode accounts for only $\sim 3\%$ of the variance (the cutoff is 5.10%).

In Fig. 7, we observe that as the amplitudes of absorption in each spectral band are highly correlated, it seems that more DoF may be realized by considering the spectral shape, i.e., using normalized spectra. Figure 7(b) shows the same as Fig. 7(a) but for a PCA applied to spectra normalized so that their average absorption across all wavelengths is $1\ \mathrm{m}^{-1}$. As expected, the variance is spread out across the modes more evenly, with the first

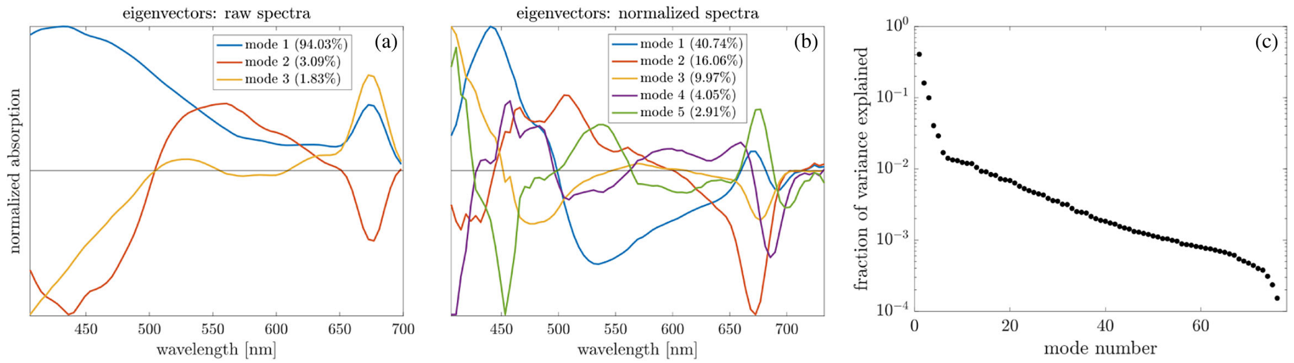

Fig. 7. (a) Spectra resulting from weighted PCA of raw spectra, with associated FVA given in legend. (b) Same for normalized spectra. (c) FVA versus mode number for all spectra from Fig. 7(b); note the different dropoffs in FVA with mode number between modes 1–5 and those >5. Solid black line indicates the cutoff point for the broken stick method.

mode accounting for nearly half the variance of that of the unnormalized case. This suggests that there is more information to be gained when considering the spectral shape and amplitude separately. This is not entirely surprising, given the relatively large dynamic range of Chl concentrations found in surface ocean waters (e.g., concentrations spanning several orders of magnitude) versus the relative variation in pigmentation per carbon or cell (a factor of six or so). The broken stick method identifies four DoF for these data. Arguably there are at most five DoF by a more relaxed criterion; the change in gradient in Fig. 7(c) suggests five DoF in these data if one uses a scree-type method of comparing cumulative variance explained versus mode number [50]. Furthermore, modes ≥5 are noisy spectral shapes that appear much more random than informative or interpretable.

As one of the foci here is to determine how much information can be obtained from hyperspectral data relative to what can be determined from a chlorophyll-based prediction, it is informative to ask: how many DoF remain after one has made a prediction for spectral shape of particulate absorption based on chlorophyll alone? To this end we used an existing power-law-based parameterization of the spectral shape as a function of chlorophyll to predict $a_p(\lambda)$ [51], subtract this prediction, and determine the DoF in the remaining residual. This is a means of addressing how well the chlorophyll concentration reflects the spectral shape of the data and how much information remains. Chlorophyll concentration [mg m^{-3}] is estimated from a line height algorithm:

$$\mathrm{Chl}_{\mathrm{ALH}} = \frac{1}{a^*_{\mathrm{LH}}}\left(a(676) - \left(\frac{a(715)-a(650)}{715-650} \times (676-650) + a(650)\right)\right), \tag{4}$$

where 650, 676, and 715 all have units of nanometers, and a^*_{LH} is the chlorophyll-specific absorption line height. From this, $a_p(\lambda)$ is estimated according to

$$a_p(\lambda) = A(\lambda)\mathrm{Chl}_{\mathrm{ALH}}^{B(\lambda)}, \tag{5}$$

where (A, B) are functions taken from Ref. [51], updated from Ref. [52]. Note that the nonlinearity of this equation may be the reason we get fewer DoF in the difference and with shapes that are easier to interpret. First, the amplitudes of the residuals indicate that this chlorophyll-based model predicts 74.3% of the variance of the data. Second, the first mode of the PCA, which accounts for most (>91%) of the remaining variance, has a shape similar to a typical NAP absorption spectrum [Figs. 8 and 9(a)]. Together these suggest, as above, that chlorophyll is a very strong predictor of overall pigment composition, and that most of the deviation from a chlorophyll-based estimation of $a_p(\lambda)$ is due to NAP absorption. The broken stick method indicates two DoF in the residuals, as the third mode accounts for 1.90% of the residuals' variance (as compared to a cutoff of 4.45%); this is consistent with the DoF in Fig. 7(c), as the functions for (A, B) above use two DoF and the residuals retain two DoF.

A tradeoff associated with PCA is that it identifies spectral shapes that account for the most variance in the data, but that these are wholly empirical and require additional information to be interpreted. An analysis of these modes that provides biogeochemical information is pigment decomposition, whereby a given $a_p(\lambda)$ spectrum is inverted to determine a set of functions that represents absorption by individual pigments or pigment groups, and NAP. One explanation of this is that phytoplankton accessory pigments covary less than phytoplankton species or representative spectra for phytoplankton taxonomic groups, meaning they can be inverted for with higher signal-to-noise ratios/more DoF. We apply the pigment decomposition algorithm from [24] to the modes identified by PCA above, with the modification that pigment concentrations are allowed to be negative, because a particular mode may represent relative deficiency of a given pigment. This analysis replaces the speculations above with quantification. For instance, the first mode of the raw spectra's PCA [Fig. 9(a)] is dominated by the signal of Chl a absorption, with relatively equal contributions of all accessory pigments and NAP in addition. Mode 2 shows positive accessory pigments (i.e., all pigments excluding Chl a) [Fig. 9(b)]. Mode 3 [Fig. 9(c)] again shows a strong influence of Chl a, possibly indicating the influence of pigment packaging, which results in different ratios of pigment absorption in the blue and red wavelengths. Five modes of the normalized spectra's PCA [Figs. 9(d)–9(h)] are represented by variable positive and negative accessory pigment absorption, and NAP in the case of modes 3–5. Figure 9(i) shows that the first mode from the residuals' PCA (Fig. 8) is best modeled as a strong contribution from NAP as well as positive pigments, and similar for the third mode [Fig. 9(k)] except that the contribution of pigments is negative. The second mode of the residuals' PCA is most strongly influenced by carotenoid and biliprotein pigments [Fig. 9(j)]. In other words, most of the variability in $a_p(\lambda)$ that is *not* explained by chlorophyll concentration is explained by the relative contribution of NAP versus pigments to particulate absorption.

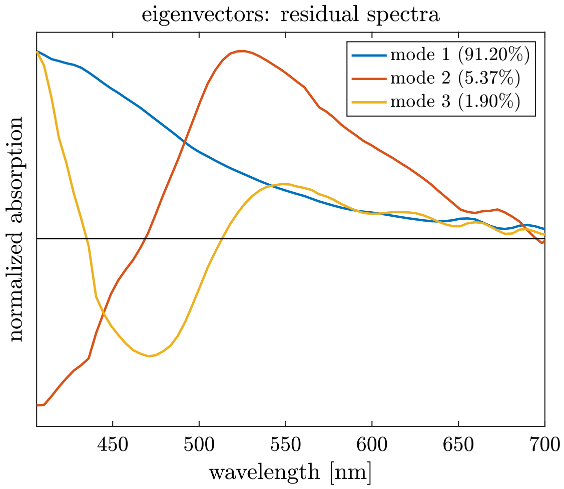

Fig. 8. Same as Fig. 7(a) but for residuals from Chl-based approximation.

Altogether these analyses demonstrate several important points:

- Hyperspectral particulate absorption spectra, despite having >80 independent spectral measurements per sample in our

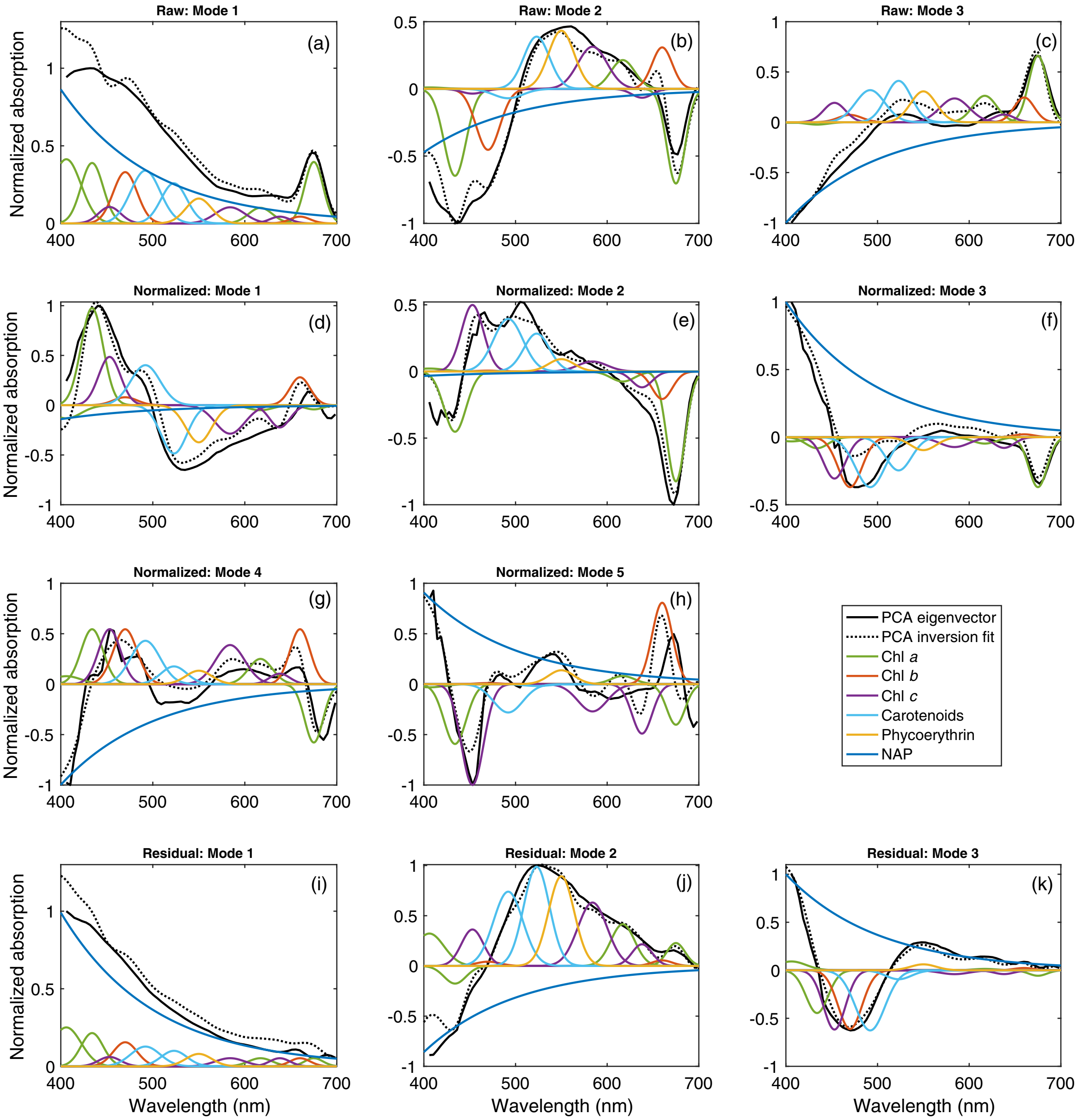

Fig. 9. Pigment decomposition as in Ref. [24] for eigenvector PCA spectra for (a)–(c) raw absorption spectra modes 1–3 shown in Fig. 7(a); (d)–(h) normalized absorption spectra modes 1–5 shown in Fig. 7(b); and (i)–(k) residual spectra modes 1–3 shown in Fig. 8. All panels show eigenvector PCA spectra in unitless normalized absorption (solid black line), the sum of the component pigment and non-algal particle (NAP) functions from the inversion of eigenvectors ("PCA inversion fit," dotted black line), and the component pigment and NAP absorption spectra (see legend for assigned colors). Carotenoids include photosynthetic and photoprotective pigments.

case, have only four to five DoF, i.e., are well described as a combination of just a few spectral shapes.

- Spectral amplitude governs most of the variability, because absorption at different wavelengths is tightly correlated.
- Chlorophyll *a* is an extremely good predictor of spectral shape, with most of the remaining variability determined by the relative contributions of NAP and accessory pigments to absorption.
- Ultimately, extracting information from hyperspectral absorption data beyond what can be inferred from the concentrations of chlorophyll and NAP requires very low measurement error and/or additional information such as UV bands. For R_{rs}, polarimetry measurements from PACE will likely yield additional useful information (see Section 4.E).

4. OTHER ANALYSES, CONSIDERATIONS, AND CONCLUSION

A. Derivative Analysis

Another commonly used technique, and one that is often cited as motivation for acquiring hyperspectral data, is derivative analysis, i.e., taking spectral derivatives of signals and analyzing/comparing these derivatives rather than the original signals (e.g., [53,54]). This can be a powerful visual tool, as it sharpens features in otherwise smoothly varying data. However, derivatives significantly amplify any measurement error that is not spectrally slow-varying—this is evidently an issue given the sensitivity to measurement error we have seen above. In practice, smoothing filters are typically applied before taking derivatives, which removes any high-frequency variability due likely to uncorrelated noise. As a derivative is, in essence, a high-pass filter and therefore serves to remove low-frequency variation, this procedure applies first a low-pass and then a high-pass filter to the original signal. How exactly this manipulation affects the signal depends on which of many possible smoothing filters is applied. It cannot provide *more* information than the original signal unless the smoothing filter is chosen to incorporate pre-existing knowledge, e.g., that a variation in phytoplankton group spectra is confined to spectral frequencies $>\delta$ nm, so that any variation in a measurement at frequencies $\leq\delta$ nm must be due to measurement errors and can be discarded. Smoothing filters are not typically chosen from this type of reasoning, however, and furthermore, there is no such simple cutoff, and errors are not confined only to specific spectral frequencies. Derivatives may be useful in maximizing information by contrasting frequency characteristics of both kernels and error, but this must be done carefully. The original signals in hyperspectral data are still useful without taking derivatives, just by providing more accurate information about the wavelength of an incoming photon. Hyperspectral measurements thus can be used to parse between signals whose absorption peaks are too similar to be distinguished by multispectral measurements, allowing for wider flexibility and range of potential inversions. Because there are necessarily fewer photons (i.e., signal) captured by narrower wavebands, individual hyperspectral wavebands will consequently have lower signal; the total signal may contain more information because of the enhanced spectral resolution. However, many inversion techniques will also require calibration using multi-spectral bands, and the extension to hyperspectral resolution will require care. Finally, we also note that hyperspectral kernels may also be useful in guiding the selection of appropriate smoothing filters by providing the aforementioned pre-existing information about absorption spectra's variation–wavelength relationships.

B. Hybrid Methods and Ancillary Data

The analyses in Sections 2 and 3 are entirely output and input based (from the perspective of the inversion), respectively, and in that sense are complementary; in general, their agreement gives us confidence that the upper limit to DoF for surface ocean absorption data is likely not significantly different from ∼4. We note that ICA and PCA are not the only available methods to address the question at hand; these are particularly intuitive and widely known, which also (importantly here) allow for the use of arbitrary/empirical spectra unlike some other approaches, and are therefore suitable for the general question at hand herein. More sophisticated methods have comparative advantages [55–57] that we would therefore recommend for investigating more specific or targeted inversion questions.

In practice, a combination of both measurement and pre-existing information, as well as the incorporation of other environmental data, can be useful for extracting the most information possible. In some cases, with higher signal-to-noise ratios such as in a coastal time series, there may be more DoF in the data. One example is discussed in Section 2—one can limit an inversion for N spectra to one for $n < N$ spectra by excluding from the N all those spectra that, e.g., correspond to organisms belonging to a different thermal niche. Then, along a temperature gradient, the same N spectra can be inverted for, but at each time and place, the actual inversion is for a reduced subset as defined by ancillary temperature data. For the dataset we analyzed here, incorporating associated temperature and salinity data did not affect the inversion; the FVAs for the normalized spectra plus temperature and salinity covariates were (40.09, 15.67, 9.73, 4.08, 2.94)% for the first five modes, compared to (40.74, 16.06, 9.97, 4.05, 2.91)% in Fig. 7(b). We interpret this to be because temperature and salinity covary strongly with [Chl], meaning their inclusion provides little additional information. This may not be true of other covariates (e.g., light availability) or for other datasets, however.

One could instead try every combination of $n < N$ spectra whose covariance permits a well-posed inversion, and select among the different combinations that fit the data best or balance goodness-of-fit to the data with community structure expectations based on other information (e.g., diatoms might be more expected to contribute significantly to the absorption spectrum during a spring bloom; dinoflagellates might be expected to be more dominant in autumn in some locations; coccolithophores might be more expected in the presence of certain water column stratification characteristics that may be remotely or autonomously observable). Uncertainty in the spectra for different organisms or phytoplankton groups can be reduced by incorporating nutrient data and knowledge of how nutrient status affects organisms' absorption spectra. There are myriad ways to set up inversions and to refine these by leveraging additional information; these are just a few examples of how one might make the most of one's data in the context of the limitations outlined in this paper. The appropriate technique for a given case will depend on what data and pre-existing knowledge are available and what question is being asked.

Another method we have not discussed here but that is used widely in inversion problems is that of regularization. Regularizing an inverse problem means applying an additional constraint on the solution so as to select among various possible solutions that fit the data equally well, i.e., picking the solution that best balances goodness-of-fit to the measurement and some additional criterion. The most obvious example here would be, when inverting for phytoplankton groups at a particular time and place, penalizing solutions for the difference between their phytoplankton group distributions and some climatological average or expectation for that time and place. One then does not require that a solution be well-posed because one has an additional means for selecting between possible solutions. This approach will of course tend to give results that look more like expectations, though with the appropriate statistical framework, an assessment can be made as to what extent the results are driven by the measurement versus, e.g., a Bayesian prior [55]. As it is not at all clear what an expected solution should look like over much of the ocean for many questions of interest to marine ecologists and biogeochemists, while regularization methods are a promising suite of tools to deal with the fundamentally underdetermined nature of ocean color remote sensing, their appropriate application requires careful consideration as to what prior information one actually has (though this can always be said to be the case). This is even more true considering that in even some of the most well-studied and supposedly homogeneous parts of the ocean, large shifts in community structure often occur unexpectedly and on very short timescales. Even so, regularization is a ubiquitously useful and mature set of techniques within the atmospheric community [55], and is involved in the atmospheric correction step. Furthermore, any semi-analytical inherent optical property (IOP) satellite algorithm currently in use assumes spectral behavior by constituents, which is a form of regularization [58]. Here we have argued that even if such algorithms improved in their fidelity, the amount of independent information they could obtain will not increase by much.

C. Error Specifics

Throughout this paper, we have highlighted the crucial role of error characteristics, and their appropriate treatment, for meaningful inversions. Error is the difference between having four to five DoF rather than >60, and the difference between being able to meaningfully invert for four spectra versus 44. Error can take different forms—errors can be relative, absolute, or a mixture (Table 1); they can be spectrally flat or vary spectrally or be correlated between different wavelengths (due to both spectral bandpass and actual correlations between different wavelengths); they can be in the measurement itself or in the spectra being inverted for. All of these can be taken into account—see Chapters 6 and 7 of [19]—but affect DoF differently, meaning a characterization of error is required to ensure inversions are meaningful. Noise is less exciting to most than signal, and error characterization can be tedious, but when inverting spectra, error has special importance. Some errors such as random electronic noise can be reduced by averaging many measurements in time or space. Others, such a bias in calibration, cannot.

D. Extension to Remote Sensing Reflectance

Remote sensing reflectance (R_{rs}) is a function of absorption and elastic and inelastic scattering, and is computed from radiometric measurements made by satellites. Each step in the calculation of R_{rs}, in addition to the finite resolution of the sensor, brings with it uncertainty or error. We have discussed the DoF in the case of particulate absorption. The DoF for remote sensing reflectance R_{rs} measurements from satellites is likely to have a similar number of DoF as the particulate absorption DoF. While R_{rs} may contain additional information associated with scattering (e.g., spectral information associated with the underlying size of particles or inelastic scattering information related to chlorophyll fluorescence), there are additional sources of errors (those associated with the procedure of removing atmospheric signals in R_{rs}) that could potentially reduce the number of DoF. One could use top-of-the-atmosphere reflectances and perform the same exact analysis as here to compute the total DoF of both atmospheric and oceanic signatures as a start. In a best-case scenario, one DoF will be sacrificed to the atmosphere and another to CDOM + NAP + atmospheric residual, then leaving the noise of the radiometer to determine how many DoF remain. NASA's aim is 5% radiometric accuracy, which is not an easy fit [59]. Additional steps (that bring with them their own errors and uncertainties) are then necessary to get from R_{rs} to IOPs [58]. In any case, a thorough uncertainty analysis will need to be understood and propagated through the calculations (see Ref. [60]) for the latest understanding of those uncertainties). These considerations are paramount for the upcoming NASA PACE mission, as a key oceanographic objective of this mission is to invert the hyperspectral data that PACE will provide to estimate multiple pigments or phytoplankton groups.

E. Beyond Chlorophyll

The analyses in Section 3 demonstrate that chlorophyll alone is a powerful predictor of absorption characteristics in the surface ocean. Chlorophyll retrievals from satellite have been the cornerstone of ocean color remote sensing for decades, and numerous remote sensing products are correlated with chlorophyll. Much of the optimism around hyperspectral data is that it will allow the ocean color community to break this ubiquitous dependence on a single variable.

While all of the optical variation in the sea cannot be said to fall along a single axis, it does appear that much of the variation in the surface covaries with [Chl]. Thus, the interest in going "beyond chlorophyll" can be considered an interest in deviations from this axis. This in turn means that one must always ask whether any additional sophistication provides a better prediction than chlorophyll alone, and by how much if so. PACE's UV and polarimetric information may be particularly useful in this regard; [10] polarization will help better separate oceanic and atmospheric contributions to the total signal, and UV will help better separate CDOM, NAP, and phytoplankton contributions to the oceanic signal. That these deviations are by definition second order—though we note emphatically that this does not make them unimportant or uninteresting!—underscores the necessity of the highest possible quality measurements for their study.

F. Judicious Use of Degrees of Freedom

It is important to note that if ultimately there is a finite number of independent pieces of ecological/biogeochemical information that can be retrieved by hyperspectral data, this mandates that great care be taken in defining what is being inverted for. This will, in general, depend on the question being asked. One must allot one's N DoF in an inversion in a way that is consistent with the environment, e.g., if diatoms are a significant component of the optical signature, they must be included in some way, because they will affect the measurement, and therefore their influence will propagate into the inversion no matter how it is constructed. One must also allot these N DoF towards a coarsened picture of reality that allows one to address the questions in which one is interested, e.g., if one is not specifically interested in diatoms, it is probably a poor use of one's DoF to invert for diatoms separately from other large cells with similar absorption characteristics (e.g., all having similar ratios of blue to red absorption). Spectra will always involve some degree of amalgamation; for instance, CDOM is not an individual compound but rather a diverse array of chemical constituents. This balance is certainly an art and rather question dependent. For a global-ocean phytoplankton inversion, it may never be possible to invert for high-light separately from low-light *Prochlorococcus* ecotypes; for a regional oligotrophic gyre-specific inversion, this may be a desirable use of a DoF. In all cases, a validation exercise needs to take place to ensure that the inversion is consistent with the environment in question (e.g., an inversion in the Arctic for *Prochlorococcus* ecotypes may provide statistically significant results even though no such phytoplantkon are in the water).

G. Conclusion

Inversion of optical measurements is a powerful methodology in optical oceanography, but its application requires careful consideration of the underlying mathematics. To this end, we investigated the DoF in hyperspectral particulate absorption data originating from the open ocean and for inversions of these data. Complementary analyses indicate that such data and inversions thereof have four to five DoF for reasonable error amplitudes. This number is likely of the same order as the number of independent pieces of information that can be retrieved from hyperspectral satellites measuring in the visible that incorporate additional sources of error such as atmospheric correction but potentially additional information due to variability scattering. As we expect from the limited DoF in these measurements, designing an inversion also requires careful consideration as to how best to treat data to get the most out of these DoF. As errors can be either absolute (e.g., instrument sensitivity) or relative (e.g., contamination of signal by scattering), and can have different spectral characteristics (e.g., different correlation structures between errors of signal at different wavelengths), these characteristics can affect the ultimate DoF differently and must be taken into account. Error characteristics and a careful treatment thereof are essential for well-posed inversions, which can also be improved by incorporating pre-existing knowledge and/or additional data sources. Chlorophyll is a very effective predictor of visible optical properties in the ocean, and any attempt to improve on a chlorophyll-based prediction must be evaluated in terms of how much of an improvement from a chlorophyll-based prediction is provided.

Funding. National Environmental Research Council (NE/R015953/1); National Aeronautics and Space Administration (NNX09AU43G, NNX11AQ14G); Horizon 2020 Framework Programme (820989).

Acknowledgment. It is a pleasure to thank Collin Roesler, Anna Hickman, Michael Behrenfeld, and Stephanie Dutkiewicz for useful discussions and data access, and all those who have helped with the *Tara Oceans* expeditions project for their efforts. This paper greatly benefited from the suggestions of three anonymous reviewers. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 820989 (project COMFORT, our common future ocean in the Earth system—quantifying coupled cycles of carbon, oxygen, and nutrients for determining and achieving safe operating spaces with respect to tipping points) and from the National Environmental Research Council under the Climate Linked Atlantic Sector Science Programme (grant No. NE/R015953/1). The work reflects only the author's/authors' view; the European Commission and their executive agency are not responsible for any use that may be made of the information the work contains. Funding for the collection and processing of the AC-S data set was provided under grants NNX11AQ14G and NNX09AU43G, and funding for the Ocean Optics class where this project was conceived was provided under grant NNX14AP66G, all from the NASA Ocean Biology and Biochemistry program.

Disclosures. The authors declare that there are no conflicts of interest related to this paper.

REFERENCES

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski, "Primary production of the biosphere: integrating terrestrial and oceanic components," Science **281**, 237–240 (1998).
2. M. J. Behrenfeld, E. Boss, D. A. Siegel, and D. M. Shea, "Carbon-based ocean productivity and phytoplankton physiology from space," Global Biogeochem. Cycles **19**, GB1006 (2005).
3. J. Uitz, H. Claustre, B. Gentili, and D. Stramski, "Phytoplankton class-specific primary production in the world's oceans: seasonal and interannual variability from satellite observations," Global Biogeochem. Cycles **24**, 1–19 (2010).
4. V. S. Saba, M. A. M. Friedrichs, D. Antoine, R. A. Armstrong, I. Asanuma, M. J. Behrenfeld, A. M. Ciotti, M. Dowell, N. Hoepffner, K. J. W. Hyde, J. Ishizaka, T. Kameda, J. Marra, F. Mélin, A. Morel, J. O'Reilly, M. Scardi, W. O. Smith, Jr., T. J. Smyth, S. Tang, J. Uitz, K. Waters, and T. K. Westberry, "An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe," Biogeosciences **8**, 489–503 (2011).
5. K. Shifrin and G. Tonna, "Inverse problems related to light scattering in the atmosphere and ocean," in *Advances in Geophysics*, R. Dmowska and B. Saltzman, eds. (Elsevier, 1993), Vol. **34**, pp. 175–252.
6. J. O'Reilly, "Vol. 11: SeaWiFS postlaunch calibration and validation analyses, part 3," NASA Tech. Memo., NASA, TM-2000-206892 (2000).
7. C. Hu, Z. Lee, and B. Franz, "Chlorophyll-a algorithms for oligotrophic oceans: a novel approach based on three-band reflectance difference," J. Geophys. Res. **117**, C01011 (2012).
8. Y. C. Agrawal and H. C. Pottsmith, "Instruments for particle size and settling velocity observations in sediment transport," Mar. Geol. **168**, 89–114 (2000).
9. A. Bricaud, "Natural variability of phytoplanktonic absorption in oceanic waters: influence of the size structure of algal populations," J. Geophys. Res. **109**, C11010 (2004).
10. J. Chowdhary, P.-W. Zhai, E. Boss, H. Dierssen, R. Frouin, A. Ibrahim, Z. Lee, L. A. Remer, M. Twardowski, F. Xu, X. Zhang, M. Ottaviani, W. R. Espinosa, and D. Ramon, "Modeling atmosphere-ocean radiative transfer: a PACE mission perspective," Front. Earth Sci. **7** (2019).
11. P. J. Werdell, M. J. Behrenfeld, P. S. Bontempi, E. Boss, B. Cairns, G. T. Davis, B. A. Franz, U. B. Gliese, E. T. Gorman, O. Hasekamp, K. D. Knobelspiesse, A. Mannino, J. Vanderlei Martins, C. R. McClain, G. Meister, and L. Remer, "The plankton, aerosol, cloud, ocean ecosystem mission: status, science, advances," Bull. Am. Meteorol. Soc. **100**, 1775–1794 (2019).
12. S. Alvain, C. Moulin, Y. Dandonneau, and F. Bréon, "Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery," Deep Sea Res. **52**, 1989–2004 (2005).
13. S. Alvain, C. Moulin, Y. Dandonneau, and H. Loisel, "Seasonal distribution and succession of dominant phytoplankton groups in the

global ocean: a satellite view," Global Biogeochem. Cycles **22**, GB3001 (2008).

14. D. Raitsos, S. Lavender, C. D. Maravelias, J. Haralabous, A. J. Richardson, and P. C. Reid, "Identifying four phytoplankton functional types from space: an ecological approach," Limnol. Oceanogr. **53**, 605–613 (2008).
15. A. Bracher, M. Vountas, T. Dinter, J. P. Burrows, R. Röttgers, and I. Peeken, "Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data," Biogeosciences **6**, 751–764 (2009).
16. Z. Ben Mustapha, S. Alvain, C. Jamet, H. Loisel, and D. Dessailly, "Automatic classification of water-leaving radiance anomalies from global SeaWiFS imagery: application to the detection of phytoplankton groups in open ocean waters," Remote Sens. Environ. **146**, 97–112 (2014).
17. P. J. Werdell, C. S. Roesler, and J. I. Goes, "Discrimination of phytoplankton functional groups using an ocean reflectance inversion model," Appl. Opt. **53**, 4833–4849 (2014).
18. O. Farikou, S. Sawadogo, A. Niang, D. Diouf, J. Brajard, C. Mejia, Y. Dandonneau, G. Gasc, M. Crepon, and S. Thiria, "Inferring the seasonal evolution of phytoplankton groups in the Senegalo-Mauritanian upwelling region from satellite ocean-color spectral measurements," J. Geophys. Res. Oceans, **120**, 2331–2349 (2015).
19. S. Twomey, *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements* (Dover Publications, 1977).
20. Z. Lee, K. Carder, R. Arnone, and M. He, "Determination of primary spectral bands for remote sensing of aquatic environments," Sensors **7**, 3428–3441 (2007).
21. R. A. Vandermeulen, A. Mannino, A. Neeley, J. Werdell, and R. Arnone, "Determining the optimal spectral sampling frequency and uncertainty thresholds for hyperspectral remote sensing of ocean color," Opt. Express **25**, A785 (2017).
22. G. Zheng and D. Stramski, "A model based on stacked-constraints approach for partitioning the light absorption coefficient of seawater into phytoplankton and non-phytoplankton components," J. Geophys. Res. Oceans **118**, 2155–2174 (2013).
23. A. M. Ciotti, M. R. Lewis, and J. J. Cullen, "Assessment of the relationships between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient," Limnol. Oceanogr. **47**, 404–417 (2002).
24. A. Chase, E. Boss, R. Zaneveld, A. Bricaud, H. Claustre, J. Ras, G. Dall'Olmo, and T. K. Westberry, "Decomposition of in situ particulate absorption spectra," Meth. Oceanogr. **7**, 110–124 (2013).
25. E. Devred, S. Sathyendranath, V. Stuart, and T. Platt, "A three component classification of phytoplankton absorption spectra: application to ocean-color data," Remote Sens. Environ. **115**, 2255–2266 (2011).
26. H. Gordon and O. Brown, "A semianalytic radiance model of ocean color," J. Geophys. Res. Atmos. **93**, 10,909–924 (1988).
27. D. Stramski, A. Bricaud, and A. Morel, "Modeling the inherent optical properties of the ocean based on the detailed composition of the planktonic community," Appl. Opt. **40**, 2929–2945 (2001).
28. E. Leymarie, D. Doxaran, and M. Babin, "Uncertainties associated to measurements of inherent optical properties in natural waters," Appl. Opt. **49**, 5415–5436 (2010).
29. R. Frouin, "In-flight calibration of satellite ocean-colour sensors," IOCCG Report 14 (IOCCG, 2013).
30. R. Frouin, D. Ramon, E. Boss, D. Jolivet, M. Compiègne, J. Tan, H. Bouman, T. Jackson, B. Franz, T. Platt, and S. Sathyendranath, "Satellite radiation products for ocean biology and biogeochemistry: needs, state-of-the-art, gaps, development priorities, and opportunities," Front. Mar. Sci. **5**, 1–20 (2018).
31. H. M. Dierssen, "Hyperspectral measurements, parameterizations, and atmospheric correction of whitecaps and foam from visible to shortwave infrared for ocean color remote sensing," Earth Sci. **7**, 14 (2019).
32. R. R. Bidigare, M. E. Ondrusek, J. H. Morrow, and D. A. Kiefer, "In-vivo absorption properties of algal pigments," Proc. SPIE **1302**, 290–302 (1990).
33. B. B. Cael, "Sinking versus suspended particle size distributions in the North Pacific Subtropical Gyre," github (2020) https://github.com/bbcael.
34. L. Duyens, "The flattening of the absorption spectrum of suspensions, as compared to that solutions," Biochim. Biophys. Acta **19**, 1–12 (1956).
35. S. Dutkiewicz, A. Hickman, O. Jahn, W. Gregg, C. Mouw, and M. Follows, "Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model," Biogeosciences **12**, 4447–4481 (2015).
36. L. Moore, "Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive," Mar. Ecol. Prog. Ser. **116**, 259–275 (1995).
37. C. S. Roesler, S. M. Etheridge, and G. C. Pitcher, "Application of an ocean color algal taxa detection model to red tides in the southern Benguela," in *10th International Conference on Harmful Algae* (Florida Fish and Wildlife Conservation Commission and Intergovernmental Oceanographic Commission, 2003).
38. B. Cael and E. Boss, "Simplified model of spectral absorption by non-algal particles and dissolved organic materials in aquatic environments," Opt. Express **25**, 25486–25491 (2017).
39. M. Babin, D. Stramski, G. M. Ferrari, H. Claustre, A. Bricaud, G. Obolensky, and N. Hoepffner, "Variations in the light absorption coefficients of phytoplankton, nonalgal particles, and dissolved organic matter in coastal waters around Europe," J. Geophys. Res. Oceans **108**, 3211 (2003).
40. J. M. Sullivan, M. S. Twardowski, J. R. V. Zaneveld, C. M. Moore, A. H. Barnard, P. L. Donaghay, and B. Rhoades, "Hyperspectral temperature and salt dependencies of absorption by water and heavy water in the 400–750 nm spectral range," Appl. Opt. **45**, 5294–5309 (2006).
41. I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philos. Trans. R. Soc. A **374**, 20150202 (2016).
42. E. Boss, M. Picheral, T. Leeuw, A. Chase, E. Karsenti, G. Gorsky, L. Taylor, W. Slade, J. Ras, and H. Claustre, "The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean; contribution of the Tara Oceans expedition," Methods Oceanogr. **7**, 52–62 (2013).
43. P. J. Werdell, S. Bailey, G. Fargion, C. Pietras, K. Knobelspiesse, G. Feldman, and C. McClain, "Unique data repository facilitates ocean color satellite validation," EOS Trans. Am. Geophys. Union **84**, 377–387 (2003).
44. Y. Koren and L. Carmel, "Robust linear dimensionality reduction," IEEE Trans. Visual Comput. Graphics **10**, 459–470 (2004).
45. O. Tamuz, T. Mazeh, and S. Zucker, "Correcting systematic effects in a large set of photometric light curves," Mon. Not. R. Astron. Soc. **356**, 1466–1470 (2005).
46. K. R. Gabriel and S. Zamir, "Lower rank approximation of matrices by least squares with any choice of weights," Technometrics **21**, 489–498 (1979).
47. L. Delchambre, "Weighted principal component analysis: a weighted covariance eigendecomposition approach," Mon. Not. R. Astron. Soc. **446**, 3545–3555 (2015).
48. M. J. Greenacre, *Theory and Applications of Correspondence Analysis* (Academic, 1984).
49. D. A. Jackson, "Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches," Ecology **74**, 2204–2214 (1993).
50. R. B. Cattell and J. Jaspers, "A general plasmode (no. 30-10-5-2) for factor analytic exercises and research," Multivariate Behavioral Research Monographs **67**, 211 (1967).
51. A. P. Chase, E. Boss, I. Cetinić, and W. Slade, "Estimation of phytoplankton accessory pigments from hyperspectral reflectance spectra: toward a global algorithm," J. Geophys. Res. Oceans **122**, 9725–9743 (2017).
52. A. Bricaud, A. Morel, M. Babin, K. Allali, and H. Claustre, "Variations of light absorption by suspended particles with chlorophyll-a concentration in oceanic (case 1) waters: Analysis and implications for bio-optical models," J. Geophys. Res. **103**, 31033–31044 (1998).
53. E. Organelli, A. Bricaud, D. Antoine, and J. Uitz, "Multivariate approach for the retrieval of phytoplankton size structure from measured light absorption spectra in the Mediterranean Sea (BOUSSOLE site)," Appl. Opt. **52**, 2257–2273 (2013).

54. D. Catlett and D. A. Siegel, "Phytoplankton pigment communities can be modeled using unique relationships with spectral absorption signatures in a dynamic coastal environment," J. Geophys. Res. Oceans **123**, 246–264 (2018).
55. C. D. Rodgers, *Inverse Methods for Atmospheric Sounding: Theory and Practice* (World Scientific, 2000), Vol. **2**.
56. X. Xu and J. Wang, "Retrieval of aerosol microphysical properties from aeronet photopolarimetric measurements: 1. Information content analysis," J. Geophys. Res. Atmos. **120**, 7059–7078 (2015).
57. T. Vukicevic, O. Coddington, and P. Pilewskie, "Characterizing the retrieval of cloud properties from optical remote sensing," J. Geophys. Res. Atmos. **115**, D20211 (2010).
58. P. J. Werdell, L. I. McKinna, E. Boss, S. G. Ackleson, S. E. Craig, W. W. Gregg, Z. Lee, S. Maritorena, C. S. Roesler, C. S. Rousseaux, D. Stramski, J. M. Sullivan, M. S. Twardowski, M. Tzortziou, and X. Zhang, "An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing," Prog. Oceanogr. **160**, 186–212 (2018).
59. C. D. Mobley, J. Werdell, B. Franz, Z. Ahmad, and S. Bailey, *Atmospheric Correction for Satellite Ocean Color Radiometry* (NASA, 2016).
60. International Ocean Colour Coordinating Group, "Uncertainties in Ocean Colour Remote Sensing," IOCCG Report Series, No. 18, F. Mélin, ed. (IOCCG, Dartmouth, Canada, 2019).