2019

Deriving inherent optical properties from decomposition of hyperspectral non-water absorption

Brice K. Grunert

Colleen B. Mouw
University of Rhode Island, cmouw@uri.edu

See next page for additional authors

This is a pre-publication author manuscript of the final, published article.

Citation/Publisher Attribution
Grunert, B. K., Mouw, C. B., Ciochett, A. B. (2019). Deriving inherent optical properties from decomposition of hyperspectral non-water absorption. *Remote Sensing of Environment*, *225*, 193-206 .doi: 10.1016/j.rse.2019.03.004

Available at: https://doi.org/10.1016/j.rse.2019.03.004

Authors
Brice K. Grunert, Colleen B. Mouw, and Audrey B. Ciochetto

Deriving inherent optical properties from decomposition of hyperspectral non-water absorption

Brice K. Grunert[1*], Colleen B. Mouw[2], Audrey B. Ciochetto[2]

[1]Michigan Technological University, Department of Geological and Mining Engineering and Sciences, 1400 Townsend Drive, Houghton, MI 49931, USA

[2] University of Rhode Island, Graduate School of Oceanography, 215 South Ferry Road, Narragansett, RI 02882, USA

[*]Corresponding author: bgrunert@mtu.edu, +1 414-322-7506

Keywords: hyperspectral, PACE, inherent optical properties, colored dissolved organic matter, phytoplankton, ocean biogeochemistry

Abstract

Semi-analytical algorithms (SAAs) developed for multispectral ocean color sensors have benefited from a variety of approaches for retrieving the magnitude and spectral shape of inherent optical properties (IOPs). SAAs generally follow two approaches: 1) simultaneous retrieval of all IOPs, resulting in pre-defined bio-optical models and spectral dependence between IOPs and 2) retrieval of bulk IOPs (absorption and backscattering) first followed by decomposition into separate components, allowing for independent retrievals of some components. Current algorithms used to decompose hyperspectral remotely-sensed reflectance into IOPs follow the first strategy. Here, a spectral deconvolution algorithm for incorporation into the second strategy is presented that decomposes $a_{t-w}(\lambda)$ from *in situ* measurements and estimates absorption due to phytoplankton ($a_{ph}(\lambda)$) and colored detrital material ($a_{dg}(\lambda)$) free of explicit assumptions. The algorithm described here, Derivative Analysis and Iterative Spectral Evaluation of Absorption (DAISEA), provides estimates of $a_{ph}(\lambda)$ and $a_{dg}(\lambda)$ over a spectral range from 350-700 nm. Estimated $a_{ph}(\lambda)$ and $a_{dg}(\lambda)$ showed an average normalized root mean square difference of <30% and <20%, respectively, from 350-650 nm for the majority of optically distinct environments considered. Estimated S_{dg} median difference was less than 20% for all environments considered, while distribution of S_{dg} uncertainty suggests that biogeochemical variability represented by S_{dg} can be estimated free of bias. DAISEA results suggest that hyperspectral satellite ocean color data will improve our ability to track biogeochemical processes affiliated with variability in $a_{dg}(\lambda)$ and S_{dg} free of explicit assumptions.

1. Introduction

Dissolved organic matter (DOM) comprises the largest pool of fixed carbon in the ocean, roughly equivalent to the reservoir of atmospheric CO_2 (~670 Pg C; Hansell et al. 2009; Ogawa et al. 2001). Yet, sources and cycling of DOM in the global ocean remain poorly constrained due to difficulty in assigning origin and tracking changes in composition to a complex mixture of organic compounds composed of up to ~20,000 molecular formulas in a sample (Andrew et al. 2013; Mentges et al. 2017; Riedel and Dittmar 2014). A portion of DOM is optically active, colored dissolved organic matter (CDOM), and displays distinct spectral variability between uniquely sourced material, namely terrestrial and marine-derived, and different degradation pathways, such as microbial or photodegradation (Catalá et al. 2016; Danhiez et al. 2017; Helms et al. 2013; Helms et al. 2008; Zhao et al. 2017). Due to its interaction with light, CDOM can be rapidly characterized using optical sensors and is observable from autonomous and satellite platforms (e.g., Siegel et al. 2005; Xing et al. 2012). These observations are crucial to adequately model ocean biogeochemical and physical processes due to the influence of CDOM on distribution and spectral quality of light in the water column and heating of the surface ocean (Chang and Dickey 2004; Dutkiewicz et al. 2015; Kim et al. 2016).

CDOM absorption ($a_g(\lambda)$, m^{-1}; λ denotes wavelength) at visible wavelengths also tracks the spectral shape of a_g (S_g) and dissolved organic carbon concentration ([DOC], $mg{\cdot}L^{-1}$) in coastal waters where a strong gradient of relatively degraded, terrestrial-derived material and conservative mixing produce a clear, observable signal across unique pools of CDOM (Cory and Kling 2018; Fichot and Benner 2011; Mannino et al. 2014; Stedmon and Markager 2003). This continuous dilution of $a_g(\lambda)$ in coastal waters presents predictive capability of CDOM molecular weight, degradation state and terrestrial biomarkers (e.g., lignin) using $a_g(\lambda)$ due to unique spectral features present in terrestrial material relative to CDOM of marine origin (Fichot et al. 2016; Fichot et al. 2013; Helms et al. 2008; Vantrepotte et al. 2015). While these relationships are strong in coastal waters, open ocean waters do not display a consistent relationship between $a_g(\lambda)$, S_g and [DOC] due to relatively low production rates and strong photodegradation in surface ocean waters (Helms et al. 2013; Nelson et al. 2010). This disconnect between single wavelength estimates of $a_g(\lambda)$ and S_g currently limits our ability to accurately track production and degradation of CDOM across broad spatial scales while also introducing significant uncertainty in estimates of $a_{ph}(\lambda)$ and derived products (e.g., chlorophyll-a concentration). Additionally, increasing observations of $a_g(\lambda)$ have shown that S_g displays significant variability and is capable of characterizing CDOM of unique source, environmental conditions and degradation state (Asmala et al. 2018; Danhiez et al. 2017; Grunert et al. 2018; Helms et al. 2008, 2013). Considering this, it is likely that this parameter contains very useful information regarding food web processes and marine carbon cycling relevant to understanding the balance of the marine DOM carbon reservoir.

Hyperspectral ocean color observations from *in situ* measurements including flow-through systems and proposed satellite sensors such as the German Aerospace Center's Environmental Mapping and Analysis Program sensor and NASA's Plankton, Aerosol, Cloud and ocean Ecosystem (PACE) sensor provide the potential to observe inherent optical properties (IOP's), including phytoplankton absorption ($a_{ph}(\lambda)$, m^{-1}), non-algal particulate (NAP) absorption ($a_d(\lambda)$, m^{-1}) and $a_g(\lambda)$, with greater accuracy across the global ocean. Hyperspectral satellite observations have the proven ability to characterize unique phytoplankton functional groups (Bracher et al. 2009; Sadeghi et al. 2012) while flow-through systems have provided an unprecedented view of phytoplankton productivity and physiology at a global scale (Chase et al. 2013; Werdell et al. 2013). Additional work including derivative analysis has also shown potential for estimating pigment concentrations, $a_g(\lambda)$, S_g, $a_d(\lambda)$ and the spectral shape of $a_d(\lambda)$ (S_d; Wang et al. 2016; Chase et al. 2017; Vandermeulen et al. 2017; Wang et al. 2018). To date, satellite algorithms use an assumed value or starting point for $S_{dg,}$ the combined spectral slope term for $a_d(\lambda)$ and $a_g(\lambda)$, based on global or regional observations and/or constrain solutions within a pre-defined space (Lee et al. 2002; Werdell et al. 2013; Dong et al. 2013; Zhang et al. 2015). These approaches are all made possible by a variety of existing inversion approaches developed for multispectral data outlined by Werdell et al. (2018).

Hyperspectral approaches are still scarce but apply bottom-up strategies on *in situ* $R_{rs}(\lambda)$ capable of estimating pigment concentrations and separating $a_g(\lambda)/S_g$ and $a_d(\lambda)/S_d$ using assumed starting points and lower/upper bounds on variables. Bottom-up strategies provide accurate solutions but result in IOP retrievals that are spectrally dependent on each other (Mouw et al. 2015). Here, we provide a top-down approach that independently estimates S_{dg}, $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ free of explicit assumptions from total non-water absorption ($a_{t\text{-}w}(\lambda)$) using derivative analysis, iterative spectral evaluation and Gaussian decomposition of total non-water absorption spectra. Beyond estimation of S_{dg} and more accurate spectral retrievals of $a_{ph}(\lambda)$, such a method provides clearer spectral features for the derivation of phytoplankton functional types, including Gaussian fitting and second or fourth derivative analysis of phytoplankton pigments (Chase et al. 2017; Vandermeulen et al. 2017; Wang et al. 2017). We focus on accurate retrieval of S_{dg} and $a_{dg}(\lambda)$ to represent biogeochemical variability in NAP and CDOM absorption represented by the spectral shape and magnitude of $a_{dg}(\lambda)$. Our results suggest the algorithm, Derivative Analysis and Iterative Spectral Evaluation of Absorption (DAISEA), will work well with future top-down hyperspectral inversion approaches.

2. Methods

2.1 Data

In situ data were accessed from NASA's SeaWiFS Bio-optical Archive and Storage System (SeaBASS, https://seabass.gsfc.nasa.gov/) on January 12, 2018 (Werdell et al. 2003). We focused our collection on data where $a_{ph}(\lambda)$, $a_d(\lambda)$ and $a_g(\lambda)$ were all measured coincidentally on a benchtop spectrophotometer within 10 m of the surface (Fig. 1). We initially quality controlled each set of absorption spectra by considering if any values were below zero for individual spectra. If the minimum value was more negative than -0.1, the spectra was discarded; if the value was greater than -0.1, an offset for the most negative value was applied to the entire spectrum. In doing so, spectral shape was retained while removing poorly defined absorption values that resulted in negative algorithm solutions. We removed any spectra where S_{dg} was less than 0.004 nm^{-1}, values unrealistic with historic observations and estimates (e.g., Siegel et al. 2002; Wang et al. 2005). Additionally, spectra that had been sampled at a resolution less than 2 nm were not considered to ensure spectral shape was maintained when downsampling. After removing poor quality spectra, a total of 4,787 spectra remained. These spectra were randomly split into training (n=3,434; Fig. 1a) and test datasets (n=1,353; Fig. 1b) so that training spectra accounted for ~75% of total spectra. All absorption spectra were subsampled to 5 nm either through direct sub-sampling or linear interpolation to avoid introducing artificial curvature, with the spectral range from 350-700 nm used (71 data points). Some spectra were not sampled down to exactly 350 nm but were measured at or below 355 nm (e.g., 350.7; n=79); for these spectra, we extrapolated to 350 nm using a discretized partial differential equation with an enhanced plate metaphor (D'Errico 2005). Typical uncertainty estimates for spectrophotometer measurements, assessed as differences among triplicate samples, ranged from ~5-10% relative difference (Mouw, unpublished data). We focused on 5 nm spectral resolution here for an assessment of performance relative to the anticipated resolution of PACE.

2.2 DAISEA Algorithm Development

Our approach for decomposing $a_{t\text{-}w}(\lambda)$ focused on estimating $a_{dg}(\lambda)$ first through derivative analysis, optimizing the fit of $a_{dg}(\lambda)$ through iterative spectral evaluation, then estimating $a_{ph}(\lambda)$ using Gaussian decomposition. Steps described in this section are summarized in a schematic and accompanied by figures illustrating the primary components of each step (Fig. 2). Steps 1-7 evaluate $a_{t\text{-}w}(\lambda)$ to optimize estimates of $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ and Step 8 is a Gaussian decomposition of $a_{t\text{-}w}(\lambda)$ using estimated $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ with constraints defined below. For a detailed discussion on general algorithm framework and empirical relationships used, we refer the reader to section 4.1.2.

Step 1

To first parameterize $a_{dg}(\lambda)$, the second derivative of $a_{t\text{-}w}(\lambda)$ was calculated as:

$$\frac{d^2 a_{t-w}(\lambda)}{d\lambda^2} \approx \frac{a_{t-w}(\lambda_i) - 2a_{t-w}\left(\lambda_j\right) + a_{t-w}(\lambda_k)}{\Delta\lambda^2} \qquad (1)$$

where $\Delta\lambda$ indicates the wavelength resolution used to measure $a_{t\text{-}w}(\lambda)$ (here, 5 nm as described above) and $\Delta\lambda=\lambda_k-\lambda_j=\lambda_j-\lambda_i$, j=i+1, k=i+2 and λ_i is the current wavelength (Tsai and Philpot 1998). Points where the second derivative equals 0 indicate inflection points of the spectrum (Fig. 2a; Lee et al. 2007). In theory, for $a_{t\text{-}w}(\lambda)$, these are points where individual phytoplankton pigments least impact the underlying exponential signal and thus are considered as the observed signal most likely representative of $a_{dg}(\lambda)$ spectral shape. These points were defined as λ_{d0} and were found by identifying where $d^2a_{t\text{-}w}(\lambda)$ was approximately 0. These points were identified by rounding second derivative values to the median magnitude of the second derivative, which is a function of the magnitude of observed absorption. For example, if the median second derivative was 0.005, a value of 0.0008 at 440 nm would be considered not zero and not included (rounded to 0.001). A value of 0.0004 at 450 nm would be considered zero (rounded to 0.000), 450 nm would be classified as a λ_{d0} wavelength and the corresponding absorption would be used in Step 2.

Step 2

Using wavelengths identified in Step 1, an initial exponential expression was fitted following

$$a_{t-w}(\lambda_{d0}) = a_{t-w}(\lambda_0)e^{-S(\lambda_{d0}-\lambda_0)} \tag{2}$$

where λ_0 is the minimum wavelength in λ_{d0} (Fig. 2b). S derived from Eq. 2 was used as the initial estimate of S_{dg} and $a_{dg}(\lambda_0)$ was estimated at 440 nm by estimating the relative contribution of $a_{dg}(440)$ to $a_{t-w}(440)$ using a piece-wise exponential relationship derived from the training dataset as follows:

$$\%\, a_{ph}(440) = 1.038e^{-0.9257\left(\frac{a_{t-w}(555)}{a_{t-w}(680)}\right)} \text{where } \frac{a_{t-w}(555)}{a_{t-w}(680)} > 0.685 \tag{3}$$

or

$$\%\, a_{ph}(440) = 2.088e^{-1.946\left(\frac{a_{t-w}(555)}{a_{t-w}(680)}\right)} \text{ where } \frac{a_{t-w}(555)}{a_{t-w}(680)} \leq 0.685 \tag{4}$$

and

$$\%\, a_{dg}(440) = 100 - \%\, a_{phy}(440) \tag{5}$$

Eq. 3 was developed on the entire dataset, and outliers were determined as residuals outside 1.5 times the interquartile range (25th and 75th quantiles). After determining outliers, a moving window of 10% $a_{ph}(440)$ contribution to $a_{t-w}(440)$ was used to assess for a significant bias in residuals derived from this relationship. Bias was defined as a median residual value an order of magnitude different (positive or negative) than median residual bias between the relationship and all data points. This threshold indicated a bias for $a_{ph}(440)$ contributions greater than 60%, corresponding to an $a_{t-w}(555)$/ $a_{t-w}(680)$ ratio of 0.685. From this, a new relationship (Eq. 4) was developed to estimate $a_{ph}(440)$ percent contribution above 60% without bias. These equations are discussed further in Section 4.1.2 and figures referenced therein.

From the previous steps, the spectra for $a_{dg}(\lambda)$ was then estimated (Fig. 2b) as follows

$$a_{dg}(\lambda) = \left(a_{t-w}(440) \cdot \%a_{dg}(440)\right)\, e^{-S_{dg}(\lambda-440)} \tag{6}$$

Step 3

To determine if the $a_{dg}(\lambda)$ estimate was acceptable, we compared it to $a_{t\text{-}w}(\lambda)$:

$$a_{residual}(\lambda) = a_{t-w}(\lambda) - a_{dg}(\lambda) \tag{7}$$

If $a_{residual}(\lambda)$ was always positive, the previous variables - λ_0, $a_{dg}(\lambda_0)$, S_{dg} – were maintained at the current estimated values (e.g., λ_0=440 nm; Fig. 2c). If $a_{residual}(\lambda)$ was negative at any point, the wavelength corresponding to the most negative residual was used as λ_0, and $a_{dg}(\lambda_0)=a_{t\text{-}w}(\lambda_0)$ to re-calculate $a_{dg}(\lambda)$ from

$$a_{dg}(\lambda) = a_{dg}(\lambda_0)e^{-S_{dg}(\lambda-\lambda_0)} \tag{8}$$

Resulting $a_{residual}(\lambda)$ was re-calculated again following Eq. 7 for the new estimated $a_{dg}(\lambda)$. This step was repeated until all $a_{residual}(\lambda)$ values were positive, with S_{dg} incrementally adjusted by +0.0001 nm^{-1} to a maximum adjustment of +0.011 nm^{-1}. If a potential solution was not found, S_{dg} was then incrementally adjusted by -0.0001 nm^{-1} to a minimum adjustment of -0.004 nm^{-1}. The difference in adjustment and focus on positive adjustment values first is discussed further in Section 4.1.2. If no valid solution was found through this routine, the initial estimate of $a_{dg}(\lambda)$ was used; if a valid solution was found, that was the new $a_{dg}(\lambda)$ estimate (e.g., Fig. 2c). At this step, negative residual values were allowed, and accounted for in Step 5. This occurred in 91 of the 1,353 spectra evaluated (6.7% of the time).

Step 4

Using the new or initial $a_{dg}(\lambda)$ estimate, $a_{ph}(\lambda)$ was estimated (Fig. 2d) following

$$a_{ph}(\lambda) = a_{t-w}(\lambda) - a_{dg}(\lambda) \tag{9}$$

Step 5

To determine if $a_{dg}(\lambda)$ was estimated reasonably well, we considered the ratio of $a_{ph}(350)$:$a_{ph}(440)$, where a value greater than 1.5 was used to indicate whether a significant portion of the $a_{dg}(\lambda)$ signal was still present in the residuals. While some waters with a significant pigment contribution below 400 nm (e.g., mycosporine-like amino acids) may have violated this rule, it was generally applicable following discussion in Section 4.1.2.

If $a_{ph}(350)$:$a_{ph}(440)$ was greater than 1.5, a blended estimate of $a_{dg}(\lambda)$ was produced by fitting residuals from 350-400 nm with an exponential model (Fig. 2e) following:

$$a_{dg\_residual}(\lambda) = a_{residual}(\lambda_0)e^{-S_{residual}(\lambda-\lambda_0)} \tag{10}$$

A new estimate of $a_{dg}(\lambda)$, denoted as $a_{dg2}(\lambda)$, was created from:

$$a_{dg2}(\lambda) = a_{dg}(\lambda) + a_{dg\_residual}(\lambda) \tag{11}$$

A new S_{dg} was re-calculated for $a_{dg2}(\lambda)$ and the next iteration of $a_{dg}(\lambda)$ was estimated from:

$$a_{dg}(\lambda) = \left(a_{t-w}(440) \cdot \%a_{dg}(440)\right) e^{-S_{dg\_new}(\lambda-440)} \tag{12}$$

The $a_{dg}(\lambda)$ estimated from Eq. 12 was then iteratively evaluated by adjusting S_{dg} and assessing whether $a_{dg}(\lambda) > a_{t-w}(\lambda)$ at any wavelength, within each iteration. If $a_{dg}(\lambda) > a_{t-w}(\lambda)$, an offset was calculated by finding the wavelength where $a_{dg}(\lambda)$ was most overestimated following

$$a_{dg}(\lambda_{ind}) = a_{t-w}(\lambda_{ind}) - [a_{dg}(\lambda_{ind}) - a_{t-w}(\lambda_{ind})] \tag{13}$$

where λ_{ind} corresponds to the wavelength where $a_{dg}(\lambda)$ was most overestimated (maximum positive value from $a_{dg}(\lambda)$-$a_{t-w}(\lambda)$). Eq. 8 was then used to re-calculate $a_{dg}(\lambda)$, with λ_0=λ_{ind} and $a_{dg}(\lambda_0)$ equivalent to $a_{dg}(\lambda_{ind})$ from Eq. 13. The offset corrects for overestimations, but the application of a new λ_0 with the current S_{dg} can allow for overestimation at a different λ. If this step was performed, $a_{dg}(\lambda_0)$ was no longer set to the empirically-derived estimate of $a_{dg}(440)$, rather, and $a_{dg}(\lambda_0)$ and S_{dg} were altered simultaneously to find a solution (i.e., $a_{dg}(\lambda_0)$ set to $a_{t-w}(\lambda)$ minus an offset, and S_{dg} to the next iterative slope value). These steps were performed in a step-wise manner until $a_{ph}(350)$:$a_{ph}(440)$ was less than 1.5 or until the maximum number of allowable iterations, currently set to 20, was reached (Fig. 2f,g). If 20 iterations were reached without a solution, the final calculated model (from the 20th iteration) was used. This allowed for negative residuals to be included in the subsequent estimate of $a_{ph}(\lambda)$, following Eq. 9. However, negative values did not impact the spectral analysis used to identify pigments for fitting in Step 7 and negative values were removed through simultaneous fitting of $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ in Step 8, described below, where a constraint of non-negative values on solutions was imposed.

Step 6

In Step 6, we identified locations and widths of Gaussian curves in $a_{ph}(\lambda)$ derived from Eq. 9 or its equivalent derived from Steps 10-13. For this, we utilized a generic version of Eq. 1 to calculate the second derivative of estimated $a_{ph}(\lambda)$ as spectral features were accentuated in the second derivative relative to $a_{ph}(\lambda)$ (Fig. 2h). The second derivative was smoothed with a linear Savitzky-Golay filter using a 10 nm smoothing window. This smoothing reduced the number of features identified that correspond to signal noise. The smoothed second derivative was inverted to allow for identification of local maxima (equivalent to where the first derivative equals 0) and direct estimation of Gaussian curves on these spectral features – this is a key distinction between our methodology and published bottom-up approaches where Gaussian width and height are constrained as initial conditions or within a fitting window. Identified peaks were then used as an initial estimate of the number of peaks and each peak's location and width (Fig. 2h,i) with each Gaussian curve modeled following

$$f(x, \varphi, \mu, \sigma) = \varphi e^{-\frac{(x-\mu)^2}{2\sigma}} \tag{14}$$

where σ (nm) is the width of the curve, φ (m^{-1}) is the height of the Gaussian curve defined as $\varphi = \frac{1}{\sigma\sqrt{2\pi}}$, consistent with Gaussian curve height defined as full width at the half maximum, and μ (nm) is the peak center position. Any Gaussian curves with a σ less than 5 nm were removed at this stage, as these features were fit to noise and not pigments when using a 5 nm spectral resolution. At this stage, φ was scaled to the second derivative requiring re-parameterization of Gaussian curve heights relative to $a_{ph}(\lambda)$. Additionally, noisy data where $\sigma > 5$ nm can result in more identified peaks than was realistic. These issues were addressed in Step 7.

Step 7

For Step 7, Gaussian curves identified in Step 6 were scaled to $a_{ph}(\lambda)$. This was done by prioritizing peaks based on their relative prominence, identified as the φ determined for each identified peak in Step 6 (scaled to the second derivative). When identified in this manner, pigments that did not overlap, or overlap little, were fitted to $a_{ph}(\lambda)$ first, following the assumption that the majority of the absorption signal in that spectral region belongs to that Gaussian component (e.g., chlorophyll-a peak at 676 nm was typically prioritized for fitting first due to little overlap with other pigments). From this, $a_{ph}(\lambda)$ was iteratively fit with each Gaussian curve, the signal from that curve was removed, and the next Gaussian curve was fit to the remaining $a_{ph}(\lambda)$ signal to get a best approximation of φ for each Gaussian curve following

$$a_{ph_i}(\lambda) = a_{ph}(\lambda) - \sum_{i=1}^{n} \varphi_i e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \tag{15}$$

where n indicates the number of peaks identified for fitting from Steps 6d and 6e (Fig. 2) and μ and σ identified in Step 6 were used for each peak. Due to the additive nature of fitting Gaussian curves, there was potential for some peaks to have a negative height. After estimating an appropriate φ for each curve, we filtered out peaks with negative heights and we limited the total possible number of peaks to 16, although fewer peaks were typically identified (mean=7.7 peaks, s.d.=2.2 peaks). Most Gaussian decomposition schemes assume the presence of ~12 peaks (e.g., Hoepffner and Sathyendranath 1993; Wang et al. 2016; Chase et al. 2017). These studies have considered similar peak locations with minor differences accounting for a total of 16 unique peak locations in the literature. From this, we assumed if more than 16 peaks were present and all had a positive peak height, some identified peaks were noise or signals not affiliated with phytoplankton pigments that had not been removed in earlier steps. We sorted for likely pigment signals by prominence, using the same method described for peak height previously, and selected the 16 most prominent identified peaks if more than 16 peaks were identified. Next, we used the σ, φ and μ values identified for each Gaussian curve as input into a least squares Gaussian decomposition model that best fit our initial $a_{ph}(\lambda)$ estimate (Eq. 10) with the initial Gaussian curve estimates and fitting constraints described in Step 8 to define an updated set of Gaussian curves (Fig. 2j) following the expression:

$$a_{ph}(\lambda) = \sum_{i=1}^{n} \varphi_i e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \tag{16}$$

Step 8

Results from Steps 1-7 provided the start point for a combined retrieval of $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ from $a_{t\text{-}w}(\lambda)$. Using the estimate of $a_{dg}(\lambda)$ from Steps 1-7 and an estimate for each identified Gaussian curve fitted to $a_{ph}(\lambda)$, a least squares fitting approach was performed using the following expression:

$$a_{t-w}(\lambda) = a_{dg}(\lambda_0)e^{-S_{dg}(\lambda-\lambda_0)} + \sum_{i=1}^{n} \varphi_i e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \tag{17}$$

Analogous to methods used for identifying poorly constrained features that deviate from an underlying exponential signal presented elsewhere (e.g., Massicotte and Markager 2016), the model decomposed $a_{t\text{-}w}(\lambda)$ by utilizing a baseline exponential (Eq. 8) accompanied by a pre-defined number of Gaussian components based on previous steps (Eq. 16). This method differs from other Gaussian decomposition methods applied to particulate absorption (a_p), in that those methods typically have a pre-defined number of Gaussian components based on analysis of separate $a_{ph}(\lambda)$ for the respective system (e.g., Chase et al. 2013; Wang et al. 2016). This methodology fits primary pigments with width estimated from spectral features identified in the second derivative of estimated $a_{ph}(\lambda)$, allowing for a constrained solution to decomposing $a_{t\text{-}w}(\lambda)$ while not assuming the presence of any specific types of phytoplankton. Parameters in Eq. 17 were constrained utilizing results from Steps 1-7: $a_{dg}(\lambda_0)$ can vary from 0 m^{-1} to $a_{t\text{-}w}(\lambda_0)$, S_{dg} can vary by -0.002 nm^{-1} to +0.003 nm^{-1} from the input estimate, Gaussian peak width can vary from input width to 3 times the input width, Gaussian peak height can vary by 0.25 times input height to 3 times input height and μ is fixed at the identified location due to high confidence in the second derivative analysis. DAISEA output was as follows: $a_{dg}(\lambda)$ was that estimated in Eq. 17, while $a_{ph}(\lambda)$ was the difference between observed $a_{t\text{-}w}(\lambda)$ and $a_{dg}(\lambda)$ from Eq. 17 (Fig. 3). Step 8 ensures coherence between the exponential signal and overlying deviations due to $a_{ph}(\lambda)$ as constrained through Steps 1-7 in a flexible manner, while not assuming that $a_{ph}(\lambda)$ can be best parameterized by 6-8 Gaussian curves. Fitting of secondary features was possible but also increases the probability of over-constraining a solution (i.e. less flexibility in adjustments to $a_{dg}(\lambda)$).

2.2.1 Low $a_{ph}(\lambda)$ waters

We found that waters dominated by $a_{dg}(\lambda)$ were best decomposed by fitting an initial exponential function and adjusting to a realistic solution following Eq. 8, 9 and 13. These cases were identified after Eq. 3 and 4; waters were considered dominated by $a_{dg}(\lambda)$ where the ratio of $a_{t\text{-}w}(555)$:$a_{t\text{-}w}(680) > 2.528$ (the empirical value indicating $a_{ph}(440) < 10\%$ of $a_{t\text{-}w}(440)$). For these situations, the algorithm opted out of the Gaussian decomposition routine and followed a simplified routine analogous to Steps 2-4, where S_{dg} was considered equivalent to S calculated for $a_{t\text{-}w}(\lambda)$ (Eq. 2), and magnitude was adjusted so that $a_{dg}(\lambda) \leq a_{t\text{-}w}(\lambda)$. We chose this threshold as forcing Eq. 17 to fit all cases resulted in significantly more error in S_{dg} estimates when $a_{ph}(440)$ contributed < 10% of $a_{t\text{-}w}(440)$. Above this threshold, using Eq. 17 to fit for $a_{ph}(\lambda)$ improved estimates of $a_{dg}(\lambda)$ and S_{dg} while also providing an estimate of $a_{ph}(\lambda)$. The exact value of 10% may not be an ideal threshold for all datasets but worked well as a threshold here and fit within our presentation scheme. Eq. 3 and 4 are empirical and follow band-ratio techniques used for fitting S_{dg} in current semi-analytical schemes (Lee et al. 2009; Matsuoka et al. 2013). Noise in this relationship was explained by variability in the exact shape of $a_{ph}(\lambda)$ due to varying phytoplankton composition, physiology and pigment packaging effects (Bricaud and Morel 1986; Bricaud et al. 1983; Ciotti et al. 2002; Johnsen et al. 1994) as well as variability in the spectral shape and features of $a_g(\lambda)$ and $a_d(\lambda)$ (Grunert et al. 2018). As the algorithm is currently optimized for a global approach, users may find that adjusting the empirical values used to initially estimate $a_{dg}(440)$ and adjusting the value of 1.5 for the ratio of $a_{ph}(350)$:$a_{ph}(440)$ (Step 5) for a value more representative of their study region results in better algorithm performance.

2.2.2 Functions

To develop DAISEA, we focused on creating a primary, custom Matlab function – daisea, an approach that utilizes derivative analysis and iterative fitting to optimize input spectra used in a least squares Gaussian decomposition scheme fitting an exponential signal and a pre-defined number of constrained Gaussian peaks. DAISEA uses a package of custom sub-functions. This package is freely available via GitHub (https://github.com/bricegrunert/daisea/tree/v1.0.0; DOI: 10.5281/zenodo.1306817). Version updates will follow Github conventions. Users are encouraged to use the most recent version for application.

2.3 Data Analysis

To assess the performance of DAISEA across a variety of water conditions, we present results as eight different categories based on the percent contribution of $a_{ph}(440)$ relative to $a_{t-w}(440)$, with the distribution of spectra within these classes shown in Fig. 1. Classes were defined as the percent contribution a_{ph} has to the overall absorption budget at 440 nm (%$a_{ph}(440)$) of 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, and >70, with n=286, 257, 303, 210, 146, 89, 34 and 28 spectra, respectively. This classification scheme emphasizes the relative, not the absolute, contribution of phytoplankton to the overall absorption signal. Thus, waters where $a_{ph}(440)$ is the dominant contributor to total absorption are not limited to highly productive waters. In this sense, algorithm performance was not assessed across classic definitions of Case 1 or Case 2 waters (Morel and Prieur 1977). Rather, the only group dominated by coastal and inland waters was 0-10

%a_{ph}(440). It should be emphasized that these categories are only used to present the data within the context of relative contribution of $a_{ph}(\lambda)$ and $a_{dg}(\lambda)$. Beyond separating 0-10 %a_{ph}(440) spectra for fitting using Eq. 2, the algorithm does not analyze spectra differently based on these categories.

To determine whether $a_{ph}(\lambda)$ or $a_{dg}(\lambda)$ was retrievable, we calculated the absolute difference in the opposing metric and compared it to the observed value. For example, if $a_{ph\_obs}(\lambda) > |a_{dg\_obs}(\lambda)-a_{dg\_est}(\lambda)|$, we consider it retrievable at that wavelength. Within each %a_{ph}(440) group, we summed the total number of instances at each wavelength where $a_{ph}(\lambda)$ or $a_{dg}(\lambda)$ was greater than the absolute difference in the opposing metric and divided by the total number of spectra to get a percent retrievable metric for that %a_{ph}(440) group. In addition to percent retrievable metrics, we calculated Bayes factors (BF_{10}, unitless) to assess fit significance (Wetzels and Wagenmakers 2012). Bayes factors represent the likelihood that the fitted model adequately represents the data relative to an alternative model. Bayes factors can be interpreted literally, so that $BF_{10}=2$ means the data are twice as likely to be explained by the fitted model than an alternative model. Here, we used a $BF_{10} \geq 3$ as the threshold for significance (Wetzels and Wagenmakers 2012). We also calculated root mean square difference (RMSD), normalized RMSD (NRMSD), bias, mean absolute difference (MAD) and unbiased absolute percent difference (UAPD) using the following expressions:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}\left[\left(x_i^{estimated}\right) - \left(x_i^{observed}\right)\right]^2}{n}} \qquad (18)$$

$$NRMSD\ (\%) = \frac{RMSD}{x_{max}^{observed} - x_{min}^{observed}} \times 100 \qquad (19)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{estimated} - x_i^{observed}\right) \tag{20}$$

$$MAD = \frac{\sum_{i=1}^{n}\left(\left|x_i^{estimated} - x_i^{observed}\right|\right)}{n} \tag{21}$$

$$UAPD\ (\%) = \frac{\left|x_{estimated} - x_{observed}\right|}{0.5(x_{estimated} + x_{observed})} \times 100 \tag{22}$$

3. Results

3.1 DAISEA Performance

Here, we present the results of DAISEA performance on the test dataset. Across all groups, $a_{dg}(\lambda)$ was retrievable >80% of the time for wavelengths < 450 nm (Fig. 4a). For waters where $a_{dg}(\lambda)$ contributed greater than 60%, it was retrievable at a rate of >80% for all wavelengths up to 650 nm. For $a_{ph}(\lambda)$, local maxima in retrieval corresponded to chlorophyll-a absorption peaks (~440 and 680 nm within DAISEA), with these wavelengths displaying >80% retrievability for waters with $\%a_{ph}(440) > 10$ (Fig. 4b). Relative difference for $a_{ph}(\lambda)$ and $a_{dg}(\lambda)$ was parameterized as NRMSD and displayed excellent performance for both parameters across most wavelengths and environments. For all conditions except $\%a_{ph}(440) > 70$, $a_{dg}(\lambda)$ had a mean difference less than 20% for wavelengths from 350-650 nm (Fig. 4c). Mean $a_{ph}(\lambda)$ difference was generally less than 20% from 350-650 nm when $\%a_{ph}(440)$ was > 10 (Fig. 4d). As seen in Fig. 4 and 5, $a_{ph}(\lambda)$ was biased to greater than observed values when it was a non-dominant contributor at 440 nm and was biased towards values less than observed when it was a dominant contributor at 440 nm, and vice versa for $a_{dg}(\lambda)$ (Fig. 4e). Mean absolute difference generally decreased as the contribution of $a_{ph}(440)$ increased (Fig. 4f).

The threshold for estimating $a_{dg}(\lambda)$ with DAISEA appears to be $\%a_{ph}(440) < 70$; for these conditions, $a_{dg}(\lambda)$ is estimated with NRMSD < 20% from 350-650 nm. NRMSD for $a_{ph}(\lambda)$ was < 20% for the majority of wavelengths between 400-650 nm when $\%a_{ph}(440)$ was > 10. This was also consistent when considering the retrievability of $a_{ph}(440)$ under different conditions and can be considered as the threshold for estimating $a_{ph}(\lambda)$. S_{dg} uncertainty increased with increasing contribution of $a_{ph}(440)$; however, performance was reasonable across all water conditions and estimates (Table 1). This was also confirmed when considering Bayes factors for fitted models. Overall, BF_{10} were quite high with 94.0% of collective model retrievals showing a $BF_{10} > 3$, our cutoff for significance, and 92.9% with a $BF_{10} > 10$, demonstrating strong confidence in our approach. The model retrieved $a_{dg}(\lambda)$ with slightly better success than $a_{ph}(\lambda)$, with $BF_{10} > 3$ for 99.5% and 88.5% of the dataset, respectively. Of models that did not fit observed $a_{ph}(\lambda)$ adequately (n=156), the majority of poor model fits occurred in waters where $\%a_{ph}(440) < 20$ (n=129). Only 6 model fits were not adequate for $a_{dg}(\lambda)$ and distribution across $\%a_{ph}(440)$ groups was random.

Using the 2.528 threshold applied to $a_{t\text{-}w}(440)$ ratios to separate low $\%a_{ph}(440)$ did present issues, particularly in the $\%a_{ph}(440)$ of 10-20 category. This threshold miscategorized spectra from this category as having $\%a_{ph}(440) < 10$ in 175 of 257 spectra, resulting in poorly resolved $a_{ph}(\lambda)$ for these. This highlights the primary drawback of utilizing empirical relationships and a weakness in our approach. Due to using the 2.528 threshold to separate $\%a_{ph}(440)$ contribution, more than half of the $a_{ph}(\lambda)$ estimates in the 10-20 $\%a_{ph}(440)$ group had negative values at wavelengths greater than 650 nm, outside of the chlorophyll-a (Chl) absorption peak at 676 nm (typically assigned to 680 nm within the algorithm framework; Fig. 5). While we attempted to account for this by using an offset from Eq. 13, moving the location of λ_0 and maintaining S_{dg} can result in negative values at a different spectral location. However, the benefit of using this scheme was evident in improved estimates of S_{dg} for correctly classified spectra (i.e., spectra where $\%a_{ph}(440)$ was <10). Without a threshold to separate these spectra, attempting to fit $a_{ph}(\lambda)$ when it contributed very little resulted in poor algorithm performance, with more spectra poorly fit than with the current approach including a threshold.

One of the primary motivators for developing DAISEA was to accurately retrieve S_{dg} without any assumptions regarding spectral shape while also independently estimating $a_{ph}(\lambda)$. Our results suggest that this is possible across a variety of optical conditions with a reasonable to excellent degree of accuracy, depending on the relative contribution of $a_{dg}(\lambda)$. Across the different groups of varying $a_{ph}(440)$ contribution, median difference in S_{dg} varied from 0.9-17.7%, with third quartile differences ranging from 2.4-39.2% (Fig. 6a; Table 1). Mean S_{dg} observed across all spectra in the test dataset was 0.0147 nm^{-1} compared to a mean estimated value of 0.0150 nm^{-1}, while median observed and estimated S_{dg} was 0.0152 and 0.0153 nm^{-1}, respectively. Across individual groups, we evaluated the differences and present anticipated accuracy for S_{dg} (Table 1). For most groups, median difference was << 0.001 nm^{-1} and absolute differences affiliated with the 1st and 3rd quantiles ranged up to -0.0037 and 0.0021 nm^{-1}, respectively, but were typically much smaller. We also considered distribution of differences in S_{dg} across all groups and it followed a predominantly normal distribution (data not shown), without an obvious bias between observed and estimated S_{dg} regardless of $\%a_{ph}(440)$ contribution (Fig. 6b).

3.2 Consistency in Gaussian features

We considered the accuracy of our Gaussian component locations within DAISEA by comparing to Gaussian component locations identified on observed $a_{ph}(\lambda)$ using the same Gaussian decomposition approach (Fig. 7). Overall, peak locations were quite similar, although DAISEA fitted more peaks (total peaks=10,394; 7.7 peaks/spectra) than for observed $a_{ph}(\lambda)$ (total peaks=8,794; 6.5 peaks/spectra). Since $a_{ph}(\lambda)$ estimated from the algorithm was derived from the smoothed residuals of $a_{t\text{-}w}(\lambda) - a_{dg\_est}(\lambda)$,the additional noise in the spectra was derived from deviations in $a_d(\lambda)$ and $a_g(\lambda)$ not accounted for by a strictly exponential fit. We discuss potential reasons for an increase in fitted peaks in DAISEA output over observed $a_{ph}(\lambda)$ in Section 4.2, as well as fitting significantly fewer peaks under our approach than other Gaussian decomposition approaches (e.g., Chase et al. 2013).

4. Discussion

4.1 DAISEA

4.1.1 Application

As evidenced here and elsewhere, hyperspectral ocean color data provides a means for estimating more variables in a less constrained manner (Bracher et al. 2009; Dierssen et al. 2015; Uitz et al. 2015; Vandermeulen et al. 2017; Wang et al. 2017). Global variability in water optical properties is significant yet the non-uniqueness of $R_{rs}(\lambda)$ hampers consistent interpretation across both empirical and semi-analytical methods (Werdell et al. 2018 and references therein). Previous concepts for working around this issue, particularly in light of multispectral limitations, have included screening $R_{rs}(\lambda)$ to most likely cases based on optical water types, non-linear spectral optimization, linear matrix inversion, bulk inversion, ensemble inversion and spectral deconvolution (Brando et al. 2012; Hieronymi et al. 2017; Mélin and Vantrepotte 2015; Trochta et al. 2015; Werdell et al. 2018). These approaches are broadly defined as bottom-up and top-down strategies (Mouw et al. 2015), where bottom-up strategies simultaneously solve for all parameters while top-down strategies allow for independent retrieval of absorbing and scattering constituents (a and b_b).

Current hyperspectral approaches capable of estimating $a_{ph}(\lambda)$, $a_d(\lambda)$ and $a_g(\lambda)$ operate with bounded ranges and relatively defined pigment locations within a bottom-up framework using $R_{rs}(\lambda)$. Notably, Chase et al. (2017) provide a global approach that performs quite well on *in situ* reflectance data through the use of assumed starting points for IOP's based on global means, Gaussian decomposition of $a_{ph}(\lambda)$ using constrained Gaussian curves and lower and upper bounds imposed on all IOP's. Here, we developed a hyperspectral decomposition approach, DAISEA, suitable for a top-down inversion strategy analogous to spectral deconvolution approaches developed for multispectral data (e.g., QAA; Lee et al. 2002; Werdell et al. 2018). Within DAISEA, spectral shapes and features of $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ are not assumed but identified through derivative analysis and comparing retrieved spectra to anticipated thresholds. This provides a means for estimating the spectral shape of $a_{dg}(\lambda)$ and phytoplankton pigment identification free of explicit assumptions while also limiting retrievals based on a suitable signal-to-noise ratio (i.e., the algorithm only fits primary spectral features after spectral smoothing with a Savitzky-Golay filter). DAISEA is anticipated to pair well with future inversion schemes designed to work with hyperspectral $R_{rs}(\lambda)$ and flow-through $a_{t-w}(\lambda)$ datasets (e.g., Twardowski and Tonizzo 2018).

Top-down approaches have been used to retrieve IOP's in an independent manner in a variety of aquatic environments (Mouw et al. 2015). We demonstrated that DAISEA works quite well for *in situ* absorption datasets. The algorithm does not currently perform well with top-down inversion strategies designed for multispectral data due to relatively high error in estimating $a_{t-w}(350)$. This is a short-coming of our approach, but future top-down hyperspectral inversion approaches are expected to have minimal error and bias across the full spectral range of PACE

(350-800 nm) (Twardowski and Tonizzo 2018). While it is not possible at this time to fully assess the compatibility of DAISEA with these developing approaches, early indications suggest that spectral accuracy of $a_{ph}(\lambda)$ and $a_{dg}(\lambda)$ could be quite reasonable and in-line with error attached to current approaches (Chase et al. 2017; Wang et al. 2018). Additionally, we did not pursue separation of $a_d(\lambda)$ and $a_g(\lambda)$ as current methods for separating within a top-down scheme rely on empirical approaches. Independent approaches for separating these features are currently being considered for future work. Considering the accuracy of estimating S_{dg} with DAISEA (Table 1) and minimal additional uncertainty in separating $a_{dg}(\lambda)$ into $a_d(\lambda)$ and $a_g(\lambda)$ in future work (5-10%), S_g could feasibly be retrieved with a median difference of 0.001-0.002 nm^{-1} across most optical conditions. This would provide an adequate resolution for estimating CDOM source, production and degradation processes as characterized in a variety of *in situ* studies.

4.1.2 General framework and empirical relationships

The general premise of DAISEA is that $a_{dg}(\lambda)$ can be accurately modeled using an exponential model and that deviations from this exponential model are solely due to $a_{ph}(\lambda)$. There are alternate explanations for both of these assumptions (e.g., Cael and Boss 2017; Catalá et al. 2016); however, there is biogeochemical significance in S_{dg}, while phytoplankton would presumably produce the largest deviation from an exponential signal as observable from satellite ocean color data. Beyond these basic assumptions, we also considered the relationship between $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ within a theoretical framework (Fig. 8). Based on this framework, it is important to recognize how varying contributions of each component will inherently lead to specific biases. For example, $S_{350:400}$ fitted to $a_{t\text{-}w}(\lambda)$ where $a_{ph}(\lambda)$ contributes will result in lower slope values due to the higher contribution of $a_{ph}(\lambda)$ at 400 nm relative to 350 nm. This is why we increased estimates of S_{dg} first, then alternated to decreasing S_{dg}, as an exponential fit of $a_{t\text{-}w}(\lambda)$ will produce lower S values when $a_{ph}(\lambda)$ contributes to the signal. Finding where the second derivative of $a_{t-w}(\lambda)$ equals 0 and fitting an exponential at these points minimizes this impact (essentially "cutting through" primary pigment features for a least squares fit); however, there was still a consistent bias towards lower S_{dg} values as $a_{ph}(440)$ contribution increased, as expected. The general framework illustrated in Fig. 8 is also the justification for setting a ratio of 1.5 to $a_{ph}(350)$:$a_{ph}(440)$; when the residual used to estimate $a_{ph}(\lambda)$ had a ratio higher than this, it was almost always indicative of a significant portion of the $a_{dg}(\lambda)$ signal remaining in the residual used to calculate $a_{ph}(\lambda)$.

In short of independent variables to validate each component of interest, some explicit assumptions are required within any algorithm framework. Here, we chose to limit our solutions by constraining initial $a_{dg}(440)$ estimates by the empirical relationship between $a_{t-w}(555)/a_{t-w}(680)$ and $\%a_{ph}(440)$ from the training dataset (Fig. 9a) and a theoretical ratio of 1.5 for $a_{residual}(350)/a_{residual}(440)$ (Eq. 7) to determine whether the contribution of $a_{dg}(\lambda)$ to $a_{t-w}(\lambda)$ from 350-400 nm had been reasonably estimated and removed. These relationships do not explicitly dictate the final product, but guide the algorithm to reasonable estimates, at which point fitting is not constrained by these specific values. They do, however, leave an impact on how results are constrained. As we discussed previously, empirical relationships can often fall short of their intended accuracy. Despite a similar optical and geographical distribution between the training and test datasets (Fig. 1), the piece-wise exponential relationship derived from the training dataset to predict $\%a_{ph}(440)$ (r^2=0.91, RMSD=0.068 for fitted points) did not predict the same relationship nearly as well for the test dataset (r^2=0.58, RMSD=0.110; Fig. 9b). We considered sensitivity to this empirical relationship on the test dataset. By using a single exponential expression fitted to the test dataset (data points in Fig. 9b), with a value of 0.779 instead of 1.038 and -0.5834 instead of -0.9257 in Eq. 3, NRMSD fell below 6% for all wavelengths, with higher values in the UV and lower at longer wavelengths (data not shown). However, the number of Gaussian curves fitted within the algorithm were different for 17% of spectra, with nearly all instances fit with one fewer Gaussian curves. This suggests that the algorithm is relatively robust across datasets but does exhibit significant sensitivity to empirical values.

We adjusted the theoretical value of 1.5 to lower values as a stricter threshold for removing a residual $a_{dg}(\lambda)$ signal from the estimate of $a_{ph}(\lambda)$ derived from Eq. 9. Algorithm results did not significantly change with values less than 1.5; however, spectra that contain pigments below 400 nm (e.g., mycosporine-like amino acids) required a value of 1.5 to adequately identify and fit these pigments. For spectra that did not contain a significant absorption signal below 400 nm, the shape of the spectra here is predominantly exponential. If a Gaussian component is erroneously assigned in this spectral region, as in the example spectra in Figs. 2 and 3, the curve can be minimized in Step 8 by fitting an adjusted exponential to $a_{t\text{-}w}(\lambda)$. This adjustment is allowable within the constraints provided and provides for consistent and stable solutions, since no Gaussian components are dropped. This approach and the empirical values used best fit our global dataset, but adjusting empirical values to a regional value is quite easy within the code available online (Section 2.2.2).

4.2 Gaussian Decomposition Approaches

Gaussian component location was consistent between observed $a_{ph}(\lambda)$ and DAISEA output (Fig. 7). We did not utilize Gaussian components to estimate the final $a_{ph}(\lambda)$ output, as we found that a smoothed residual of $a_{t\text{-}w}(\lambda)$-$a_{dg\_est}(\lambda)$ more accurately represented observed $a_{ph}(\lambda)$. This is likely due to fitting fewer Gaussian components than needed to accurately model $a_{ph}(\lambda)$, as DAISEA fits fewer peaks than alternate Gaussian decomposition schemes due to a difference in methodologies (Hoepffner and Sathyendranath 1993; Chase et al. 2013). These algorithms typically identify pigments from first and second derivative analysis of an existing database of phytoplankton spectra then assign windows around these points (typically 12 peaks). Our approach focuses on identifying primary pigment features to best fit observed $a_{t\text{-}w}(\lambda)$ without assuming the locations of pigments, resulting in fewer identified peaks (~7 peaks). There is potential to increase the sensitivity of the peak finding step. Our focus was to retrieve $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ accurately, including spectral shape, rather than individually parameterizing phytoplankton pigments. It is possible to utilize the $a_{ph}(\lambda)$ output in a separate Gaussian decomposition scheme, or other approach that identifies phytoplankton pigments. However, it should be noted that derivative analysis of the final $a_{ph}(\lambda)$ output, even after smoothing, resulted in more identified peaks than the observed $a_{ph}(\lambda)$ using our scheme. This is very likely due to the inclusion of chromophores in $a_g(\lambda)$ and $a_d(\lambda)$ that result in deviations from the typical exponential expression used to model these parameters, features visibly apparent in many of the $a_{dg}(\lambda)$ spectra. While often overlooked, these features have been recognized for some time (Babin et al. 2003; Schwarz et al. 2002) and a recent methodology for fitting these peaks provides a means of both quantifying them and more accurately modeling the underlying exponential signal (Catalá et al. 2016; Massicotte and Markager 2016; Grunert et al. 2018). This approach is useful for *in situ* data, but not practical for our proposed methodology and likely a non-factor when considering $a_{t\text{-}w}(\lambda)$ derived from satellite $R_{rs}(\lambda)$.

5. Conclusions

We show that across most optical conditions considered, DAISEA can accurately estimate $a_{dg}(\lambda)$, S_{dg} and $a_{ph}(\lambda)$ magnitude and spectral features for all water types where $\%a_{ph}(440) > 10$.

We parameterized the percent of $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ estimates that were retrievable by comparing the signal observed for one IOP to the difference between estimated and observed values obtained for the other IOP. Consistent with the general accuracy of DAISEA, primary features (i.e., chlorophyll-a absorption peaks) of $a_{ph}(\lambda)$ were retrievable for greater than 80% of spectra across environments where $\%a_{ph}(440) > 10$; $a_{dg}(\lambda)$ was retrievable for at least 80% of spectra from 350-650 nm when $\%a_{ph}(440) < 70$. NRMSD metrics suggest strong algorithm performance across most optical variability from 350-650 nm. Algorithm bias shows a tendency to overestimate $a_{ph}(\lambda)$ when $\%a_{ph}(440) < 40$ and to underestimate $a_{ph}(\lambda)$ when $\%a_{ph}(440) > 60$.

Currently, coincident hyperspectral measurements of $R_{rs}(\lambda)$, $b_{bp}(\lambda)$, $a_{ph}(\lambda)$, $a_d(\lambda)$ and $a_g(\lambda)$ observed down to a minimum wavelength of 350 nm, the proposed lower wavelength limit of PACE, are quite uncommon relative to coincident measurements at wavelengths ≥ 400 nm. This, along with limited hyperspectral inversion approaches free of spectral bias, limited our ability to fully assess how well DAISEA will perform in the context of a top-down spectral deconvolution approach. Despite empirical schemes for separation of $a_{dg}(\lambda)$ and S_{dg} into the component parts (NAP and CDOM; e.g., Dong et al. 2013), we did not pursue separation here. Considering current algorithm performance, we anticipate that a well-performing scheme to separate $a_{dg}(\lambda)$ into its component parts will allow for appropriate resolution in S_g to estimate source and degradation state of CDOM in the surface ocean.

Acknowledgements

We gratefully acknowledge contributors to the SeaBASS data set (https://seabass.gsfc.nasa.gov/). Thank you to Piotr Kowalczuk, Frédéric Mélin and two anonymous reviewers for comments that greatly improved the final version of this manuscript. Algorithm development was funded by a NASA Earth and Space Science Fellowship awarded to Grunert.

References

Andrew, A.A., Del Vecchio, R., Subramaniam, A., & Blough, N.V. (2013). Chromophoric dissolved organic matter (CDOM) in the Equatorial Atlantic Ocean: Optical properties and their relation to CDOM structure and source. *Marine Chemistry, 148,* 33-43

Babin, M., Stramski, D., Ferrari, G.M., Claustre, H., Bricaud, A., Obolensky, G., & Hoepffner, N. (2003). Variations in the light absorption coefficients of phytoplankton, nonalgal particles, and dissolved organic matter in coastal waters around Europe. *Journal of Geophysical Research, 108*

Behrenfeld, M.J., Hu, Y., Hostetler, C.A., Dall'Olmo, G., Rodier, S.D., Hair, J.W., & Trepte, C.R. (2013). Space-based lidar measurements of global ocean carbon stocks. *Geophysical Research Letters, 40,* 4355-4360

Behrenfeld, M.J., Hu, Y., O'Malley, R.T., Boss, E.S., Hostetler, C.A., Siegel, D.A., Sarmiento, J.L., Schulien, J., Hair, J.W., Lu, X., Rodier, S., & Scarino, A.J. (2016). Annual boom–bust cycles of polar phytoplankton biomass revealed by space-based lidar. *Nature Geoscience, 10,* 118-122

Bracher, A., Vountas, M., Dinter, T., Burrows, J.P., Rüttgers, R., & Peeken, I. (2009). Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data. *Biogeosciences, 6,* 751-764

Brando, V., Dekker, A., Park, Y.J., & Schroeder, T. (2012). Adaptive semianalytical inversion of ocean color radiometry in optically complex waters. *Appl Opt, 51*

Bricaud, A., Babin, M., Morel, A., & Claustre, H. (1995). Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: Analysis and parameterization. *Journal of Geophysical Research, 100,* 13,321-313,332

Bricaud, A., & Morel, A. (1986). Light attenuation and scattering by phytoplanktonic cells: a theoretical modeling. *Appl Opt, 25,* 571-580

Bricaud, A., Morel, A., & Prieur, L. (1983). Optical efficiency factors of some phytoplankters. *Limnology & Oceanography, 28,* 816-832

Cael, B.B., & Boss, E. (2017). Simplified model of spectral absorption by non-algal particles and dissolved organic materials in aquatic environments. *Opt Express, 25,* 25486-25491

Catalá, T.S., Reche, I., Ramón, C.L., López-Sanz, À., Álvarez, M., Calvo, E., & Álvarez-Salgado, X.A. (2016). Chromophoric signatures of microbial by-products in the dark ocean. *Geophysical Research Letters, 43,* 7639-7648

Chang, G.C., & Dickey, T.D. (2004). Coastal ocean optical influences on solar transmission and radiant heating rate. *Journal of Geophysical Research, 109*

Chase, A., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., Dall'Olmo, G., & Westberry, T.K. (2013). Decomposition of in situ particulate absorption spectra. *Methods in Oceanography, 7,* 110-124

Chase, A.P., Boss, E., Cetinić, I., & Slade, W. (2017). Estimation of Phytoplankton Accessory Pigments From Hyperspectral Reflectance Spectra: Toward a Global Algorithm. *Journal of Geophysical Research: Oceans, 122,* 9725-9743

Ciotti, A.M., Lewis, M., & Cullen, J.J. (2002). Assessment of the relationship between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient. *Limnology & Oceanography, 47,* 404-417

Cory, R.M., & Kling, G.W. (2018). Interactions between sunlight and microorganisms influence dissolved organic matter degradation along the aquatic continuum. *Limnology and Oceanography Letters, 3*, 102-116

Danhiez, F., Vantrepotte, V., Cauvin, A., Lebourg, E., & Loisel, H. (2017). Optical properties of chromophoric dissolved organic matter during a phytoplankton bloom. Implication for DOC estimates from CDOM absorption. *Limnology and Oceanography, 62*, 1409-1425

Dutkiewicz, S., Hickman, A.E., Jahn, O., Gregg, W.W., Mouw, C.B., & Follows, M.J. (2015). Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model. *Biogeosciences, 12*, 4447-4481

D’Errico, J. (2005). Surface Fitting using gridfit. In. (https://www.mathworks.com/matlabcentral/fileexchange/8998-surface-fitting-using-gridfit): MATLAB Central File Exchange

Fichot, C.G., & Benner, R. (2011). A novel method to estimate DOC concentrations from CDOM absorption coefficients in coastal waters. *Geophysical Research Letters, 38*

Fichot, C.G., Benner, R., Kaiser, K., Shen, Y., Amon, R.M.W., Ogawa, H., & Lu, C.-J. (2016). Predicting Dissolved Lignin Phenol Concentrations in the Coastal Ocean from Chromophoric Dissolved Organic Matter (CDOM) Absorption Coefficients. *Frontiers in Marine Science, 3*

Fichot, C.G., Kaiser, K., Hooker, S.B., Amon, R.M., Babin, M., Belanger, S., Walker, S.A., & Benner, R. (2013). Pan-Arctic distributions of continental runoff in the Arctic Ocean. *Sci Rep, 3*, 1053

Grunert, B.K., Mouw, C.B., & Ciochetto, A.B. (2018). Characterizing CDOM Spectral Variability Across Diverse Regions and Spectral Ranges. *Global Biogeochemical Cycles, 32*, 57-77

Hansell, D.A., Carlson, C.A., Repeta, D.J., & Schlitzer, R. (2009). Dissolved organic matter in the ocean: A controversy stimulates new insights. *Oceanography, 22*, 202-211

Helms, J.R., Stubbins, A., Perdue, E.M., Green, N.W., Chen, H., & Mopper, K. (2013). Photochemical bleaching of oceanic dissolved organic matter and its effect on absorption spectral slope and fluorescence. *Marine Chemistry, 155*, 81-91

Helms, J.R., Stubbins, A., Ritchie, J.D., Minor, E.C., Kieber, D.J., & Mopper, K. (2008). Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter. *Limnology and Oceanography, 53*, 955-969

Hieronymi, M., Müller, D., & Doerffer, R. (2017). The OLCI Neural Network Swarm (ONNS): A Bio-Geo-Optical Algorithm for Open Ocean and Coastal Waters. *Frontiers in Marine Science, 4*

Johnsen, G., Nelson, N., Jovine, R.V.M., & Prezelin, B. (1994). Chromoprotein- and pigment dependent modeling of spectral light absorption in two dinoflagellates, Prorocentrum minimum and Heterocapsa pygmaea. *Marine Ecology Progress Series, 114*, 245-258

Kim, G.E., Gnanadesikan, A., & Pradal, M.-A. (2016). Increased Surface Ocean Heating by Colored Detrital Matter (CDM) Linked to Greater Northern Hemisphere Ice Formation in the GFDL CM2Mc ESM. *Journal of Climate, 29*, 9063-9076

Lee, Z., Carder, K.L., Arnone, R., & He, M. (2007). Determination of primary spectral bands for remote sensing of aquatic environments. *Sensors, 7*, 3428-3441

Lee, Z., Carder, K.L., & Arnone, R.A. (2002). Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. *Appl Opt, 41*, 5755-5772

Mannino, A., Novak, M.G., Hooker, S.B., Hyde, K., & Aurin, D. (2014). Algorithm development and validation of CDOM properties for estuarine and continental shelf waters along the northeastern U.S. coast. *Remote Sensing of Environment, 152*, 576-602

Mentges, A., Feenders, C., Seibt, M., Blasius, B., & Dittmar, T. (2017). Functional Molecular Diversity of Marine Dissolved Organic Matter Is Reduced during Degradation. *Frontiers in Marine Science, 4*, 194

Morel, A., & Prieur, L. (1977). Analysis of variations in ocean color. *Limnology & Oceanography, 22*, 709-722

Mouw, C.B., Greb, S., Aurin, D., DiGiacomo, P.M., Lee, Z., Twardowski, M., Binding, C., Hu, C., Ma, R., Moore, T., Moses, W., & Craig, S.E. (2015). Aquatic color radiometry remote sensing of coastal and inland waters: Challenges and recommendations for future satellite missions. *Remote Sensing of Environment, 160*, 15-30

Mélin, F., & Vantrepotte, V. (2015). How optically diverse is the coastal ocean? *Remote Sensing of Environment, 160*, 235-251

Nelson, N.B., Siegel, D.A., Carlson, C.A., & Swan, C.M. (2010). Tracing global biogeochemical cycles and meridional overturning circulation using chromophoric dissolved organic matter. *Geophysical Research Letters, 37*

Ogawa, H., Amagai, Y., Koike, I., Kaiser, K., & Benner, R. (2001). Production of refractory dissolved organic matter by bacteria. *Science, 292*, 917-920

Riedel, T., & Dittmar, T. (2014). A method detection limit for the analysis of natural organic matter via Fourier transform ion cyclotron resonance mass spectrometry. *Analytical chemistry, 86*, 8376-8382

Sadeghi, A., Dinter, T., Vountas, M., Taylor, B., Altenburg-Soppa, M., & Bracher, A. (2012). Remote sensing of coccolithophore blooms in selected oceanic regions using the PhytoDOAS method applied to hyper-spectral satellite data. *Biogeosciences, 9*, 2127-2143

Schwarz, J.N., Kowalczuk, P., Kaczmarek, S., Cota, G., Mitchell, B.G., Kahru, M., Chavez, F.P., Cunningham, A., McKee, D., Gege, P., Kishino, M., Phinney, D.A., & Raine, R. (2002). Two models for absorption by coloured dissolved organic matter (CDOM). *Oceanologia, 44*, 209-241

Stedmon, C.A., & Markager, S. (2003). Behaviour of the optical properties of coloured dissolved organic matter under conservative mixing. *Estuarine, Coastal and Shelf Science, 57*, 973-979

Trochta, J.T., Mouw, C.B., & Moore, T.S. (2015). Remote sensing of physical cycles in Lake Superior using a spatio-temporal analysis of optical water typologies. *Remote Sensing of Environment, 171*, 149-161

Vandermeulen, R.A., Mannino, A., Neeley, A., Werdell, J., & Arnone, R. (2017). Determining the optimal spectral sampling frequency and uncertainty thresholds for hyperspectral remote sensing of ocean color. *Opt Express, 25*, A785-A797

Vantrepotte, V., Danhiez, F.P., Loisel, H., Ouillon, S., Meriaux, X., Cauvin, A., & Dessailly, D. (2015). CDOM-DOC relationship in contrasted coastal waters: implication for DOC retrieval from ocean color remote sensing observation. *Opt Express, 23*, 33-54

Wang, G., Lee, Z., Mishra, D.R., & Ma, R. (2016). Retrieving absorption coefficients of multiple phytoplankton pigments from hyperspectral remote sensing reflectance measured over cyanobacteria bloom waters. *Limnology and Oceanography: Methods, 14*, 432-447

Wang, G., Lee, Z., & Mouw, C. (2017). Multi-Spectral Remote Sensing of Phytoplankton Pigment Absorption Properties in Cyanobacteria Bloom Waters: A Regional Example in the Western Basin of Lake Erie. *Remote Sensing, 9*

Wang, G., Lee, Z., & Mouw, C. (2018). Concentrations of Multiple Phytoplankton Pigments in the Global Oceans Obtained from Satellite Ocean Color Measurements with MERIS. *Applied Sciences, 8*

Wang, P., Boss, E., & Roesler, C.S. (2005). Uncertainties of inherent optical properties obtained from semianalytical inversions of ocean color. *Appl Opt, 44*, 4074-4085

Werdell, P.J., Bailey, S.W., Fargion, G.S., Pietras, C., Knobelspiesse, K.D., Feldman, G.C., & McClain, C.R. (2003). Unique data repository facilitates ocean color satellite validation. *EOS Trans. AGU, 84*, 377

Werdell, P.J., McKinna, L.I.W., Boss, E., Ackleson, S.G., Craig, S.E., Gregg, W.W., Lee, Z., Maritorena, S., Roesler, C.S., Rousseaux, C.S., Stramski, D., Sullivan, J.M., Twardowski, M.S., Tzortziou, M., & Zhang, X. (2018). An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing. *Progress in Oceanography, 160*, 186-212

Werdell, P.J., Proctor, C.W., Boss, E., Leeuw, T., & Ouhssain, M. (2013). Underway sampling of marine inherent optical properties on the Tara Oceans expedition as a novel resource for ocean color satellite data product validation. *Methods in Oceanography, 7*, 40-51

Wetzels, R., & Wagenmakers, E.J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychon Bull Rev, 19*, 1057-1064

Xing, X., Morel, A., Claustre, H., D'Ortenzio, F., & Poteau, A. (2012). Combined processing and mutual interpretation of radiometry and fluorometry from autonomous profiling Bio-Argo floats: 2. Colored dissolved organic matter absorption retrieval. *Journal of Geophysical Research: Oceans, 117*

Zhao, Z., Gonsior, M., Luek, J., Timko, S., Ianiri, H., Hertkorn, N., Schmitt-Kopplin, P., Fang, X., Zeng, Q., & Jiao, N. (2017). Picocyanobacteria and deep-ocean fluorescent dissolved organic matter share similar optical properties. *Nature communications, 8*, 15284

Tables

Table 1.

Median and distribution of observed S_{dg} (1^{st} and 3^{rd} quartile) delineated by percent $a_{ph}(440)$ contribution. Relative accuracy of estimated S_{dg} is presented as the median and distribution of absolute difference (estimated S_{dg} – observed S_{dg}).

Observed S_{dg} (nm^{-1})				***Relative estimated S_{dg} accuracy (nm^{-1})***		
1^{st} quartile	**Median**	**3^{rd} quartile**	**$a_{ph}(440)$**	**1^{st} quartile**	**Median**	**3^{rd} quartile**
0.0146	0.0153	0.0161	**<10%**	-0.0003	-0.0001	-0.0001
0.0143	0.0165	0.0176	**10-20%**	-0.0015	-0.0010	+0.0004
0.0141	0.0156	0.0175	**20-30%**	-0.0010	-0.0001	+0.0013
0.0127	0.0142	0.0159	**30-40%**	-0.0015	+0.0001	+0.0017
0.0126	0.0140	0.0150	**40-50%**	-0.0024	-0.0005	+0.0016
0.0128	0.0146	0.0160	**50-60%**	-0.0018	-0.0003	+0.0021
0.0120	0.0138	0.0167	**60-70%**	-0.0024	-0.0007	+0.0005
0.0139	0.0191	0.0211	**>70%**	-0.0037	-0.0022	-0.0002

Figure Captions

1. Locations of spectra utilized in the (a) training dataset and (b) test dataset where color and size represent spectra grouped by varying $a_{ph}(440)$ percent contribution to total non-water absorption.
2. Schematic and figures illustrating primary steps for the Gaussian decomposition algorithm. This schematic is provided to aid in visualizing and organizing the steps detailed in the accompanying text (Section 2.2). Each figure illustrates the step as indicated for an example spectra. Not all spectra require all the steps depicted, while some spectra walk through all the steps (e.g., Fig. 2c shows a successful first guess, while some spectra required an iteration at this step).
3. Algorithm output for the example spectra depicted in Fig. 2. Gray dashed lines indicate the estimated (a) $a_{dg}(\lambda)$ and (b) $a_{ph}(\lambda)$ used as input into the least squares Gaussian decomposition of observed $a_{t-w}(\lambda)$ and black dashed lines indicate the respective observed IOP. For (a) and (b), respective colored lines display algorithm output. For (c), the brown line represents $a_{dg}(\lambda)$ algorithm output, the green line represents $a_{dg}(\lambda) + a_{ph}(\lambda)$ algorithm output and the black line with circles indicates observed $a_{t-w}(\lambda)$. This example shows how a Gaussian component can be fitted to the residuals derived from Step 5 (Fig. 2), but is minimized due to a better fit of observed $a_{t-w}(\lambda)$ with an exponential curve.
4. Performance metrics for each group delineated by $a_{ph}(440)$ percent contribution to total non-water absorption (indicated by color, from tan to dark green). Each plot corresponds to (a) percent retrievable $a_{ph}(\lambda)$, (b) percent retrievable $a_{dg}(\lambda)$, (c) $a_{ph}(\lambda)$ %NRMSD, (d) $a_{dg}(\lambda)$ %NRMSD, (e) $a_{ph}(\lambda)$ bias ($a_{dg}(\lambda)$ bias represented as inverse of each line) and (f)

mean absolute difference for both $a_{ph}(\lambda)$ and $a_{dg}(\lambda)$ (equivalent value by nature of the metric).

5. Mean performance of the algorithm for all test spectra within each group of spectra delineated by $a_{ph}(440)$ percent contribution to total non-water absorption relative to mean observed (a,c,e,g,i,k,m,o) $a_{ph}(\lambda)$ and (b,d,f,h,j,l,n,p) $a_{dg}(\lambda)$. The number of spectra within each group was: (a,b) n=286; (c,d) n=257; (e,f) n=303; (g,h) n=210; (i,j) n=146; (k,l) n=89; (m,n) n=34; (o,p) n=28.
6. (a) Unbiased absolute percent difference of S_{dg} for each grouping delineated by $\%a_{ph}(440)$, indicated by the color (see legend) and (b) distribution and relationship between observed and estimated S_{dg}, with marker color indicating $\%a_{ph}(440)$ and the dashed black line (--) representing the 1:1 line.
7. Distribution of identified peak locations for (a) observed $a_{ph}(\lambda)$ and (b) $a_{ph}(\lambda)$ estimated from $a_{t-w}(\lambda)$. Overall, identified peaks were quite consistent between the two signals displaying the strength of the scheme for initial estimates and constraints used for the Gaussian decomposition model.
8. A theoretical representation of varying spectral shape of $a_{t-w}(\lambda)$ under varying contributions of $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$. The base $a_{dg}(\lambda)$ and $a_{ph}(\lambda)$ spectra used for each curve are taken from measured spectra. We utilized this theoretical framework to develop the algorithm, namely understanding how changes in $a_{ph}(\lambda)$ percent contribution will inherently impact estimates of S_{dg}, how this inherent bias is impacted by wavelengths used and how to assess whether $a_{dg}(\lambda)$ has been accurately retrieved from $a_{t-w}(\lambda)$ free of an empirical relationship.
9. Relationship between $a_{t-w}(555)/a_{t-w}(680)$ and $a_{ph}(440)$ contribution for the (a) training dataset, where the piecewise exponential relationship from Eq. 3 and 4 is represented by

the red line, blue points indicate fitted data and gray points indicate values excluded from model fitting (r^2=0.91, RMSD=0.068). Outliers were defined as 1.5·1st /3rd quartile and were used to remove the influence of the large spread in data points with %a_{ph}(440) < 10, as these points represented nearly 25% of the dataset. (b) Test dataset points relative to the piecewise exponential relationship derived from the training dataset, displaying the primary weakness in empirical relationships (r^2=0.58, RMSD=0.110).