Research Article | Vol. 26, No. 6 | 19 Mar 2018 | OPTICS EXPRESS 7404
Optics EXPRESS

Performance metrics for the assessment of satellite data products: an ocean color case study

Bridget N. Seegers,[1,2,*] Richard P. Stumpf,[3] Blake A. Schaeffer,[4] Keith A. Loftin,[5] and P. Jeremy Werdell

[1]*NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Greenbelt, MD 20771, USA*
[2]*Universities Space Research Association (USRA), Columbia, MD, USA*
[3]*NOAA, National Ocean Service, Silver Spring, MD, USA*
[4]*US Environmental Protection Agency, Office of Research and Development, Durham, NC, USA*
[5]*US Geological Society, Kansas Water Science Center, Lawrence, KS, USA*
[*]*bridget.n.seegers@nasa.gov*

Abstract: Performance assessment of ocean color satellite data has generally relied on statistical metrics chosen for their common usage and the rationale for selecting certain metrics is infrequently explained. Commonly reported statistics based on mean squared errors, such as the coefficient of determination (r^2), root mean square error, and regression slopes, are most appropriate for Gaussian distributions without outliers and, therefore, are often not ideal for ocean color algorithm performance assessment, which is often limited by sample availability. In contrast, metrics based on simple deviations, such as bias and mean absolute error, as well as pair-wise comparisons, often provide more robust and straightforward quantities for evaluating ocean color algorithms with non-Gaussian distributions and outliers. This study uses a SeaWiFS chlorophyll-a validation data set to demonstrate a framework for satellite data product assessment and recommends a multi-metric and user-dependent approach that can be applied within science, modeling, and resource management communities.

OCIS codes: (010.0010) Atmospheric and oceanic optics; (010.0280) Remote sensing and sensors; (000.5490) Probability theory, stochastic processes, and statistics.

References and links

1. C. R. McClain, "A Decade of Satellite Ocean Color Observations," Annu. Rev. Mar. Sci. **1**(1), 19–42 (2009).
2. IOCCG, "Why Ocean Colour? The Societal Benefits of Ocean-Colour Technology," T. Platt, N. Hoepffner, V. Stuart and C. Brown, (eds.), Reports of the International Ocean-Colour Coordinating Group, No. 7, IOCCG, Dartmouth, Canada. (2008).
3. IOCCG, "Remote Sensing in Fisheries and Aquaculture," M. H. Forget, V. Stuart and T. Platt, (eds.), Reports of the International Ocean-Colour Coordinating Group, No. 8, IOCCG, Dartmouth, Canada. (2009).
4. S. Dutkiewicz, A. E. Hickman, O. Jahn, W. W. Gregg, C. B. Mouw, and M. J. Follows, "Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model," Biogeosciences **12**(14), 4447–4481 (2015).
5. A. Gnanadesikan, K. Emanuel, G. A. Vecchi, G. W. Anderson, and R. Hallberg, "How ocean color can steer Pacific tropical cyclones," Geophys. Res. Lett. **37**(18), L18802 (2010).
6. C. S. Rousseaux and W. W. Gregg, "Recent decadal trends in global phytoplankton composition," Global Biogeochem. Cycles **29**(10), 1674–1688 (2015).
7. B. A. Schaeffer, K. Loftin, R. P. Stumpf, and P. J. Werdell, "Agencies collaborate, develop a cyanobacteria assessment network," Eos (Wash. D.C.) **96**, ••• (2015).
8. B. A. Walther, J. L. Moore, "The definitions of bias, precision, and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance," Ecography **28**(6), 815–829 (2005).
9. R. J. W. Brewin, S. Sathyendranath, D. Müller, C. Brockmann, P. Y. Deschamps, E. Devred, R. Doerffer, N. Fomferra, B. Franz, M. Grant, S. Groom, A. Horseman, C. Hu, H. Krasemann, Z. Lee, S. Maritorena, F. Melin, M. Peters, T. Platt, P. Regner, T. Smyth, F. Steinmetz, J. Swinton, J. Werdell, and G. N. White III, "The ocean colour climate change initiative: III. a round-robin comparison on in-water bio-optical algorithms," Remote Sens. Environ. **162**, 271–294 (2015).

#315013 https://doi.org/10.1364/OE.26.007404
Journal © 2018 Received 8 Dec 2017; revised 1 Mar 2018; accepted 2 Mar 2018; published 14 Mar 2018

Optics EXPRESS

10. C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Clim. Res. **30**, 79–82 (2005).
11. C. J. Willmott, K. Matsuura, and S. M. Robeson, "Ambiguities inherent in sums-of-squares-based error statistics," Atmos. Environ. **43**(3), 749–752 (2009).
12. C. J. Willmott, S. M. Robeson, and K. Matsuura, "Climate and Other Models May Be More Accurate Than Reported," Eos (Wash. D.C.) **98**, ••• (2017).
13. J. S. Armstrong, *Evaluating Forecasting Methods. In Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer, 2001).
14. M. H. Birnbaum, "Reply to the Devil's advocates: Don't confound model testing and measurement," Psychol. Bull. **81**(11), 854–859 (1974).
15. T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," Geosci. Model Dev. **7**(3), 1247–1250 (2014).
16. C. A. Stow, J. Jolliff, D. J. McGillicuddy, Jr., S. C. Doney, J. I. Allen, M. A. M. Friedrichs, K. A. Rose, and P. Wallhead, "Skill assessment for coupled biological/physical models of marine systems," J. Mar. Syst. **76**(1-2), 4–15 (2009).
17. F. Mélin and B. A. Franz, "Assessment of satellite ocean colour radiometry and derived geophysical products," in Optical Radiometry for Oceans Climate Measurements, chap. 6.1, G. Zibordi, C. Donlon, and A. Parr, eds., Academic, Experimental Methods in the Physical Sciences **47**, 609–638 (2014).
18. S. C. Doney, I. Lima, J. K. Moore, K. Lindsay, M. J. Behrenfeld, T. K. Westberry, N. Mahowald, D. M. Glover, and T. Takahashi, "Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data," J. Mar. Syst. **76**(1-2), 95–112 (2009).
19. J. E. O'Reilly, S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain, "Ocean color chlorophyll algorithms for SeaWiFS," J. Geophys. Res. **103**(C11), 24937–24953 (1998).
20. S. Maritorena, D. A. Siegel, and A. R. Peterson, "Optimization of a semianalytical ocean color model for global-scale applications," Appl. Opt. **41**(15), 2705–2714 (2002).
21. C. Hu, Z. Lee, and B. Franz, "Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference," J. Geophys. Res. **117**(C1), C01011 (2012).
22. S. W. Bailey and P. J. Werdell, "A multi-sensor approach for the on-orbit validation of ocean color satellite data products," Remote Sens. Environ. **102**(1-2), 12–23 (2006).
23. B. A. Franz, P. J. Werdell, G. Meister, S. W. Bailey, R. E. Eplee, Jr., G. C. Feldman, E. Kwiatkowska, C. R. McClain, F. S. Patt, and D. Thomas, "The continuity of ocean color measurements from SeaWiFS to MODIS," Proc. SPIE **5882**, 58820W (2005).
24. J. L. Mueller, R. R. Bidigare, C. Trees, W. M. Balch, J. Dore, D. T. Drapeau, D. Karl, L. Van Heukelem and J. Perl, "Ocean optics protocols for satellite ocean color sensor validation, revision 5, volume V: Biogeochemical and bio-optical measurements and data analysis protocols," NASA Tech. Memo. 2003–211621, NASA Goddard Space Flight Center, Greenbelt, Maryland (2003).
25. S. B. Hooker, L. Clementson, C. S. Thomas, L. Schlüter, M. Allerup, J. Ras, H. Claustre, C. Normandeau, J. Cullen, M. Kienast, W. Kozlowski, M. Vernet, S. Chakraborty, S. Lohrenz, M. Tuel, D. Redalje, P. Cartaxana, C. R. Mendes, V. Brotas, S. G. P. Matondkar, S. G. Parab, A. Neeley, and E. S. Egeland, "The Fifth SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-5),". NASA Tech. Memo 2012–217503, NASA Goddard Space Flight Center, Greenbelt, Maryland (2012).
26. P. J. Werdell, L. I. W. McKinna, E. Boss, S. G. Ackleson, S. E. Craig, W. W. Gregg, Z. Lee, S. Maritorena, C. S. Roesler, C. S. Rousseaux, D. Stramski, J. M. Sullivan, M. S. Twardowski, M. Tzortziou, and X. Zhang, "An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing," Prog. Oceanogr. **160**, 186–212 (2018).
27. T. S. Kostadinov, D. A. Siegel, S. Maritorena, and N. Guillocheau, "Ocean color observations and modeling for an optically complex site: Santa Barbara Channel, California, USA," J. Geophys. Res. **112**(C7), C07011 (2007).
28. P. J. Werdell, B. A. Franz, S. W. Bailey, G. C. Feldman, E. Boss, V. E. Brando, M. Dowell, T. Hirata, S. J. Lavender, Z. Lee, H. Loisel, S. Maritorena, F. Mélin, T. S. Moore, T. J. Smyth, D. Antoine, E. Devred, O. H. d'Andon, A. Mangin, and A. Mangin, "Generalized ocean color inversion model for retrieving marine inherent optical properties," Appl. Opt. **52**(10), 2019–2037 (2013).
29. IOCCG, "Remote Sensing of Inherent Optical Properties: Fundamentals, Test of Algorithms, and Applications," Z.-P. Lee. (eds.), Reports of the International Ocean-Colour Coordinating Group, No. 5, IOCCG, Dartmouth, Canada. (2006).
30. J. W. Campbell, "The lognormal distribution as a model for bio-optical variability in the sea," J. Geophys. Res. **100**(C7), 13237–13254 (1995).
31. International vocabulary of metrology – Basic and general concepts and associated terms (VIM 3rd edition) (2012).
32. H. P. Young, "Condorcet's theory of voting," Am. Polit. Sci. Rev. **82**(4), 1231 (1988).
33. S. B. Broomell, V. David, and H. H. Por, "Pair-wise comparisons of multiple models," Judgm. Decis. Mak. **6**(8), 821 (2011).
34. C. D. Mobley, J. Werdell, B. Franz, Z. Ahmad, and S. Bailey, "Atmospheric correction for satellite ocean color radiometry," NASA Technical Memorandum, 217551, **85** (2016).

Optics EXPRESS

35. P. J. Werdell, S. Bailey, B. Franz, L. Harding, Jr., G. C. Feldman, and C. R. McClain, "Regional and seasonal variability of chlorophyll-a in Chesapeake Bay as observed by SeaWiFS and MODIS-Aqua," Remote Sens. Environ. **113**(6), 1319–1330 (2009).
36. E. J. Kwiatkowska, B. A. Franz, G. Meister, C. R. McClain, and X. Xiong, "Cross calibration of ocean-color bands from Moderate Resolution Imaging Spectroradiometer on Terra platform," Appl. Opt. **47**(36), 6796–6810 (2008).
37. G. Meister, B. A. Franz, E. J. Kwiatkowska, and C. R. McClain, "Corrections to the calibration of MODIS Aqua ocean color bands derived from SeaWiFS data," IEEE Trans. Geosci. Remote Sens. **50**(1), 310–319 (2012).
38. F. J. Anscombe, "Graphs in statistical analysis," Am. Stat. **27**(1), 17–21 (1973).
39. F. Mosteller and J. W. Tukey, *Data analysis and Regression: A Second Course in Statistics* (Addison-Wesley, 1977).
40. J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis* (Wadsworth, 1983).
41. E. Tufte, *The Visual Display of Quantitative Information* (Graphics, 1983)
42. W. S. Cleveland, *Visualizing Data* (Hobart, 1993)
43. B. E. J. Cisneros, T. Oki, N. W. Arnell, G. Benito, J. G. Cogley, P. Döll, T. Jiang, and S. S. Mwakalila, Economic and Related Instruments to Provide Incentives (chapter 17), in: Climate Change 2014: impacts, adaptation, and vulnerability. Part A: global and sectoral aspects. contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change 2014.
44. C. J. Willmott, "On the validation of models," Phys. Geogr. **2**(2), 184–194 (1981).

1. Introduction

The development and refinement of algorithms to derive geophysical variables from satellite measurements of ocean color has been pursued for decades [1]. These data records play a key role in furthering our scientific understanding of the spatial and temporal distributions of marine phytoplankton and other biogeochemical parameters on regional to global scales. Such parameters provide proxy (surrogate) indicators of marine ecosystem health and link to economically important measures, such as fisheries production, water quality, and recreational opportunities [2-3]. In the four decades since the advent of satellite ocean color, the number of algorithms and approaches to produce geophysical data products has increased substantially given improved knowledge of ocean optics, advances in and an increased volume of *in situ* measurements, improvements in computing power, and open access to satellite data records. Satellite measurements of ocean color now play an important role in scientific Earth system modeling [4–6] and resource management decision support [7]. This growing demand for satellite ocean color data products has necessitated the development and expansion of algorithms to accommodate user demands and requirements that span oceans, coastal marine waters, estuaries, lakes, reservoirs, and large rivers. Accommodating this influx of new and enhanced end-user needs subsequently resulted in a growing difficulty in assessing how algorithm refinements or algorithm implementation across (new) missions ultimately results in any meaningful or constructive improvement in the accuracy and precision of derived satellite data products. This difficulty partly results from the ocean color science community traditionally relying on a small set of statistical tools for algorithm assessment that provide metrics of overall performance that are not unequivocally easily interpreted or are appropriate for some, but not all, data sets or missions (and, thus, not appropriate across regions or missions).

Estimating the performance of an algorithm requires metrics for accuracy, bias, and, ideally, variability (precision) [8]. The ocean color community frequently assesses algorithm performance using ordinary least squares metrics, in particular the root mean square error of the regression (RMSE), the coefficient of determination (r^2), and the regression slope (see [9] for additional review). RMSE provides an appropriate metric for validation exercises when error distributions are Gaussian [15] and when the goal of an investigation is highlighting sensitivity to outliers (conceivable when testing a model). However, Gaussian data sets without outliers are not ubiquitous across all ocean color data sets to be validated, rendering these metrics occasionally informatively inferior to metrics without as much sensitivity to outliers and non-Gaussian distributions [10–14]. More commonly, error distributions in ocean color validation data sets have long tails (outliers) (Fig. 1), and RMSE estimates do not

Optics EXPRESS

capture the average error. The potentially misinterpreted results associated with sum of squares-based metrics has led to recommendations of metrics based on absolute deviations or errors [10-11, 13]. Mean absolute error (MAE), sometimes referred to as mean absolute deviation (MAD), and RMSE, also referred to as root mean square deviation (RMSD) take the form:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |M_i - O_i|}{n} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (M_i - O_i)^2}{n}} \tag{2}$$

where M, O, and n represent the modeled value, the observation, and the sample size, respectively. RMSE varies not only with the average error, but also with variability in the error magnitudes (through their squaring) and the square root of the number of samples. In other words, RSME differs from MAE through its additional dependence on the distribution of error magnitudes and the sample size, both of which underscore its additional sensitivity to data set distributions and outliers [10,12].

While both r^2 and regression slopes have their merits, they provide incomplete descriptions of algorithm performance (and slope is not an error metric). Reporting both has value (albeit not in isolation), but for completeness and to encourage community discussion of their interpretation, a review of their limitations follows. Regarding the former, r^2 is not only sensitive to outliers, but is also: (1) inconsistently interpretable across varied data sets, as the prediction variance is normalized to the total variance and, thus, a model with a fixed error will report different r^2 results when applied to areas with narrow versus wide data ranges [14]; and (2) can overstate variable relationships even with randomly selected variables [13]. With regards to the latter, the regression slope remains particularly unreliable for data sets with outliers as it employs squaring that can under- or over-emphasize the outliers, unless weighting or other complex methods are used to remove points with leverage on the relationship [16]. While slopes may be useful in assessing model performance over wide data ranges, they can also easily report a value of unity for a strongly biased, low-precision model, thereby complicating their interpretation and utility [14]. If error varies linearly across the data range, a slope (on the error residuals) may provide insight into such trends. Ultimately, r^2 and slope provide useful metrics for ocean color validation activities, but only with cautious interpretation and in combination with additional error metrics. Table 1 provides an additional summary and comparison between selected and historically used statistics.

The urgency in developing robust (and, perhaps more importantly, broadly community endorsed) approaches for remote sensing algorithm assessment is evident through international efforts such as the Ocean Colour Climate Change Initiative that present comprehensive approaches to algorithm analysis [9]. In addition, agency laboratories such as the NASA Ocean Biology Processing Group (OBPG; https://oceancolor.gsfc.nasa.gov) require performance metrics that can be consistently applied to multiple missions of varied duration and availability of field validation data – without which comparisons of algorithms within a mission and of data products across missions become very difficult to interpret given spatial and temporal biases in field sampling and varied numbers of satellite-to-*in situ* matchup pairs. The importance of standardized methods, common assessment approaches and limitations, along with challenges associated with gathering high quality *in situ* validation data are discussed by Mélin and Franz in their assessment of ocean color satellite radiometry and geophysical products [17].

Optics EXPRESS

Table 1. Summary of performance metric statistics suggested used by the manuscript and others commonly used by the ocean color satellite data products community highlighting the advantages and disadvantages of different metrics.

Measurement	Frequently Used Metrics	Why or Why Not for Ocean Color	Notes
Accuracy	RMSE	• Distribution sensitive (assumes Gaussian) • Often misinterpreted to be a simple estimate of average error • No consistent relationship with average error magnitudes	Other Sum of Squares based measures have same problems, such as standard deviation, standard error.
Goodness of fit	r^2	• Can be misinterpreted if not given in context, because it lacks a response to bias and is sensitive to outliers • Can misrepresent error when the range is small • Can overstate variable relationships even with apparently random error	
	Slope	• Can be misinterpreted, by reporting a good value for strongly-biased, low-precision models. • Leverages (biased errors on either end) produce meaningless slopes • Cannot address non-linear error • Can allow tuning of a model to fit a particular region	Common least squares regression gives biased slope when the x variables contain errors [9]
	Suggested Metrics		
Bias	Bias	• Quantifies the average difference between this estimator and expected value • Estimates systematic error	Often based on mean, however median error can also be used if a more robust metric is needed
Accuracy	MAE	• Does not amplify outliers • Accurately reflects error magnitude	Compared to mean, median absolute estimates are less sensitive to outliers. Similar metrics include mean/ median absolute percent error
	New Approaches		
Point by point accuracy	% wins (Residuals)	• Considers model failures • Provides consistent head-to-head comparison of algorithms	Pairwise comparison Decision support metric
Temporal stability	CV Intra-pixel	• Estimates imagery pixel stability. • Estimates algorithm spatial and temporal performance. • Does not require satellite-to-*in situ* match-ups	

Again, highlighting the interest and need for community discussion of algorithm assessment. While not necessarily related to ocean color, validation methods are also being examined in greater detail in other areas of oceanography [e.g., 16-17]. Stow et al. [16] reviewed the statistical metrics used to assess model skill in 142 papers from oceanographic journals from 2000 until 2007. They found that most studies relied on simple visual assessments, used subjective language such as "reasonable" to assess model performance, and rarely employed quantitative and objective statistics such as residuals (<20% of the time), all of which suggests a need for more rigorous methods. Stow et al. [16] also summarized a variety of statistical metrics for assessment, including approaches to compare spatial maps. Similarly, Doney et al. [18] examined the need for a standardized set of performance metrics to allow for ease in inter-comparing ecosystem-biogeochemistry model performance. They ultimately suggested a set of quantitative metrics and encouraged the adoption of a community-wide systematic standardized approach. Other Earth system disciplines have considered forecast evaluations methods, with discussions ranging from general assessment strategies for forecast models [e.g., 13] to specific methodologies, such as improved selection and interpretation of error metrics [12,14].

Optics EXPRESS

Ultimately, given the influx of new and revised algorithms and new missions and increasing dynamic ranges of interest and expertise, the ocean color community needs consistent, meaningful, and community-endorsed statistical approaches for algorithm assessment that accommodate varied data set sizes and can be equally effectively applied to (that is, are scalable to) global, regional and local applications. This study presents an exploration of metrics to assess algorithm performance and proposes approaches to combine metrics for comprehensive algorithm evaluation. It also presents a recommended set of performance metrics that includes spatial and temporal assessments, which have often been overlooked with previous methods. The goals of this study are to: (1) identify and demonstrate a simple, reliable suite of statistical methods that are easy and appropriate for use by the science and end-user communities to assess remote sensing algorithms without a priori assumptions of data distributions; and (2) illustrate the pressing need to think critically about statistical analysis and move beyond the statistical metrics the ocean color community traditionally relies upon that can be regularly misinterpreted and therefore misleading. As a case study, this paper focuses on a well published and peer-reviewed satellite ocean color data product, the near-surface concentration of the photosynthetic pigment chlorophyll-a (*Chl*; mg m^{-3}) [19–21]. This paper does not provide a definitive study that represents all water masses, data products, and user needs at all times, but rather highlights a set of metrics, graphics, and a strategy for algorithm assessment using some example global applications and reinforces the need to be analytical about model performance evaluation.

2. Methods

2.1 Data and algorithms

Coincident satellite-to-*in situ Chl* match-ups for the NASA Sea-viewing Wide Field-of-view Sensor (SeaWiFS; 1997-2010) were acquired from the NASA/OBPG SeaWiFS Bio-optical Archive and Storage System (SeaBASS) [22]. This satellite data product and *in situ* data set were selected because: (1) both are well characterized [1,22]; (2) both provided a wide dynamic range of observations (0.012 to 72 mg m^{-3} *in situ*); and, (3) the satellite retrievals of *Chl* from the multiple algorithms under consideration have very subtle differences that result in their performance being difficult to compare (thus, offering a desirably challenging data set with which to vet this approach). The match-ups were executed using a 5x5 satellite pixel box centered on the location of the *in situ* measurement and quality control of the match-ups followed methods detailed in Bailey and Werdell [22]. Briefly, (1) coincidence was considered as <3 hours between the satellite and *in situ* observation; (2) matches with more than half of marine pixels masked in a 5x5 satellite pixel box were excluded; (3) matches with coefficients of variation of the remaining unmasked pixels in the box exceeding 0.15 were excluded; and (4) *Chl* was reported as the filtered median of the remaining unmasked pixels in the box. The final sample size was 2,161 satellite-to-*in situ* pairs. These pairs were stratified into three trophic regions, defined using the mission-long SeaWiFS *Chl* climatology as oligotrophic ($Chl \leq 0.1$mg m^{-3}), mesotrophic ($0.1 < Chl \leq 1$ *Chl* mg m^{-3}), and eutrophic ($Chl > 1$ mg m^{-3}) [23]. A range of uncertainties accompany the *in situ* data used as reference data, a deep exploration of which exceeds the scope of this manuscript. Briefly, however, definition of these uncertainties has been pursued or cataloged [24–26]. Therefore, type II linear regression with the reduced major axis (RMA) approach was used, accounting for uncertainties in both the dependent and independent variables [9, 27]. MAE, for example, can be scaled into an unbiased percentage by scaling the model-*in situ* difference by the mean of the model and *in situ* observations [17].

Three approaches to derive SeaWiFS Chl were considered, namely the OC3, OCI, and GSM algorithms. Briefly, ocean color satellite instruments measure top-of-atmosphere radiances at discrete visible and near-infrared wavelengths. Atmospheric correction algorithms are applied to these radiances to remove the contributions of the atmosphere and

Optics EXPRESS

derive estimates of spectral remote-sensing reflectances ($R_{rs}(\lambda)$; sr^{-1}), the light exiting the water column normalized to the incident surface irradiance [28]. Bio-optical algorithms are then applied to the $R_{rs}(\lambda)$ to generate estimates of geophysical data products of interest, such as *Chl*. OC3 estimates *Chl* following the band ratio approach of O'Reilly et al. [19], where a blue-to-green ratio of $R_{rs}(\lambda)$ statistically relates to *Chl* via a polynomial expression (see also https://oceancolor.gsfc.nasa.gov/atbd/chlor_a/). Within OC3, the numerator is designated as the greater of $R_{rs}(443)$, $R_{rs}(490)$ and the denominator is $R_{rs}(555)$. The ocean chlorophyll index (OCI) estimates *Chl* following Hu et al. [21], which blends two algorithms: (1) OC4, another band-ratio approach that differs from OC3 in that the numerator is designated as the greatest of $R_{rs}(443)$, $R_{rs}(490)$ and $R_{rs}(510)$ for a given satellite pixel; and (2) an independent chlorophyll index (CI) derived as a spectral $R_{rs}(\lambda)$ line height of reflectance at 555 nm above a baseline drawn from 443 to 670. OCI uses CI exclusively for pixels where *Chl* < 0.15 mg m^{-3}, OC4 exclusively where *Chl* > 0.2 mg m^{-3}, and a weighted transition from CI to OC4 where 0.15 < *Chl* 0.2 mg m^{-3}. While these latter two algorithms strictly adopt empirical relationships between $R_{rs}(\lambda)$ and *Chl*, the final algorithm employs the semi-analytical approach of Maritorena et al. [20]. GSM (Garver, Siegel, Maritorena) uses a simplified form of the radiative transfer equation and a non-linear spectral matching optimization to derive *Chl* from $R_{rs}(\lambda)$ [28-29].

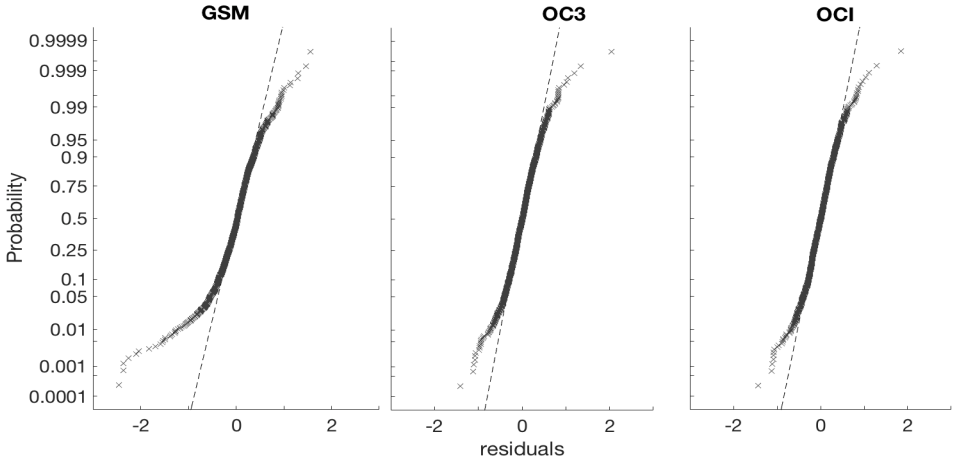

Fig. 1. Normality probability plots of the error distributions for several ocean color models (described in section 2.1). Gaussian distributions would fall onto the dashed 1:1 line, the error distributions have long tails and therefore are non-Gaussian.

Normality plots (Fig. 1) reveal that the error distribution of this SeaWiFS validation data set to be non-Gaussian with long tails, suggesting that mean square error metrics may be undesirable. This data set, being global is comprehensive, has a sample size that exceed 2,000 matches. To our knowledge, this is the largest ocean color validation data set in the ocean color community – thus, providing a best-case scenario – and, yet, its distribution remains non-Gaussian. This does not challenge previous demonstrations that global, log-transformed *Chl* is nearly normally distributed [30], but rather indicates that the accumulated ground truth samples do not represent this normal distribution. Naturally, the sample sizes decrease when this data set is broken into subsets by trophic region and normality is never achieved. At the time of this writing, similar validation data sets available from SeaBASS include far fewer satellite-to-*in situ* pairs (e.g., <200 for the Suomi NPP Visible Infrared Imaging Radiometer Suite (VIIRS); see https://seabass.gsfc.nasa.gov) and all demonstrate non-Gaussian distribution behavior (not shown).

Optics EXPRESS

2.2 Selection of recommended statistics

A variety of statistical performance metrics for algorithm performance assessment exist. These metrics cross numerous scientific communities, but their appropriateness for specific data sets varies [13]. Broadly speaking, identification of a meaningful metric depends on the intersection of the statistics appropriate for the characteristics of the modeled products and the statistics appropriate for the application of those products. User considerations when selecting performance metrics often include the impact of: (1) outliers; (2) the full dynamic range of the data versus a specific, narrow data range (e.g., performance in the global ocean versus in a single estuary or lake); (3) the temporal and/or spatial stability of an algorithm; (4) the spatial coverage provided by an algorithm; (5) allowable uncertainties; and, (6) allowable biases. Clarifying such considerations enables selection of performance metrics a priori [13]. Questions relating to trends, for example, may be better addressed by emphasizing model biases and long-term consistency as a priority over absolute model accuracies.

2.2.1 Error metrics

Core performance metrics for algorithm evaluation include bias (systematic error), variability (random error, precision), and accuracy that combines bias and variability [8,31]. Typically, systematic bias and accuracy metrics are calculated, and random error is inferred [10,13,14], even when it can be calculated from RMSE and bias [8,12]. Bias has long been a reported value in ocean color algorithm assessment and offers a simple description of the systematic direction of the error, as either over- or under- estimating the prediction on average [8]. MAE is an appropriate metrics of accuracy for non-Gaussian distributions. Random error provides an estimate of precision and isolates the contribution of random variability produced by the measurement from the overall algorithm error [8]. As such, the International Vocabulary of Metrology (VIM) defines random measurement error as equal to measurement error less the total systematic measurement error (bias) [31]. While methods to remove systematic error from total error exist under a Gaussian assumption exist, approaches to quantify the random error component of MAE for known non-Gaussian or unknown distributions are less developed [8,13]. The advantages of developing such an approach for ocean color algorithms will be covered further in the discussion. The remainder of this study focuses on bias and MAE, defined as:

$$\text{bias} = 10\,\hat{}\left(\frac{\sum_{i=1}^{n}\log_{10}(M_i)-\log_{10}(O_i)}{n}\right) \tag{3}$$

$$\text{MAE} = 10\,\hat{}\left(\frac{\sum_{i=1}^{n}\left|\log_{10}(M_i)-\log_{10}(O_i)\right|}{n}\right) \tag{4}$$

Note that the observations are log-transformed (e.g., such that Eq. (4) differs from Eq. (1)). Many marine geophysical variables are conventionally log-transformed prior to calculation of error metrics as uncertainty and variance are proportional to the concentration, and the data values frequently span multiple orders of magnitude (Fig. 2). The end result of this log-transformation is the conversion of the metric from linear to multiplicative space. Generally speaking, the use of either linear or multiplicative metrics depends on the characteristics of the model, the variable of interest, and their uncertainties. Those with constant uncertainties (homoscedastic), such as water temperature, benefit from evaluation with linear metrics. Those with uncertainty that varies proportionally with data value, such as *Chl*, benefit from assessment with multiplicative metrics. Linear metrics have the same units as the variable examined, whereas multiplicative metrics are dimensionless. A multiplicative bias of 1.2 indicates that the model is 1.2x (20%) greater on average than the observed variable. Multiplicative MAE always exceed unity, such that a MAE of 1.5 indicates relative

measurement error of 50%. Here, multiplicative forms of the metrics were used, as *Chl* error is proportional to its concentration and spans over four decades in magnitude in the SeaWiFS validation data set (Fig. 2). Accordingly, the statistics were calculated in $\log_{10}$ space, then converted out of $\log_{10}$ space prior to interpretation of the results. The back-transformation from $\log_{10}$ space results in bias values closest to unity being the least biased and bias less than unity indicating a negative bias. The ocean color community has not typically transformed metrics from $\log_{10}$ space. This back transformation minimizes potential misinterpretation of reported error; for example, a reported $\log_{10}$ value of 0.3 does not indicate 30% uncertainty, but rather approximately a 100% uncertainty ($10^{0.3} = 1.995$), suggesting a preferred practice of reporting 1.995 in lieu of 0.3. The r^2 and the regression slope were also calculated for the analysis using $\log_{10}$-transformed *Chl*.

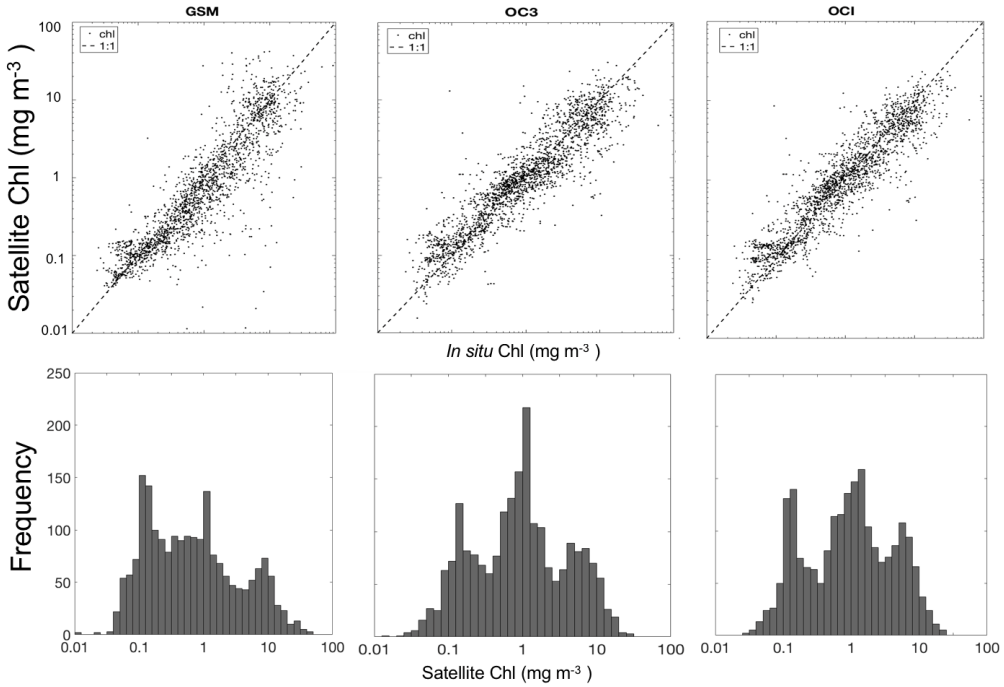

Fig. 2. The top row are SeaWiFS-GSM, OC3, and OCI derived *Chl* to *in situ Chl* match-up scatterplot comparisons. The bottom row histograms shows the distribution of SeaWiFS-GSM, OC3, and OCI derived *Chl* values. Data were $\log_{10}$ transformed for display.

2.2.2 Decision metrics

Decision metrics enable additional comparison and selection of algorithms. Decision metrics date back to the 18th century mathematician Condorcet and are often described as "voting" methods [32]. One immediate practical approach is the pair-wise comparison based on Condorcet [33]. Pair-wise comparisons operate sequentially on each observation: (1) for a given observation, the model-observation differences are calculated for every model under consideration; (2) the model with the minimal difference is designated the winner for that given observation; (3) the number of wins per model are tabulated for all observations; and, (4) the model with the most wins is designated the best performing model. Unlike many other error metrics, the pair-wise comparison directly considers model failures – when model *A* provides a valid retrieval for a given observation but model *B* does not, only model *A* remains in the pool of potential winners for that observation. This metric will penalize a model that fails frequently, but performs well when it works. In this study, we adopted the pair-wise comparison of algorithm residuals (= model – observation), with the lowest residual designated as the winner. Results of this analysis were reported in terms of *percent wins*.

Optics EXPRESS

2.2.3 Spatially and temporally mapped metrics

Spatial and temporal performance of an algorithm may further inform the performance assessment, as coincident match-ups between satellite and *in situ* data cannot ubiquitously capture model performance under all conditions at all times. In general, satellite and *in situ* match-up data sets remain sparsely populated on large temporal and spatial scales [22]. Time-series analysis and population statistics provide one means of exploring spatial and temporal performance when sufficient *in situ* data exist [27,35]. Satellite imagery analysis provides another complementary – and, to our knowledge, largely unexplored – means of assessing algorithm behavior and consistency in space and time. Using satellite imagery to evaluate algorithm spatial extent of valid retrievals, temporal (e.g., day-to-day or week-to-week) consistency in retrievals, and spatiotemporal distributions of error metrics from satellite pixels may provide an additional decision discriminator when traditional model-versus-observation error metrics are otherwise limited. Such analyses may also be informative where decision support activities prioritize consistent and broad satellite coverage over model bias or accuracy. Furthermore, satellite imagery assessment informs on the effects of satellite data processing (through flagging or masking of questionable retrievals) on the algorithms, as elements of processing also vary in performance in space and time [34].

An approach that builds upon existing concepts used for on-orbit satellite calibration and validation activities was adopted to assess algorithm spatial and temporal performance [23, 36-37]. SeaWiFS 14-day global composites (1-15 September 2007) were produced at 9-km spatial resolution using an equal-area sinusoidal projection using SeaDAS software. Derived products included *Chl* mean, *Chl* standard deviation, and the number of observations per spatial bin included in the mean and standard deviation. The mean and the standard deviation were calculated from pixels that contribute to the 9km bin both spatially and temporally throughout the 14-day window. Trophic regions as described in Section 2.1 were used for comparison. For each spatial bin in the 14-day composite with greater than one observation, the coefficient of variation (CV) was calculated as the ratio of the mean of the standard deviation to the mean, which is a normalized estimate of data spread around the mean. The CV was used as an estimate of intra-pixel stability and an indicator of temporal consistency.

2.2.4 Decision graphics

In addition to statistical metrics, plots and graphics have long been demonstrated as necessary for understanding model performance and uncertainties. Two basic plots are common in model assessment, namely, scatterplots of modeled versus reference values and residual plots of the difference between model and reference versus reference values [38-39]. Additionally, a variety of plots can be used to compare multivariate data and aid in model comparison such as scatterplot matrixes, parallel coordinate or profile symbol plots, and star plots [40–42]. Star plots (also known as radar plots) are used in this study to provide an example of an effective graphical approach for evaluating the behavior of algorithms across multiple error metrics [40–43]. A star plot visually displays and compares multiple metrics and, with appropriate scaling, highlights differences in the metrics [41,43]. In general, the plot center represents values that indicate unacceptable algorithm performance, such that values on a spoke (or ray) nearer to the center identify the poorer performing approaches. The maximum length of each spoke reveals more optimal performance of an algorithm, such that the best performing instance reaches farthest from the center. Star plots were generated to visually display and compare algorithm performance assessment using the bias, MAE, pair-wise comparison, and CV metrics, with their values scaled from zero to one. Maximum and minimum values must be assigned for each variable to create the range for normalization. Note that normalizing over the range of values requires attention to avoid exaggerating trivial differences between modeled retrievals [40]. For normalization in this case, zero was used for all minimum values and maximums were created by adding 0.1 to each variable's absolute max value, with the exception of percent wins, for which was assigned a max value of 90%. Lower values for

Optics EXPRESS

many metrics (e.g., bias, MAE and CV) indicate better performance and, therefore, for the purposes of star plot normalization and visualization, we subtracted these metric values from a number greater than their maximum absolute value before normalizing. This transformation resulted in all of the best performing metrics visually reporting the largest values in the star plots, near the end of the spokes.

3. Results

Satellite-to-*in situ* match-ups were executed for analyses on the full data set (Fig. 2). Qualitatively, the scatter plots show reasonably equivalent performance across the full dynamic range of *Chl*. The GSM regression slope was closest to unity (Table 2) despite showing the most scatter and outliers and the least visually linear relationship across the dynamic range of *Chl* (Fig. 2).

Table 2. Statistical output comparing algorithm performance of the SeaWIFS-to-*in situ* *Chl* validation data set. The highlights indicate which algorithm best performed for each statistical comparison. If results were within 0.02 of best performing they were highlighted simply to emphasize similarly performing algorithms. It is possible to compare suggested approach on the left in addition to r^2 and regression slop on the right.

	Suggested Metrics					Other	
Water Type Algorithm	**n**	**bias**	**MAE** Accuracy	**Overall Wins (%)**	**CV**	**r^2**	**slope**
Across All							
GSM	2037	0.79	1.76	41.4	0.59	0.78	0.99
OC3	2161	1.03	1.63	49.5	0.55	0.84	0.90
OCI	2161	1.03	1.61	53.8	0.45	0.85	0.90
Oligotrophic							
GSM	247	1.39	1.47	67.7	1.05	0.14	1.41
OC3	248	1.66	1.82	30.3	1.62	0.11	2.08
OCI	248	1.72	1.81	58.7	1.06	0.14	1.87
Mesotrophic							
GSM	864	0.79	1.58	47.1	0.85	0.51	1.24
OC3	901	1.21	1.52	59.9	0.70	0.59	1.24
OCI	901	1.18	1.54	40.7	0.63	0.60	1.30
Eutrophic							
GSM	926	0.67	2.05	30.6	0.43	0.41	1.45
OC3	1011	0.80	1.68	44.5	0.34	0.53	1.08
OCI	1011	0.81	1.62	59.2	0.34	0.55	1.01

Inspection of satellite-*in situ* residuals confirms the equivalent performance shown in the scatterplots and highlights the long tail of the GSM residual distribution (Fig. 3). Bias and MAE were calculated for the full data set and stratified by trophic level (Table 2). GSM reported slightly fewer successful match-ups (5.7%) than the OC3 and OCI. Semi-analytical algorithms such as GSM – and spectral matching approaches in general – are more sensitive to spectrally-dependent errors in radiometric data than those that employ band ratios and band differences and, following, fail to provide a retrieval more frequently.

For the full data set, OCI and OC3 reported the lowest biases, with indistinguishable values of 1.03 (~3%). Recall that bias values closer to unity indicate less biased results and values less than one indicate negative biases, per the back transformation from $\log_{10}$ space. GSM reported the only negative bias of 0.79 (–21%). OCI and OC3 reported the lowest MAE with values of 1.6 indicating variability of 60% across all *Chl*. Collective consideration of bias and MAE designates OCI as the best performer for the full data set. While the r^2 also indicates this, it does not provide ample additional information. Exploring this briefly, when two data sets have the same data range, their r^2 provide qualitatively similar, and redundant,

Optics EXPRESS

information compared to MAE and RMSE. But, both MAE and RMSE, however, provide a quantification of the error, whereas r^2 does not.

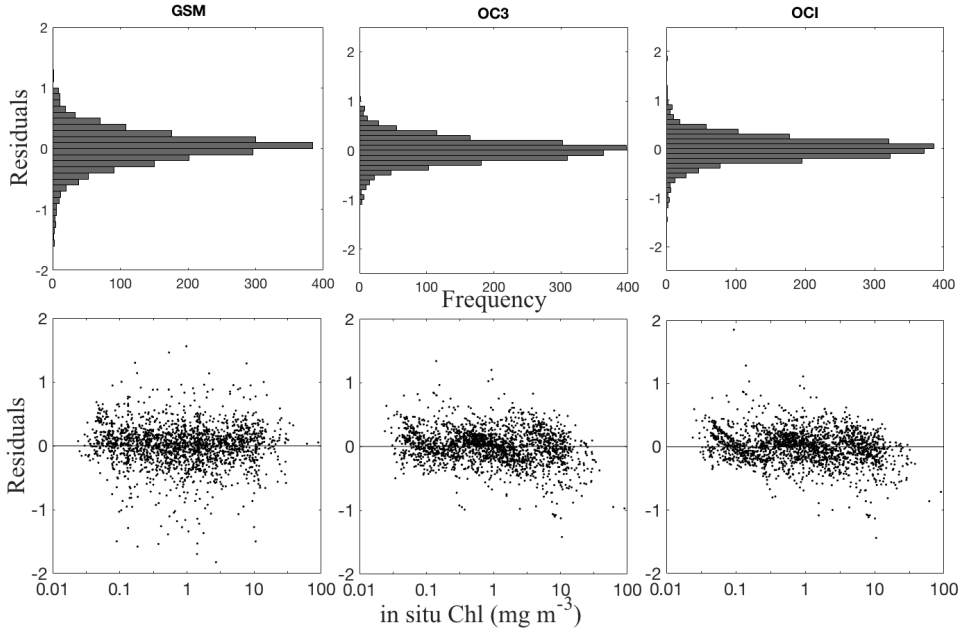

Fig. 3. Log_{10} residuals histograms and scatterplots the SeaWiFS-to-*in situ Chl* match-ups. The top row are histograms of log_{10} summarizing the error distribution of GSM, OC3, and OCI algorithms. The bottom panels are residual plots of the difference between model satellite *Chl* and the reference *in situ* values versus reference values. The plots were created with log_{10} values, but the axes are in *Chl* units (mg m^{-3}).

Algorithm performance varied for each trophic level (Table 2). A detailed discussion of mechanisms for this variation exceeds the scope of this paper, but briefly, causes include trophic-level-specific variations in atmospheric correction and *Chl* algorithm performance, *in situ* data sampling and processing (*in situ* measurement uncertainties can vary with water type), and spatial and temporal representativeness. GSM emerged as the best performer for oligotrophic water. Oligotrophic values of r^2 and regression slope are not unequivocally informative, largely resulting from a small *Chl* range that spans only from 0.02 to 0.1 mg m^{-3} for this trophic level. The extremely low r^2, combined with the large slopes, might lead to a conclusion that these models perform most poorly in this trophic region. Yet, their accuracies in oligotrophic waters exceed those in eutrophic waters, and GSM in oligotrophic water reports the best MAE accuracy of any application presented in this study. For mesotrophic waters, similarities in reported error metrics confound performance assessment, as the biases, MAE, r^2 and regression slope differ only slightly across algorithms. Depending on the end user requirements, an evaluator may be forced to simply prioritize bias (OCI) versus accuracy (OC3) or vice versa. For eutrophic waters, OCI emerged as the best performer across all metrics.

Optics EXPRESS

Table 3. Chl algorithm performance assessed point by point across all water types and by individual water type. The "winner" was the algorithm with the smallest absolute residual in a pair to pair comparison.

	Percent Wins		
Algorithm	**GSM**	**OC3**	**OCI**
	Across Water Types n = 2161		
GSM	X	57.5	59.0
OC3	42.5	x	52.0
OCI	41.0	48.0	x
Overall Wins	41.8	52.7	55.5
GSM Failure	124 (5.7%)		
	Oligotrophic n = 248		
GSM	X	29.0	40.7
OC3	71.0	x	67.7
OCI	59.3	32.3	x
Overall Wins	65.1	30.7	54.2
GSM Failure	1 (0.4%)		
	Mesotrophic n = 901		
GSM	X	54.6	51.8
OC3	45.4	x	31.9
OCI	48.2	68.1	x
Overall Wins	46.8	61.4	41.8
GSM Failure	37 (4.1%)		
	Eutrophic n = 1011		
GSM	X	67.3	71.2
OC3	32.7	x	59.7
OCI	28.8	40.3	x
Overall Wins	30.8	53.8	65.5
GSM Failure	85 (8.4%)		

Results from the pair-wise comparisons provide additional discriminators in support of the previously reported error metrics (Tables 2 and 3). For the full data set, OCI won most frequently (~54% wins), supporting the error metric identification of this algorithm as the best performer. This performance is not uniformly distributed across water types. For oligotrophic waters, GSM won most frequently (65.1%), supporting its error metric identification as the best performer. For mesotrophic water, pair-wise comparison provides perhaps the most discriminating assessor of algorithm performance. The error metrics presented above identified OC3 and OCI as candidate best performers for this mesotrophic subset, however, the pair-wise comparison reported OC3 won most frequently across all algorithms overall (61.4%) and when compared one-on-one with OCI, OC3 outperformed 68.1% of the time. For eutrophic water, OCI emerged as the best performer (61.4%), which also reported slightly better bias and MAE. In all subsets, GSM reported slightly smaller sample sizes, with its frequency of failure systematically increasing from oligotrophic (0.4% failure rate) to eutrophic (8.4% failure rate) waters. This difference can partially explain the lower percent wins for GSM, but not enough to explain the substantial differential in wins between GSM and the other algorithms in eutrophic water.

While not executed fully here, one might also compare only common satellite-*in situ* pairs (that is, only those where all approaches provided a valid match-up). In some situations, additional information on algorithm performance may be revealed through evaluation of results across common ranges of applicability. That said, within the context of this study, a reanalysis across all algorithms considering only the 2,037 GSM match-ups led to minimal differences relative to the values reported in Table 2, with OCI and OC3 bias and MAE shifting by <0.02.

Consideration of the temporal patterns in satellite imagery offers an additional discriminator for the previously reported error metrics. The 14-day composites of OCI, OC3, and GSM show similar patterns in the global spatial distribution of *Chl* (Fig. 4). In the open

Optics EXPRESS

ocean gyres, GSM and OCI maintain lower intra-pixel CVs of 1.05 and 1.06, respectively, compared to 1.62 for OC3 suggesting greater temporal stability in their retrievals (Table 2).

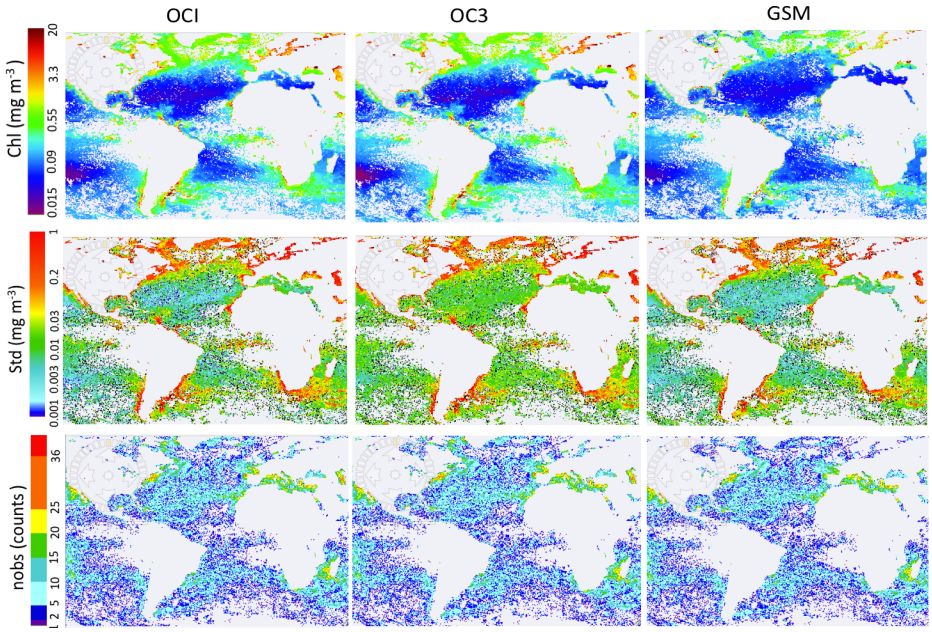

Fig. 4. Global image of SeaWiFS 14-day OCI-, OC3-, GSM-derived mean *Chl*, standard deviation, and number of observations (nobs). Satellite imagery analysis provides a means of assessing algorithm behavior and consistency in time and space. The satellite imagery can be used to evaluate algorithm spatial extent of valid retrievals, temporal (e.g., day-to-day or week-to-week) consistency in retrievals, and spatiotemporal distributions of error metrics from compiled satellite pixels. *nobs* can be used to compare the spatial coverage consistency of the algorithms. These analyses may also be informative where decision support activities prioritize consistent and broad satellite coverage.

In mesotrophic water, where the other error metrics do not unequivocally identify a best performer, the intra-pixel CV identifies OCI (0.63) as a somewhat better performer than OC3 (0.7), which provides a useful metric for decision support prioritizing temporal algorithm stability over overall algorithm variability. The intra-pixel CV of OCI falls below the other algorithms for the full, mesotrophic, and eutrophic data sets and just above that of GSM alone for the oligotrophic subset. However, in complex and dynamic waters, large natural spatial and temporal variability might be expected and, in those cases CV, cannot be as effectively used as a guidance for model performance. In addition, algorithm saturation at their lowest and/or highest ends (that is, at the boundaries of which retrievals are provided) could also provide misleadingly low CVs.

4. Discussion and conclusions

Restating the specific goals of this study, it aimed to: (1) demonstrate a simple, reliable suite of statistical methods that are appropriate for assessing remote sensing algorithms without *a priori* assumptions of data distributions; and (2) reiterate the need to think critically about statistical analysis and to move beyond the statistical metrics the ocean color community traditionally relies upon that are regularly misinterpreted and sometimes misapplied. While a modern, global evaluation of common SeaWiFS *Chl* data products emerged naturally as a secondary study deliverable, the forthcoming discussion primarily explores goals (1) and (2). A major component of this work is the suggested use of error metrics that avoid sum-of-square error measures, in favor of simple deviation metrics, because of the non-Gaussian error distribution of the case-study data set (and others commonly used in satellite data

Optics EXPRESS

product validation activities) and the desire to minimize the impact of outliers on such analyses [e.g., 11–15]. Although RMSE can often provide similar results to MAE, it will deviate more strongly in the presence of greater extremes in outliers (noting, of course, the utility of RMSE when there is specific interest in the relative error of the outliers). As community interest often focuses on bulk errors in satellite retrievals – most notably, space agencies with requirements to produce the best possible globally-representative data products from multiple satellite missions – this study highlighted MAE and its portability across data sets in lieu of RMSE.

In principle, MAE, RMSE, bias, r^2, and regression slopes all provide useful information for algorithm performance assessment when applied appropriately and interpreted conscientiously. In practice, however, misuse and misinterpretation exist and, following, additional community dialog on error metric best practices and proper reporting and interpretation remains prudent. This work serves only to contribute to a larger conversation to be conducted within the ocean color community. Table 2 reports validation results from our recommended error metrics (bias, MAE, CV, and percent wins), as well as from the commonly adopted metrics of r^2 and regression slope. Put forth here simply as an instructional example, a deficiency of regression slope as a metric emerges in Table 2. The slope for GSM for the combination of all water types, for example, approaches unity (0.99), yet within each water type it exceeds unity (1.4, 1.24, and 1.45 for the oligotrophic, mesotrophic, and eutrophic subsets, respectively). In addition, and as noted earlier, some slopes reported as near unity (e.g., the latter GSM case and OC3 and OCI in eutrophic water) accompany severe biases and poor MAEs. Similarly, while r^2 provides a useful relative ranking, its values can be misinterpreted even for algorithms that perform well, such as in oligotrophic waters as discussed previously. Acknowledging the pedantry of the suggestion, it remains critical for the ocean color community to avoid reporting these values in isolation (that is, without additional metrics such as bias and MAE).

While it may be tempting to continue with MAE, RMSE, and their equivalents (e.g., mean absolute percent error), as well as r^2 and regression slopes, for legacy purposes, the use of too many metrics introduces confusion and redundant metrics can lead to decision partiality [13]. For example, a series of metrics estimating the same performance aspect of an algorithm will repeatedly favor the same algorithm (e.g., MAE and RMSE, which differ primarily through the latter's use of squaring). Redundant metrics also tend to lead decision making towards algorithms that perform best at variability, as there are more metrics for variability than for bias.

Ocean color end-users rely heavily on satellite imagery. The recommended metrics of percent wins and intra-pixel CV provided new insights into algorithm performance by considering algorithm failure and temporal stability in satellite imagery as part of their performance evaluation. Percent wins, the pairwise match-up of residuals, incorporated failure into the computation of relative performance (Table 3). The CV of composite satellite pixels provides a way for considering the coverage and temporal stability of an algorithm (Fig. 4, Table 2). Naturally, their application will vary depending on the end-user requirements. For example, the number of observations might provide a priority metric for certain needs. The intra-pixel CV temporal stability comparisons offer a key piece of information that can address algorithm quality prior to examining satellite-to-*in situ* matchups. If the end-user concern is image products, then spatial/temporal metrics may take precedence over bias or accuracy metrics. This image analysis does not require field observations, such that algorithms can be examined for consistency and variability with satellite imagery alone. Furthermore, the approach has particular value for revealing algorithm differences regarding satellite data processing flags and masking of invalid data, a task that is nearly impossible using *in situ* matchups. The proposed global spatial and temporal metrics and analyses also scale easily to localized regions (Fig. 5), making them especially useful in places where regional satellite-to-*in situ* match-ups remain limited. The

Optics EXPRESS

case study for this focuses on productive water offshore of the U.S. east coast (Fig. 5). Generally speaking, patterns of OCI and GSM *Chl* behave similarly, although some of the potential additional variability in GSM discussed in Section 3 reveals itself (e.g., the spurious high (yellow) value in the upper left corner). Both algorithms report nearly identical sample sizes per spatial bin, yet the patterns in standard deviation differ somewhat strikingly, providing a potentially useful discriminator when prioritizing temporal stability. Note also that this meso-to-eutrophic case study visually suggests somewhat similar performance between OCI and GSM, which is in conflict with the error metrics reported in Section 3. Ultimately, such a conflict indicates a potential regional influence on algorithm performance, thereby reinforcing the importance and value of consistent assessment metrics that scale globally to regionally. Again, it must be kept in mind that in naturally dynamic regions of the ocean, the CV will not give meaningful insight in to algorithm performance, because the real system variability will increase the CV.

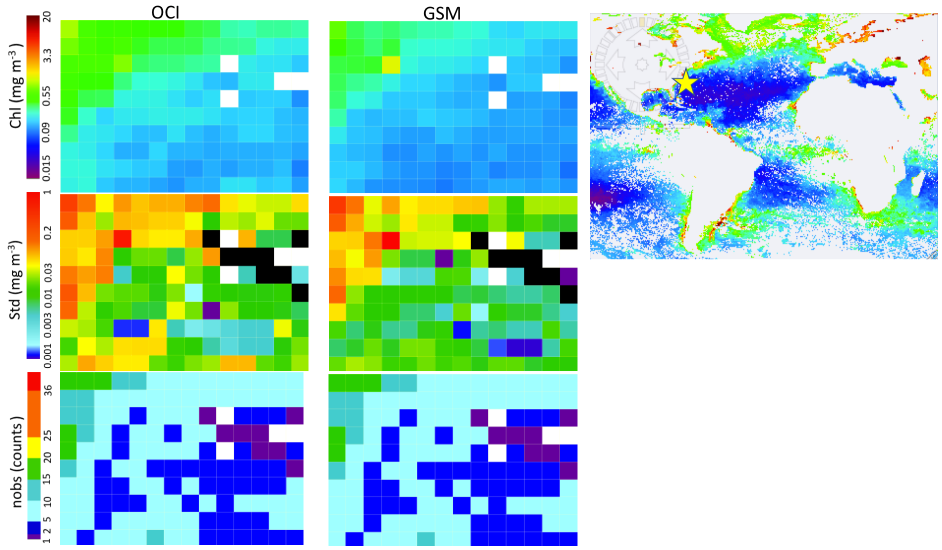

Fig. 5. US central east coast regional images of a SeaWiFS14 day OCI and GSM mean *Chl*, standard deviation, and number of observations (nobs). Additional details in the Fig. 4 caption. The regional image can be helpful in assessing algorithm features at a local level.

Sound graphical analysis remains a key part of any data analysis [39-40]. Scatterplots and residual error plots capture individual metrics. Star plots, however, provide a tool to visually consolidate algorithm performance assessment across all recommended metrics (bias, MAE, percent wins, and CV, and across water types), thus providing a powerful and convenient resource for visual comparison of results and differences (Fig. 6). While the algorithms report comparable performance visually across the full data set, their performance differences reveal themselves more readily for the individual trophic levels. For example, the star plots clearly demonstrate the superior performance of GSM in oligotrophic waters and its lesser performance in eutrophic waters. They also highlight the two dominant discriminators for the mesotrophic subset, namely the pair-wise comparisons and intra-pixel CV. Furthermore, the star plots demonstrate the advantage of moving beyond only reporting generic satellite-to-*in situ* scatter plots. Algorithm performance details remain hidden in scatterplots when data are not examined in additional detail, for example by water types or season, as is often the case in regions with small dynamic ranges of observations [e.g., 35]. The additional visual exploration provided by star plots, for example, assists with identification of patterns in results that might otherwise be

Optics EXPRESS

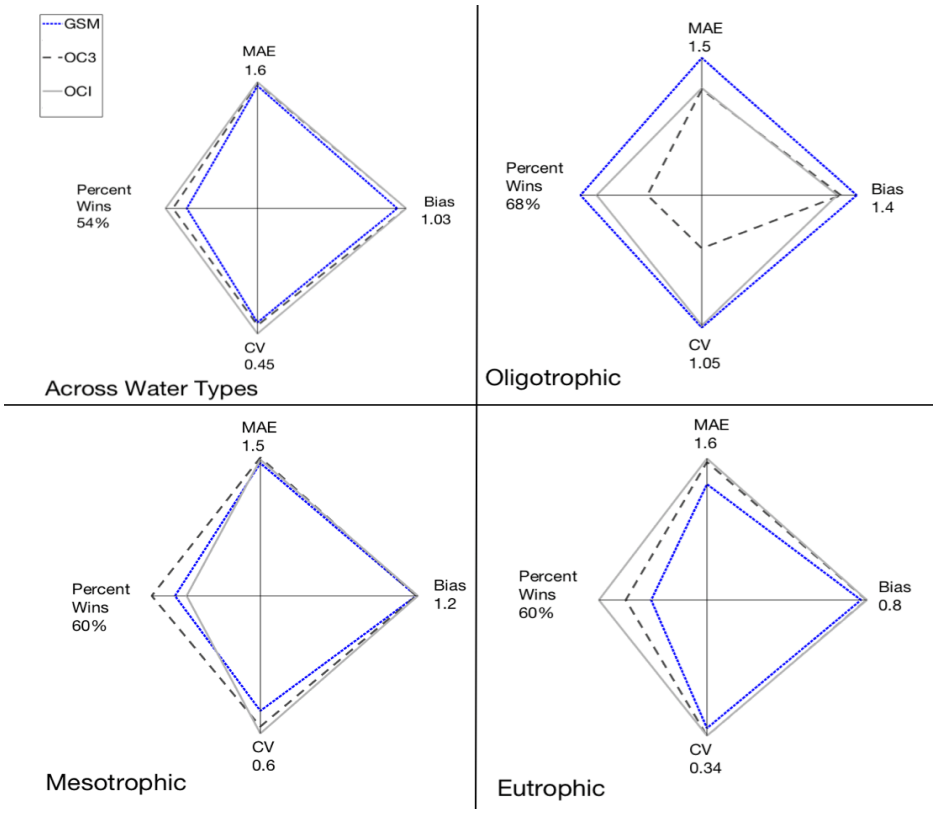

Fig. 6. Comparison of the metrics results of bias, MAE, pairwise percent wins and coefficient of variation summarized in star plots across all water types. The plot center represents values that indicate poor algorithm performance, while farthest from center represents the best performance. The numbers represent the value of the best performing algorithm value for each metric.

overlooked without such a synoptic view. To reiterate a nuance, the importance of thoughtfully scaling the star plot spokes is essential, such that minor differences are not exaggerated, or conversely, that significant, yet small differences, are not overlooked.

Ultimately, the ocean color community desires development and standardization of an objective classification system for algorithm performance that makes use of multiple performance metrics. Again, ideally, metrics with consistent applicability to a range of sample sizes, outliers, and error distributions provide the greatest utility to support space agency validation of satellite data products across missions and regional activities with limited ground-truth data. Brewin et al. [9] developed an objective assessment method assigning points based on algorithm performance with algorithms compared to one another through a suite of statistical measures, acknowledging the limitations and uncertainties of their approach. The use of redundant statistics remains one source of decision partiality in their approach, therefore best practices may evolve to identification of one metric each for statistical bias, precision, and accuracy.

Other objective measures, such as pair-wise comparisons and the use of graphical displays, like star plots, provide additional, independent, and consistent methods for evaluating multiple algorithms without the use of redundant statistics. Stow et al. [16] and Doney et al. [18] also proposed a standard set of metrics for performance evaluation for ecological models with a similar purpose of challenging the field to routinely use a recommended and standardized approach to model assessment. Some metrics overlap with those suggested here (e.g. bias and MAE) however, they diverged from this paper with their

Optics EXPRESS

suggestions of square error metrics (r^2 and RMSE), which may be appropriate for such time-dependent modeling. A consistent set of robust metrics will improve the quality of the analysis and simplify the community assessment of algorithms and their potential utility. Brewin et al. [9] proposed meta-analysis of parameters for ocean color. The pair-wise method can fit nicely into this approach, potentially allowing comparison of the metrics between algorithms.

Precision is not often reported in ocean color methods, most likely because RMSE tends to track random error (more closely than bias). A precision metric consistent with MAE would provide more insight into patterns when the biases are relatively high, but this involves research beyond the scope of this paper. Generally speaking, there has been a lack of discussion of precision metrics for non-Gaussian distributions, although the assessment of random error could be useful to algorithm performance assessment. In contrast, precision metrics for distributions with Gaussian errors exist. Briefly, mean square error can be readily partitioned as the sum of bias^2 + precision^2 (with similar portioning available for related statistics based on standard error). Currently, however, there is no equivalent method for determining precision metrics reported for MAE [8,44].

Finally, we recommend that all future validation and algorithm comparison studies clearly and unequivocally explain the rationale for the *a priori* selection of metrics (in particular, the consideration of redundant metrics if pursued) and the path and/or steps used to identify the best performer. This will provide additional clarity to the subsequent reader, and also reinforce critical interpretation of results by the researchers. The end-user, based on his/her research question(s) and available resources, may need to select specific metrics that are higher or lower priority. One algorithm might be incrementally superior to another in reported metrics, but the performance superiority may be too small to be of consequence, especially if differences exist in the implementation of the algorithm (e.g., its computational efficiency). Furthermore, nuances remain in the choice of metrics that require user evaluation. Data sets of limited size but large dynamic range, for example, may require alternate metric formulations. In this case, biases might be reported as the median of the differences instead of the mean, just as, the median of the absolute error might be used instead of the mean. Finally, constraining and preselecting the metrics will also reduce the risk of decision prejudice, thus avoiding the selection of the metric that favors a preferred result.

In summary, a generic summary of recommendations for ocean color validation activities:

(1) Identify the end-user/application criteria to be used and priorities to be applied in performance assessment and best performer identification in order to identify the appropriate metrics.

(2) Subsequently, report the rationale for all decisions and metrics when documenting results.

(3) Apply quality assurance and control best practices to the data sets, both reference and model.

(4) Use error metrics that are statistically robust for non-Gaussian data, such as metrics based on absolute deviation rather than those based on mean square error (or slope). Or, demonstrate the appropriateness of other metrics.

(5) Select no more than one metric for each estimate of bias, accuracy, and precision to reduce the likelihood of decision bias caused by redundant metrics.

(6) Use metrics that inform on algorithm temporal and spatial stability, which are not typically captured in #4.

(7) Include additional objective metrics such as pair-wise comparisons (percent wins) to capture relative performance and aid in decision support (e.g., Brewin et al., 2015).

Optics EXPRESS

(8) Generate decision graphics, such as star plots, in addition to traditional scatter (and residual error) plots to offer synoptic visualization of all considered metrics.

In conclusion, traditional approaches for ocean color algorithm performance assessment rely heavily on commonly used, but not necessarily appropriate, statistical metrics. A methodology combining metrics and graphics is essential in addition to considering end-users criteria for the assessment. No single metric covers all performance criteria and therefore combining metrics is necessary. This study demonstrated and provides recommendations for an alternative, straightforward and robust approach for evaluating and comparing ocean color algorithms, specifically bias, MAE, percent wins, and intra-pixel CV. The goal of this study was not to provide the final word on metrics for satellite validation, but rather contribute to an emerging, larger community-wide conversation on satellite algorithm performance assessment. and underline the necessity to critically think about model evaluation.

Funding

This project has been funded by the NASA Ocean Biology and Biogeochemistry Program/Applied Sciences Program under proposal 14-SMDUNSOL14- 0001 and by EPA, NOAA, and USGS.

Acknowledgments

We thank Sean Bailey, Jason Lefler, Tommy Owens, Daniel Knowles Jr., Christopher Proctor, and Joel Scott for their invaluable advice, insight, and assistance. We also thank three anonymous reviewers for their detailed and conscientious comments. This article has been reviewed by the National Exposure Research Laboratory and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the U.S. Government. The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.