LIMNOLOGY and OCEANOGRAPHY

ASLO

Limnol. Oceanogr. 9999, 2025, 1–17

doi: 10.1002/lno.70098

RESEARCH ARTICLE

On the fundamental additive modes of ocean color absorption

J. Xavier Prochaska [1,2*] **Patrick Gray** [3,4]

[1]Department of Ocean Sciences, University of California, Santa Cruz, California, USA; [2]Department of Astronomy and Astrophysics, University of California, Santa Cruz, California, USA; [3]School of Marine Sciences, University of Maine, Orono, Maine, USA; [4]Department of Marine Geosciences, Charney School of Marine Sciences, University of Haifa, Haifa, Israel

Abstract

Previous principal component analyses of ocean color absorption coefficient spectra $a(\lambda)$ have shown the variation in these data is captured by a few eigenfunctions. Here, we perform an unsupervised, non-negative matrix factorization (NMF) of $a(\lambda)$ to derive their fundamental and physically interpretable modes. When applied independently to two large datasets—one semi-empirical and one from inline measurements of the *Tara* Microbiome expedition—we find that four NMF basis functions describe $>99.9\%$ of the variance in each. Furthermore, despite significant differences between the datasets in methodology and by geographic and temporal acquisition, the two sets of basis functions show very similar features at wavelengths $\lambda \approx 400-750$ nm. Two of the modes capture the amplitude and spectral slope of absorption by color dissolved organic matter and/or detritus. The other two describe absorption by phytoplankton ($a_{\rm ph}$) separated into the pigments that couple tightly to the chlorophyll *a* (Chl *a*) 675 nm feature and another that captures $a_{\rm ph}$ variability at $\approx$450 nm. Together, the majority of ocean color absorption is physically described by these four fundamental modes. We present several applications of the NMF analysis including the exploration of geographic trends in particulate composition, the search for outlier absorption spectra, and the application of a new, additive decomposition of $a_{\rm ph}$. Lastly, we detail the limitations of this technique, especially in the context of mechanistic approaches more commonly adopted in the literature.

The constituents near the ocean's surface—for example, sea water, phytoplankton, dissolved organics, minerals, detrital particles—absorb and scatter sunlight to give the ocean its apparent color. These constituents play a variety of physical and biogeochemical roles in the ocean system and tracking them through space and time is critical to our understanding of key oceanographic processes (e.g., Ismail and Al-Shehhi 2023; Litchman et al. 2015; McClain 2009). As such, a primary focus of ocean color observations by oceanographers is to determine the constituents in sea water across the Earth's ocean including coastal waters (e.g., Loisel et al. 2018; Werdell et al. 2013). These measurements may then be used operationally to monitor hazards such as red tides or to understand biogeochemical processes ranging from discharged river sediments to global phytoplankton growth and composition (e.g., IOCCG Protocol Series 2008).

A standard approach to any such analysis is to consider one or more inherent optical properties (IOPs), properties of the medium that are invariant to the ambient light field. Inherent optical properties like absorption coefficient spectra $a(\lambda)$ are set by the components of ocean water—phytoplankton, colored dissolved organic matter (CDOM), pure (salt) water—and offer a path to focus the analysis on the items of scientific interest as opposed to other factors that impact the observations (e.g., the atmosphere or the Sun's angle).

Despite the complexity of the living ocean, previous research has emphasized that the global variations in IOPs like the

*Correspondence: jxp@ucsc.edu

Associate editor: David Antoine

Data Availability Statement: All of the code and results presented here are available on GitHub in these two repositories: https://github.com/AI-for-Ocean-Science/cnmf, https://github.com/AI-for-Ocean-Science/ocean-color. All Tara Microbiome optical data are available in raw form at NASA's SeaBASS archive via the search keyword "Tara_Microbiome." All data prepared and formatted for this study is available on GitHub https://github.com/patrickcgray/spatial_patchiness_tara, easily ingestable as a geofeather at multiple stages of processing.

absorption coefficient may be reduced to a few (mathematical) eigenvectors (e.g., Cael et al. 2020a; Garver et al. 1994). This reflects several properties of the ocean and the datasets: (1) the relative similarity of ocean constituents across the open ocean (e.g., Bricaud et al. 1995); (2) physical correlations between the components that are present even in complex waters (Cael et al. 2020a; Morel 1988); and (3) the similarity of absorption profiles between several components, for example, CDOM and detritus (Stramski et al. 2001) or phytoplankton pigments (A. Chase et al. 2013). These properties motivate decompositions of IOPs to a few basis functions that contain (nearly) all of the information content.

In oceanography and other sciences, a common approach to data decomposition is to perform a principal component analysis (PCA; Jolliffe and Cadima 2016, often referred to as empirical orthogonal functions or EOF). A PCA decomposes a dataset into a series of orthonormal eigenvectors which minimize the variance in the data relative to their mean. It is optimal and rigorously mathematical, but the resultant eigenspectra need not on their own have scientifically relevant shapes. Therefore, the reduced representation described by the derived eigenvectors may offer limited interpretability. One is thus motivated (when possible) to consider other, more physically relevant decompositions.

Indeed, the ocean color literature includes many examples of mechanistic decompositions of IOPs, especially of absorption coefficient spectra (e.g., Garver and Siegel 1997; Kehrli et al. 2024; Werdell et al. 2013; Zhang et al. 2015). These models adopt empirically derived or motivated components from laboratory work and/or analyses of isolated components of seawater (e.g., CDOM). While these models are more physical than PCA and therefore highly interpretable, they are not guaranteed to describe the great diversity of absorption spectra. Or, if they have sufficient complexity (e.g., many components), they may incur significant degeneracy and correlations that limit applicability. In this manuscript, we seek a decomposition that compactly describes the variance in absorption spectra while maintaining interpretability.

To this end, we introduce a new technique for ocean color applications known as Non-negative matrix factorization (NMF) analysis (Lawton and Sylvestre 1971). Similar to PCA, the NMF decomposition is unsupervised, that is, the algorithm "learns" the dominant correlations in the data to optimally describe the dataset. Furthermore, it is compact, that is, the NMF decomposition is designed to describe variations in the dataset with the fewest number of modes. The primary and critical difference from PCA is that the basis functions (akin to PCA eigenvectors), coefficients, and input data are all non-negative.

Mathematically, NMF is a regularized matrix decomposition of non-negative data into a set of non-negative basis functions and positive coefficients. The NMF algorithm was introduced originally to describe chemical systems (Lawton and Sylvestre 1971), and has been applied in a variety of fields ranging from the flux of astronomical sources (G. Zhu, unpublished) to bioinformatics (Taslaman and Nilsson 2012). It is well suited to IOPs like $a(\lambda)$ where the quantity is non-negative. Furthermore, NMF is advantageous relative to PCA when the mean of the dataset is not relevant or is itself non-physical.

In the following, we perform an NMF decomposition of absorptions spectra from two distinct and large datasets. First and foremost, we seek a compact set of highly interpretable basis functions that describe the primary variations in absorption spectra across the global ocean. In turn, we will assess the number of modes required to explain a high percentage ($\gg 99\%$) of the variance while maintaining interpretability. We will then explore geographic trends in the NMF coefficients calculated from the decomposed absorption spectra and examine spectra that are poorly fitted by the basis functions (aka outliers). Last, by demonstrating the value of NMF analysis on absorption coefficient spectra, we may promote like-studies on other non-negative and frequently measured quantities in ocean optics: backscattering, diffuse attenuation coefficients, remote-sensing reflectances.

Materials

Data

We seek a decomposition of the absorption coefficient that is broadly applicable, that is, applies to both Case I (phytoplankton dominated, and typically but not exclusively, in clear and open ocean waters) and Case II (turbid, complex, coastal) waters. To this end, we considered absorption spectra from the literature obtained across a range of geographic locations and sampling a range of water types. Furthermore, we have restricted to datasets with a high quality control and those with spectral sampling no coarser than 10 nm.

For this manuscript, we selected the absorption spectra from Loisel et al. (2023) (hereafter L23) and the *Tara* Mission Microbiomes AtlantECO (hereafter just; *Tara*; Gray et al. 2025). There are, of course, additional datasets available that satisfy our criteria (e.g., other *Tara* missions; Jordan et al. 2024; Valente et al. 2022). However, we considered the two datasets examined here to be sufficient as they span the global ocean and were derived with different protocols. Geographic maps of each dataset may be found in Fig. 1 of L23 and Fig. 9 of this manuscript (*Tara*).

Indeed, the instrumentation and processing methodology for these two datasets is entirely independent. The $a_{\rm ph}$ spectra from L23 was obtained with the filter-pad spectrophotometric method where one first measures the total particulate absorption spectrum $a_{\rm p}(\lambda)$ and then subtracts off an independent estimate of the non-algal particulate absorption to infer $a_{\rm ph}$ (IOCCG Protocol Series 2018). The $a_{\rm ph}$ spectra collected and collated by L23 were then combined with simulated absorption by CDOM and detritus (each an exponential function

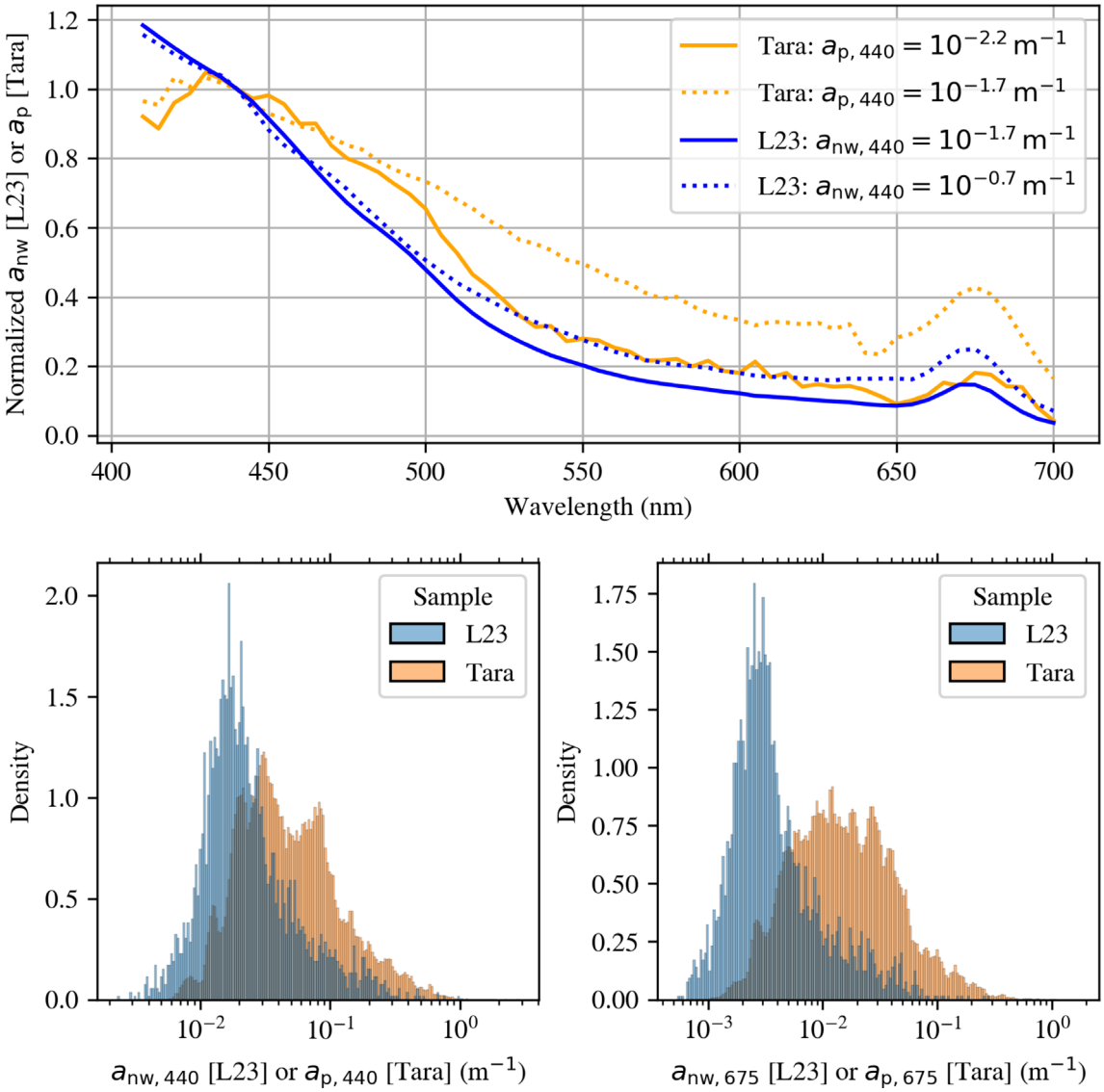

Fig. 1. The top panel shows four representative absorption spectra (with varying magnitude) from the two major datasets examined in this manuscript: *Tara* (orange; a_p) and L23 (blue; a_{nw}). We have selected examples with high (solid) and low (dotted) absorption at 440 nm as described by the legend. All of these spectra have been normalized at that wavelength, for presentation purposes only. The lower panels describe the distributions of absorption at 440 nm (left, a_{440}) and 675 nm (right, a_{675}) of the two datasets (non-water absorption for L23 and particulate absorption for *Tara*). It is evident that the *Tara* dataset includes a larger fraction of spectra with higher absorption at all wavelengths $\lambda > 440$ nm stemming from the coastal focus of this expedition.

with unique amplitude and shape). For the NMF analysis that follows, we use the non-water absorption spectra $a_{nw}(\lambda)$ of L23 as one set of inputs. We restrict to wavelengths $\lambda = 410 - 700$ nm and adopt the 5 nm sampling of their dataset.

The absorption spectra from *Tara* was collected via an AC-s absorption meter (Seabird Scientific) running continuously on a ship-based underway flow-through system. Absorption by sea water and dissolved matter (i.e., CDOM) has already been removed from the *Tara* dataset as part of their standard methodology (Slade et al. 2010) which subtracts the dissolved component (operationally defined as absorption with a 0.2-μm filter attached to the instrument's intake) from the total yielding the particulate absorption spectrum:

$$a_p(\lambda) \equiv a(\lambda) - a_{0.2\mu m}(\lambda) \quad (1)$$

This also helps remove instrument drift and biofouling. In this manuscript, we analyze a newly reprocessed version of *Tara* Microbiome (Gray et al. 2025). Regarding additional processing, we removed any spectrum containing one or more

negative values on concerns that the data were not properly subtracted. We then rebin the *Tara* spectra onto the same 410–700 nm grid as L23 with 5 nm sampling using nearest-neighbor interpolation. This sampling is close to the native sampling of the AC-s absorption meter, and these resampled, particulate absorption spectra $a_p(\lambda)$ are the 2nd input dataset for our NMF analysis.

Altogether, there are 3320 (239,880) spectra for L23 (*Tara*). Figure 1 shows a set of representative data taken from water with a distinct set of properties characterized by the absorption at 440 nm, a_{440}. One recognizes the strong and increasing absorption at $\lambda < 500$ nm associated with CDOM and/or detritus layered on top of phytoplankton pigment absorption and the "bump" at $\lambda \approx 675$ nm characteristic of Chl *a* (Bricaud et al. 2004). These are common across spectra, but the heights and slopes of these features vary. Our formalism seeks to describe such variations in the dataset with a limited number of basis functions.

The lower panels of Fig. 1 show the distribution of absorption at 440 nm and 675 nm. The two datasets overlap although the L23 distributions tend toward lower values that are more characteristic of open waters. The other point to emphasize is that the incidence of highly turbid waters (e.g., $a_{440} \sim 1\ m^{-1}$) is rare in both datasets. Future work may consider the analysis of a balanced dataset that samples more uniformly in a metric like a_{440}. Future work could also include additional datasets that emphasize coastal waters and/or regions not well covered by the ones studied here.

Formalism

The NMF algorithm was introduced by Lawton and Sylvestre (1971) and then refined by Paatero and Tapper (1994); it has since been employed in a range of fields. Provided a dataset with dimensions $N \times M$ (number of spectra, number of features [i.e., number of wavelength channels]), one seeks solutions to the matrix equation:

$$X = WH \tag{2}$$

where W is a matrix of basis functions (akin to PCA eigenvectors) and H are the coefficients that encapsulate the decomposition of each input spectrum. Like PCA, the NMF algorithm is additive and linear, but unlike a PCA decomposition all elements of X, H, and W are required to be non-negative. Also similar to PCA, one may construct W to have a lower dimensionality m than the feature space (i.e., $m < M$). In this case, the matrices W and H have shapes $M \times m$ and $N \times m$, respectively, and the m basis functions provide a reduced (i.e., compact) representation of the dataset.

Because there is typically no exact solution to Eq. 2 (even with $m = M$), one instead seeks solutions that minimize a cost function $\mathcal{L}$. The standard function resembles the χ^2 function,

$$\mathcal{L} = \|(X - WH)\|^2 \tag{3}$$

which assumes homoscedastic errors. Zhu extended this formalism to include heteroscedastic uncertainties and masked data encapsulated in a weight matrix V:

$$\mathcal{L} = \left\| V^{1/2} \cdot (X - WH) \right\|^2 \tag{4}$$

They also provided a Python package (Zhu 2023) to solve for X adopting the technique of Lee and Seung (1999).

We have taken their code, and the follow-up work of Ren et al. (2018) for our NMF decomposition analysis. One modification to this standard NMF treatment is that we normalize the basis functions to sum to unity:

$$\sum_i W_{ij} = 1 \text{ (for all } j) \tag{5}$$

This gives meaning to the relative values of the coefficients. The other algorithmic advance implemented is to allow the user to fix one or more of the basis functions in the W matrix, that is to explicitly specify one or more of the modes. We refer to this option as a constrained NMF.

Results

In this section, we decompose the absorption coefficient spectra for the L23 (specifically, non-water absorption spectra a_{nw}) and the *Tara* (particulate absorption, a_p) datasets. For each dataset, the NMF analysis yields a set of basis functions (W^{L23}, W^{Tara}) using the methodology presented in the previous section. At the same time, we derive coefficients (H^{L23}, H^{Tara}) for each spectrum of each dataset.

Non-negative matrix factorization decomposition of L23

With the $a_{nw}(\lambda)$ spectra from L23 as the input, we performed an NMF decomposition with $m_{NMF} = 2$ to 10 basis functions. For the calculation of the cost function given by Eq. 4, we assumed a constant error of $0.05\ m^{-1}$ to provide equal weighting of the data. The results are invariant to the value provided it is a constant and non-zero. Figure 2 shows the unexplained variance as a function of m_{NMF} and these are compared against a similar evaluation using a PCA decomposition. For $m \leq 5$, the two approaches explain nearly the same portion of variance per component. Similar to previous work (e.g., Cael et al. 2020a; Garver et al. 1994), we find that only $m = 3$ components are required to explain $\approx 99.9\%$ of the variance and $m = 5$ components explain $\approx 99.99\%$. Furthermore, the variance per mode declines more steeply than a -2 power-law slope indicating the information content in the higher order modes is very low. The curves depart slightly at $m > 5$ which we hypothesize is due to the PCA better describing small fluctuations in the data (e.g., noise) owing to its partially negative basis functions.

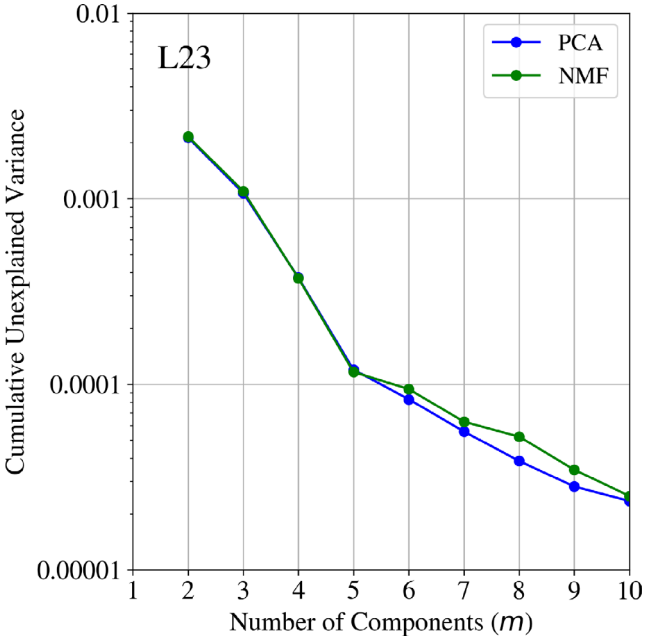

Fig. 2. The unexplained variance (i.e., 1 − explained variance) for principal component analysis (PCA, blue) and non-negative matrix factorization (NMF, green) decompositions of the L23 dataset as a function of the number of components (m). In both cases, the decompositions explain $\approx 99.9\%$ of the variance with only $m = 3$ components and $\approx 99.99\%$ of the variance with $m = 5$.

For the remainder of the manuscript, we will focus on NMF models with $m_{\mathrm{NMF}} = 4$ which we will argue represent the fundamental modes of ocean color absorption. The choice of $m_{\mathrm{NMF}} = 4$ was motivated by two factors. First, we found that an $m_{\mathrm{NMF}} = 3$ model insufficiently captures variations related to phytoplankton absorption. These variations are small but of significant scientific interest to warrant inclusion. Second, we found that a 5$^{\mathrm{th}}$ basis function ($m_{\mathrm{NMF}} = 5$) shows high frequency features that are more characteristic of noise than true absorption. Together, this led us to adopt $m_{\mathrm{NMF}} = 4$ decompositions for the L23 and *Tara* datasets.

Figure 3 presents a comparison of the PCA and NMF decompositions of the L23 dataset with $m_{\mathrm{NMF}} = 4$. The former is explicitly orthonormal and therefore exhibit negative "absorption" features which are non-physical: no true absorption spectra can take on negative values. Indeed, we may struggle to interpret anything other than the first PCA mode. In contrast, the NMF basis functions are explicitly non-negative, and the first two (W_1, W_2) are easily interpreted: these resemble absorption by CDOM and Chl *a*, respectively. In "Physical interpretation of the NMF basis functions" section, we interpret the other two basis functions (W_3, W_4) as variations in the shape of CDOM/detritus and phytoplankton absorption.

Supporting Information Fig. S1 shows NMF decompositions for two representative non-water absorption spectra from the

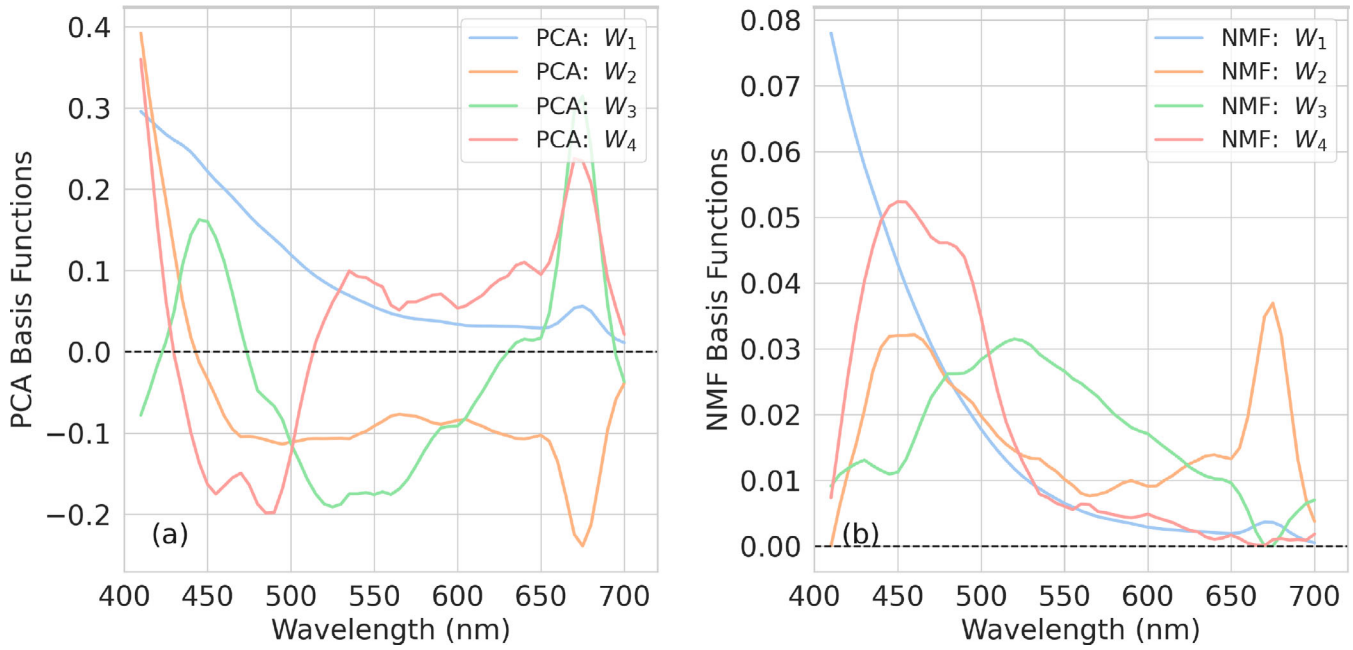

Fig. 3. Basis functions for the L23 dataset from the principal component analysis (PCA, left) and non-negative matrix factorization (NMF, right). While the first two PCA eigenfunctions bear some resemblance to known $a(\lambda)$ spectra, not even these follow physical functions closely. For example, W_2 for the PCA shows strong negative "absorption" at the 670 nm peak of Chl *a*. In contrast, the first two NMF basis functions (W_1, W_2) strongly resemble those of colored dissolved organic matter (CDOM) and phytoplankton respectively. We discuss these and further interpretations in "Physical Interpretation of the NMF Basis Functions" section.

L23 dataset. In both cases, the NMF model closely matches the data, confirming the NMF basis functions capture the detailed features of absorption spectra. Examining the decompositions, we note the absorption blueward of ≈ 550 nm is dominated by the 1st basis function W_1^{L23} and the H_1^{L23} coefficient is correspondingly large. Variations in the other components suggest differences in phytoplankton absorption and/or differences in the slope of the CDOM/detritus absorption. We will explore these later in the manuscript.

Our NMF analysis generated a decomposition for every L23 absorption spectrum and the NMF coefficients for all of these are provided in Supporting Information Table S1, indexed by the row number of the L23 dataset. The basis functions W^{L23}, meanwhile, are provided in the data repository on GitHub that accompanies this manuscript.

Non-negative matrix factorization decomposition of the Tara dataset

Consider next an NMF analysis of the *Tara* dataset. First, we fit the *Tara* dataset using the four basis functions derived from the L23 data (W^{L23}; Fig. 3b) to test the extent to which these basis functions generalize. We refer to the resultant coefficients as H^{L23}_{Tara}, that is, L23 basis functions applied to *Tara* data. Even though the L23 basis functions were derived from non-water absorption spectra $a_{nw}(\lambda)$ and then applied to the particulate absorption spectra of *Tara* we find the fits explain $>99.5\%$ of the variance in the *Tara* dataset. In this respect, the NMF basis functions may be considered a general representation of non-water absorption spectra.

We also find that for each of the four basis functions, the distribution of H^{L23}_{Tara} values are shifted to higher values. This is especially true for $H^{L23}_{Tara,2}$ and $H^{L23}_{Tara,4}$ which are nearly an order of magnitude higher on average. This follows expectation given that *Tara* Microbiome surveyed a higher fraction of coastal waters than the more open-ocean waters considered by L23. Although the *Tara* experiment has subtracted off absorption due to dissolved organic matter (with size $<0.2\,\mu$m; Slade et al. 2010), the signal for the CDOM-like basis function (W_1^{L23}) also exceeds that of the L23 dataset. We expect, that these high $H^{L23}_{Tara,1}$ values are due to non algal particles NAP (*see* "W_1 represents the degree of CDOM and/or detritus absorption" section).

We have performed a separate and independent $m_{NMF}=4$ NMF analysis of the *Tara* dataset, deriving a unique set of basis functions W^{Tara} and a complete set of coefficients H^{Tara} for the 239,880 absorption spectra (listed in Supporting Information Table S2, and indexed by the Unix nanosecond timestamp of each observation). Figure 4 compares the derived basis functions from *Tara* (W^{Tara}) with those from L23 (W^{L23}). The first, somewhat striking result is that with the exception of W_3 the basis functions closely resemble one another. The 1st basis function for each (W_1) shows a nonlinear increase in absorption to the blue with the primary difference being a somewhat shallower slope for *Tara* (measured in "W_1 represents the degree of CDOM and/or detritus absorption" section). The W_2^{Tara} with W_2^{L23} profiles are also very similar; the small differences are a trade-off between weaker absorption in *Tara* at $\lambda \approx 450$ and ≈ 600 nm with higher absorption at ≈ 525 and 675 nm. Lastly, the W_4 basis functions of each dataset are dominated by absorption at $\lambda \approx 430-550$ nm. Given the significant differences in methodology as well as in the geographic and temporal acquisition, the commonality between the two datasets is remarkable and implies these are fundamental basis functions for absorption in the global ocean.

The greatest difference between the NMF decompositions is in the 3rd basis function, W_3. The gross shapes are similar and each exhibits a (non-physical) absorption "trough" at $\lambda > 675$ nm, but the W_3 absorption is shifted to redder wavelengths ($\lambda \approx 500-650$ nm). Below, we argue that this difference in W_3 primarily reflects the dominance of CDOM in L23 and NAP in *Tara*.

Discussion

Physical interpretation of the NMF basis functions

The previous section presented the NMF decompositions independently for the L23 and *Tara* datasets. Figure 4 compares the two and we reemphasize the commonality of the derived basis functions. Unlike previous unsupervised decompositions of $a(\lambda)$ spectra (e.g., PCA; Cael et al. 2020b), these basis functions are non-negative and therefore, in principle, more interpretable than previous decompositions. In this section, we proceed to connect each basis function to a distinct and physically meaningful aspect of ocean color absorption.

W_1 represents the degree of CDOM and/or detritus absorption

Readers familiar with the absorption coefficient will recognize that W_1 from each dataset exhibits a steeply rising (i.e., nonlinear) absorption profile representative of CDOM and/or NAP. Previously, these components have been modeled with either an exponential or power-law functional form. Figure 5 shows separate functional fits to the W_1^{L23} and W_1^{Tara} basis functions where we have assumed a power-law,

$$a(\lambda) = A_\beta \lambda^{-\beta} \tag{6}$$

and an exponential

$$a(\lambda) = A_S \exp[-S(\lambda - 440\ \text{nm})] \tag{7}$$

restricted to the interval $\lambda = 410-530$ nm where CDOM/detrital slopes are commonly assessed (e.g., Kehrli et al. 2023, although our results are largely insensitive to the range). Adopting standard techniques and assuming constant weighting at each wavelength of the basis function, we derive best-fit shape parameters for these functional forms: (1) power-law: $\beta^{L23}=7.3$ and $\beta^{Tara}=5.5$ for L23 and *Tara*, respectively, and (2) exponential: $S^{L23}=0.016$ and $S^{Tara}=0.012\,\text{nm}^{-1}$. For W_1^{L23}, the power-law exponent and

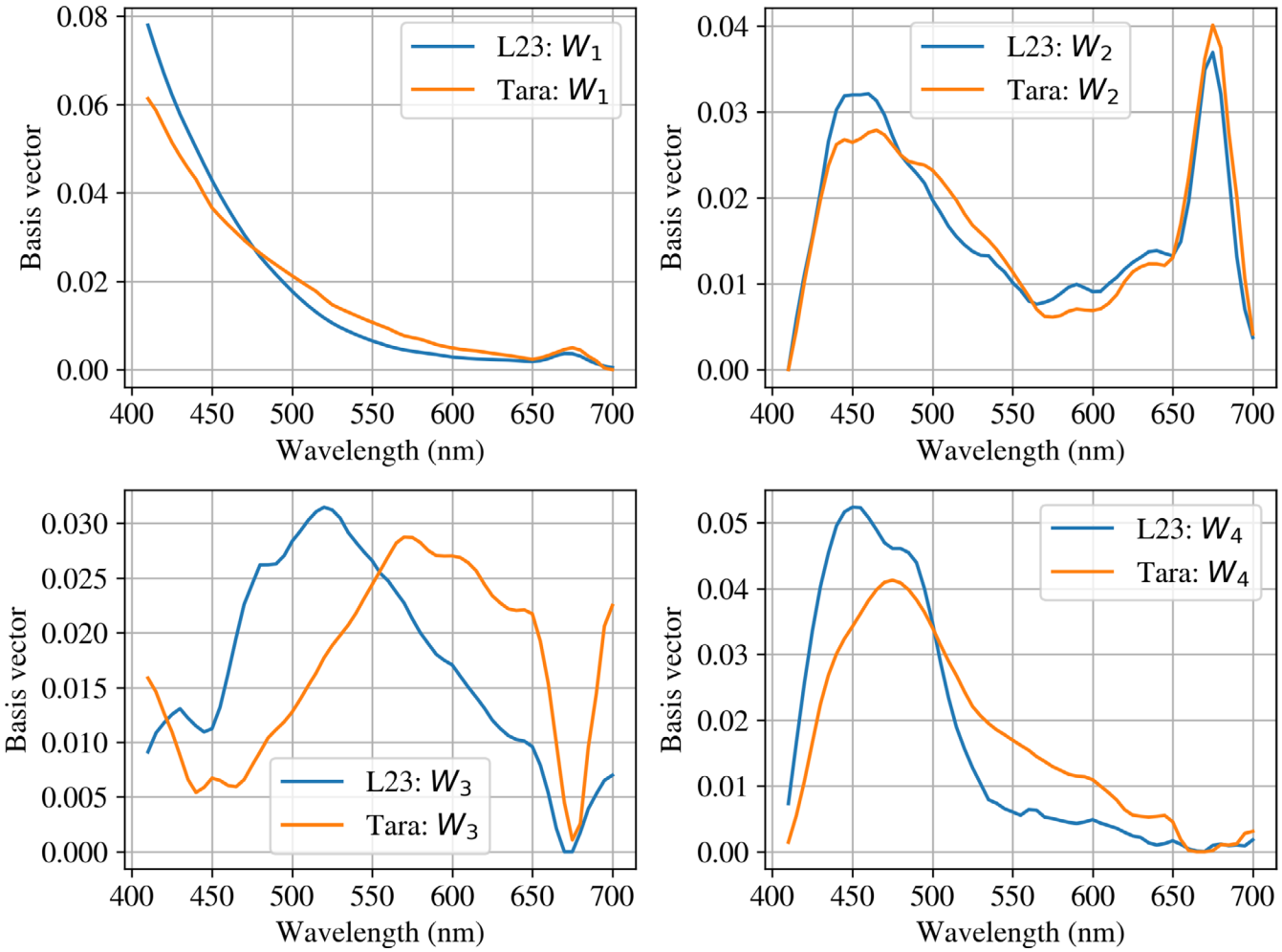

Fig. 4. Comparison of the non-negative matrix factorization (NMF) basis functions derived independently for the L23 (blue; same as Fig. 3b) and *Tara* (orange) datasets. Qualitatively, the first two and last basis functions—W_1, W_2, W_4—are remarkably similar. The $W_1^{\rm Tara}$ basis function shows a nonlinearly increasing absorption to shorter wavelengths similar to the colored dissolved organic matter (CDOM) and detritus driven $W_1^{\rm L23}$ profile, but shallower and more consistent with pure non-algal particle absorption. The $W_2^{\rm Tara}$ basis functions are very similar indicating the pigments that co-vary with Chl *a* are similar between the two experiments. We draw a similar conclusion for W_4 which we infer tracks additional phytoplankton absorption that does not co-vary strictly with Chl *a*. The largest differences lie in the W_3 basis functions which compensate for differences in slope for CDOM and/or detritus.

exponential constant compare favorably to literature values for CDOM (Stramski et al. 2001) and the parameters for CDOM adopted by L23. This slope has been used previously as an indicator for CDOM composition, where steeper slopes suggest older (refractory) compounds and flatter slopes suggest newer (labile) compounds (Carder et al. 1989).

Comparing the slopes for *Tara* with those for the L23 basis function, we find the $W_1^{\rm Tara}$ basis function has a shallower slope ($\beta^{\rm Tara} < \beta^{\rm L23}$, $S^{\rm Tara} < S^{\rm L23}$) which is more reflective of NAP than CDOM (Iturriaga and Siegel 1989). The estimated slope for $W_1^{\rm Tara}$ is also consistent with the NAP spectrum adopted by previous fits to *Tara* spectra (A. Chase et al. 2013). We can further test the inference that $W_1^{\rm Tara}$ primarily expresses NAP absorption by comparing the $H_1^{\rm Tara}$ coefficients for the *Tara* spectra against an estimate of CDOM concentration from the *Tara* project via a concurrent measurement from a UV fluorometer. We find the $H_1^{\rm Tara}$ coefficients are not tightly correlated with the CDOM estimate (Kendall's $\tau = 0.2$), and we conclude that the absorption expressed by $W_1^{\rm Tara}$ is primarily NAP.

We emphasize that the $W_1^{\rm L23}$ basis function simultaneously captures the absorption from CDOM and NAP in the L23 dataset (their a_g and a_d). This is illustrated in the right-hand panel of Fig. 5 which compares $H_1^{\rm L23}$ against the summed absorption of CDOM and detritus ($a_{\rm dg}$) evaluated at 405 nm. These are tightly correlated and we confirm that the $W_1^{\rm L23}$ basis function expresses the majority of the absorption associated with these two components. We further emphasize that because $W_1^{\rm L23}$ is the only basis function with rising absorption at the bluest wavelengths, the individual components are

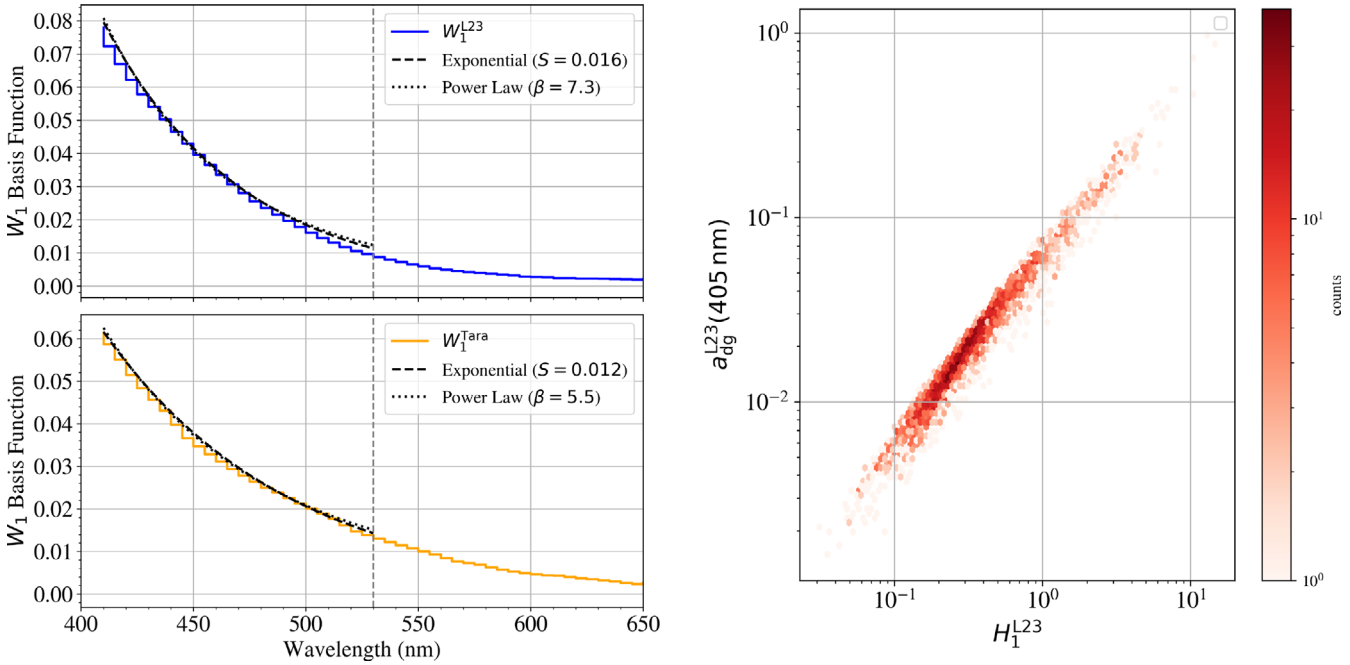

Fig. 5. (Left) Fits to the first non-negative matrix factorization (NMF) basis functions from the L23 and *Tara* datasets: an exponential model with scaling parameter S (black dashed) and a power-law model with exponent β (black, dotted). Each of these is a good description of the data and we find that the *Tara* profile is systematically shallower at these wavelengths. This follows from the generally stronger absorption by colored dissolved organic matter (CDOM) relative to non-algal particles in the L23 spectra. (Right) Comparison of the H_1^{L23} coefficient with the combined absorption by CDOM and detritus a_{dg} evaluated at 405 nm. The very tight correlation between the two demonstrates that H_1^{L23} successfully describes the absorption from these two constituents.

degenerate, that is, our NMF analysis indicates that it is very difficult to differentiate between these two components based on absorption alone.[1]

One also notes that the W_1 basis function for each dataset exhibits a weak absorption feature at $\approx$675 nm that is not present in any real CDOM or detritus spectrum. We expect this is due to co-variance between Chl *a* and CDOM/detritus absorption. It could be removed (e.g., one could interpolate across the feature) and new basis functions with that modification could be performed with the constrained NMF method described in "Formalism" section.

W_3 represents the slope of CDOM and/or detritus absorption

For both L23 and *Tara*, the W_3 basis function expresses absorption across a majority of wavelengths in the analysis window (Fig. 4). Furthermore, unlike the other basis functions, W_3 does not resemble any common features of phytoplankton. Our examination of W_3 indicates it primarily captures diversity in the spectral slope of CDOM and/or detritus absorption.

As one example, Fig. 6a shows a fit from the *Tara* dataset with very strong detrital absorption. In this case, the spectral slope at $\lambda < 600$ nm is much shallower than the W_1^{Tara} basis function that generally captures NAP absorption. To compensate, the fit includes a large H_3^{Tara} component (and an unphysically large H_2^{Tara} component). This is one of many examples where H_3^{Tara} is driven to large values to yield detritus absorption with a shallow slope.

This inference that W_3 accounts for shallow CDOM and/or detritus absorption is further supported by Fig. 6b. Here, we plot a measure of the slope of a_{dg} from L23 using the ratio of absorption at 500 to 405 nm: $^{L23}(500)/^{L23}(405)$. For spectra with the lowest H_3^{L23} values (here plotted at 10^{-3} for presentation purposes), the $^{L23}(500)/^{L23}(405)$ values are smallest, that is, the a_{dg} absorption is steepest. As H_3^{L23} increases, the ratio increases albeit with significant scatter. A Kendall's tau rank correlation test rules out the null hypothesis of an absence of association at very high confidence level (>99.9999%). We conclude that W_3 primarily tracks the slope of CDOM and/or detritus absorption in the ocean.

W_2 and W_4 represent absorption by phytoplankton

An inspection of W_2 gives the impression of absorption generally attributed to Chl *a* and related pigments (e.g., Mobley 2022). Of the four basis functions, it is also the most expressive of the 675 nm absorption feature of Chl *a*. This is further emphasized in Fig. 7 which compares the H_2^{Tara} values against the standard metric for Chl *a* from the *Tara* experiment, the 675 nm line height (*Tara* LH; Boss et al. 2001; Roesler and Barnard 2013). These two metrics are highly

[1]We note that extending the analysis to $\lambda < 400$ nm yields similar results and degeneracy.

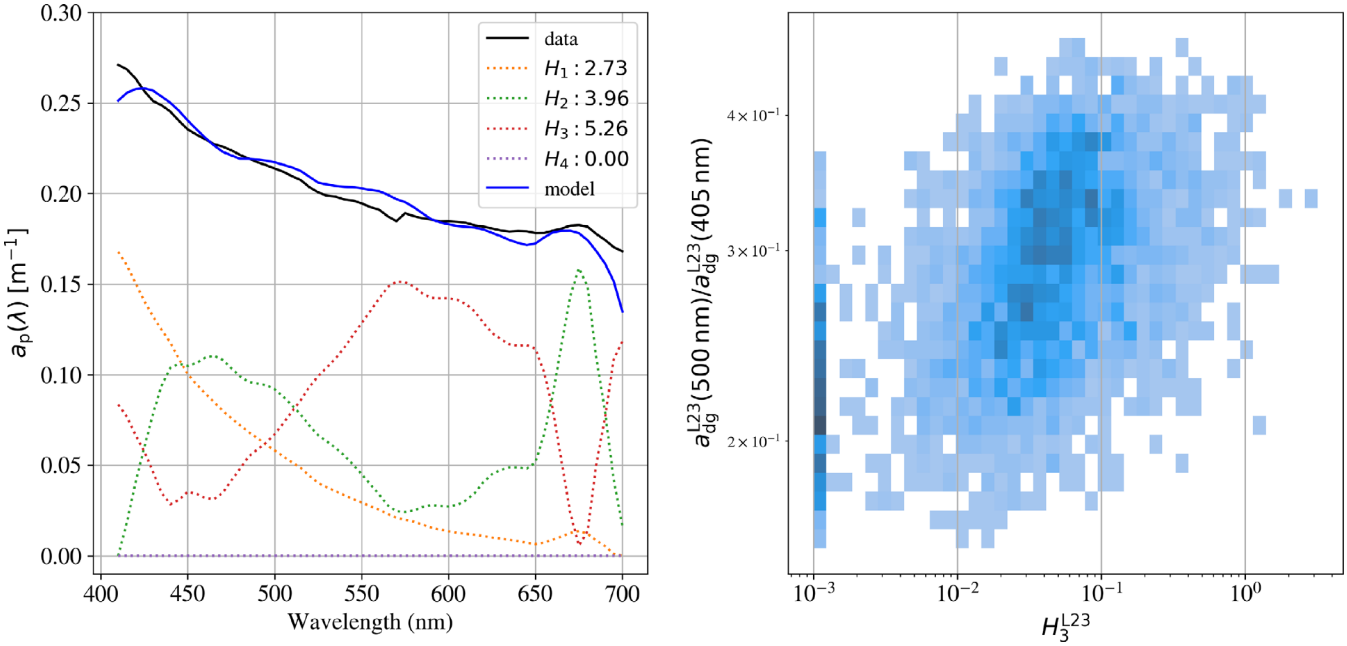

Fig. 6. (Left) The black curve shows a particulate absorption coefficient spectrum $a_p(\lambda)$ from the *Tara* Expedition with very strong detritus absorption (observed at 27.04 S, 48.50 W on November 19, 2021). The colored, dotted curves show the decomposition of $a_p(\lambda)$ into the four non-negative matrix factorization (NMF) basis functions with the total model given by the solid, blue curve. Because the detrital absorption is much shallower than the W_1^{Tara} basis function, the decomposition also includes a strong W_3^{Tara} component which captures the detrital absorption at $\lambda \approx 600$ nm. Examples like this indicate the H_3^{Tara} coefficient primarily expresses strong and shallow detrital absorption. (right) Binned histogram of H_3^{L23} coefficients vs. an estimate of the slope of colored dissolved organic matter (CDOM) and detritus absorption a_{dg} based on the ratio of absorption at 500 nm relative to 405 nm (shallower slopes have higher values). The two are highly correlated; a Kendall's tau test rules out the null hypothesis for no correlation at greater than 99.999% confidence. Similar to the *Tara* example, higher H_3^{L23} coefficients indicate shallower absorption by CDOM and detritus. Note that for presentation purposes, we set a minimum value to H_3^{L23} of 10^{-3}.

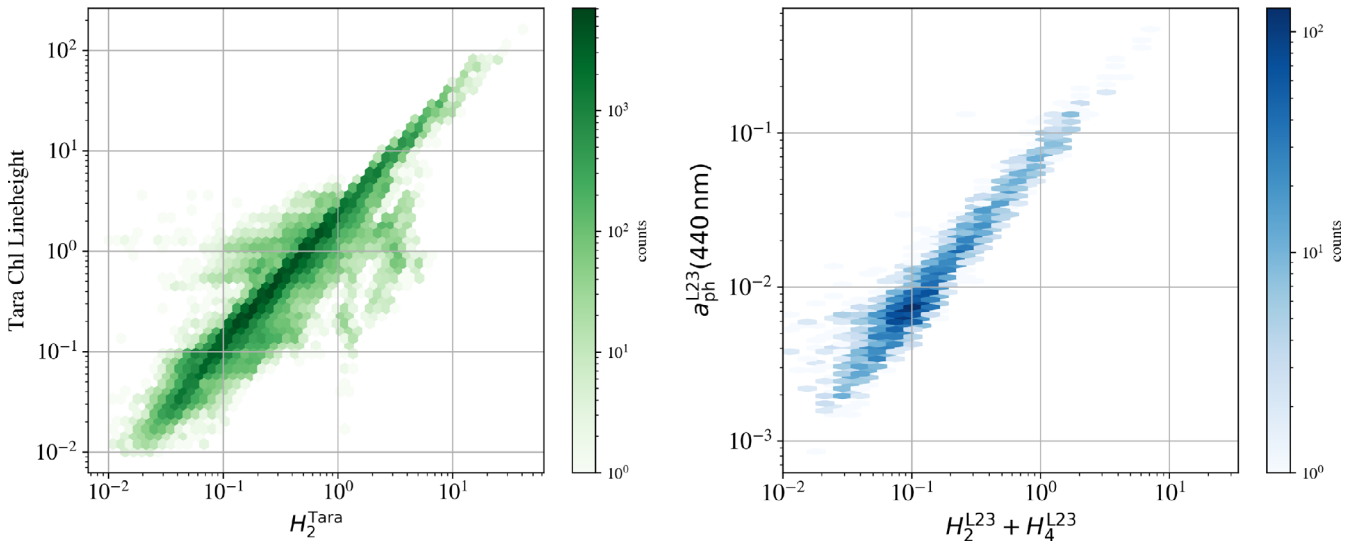

Fig. 7. (Left) Histogram of the Chl *a* *Tara* LH measurement at ≈ 675 nm against the H_{Tara}^2 coefficient which most expresses the 675 nm feature. The observed, tight correlation between the two quantities is expected and indicates the W_2 basis function tracks Chl *a* absorption and the pigments most closely associated with Chl *a*. (Right) Histogram of the absorption by phytoplankton at 440 nm $a_{ph}(440)$ from the L23 dataset against the sum of the coefficients $(H_{L23}^2 + H_{L23}^4)$ for the basis functions that we argue primarily describe phytoplankton absorption. The tight correlation observed is significantly poorer if we consider only one of the two basis functions.

correlated (Kendall's tau statistic of 0.88; p-value $< 10^{-5}$) and the correlation does not improve by including any of the other NMF basis functions. On its own, the $H_2^{\rm Tara}$ coefficient provides a direct estimation of the Chl a concentration as assessed by *Tara* LH. Assuming a linear relationship between the two quantities, we find $\log Tara\ {\rm LH} \approx 1.25 \log {\rm H}_2^{\rm Tara} + 0.28$, which recovers *Tara* LH with a mean absolute error of $\approx 23\%$.

The W_2 basis function also shows absorption at $\lambda \approx 450$ nm characteristic of Chl b and other pigments known to covary with Chl a. The absorption at these bluer wavelengths, however, is less than that typically observed in published absorption spectra of phytoplankton (e.g., Lomas et al. 2024; Stramski et al. 2001). This "missing" absorption is captured by W_4 which we infer defines pigments, functional families and/or packaging that correlates with Chl a but not strictly. In short, we argue that W_2 and W_4 together define the absorption from phytoplankton with W_2 capturing the absorption tightly correlated with the 675 nm feature and W_4 describes variability in $a_{\rm ph}$ at ≈ 450 nm due to pigment variations and packaging effects.

To test this hypothesis, we have also performed an NMF decomposition of the phytoplankton absorption spectra $a_{\rm ph}$ from the L23 dataset (*see* "Modeling variations in $a_{\rm ph}$ with NMF" section for details). We find the resultant two basis functions closely resemble those of W_2 and W_4. Second, Fig. 8 shows a least-squares fit to the sum of $W_2^{\rm L23}$ and $W_4^{\rm L23}$. For this fit, we take the following pigments from Bricaud et al. (2004): Chl a, Chl b, Chl $c1$, Peri, β-Car. These were chosen primarily because they show significant absorption at $\lambda \approx 400 - 600$ nm. Aside from the data at $550\,{\rm nm} < \lambda < 650$ nm; the fit offers a reasonable description for the sum of these basis functions, that is, further evidence that together they describe $a_{\rm ph}$.

Lastly, we compare the sum of the coefficients for the W_2 and W_4 basis functions ($H_2^{\rm L23} + H_4^{\rm L23}$) against the known $a_{\rm ph}(440)$ values of the L23 spectra. These quantities are tightly correlated (Fig. 7b; Kendall's tau p-value $< 10^{-5}$) and we conclude that together W_2 and W_4 describe variations in phytoplankton absorption across the global ocean.

Previous studies of $a_{\rm ph}$ have described its variation with Chl a concentration Chl a (Bricaud et al. 1995; Bricaud and Stramski 1990). The dominant modulation is the flattening in absorption at blue-green wavelengths ($\lambda \approx 420 - 525$ nm) as Chl a increases. Bricaud et al. (1995) and previous works attributed these variations to both the degree of pigment packaging and the covariation of additional pigments with Chl a. Furthermore, they introduced a functional form for $a_{\rm ph}$ which has a power-law dependence on Chl a, that is, $a_{\rm ph} \propto {\rm Chl}\ a^{\rm B}$. Our data-driven decomposition of the absorption coefficient has captured in W_2 and W_4 two modes of an additive model without any explicit dependence on Chl a. We offer further comparison between the NMF decomposition and the power-law expression of Bricaud et al. (1995) in "Modeling variations in $a_{\rm ph}$ with NMF" section, finding that the additive model statistically offers a better description of $a_{\rm ph}$.

Applications

In the previous section, we scientifically interpreted the four basis functions that successfully describe natural variability in the non-water absorption spectra of the global ocean. We now introduce several example applications that leverage these basis functions, the decompositions (i.e., coefficients) of the individual absorption spectra, and/or additional NMF analysis. We also comment on the limitations of the

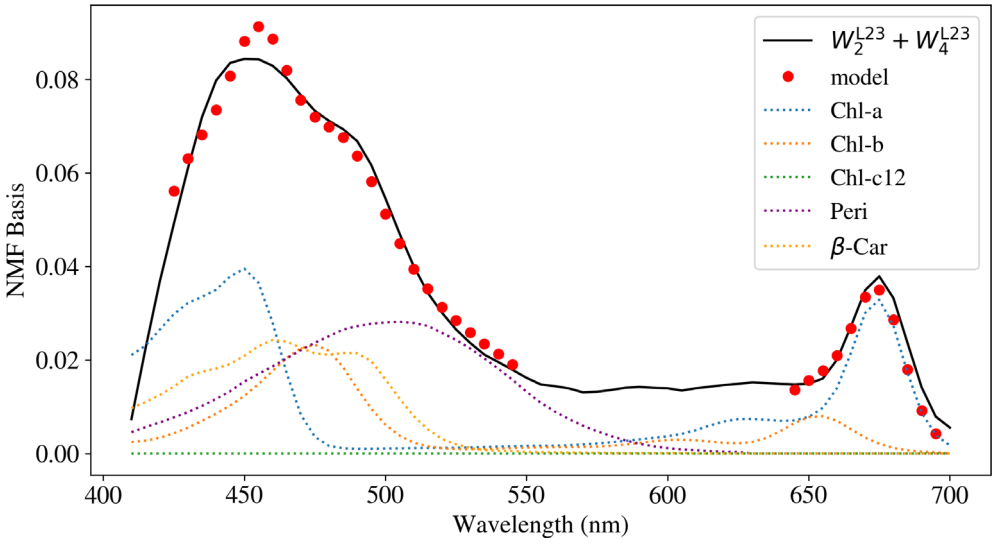

Fig. 8. Fit to the sum of the $W_2^{\rm L23}$ and $W_4^{\rm L23}$ basis functions derived from the L23 dataset. Here, we considered a linear sum of 5 pigments associated with phytoplankton as described in the legend. These can describe $W_2^{\rm L23}$ at the majority of wavelengths with the clear exception of $550 < \lambda < 650$ nm where additional pigments must contribute.

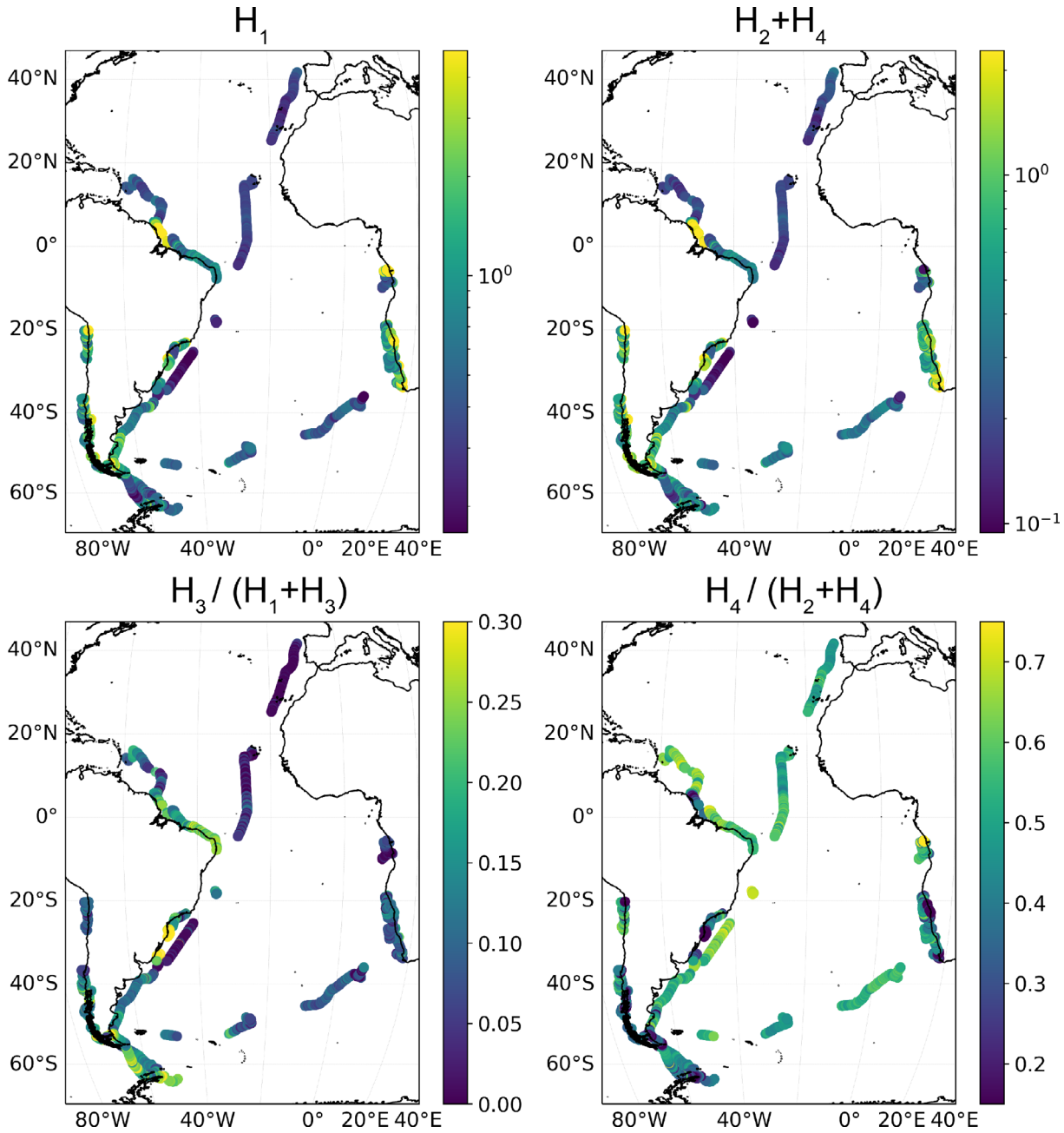

Fig. 9. Spatial distribution of the non-negative matrix factorization (NMF) decompositions (coefficients) for the *Tara* Microbiome dataset. Clockwise from the top left is $H_1^{\rm Tara}$ which primarily traces primarily non-algal particles, $H_2^{\rm Tara} + H_4^{\rm Tara}$ which assesses phytoplankton absorption, $H_4^{\rm Tara}/(H_2^{\rm Tara} + H_4^{\rm Tara})$ which describes phytoplankton spectra modulation, and $H_3^{\rm Tara}/(H_1^{\rm Tara} + H_3^{\rm Tara})$ which primarily emphasizes the NAP slope modulation. Note in the bottom row that higher values indicate more deviation from the base eigenvectors of H_1 and H_2.

NMF technique, especially in comparison to other approaches for the decomposition of absorption spectra.

Non-negative matrix factorization coefficients for particulate composition across regions

One application is to examine the values of the decompositions (i.e., H coefficients) from the individual absorption spectra as a function of geographic location to infer trends in CDOM, NAP, and phytoplankton. We note that a benefit here over other decompositions is that we are less likely to read into noisy and uncertain derived values and thus over-interpret a less than trustworthy biogeochemical signal. Figure 9 shows plots of the spatial distributions of the $H^{\rm Tara}$ coefficients from the *Tara* Microbiome expedition. The top panels describe absorption by NAP ($H_1^{\rm Tara}$) and phytoplankton ($H_2^{\rm Tara} + H_4^{\rm Tara}$), and we find generally expected patterns in NAP

and Chl *a* (both are higher in coastal regions relative to the open ocean and there is a strong correlation between the two).

The lower two panels of Fig. 9 examine modulations in the slope of detrital absorption $H_3^{\rm Tara}/(H_1^{\rm Tara}+H_3^{\rm Tara})$ and variations in the phytoplankton absorption $H_4^{\rm Tara}/(H_2^{\rm Tara}+H_4^{\rm Tara})$. We use normalized ratios here to emphasize the differences in the coefficients since generally most of the coefficients are correlated across space and time. As described in "W_3 represents the slope of CDOM and/or detritus absorption" section, the W_3 basis function appears to correspond to the shape of NAP absorption where an increase in H_3 results in flatter NAP spectra. Our expectation is broadly that a shallower NAP slope may indicate mineral dominated assemblages and a steeper slope indicates more organic assemblages (Babin et al. 2003). Examining Fig. 9, open ocean regions show very little contribution of the "slope flattening" (H_3 is low relative to H_1), matching our expectation for organic-dominated particle assemblages in oligotrophic regions. In some coastal regions, the $H_3^{\rm Tara}$ coefficients are $\sim 50\%$ of $H_1^{\rm Tara}+H_3^{\rm Tara}$ suggesting a much flatter slope and potentially higher mineral content in the particle population. Regions where this is observed include the Amazon River outflow, the Straits of Magellan, the Western Antarctic Peninsula, and near Florianópolis, Brazil during a known diatom bloom. In this last case of a diatom bloom, which is nearly a global max in the $H_3^{\rm Tara}/(H_1^{\rm Tara}+H_3^{\rm Tara})$ ratio, we observe distinct shoulders in the absorption spectra and speculate the higher value is due to pigments not well captured by the first four components and the higher H_3 is the best alternative to improve the fit and not necessarily related to a flatter NAP slope.

Now consider variations in phytoplankton absorption expressed by $H_4^{\rm Tara}/(H_2^{\rm Tara}+H_4^{\rm Tara})$, where we remind the reader that the W_2 basis function tracks absorption most related to Chl *a*, while W_4 is interpreted as modulating this to capture the primary variations in accessory pigments and pigment packaging effects. Viewing the spatial distribution of $H_4^{\rm Tara}/(H_2^{\rm Tara}+H_4^{\rm Tara})$ we find that the "accessory pigment indicator" $H_4^{\rm Tara}$ is primarily elevated offshore and it is generally negatively correlated with *Tara*LH though with substantial spread. This metric has a global peak near the mouth of the Congo River where we appear to have considerable CDOM, low salinity ($\sim$ 22 PSU), and possibly a set of phytoplankton pigments not well captured by the W_2 basis function.

Searching for outliers

An application that primarily makes use of the fundamental NMF basis functions is to identify non-water absorption spectra that are poorly modeled by these functions. In turn, these would represent outliers, that is, anomalous spectra that fall off the manifold defined by the large, global ocean datasets analyzed here. This can be performed on our training datasets (as done below) or any other hyperspectral samples with comparable wavelength coverage and sampling.

Figure 10 presents example absorption spectra from L23 and *Tara* whose NMF models have among the highest absolute and relative root-mean-square errors (RMSEs) when compared with the true spectra. In the evaluation of RMSE, we have ignored measurements with values less than 0.003 $\rm m^{-1}$ to minimize the influence of noise. The top panels in Fig. 10 show examples with large absolute RMSE, where the NMF model deviates by ≈ 0.01 $\rm m^{-1}$ at most wavelengths. In the *Tara* example, the data show a "shoulder" in the primary absorption peak at blue wavelengths suggestive of strong absorption by dinoflagellates or a monospecific bloom. A smaller feature may also be present in the L23 example.

The lower panels of Fig. 10 have among the highest relative RMSE in the datasets. The L23 has very weak non-water absorption and the spectrum is especially noisy. It is possible this outlier is primarily the result of poor data quality. The *Tara* example, meanwhile, is almost entirely modeled by the 1$^{\rm st}$ basis function (W_1) yet the spectrum is slightly steeper yielding the high relative RMSE.

We emphasize that the examples in Fig. 10 have among the highest RMSEs in the two datasets. Therefore, the relatively small differences between model and data in even these "outliers" demonstrate the near-universal representation of absorption provided by the NMF basis functions for the global ocean. We may anticipate, however, that very rare events (e.g., algal blooms in coastal waters) will show features not well captured by the NMF basis functions and notably higher RMSE values. We will explore this hypothesis in future work.

Inherent optical property retrievals using NMF basis functions

Another motivation of this work was to develop a new formalism for the parameterization of absorption spectra at hyperspectral resolution in anticipation of the Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission. Plankton, Aerosol, Cloud, ocean Ecosystem mission is providing the first global, nearly daily cadence dataset of hyperspectral remote-sensing reflectances $R_{\rm rs}$. A standard practice is to retrieve from these $R_{\rm rs}$ data estimates for the absorption and backscattering IOPs of the water (e.g., Werdell et al. 2013). These inference algorithms for IOP retrievals require a parameterization of $a(\lambda)$ and the backscattering coefficient spectra $b_b(\lambda)$ (e.g., Mobley 2022). Ideally, for the parameterization of IOP retrievals one adopts an approach that maximizes scientific description while minimizing the number of free parameters.

We have argued that the $m_{\rm NMF}=4$ NMF decomposition presented here provides just such a prescription for absorption spectra. Furthermore, the NMF analysis is easily extended to include absorption in more turbid conditions, for example, coastal and/or inland waters. However, we may advocate for IOP retrievals that one consider a hybrid combination of the NMF decomposition with a power-law or exponential model of CDOM and/or detritus as frequently implemented in other

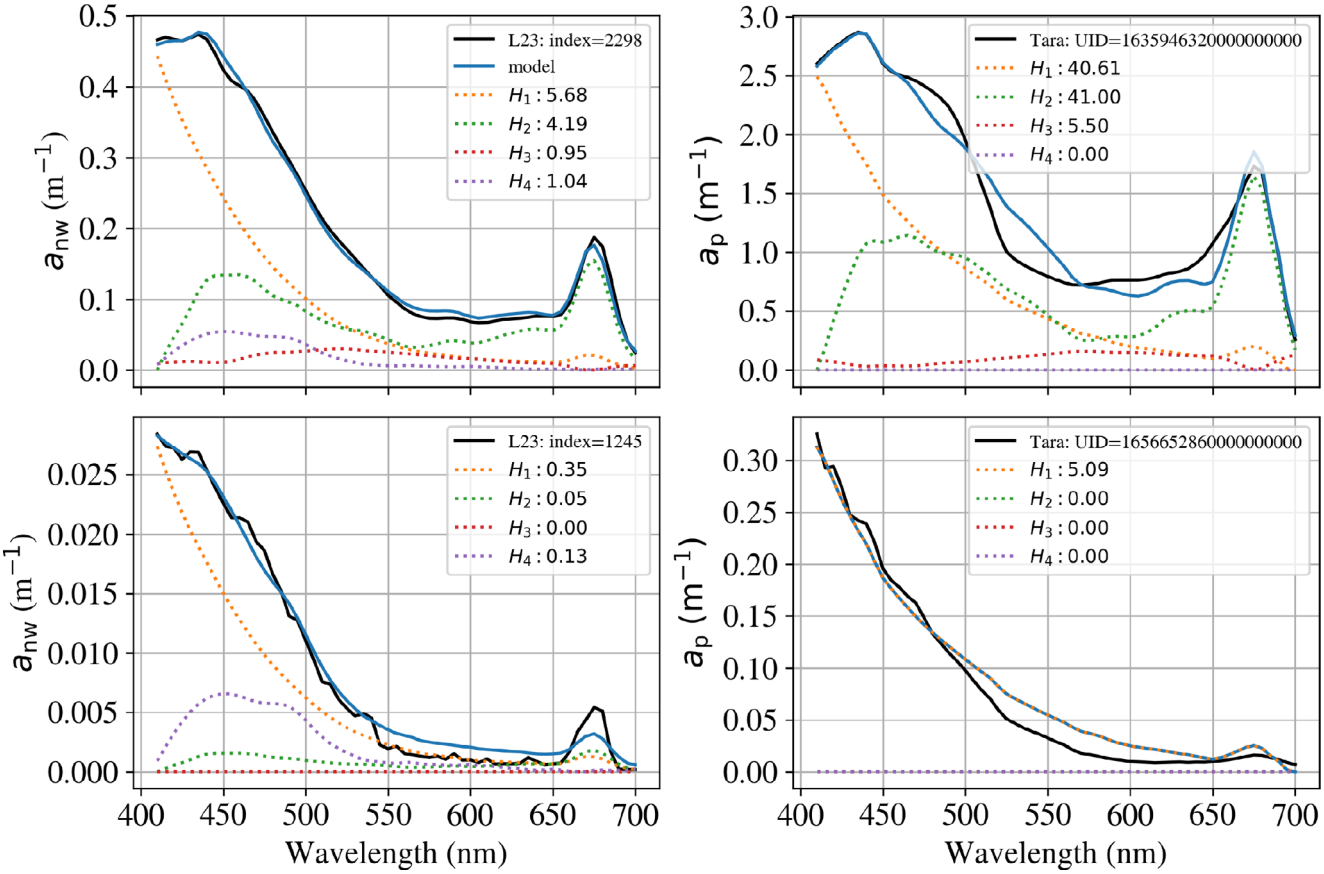

Fig. 10. The top two panels show spectra from (left) L23 and (right) *Tara* with among the highest absolute root-mean-square error (RMSE) between the data (black) and non-negative matrix factorization (NMF) fit (blue). The *Tara* spectra exhibit significant absorption at $\lambda \approx 475$ nm that may result from nearly monospecific bloom conditions. This rare feature is not captured by the four NMF basis functions, that is, this is an outlier. The L23 spectrum exhibits a similar departure, albeit weaker. The bottom two panels show examples for two of the highest relative RMSE. The L23 (bottom left) spectrum is likely corrupted by noise, that is, measurement error. The Tara (bottom right) spectrum shows NAP absorption that is steeper than the W_1 derived from the *Tara* dataset, again a notable outlier.

algorithms (e.g., Garver and Siegel 1997). Such an approach will be advantageous if the underlying CDOM/NAP absorption is truly exponential or very close to it.

Modeling variations in a_{ph} with NMF

Now consider an application of the NMF technique to phytoplankton absorption alone. As described in "W_2 and W_4 represent absorption by phytoplankton" section, the NMF decomposition introduced in this paper includes two basis functions (W_2, W_4) that capture variations in phytoplankton absorption a_{ph}. Such variations stem from (i) differences in the pigment packaging effect; and (ii) the inclusion/omission of pigments and specific phytoplankton families. These variances in a_{ph} in ocean water were recognized decades ago (Bricaud and Stramski 1990; Mitchell and Kieper 1988), and Bricaud et al. (1995) introduced a model to describe the average variation:

$$a^*_{ph} = A(\lambda)\text{Chl}\ a^{B(\lambda)} \quad (8)$$

with $a^*_{ph} \equiv a_{ph}/\text{Chl}\ a$ and Chl a the concentration of Chl a and A and B the wavelength-dependent coefficients. In their analysis, Bricaud et al. (1995) reported that this nonlinear model offered the best fit ("highest determination coefficient") to the $\sim$800 a_{ph} spectra obtained on six cruises.

As an alternative to the Bricaud formalism, we performed an $m_{NMF} = 2$ NMF decomposition of the a_{ph} absorption spectra provided in the L23 dataset. The resultant basis functions (termed W^{ph}) are presented in Supporting Information Fig. S2. The two basis functions closely resemble W_2 and W_4 derived from the total, non-water absorption spectra (Fig. 4), which confirms their primary role in describing a_{ph}. Examining Fig. S2, we identify the 1st basis function W^{ph}_1 as the pigments that most tightly vary with Chl a, including absorption at $\lambda \approx 500-650$ nm which resembles prymnesiophytes (e.g., coccolithophores and other golden-brown flagellates). Conversely, the 2nd basis function W^{ph}_2 captures pigments with dominant absorption at blue wavelengths (e.g., picoplankton) and/or packaging effects.

Now compare the two-component NMF model with the standard Bricaud formalism. Figure S3, meanwhile, presents two sets of fits to a_{ph}: one with the Bricaud function (Eq. 8) updated by Bricaud et al. (1998), which is a one-parameter fit (Chl *a*), and the other using our $m_{NMF} = 2$ NMF decomposition (Figure S3). Clearly, the NMF model yields a better description of a_{ph} at nearly all wavelengths.

Building on the examples shown in Fig. S3 of the Supplemental Information, we have evaluated the root-mean-square error (RMSE) between the a_{ph} data of L23 and (i) the two-component NMF model for a_{ph} and (ii) the Bricaud model. At all $a_{ph}(440)$, and therefore Chl *a*, the NMF models have lower RMSE (the median ratio is 2.5). Therefore, we find that a relatively simple, two-component additive (i.e., linear) model for a_{ph} provides a more accurate description of these data than the one-parameter, nonlinear function that is frequently adopted. It would be straightforward to implement in IOP retrievals, and the coefficients directly describe variations in a_{ph} absorption.

Limitations

It is also important to recognize the limitations of the NMF decompositions. While the four basis functions presented here successfully reproduce the variations in large datasets of absorption spectra and with greater interpretability than other statistical techniques (e.g., PCA), the basis functions are unlikely to map directly to key quantities of interest of the ocean color community, for example, Chl *a* concentration. That is, to the extent that one wishes to estimate the concentration of CDOM or search for absorption from a particular phytoplankton species, one would need to calibrate the NMF decompositions against a dataset of such measurements (e.g., Chl *a*, as described in "W_2 and W_4 represent absorption by phytoplankton" section).

Indeed, applications that have unique needs (e.g., identifying a specific algal bloom) may benefit from a more mechanistic approach (e.g., Anderson et al. 2011). Such techniques seek to isolate specific, scientifically desired measures, for example, phytoplankton size distributions (Zhang et al. 2015), while ignoring other aspects of the data. Other methods aim to separate the total absorption into bulk components, for example, phytoplankton, CDOM, and NAP (e.g., Lin et al. 2013; Stramski et al. 2019). Mechanistic approaches have the added benefit of allowing one to perform analyses with empirical motivated basis functions, derived from direct measurements of known constituents from in situ or laboratory samples. Examples include an exponential function for CDOM or adopting absorption spectra of specific phytoplankton pigments (e.g., A. P. Chase et al. 2017).

On the other hand, mechanistic models have their own drawbacks. If designed for a specific application, they may not be sufficiently generic to describe the full variability in a given dataset. In this case, techniques like χ^2 minimization may drive solutions to erroneous values and/or generate bias. Or, if the mechanistic model does have high complexity, it may exhibit undesirable correlations between fitted parameters. Indeed, from the results presented here (and *see* Cael et al. 2020b; J. X. Prochaska and R. Frouin, submitted), we contend that any model of non-water absorption can only recover four (or five) distinct parameters and even these are apt to show correlations.

Another limitation of the NMF methodology stems from its core strength. The high interpretability of NMF is due in part from the fact that it is a linear, additive model. For absorption spectra, higher values of any given coefficient implies greater contributions from the components described by its basis function (e.g., Chl *a* for the 2nd basis function W_2 presented here). However, for features that have nonlinear variability, for example, the slope of the near-exponential behavior of CDOM or detritus, the NMF approach will not optimally capture the variations. Here, we found differences in the slope for CDOM was instead primarily accounted for by the 3rd basis function (W_3, "W_3 represents the slope of CDOM and/or detritus absorption" section).

Conclusions

We have presented a NMF of absorption coefficients $a(\lambda)$ drawn from two large, methodologically independent datasets spanning the coastal and open ocean. The decomposition of each into four non-negative basis functions describes over 99.9% of the data variance. The basis functions derived from each dataset have remarkable similarity despite significant differences in the methodology and sampling strategies (e.g., CDOM absorption is absent in the *Tara* data). We argue, therefore, that these basis functions reveal the fundamental modes of ocean color absorption.

We showed that two of the modes describe the amplitude and slope of CDOM and/or detrital absorption a_{dg}, while the other two provide an additive, linear breakdown of phytoplankton absorption a_{ph}. We further demonstrate that this a_{ph} decomposition offers a better (lower RMSE) model of the variance in phytoplankton absorption due to packaging and/or pigment variations than previous, nonlinear models. Applications of the NMF method include examining geographical trends in the features expressed by the basis functions (e.g., detrital absorption, phytoplankton packaging) and implementing the functions in IOP retrieval algorithms.

We have also discussed the limitations of the NMF methodology, in particular that one must perform additional analysis to relate the NMF decompositions to specific physical quantities (e.g., Chl *a* concentration). On the other hand, with these NMF basis functions one may perform data-driven exploration of correlative features (e.g., pigment groupings; Kramer and Siegel 2019) that may then be related to key aquatic biogeochemical processes.

Future work may develop NMF decompositions for other bodies of water and/or other non-negative IOPs

(e.g., backscattering) or apparent optical properties (e.g., reflectances). As with the absorption spectra, we anticipate these will provide a highly interpretable, compact description of the primary features in ocean color observations.

Author Contributions

J. Xavier Prochaska formulated the idea, led the analysis, and led the writing. Patrick Gray provided the Tara data, offered ocean optics expertise, and participated in writing and editing.

Acknowledgments

J. Xavier Prochaska thanks Dariusz Stramski and Rick Reynolds at Scripps Institution of Oceanography for their input, especially on the L23 dataset. We thank Norm Nelson for input on interpreting the basis functions that describe phytoplankton and Heidi Dierssen for helpful comments. We particularly thank Guillaume Bourdin for his effort processing, quality controlling, and openly sharing the Tara Microbiome dataset in advance of publication. We acknowledge support from the Zuckerman STEM Leadership Program to Patrick Gray. The optical inline dataset was collected and analyzed with support from NASA Ocean Biology and Biogeochemistry program under grants NNX13AE58G and NNX15AC08G, 80NSSC21K0783, and 80NSSC20K1641 to the University of Maine. We wish to thank the Tara Ocean Foundation, the SV Tara crew, and all those who participate in Mission Microbiomes AtlantECO and adopt its Data Sharing and Publication Best Practices (https://zenodo.org/communities/mission-microbiomes-atlanteco/). This publication has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 862923 (project AtlantECO). This output reflects only the author's view and the European Union cannot be held responsible for any use that may be made of the information contained therein. We are keen to thank the commitment of the following institutions for their financial and scientific support that made Mission Microbiomes AtlantECO possible: Stazione Zoologica Anton Dohrn, European Bioinformatics Institute (EMBL-EBI), Centre national de la recherche scientifique (CNRS), Centre National de Séquençage (CNS, Genoscope), agnès b., BIC, Capgemini Engineering, Fondation Groupe EDF, Compagnie Nationale du Rhône, L'Oréal, Biotherm, Région Bretagne, Lorient Agglomération, Billerudkorsnas, Havas Paris, Fondation Rothschild, Office Français de la Biodiversité, AmerisourceBergen, Philgood Foundation, UNESCO-IOC, Etienne Bourgois. J. Xavier Prochaska acknowledges support from a Simons Pivot Fellowship.

Conflicts of Interest

None declared.

References

Anderson, C. R., R. M. Kudela, C. Benitez-Nelson, et al. 2011. "Detecting Toxic Diatom Blooms From Ocean Color and a Regional Ocean Model." *Geophysical Research Letters* 38, no. 4: L04603. https://doi.org/10.1029/2010GL045858.

Babin, M., D. Stramski, G. M. Ferrari, et al. 2003. "Variations in the Light Absorption Coefficients of Phytoplankton, Nonalgal Particles, and Dissolved Organic Matter in Coastal Waters Around Europe." *Journal of Geophysical Research: Oceans* 108, no. C7: 321–13231. https://doi.org/10.1029/2001JC000882.

Boss, E., M. S. Twardowski, and S. Herring. 2001. "The Shape of the Beam Attenuation Spectrum and Its Relation to the Size Distribution of Oceanic Particles." *Applied Optics* 40: 4885–4893. https://doi.org/10.1364/AO.40.004885.

Bricaud, A., M. Babin, A. Morel, and H. Claustre. 1995. "Variability in the Chlorophyll-Specific Absorption Coefficients of Natural Phytoplankton: Analysis and Parameterization." *Journal of Geophysical Research* 100: 13321–13332. https://doi.org/10.1029/95JC00463.

Bricaud, A., H. Claustre, J. Ras, and K. Oubelkheir. 2004. "Natural Variability of Phytoplanktonic Absorption in Oceanic Waters: Influence of the Size Structure of Algal Populations." *Journal of Geophysical Research: Oceans* 109, no. C11: 1–23. https://doi.org/10.1029/2004JC002419.

Bricaud, A., A. Morel, M. Babin, K. Allali, and H. Claustre. 1998. "Variations of Light Absorption by Suspended Particles With Chlorophyll a Concentration in Oceanic (Case 1) Waters: Analysis and Implications for Bio-Optical Models." *Journal of Geophysical Research* 103: 31033. https://doi.org/10.1029/98JC02712.

Bricaud, A., and D. Stramski. 1990. "Spectral Absorption Coefficients of Living Phytoplankton and Nonalgal Biogenous Matter: A Comparison Between the Peru Upwelling Area and the Sargasso Sea." *Limnology and Oceanography* 35: 562. https://doi.org/10.4319/lo.1990.35.3.0562.

Cael, B. B., A. Chase, and E. Boss. 2020. "Information Content of Absorption Spectra and Implications for Ocean Color Inversion." *Applied Optics* 59: 3971. https://doi.org/10.1364/AO.389189.

Carder, K. L., R. G. Steward, G. R. Harvey, and P. B. Ortner. 1989. "Marine Humic and Fulvic Acids: Their Effects on Remote Sensing of Ocean Chlorophyll." *Limnology and Oceanography* 34: 68–81. https://doi.org/10.4319/lo.1989.34.1.0068.

Chase, A., E. Boss, R. Zaneveld, et al. 2013. "Decomposition of In Situ Particulate Absorption Spectra." *Methods in Oceanography* 7: 110. https://doi.org/10.1016/j.mio.2014.02.002.

Chase, A. P., E. Boss, I. Cetinic, and W. Slade. 2017. "Estimation of Phytoplankton Accessory Pigments From Hyperspectral Reflectance Spectra: Toward a Global Algorithm." *Journal of Geophysical Research, Oceans* 122: 9725. https://doi.org/10.1002/2017JC012859.

Garver, S. A., and D. A. Siegel. 1997. "Inherent Optical Property Inversion of Ocean Color Spectra and Its Biogeochemical Interpretation: 1. Time series From the Sargasso Sea." *Journal of Geophysical Research* 102: 18–607. https://doi.org/10.1029/96JC03243.

Garver, S. A., Siegel, D. A., & B. Greg, M. 1994, Variability in Near-Surface Particulate Absorption Spectra: What Can a Satellite Ocean Color Imager See? *Limnology and Oceanography*, 39, 1349. https://doi.org/10.4319/lo.1994.39.6.1349

Gray, P. C., E. Boss, G. Bourdin, et al. 2025. "Emergent Patterns of Patchiness Differ between Physical and Planktonic Properties in the Ocean." *Nature Communications* 16: 1808. https://doi.org/10.1038/s41467-025-56794-x.

IOCCG Protocol Series. 2008. Why Ocean Colour? The Societal Benefits of Ocean-Colour Technology, Ocean Optics and Biogeochemistry Protocols for Satellite Ocean Colour Sensor Validation. Dartmouth, NS, Canada: IOCCG.

IOCCG Protocol Series. 2018. Inherent Optical Property Measurements and Protocols: Absorption Coefficient, Ocean Optics and Biogeochemistry Protocols for Satellite Ocean Colour Sensor Validation. Dartmouth, NS, Canada: IOCCG. https://doi.org/10.25607/OBP-119.

Ismail, K. A., and M. R. Al-Shehhi. 2023. "Applications of Biogeochemical Models in Different Marine Environments: A Review." *Frontiers in Environmental Science* 11: 126. https://doi.org/10.3389/fenvs.2023.1198856.

Iturriaga, R., and D. A. Siegel. 1989. "Microphotometric Characterization of Phytoplankton and Detrital Absorption Properties in the Sargasso Sea." *Limnology and Oceanography* 34: 1706. https://doi.org/10.4319/lo.1989.34.8.1706.

Jolliffe, I. T., and J. Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374: 20150202. https://doi.org/10.1098/rsta.2015.0202.

Jordan, T. M., G. Dall'Olmo, G. Tilstone, et al. 2024. "A Compilation of Surface Inherent Optical Properties and Phytoplankton Pigment Concentrations From the Atlantic Meridional Transect." *Earth System Science Data Discussions* 2024: 1. https://doi.org/10.5194/essd-2024-267.

Kehrli, M. D., D. Stramski, R. A. Reynolds, and I. D. Joshi. 2023. "Estimation of Chromophoric Dissolved Organic Matter and Non-algal Particulate Absorption Coefficients of Seawater in the Ultraviolet by Extrapolation From the Visible Spectral Region." *Optics Express* 31: 17450. https://doi.org/10.1364/OE.486354.

Kehrli, M. D., D. Stramski, R. A. Reynolds, and I. D. Joshi. 2024. "Model for Partitioning the Non-Phytoplankton Absorption Coefficient of Seawater in the Ultraviolet and Visible Spectral Range Into the Contributions of Non-algal Particulate and Dissolved Organic Matter." *Applied Optics* 63: 4252. https://doi.org/10.1364/AO.517706.

Kramer, S. J., and D. A. Siegel. 2019. "How Can Phytoplankton Pigments Be best Used to Characterize Surface Ocean Phytoplankton Groups for Ocean Color Remote Sensing Algorithms?" *Journal of Geophysical Research, Oceans* 124: 7557–7574. https://doi.org/10.1029/2019JC015604.

Lawton, W. H., and E. A. Sylvestre. 1971. "Self Modeling Curve Resolution." *Technometrics* 13: 617. https://doi.org/10.1080/00401706.1971.10488823.

Lee, D. D., and H. S. Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401: 788. https://doi.org/10.1038/44565.

Lin, J., W. Cao, G. Wang, and S. Hu. 2013. "Approach for Determining the Contributions of Phytoplankton, Colored Organic Material, and Nonalgal Particles to the Total Spectral Absorption in Marine Waters." *Applied Optics* 52: 4249. https://doi.org/10.1364/AO.52.004249.

Litchman, E., P. de Tezanos Pinto, K. F. Edwards, C. A. Klausmeier, C. T. Kremer, and M. K. Thomas. 2015. "Global Biogeochemical Impacts of Phytoplankton: A Trait-Based Perspective." *Journal of Ecology* 103: 1384. https://doi.org/10.1111/1365-2745.12438.

Loisel, H., D. Schaffer Ferreira Jorge, R. A. Reynolds, and D. Stramski. 2023. "A Synthetic Optical Database Generated by Radiative Transfer Simulations in Support of Studies in Ocean Optics and Optical Remote Sensing of the Global Ocean." *Earth System Science Data* 15: 3711. https://doi.org/10.5194/essd-15-3711-2023.

Loisel, H., D. Stramski, D. Dessailly, C. Jamet, L. Li, and R. A. Reynolds. 2018. "An Inverse Model for Estimating the Optical Absorption and Backscattering Coefficients of Seawater From Remote-Sensing Reflectance Over a Broad Range of Oceanic and Coastal Marine Environments." *Journal of Geophysical Research, Oceans* 123: 2141. https://doi.org/10.1002/2017JC013632.

Lomas, M. W., A. R. Neeley, R. Vandermeulen, et al. 2024. "Phytoplankton Optical Fingerprint Libraries for Development of Phytoplankton Ocean Color Satellite Products." *Scientific Data* 11: 168. https://doi.org/10.1038/s41597-024-03001-z.

McClain, C. R. 2009. "A Decade of Satellite Ocean Color Observations." *Annual Review of Marine Science* 1: 19–42. https://doi.org/10.1146/annurev.marine.010908.163650.

Mitchell, G. B., and D. A. Kieper. 1988. "Deep Sea Research Part A. Oceanographic Research Papers." 35: 665. https://doi.org/10.1016/0198-0149(88)90025-8.

Mobley, C. D., ed. 2022. The Oceanic Optics Book, 924. Dartmouth, NS, Canada: International Ocean Colour Coordinating Group (IOCCG). https://doi.org/10.25607/OBP-1710.

Morel, A. 1988. "Optical Modeling of the Upper Ocean in Relation to Its Biogenous Matter Content (Case I Waters)." *Journal of Geophysical Research* 93, no. 10: 10749. https://doi.org/10.1029/JC093iC09p10749.

Paatero, P., and U. Tapper. 1994. "Positive Matrix Factorization: A Non-Negative Factor Model With Optimal Utilization of Error Estimates of Data Values." *Environmetrics* 5: 111. https://doi.org/10.1002/env.3170050203.

Ren, B., L. Pueyo, G. B. Zhu, J. Debes, and G. Duchêne. 2018. "Non-Negative Matrix Factorization: Robust Extraction of Extended Structures." *Astrophysical Journal* 852: 104. https://doi.org/10.3847/1538-4357/aaa1f2.

Roesler, C. S., and A. H. Barnard. 2013. "Optical Proxy for Phytoplankton Biomass in the Absence of Photophysiology: Rethinking the Absorption Line Height." *Methods in Oceanography* 7: 79. https://doi.org/10.1016/j.mio.2013.12.003.

Slade, W. H., E. Boss, G. Dall'Olmo, et al. 2010. "Underway and Moored Methods for Improving Accuracy in Measurement of Spectral Particulate Absorption and Attenuation." *Journal of Atmospheric and Oceanic Technology* 27: 1733. https://doi.org/10.1175/2010JTECHO755.1.

Stramski, D., A. Bricaud, and A. Morel. 2001. "Modeling the Inherent Optical Properties of the Ocean Based on the Detailed Composition of the Planktonic Community." *Applied Optics* 40: 2929. https://doi.org/10.1364/AO.40.002929.

Stramski, D., L. Li, and R. A. Reynolds. 2019. "Model for Separating the Contributions of Non-Algal Particles and Colored Dissolved Organic Matter to Light Absorption by Seawater." *Applied Optics* 58: 3790. https://doi.org/10.1364/AO.58.003790.

Taslaman, L., and B. Nilsson. 2012. "A Framework for Regularized Non-negative Matrix Factorization, With Application to the Analysis of Gene Expression Data." *PLoS One* 7: e46331. https://doi.org/10.1371/journal.pone.0046331.

Valente, A., S. Sathyendranath, V. Brotas, et al. 2022. "A Compilation of Global Bio-Optical In Situ Data for Ocean Colour Satellite Applications – Version Three." *Earth System Science Data* 14: 5737. https://doi.org/10.5194/essd-14-5737-2022.

Werdell, P. J., B. A. Franz, S. W. Bailey, et al. 2013. "Generalized Ocean Color Inversion Model for Retrieving Marine Inherent Optical Properties." *Applied Optics* 52: 2019. https://doi.org/10.1364/AO.52.002019.

Zhang, X., Y. Huot, A. Bricaud, and H. M. Sosik. 2015. "Inversion of Spectral Absorption Coefficients to Infer Phytoplankton Size Classes, Chlorophyll Concentration, and Detrital Matter." *Applied Optics* 54: 5805. https://doi.org/10.1364/AO.54.005805.

Zhu, G. B. 2023. "Guangtunbenzhu/NonnegMFPy: Zenodo (Version 20231231)." *Zenodo*. https://doi.org/10.5281/zenodo.10447678.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Submitted 15 June 2024
Revised 27 December 2024
Accepted 16 May 2025