JGR Oceans

RESEARCH ARTICLE

10.1029/2021JC017231

Key Points:

- A reflectance line height metric was used as a predictor of the particulate backscattering coefficient at 555 nm
- The degree of overlap metric was used to correct validation skill metrics for measurement uncertainties in modeled and observed data
- Bland–Altman and zeta-score plots were explored as alternatives to one-to-one scatter plots routinely used for algorithm validation

Correspondence to:

L. I. W. McKinna,
lachlan.mckinna@go2q.com.au

Citation:

McKinna, L. I. W., Cetinić, I., & Werdell, P. J. (2021). Development and validation of an empirical ocean color algorithm with uncertainties: A case study with the particulate backscattering coefficient. *Journal of Geophysical Research: Oceans*, *126*, e2021JC017231. https://doi.org/10.1029/2021JC017231

Received 1 FEB 2021
Accepted 10 APR 2021

Development and Validation of an Empirical Ocean Color Algorithm with Uncertainties: A Case Study with the Particulate Backscattering Coefficient

Lachlan I. W. McKinna[1], Ivona Cetinić[2,3], and P. Jeremy Werdell[3]

[1]Go2Q Pty Ltd, Sunshine Coast, QLD, Australia, [2]GESTAR/USRA, Columbia, MD, USA, [3]NASA Goddard Flight Center, Greenbelt, MD, USA

Abstract We explored how algorithm (model) and *in situ* measurement (observation) uncertainties can effectively be incorporated into empirical ocean color model development and assessment. In this study we focused on methods for deriving the particulate backscattering coefficient at 555 nm, $b_{bp}(555)$ (m^{-1}). We developed a simple empirical algorithm for deriving $b_{bp}(555)$ as a function of a remote sensing reflectance line height (LH) metric. Model training was performed using a high-quality bio-optical dataset that contains coincident *in situ* measurements of the spectral remote sensing reflectances, $R_{rs}(\lambda)$ (sr^{-1}), and the spectral particulate backscattering coefficients, $b_{bp}(\lambda)$. The LH metric used is defined as the magnitude of $R_{rs}(555)$ relative to a linear baseline drawn between $R_{rs}(490)$ and $R_{rs}(670)$. Using an independent validation dataset, we compared the skill of the LH-based model with two other models. We used contemporary validation metrics, including bias and mean absolute error (MAE), that were corrected for model and observation uncertainties. The results demonstrated that measurement uncertainties do indeed impact contemporary validation metrics such as mean bias and MAE. Zeta-scores and *z*-tests for overlapping confidence intervals were also explored as potential methods for assessing model skill.

Plain Language Summary If we repeat a scientific measurement multiple times, we expect to record slightly different values each time. The average of these data is reported as the measurement and the spread of data either side of the average as the measurement uncertainty. With the knowledge of measurement uncertainty sources, such as a measurement sensor's internal instability, we can transfer the uncertainty through mathematical models. Satellite sensors measure light reflected from the ocean. The "ocean color" reflectance signal contains information about seawater optical properties. In this research, we explored how measurement uncertainties can be treated when ocean color models, that decipher the reflectance signal, are constructed and evaluated. In a case study, we developed an ocean color model to predict the optical particulate backscattering coefficient; a quantity that describes how particles scatter light in a backwards direction. In a process called validation we assessed the skill of our model by comparing model-estimated values with directly observed, or "sea-truth," values. We found that when measurement uncertainties were considered, the validation results changed. The line height model was also compared with two other existing methods and found to perform with similar skill in the open ocean and potentially better skill in murky coastal waters.

1. Introduction

Ocean color sensors measure spectral top-of-atmosphere radiances, $L_t(\lambda)$ (W m^{-2} sr^{-1} nm^{-1}), which are routinely separated into atmospheric and oceanic components using atmospheric correction (AC) algorithms (Frouin et al., 2019). The derived spectral water-leaving radiance signal, $L_w(\lambda)$ (W m^{-2} sr^{-1} nm^{-1}), in the visible domain (400–700 nm) is directly attributable to the types and relative concentrations of optically active matter present in the ocean's near-surface. For NASA's standard bio-optical algorithms, the radiometric quantity known as remote sensing reflectance signal, R_{rs} (sr^{-1}), is typically used as a model input. Where R_{rs} is defined as the ratio of the water-leaving radiance signal, L_w (W m^{-2} sr^{-1} nm^{-1}), to down-welling planar irradiance signal at the sea surface, E_d (W m^{-2} nm^{-1}).

A range of AC and bio-optical algorithms have been developed that allow marine geophysical parameters to be derived from sensor-observed radiometry. Over the last two decades, synoptic near-daily spatiotemporal observations collected by ocean color sensors have greatly improved our understanding of near-surface physical, biological, and biogeochemical oceanic processes. Indeed, some ocean color satellite-observed variables, such as chlorophyll *a* pigment concentration, *Chla* (mg m^{-3}), are now considered essential climate variables (Franz et al., 2017).

Most legacy ocean color algorithms used for deriving marine parameters such as *Chla*, a proxy for phytoplankton abundance, are typically empirical band ratio-type algorithms (O'Reilly & Werdell, 2019). Such algorithms rely on statistical relationships between the ratio of blue/green sensor bands and *in situ* measurements of *Chla*. Thus, from sensor-observed blue and green R_{rs} *Chla* can be quantified. Whilst such empirical *Chla* algorithms have mostly met mission accuracy objectives (McClain, 2009), they are best suited to oceanic waters, are not ubiquitously robust in optically complex (e.g., highly turbid and hypereutrophic) coastal and shelf waters, and can even be limited in oligotrophic waters (Hu et al., 2012). In such locations, alternative algorithms are necessary. Addressing this need are semi-analytical algorithms (SAAs) that make use of simplified radiative transfer theory as well as empiricism.

SAAs are radiative transfer-based and derive water-column optical properties directly from R_{rs} (Werdell et al., 2018). Once determined by an SAA, the total absorption, a (m^{-1}), and backscattering, b_b (m^{-1}), coefficients, collectively referred to as the inherent optical properties (IOPs), can be separated into optically distinct non-water sub-components (Werdell et al., 2018). NASA's standard SAA for deriving IOPs is the Generalized Inherent Optical Properties algorithm framework (GIOP) (Werdell et al., 2013) which has a modular structure thereby allowing end-users to select their own SAA parameterization. We note that a default configuration of the GIOP is used to produce NASA's standard IOP data products.

A number of key biogeochemical parameters used to study phytoplankton ecology, marine biogeochemistry, and ecosystem responses to climate change can be derived from IOPs. These so-called "IOP-based" data products depend on the accuracy of derived IOPs. One such parameter, particulate organic carbon (POC) (mg m^{-3}), can be used to study oceanic carbon fluxes and can be modeled as a function of the spectral particulate backscattering coefficient, $b_{bp}(\lambda)$ (m^{-1}) (Evers-King et al., 2017). In oligotrophic oceanic waters, where *Chla* < 0.05 mg m^{-3}, very low abundances of phytoplankton and sub-micron matter contribute significantly to $b_{bp}(\lambda)$ (Dall'Olmo et al., 2009; Stramski et al., 2004; Zhang et al., 2020). In these locations, SAA retrievals of $b_{bp}(\lambda)$ are often biased high (Lee & Huot, 2014) even when the corrections for inelastic Raman scattering are applied (McKinna et al., 2016). Sub-optimal $b_{bp}(\lambda)$ retrievals in oligotrophic gyres, which represent <40% of the global ocean, may impede the accuracy of b_{bp}-based models for estimating POC (Evers-King et al., 2017).

Several studies have demonstrated *Chla*-based empirical models for deriving $b_{bp}(\lambda)$ (Antoine et al., 2011; Brewin et al., 2012; Huot et al., 2008; Morel, 1988; Morel & Maritorena, 2001). This approach is particularly attractive for use in oligotrophic waters where SAA models can underperform. To utilize *Chla*-based $b_{bp}(\lambda)$ models first requires accurate satellite derivation of *Chla*, which can be challenging in oligotrophic waters where legacy band-ratio type algorithms perform sub-optimally. Hu et al. (2012) demonstrated that a three-band color-index (CI) difference metric, or reflectance line height (LH), -based approach to estimate *Chla* in oligotrophic waters is equally accurate to the blue-green band ratio models. Because the LH-based model is based on a reflectance difference, the approach is more robust to residual sunglint contamination, unknown errors from AC, and straylight contamination than reflectance ratio-based models (Hu et al., 2012, 2019). To reduce model complexity, we propose using a LH metric as an empirical predictor of $b_{bp}(\lambda)$ as opposed to using a *Chla*-based approach that requires one to estimate LH in an intermediate calculation. We note that Hu et al. (2012) mathematically showed that the magnitude of LH is more sensitive to changes in absorption in oligotrophic waters rather than $b_{bp}(\lambda)$. To that end, we will consider over what range of trophic conditions a LH-based $b_{bp}(\lambda)$ model is feasible and consider its expected limitations in phytoplankton-dominated, low *Chla*, oceanic waters.

Ocean color algorithms are routinely validated via "matchup" studies. These analyses are pair-wise comparisons of satellite-derived (M_i) with *in situ* observations (O_i) of the parameter of interest (e.g., *Chla* or b_{bp}). Aside from assessing a single algorithm's skill, matchup analyses can also be extended to inter-comparison studies that assist in algorithm selection (Brewin et al., 2015; Seegers et al., 2018). For continuous variables, commonly used matchup metrics include, but are not limited to, the coefficient of determination (R^2), type II linear regression metrics, mean bias, and mean absolute error (MAE). As oceanic biogeochemical variables predominantly follow log-normal distributions (Campbell, 1995), validation metrics are often calculated using $\log_{10}$-transformed data. Recently, Seegers et al. (2018) suggested that mean bias and MAE were robust skill assessment metrics and were adopted by NASA's OB.DAAC for standard ocean color data product validation (https://seabass.gsfc.nasa.gov/).

Uncertainties have traditionally been overlooked during ocean color algorithm development. However, the ocean color community does recognize the importance of model and *in situ* observation uncertainty provenance and has recently provided detailed guidance on the topic (IOCCG, 2020). Nonetheless, uncertainties are rarely considered during model validation. This is likely due to previously limited knowledge of model and observation uncertainties. In other disciplines, such as watershed and climate modeling, progress has been made toward incorporating model and observation uncertainties into model skill assessment (Eyring et al., 2019; Harmel et al., 2010). As we continue to improve our understanding of satellite sensor and *in situ* observation uncertainties it is critical that: (i) ocean color algorithms with empirical aspects account for uncertainties in the data sets used to train the model, (ii) algorithms are capable of estimating derived product uncertainties, and (iii) we develop techniques that consider uncertainties during model validation analyses. Here, we use our LH-based $b_{bp}(\lambda)$ empirical modeling exercise as a case study to demonstrate how uncertainties can be incorporated into model development, validation, and inter-comparison.

The objective of this study was twofold: (i) determine if an empirical LH-based model can be used to derive $b_{bp}(555)$ and associated standard uncertainties $u(b_{bp}(555))$ and (ii) explore how measurement uncertainties might be used in ocean color algorithm validation. We perform exploratory analysis and model development using two in situ datasets: the DS3 dataset (Stramski & Reynolds, 2018) and the Ocean Color Climate Change Initiative (OC-CCI) dataset (Valente et al., 2019). Specifically, we use the DS3 dataset to train the LH-based model and then the OC-CCI dataset is used to validate it. The skill of the GIOP and the Huot et al. (2008) *Chla*-based model are also assessed using the OC-CCI dataset. In our validation studies, we correct difference metrics for measurement uncertainty and consider how one might use skill assessment metrics to guide algorithm selection.

2. Data and Methods

2.1. Reflectance LH Metric

Reflectance line height metrics quantify the magnitude of a sensor-observed radiometric observation (e.g., R_{rs}) at a given band relative to a linear baseline interpolated between two adjacent bands. Some LH metrics used in ocean color remote sensing include, but are not limited to, the maximum chlorophyll index (MCI) (Gower et al., 2008), the normalized fluorescent line height (Behrenfeld et al., 2009), the floating algae index (Hu, 2009), the cyanobacteria index (Lunetta et al., 2015), the maximum peak height (Matthews & Odermatt, 2015), the color difference metric (Mitchell et al., 2017), and the CI (Hu et al., 2012).

A LH metric can generally be expressed as:

$$\text{LH} = R_{rs}\left(\lambda_g\right) - \left[R_{rs}\left(\lambda_b\right) + \frac{\lambda_g - \lambda_b}{\lambda_r - \lambda_b}\left(R_{rs}\left(\lambda_r\right) - R_{rs}\left(\lambda_b\right)\right)\right], \tag{1}$$

where λ_b, λ_g, and λ_r denotes band-center wavelengths of sensor-specific blue, green, and red sensor bands, respectively. The term inside the square bracket is a linear interpolation between λ_b and λ_r. We note that the LH metric used in Hu et al. (2012), for example, was developed for NASA's Sea-viewing Wide Field-of-view Sensor (SeaWiFS) in which case λ_b, λ_g, and λ_r correspond to 443, 555, and 670 nm, respectively.

Following McKinna et al. (2019), we have used the first-order first-moment formulations to estimate uncertainties due to random radiometric error. This approach is valid when the uncertainty is small relative to the measurement. For the LH metric, our estimated uncertainty, u(LH), is calculated as follows:

$$u\left(LH\right)=\left\{u\left(R_{rs}\left(\lambda_b\right)\right)^2+\left[\frac{\lambda_g-\lambda_b}{\lambda_g-\lambda_b}-1\right]^2u\left(R_{rs}\left(\lambda_b\right)\right)^2+\left[-\frac{\lambda_g-\lambda_b}{\lambda_g-\lambda_b}\right]^2u\left(R_{rs}\left(\lambda_r\right)\right)^2\right.$$
$$+2\left[\frac{\lambda_g-\lambda_b}{\lambda_g-\lambda_b}-1\right]u\left(R_{rs}\left(\lambda_g\right),R_{rs}\left(\lambda_b\right)\right)+2\left[-\frac{\lambda_g-\lambda_b}{\lambda_g-\lambda_b}\right]u\left(R_{rs}\left(\lambda_g\right),R_{rs}\left(\lambda_r\right)\right) \tag{2}$$
$$\left.+2\left[\frac{\lambda_g-\lambda_b}{\lambda_g-\lambda_b}-1\right]\left[-\frac{\lambda_g-\lambda_b}{\lambda_g-\lambda_b}\right]u\left(R_{rs}\left(\lambda_b\right),R_{rs}\left(\lambda_r\right)\right)\right\}^{\frac{1}{2}},$$

where $u(R_{rs}(\lambda))^2$ is the variance in the remote sensing reflectance for a given sensor band. The covariance terms for the i^{th} and j^{th} remote sensing reflectances are denoted as $u(R_{rs}(\lambda_i),R_{rs}(\lambda_j))$. Note, in this study we have not included the covariances terms in our LH uncertainty estimates as these are unknown, however, we acknowledge that they have important implications for estimating data product uncertainties (Lamquin et al., 2013; McKinna et al., 2019).

2.2. Model Development Dataset

2.2.1. DS3 Dataset

We used the DS3 Ocean Optics Dataset (Stramski & Reynolds, 2018) for model development. The DS3 dataset comprises *in situ* IOP and radiometry measurements collected and processed in a consistent manner by a single institute, the Scripps Institute of Oceanography, and was previously used in the development of the LS2 inverse bio-optical model (Loisel et al., 2018). The DS3 contains 243 data records (rows) and is well-suited for bio-optical model development as it is representative of a range of oceanic conditions including very clear waters of the South Pacific Gyre. We note that all b_{bp} data in DS3 were collected with HOBI Labs HydroScat fixed-angle volume scattering function meters. Spectral dependency for R_{rs} and b_{bp} is hereafter implied.

Each spectral R_{rs} record in DS3 has six data fields corresponding to six SeaWiFS spectral bands centered on 412, 443, 490, 510, 555, and 670 nm, respectively. The DS3 dataset has four spectral b_{bp} data fields corresponding to 442, 510, 550, and 671 nm, respectively. During model development, data records were excluded if the LH value could not be calculated (i.e., where one or more of the required R_{rs} data fields were missing). Similarly, data records were excluded if less than three valid spectral b_{bp} fields were present. Where there were three or more valid spectral b_{bp} records present, a curve was fit through the data in log-linear space using a power law model of the form:

$$b_{bp}\left(\lambda\right)=b_{bp}\left(\lambda_0\right)\left(\frac{\lambda}{\lambda_0}\right)^{\gamma}. \tag{3}$$

From the model fit, values of $b_{bp}(555)$ were derived as well as the spectral slope coefficient, γ.

2.2.2. OC-CCI Dataset

For model validation, we used the ESA OC-CCI bio-optical dataset (Valente et al., 2019). The OC-CCI is a large merged dataset that comprises 143,935 *in situ* data records, including the NASA bio-Optical Marine Algorithm Dataset (NOMAD; Werdell & Bailey, 2005). We note that not all records have coincident R_{rs} and b_{bp} measurements. The dataset encompasses a wide variety of optical conditions and has spectral R_{rs} data fields that are consistent with several ocean color sensors including the Medium Resolution Imaging Spectroradiometer, the Moderate Resolution Imaging Spectroradiometer (MODIS), the Visible Infrared Imaging Radiometer Suite, the Ocean and Land Color Instrument (OLCI), and SeaWiFS. Unlike the DS3 dataset, the sensors used to measure b_{bp} are varied and are sourced from multiple institutes. To maintain its independence from the training data, the OC-CCI dataset was screened and any data records present in the DS3 dataset were removed.

We used the "satbands6" tables of the OC-CCI dataset meaning that the closest R_{rs} and b_{bp} spectra measured within ±6 nm of the SeaWiFS band centers. A total of 340 OC-CCI data records were available for model validation. We used Equation 3 to fit a power law model through OC-CCI b_{bp} records where three or more valid spectral measurements were available.

2.2.3. Satellite Data

We demonstrate the LH-based model with SeaWiFS imagery of two different regions: (i) the North Pacific Ocean adjacent to Hawaii and (ii) the Chesapeake Bay, USA. The first region is characterized by oligotrophic, low scattering waters, while the second region is characterized as optically complex, with highly scattering waters. The image of the Hawaiian region was captured on December 1, 2000, and the image of the Chesapeake Bay region was captured on April 23, 2003. SeaWiFS level-1 files were downloaded from NASA OB.DAAC (NASA Goddard Space Flight Center, 2010) and processed using the l2gen module of NASA's Ocean Color Science Software (https://oceandata.sci.gsfc.nasa.gov/ocssw/). NASA's standard AC was applied and the following level-2 data products were produced: R_{rs}(490), R_{rs}(555), R_{rs}(670), *Chla*, and b_{bp}(555). Rayleigh-corrected reflectances, used for generating quasi-true color images, were also produced at 490, 555, and 670 nm. *Chla* was derived using the standard NASA algorithm (Hu et al., 2012; O'Reilly et al., 1998) and b_{bp}(555) was derived using the default configuration of the GIOP algorithm (Werdell et al., 2013) with the empirical Raman scattering correction of Lee et al. (2013) applied.

Data visualization and analysis were performed using NASA's SeaWiFS Data and Analysis Software package (SeaDAS; https://seadas.gsfc.nasa.gov/). Quasi-true color images of each region were generated from Rayleigh-corrected reflectances using SeaDAS' built-in RGB functions. For our comparisons, we did not reproject/map L2 images.

2.3. Model Fitting

We used Python 3.7.0 for model development. For curve fitting, we selected orthogonal distance regression (ODR) as distributed in Python's Scientific library (SciPy). We selected ODR (a type-II regression method), as opposed the more traditional ordinary least squares because it considers measurement uncertainty in both the dependent and independent variables. After exploratory analyses of the DS3 dataset, we decided to model b_{bp}(555) as a log-linear function of LH:

$$b_{bp}\left(555\right) = 10^{a_0 + a_1 \mathrm{LH}} \tag{4}$$

where the unknown coefficients a_0 and a_1 were determined by bootstrapped ODR curve fitting. Incidentally, this model is of similar mathematical form as the Hu et al. (2012) LH-based *Chla* model.

Standard uncertainties in a_0 and a_1, denoted as $u(a_0)$ and $u(a_1)$, respectively, were estimated using bootstrap curve fitting. Specifically, 80% of the DS3 dataset was randomly selected and ODR curve fitting was performed to derive a_0 and a_1. This process was repeated 1,000 times to generate distributions of a_0 and a_1 from which the mean was computed. From the covariance matrix of a_0 and a_1, standard deviations (i.e., the standard uncertainties $u(a_0)$ and $u(a_1)$) and covariance term $u(a_0, a_1)$ were estimated.

From Equation 4, we estimated the uncertainty in derived b_{bp}(555) as:

$$\begin{aligned} u\left(b_{bp}\left(555\right)\right) \; = \; & \Big\{\left[\ln\left(10\right)u\left(a_0\right)10^{\nu}\right]^2 + \left[\mathrm{LH}\ln\left(10\right)u\left(a_1\right)10^{\nu}\right]^2 + \left[u\left(\mathrm{LH}\right)a_1\ln\left(10\right)10^{\nu}\right]^2 \\ & +2u\left(a_0, a_1\right)\mathrm{LH}\left[\ln\left(10\right)10^{\nu}\right]^2\Big\}^{1/2} \end{aligned} \tag{5}$$

where $\nu = a_0 + a_1$LH and u(LH) is computed using Equation 2.

2.4. Uncertainties

Historically, *in situ* measurements do not always have accompanying uncertainty estimates and for this study we have made assumptions about the standard uncertainties in R_{rs} and b_{bp}(555). We assumed 5% relative standard uncertainty in DS3 and OC-CCI spectral R_{rs} values (IOCCG Protocol Series, 2019) and

5% relative standard uncertainty in b_{bp} measurements due to calibration uncertainty (Sullivan et al., 2013). From these relative uncertainties, we computed the standard uncertainty for each quantity. Standard uncertainties in GIOP-derived $b_{bp}(555)$ were estimated following McKinna et al. (2019) while standard uncertainties in Huot-derived $b_{bp}(555)$ were also estimated using a first-order analytical methodology (see Appendix A for detail).

We note these relative uncertainties may be somewhat optimistic. Indeed, fixed angle volume scattering function meters have been reported as having larger, spectrally dependent uncertainties greater than 5% (Dall'Olmo et al., 2009) and McKee et al. (2009) reported b_{bp} uncertainties that do not scale with magnitude. However, we believe 5% is still a useful starting point to explore how uncertainties might be considered in model skill assessment (validation) metrics. In a similar style to work of McKinna et al. (2019), the model skill assessments we present may be repeated or expanded to other models provided one has reasonable knowledge (or estimate) of observation and model uncertainties.

2.5. Model Skill Assessment Metrics

To evaluate the predictive skill of our model(s), we compared model-derived $b_{bp}(555)$ with *in situ* observed values. This approach is also referred to as "model validation." Our model validation was conducted using the OC-CCI dataset. Typically, in ocean color remote sensing, linear regression statistics are reported such as R^2, slope, intercept, and root mean squared error. However, Seegers et al. (2018) demonstrated that pairwise comparison metrics such as the mean bias and mean absolute error (MAE) are robust model assessment metrics, particularly when working with datasets that do not follow Gaussian distribution and have outliers present.

In this study, we computed the mean bias and MAE as follows:

$$\text{bias} = \frac{1}{N}\sum_{i=1}^{N} D_i \tag{6}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left|D_i\right| \tag{7}$$

where M_i and O_i are the i^{th} modeled (derived) and observed (*in situ*) data points, respectively, and the difference, D_i, is equal to $M_i – O_i$. We also computed these metrics for $\log_{10}$-transformed data, following Seegers et al. (2018), as ocean color datasets are often log-normally distributed (Campbell, 1995):

$$\text{bias}_{\text{log}} = 10^{\left\{\frac{1}{N}\sum_{i=1}^{N} M_i' - O_i'\right\}} \tag{8}$$

$$\text{MAE}_{\text{log}} = 10^{\left\{\frac{1}{N}\sum_{i=1}^{N} \left|M_i' - O_i'\right|\right\}}. \tag{9}$$

where $M_i' = \log_{10}\left(M_i\right)$ and $O_i' = \log_{10}\left(O_i\right)$. Note, for some calculations the standard uncertainty of $\log_{10}$-transformed data was required. We denote modeled and observed $\log_{10}$-transformed standard uncertainties as $u\left(M_i'\right)$ and $u\left(O_i'\right)$, respectively and were estimated as:

$$u\left(M_i'\right) = \frac{u\left(M_i\right)}{\ln\left(10\right)M_i} \tag{10}$$

$$u\left(O_i'\right) = \frac{u\left(O_i\right)}{\ln\left(10\right)O_i}. \tag{11}$$

One benefit of using $\log_{10}$-transformed mean bias and MAE is that the metrics have been transformed from linear to multiplicative space. For example, a $\log_{10}$-transformed mean bias value of 1.1 means the model on average overestimates by 10% relative to *in situ* measurements. To facilitate historical comparisons, we present data in scatter plots and report slope, intercept, and R^2 linear regression statistics in $\log_{10}$-space using reduced major axis (RMA) regression.

2.6. Incorporating Model and Observation Uncertainties

We explore three ways one might incorporate uncertainties into model skill assessment (validation): (i) independent pair parametric testing based on confidence intervals, (ii) corrected skill metrics based on confidence interval overlap, and (iii) zeta-scores.

2.6.1. Pairwise Independent Sample Z-Testing

We assume that the i^{th} modeled and observed values of $b_{bp}(555)$ have means of M_i and O_i, respectively, that are normally distributed with known standard errors (i.e., the standard uncertainties) of $u(M_i)$ and $u(O_i)$. We performed two-tailed z-tests for independent samples with a null hypothesis H_0: $M_i = O_i$ and alternative hypothesis H_a: $M_i \neq O_i$ at the significance level $\alpha = 0.01$. We tallied the proportion of all M_i *and* O_i pairs where the null hypothesis was accepted.

Extending the formulation presented in Austin and Hux (2002), we can express the i^{th} *z-test* metric as:

$$z_{test,i} = \frac{2.576\left[DO_c\left(u\left(M_i\right)+u\left(O_i\right)\right)-\sqrt{u\left(M_i\right)^2+u\left(O_i\right)^2}\right]}{\left[u\left(M_i\right)+u\left(O_i\right)\right]} \tag{12}$$

where DO_c, is the critical degree of overlap of the two 99% confidence intervals. If the actual degree of overlap is less than DO_c then the null hypothesis is rejected. Values of DO_c must be computed for each pair of M_i and O_i. As an example, if $u(M_i) = 0.00095\ \mathrm{m}^{-1}$ and $u(O_i) = 0.00035\ \mathrm{m}^{-1}$, then DO_c would be 0.22. Detail on how to compute the actual degree of overlap is given next.

2.6.2. Degree of Overlap

Let us consider that the i^{th} pair of modeled and observed $b_{bp}(555)$ data points, M_i and O_i, represent the mean of the probability distribution functions $p_m(m_i)$ and $p_o(o_i)$, respectively, whose dispersion is described by the standard uncertainties $u(M_i)$ and $u(\mathrm{O}_i)$, respectively. The degree of overlap (DO_i) of $p_m(m_i)$ and $p_o(o_i)$ can be expressed as per Equation 7 in Harmel et al. (2010) as:

$$DO_i = \int_{M_{i,min}}^{M_{i,max}} p_m\left(m_i\right)dm \cdot \int_{O_{i,min}}^{O_{i,max}} p_o\left(o_i\right)do \tag{13}$$

$$DO_i = \left[prob\left(o_i < M_{i,max}\right) - prob\left(o_i < M_{i,min}\right)\right]\cdot\left[prob\left(m_i < O_{i,max}\right) - prob\left(m_i < O_{i,min}\right)\right]. \tag{14}$$

where the $M_{i,min}$ and $M_{i,max}$ represent the uncertainty (or confidence) boundaries for $p_m(m_i)$ and $O_{i,min}$ and $O_{i,max}$ represent the uncertainty boundaries for $p_o(o_i)$. These lower and upper boundaries are user defined and may, for example, be set to 0.05 and 0.95 for a 90% confidence level.

To calculate DO_i for each O_i and M_i pair, we used Python 3.7.0 code and the SciPy scientific and engineering package. First, the values $M_{i,min}$, $M_{i,max}$, $O_{i,min}$, and $O_{i,max}$ were computed with the function *scipy.stats.norm.pdf* for a given uncertainty boundary. Next, the *scipy.stats.norm.cdf* function was used to compute the probabilities in Equation 14. We note that both functions required a mean and standard deviation as inputs. For model and observation data we used M_i and $u(M_i)$, and O_i and $u(O_i)$, respectively. For $\log_{10}$-transformed model and observation data we used $M_i^{'}$ and $u\left(M_i^{'}\right)$, and $O_i^{'}$ and $u\left(O_i^{'}\right)$, respectively.

2.6.3. Corrected Difference Metrics

To account for uncertainties in pair-wise comparisons of M_i and O_i, we followed the method of Harmel et al. (2010) and computed a correction factor, CF_i:

$$CF_i = 1 - DO_i. \tag{15}$$

The corrected pair-wise difference, CD_i was then calculated as:

$$CD_i = CF_i \times D_i. \tag{16}$$

Corrected mean bias and mean absolute error was next calculated as:

$$\text{bias} = \frac{1}{N}\sum_{i=1}^{N} CD_i \tag{17}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left| CD_i \right| \tag{18}$$

Less weight is applied to D_i when DO_i approaches 1. Essentially, for completely overlapping $p_m(m_i)$ and $p_o(o_i)$, where $DO_i = 1$, the value of CF_i will be zero as statistically no difference can be discerned between the two overlapping probability distribution functions.

Corrected mean bias and MAE for log10-transformed data was calculated as:

$$\text{bias}_{\log} = 10^{\left\{\frac{1}{N}\sum_{i=1}^{N} CF_i\left(M_i' - O_i'\right)\right\}}, \tag{19}$$

$$\text{MAE}_{\log} = 10^{\left\{\frac{1}{N}\sum_{i=1}^{N} \left|CF_i\left(M_i' - O_i'\right)\right|\right\}}. \tag{20}$$

2.6.4. Zeta-Scores and Bland–Altman Plots

Bland–Altman plots are useful for comparing agreement between M_i and O_i (Bland & Altman, 1986). The Bland–Altman plot is a scatter plot with D_i on the vertical axis and the average of M_i and O_i on the horizontal axis. A statistical confidence region (e.g., 95% confidence interval) for D_i is also usually plotted. Recently, the aerosol remote sensing community has demonstrated Bland–Altman plots as a useful tool for visualizing sensor-to-sensor evaluations (Fu et al., 2020; Knobelspiesse et al., 2019). Similar to Knobelspiesse et al. (2019), we explored a modified Bland–Altman-type plot where D_i is normalized by model and observation uncertainties to give the zeta-score metric. The i^{th} zeta-score, ζ_i, is computed as:

$$\zeta_i = \frac{D_i}{\sqrt{u\left(M_i\right)^2 + u\left(O_i\right)^2}}. \tag{21}$$

For comparing methods, values of $|\zeta| \leq 2$ are considered satisfactory, $2 \leq |\zeta| \leq 3$ are considered questionable, and $|\zeta| \geq 3$ should be considered as unsatisfactory (Analytical Methods Committee Amctb No. 74, 2016). When plotting zeta scores in a Bland–Altman style, we color-coded the aforementioned regions in a traffic light (green-orange-red) style to assist with interpretation.

We also explored Bland–Altman plots with corrected differences, as per Equation 16, and also corrected zeta scores, ζ_i', plots where the score is computed as:

$$\zeta_i' = \frac{CF_i D_i}{\sqrt{u\left(M_i\right)^2 + u\left(O_i\right)^2}}. \tag{22}$$

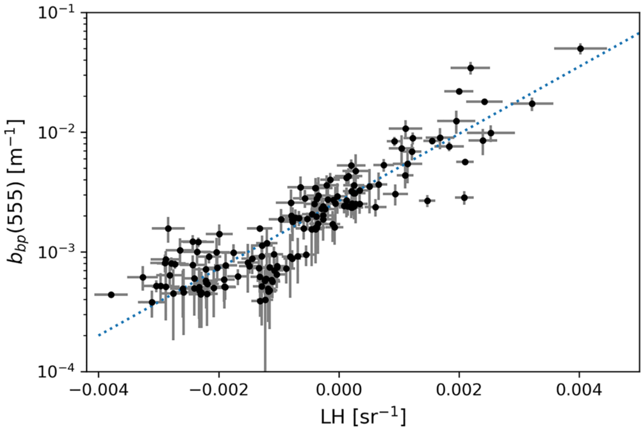

Figure 1. Scatter plot of LH versus $b_{bp}(555)$ ($N = 153$) shows a strong log-linear relationship. Gray bars represent estimated standard uncertainties.

3. Results and Discussion

3.1. LH-Based Model Fit

Using the DS3 dataset, we found there to be a strong log-linear relationship between LH and $b_{bp}(555)$ when λ_b, λ_g, and λ_r were set to 490, 555, and 670 nm, respectively. This can be visualized in Figure 1 and quantified with a R^2 of 0.88. Using bootstrap model fitting, we determined that the best fit model coefficients for Equation 4 were $a_0 = -2.5770$ and $a_1 = 281.27$, with associated standard uncertainties of $u(a_0) = 2.4819 \times 10^{-2}$ and $u(a_1) = 20.777$, and a covariance term of $u(a_0, a_1) = 0.24852$. We note that in Figure 1 a small cluster of 11 data points fell below the best fit line (between LH values of −0.002 and 0). We found that these corresponded to nine MALINA cruise stations 9, 10, 21,23, 25, 45, and 50, and three ICESCAPE 2011 stations 23, 24 and 32 which were all sampled in Artic waters of the Beaufort Sea.

Using the bootstrap resampling approach, we also generated cross-validated model validation statistics. These statistics are summarized in Table 1 and indicated that the model performed with good predictive skill with a R^2 of 0.787, slope of 1.08, a positive bias of 4%, and a mean absolute error 47%. We next used the separate OC-CCI dataset to further evaluate model skill.

3.2. Scatter Plots and Validation Metrics

We used the OC-CCI dataset to validate the LH-based model. For comparative purposes, we also derived $b_{bp}(555)$ using the GIOP model and the empirical *Chla*-based model of Huot et al. (2008) where *Chla* and its standard uncertainty, $u(Chla)$, were derived first as an intermediate product with NASA's standard empirical algorithm (Hu et al., 2012; O'Reilly & Werdell, 2019). We hereby refer to the Huot et al. (2008) model as “Huot.”

The scatter plots shown in Figure 2 are a common tool used to visually interpret ocean color algorithm predictive skill. Over the full dynamic range, the scatter plots indicate that model-derived $b_{bp}(555)$ values agree reasonably well with *in situ* observed values. However, when observed $b_{bp}(555) < 0.00125$ m^{-1}, the scatter plots indicate that the LH and GIOP models overestimated whereas the Huot model showed much better agreement with observed values. Conversely, when observed $b_{bp}(555) \geq 0.00125$ m^{-1} the GIOP and LH models showed good agreement with observed values whereas the Huot model tended to underestimate. Visually, the GIOP approach appears to be a better predictor of $b_{bp}(555)$ over the full dynamic range.

Table 2 displays validation metrics for the LH, GIOP, and Huot models. We computed these statistics for the full dataset ($N = 326$) and two arbitrary subsets. The first subset, referred to as the “low-value” subset ($N = 60$), was partitioned based on O_i values of $b_{bp}(555) < 0.00125$ m^{-1}. The second subset, referred to as the “high-value” subset ($N = 266$), was partitioned where O_i values of $b_{bp}(555) \geq 0.00125$ m^{-1}. We computed bias, MAE, bias$_{\log}$, and MAE$_{\log}$ using both standard and corrected differences. Metrics in Table 2 with a prime (′) symbol indicate they were computed with correction factors applied to account for uncertainties in both measured and observed quantities. The final column in Table 2 is a tally of the “ wins” to assist with

Table 1
Cross-Validation Results for the LH-Based Model for $b_{bp}(555)$

DS3 median (m^{-1})	DS3 std (m^{-1})	DS3 range (m^{-1})	R^2*	Slope*	Bias (m^{-1})	MAE (m^{-1})	bias$_{\log}$ (unitless)	MAE$_{\log}$ (unitless)
0.00158	5.61×10^{-3}	3.79×10^{-4}–4.98×10^{-2}	0.787 (0.0342)	1.08 (0.122)	−2.63 × 10−4 (4.59 × 10−4)	0.0116 (2.61 × 10−4)	1.04 (0.0896)	1.47 (0.0432)

Note. The mean and standard deviation of each bootstrapped validation metric distribution is reported. Standard deviations are in parentheses. To contextualize the bias metrics, the mean and range of $b_{bp}(555)$ from the DS3 dataset are reported. *Computed in $\log_{10}$–$\log_{10}$ space.

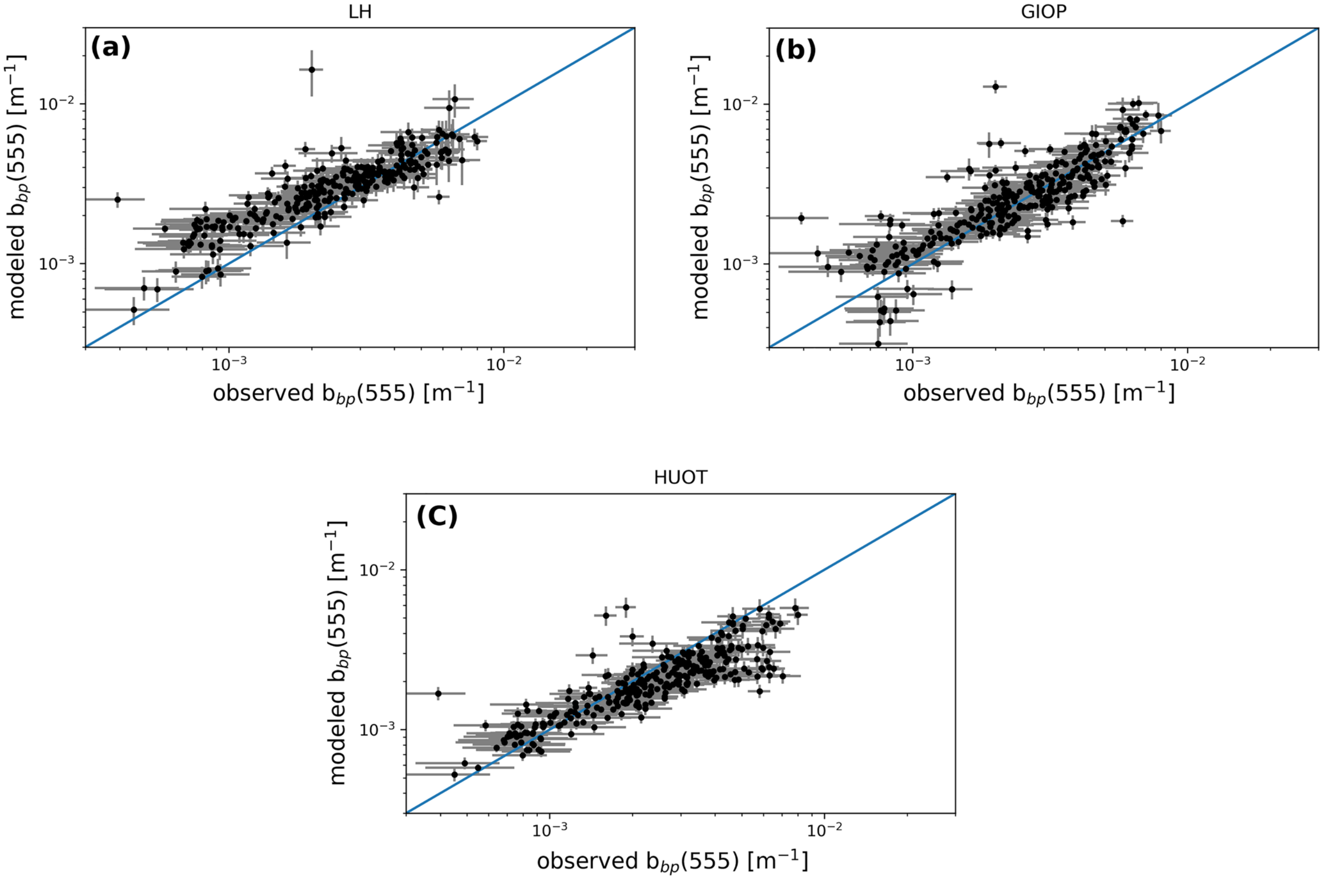

Figure 2. Scatter plots comparing $b_{bp}(555)$ derived from radiometry using a model with *in situ* measurements. Subplots (a)–(c) correspond to the LH, GIOP, and Huot models, respectively.

comparing the LH, GIOP, and Huot models. We define a "win" as the best inter-model performance for a given validation metric category.

Results for "All data" in Table 2 show that the three models performed with similar predictive skill. The GIOP was considered "best" with 10 wins and outperformed the LH and Huot models. Based on the MAE

Table 2
Model Difference Statistics Comparing Three Models: LH-Based Model, GIOP, and Huot

$b_{bp}(555)$ range	Model	N	R^2*	Slope*	Bias (m^{-1})	Bias′ (m^{-1})	MAE (m^{-1})	MAE′ (m^{-1})	bias$_{\log}$ (unitless)	bias′$_{\log}$ (unitless)	MAE$_{\log}$ (unitless)	MAE′$_{\log}$ (unitless)	No. wins
All data	LH	326	0.730	1.35	3.90×10^{-4}	2.61×10^{-4}	8.01×10^{-4}	3.96×10^{-4}	1.21	1.12	1.33	1.16	0
	GIOP	326	**0.733**	**1.04**	$\mathbf{1.52 \times 10^{-4}}$	$\mathbf{9.51 \times 10^{-5}}$	$\mathbf{6.75 \times 10^{-4}}$	$\mathbf{3.86 \times 10^{-4}}$	**1.06**	**1.03**	**1.27**	**1.15**	**10**
	Huot	326	0.699	1.40	-7.38×10^{-4}	-5.49×10^{-4}	9.15×10^{-4}	6.59×10^{-4}	0.812	0.834	1.37	1.27	0
<1.25E−3 m^{-1}	LH	60	0.225	0.764	6.72×10^{-4}	5.19×10^{-4}	6.75×10^{-4}	5.19×10^{-4}	1.73	1.55	1.73	1.55	0
	GIOP	60	0.049	0.614	2.65×10^{-4}	1.77×10^{-4}	3.81×10^{-4}	2.30×10^{-4}	1.24	1.15	1.48	1.29	0
	Huot	60	**0.235**	**1.01**	$\mathbf{1.51 \times 10^{-4}}$	$\mathbf{1.47 \times 10^{-4}}$	$\mathbf{1.96 \times 10^{-4}}$	$\mathbf{1.52 \times 10^{-4}}$	**1.17**	**1.09**	**1.23**	**1.10**	**10**
≥1.25E−3 m^{-1}	LH	266	0.548	1.24	3.25×10^{-4}	2.03×10^{-4}	8.29×10^{-4}	3.69×10^{-4}	1.12	1.05	1.26	**1.09**	1
	GIOP	266	**0.602**	**0.947**	$\mathbf{1.27 \times 10^{-4}}$	$\mathbf{7.73 \times 10^{-5}}$	$\mathbf{7.41 \times 10^{-4}}$	4.20×10^{-4}	**1.02**	**1.00**	**1.24**	1.12	**8**
	Huot	266	0.448	1.28	-9.39×10^{-4}	-1.38×10^{-4}	1.08×10^{-3}	**2.64E−4**	0.748	0.786	1.41	1.31	1

Note. Bold text indicates best performance for each skill metric. No. wins (last column) indicates number of statistical tests in which respective dataset outperformed others. *Computed in log$_{10}$–log$_{10}$ space. Difference metrics with correction factor applied.

and $\mathrm{MAE}_{\log}$ metrics, the empirical LH and Huot models perform similarly. However, based on the bias and $\mathrm{bias}_{\log}$ metrics, LH-derived values are on average overestimated by 21%, while the Huot-derived values are underestimated by 19%. It is important to note that the correction factor had a noticeable effect on the skill metrics. For example, the LH model's $\mathrm{MAE}_{\log}$ value was 1.33 and the corresponding corrected value, $\mathrm{MAE}'_{\log}$, was 1.16.

The low-value subset metrics indicated that the *Chla*-based Huot model performed better than LH and GIOP with 10 wins. This is not surprising given that the *Chla*-based Huot model was developed using *in situ* data collected in the South Pacific Gyre, an area considered to have the "clearest" oceanic waters (Huot et al., 2008; Morel et al., 2007). This also suggests that the Hu et al. (2012) model is performing well in context of deriving oligotrophic *Chla* as an intermediate product needed as an input to the Huot model. The high-value subset metrics indicated that the GIOP performed better than the LH and Huot models with eight wins. We note that the LH model had bias and MAE metrics, including $\log_{10}$-scaled and corrected values, similar to the GIOP model. This result is encouraging given the relative simplicity of the LH model compared with the more mathematically complex GIOP model.

Figure 3 shows Bland–Altman-type scatter plots for the LH, GIOP, and Huot models. Panels on the left-hand side of Figure 3 show the uncorrected difference between model and observed $b_{bp}(555)$, D_i, on the y-axis and the method average value on the x-axis. Panels on the right-hand side of Figure 3 show corrected D_i scaled by CF_i. The plots of uncorrected D_i show that the LH model typically overestimates $b_{bp}(555)$ values less than 0.002 m^{-1} with most D_i values lying inside the 97.5% confidence interval. However, when we look at the corrected Bland–Altman plot for LH (Figure 3d), the model appears to have much better skill with many more data points in the plot falling closer to zero. We see the same effect for the GIOP and Huot models. Notably, both Bland–Altman plots (Figures 2c and 2f) show signs of the Huot model underestimating larger values of $b_{bp}(555)$, which is consistent with Figure 2 and statistics in Table 2.

3.3. Zeta Score Plots

Zeta score plots are shown in Figure 4. The left-hand panel are standard zeta scores while the right-hand side are zeta scores computed with corrected D_i values. We have color coded the plots, green-yellow-red, to assist in visualizing where acceptable, questionable, and poor agreement occur, respectively. Uncorrected zeta scores in Figures 4a–4c generally show that the majority of D_i values fall within the green zone, meaning they are acceptable. Upon careful inspection we note that D_i values are mostly greater than zero for LH, seem distributed evenly about zero for GIOP, and often less than zero for Huot. This pattern remains, to a lesser extent, in plots of corrected D_i values (Figures 4d–4f).

Table 3 shows summary statistics of zeta scores. Tallies of how many zeta scores, both corrected and uncorrected, fall within the green, yellow, and red regions are also given. Performance was judged best when zeta scores are close to zero and fall mostly within the green zone. Similar to Table 2, we consider zeta scores for the entire dataset, the low-value subset, and the high-value subset. The statistics for all data indicate that the GIOP performs best with 10 wins. For the low-value subset, the Huot model performs best with 8 wins and the LH model narrowly outperforms the GIOP for the high-value subset.

This brief example demonstrates how zeta score plots might complement existing linear regression, mean bias, and MAE metrics used in ocean color validation studies if model and observation standard uncertainties are known. Of particular benefit is their ease of interpretability with the "traffic light" color-coded plots. The tallied "wins" in Table 3 are similar to those in Table 2 for all data (GIOP performs best) and for the low-value subset (Huot model performs best). However, the zeta scores suggest that the LH model performs best for the high-value subset.

3.4. Confidence Interval Z-Tests

We performed multiple two-tailed z-tests for independent samples with H_0: $M_i = O_i$ and H_a: $M_i \neq Oi$ at a significance level of $\alpha = 0.01$. In Table 4, we tallied the results where the null hypothesis was retained and was considered a "success". As with previous analyses, we tallied results for the entire dataset, the

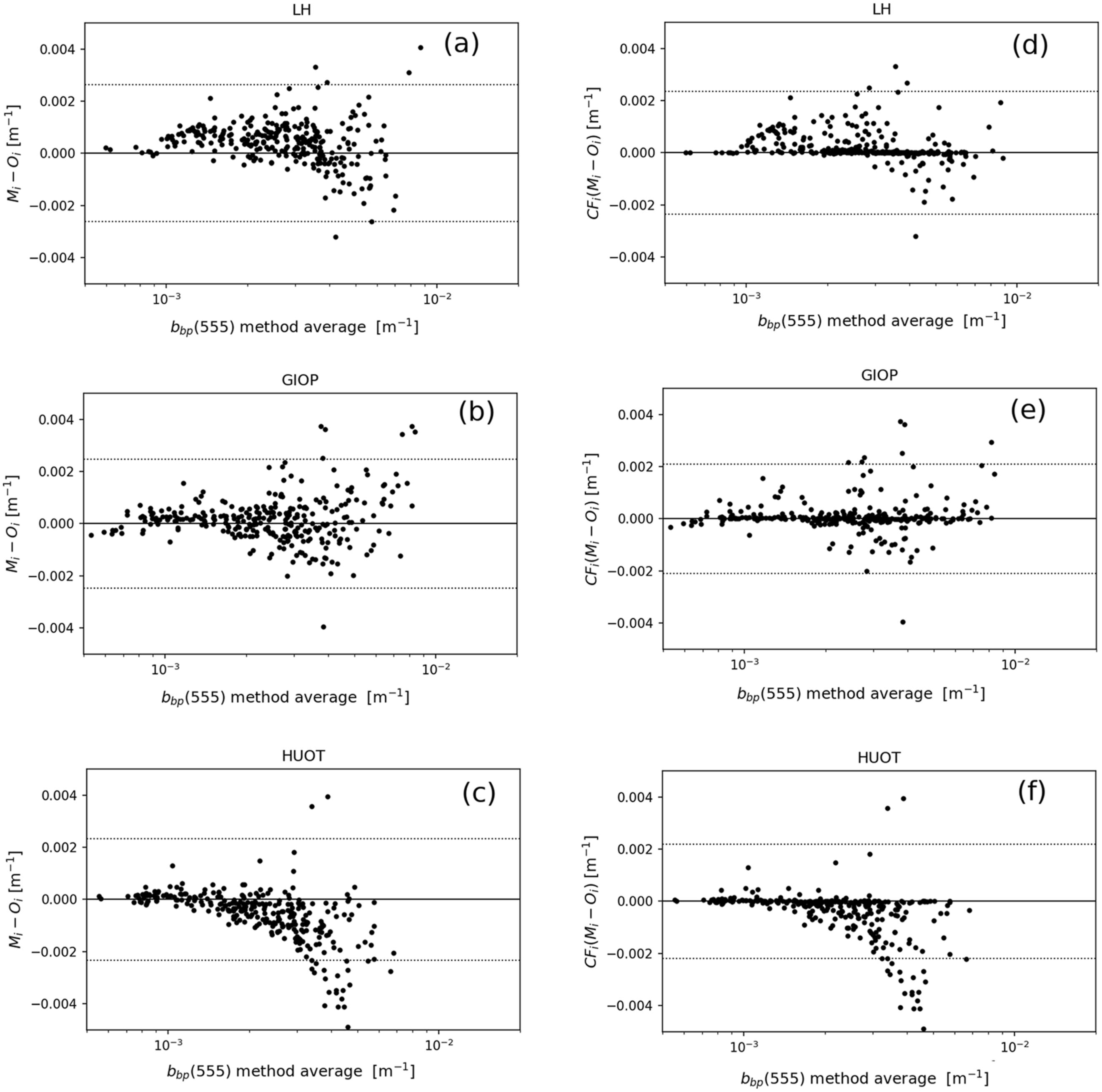

Figure 3. Bland–Altman plots of differences between modeled and observed b_{bp}(555) varying with the method average values of b_{bp}(555). Subplots (a)–(c) correspond to LH, GIOP, and Huot models, respectively. Subplots (d)–(f) are Bland–Altman plots with corresponding differences for the LH, GIOP, and Huot models, respectively. Dashed horizontal lines represent 97.5% confident interval about zero.

low-value subset, and the high-value subset. We repeated the analysis in the case where the data had been $\log_{10}$-transformed.

When considering the untransformed data, the GIOP had the most successes (73%) for the full dataset, the Huot model had the most successes (87%) for the low-value subset, and the LH model had the most successes (79%) for the high-value subset. For the $\log_{10}$-transformed data the LH model had the most successes for all data (67%) and the high-value subset (77%) while the Huot model had the most successes for the low-value subset (64%). These results are consistent with previous analyses with the exception of the

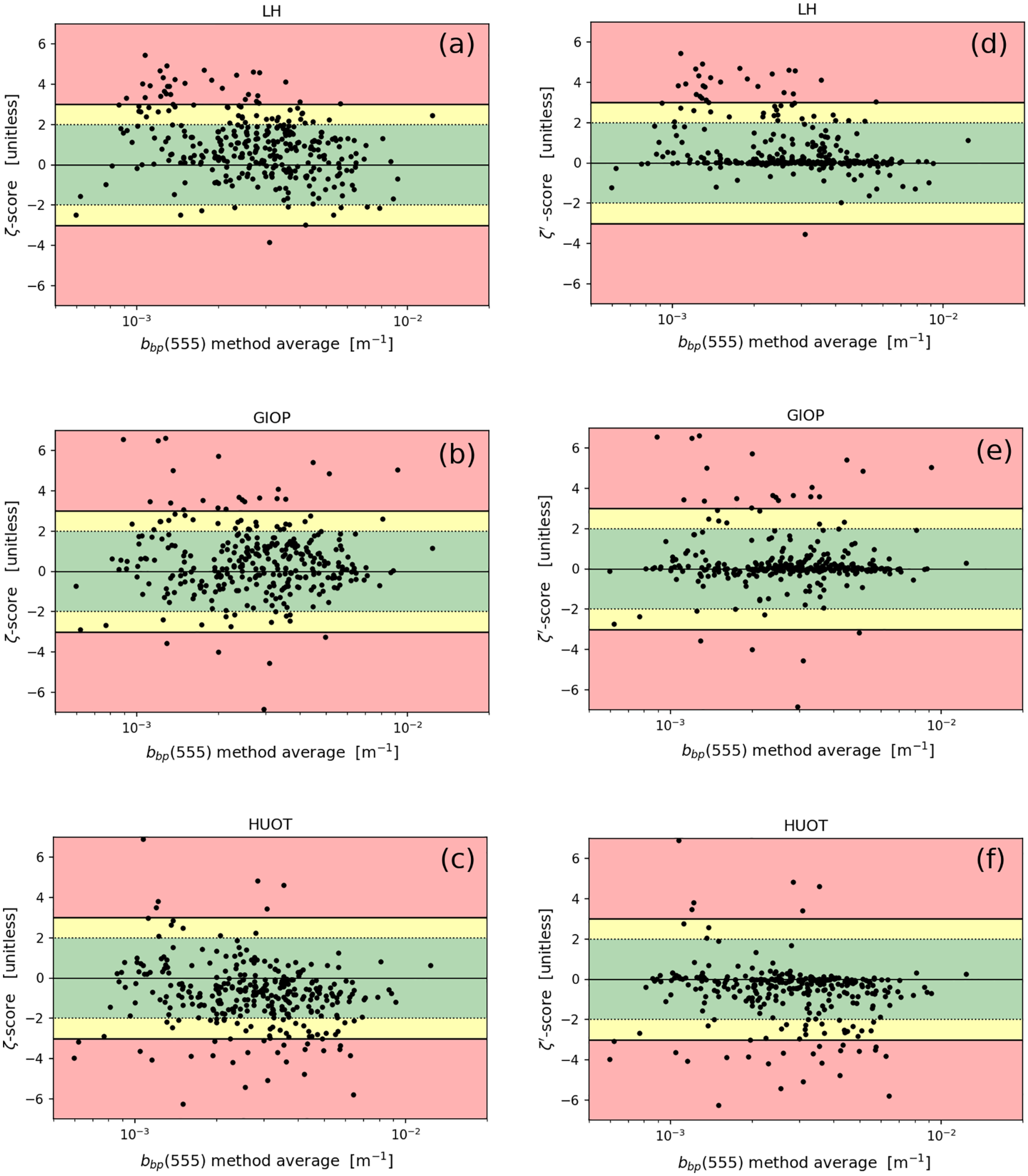

Figure 4. Zeta score plots comparing modeled and observed b_{bp}(555) varying with the method average values of b_{bp}(555). Subplots (a)–(c) correspond to LH, GIOP, and Huot models, respectively. Subplots (d)–(f) use corrected zeta scores for the LH, GIOP, and Huot models, respectively.

Table 3
Zeta-Score Statistics and Tallies for Three Models: LH, GIOP, and Huot

$b_{bp}(555)$ range	Model	N	Mean ζ (std)	Mean ζ' (std)	Tally of $\lvert\zeta\rvert < 2$	Tally of $\lvert\zeta'\rvert < 2$	Tally of $2 \le \lvert\zeta\rvert < 3$	Tally of $2 \le \lvert\zeta'\rvert < 3$	Tally of $\lvert\zeta\rvert \ge 3$	Tally of $\lvert\zeta'\rvert \ge 3$	No. wins
All	LH	326	0.888 (1.62)	0.561 (1.36)	249	278	45	22	32	26	0
	GIOP	326	**0.348** (1.78)	**0.234** (1.55)	**265**	**287**	**35**	**14**	**26**	**25**	**8**
	Huot	326	−0.814 (1.63)	−0.586 (1.48)	254	272	38	23	34	31	0
<1.25E−3 m^{-1}	LH	60	2.54 (1.31)	1.92 (1.60)	19	31	19	12	22	17	0
	GIOP	60	1.08 (1.94)	0.739 (1.79)	45	50	8	**3**	7	7	1
	Huot	60	**0.756** (1.22)	**0.436** (1.10)	**53**	**55**	**5**	**3**	**2**	**2**	**8**
≥1.25E−3 m^{-1}	LH	266	0.510 (1.42)	0.252 (1.06)	**230**	**247**	**26**	10	**10**	**9**	**5**
	GIOP	266	**0.179** (1.72)	**0.119** (1.48)	220	237	28	**8**	19	18	3
	Huot	266	−1.17 (1.49)	−0.820 (1.45)	201	217	33	20	32	29	0

Note. Bold typeface indicates best performance. Prime symbol (′) indicates corrected difference metrics.

LH model performing best for all data when $\log_{10}$-transformed. We do note, however, in Table 2 that for all data the $\log_{10}$-transformed LH skill metrics, aside from bias$_{\log}$, were generally similar to those of the GIOP.

4. Discussion

4.1. Summary of LH Model

In this study, we developed a LH-based ocean color model for estimating $b_{bp}(555)$. Often measurement uncertainties are not considered during empirical ocean color algorithm development. Thus, the objective for this exercise was to demonstrate how one might develop an empirical model that takes into account the uncertainties in training, validation, and model input data. The inter-comparison of LH with Huot and GIOP models primarily allowed us to determine if the LH model was performing with similar "in-family" predictive skill relative to the established models. However, the inter-comparison also served as an opportunity to benchmark the three models using a consistent validation dataset that included assumptions about measurement uncertainties.

Regression and difference metrics (Table 2), zeta-scores (Table 3), and confidence interval z-tests (Table 4) indicated that the GIOP performed best when the full validation data set was used. Qualitatively, the scatter plots (Figure 2) tend to confirm this result. However, after partitioning the validation dataset into low- and high-value subsets, the results revealed that the Huot model consistently outperformed the LH and GIOP for the low-value subset (i.e., where $b_{bp}(555) < 1.25 \times 10^{-3}$ m^{-1}), suggesting accuracy in model-derived *Chla*. The LH model outperformed Huot and marginally outperformed GIOP, for the high-value subset (i.e., where $b_{bp}(555) \ge 1.25 \times 10^{-3}$ m^{-1}). We note that the LH model did not show particularly good performance for the low-value subset. This result is not surprising considering Hu et al. (2012) showed that absorption, not backscattering, is expected to dominate a LH metric signal in oligotrophic waters. The fact the LH model performed well in high-value subset is, however, a promising result as SAAs such as the GIOP can have difficulty converging to a valid solution in highly turbid, optically complex environments.

While the LH model may not replace existing physics-based SAAs such as the GIOP, it may prove useful as a computationally efficient sanity check tool or perhaps serve to improve computational efficiency by (i) providing

Table 4
Tallies of Statistically Significant Overlap of Modeled and Observed 99% Confidence Intervals

$b_{bp}(555)$ range	Model	N	Tally (%)	Tally*
All	LH	326	230 (70%)	**220 (67%)**
	GIOP	326	**238 (73%)**	199 (61%)
	Huot	326	221 (68%)	145 (44%)
<1.25E−3 m^{-1}	LH	60	20 (33%)	15 (25%)
	GIOP	60	40 (67%)	29 (48%)
	Huot	60	**52 (87%)**	**39 (64%)**
≥1.25E−3 m^{-1}	LH	266	**210 (79%)**	**205 (77%)**
	GIOP	266	198 (74%)	170 (64%)
	Huot	266	169 (64%)	106 (40%)

Note. Percentage of total number is also given. Bold values indicate best within-group tally. *$\log_{10}$-transformed data.

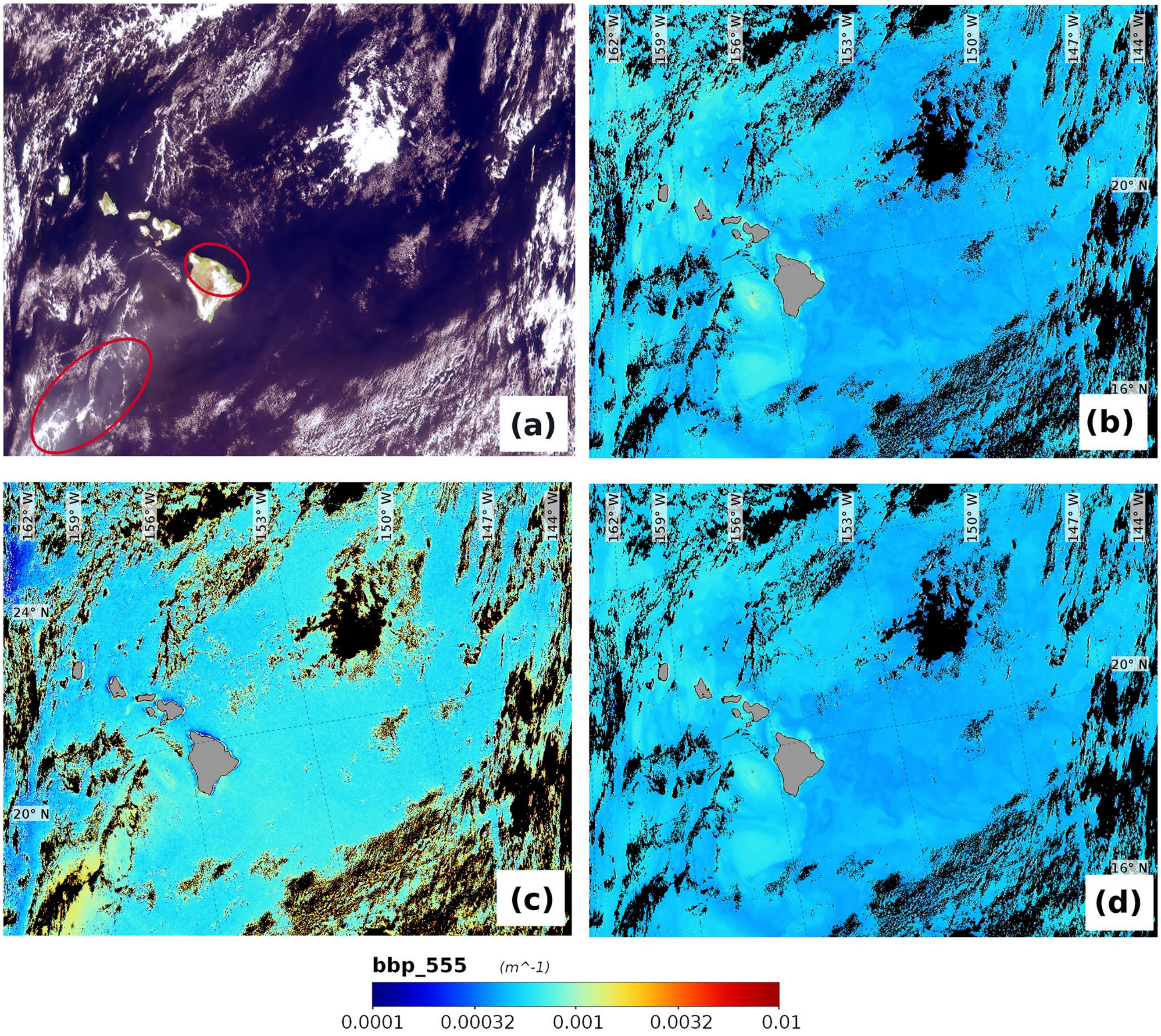

Figure 5. SeaWiFS imagery of the Hawaiian Island region of the North Pacific Ocean captured on December 1, 2000. Panel (a) is a quasi-true color image. Panels (b)–(d) depict $b_{bp}(555)$ derived using the LH, GIOP, and Huot models, respectively. Red ellipses denote regions where the GIOP exhibits artifacts in the retrievals. Cloud contaminated pixels are masked in black.

inverse models a good first guess for $b_{bp}(555)$ and/or (ii) helping to constrain the solution space. Similarly, the *Chla*-based Huot model may prove to be a useful in oligotrophic waters where SAAs are also known to underperform.

4.2. Application of LH Model to Satellite Imagery

We applied the LH, GIOP, and Huot models to two sample SeaWiFS scenes. By doing so we could visually determine if each model resolves oceanographic features in an expected manner or, alternatively, generates unwanted spatial artifacts and/or returns an unexpected number of invalid pixels (product failures). The first scene shown in Figure 5 is an oligotrophic region of the North Pacific Ocean adjacent to the Hawaiian Islands. In such a region, we expect that the oceanic b_{bp} signal is driven primarily by phytoplankton biomass. Qualitatively, the LH and Huot models gave very similar retrievals in oligotrophic waters of the North

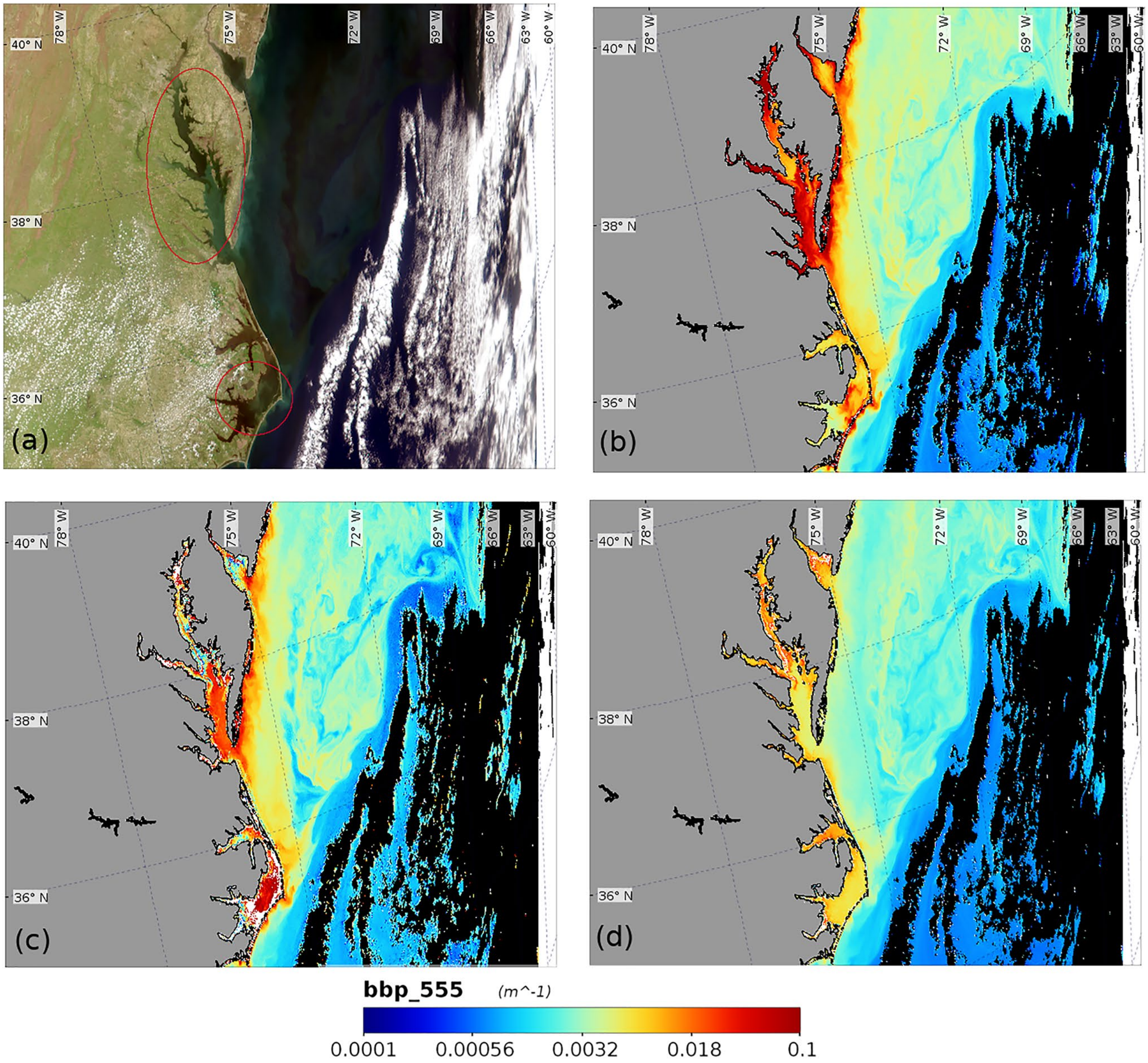

Figure 6. SeaWiFS imagery of the Chesapeake Bay region captured on the 28 April 2003. Panel (a) is a quasi-true color image. Panels (b)–(d) depict $b_{bp}(555)$ derived using the LH, GIOP, and Huot models, respectively. Red ellipses denote spatial features visible in the quasi-true color image that LH model resolves. Cloud contaminated pixels are masked in black.

Pacific Ocean (Figures 5b and 5d) with scene-wide median values of 4.83×10^{-4} m^{-1} and 4.75×10^{-4} m^{-1}, respectively. For reference, the GIOP scene-wide median was 6.75×10^{-4} m^{-1}.

The LH and Huot models resolved spatial features that were not well-distinguished in the GIOP retrieval such as eddies to the southwest of the Island of Hawaii (the largest island) and regions of low $b_{bp}(555)$ to the east of Hawaii. In addition, the LH model seemed robust to cloud edge, and straylight from land–areas where GIOP algorithm gives unusual retrievals (e.g., the red ellipses in Figure 5). Good performance of the LH approach in those areas is not surprising as Hu et al. (2012) demonstrated that LH metrics are robust to image artifacts such as cloud edge, straylight, and sunglint.

The second SeaWiFS scene shown in Figure 6 is of the Chesapeake Bay and the Mid-Atlantic Bight region. In the quasi-true color image (Figure 6a), the upper and lower red ellipses indicate the positions of the Chesapeake Bay and the Pamlico Sound, respectively. These two areas are complex bodies of water where the optical properties are driven by suspended mineral sediments, colored dissolved organic matter (CDOM), and high phytoplankton abundance. Dark-colored patches of water were likely dominated by CDOM. In addition, offshore phytoplankton blooms can be seen as green patches in the quasi-true color image.

The LH, GIOP, and Huot models resolve offshore $b_{bp}(555)$ similarly. The distinct gradient in $b_{bp}(555)$ is indicative of the edge of the Gulf Stream current. In the Chesapeake Bay and Pamlico Sound the LH model resolves high values of $b_{bp}(555)$, corresponding to bright features in the quasi-true color images, that are likely to be sediment or phytoplankton. In addition, the LH model resolves dark-colored patches of water, likely to be CDOM-dominated, as having lower $b_{bp}(555)$. The GIOP and Huot models do not retrieve as many valid pixels as the LH model nor do they resolve features visible in the quasi-true color image. While we cannot comment on the absolute accuracy of the LH model retrievals for that sample image due to a lack of validation data, the results suggest the model may be robust in optically complex waters.

4.3. Uncertainties and Skill Assessment Methods

The second objective of this study was to explore how measurement uncertainties might be incorporated into contemporary ocean color algorithm validation and we believe this work represents one of the first attempts to examine this. In the current validation paradigm, data pairs M_i and O_i are typically treated as exact values and their intrinsic standard uncertainties $u(M_i)$ and $u(O_i)$ are not considered. In an attempt to address this, we explored corrected difference metrics (mean bias and MAE), Bland–Altman and zeta-score plots, and confidence interval overlap testing.

In this study, we assumed 5% relative uncertainties in both $R_{rs}(\lambda)$ and $b_{bp}(555)$. In doing so we treated $u(M_i)$ and $u(O_i)$ as though they scale with the magnitude of M_i and O_i, respectively, in equal proportion (i.e., 5%) at all sensor wavelengths. We concede this assumption may not necessarily hold true but was still useful for demonstrative purposes. Indeed, Hu et al. (2013) demonstrated that relative uncertainties in SeaWiFS and MODIS $R_{rs}(\lambda)$ vary with both wavelength and bio-optical complexity, while McKee et al. (2009) reported $b_{bp}(\lambda)$ uncertainties that did not scale with magnitude. Furthermore, by assuming 5% relative uncertainties globally, one may underestimate absolute uncertainties in low-signal waters (e.g., b_{bp} in oligotrophic waters) and overestimate absolute uncertainties in high-signal waters (e.g., b_{bp} in turbid bays and estuaries). Thus, routine reporting of radiometric and IOP absolute uncertainties would be beneficial for model development and validation purposes.

By using alternative values for $u(M_i)$ and $u(O_i)$ the model skill results presented in this study are likely to vary. Nonetheless, our model development and validation framework is still valid and easily extendable to situations where improved estimates of $u(M_i)$ and $u(O_i)$ are available. From a metrological perspective, no measurement is complete without being reported along with its associated uncertainty and reliable estimates of uncertainties are better than having none. Furthermore, uncertainties in climate data records measured by Earth observation satellites should be computed and validated in a manner that follows metrological practice (Merchant et al., 2017). Thus, it is critical for the ocean color community to continue efforts to routinely characterize and report measurement uncertainties, including covariances, in both satellite and *in situ* datasets. Such characterization would support both algorithm development and satellite data product performance assessment activities, as well as use and interpretation of satellite and *in situ* data records in climate modeling studies.

The results in Table 2 indicate that application of the correction factor defined in Equation 15 did indeed change the values of the difference metrics and generally improved them. For example, if we consider the LH model's $\log_{10}$-transformed difference metrics for "All data," when the correction factor was applied the mean bias reduced from 1.22 to 1.13 and MAE reduced from 1.24 to 1.16. The effect of applying the correction factor to the LH model performance metrics can be visualized in the Bland–Altman plots where the corrected differences (Figure 3d) exhibit less variability about zero than uncorrected differences (Figure 3a).

We suggest the Bland–Altman and zeta-score plots may provide clearer graphical representations of model skill than traditional one-to-one scatter plots; a finding consistent with recent work by Knobelspiesse et al. (2019). For example, the Bland–Altman plots showed the Huot model performed well when $b_{bp}(555) < 1.1 \times 10^{-3}\ \mathrm{m}^{-1}$ after which it began to underestimate values. The color scheme of the zeta-score plots makes them particularly easy to interpret. Indeed, we envisage a "traffic light" classification scheme as a way to improve the communication of validation analyses to end-users. Furthermore, by tallying zeta-score class size it is possible to quantitatively interpret the zeta-score plots without adding undue complexity.

As a final example of how uncertainties might complement existing validation metrics, we considered multiple z-tests. The number of cases where H_0 was retained were reported (Table 4). These z-test results were generally consistent with the other validation assessments performed. We note that parametric testing requires an assumption that M_i and O_i are normally distributed with known variances. This may not always be a valid assumption and as ocean color variables typically follow a log-normal distribution and $\log_{10}$-transform of M_i, O_i, $u(M_i)$, and $u(O_i)$ may be required. Nonetheless, well-known z-tests may still serve as a cursory way of extending our understanding of agreement between M_i and O_i.

One caveat when considering these validation analyses are the magnitude of $u(M_i)$ and $u(O_i)$. The CF_i metric by definition is dependent on the degree of overlap of $p_m(m_i)$ and $p_o(o_i)$ whose dispersion we defined as $u(M_i)$ and $u(O_i)$, respectively. If these standard uncertainties are very large, the degree of overlap of $p_m(m_i)$ and $p_o(o_i)$ may be so close to 1 that our ability to compute meaningful difference metrics is encumbered. This also applies to our zeta-score calculations where large uncertainties in the denominator term may result in very small zeta-score values. As such, it may be prudent to interpret corrected validation difference metrics with thought given to the magnitude of the measurement uncertainties. While this may be challenging to visualize, graphical presentations such as PomPlots (Spasova et al., 2007) may be useful.

5. Conclusion

An empirical ocean color algorithm was developed for deriving $b_{bp}(555)$ using LH as the predictor variable. Using the simple LH empirical model as a test case, we performed end-to-end algorithm development and validation with assumed uncertainties in training, validation, and model input data. Once developed, the LH model was compared with the GIOP and Huot models. The LH model showed reasonable predictive skill across the entire dynamic range of the validation dataset with its best performance occurring when $b_{bp}(555) \geq 1.25 \times 10^{-3}$ m^{-1}.

By considering $u(M_i)$ and $u(O_i)$ we also demonstrated how standard uncertainties might be incorporated into ocean color validation. Importantly, our results clearly indicate that validation difference metrics (mean bias and MAE) were improved when corrected for measurement uncertainties. We also presented Bland–Altman and zeta-score plots as alternative methods to the traditional one-to-one scatter plots commonly used for validation. The zeta-score plots are particularly promising as their color-coded appearance makes them simpler to interpret. Overall, the study underscores the importance of on-going efforts by the ocean color community to characterize both model and observation uncertainties.

We acknowledge there are a number of other models capable of deriving $b_{bp}(555)$ (IOCCG, 2006; Werdell et al., 2018) that were not considered as this was beyond the scope of this research. However, a suitable framework for benchmarking newly developed ocean color models relative to established ones (e.g., GIOP) is particularly relevant to NASA's upcoming Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission (Werdell et al., 2019) which has a Science and Applications Team actively developing novel ocean color algorithms that take advantage of the PACE mission's hyperspectral and polarimetric capabilities. We expect methods reported here will complement existing approaches for model inter-comparisons (Brewin et al., 2015; Seegers et al., 2018).

Appendix A: Estimating Huot Backscattering Coefficient Model Uncertainties

The Huot et al. (2008) model derives $b_{bp}(555)$ as a function of $Chla$:

$$b_{bp}\left(555\right) = \alpha Chla^{\beta}. \tag{A1}$$

The coefficients α and β are calculated from Equations 8a and 8b in Huot et al. (2008) as:

$$\alpha = 2.267 \times 10^{-3} - 5.058 \times 10^{-6}\left(555 - 550\right), \tag{A2}$$

$$\beta = 0.565 - 4.86 \times 10^{-4}\left(555 - 550\right). \tag{A3}$$

We estimated the uncertainty associated with $b_{bp}(555)$ derived using Equation A1 as:

$$u\left(b_{bp}\left(\lambda\right)\right) = \left\{\left[\left[u\left(\alpha\right)Chla^{\beta}\right]^{2} + u\left(\beta\right)\alpha\ln\left(Chla\right)Chla^{\beta}\right]^{2} + \left[u\left(Chl\right)\alpha\beta Chla^{(\beta-1)}\right]^{2}\right\}^{0.5}. \tag{A4}$$

The standard uncertainties of model coefficients was estimated by dividing the 95% confidence intervals values reported in Table 1 of Huot et al. (2008) by 1.96 to give $u\left(\alpha\right) \approx 1 \times 10^{-4}$ and $u\left(\beta\right) \approx 0.02$. Uncertainty in chlorophyll-a pigment concentration, $u(Chla)$, was estimated per McKinna et al. (2019) with relative uncertainties in $R_{rs}(\lambda)$ set to 5%. We have not considered covariance terms for Equation A1 as there are unreported.

Data Availability Statement

The DS3 dataset is publicly available at doi: 10.1594/PANGAEA.886619 (Stramski & Reynolds, 2018) with further description in Loisel et al. (2018). The OC-CCI dataset is available at doi: 10.1594/PANGAEA.854832 (Valente et al., 2015) with further description in Valente et al. (2019). SeaWiFS level-1 used in this study is publicly available at doi: 10.5067/ORBVIEW-2/SEAWIFS/L1/DATA/1.

Acknowledgments
Support for this work was provided by a NASA Science of Terra, Aqua, and Suomi NPP Award and by the NASA Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) Project. We also wish to thank the three reviewers for their constructive comments.

References

Analytical Methods Committee Amctb No 74 (2016). z-Scores and other scores in chemical proficiency testing—their meanings, and some common misconceptions. *Analytical Methods*, *8*, 5553–5555. http://dx.doi.org/10.1039/C6AY90078J

Antoine, D., Siegel, D. A., Kostadinov, T., Maritorena, S., Nelson, N. B., Gentili, B., et al. (2011). Variability in optical particle backscattering in contrasting bio-optical oceanic regimes. *Limnology & Oceanography*, *56*(3), 955–973. https://doi.org/10.4319/lo.2011.56.3.0955

Austin, P. C., & Hux, J. E. (2002). A brief note on overlapping confidence intervals. *Journal of Vascular Surgery*, *36*(1), 194–195. https://doi.org/10.1067/mva.2002.125015

Behrenfeld, M. J., Westberry, T. K., Boss, E. S., O'Malley, R. T., Siegel, D. A., Wiggert, J. D., et al. (2009). Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. *Biogeosciences*, *6*(5), 779–794. https://doi.org/10.5194/bg-6-779-2009

Bland, M. J., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Brewin, R. J. W., Dall'Olmo, G., Sathyendranath, S., & Hardman-Mountford, N. J. (2012). Particle backscattering as a function of chlorophyll and phytoplankton size structure in the open-ocean. *Optics Express*, *20*(16), 17632–17652. https://doi.org/10.1364/OE.20.017632

Brewin, R. J. W., Sathyendranath, S., Müller, D., Brockmann, C., Deschamps, P.-Y., Devred, E., et al. (2015). The Ocean Color Climate Change Initiative: III. A round-robin comparison on in-water bio-optical algorithms. *Remote Sensing of Environment*, *162*, 271–294. https://doi.org/10.1016/j.rse.2013.09.016

Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research*, *100*(C7), 13237–13254. https://doi.org/10.1029/95JC00458

Dall'Olmo, G., Westberry, T. K., Behrenfeld, M. J., Boss, E., & Slade, W. H. (2009). Significant contribution of large particles to optical backscattering in the open ocean. *Biogeosciences*, *6*(6), 947–967. https://doi.org/10.5194/bg-6-947-2009

Evers-King, H., Martinez-Vicente, V., Brewin, R. J. W., Dall'Olmo, G., Hickman, A. E., Jackson, T., et al. (2017). Validation and intercomparison of ocean color algorithms for estimating particulate organic carbon in the oceans. *Frontiers in Marine Science*, *4*(251). https://doi.org/10.3389/fmars.2017.00251

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Franz, B. A., Behrenfeld, M. J., Siegel, D. A., & Signorini, S. R. (2017). Global ocean phytoplankton [in: State of the Climate in 2016]. *Bulletin of the American Meteorological Society*, *99*(8), S94–S96. https://doi.org/10.1175/2017BAMSStateoftheClimate.1

Frouin, R. J., Franz, B. A., Ibrahim, A., Knobelspiesse, K., Ahmad, Z., Cairns, B., et al. (2019). Atmospheric correction of satellite ocean-color imagery during the PACE era. *Frontiers of Earth Science*, *7*, 145. https://doi.org/10.3389/feart.2019.00145

Fu, G., Hasekamp, O., Rietjens, J., Smit, M., Di Noia, A., Cairns, B., et al. (2020). Aerosol retrievals from different polarimeters during the ACEPOL campaign using a common retrieval algorithm. *Atmospheric Measurement Techniques*, *13*(2), 553–573. https://doi.org/10.5194/amt-13-553-2020

Gower, J., King, S., & Goncalves, P. (2008). Global monitoring of plankton blooms using MERIS MCI. *International Journal of Remote Sensing*, *29*(21), 6209–6216. https://doi.org/10.1080/01431160802178110

Harmel, D. R., Smith, K. P., & Migliaccio, W. K. (2010). Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. *Transactions of the ASABE*, *53*(1), 55–63. https://doi.org/10.13031/2013.29502

Hu, C. (2009). A novel ocean color index to detect floating algae in the global oceans. *Remote Sensing of Environment*, *113*(10), 2118–2129. https://doi.org/10.1016/j.rse.2009.05.012

Hu, C., Feng, L., & Lee, Z. (2013). Uncertainties of SeaWiFS and MODIS remote sensing reflectance: Implications from clear water measurements. *Remote Sensing of Environment*, *133*, 168–182. https://doi.org/10.1016/j.rse.2013.02.012. https://doi.org/10.1016/j.rse.2013.02.012

Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., & Proctor, C. W. (2019). Improving satellite global chlorophyll a data products through algorithm refinement and data recovery. *Journal of Geophysical Research: Oceans*, *124*(3), 1524–1543. https://doi.org/10.1029/2019JC014941

Hu, C., Lee, Z., & Franz, B. (2012). Chlorophyll algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, *117*(C1). http://dx.doi.org/10.1029/2011JC007395

Huot, Y., Morel, A., Twardowski, M. S., Stramski, D., & Reynolds, R. A. (2008). Particle optical backscattering along a chlorophyll gradient. in the upper layer of the eastern South Pacific Ocean. *Biogeosciences*, *5*(2), 495–507. https://doi.org/10.5194/bg-5-495-2008

IOCCG. (2006). *Remote sensing of inherent optical properties: Fundamentals, tests of algorithms, and applications* (18). IOCCG. https://doi.org/10.25607/OBP-96

IOCCG. (2020). *Uncertainties in ocean color remote sensing* (18). IOCCG. https://doi.org/10.25607/OBP-696

IOCCG Protocol Series (2019). *Protocols for satellite ocean color validation: In situ optical radiometry*. Dartmouth. https://doi.org/10.25607/OBP-691

Knobelspiesse, K., Tan, Q., Bruegge, C., Cairns, B., Chowdhary, J., van Diedenhoven, B., et al. (2019). Intercomparison of airborne multi-angle polarimeter observations from the Polarimeter Definition Experiment. *Applied Optics*, *58*(3), 650–669. https://doi.org/10.1364/AO.58.000650

Lamquin, N., Mangin, A., Mazeran, C., Bourg, B., Bruniquel, V., & D'Andon, O. F. (2013). *OLCI L2 pixel-by-pixel uncertainty propagation in OLCI clean water branch*. ESA ATBD ref. S3-L2-SD-01-C01-ACR-TN.

Lee, Z., Hu, C., Shang, S., Du, K., Lewis, M., Arnone, R., & Brewin, R. (2013). Penetration of UV-visible solar radiation in the global oceans: Insights from ocean color remote sensing. *Journal of Geophysical Research: Oceans*, *118*(9), 4241–4255. https://doi.org/10.1002/jgrc.20308

Lee, Z., & Huot, Y. (2014). On the non-closure of particle backscattering coefficient in oligotrophic oceans. *Optics Express*, *22*(23), 29223–29233. https://doi.org/10.1364/OE.22.029223

Loisel, H., Stramski, D., Dessailly, D., Jamet, C., Li, L., & Reynolds, R. A. (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*(3), 2141–2171. https://doi.org/10.1002/2017JC013632

Lunetta, R. S., Schaeffer, B. A., Stumpf, R. P., Keith, D., Jacobs, S. A., & Murphy, M. S. (2015). Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA. *Remote Sensing of Environment*, *157*, 24–34. https://doi.org/10.1016/j.rse.2014.06.008

Matthews, M. W., & Odermatt, D. (2015). Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters. *Remote Sensing of Environment*, *156*, 374–382. https://doi.org/10.1016/j.rse.2014.10.010

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

McKee, D., Chami, M., Brown, I., Calzado, V. S., Doxaran, D., & Cunningham, A. (2009). Role of measurement uncertainties in observed variability in the spectral backscattering ratio: A case study in mineral-rich coastal waters. *Applied Optics*, *48*(24), 4663–4675. https://doi.org/10.1364/AO.48.004663

McKinna, L. I. W., Cetinić, I., Chase, A. P., & Werdell, P. J. (2019). Approach for propagating radiometric data uncertainties through NASA ocean color algorithms. *Frontiers of Earth Science*, *7*, 176. https://doi.org/10.3389/feart.2019.00176

McKinna, L. I. W., Werdell, P. J., & Proctor, C. W. (2016). Implementation of an analytical Raman scattering correction for satellite ocean-color processing. *Optics Express*, *24*(14), A1123–A1137. https://doi.org/10.1364/OE.24.0A1123

Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., et al. (2017). Uncertainty information in climate data records from Earth observation. *Earth System Science Data*, *9*(2), 511–527. https://essd.copernicus.org/articles/9/511/2017/

Mitchell, C., Hu, C., Bowler, B., Drapeau, D., & Balch, W. M. (2017). Estimating particulate inorganic carbon concentrations of the global ocean from ocean color measurements using a reflectance difference approach. *Journal of Geophysical Research: Oceans*, *122*(11), 8707–8720. https://doi.org/10.1002/2017JC013146

Morel, A. (1988). Optical modeling of the upper ocean in relation to its biogenous matter content (case I waters). *Journal of Geophysical Research*, *93*(C9), 10749–10768. https://doi.org/10.1029/JC093iC09p10749

Morel, A., Gentili, B., Claustre, H., Babin, M., Bricaud, A., Ras, J., & Tièche, F. (2007). Optical properties of the "clearest" natural waters. *Limnology & Oceanography*, *52*(1), 217–229. https://doi.org/10.4319/lo.2007.52.1.0217

Morel, A., & Maritorena, S. (2001). Bio-optical properties of oceanic waters: A reappraisal. *Journal of Geophysical Research*, *106*(C4), 7163–7180. https://doi.org/10.1029/2000JC000319

NASA Goddard Space Flight Center, O. E. L. Ocean Biology Processing Group. (2010).*Sea-viewing wide field-of-view sensor (SeaWiFS) L1. data*. https://doi.org/10.5067/ORBVIEW-2/SEAWIFS/L1/DATA/1

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., et al. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research*, *103*(C11), 24937–24953. http://dx.doi.org/10.1029/98JC02160

O'Reilly, J. E., & Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors—OC4, OC5 & OC6. *Remote Sensing of Environment*, *229*, 32–47. https://doi.org/10.1016/j.rse.2019.04.021

Seegers, B. N., Stumpf, R. P., Schaeffer, B. A., Loftin, K. A., & Werdell, P. J. (2018). Performance metrics for the assessment of satellite data products: An ocean color case study. *Optics Express*, *26*(6), 7404–7422. https://doi.org/10.1364/OE.26.007404

Spasova, Y., Pommé, S., & Wätjen, U. (2007). Visualisation of interlaboratory comparison results in PomPlots. *Accreditation and Quality Assurance*, *12*(12), 623–627. https://doi.org/10.1007/s00769-007-0319-9

Stramski, D., Boss, E., Bogucki, D., & Voss, K. J. (2004). The role of seawater constituents in light backscattering in the ocean. *Progress in Oceanography*, *61*(1), 27–56. https://doi.org/10.1016/j.pocean.2004.07.001

Stramski, D., & Reynolds, R. A. (2018). *DS3 ocean optics dataset*. https://doi.org/10.1594/PANGAEA.886619

Sullivan, J. M., Twardowski, M. S., Ronald, J., Zaneveld, V., & Moore, C. C. (2013). Measuring optical backscattering in water. In A. A. Kokhanovsky (Ed.), Ed., *Light scattering reviews 7: Radiative transfer and optical properties of atmosphere and underlying surface* (pp. 189–224). Springer. https://doi.org/10.1007/978-3-642-21907-8_6

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Taberner, M., et al. (2015). A compilation of global bio-optical in situ data for ocean-color satellite applications. Supplement to: Valente, A et al. (2016): A compilation of global bio-optical in situ data for ocean-colour satellite applications. *Earth System Science Data 8*, 235–252. https://doi.org/10.5194/essd-8-235-2016

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Taberner, M., et al. (2019). A compilation of global bio-optical in situ data for ocean-color satellite applications—Version two. *Earth System Science Data*, *11*(3), 1037–1068. https://doi.org/10.5194/essd-11-1037-2019

Werdell, P. J., & Bailey, S. W. (2005). An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sensing of Environment*, *98*(1), 122–140. https://doi.org/10.1016/j.rse.2005.07.001

Werdell, P. J., Behrenfeld, M. J., Bontempi, P. S., Boss, E., Cairns, B., Davis, G. T., et al. (2019). The plankton, aerosol, cloud, ocean ecosystem mission: Status, science, advances. *Bulletin of the American Meteorological Society*, *100*(9), 1775–1794. https://doi.org/10.1175/BAMS-D-18-0056.1

Werdell, P. J., Franz, B. A., Bailey, S. W., Feldman, G. C., Boss, E., Brando, V. E., et al. (2013). Generalized ocean color inversion model for retrieving marine inherent optical properties. *Applied Optics*, *52*(10), 2019–2037. https://doi.org/10.1364/AO.52.000027

Werdell, P. J., McKinna, L. I. W., Boss, E., Ackleson, S. G., Craig, S. E., Gregg, W. W., et al. (2018). An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing. *Progress in Oceanography*, *160*, 186–212. https://doi.org/10.1016/j.pocean.2018.01.001

Zhang, X., Hu, L., Xiong, Y., Huot, Y., & Gray, D. (2020). Experimental estimates of optical backscattering associated with submicron particles in clear oceanic waters. *Geophysical Research Letters*, *47*(4). e2020GL087100. https://doi.org/10.1029/2020GL087100