

Comparing level-2 and level-3 satellite ocean color retrieval validation methodologies

Joel P. Scott^{1,2*} and P. Jeremy Werdell¹

¹ NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Greenbelt, Maryland 20771, USA

² Science Applications International Corporation (SAIC), Reston, VA, USA

* joel.p.scott@nasa.gov

Abstract: Many ocean color data applications leverage global spatially composited Level-3 (L3) satellite data, because of their regular Earth-grid frame of reference. However, ocean color satellite retrieval performance is routinely evaluated on Level-2 (L2) data at the native satellite swath resolution and geometries. This study assesses how accurately binned and gridded L3 data represent L2 satellite data products via satellite-to-in situ match-up activities. L2 and L3 satellite data retrievals of the photosynthetic pigment chlorophyll-a are compared with a common in situ data set, revealing similar L2 and L3 satellite-to-in situ performance for both MODIS-Aqua and VIIRS-SNPP. This agreement implies that L2 validation results are generally applicable to L3 data. However, uncertainties are introduced during the generation of L3 data from L2 data. L3 data comparisons introduce a wider temporal window between the time of in situ measurement and the time of the satellite observation, which can unintentionally reflect on the quality of the satellite retrieval or algorithm performance. The choice of L3 map projection may introduce additional uncertainty by spatially distorting the true location of the satellite retrievals. Each manipulation of satellite data beyond the instrument's native spatiotemporal reference (L2) reduces the applicability of L2 validation results to higher data processing levels.

© 2019 Optical Society of America under the terms of the [OSA Open Access Publishing Agreement](#)

OCIS codes: (010.0010) Atmospheric and oceanic optics; (010.0280) Remote sensing and sensors

1. Introduction

The continuous global data record from ocean color satellites now exceeds twenty years. In this time, the oceanographic community invested heavily in the application, improvement, and evaluation of this data record [1-3]. In brief, ocean color measurements involve the estimation of water-leaving reflectance spectra from which geophysical and spectral optical quantities of interest can be derived [4]. Example derived parameters include concentrations of the photosynthetic pigment chlorophyll-a (*Chl*; $\mu\text{g/L}$), aquatic optical properties, and metrics of light penetration and water quality. Space-based estimates of ocean color provide frequent, continuous, and high-resolution measurements of these geophysical parameters on regional and global scales. Such satellite data records facilitate addressing both research and operational applications associated with aquatic ecosystem dynamics, carbon exchanges and export, phytoplankton community composition, and fisheries and watershed management, to name only a few [5-6]. Naturally, inclusion of satellite-derived data products into any such activity requires routine performance assessment and evaluation, particularly as satellite instruments age, as algorithms and suites of science data products evolve, and as new ocean color missions launch and join the growing constellation of Earth system satellites, often with differing design concepts.

The satellite data used in this study are from the NASA Moderate Resolution Imaging Spectroradiometer onboard Aqua (MODIS-Aqua) [7-9] and the Visible Infrared Radiometer Suite onboard Suomi NPP (VIIRS-SNPP) [10-12]. In this paper onward, all references to

MODIS and VIIRS will refer to MODIS-Aqua and VIIRS-SNPP, respectively. As an overview, both satellite instruments observe visible and near-infrared radiances at several discrete wavelengths at the top-of-the-atmosphere. Contributions to this signal, from the atmosphere and sea surface (e.g., Sun glint and white caps), are removed via atmospheric correction algorithms to calculate spectral remote sensing reflectance ($R_{rs}(\lambda)$; sr^{-1}), the light exiting the water column normalized to the incident surface irradiance [4]. Geophysical algorithms are applied to the $R_{rs}(\lambda)$ to derive estimates of bio-optical parameters, such as *Chl*. In standard NASA processing, the ocean chlorophyll index (OCI) approach [13] is used to produce MODIS and VIIRS *Chl* retrievals.

Ocean color satellite data processing levels are typically defined as:

- Level 0: Uncalibrated science data (engineering units) in a raw format
- Level-1A: Uncalibrated science data in an archive format
- Level-1B: Calibrated, geolocated science data, as observed
- Level-2: Derived geophysical science products, as observed
- Level-3: Derived geophysical science products, temporally/spatially composited

In their standard ocean color data processing, NASA (as well as other space agencies) executes the atmospheric correction process on Level-1B data to derive Level-2 $R_{rs}(\lambda)$, *Chl*, and other geophysical science data products. These Level-2 (L2) data remain in the native satellite swath coordinate system, not having yet been projected onto a standard grid, nor having been temporally composited. Most common satellite-to-in situ validation exercises and performance assessments evaluate L2 data products [14-16]. The NASA SeaWiFS Bio-optical Archive and Storage System (SeaBASS) [17] provides one example of a data repository that routinely provides L2 satellite-to-in situ validation results for $R_{rs}(\lambda)$, *Chl*, and other science data products.

Level-3 (L3) satellite data refer to geophysical satellite data retrievals that have been temporally and spatially aggregated onto a common geo-referenced Earth grid over a defined period of time. NASA's aggregation methodology follows the protocol set forth for the SeaWiFS ocean color mission [18-19]. For a L3 grid cell to be assigned a value, at least one L2 geophysical retrieval data point must pass a set of quality control thresholds, fall spatially within the defined bounds of the L3 grid cell, and locate temporally within the assigned aggregation time-window, which is defined as daily in this study. The set of quality control thresholds applied to generate the L3 data is defined by the flags set in the ocean color data product *l2flags* created during the NASA atmospheric correction (Level-1B to Level-2) process (Table 1). NASA L3 files report summary statistics for grid cells that have multiple contributing observations, including arithmetic means, standard deviations, observation counts, and scene counts.

NASA's ocean color L3 process creates two types of files: an intermediate file (L3b) is created from one or more L2 files, and a standard map (L3m) from the L3b file. L3b files are georeferenced onto a sinusoidal Earth grid (Fig. 1), as an approximate attempt to generate an equal-area grid for the Earth that is most representative of the satellite retrievals and the geographic and spatial area the observations describe. The meridians of this type of grid are defined to be sinusoids with equally-spaced parallels, such that the distances along the parallels are conserved [20]. There are a few defining characteristics of the ocean color sinusoidal L3b Earth-grid [18]: 1.) the individual grid cells are aligned as zonal rows with parallels of latitude defining the north and south edges; 2.) rows of cells begin and end along the 180° meridian; 3.) each row has a discrete number of cells that varies with the row's cosine of the central latitude, with the two polar rows being defined to each have three cells; and, 4.) there are an even number of rows so that the northern and southern hemispheres are

separable at the equator. Additional information on the NASA L3b data format can be found at <https://oceancolor.gsfc.nasa.gov/docs/format/l3bins/>.

Table 1. Summary of l2flags applied when generating comparison match-ups with the L2 data and when generating L3b and L3m data products.

Flag Name*	Applied to Level-2 Validation	Applied to Level-3 Construction	Flag description
ATMFAIL	X	X	Atmospheric correction failure
LAND	X	X	Pixel is over land
HIGLINT	X	X	Sunglint detected via reflectance exceeding threshold
HILT	X	X	Observed radiance very high or saturated
HISATZEN	X	X	Sensor view zenith angle exceeds threshold of 60-degrees
STRAYLIGHT	X	X	Probable stray light contamination
CLDICE	X	X	Probable cloud or ice contamination
COCCOLITH		X	Coccolithophores detected
HISOLZEN	X	X	Solar zenith exceeds threshold of 70-degrees
LOWLW	X	X	Very low water-leaving radiance
CHLFAIL	X	X	Satellite chlorophyll algorithm failure
NAVWARN	X	X	Satellite navigation quality is suspect
ABSAER		X	Absorbing Aerosols determined
MAXAERITER		X	Maximum iterations reached for NIR iteration
CHLWARN		X	Satellite-retrieved chlorophyll out-of-bounds
ATMWARN		X	Atmospheric correction is suspect
NAVFAIL	X	X	Satellite navigation failure
FILTER	X	X	Insufficient data for smoothing filter

* this is not a comprehensive list of all l2flags

L3m data are spatial re-projections of L3b data [18]. L3m data can exist in multiple Earth-grid representations, which provide customizable utility for Earth system modeling studies [21-22]. In standard NASA processing, L3m files are georeferenced onto a Plate Carrée map projection instead of the L3b sinusoidal Earth-grid (Fig. 1). While this type of Earth-grid is often called an equidistant cylindrical or rectangular grid, it should be noted that a Plate Carrée grid is a special type of equidistant cylindrical Earth-grid that has the equator as the standard parallel. The Plate Carrée map projection is desirable since the distances along the meridians are conserved [20]. However, the L3m Earth-grid distorts the representation of spatial area, increasing as a function of latitude, providing a less-representative geographic portrayal of the Earth toward the poles [23].

Most validation exercises focus on L2 products and the statistics from such exercises are those most often reported. Additionally, performance requirements for heritage and future ocean color missions (e.g., for the upcoming NASA Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission) [24] are often defined for L2 products. However, many critical Earth system analyses rely on global L3 maps, which are also the most accessible satellite data products for non-expert users, and such studies require some assessment of satellite L3 data product performance. While some L3 data product performance assessments exist [25], this paper further explores whether L2 validation statistics adequately represent L3 data product performance. This is accomplished by comparing conventional L2 satellite-to-in situ match-up results [14] with those similarly derived using L3 data. The MODIS and VIIRS OCI-derived *Chl* data products are used as the case-study, noting that a rigorous evaluation of their individual and cross-instrument performances is not our goal. Rather, this paper focuses

on evaluation of satellite data product evaluation methodologies and, specifically, the adequacy of including L2 performance metrics into studies using L3 data products.

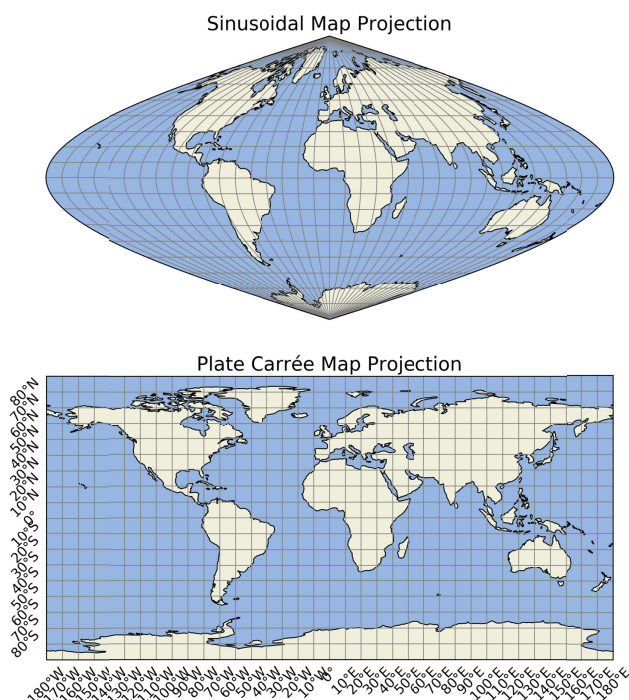


Fig. 1. A sinusoidal map-projection is shown in the upper panel, and a Plate Carrée equidistant cylindrical map projection is shown in lower panel. Note that both earth-grids are centered on the prime meridian (0-degrees longitude) and have parallels and meridians spaced 10-degrees apart at the equator. While these grids are much coarser than the 4.62-km spacing used in the L3b and L3m data under consideration by this study, the difference in spacing between the meridians at middle and high latitudes illustrate the spatial distortion that occurs for the grid cells in the Plate Carrée projection at high latitudes compared to the sinusoidal grid cells.

2. Methods

2.1 Data acquisition

The MODIS and VIIRS *Chl* data products used in this study were acquired from the NASA Ocean Biology Processing Group (OBPG). All satellite data were processed using their latest 2018 reprocessing configurations. The in situ *Chl* data were acquired from SeaBASS [26]. These in situ data were collected during the Atlantic Meridional Transect (AMT24) cruise that began off the coast Europe on 2014-09-25 09:39 UTC at (49.83° N, 4.76° W) and concluded on 2014-11-01 18:51 UTC at (50.40° S, 54.83° W), sampling both northern and southern Atlantic Ocean basins (Fig. 2; data courtesy of William Balch, Bigelow Laboratory for Ocean Sciences). This study specifically considers chlorophyll fluorescence measurements from a flow-through bio-optical system fed with de-bubbled water from the ship sea chest. AMT24 fluorometric data were chosen for several key reasons: 1.) they offer high temporal sampling from the flow-through system, at approximately 10-minute intervals, providing a high density of in situ measurements to serve as targets for generating satellite match-ups; 2.) they encompass a wide dynamic range of observed *Chl*, from oligotrophic ($Chl \leq 0.1 \mu\text{g/L}$) to

mesotrophic ($0.1 < Chl \leq 1 \mu\text{g/L}$) conditions, and variety of latitudes in multiple ocean basins, including the tropics and mid-latitudes, as well as both the Northern and Southern Hemispheres; and, 3.) they coincide with both MODIS and VIIRS retrievals, with sufficient cruise duration (37 days) to potentially generate a high number of validation match-ups with both satellite instruments. We intentionally limited our in situ data to a single cruise to avoid cruise-to-cruise differences in data collection and/or quality assurance influencing our results.

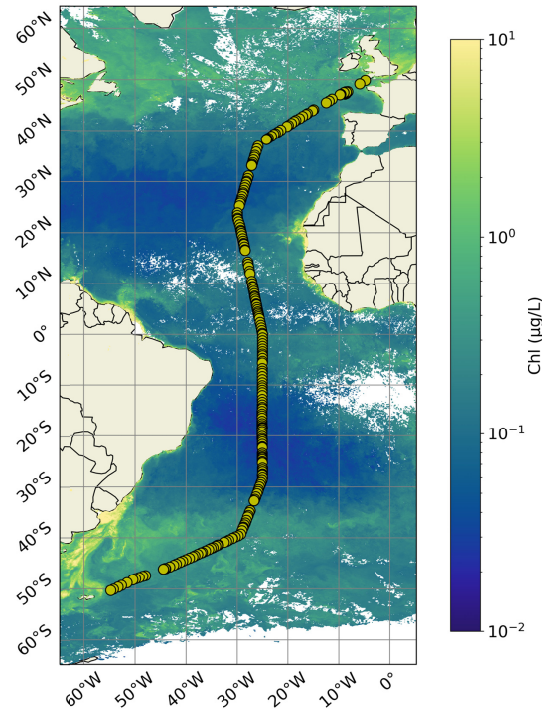


Fig. 2. Locations of AMT24 in situ chlorophyll fluorescence sampling. The background shows October 2014 monthly L3m MODIS-Aqua *Chl*.

We acknowledge that flow-through fluorometric *Chl* measurements may not be considered as ideally-suited for a formal validation study for two reasons. First, this type of measurement, during cruises in excess of 30 days, often experiences biofouling, which is the deterioration of fluorometric optical instruments' performance due to continuous growth of biofilm on the optical windows [27]. Second, fluorometric *Chl* measurements are susceptible to non-photochemical quenching (NPQ) of the fluorescence signal, where the quantum yield of chlorophyll fluorescence is lower than nighttime values due to high diel irradiance causing a contraction of phytoplankton chloroplasts [28-29]. NPQ typically occurs during the midday when a potential satellite validation matchup might occur, with the overhead passage of both MODIS-Aqua and VIIRS-SNPP occurring at approximately 13:30 local time. That said, as these in situ data serve as the common comparison standard for all satellite-to-in situ matchups, both temporal drifts due to biofouling and diel effects of NPQ can be ignored for the purposes of this comparative study since both signals will be consistently present across all satellite-to-in situ comparisons. This study should not be considered as a validation analysis that assesses satellite and algorithm performance. Rather, the authors intend this work to serve simply as a comparison of validation methodologies and performance assessment metrics for L2 and L3 satellite data.

2.2 Satellite-to-in situ match-ups

The L2 satellite-to-in situ *Chl* match-ups were generated following the methods and exclusion criteria [14], namely: 1.) temporal coincidence was defined as ± 3 hours of the time of collection for the in situ target; 2.) satellite values were the filtered mean of all unmasked pixels in a 5×5 box centered on the in situ target; and 3.) satellite values were excluded when the coefficient of variation within this box exceeded 0.15 or fewer than 13 pixels (that is, $< 50\%$ of the original 25 pixels) remained unmasked. Satellite pixels were masked under any of the following circumstances: 1.) the trigger of certain l2flags (see Table 1); 2.) the solar zenith angle exceeding 70° ; and, 3.) the satellite zenith angle exceeding 60° . In situ measurements were averaged when multiple measurements were found to fall spatially within a given matching satellite pixel box. While in situ sampling patterns that result in multiple measurements per satellite pixel are useful for quantifying sub-pixel variability, that topic is beyond the scope of this study.

L3 satellite-to-in situ *Chl* match-ups were generated following a similar, but subtly different, process. Exclusion criteria were not applied to the satellite retrievals, as quality control metrics are applied in the generation of standard NASA L3 files in the form of l2flag masking (see Table 1). L3 grid cells, containing an in situ target, cover roughly 21.3 km^2 and were extracted individually. In practice, this resulted in a simpler satellite-to-in situ matchup strategy, with only three basic steps: 1.) the in situ target date and time were used to identify the corresponding L3 daily file; 2.) the appropriate Earth-grid cell containing the in situ target latitude and longitude was identified using a grid-definition look-up table; and 3.) L3 summary statistics (e.g., *Chl* mean and standard deviation for each grid cell, the number of observations included in each grid cell (N_{obs}), and the number of input scenes included in each grid cell (N_{scenes})) were extracted via l3bindump, a component of the Ocean Color Science SoftWare (OCSSW) package publicly distributed within the NASA SeaWiFS Data Analysis System (SeaDAS).

The satellite-to-in situ matchup methodology remained identical for both L3b and L3m data, with the only differences in results arising from the related Earth-grids assigned to each file; sinusoidal versus Plate Carrée, respectively (Fig 1.). As for the L2 match-ups, in situ measurements were averaged when multiple targets were found to fall spatially within a given matching satellite grid cell.

2.3 Performance metrics

Two common error metrics for satellite performance assessment were calculated: the mean bias and the mean absolute error (MAE) [16]. The mean bias serves as a descriptor of systematic error and provides a useful assessment of whether, on average, the satellite retrievals under- or over-estimated the geophysical retrievals [30]. The MAE describes the absolute errors between the in situ and satellite values, which we note have non-Gaussian distributions (Fig. 3). Mean bias and MAE are defined as:

$$\text{mean bias} = 10^{\left(\frac{\sum_{i=1}^n \log_{10}(R_i) - \log_{10}(O_i)}{n} \right)} \text{ and} \quad (1)$$

$$\text{MAE} = 10^{\left(\frac{\sum_{i=1}^n |\log_{10}(R_i) - \log_{10}(O_i)|}{n} \right)}, \quad (2)$$

where R represents the remotely-sensed satellite retrieval, O represents the in situ observation, and n represents the sample size. We also define unbiased percent difference (UPD) as:

$$\text{UPD} = 200\% \left(\frac{\text{value}_2 - \text{value}_1}{\text{value}_2 + \text{value}_1} \right), \quad (3)$$

which we use to compare performance metrics.

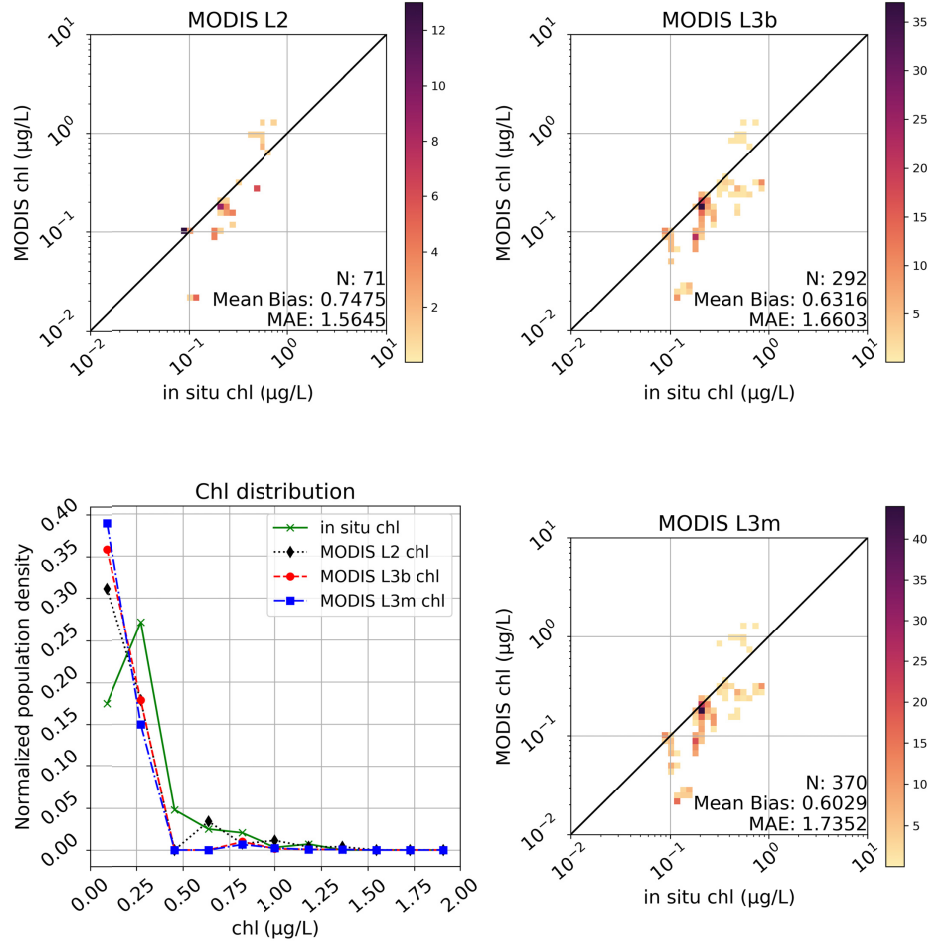


Fig. 3. Comparisons of MODIS and in situ *Chl*. The upper right, upper left, and lower left panels show satellite-to-in situ scatterplots and comparison metrics for L2, L3b, and L3m data, respectively. The solid black lines indicate a 1:1 relationship. Colors report the numerical density of match-ups in a region of the scatter plot. The lower left panel shows normalized, binned population distributions for the entire in situ *Chl* data set (green X's) and the matched-up MODIS L2 (black diamonds), L3b (red circles), and L3m (blue squares) *Chl* retrievals.

Chl is commonly presented in log-transformed space, as its data value distribution exists across several orders of magnitude and its associated uncertainties vary proportionally with data value [31, 16]. Since *Chl* values undergo transformation into $\log_{10}$ -space before error metrics are computed Eqs. (1)-(2), the metrics themselves convert from their linear form to their multiplicative form, which is dimensionless and indicates relative error as a percentage of the measurement. For mean bias, values closer to unity show the least bias, with values below unity corresponding to a negative bias (e.g., a mean bias 0.89 represents an 11% negative bias, whereas a mean bias of 1.11 represents an 11% positive bias). MAE are always positive, with a value of 1.5, for example, indicating that the satellite observations are 1.5x (50% greater) on average than the in situ observations.

3. Results

Satellite-to-in situ L2 and L3b *Chl* match-ups for both MODIS and VIIRS show reasonable agreement regarding data distribution and spread with mean biases and MAEs of similar orders of magnitude (Figs. 3 and 4, Table 2). While this should be somewhat expected, as L3b data are derived from L2 satellite retrievals, the influence of different sample sizes merits further exploration. The number of L3b match-ups (MODIS: 292, VIIRS: 283) exceeds the number of L2 match-ups (MODIS: 71, VIIRS: 67) by over a factor of four, due in part to the wider temporal window allowed by L3b daily data. This influences interpretation of performance metrics, as our selected metrics are sensitive to the number of samples and where those samples fall within the full dynamic range of *Chl*. Both MODIS and VIIRS L3b results include far more match-ups near the center of the dynamic range (~ 0.1 to ~ 0.2 $\mu\text{g/L}$), many of which show clear negative biases. Relative to L2, this results in mean biases farther from unity. MODIS and VIIRS biases increased from 25.2% for L2 to 36.8% for L3b and from 30.7% for L2 to 33.4% for L3b, respectively. For MODIS, this also results in increased MAEs, rising from 56.5% for L2 to 66.0% for L3b. VIIRS MAEs, however, fall from 62.4% for L2 to 57.0% for L3b. Ultimately, considering all available samples, the MODIS and VIIRS L2 and L3b results show somewhat contrasting performances. On average, the UPDs between MODIS L2 and L3b percent biases and MAEs differ by $\sim 37\%$ and $\sim 16\%$, respectively. For comparison, the UPDs between VIIRS L2 and L3b percent biases and MAEs differ by $\sim 8\%$ and $\sim 9\%$, respectively.

Table 2. Summary of validation comparison metrics.

		Fig/Panel	#	Mean bias	MAE
MODIS-Aqua	in situ vs L2	3A	71	0.7475	1.5645
	in situ vs L3b	3B	292	0.6316	1.6603
	in situ vs L3m	3D	370	0.6029	1.7352
	in situ vs L2 (where L2/L3b overlap)	5A	47	0.7540	1.6354
	in situ vs L3b (where L2/L3b overlap)	5B	47	0.7540	1.6346
	in situ vs L3b (where L2 failed)	5D	245	0.6115	1.6646
	in situ vs L3b (where L3b/L3m overlap)	6A	252	0.6326	1.6668
	in situ vs L3m (where L3b/L3m overlap)	6B	252	0.6319	1.6703
	in situ vs L3m (where L3b failed)	6C	118	0.5467	1.8911
VIIRS-SNPP	in situ vs L2	4A	67	0.6930	1.6242
	in situ vs L3b	4B	283	0.6656	1.5705
	in situ vs L3m	4D	357	0.6355	1.6498

The subset of MODIS match-ups with both a L2 and L3b retrieval (N=47) yielded nearly identical performance metrics for both data types, namely a bias of 24.6% and a MAE of 63.5% (Fig. 5, Table 2). The subset of MODIS L3b match-ups without a corresponding L2 match-up (N=245), however, showed considerably inferior performance relative to the common match-ups. These L3b match-ups yielded a much higher mean bias of 38.9% and a slightly higher MAE of 66.5%. Eighty-nine percent of these 245 targets failed to generate a L2 match-up due to time differences that exceeded the ± 3 -hour L2 match-up window (Fig. 5). The remaining 11% failed because the exclusion criteria applied to a given L2 5x5 satellite pixel box eliminated more than 13 pixels (that is, $>50\%$ of the available 25 pixels in the box), thereby disqualifying the satellite retrieval for consideration as a valid match-up. As MODIS and VIIRS performed similarly in this analysis, only MODIS is presented for brevity.

Satellite-to-in situ L3b and L3m *Chl* match-ups also show reasonable agreement regarding data distribution and spread with similar performance metrics (Figs. 3 and 4, Table 2). The sample sizes for L3m match-ups (MODIS: 370, VIIRS: 357) exceed those for the L3b match-

ups, but only by ~26%. MODIS and VIIRS biases modestly increased from 36.8% for L3b to 39.7% for L3m and from 33.4% for L3b to 36.5% for L3m, respectively. MODIS and VIIRS MAEs also increased from 66.0% for L3b to 73.5% for L3m and from 57.0% for L3b to 65.0% for L3m, respectively. Considering all available samples, the MODIS and VIIRS L2 and L3b results show somewhat similar performances, unlike those for the L2 versus L3b results. On average, the UPDs between MODIS L3b and L3m percent biases and MAEs differ by only ~8 and ~11%, respectively. Similarly, the UPDs between VIIRS L3b and L3m percent biases and MAEs only differ by ~10 and ~13%, respectively.

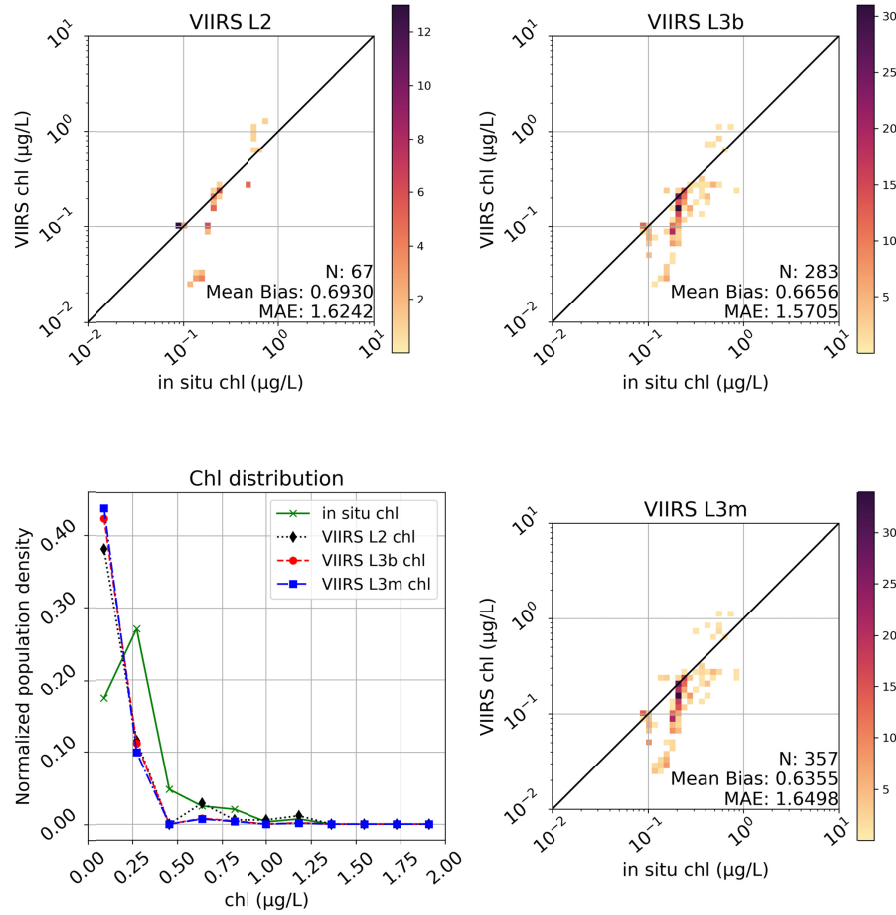


Fig. 4. As in Fig. 3, but for VIIRS.

The subset of MODIS match-ups with both a L3b and L3m retrieval (N=252) yielded nearly identical performance metrics for both data types, with percent biases and MAEs agreeing to within 0.1% and 0.3%, respectively (Fig. 6, Table 2). The subset of MODIS L3m match-ups without a corresponding L3b match-up (N=118), however, showed considerably inferior performance relative to the common match-ups. These L3m match-ups yielded a somewhat higher bias of 45.2% compared to 36.7% and 36.8% for the common match-ups and a much higher MAE of 88.6% compared to 66.7% and 67.0% for the common match-ups.

4. Discussion

Our discussion of results applies strictly to the L2 and L3 satellite-to-in situ match-up and spatial re-gridding and re-projecting paradigms described above. The application of alternate

temporal coincidence windows, homogeneity tests (described here through the coefficient of variation of a satellite pixel box), or other exclusion criteria will influence the performance metrics realized in a L2 satellite-to-in situ match-up analysis. Similarly, the use of alternate map-grids, statistics, or averaging techniques within a spatial or temporal compositing scheme will influence the values reported in a L3 data file. That said, our results can effectively be used to introduce and generally discuss the differences that may be realized when comparing performance metrics derived from native swath coordinate systems (L2 data) with those from spatially re-gridded or temporally composited data streams (L3 data).

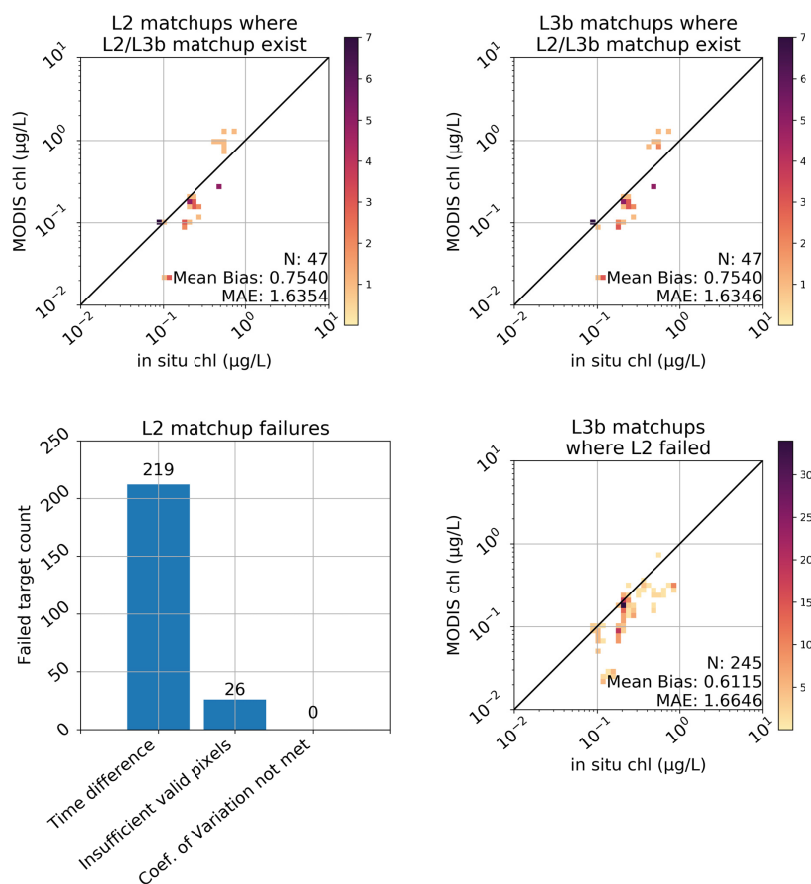


Fig. 5. Subsets of the MODIS and in situ *Chl* comparisons. The upper left panel shows only those L2 match-ups where the in situ targets also generate L3b match-ups. The upper right panel shows only those L3b match-ups where the in situ targets also generate L2 match-ups. The lower right panel shows only those L3b match-ups where the in situ targets did not generate an L2 match-up. The solid black lines indicate a 1:1 relationship. Colors report the numerical density of match-ups in a region of the scatter plot. The lower left panel shows a histogram of L2 matchup exclusion criteria that led to the in situ targets shown in the lower right panel failing to generate a L2 match-up.

Broadly speaking, performance metrics derived from L2 satellite-to-in situ match-ups are portable to daily L3 data, as demonstrated by evaluation of common in situ targets (Fig. 5). This effectively forces the ± 3 -hour temporal criteria of the L2 matchup methodology [14] onto the L3 match-ups and suggests that, at least for the dataset considered in this analysis, the L2 5x5 pixel box and individual L3 grid cell similarly represent each target's spatial footprint and geophysical conditions over this temporal range.

However, performance metrics derived from consideration of all available L3 satellite-to-in situ match-ups do not unequivocally represent those derived from L2 analyses. For MODIS, the L3b bias and MAE increased by 11.6% and 9.5%, respectively, relative to the L2 results. Whereas for VIIRS, the L3b bias increased by 2.7%, but the MAE decreased by 5.4%. This suggests that L3 match-up analyses do not consistently replicate the systematic and random errors quantified in L2 match-up analyses.

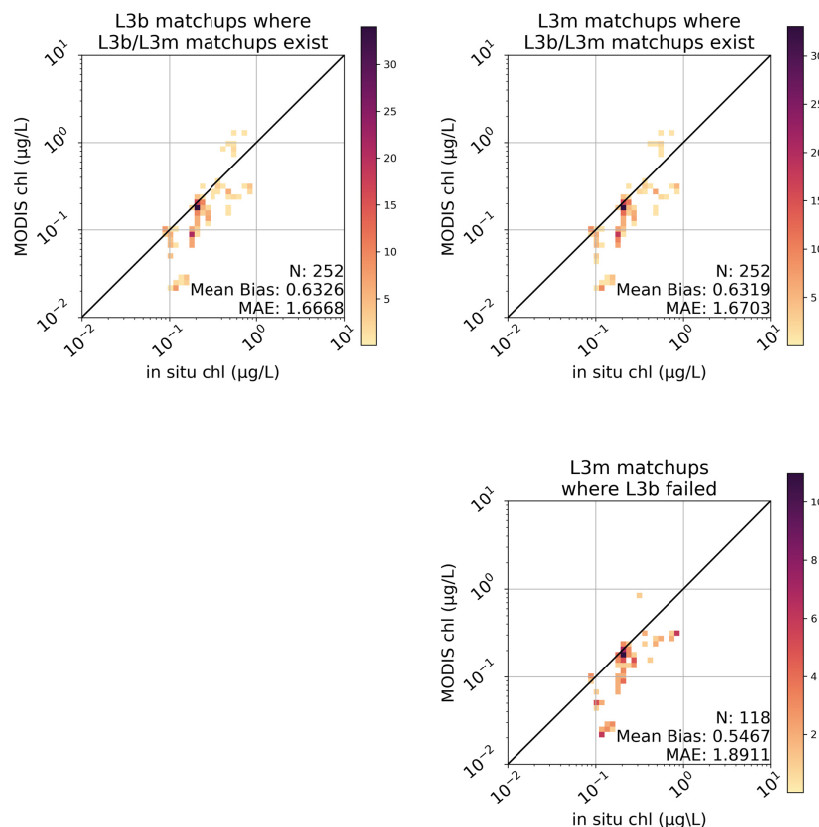


Fig. 6. Subsets of the MODIS and in situ Chl comparisons. The upper left panel shows only the L3m match-ups where the in situ targets did not generate L3b match-ups. The upper right panel shows only the L3b match-ups where the in situ targets also generate L3m match-ups. The lower right panel shows only the L3m match-ups where the in situ targets also generate L3b match-ups. The solid black lines indicate a 1:1 relationship. Colors report the numerical density of match-ups in a region of the scatter plot.

The differences in L2 and L3b performance metrics ultimately resulted from the inclusion of additional in situ targets in the L3 analyses. Additional L3 match-up targets in this study resided in both common and different geographic regions relative to the L2 match-up targets (Fig. 7). For both MODIS and VIIRS, the 30° to 50°N latitudinal band realized the majority of additional L3 match-ups – a region with many L2 match-ups as well. Additional L3 match-up targets were also identified in regions without any L2 match-ups, for example, from 25° to 30°N and -15° to -20°S for MODIS and from -5° to -20°S for VIIRS.

Additional L3 match-ups are generated for two reasons. First, consideration of all available L3 satellite-to-in situ match-ups effectively opens the allowed time window to same-day. We considered daily L3 data in this analysis; naturally, the use of longer temporal composites would open the time window further. Second, components of the exclusion

criteria applied to L2 satellite-to-in situ match-ups in NASA processing [14] cannot be easily replicated in L3 match-up analyses – in particular those criteria requiring the validity of >50% of satellite pixels in the targeted 5x5 box.

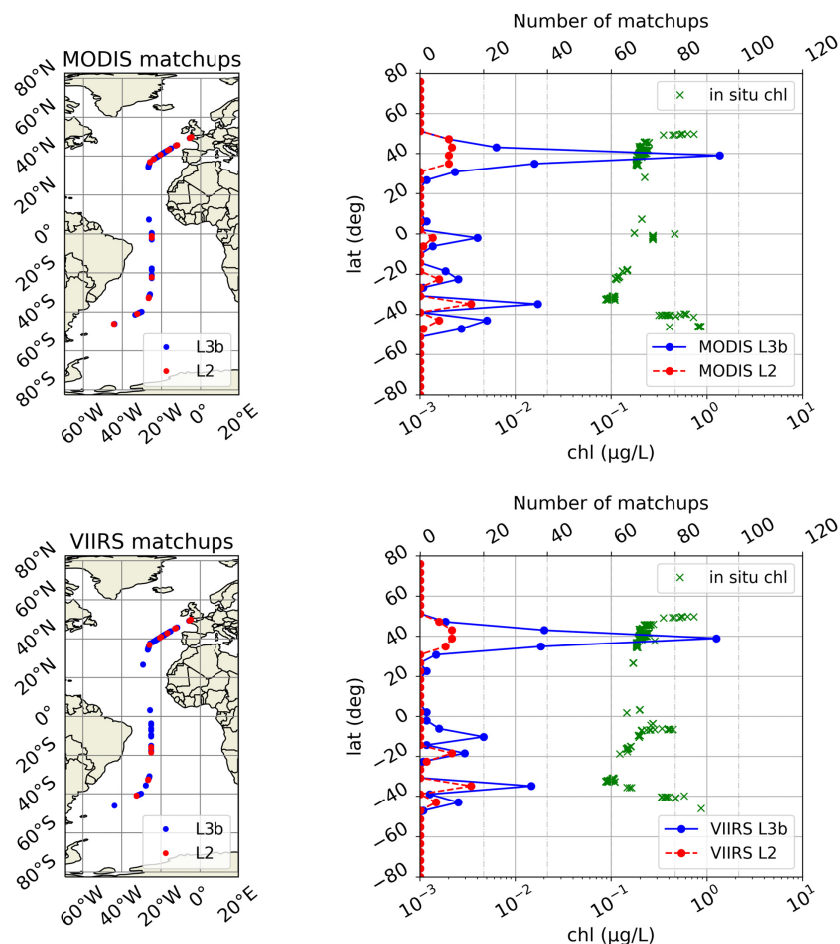


Fig. 7. Geographic distributions of MODIS and VIIRS satellite-to-in situ match-ups. The upper two panels show a comparison of the spatial location (left panel) and latitudinal frequency distribution (right panel) of MODIS L2 and L3b match-ups. The lower panels show the same for VIIRS. Binned in situ *Chl* data are shown on the secondary X-axis as green x's of the right panels.

With regards to the first consideration, 89% (N=219) valid L3b match-ups without a corresponding L2 match-up had satellite-to-in situ time differences in excess of 3 hours (thus, triggering one of the L2 match-up exclusion criteria). While one temporal match-up exclusion criterion cannot universally represent all appropriate time scales for all biogeophysical processes across all water masses, it remains conceivable that the consideration of two measurements that differ in time by 3 or more hours will introduce additional uncertainties in the match-up analyses. As an example, relaxing the MODIS L2 temporal exclusion to a same-day threshold increased the satellite-to-in situ bias and MAE by 19.9% and 39.0%, respectively.

With regards to the second consideration that there are more L3 matchups than L2, the remaining 11% of L2 match-up failures (relative to L3b) arose from L2 exclusion criteria

requiring that at least 50% of the 5x5 pixel extract of satellite data be valid and pass l2flag criteria (Table 1, Fig. 5). By contrast, the L3b compositing methodology only requires a single swath-level satellite pixel to pass the l2flag filtering criteria in order to create a valid value in a L3b grid cell, which is then available to become a L3b match-up with an in situ target. In the most extreme case, this is analogous to only considering a single swath-level satellite pixel within a 2-pixel radius to be valid for use in a L2 match-up analysis. For those L3b match-ups without a corresponding L2 match-up, N_{obs} ranged from 1 to 20, with over half having $N_{obs} < 12$. This suggests that in situ targets that create an L3b matchup, but not an L2, are more likely to be made under less optimal conditions for a satellite retrieval given potential influences by clouds and other adjacency effects and non-homogenous waters in the surrounding pixels. Each L3 grid cell contains a measure of the variance of the binned retrievals, reported as a standard deviation, which may be used to identify or filter out cells with noisy or non-homogenous input data.

Interpretation of L3 imagery can be further complicated by instrument differences, such as swath width or orbital geometry. Both MODIS and VIIRS complete 14-15 orbits per day with sun-synchronous orbit geometry that views the same regions of Earth at higher latitudes in successive, overlapping swaths (Fig. 8). Given an orbital duration of roughly 90 minutes, multiple L2 swaths contributing to the same daily L3b grid cell not only combines two views of the ocean separated by a time scale, which may have biological or geophysical relevance, but also distorts the time difference of data collection relative to a potential in situ target. These results vary slightly between MODIS and VIIRS due to differences between the instruments' scanning patterns and orbital geometry, with VIIRS having a wider swath that creates more overlap between successive orbits (Fig. 8). Despite the potential for multiple swaths to contribute to a single L3 grid cell, this did not occur for any of the MODIS L3b match-ups and for fewer than ten of the VIIRS L3b match-ups used in this analysis.

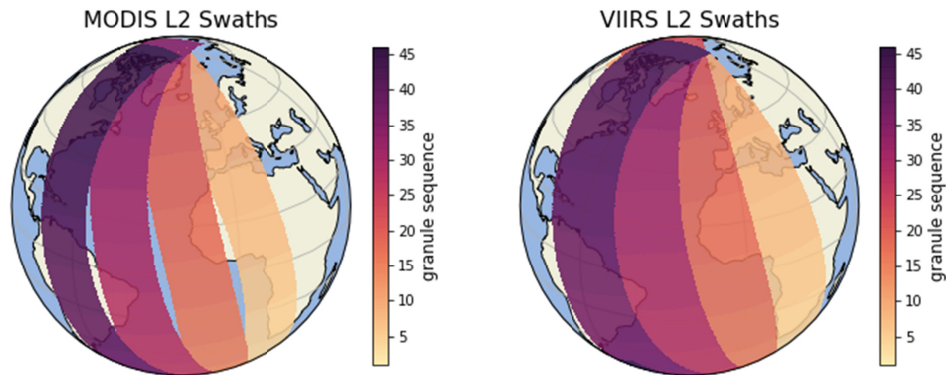


Fig. 8. Geographic coverage of successive L2 swaths for MODIS (left) and VIIRS (right).

The Earth-grid selected for the L3 data further affects the applicability of L2 validation results to the re-projected data. Two significant differences exist between L3b and L3m Earth-grids: 1) middle and high latitude cells are distorted in the L3m Plate Carrée projection, which was briefly discussed in Section 1, and 2) L3b and L3m Earth-grids have different grid cell definition and spacing. L3b grid cells are designed to be roughly equal area, whereas L3m grid cells are designed to have uniformly spaced meridians. This creates additional L3m grid cells at higher latitudes compared to L3b, which also leads to additional L3m match-ups being generated. At the equator, differences between L3b and L3m cell areas are minimal. However, L3m grid cells cover less area than L3b with increasing latitude. This can be illustrated via a simple comparison: Greenland looks to be nearly the same size as the continent of Australia in

the lower panel of Fig. 1, when in fact, Greenland ($2.2\text{e}6 \text{ km}^2$) is less than one third of the area of Australia ($7.7\text{e}6 \text{ km}^2$).

L3b and L3m Earth-grids differ in cell definition and spacing. Figure 9 was generated by calculating the center location of each L3m grid pixel, identifying the corresponding sinusoidal L3b grid cell, and computing the great-circle distance for each corner of the paired, coincident L3b and L3m grid cells. Stratifying by latitude, we see that L3b and L3m grid cell corners can differ by as much as half the grid spacing ($\sim 2.1 \text{ km}$) in the low latitudes, up to an entire grid cell (4.62 km) at the poles. Therefore, L3b data projected to a L3m Plate Carrée Earth-grid introduces a spatial ambiguity to the data of up to a few kilometers, increasing as a function of latitude, from the true satellite retrieval's L2 location. This will affect which L2 satellite values are being averaged and placed into each Level-3 grid cell, with potentially different values falling into the L3b grid cells versus the L3m. Consequently, this will create differences between L3b and L3m satellite values being matched up to the in situ targets.

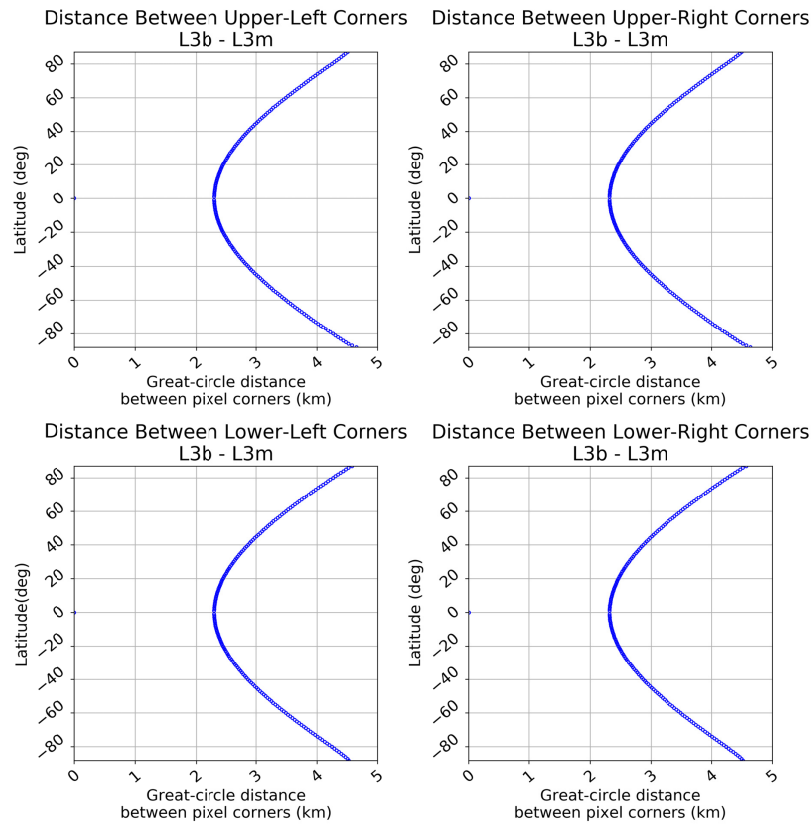


Fig. 9. Maximum great-circle distances between corresponding L3b and L3m grid cell corners, inclusive of the upper-left corners (upper left panel), upper-right corners (upper right panel), lower-left corners (lower left panel), and lower-right corners (lower right panel), stratified into 1° latitude bins on the y-axis.

Differences between the L3b and the L3m grids, arising from the choice of map projection, lead to different sets of in situ targets generating L3b and L3m match-ups. The combined effect of L3m grid cell area decreasing with latitude and number of cells increasing with latitude contributes to more match-up possibilities for L3m data in the middle and high latitudes relative to L3b, as evidenced by there being 26% more match-ups for the L3m data compared to the L3b data (Figs. 3 and 4, Table 2). This combined effect is also the reason for

the additional 118 MODIS L3m match-ups without a corresponding L3b match-up (Fig. 6). For our dataset, the mid-latitudes realized an increased number of L3m match-ups for both MODIS and VIIRS (Fig. 10). Additionally, differences between L3b and L3m grid cell corner definitions lead to a spatial shift of the L2 satellite data products relative to the original *Chl* retrieval's location. This spatial shift more strongly influences MODIS, which has larger gaps between successive swaths in the low latitudes than VIIRS (Fig. 8), creating additional L3m match-ups in the low latitudes relative to the number of L3b match-ups (Fig. 10). Conversely, the overlap between successive VIIRS swaths (Fig. 8) masks this effect, creating just as many L3b match-ups in the low latitudes as in the L3m data, but with fewer overpasses populating those L3b grid cells (Fig. 10).

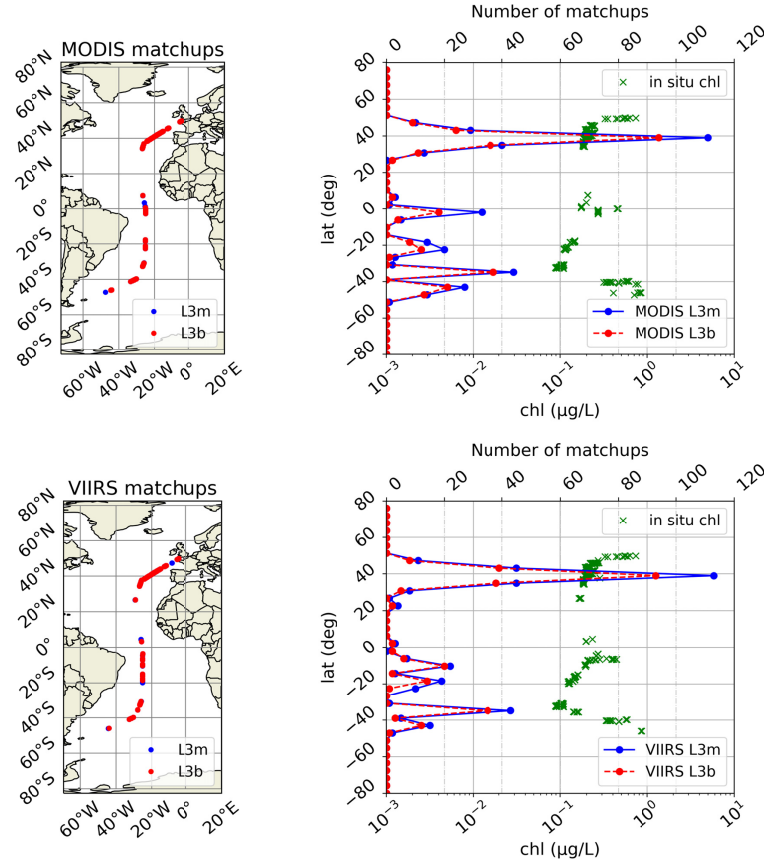


Fig. 10. As in Fig. 7, but for L3b and L3m data.

The spatial ambiguity introduced by using L3m data on the Plate Carrée Earth-grid is evidenced in its satellite-to-in situ performance metrics (Figs. 3 and 4, Table 2). L3m match-ups from both MODIS and VIIRS have their respective highest MAEs, indicating a greater level of error associated with L3m match-ups compared to L3b match-ups. Since the only differences between the two Level-3 products is their Earth-grids, the higher error associated with L3m match-ups can be attributed to differences in grid spacing and definition. Furthermore, the comparison of in situ data to the MODIS L3b satellite data generate a mean bias closer to unity compared to the L3m match-ups. Ultimately, this all suggests that the L3b satellite-to-in situ match-ups show stronger agreement than those for L3m due to the former being built on a grid that more accurately represents the Earth and introduces less spatial ambiguity compared to the latter.

5. Conclusions

Many ocean color data users prefer the simplicity of L3 satellite data. However, ocean color satellite retrieval performance is evaluated on L2 data. When compared to a common in situ data set, L2, L3b, and L3m satellite data retrievals of *Chl*, show similar match-up results for both MODIS and VIIRS instruments. These similarities indicate that L2 validation results are largely extensible to L3 data and that in situ data remain the standard against which remotely-sensed satellite data, both L2 and L3 data products, must be validated. Differences between L2 and L3 performance metrics relative to a common in situ data set arise due to the introduction of uncertainties during the process of converting satellite data from L2 to L3. These uncertainties are dependent on the location under consideration (e.g., low-, mid-, or high-latitudes) and the Earth-grid selected.

While L3 validation results largely resemble L2 comparison metrics, validation-focused studies should be conducted on L2 retrievals. L2 data remain on the native satellite swath and observing geometries and allow for a more rigorous set of exclusion criteria to be applied, compared to L3 match-ups. L3 performance metrics do not exactly replicate L2 metrics primarily due to satellite-to-in situ time differences. L3 match-ups allow a wider temporal window to exist between in situ and satellite observations, which calls into question the quality of the L3 match-ups being derived. The use of less-coincident data does not implicitly reflect on the quality of the satellite retrieval or algorithm performance. The results presented in this study are for *daily* L3 data only, performance metric differences between L2 and L3 match-ups are expected to be exacerbated when larger L3 temporal composites are considered (i.e., weekly, monthly, etc).

Projecting L3 data onto an Earth-grid may introduce further uncertainty, evidenced by discrepancies between L2 and L3 performance metrics. By projecting L3b data onto an L3m Earth-grid, spatial ambiguity is introduced since the L3m grid is not defined to have equal-area grid cells, leading to geographic distortion of the satellite data toward the poles. The signal and shape of the *Chl* distribution in the L2 data are lost with spatiotemporal binning and averaging of the L3b and L3m data sets, particularly for the contributions of smaller population groups. Each manipulation of satellite data beyond the native satellite's spatiotemporal reference (L2) makes results of validation studies performed on L2 data less applicable to higher data levels (L3b and L3m).

Funding

This work was funded and supported by the National Aeronautics and Space Administration (NASA) Earth Sciences Division Ocean Biology and Biogeochemistry Program.

Acknowledgments

We thank Barney Balch at the Bigelow Laboratory for Ocean Sciences for the collection and processing of the in situ AMT data set. We thank Sean Bailey, Norman Kuring, and Fred Patt for providing insight and advice. We thank the NASA SeaBASS team for in situ data access and support. We thank the Ocean Biology Processing Group, SeaDAS and OCSSW developers for providing ocean color data and software.

References

1. C. R. McClain, "A Decade of Satellite Ocean Color Observations," *Annu. Rev. Mar. Sci.* **1**(1), 19–42 (2009).
2. NRC, *Assessing requirements for sustained ocean color research and operations*. Washington, D.C., (National Academies, 2011).
3. S. Sathyendranath, R. J. W. Brewin, T. Jackson, F. Mélin, and T. Platt, "Ocean-colour products for climate-change studies: What are their ideal characteristics?," *Remote Sens. Environ.* **203**, 125–138 (2017).
4. C. D. Mobley, J. Werdell, B. Franz, Z. Ahmad, and S. Bailey, "Atmospheric Correction for Satellite Ocean Color Radiometry," NASA Tech. Memo. 73 (2016).

5. 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).
6. 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).
7. NASA Goddard Space Flight Center Ocean Biology Processing Group, "Moderate Resolution Imaging Spectroradiometer onboard the Aqua Earth-observing satellite (MODIS-Aqua) Ocean Color Data," NASA OB.DAAC (2018) [retrieved 07 Sep 2018], <http://doi.org/10.5067/AQUA/MODIS/L2/OC/2018>.
8. NASA Goddard Space Flight Center Ocean Biology Processing Group, "Moderate Resolution Imaging Spectroradiometer onboard the Aqua Earth-observing satellite (MODIS-Aqua) Ocean Color Data," NASA OB.DAAC (2018) [retrieved 07 Sep 2018] <http://doi.org/10.5067/AQUA/MODIS/L3B/CHL/2018>.
9. NASA Goddard Space Flight Center Ocean Biology Processing Group, "Moderate Resolution Imaging Spectroradiometer onboard the Aqua Earth-observing satellite (MODIS-Aqua) Ocean Color Data," NASA OB.DAAC (2018) [retrieved 07 Sep 2018] <http://doi.org/10.5067/AQUA/MODIS/L3M/CHL/2018>.
10. NASA Goddard Space Flight Center Ocean Biology Processing Group, "Visible Infrared Radiometer Suite onboard the Suomi National Polar-orbiting Partnership weather satellite (VIIRS-SNPP) Ocean Color Data," NASA OB.DAAC (2018) [retrieved 07 Sep 2018] <http://doi.org/10.5067/NPP/VIIRS/L2/OC/2018>.
11. NASA Goddard Space Flight Center Ocean Biology Processing Group, "Visible Infrared Radiometer Suite onboard the Suomi National Polar-orbiting Partnership weather satellite (VIIRS-SNPP) Ocean Color Data," NASA OB.DAAC (2018) [retrieved on 07 Sep 2018] <http://doi.org/10.5067/NPP/VIIRS/L3B/CHL/2018>.
12. NASA Goddard Space Flight Center Ocean Biology Processing Group, "Visible Infrared Radiometer Suite onboard the Suomi National Polar-orbiting Partnership weather satellite (VIIRS-SNPP) Ocean Color Data," NASA OB.DAAC (2018) [retrieved on 07 Sep 2018] <http://doi.org/10.5067/NPP/VIIRS/L3M/CHL/2018>.
13. 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).
14. 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).
15. G. Zibordi, J. F. Berthon, F. Mélin, D. D'Alimonte, and S. Kaitala, "Validation of satellite ocean color primary products at optically complex coastal sites: Northern Adriatic Sea, Northern Baltic Proper and Gulf of Finland," *Remote Sens. Environ.* **113**, 2574–2591 (2009).
16. B. N. Seegers, R. P. Stumpf, B. A. Schaeffer, K. A. Loftin, and P. J. Werdell, "Performance metrics for the assessment of satellite data products: an ocean color case study," *Opt. Express* **26**, 7404 (2018).
17. P. J. Werdell and S. W. Bailey, "The SeaWiFS Bio-Optical Archive and Storage (SeaBASS): Current Architecture and Implementation," NASA Tech. Memo. 1–50 (2002).
18. J. W. Campbell, J. M. Blaisdell, and M. Darzi, "Volume 32 , Level -3 SeaWiFS Data Products : Spatial and Temporal Binning Algorithms," *SeaWiFS Tech. Rep. Ser.* **32**, (1995).
19. W. B. Rossow and L. Garder, "Selection of a Map Grid for Data Analysis and Archival," *J. Clim. Appl. Meteorol.* **23**, 1253–1257 (1984).
20. K. A. Mulcahy and K. C. Clarke, "Symbolization of Map Projection Distortion: A Review," *Cartogr. Geogr. Inf. Sci.* **28**, 167–182 (2008).
21. S. C. Doney, D. M. Glover, S. J. McCue, and M. Fuentes, "Mesoscale variability of Sea-viewing Wide Field-of-view Sensor (SeaWiFS) satellite ocean color: Global patterns and spatial scales," *J. Geophys. Res. Ocean.* **108**, 1–15 (1998).
22. W. W. Gregg and N. W. Casey, "Modeling coccolithophores in the global oceans," *Deep. Res. Part II Top. Stud. Oceanogr.* **54**, 447–477 (2007).
23. D. M. Goldberg and J. R. Gott, "Flexion and Skewness in Map Projections of the Earth," *Cartogr. Int. J. Geogr. Inf. Geovisualization* **42**, 297–318 (2007).
24. P. J. Werdell, M. J. Behrenfeld, P. S. Bontempi, E. Boss, B. Cairns, G. T. Davis, B. A. Franz, U. B. Gliese, E. T. Gorman, O. Hasekamp, K. D. Knobelspiesse, A. Mannino, J. V. Martins, C. R. McClain, G. Meister, and L. A. Remer, "The Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: Status, science, advances," *Bull. Am. Meteorol. Soc.* <http://dx.doi.org/10.1175/BAMS-D-18-0056.1> (2019).
25. W. W. Gregg and N. W. Casey, "Global and regional evaluation of the SeaWiFS chlorophyll data set," *Remote Sens. Environ.* **93**, 463–479 (2004).
26. Balch, William, "AMT," NASA SeaWiFS Bio-optical Archive and Storage System (SeaBASS), (2014) [retrieved 06 Sep 2018] <http://dx.doi.org/10.5067/SeaBASS/AMT/DATA001>.
27. M. Lehaire and C. Compère, "Biofouling and Underwater Measurements," *Real-Time Coast. Obs. Syst. Mar. Ecosyst. Dyn. Harmful Algal Bloom. Theory, Instrum. Model.* 463–493 (2008).
28. P. G. Falkowski and Z. Kolber, "Variations in chlorophyll fluorescence yields in phytoplankton in the world oceans," *Austrian J. of Plant Phys.* **22**, 341–355, (1995).
29. D. A. Kiefer, "Fluorescence properties of natural phytoplankton populations," *Mar. Biol.* **22**, 263–269 (1973).
30. 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).
31. J. W. Campbell, "The lognormal distribution as a model for bio-optical variability in the sea," *J. Geophys. Res.* **100**(C7), 13237–13254 (1995).