Validation of VIIRS and MODIS reflectance data in coastal and oceanic waters: an assessment of methods

Brian B Barnes[a]*, Jennifer P Cannizzaro[a], David C English[a], Chuanmin Hu[a]

[a] College of Marine Science, University of South Florida, 140 7th Ave S, St Petersburg, FL 33701
* Corresponding author, bbarnes4@mail.usf.edu

Keywords: MODIS, VIIRS, validation, quality control

Abstract

Satellite ocean color datasets have vast potentials for assessing and monitoring of marine environments. However, with the MODIS sensor aging and the VIIRS sensor reaching maturity, it is important to continuously evaluate the quality of reflectance data from both instruments. Here, we critically assess the statistical performance of both MODIS and VIIRS, including analysis of two separate (and commonly used) VIIRS processing routines. In addition, we note variability in the literature as to the methods used to identify and remove low-quality data during similar validation exercises. Although most studies use some implementation of satellite quality flags (L2 flags) and many exclude data based on spatial heterogeneity or large temporal gap from satellite overpasses, critical assessment of these methods indicates variable performance. Indeed, we found little improvement in validation statistics after implementation of these data culling techniques, with substantial variability in effectiveness between wavebands and sensors. Overall, these findings highlight the need to critically assess the impact (on both data quantity and quality) of exclusion criteria, toward more effective techniques to ensure quality and consistency of satellite ocean color datasets.

1. Introduction

Over the past few decades, satellite ocean color sensors have proven their vast utility in assessment and monitoring of oceanic and coastal marine systems – providing high quality geophysical data products at scales unattainable using traditional sampling. The spatiotemporally synoptic data streams from these sensors can elucidate otherwise hidden ocean features and patterns while reducing reliance on the more costly ship-borne measurements. To ensure the quality and consistency of the data from mainstream ocean color sensors [such as NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) on the satellite Aqua (MODISA) and the Visible Infrared Imaging Radiometer Suite (VIIRS) on the joint NASA/NOAA Suomi National Polar-orbiting Partnership satellite (Suomi-NPP)], it is important to regularly validate these data products against those measured at the water surface. This is especially true for newer (e.g., VIIRS, 2012-present) and aging instruments. In particular, MODISA (2002-present) is currently over 16 years old (design life of 6 years), and has recently shown some associated degradation (Meister et al., 2012; Meister and Franz, 2014), making it important to continually ensure accuracy of derived products and assess cross-sensor agreement (Barnes and Hu, 2015; Hu and Le, 2014).

In the context of MODIS and VIIRS data, multispectral normalized water leaving radiance (*nLw*; mW cm^{-2} um^{-1} sr^{-1}) and remote sensing reflectance (*Rrs*; sr^{-1}) products are the primary geophysical parameters from which most other products [e.g., chlorophyll a concentration (C_a; mg m^{-3})] are derived. These two products are equivalent as one can be derived from the other through *Rrs* = *nLw* / *F0* where *F0* is the mean extraterrestrial solar irradiance (a constant for a given wavelength). For brevity, wavelength dependence for *Rrs* and *nLw* is omitted here. *Rrs* is notoriously difficult to quantify, even *in situ*. In practice, *in situ Rrs* derivation from an above-water radiometer requires collection of multiple scans of upward radiance, diffuse downwelling irradiance, and sky radiance, followed by correction for skylight and sunglint (e.g., Lee et al., 2010) by an experienced analyst. This process can differ by research group, with sometimes variable outcomes (Garaba and Zielinski, 2013; Hooker et al., 2002; Toole et al., 2000). Similarly, *in situ Rrs* derivation from a submersible radiometer requires data reduction from depth to surface and from below surface to above surface, resulting in uncertainties in the final product (Antoine et al., 2008; Hooker et al., 2002).

Aside from the uncertainties associated with *in situ Rrs* data, comparing satellite-derived data to *in situ* measurements presents additional complications with respect to scale (Blackwell et al., 2008; Salama and Su, 2011). At nadir, MODIS and VIIRS pixels have approximate spatial resolutions of 1 km and 750 m, respectively. Given the spatial heterogeneity of ocean color (especially for nearshore environments), integrated *Rrs* measures over such large areas are not necessarily well represented by an *in situ* point measurement. Additionally, while simultaneous *in situ* / satellite measurements may be possible (e.g., from a buoy platform), temporal gaps between satellite and shipborne *in situ* validation datasets are much more common. Temporal instability of *Rrs* thus can reduce validation statistics, especially in nearshore environments (e.g., those modulated by tides).

Atmospheric correction provides yet another layer of uncertainty for validation of satellite-derived *Rrs*. While the default procedures to perform atmospheric correction in MODIS and VIIRS data streams are truly state-of-the-art, absorbing atmospheric aerosols can cause large uncertainties in *Rrs* retrievals, especially in coastal environments (Gordon et al., 1997). Additionally, the two primary distributors of VIIRS data (NOAA and NASA) each use a different implementation of atmospheric correction, sensor calibration, and treatment for straylight adjacent to bright targets. Very briefly, one of the largest discrepancies between the NASA (via the software package SeaDAS, within which Level-1 to Level-2 processing is performed using L2GEN) and NOAA (via MSL12) processing routines involves accounting for deviations to the black-pixel assumption (Gordon and Clark, 1981; Siegel et al., 2000), which is a pervasive problem for turbid coastal environments. In L2GEN, atmospheric correction over turbid coastal waters (non-black pixels) is through an iterative approach, whereby modeled inherent optical properties (IOPs) are used to estimate the non-zero *Rrs* in the near-infrared (NIR) wavebands (Bailey et al., 2010; Gordon and Wang, 1994; Mobley et al., 2016; Stumpf et al., 2003). In MSL12, atmospheric correction over the same turbid coastal waters is through a combination of the Bailey et al., (2010), Ruddick et al., (2000), and Wang et al., (2012) approaches, with the former algorithm being used to estimate the aerosol single scattering reflectance ratios and the latter two algorithms being used to carry out atmospheric correction (Jiang and Wang, 2014). Traditionally, quality of satellite pixels is established via Level-2 Processing Flags (L2 Flags; Patt et al. 2003), with the goal of identifying pixels contaminated by sources of *Rrs* uncertainty (or invalidation), including clouds, sun glint, absorbing aerosols, and sensor geometry issues, among many others. Recently, Wei et al. (2016) provided an additional quality assessment method for *in situ*- and satellite-derived *Rrs*, which has been adopted within the MSL12 processing.

Due to this multitude of uncertainties, mismatches, and sources of error, validation of satellite *Rrs* and *nLw* datasets requires accurate and robust *in situ* datasets covering a wide dynamic range of water properties, which take a significant amount of time and resources to collect. Even with a robust validation dataset, however, only a fraction of *in situ Rrs* will have matchups (collocated and coincident measurements) with satellite *Rrs*. This is especially true after the satellite data have been screened for the presence of clouds, sun glint, straylight, and other factors that reduce quality (or prevent calculation) of satellite-derived *Rrs*.

Nevertheless, several studies have provided validation of MODISA and VIIRS *Rrs* data (Table 1). Overall, the majority of these studies have shown *Rrs* products provide consistent estimates (percent difference for green band *Rrs* matchups < 20%), which agrees with similar analyses using cross-validation between sensors (Barnes and Hu, 2016; Hu et al., 2015; Hu and Le, 2014; Li et al., 2015; Uprety et al., 2013). Matchup statistics are generally reduced (i.e., larger uncertainties) in the blue and red bands due to atmospheric correction uncertainties and strong water absorption, respectively (Antoine et al., 2008;

Franz et al., 2007). Note that target uncertainties for satellite retrievals of blue band nLw for very clear waters are 5% (Hooker et al., 1992; Hooker and Esaias, 1993). However, the uncertainties of *in situ* data can be at least that large (Bailey and Werdell, 2006; Hooker and Maritorena, 2000), making it difficult to disentangle uncertainties from these two sources unless uncertainties are evaluated using stable ocean targets instead of *in situ* measurements, for example over ocean gyres (Hu et al., 2013). Additionally, many validation efforts to date have focused on data from fixed platforms (see Table 1), meaning certain environments may be undersampled, including blooms, river plumes, and shallow waters (< 10m) with variable bottom types and optical depths.

Table 1: Summary of selected *Rrs* validation methods and results for MODISA and VIIRS sensors

Citation	Platform	Environment	Sensor	Processing software, calibration version*	CV Threshhold (box size)	Temporal overlap (hr)	Accuracy Statistic (547 or 551 nm)
Mélin et al., 2007	Fixed	Coastal	MODIS	L2GEN, ~2005.1	0.2 (3x3)	3.5	MAPD = 14%
Antoine et al., 2008	Fixed	Oceanic	MODIS	L2GEN, 2005.0	- (5x5)	3	MAPD = 17%
Zibordi et al., 2009	Fixed (AERONET)	Coastal	MODIS	L2GEN, ~2005.1	0.2 (3x3)	2	MAPD = 10%
Maritorena et al., 2010	Ship & Fixed (AERONET)	Coastal & Oceanic	MODIS	L2GEN, 2005.1	- (-)	-	MR = 1.006
Hlaing et al., 2013	Fixed (AERONET)	Coastal	MODIS	L2GEN, 2012.0	0.2 (3x3)	2	MAPD ~ 12%
Brando et al., 2016	Ship & Fixed (AERONET)	Coastal & Oceanic	MODIS	L2GEN, 2014.0.1	- (3x3)	2	MAPD ~ 12%
Wang et al., 2013	Fixed (MOBY)	Oceanic	VIIRS	MSL12	- (11x11)	-	MR = 0.98
Hlaing et al., 2013	Fixed (AERONET)	Coastal	VIIRS	L2GEN, 2012.2	0.2 (3x3)	2	MAPD ~ 12%
Hlaing et al., 2013	Fixed (AERONET)	Coastal	VIIRS	MSL12, IDPS v6.6	0.2 (3x3)	2	MAPD = 14%
Ahmed et al., 2013	Fixed (AERONET)	Coastal	VIIRS	L2GEN, 2013.0	0.2 (3x3)	2	MAPD = 10 - 15%
Brando et al., 2016	Fixed (AERONET)	Coastal	VIIRS	MSL12	0.2 (3x3)	2	MAPD = 14%
Wang et al., 2014	Fixed (MOBY)	Oceanic	VIIRS	MSL12	- (5x5)	8	MR = 0.992
Vandermeulen et al., 2015	Ship & Fixed (AERONET)	Coastal & Oceanic	VIIRS	NRL-APS v5.1	- (-)	3	RMSD = 0.160 $mW/cm^2/m/sr$
Wang et al., 2015	Fixed (MOBY)	Oceanic	VIIRS	MSL12	- (5x5)	-	MR = 1.0157
Wang et al., 2016	Fixed (MOBY)	Oceanic	VIIRS	MSL12	- (5x5)	-	MR = 1.0148
Brando et al., 2016	Ship & Fixed (Aeronet)	Coastal & Oceanic	VIIRS	L2GEN, 2014.0.1	- (3x3)	2	MAPD ~ 12%

MR = Mean Ratio, MAPD = Mean Absolute Percent Difference, RMSD = Root Mean Squared Difference, - = not performed or not reported, * Where not specified, approximate processing or calibration version reported

Furthermore, there is a general lack of consensus among these validation studies on the method used to assess satellite *Rrs* quality and remove low (or questionable) quality matchups. For example, the coefficient of variation (CV = standard deviation / mean) of an *n* x *n* pixel box (with the *in situ* sample location at the center) is often used to assess spatial homogeneity of the matchup location. The concept is that a highly variable environment (at the scale of satellite pixels) would more likely foster mismatches between the satellite and *in situ* targets. However, CV thresholds used for such data culling vary, with commonly used values including 0.4 (Harding et al., 2005; Le et al., 2013a, 2013b; Le and Hu, 2013), 0.2 (Ahmed et al., 2013; Hlaing et al., 2013; Zibordi et al., 2009), 0.15 (Brown et al., 2008; Weeks et al., 2012; Werdell et al., 2009) and 0.1 (Barnes et al., 2013). Mélin et al. (2007) reported minimal degradation of *Rrs* matchup statistics for a coastal environment after relaxing the CV threshold.

The average (or median) of the n x n pixel box can also be used to filter sensor and algorithm noise (Hu et al., 2001), particularly for those studies focused on oceanic waters. This can be performed in lieu of a CV threshold, or in addition to it. However, there is no consensus on the size of the box (for either the CV or box-mean approaches), with sizes including 3x3 (Ahmed et al., 2013; Brando et al., 2016; Hlaing et al., 2014), 5x5 (Antoine et al., 2008; Wang et al., 2016, 2015, 2014), and even 11x11 (Wang et al., 2013). Indeed, even though Bailey and Werdell (2006) provide a comprehensive calculation to justify a 7x7 box for SeaWiFS data, statistics are reported using a 5x5 box with no degradation.

Note that regardless of the method to address spatial heterogeneity, there is variability between studies on the method used to extract satellite data, with many extracting from Level-2 (unmapped) data (Ahmed et al., 2013; Antoine et al., 2008; Brando et al., 2016; Wang et al., 2015), while others use Level-3 (mapped) products (Barnes and Hu, 2016; Wang et al., 2013). For the latter, a cylindrical equidistant projection is typically used, with the spatial resolution of the grid being the sensor-specific spatial resolution at nadir (~1 km for MODIS, 750 m for VIIRS). As the footprint of Level-2 pixels expands at the swath edge (MODIS Level-2 pixels at the scan edge are approximately 5 x 2 km, while VIIRS scan edge pixels are approximately 1.6 x 1.6 km), mapping can result in a single Level-2 pixel covering several "pixels" in the Level-3 grid. This presents obvious ramifications for either of the n x n pixel methods used to address spatial heterogeneity. Even for unmapped (Level-2) products, pixel area expansion at the scan edge causes larger spatial areas to be assessed, while the bowtie effect can cause spatial overlaps in the n x n pixel region, especially at the boundaries between scan lines (Figs. 1-2). These impacts manifest differently for MODIS (Fig. 1) and VIIRS (Fig. 2) data due to the VIIRS pixel aggregation scheme, which results in "deleted" pixels (Cao et al., 2013) on the scan line edges at higher sensor zenith angles. Nevertheless, for both sensors, these impacts mean that n x n spatial heterogeneity procedures are not always considering exactly what is expected.

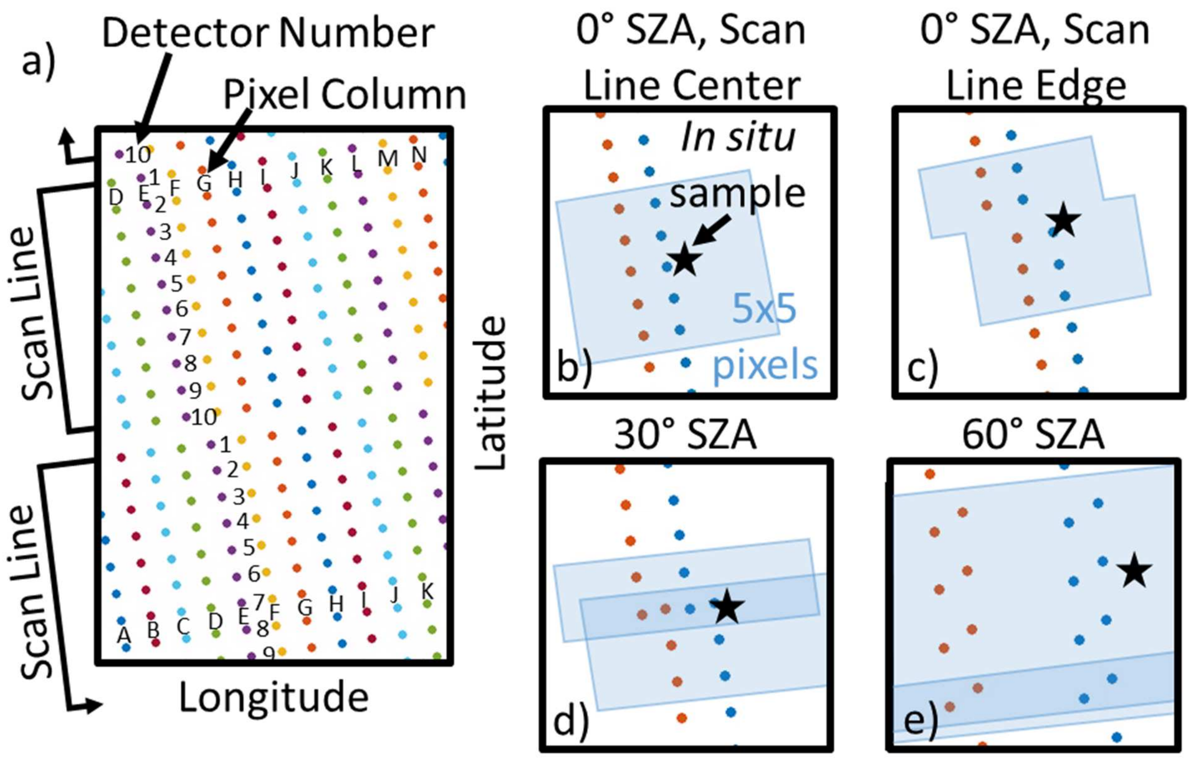

Figure 1: Spatial extents of 5 x 5 pixel boxes in a MODIS Level-2 granule as used for spatial heterogeneity testing (or box averaging) at various sensor zenith angles (SZA). (a) Geographic pixel centers of L2 data products at 0° SZA, with scan lines (10 detectors for MODIS) designated and pixel columns separated by color. (b-e) Approximate spatial extent of 5 x 5 pixel pox (blue) for arbitrarily selected *in situ* sample locations (stars) – only two pixel columns are shown for clarity. (b) For matchups near the scan line center at 0° SZA (or at any other SZA), the 5 x 5 pixel box is a rectangle oriented parallel to the pixel column. At the scan line edge (c-e), however, incongruities in the pixel centers can cause non-rectangular shapes (c), while the bowtie effect can cause overlap in the 5 x 5 pixel area (d-e). Enlargement of pixel area at higher SZA means a larger area is considered in the 5 x 5 pixel boxes. Panels b-e have the same scale. Pixel centers from approximately 30° N latitude.

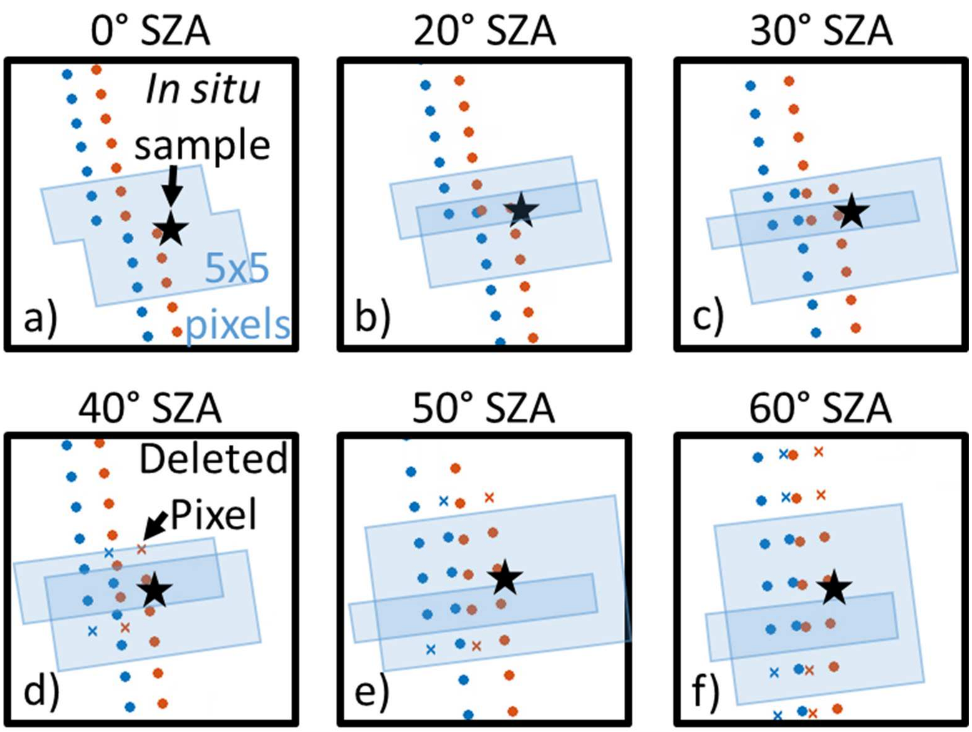

Figure 2: Similar to Fig. 1, showing approximate VIIRS 5 x 5 pixel box areas (blue) surrounding *in situ* samples (stars) placed near scan line boundaries at various sensor zenith angles (SZA). Only two pixel columns (red dots and blue dots) shown in each panel for clarity. Deleted pixels (d-f) resulting from VIIRS pixel aggregation scheme contain no geophysical data and are not considered in the 5 x 5 boxes – their geographic locations are represented by ‘x’. All panels have the same scale. Pixel centers from approximately 30° N latitude.

For temporal overlap, most studies require satellite / *in situ* matchups used in validation analyses to be either same day (Wang et al., 2014), within 3 hours (Antoine et al., 2008; Vandermeulen et al., 2015), or within 2 hours (Ahmed et al., 2013; Brando et al., 2016; Zibordi et al., 2009). Nevertheless, Mélin et al. (2007) and Barnes and Hu (2015) note no difference in matchup statistics with variable temporal overlap thresholds. Finally, for the few studies that directly list them, the specific Level-2 processing flags used to discard low-quality satellite data can vary between studies (Table 2). Despite this variability in methods, few studies statistically justify the specific thresholds (or flagging schemes) used, or provide any assessment of the impact of these particular values on the validation statistics.

Table 2: Level-2 Processing Flags (from http://oceancolor.gsfc.nasa.gov/atbd/ocl2flags/ and Wang et al., 2017).

Bit position	Default Mask	L3 Mask*	"Current" Mask‡	Bailey and Werdell (2006) †	Hlaing et al. (2013) §	Name (L2GEN)	Name (MSL12)	L2GEN Description [MSL12 description]
0		X	X	X	X	ATMFAIL	ATMFAIL	Atmospheric correction failure
1	X	X	X	X	X	LAND	LAND	Pixel is over land
2						PRODWARN	PRODWARN	Warning from ≥ 1 product algorithms
3		X	X	X	X	HIGLINT	HIGLINT	Sunglint: reflectance exceeds threshold
4	X	X	X	X	X	HILT	HILT	Radiance very high or saturated
5		X	X	X	X	HISATZEN	HISATZEN	Sensor zenith angle exceeds threshold
6						COASTZ	COASTZ	Pixel is in shallow water
7						Spare	LANDADJ	[Probable land-adjacent contamination]
8		X	X	X	X	STRAYLIGHT	STRAYLIGHT	Probable stray light contamination
9	X	X	X	X	X	CLDICE	CLOUD	Probable cloud or ice contamination
10		X				COCCOLITH	COCCOLITH	Coccolithophores detected
11						TURBIDW	TURBIDW	Turbid water
12		X	X	X	X	HISOLZEN	HISOLZEN	Solar zenith angle exceeds threshold
13						Spare	HITAU	[High Aerosol Optical Thickness]
14		X	X	X	? §	LOWLW	LOWLW	Very low water-leaving radiance
15		X		? †		CHLFAIL	CHLFAIL	Chlorophyll algorithm failure
16		X	X		?	NAVWARN	NAVWARN	Navigation quality is suspect
17		X				ABSAER	ABSAER	Absorbing Aerosols determined
18						Spare	CLDSHDSTL	[Cloud straylight or shadow]
19		X	X			MAXAERITER	MAXAERITER	NIR iteration limit reached
20				?	X	MODGLINT	MODGLINT	Moderate sun glint
21				? †		CHLWARN	CHLWARN	Chlorophyll out-of-bounds
22		X	X			ATMWARN	ATMWARN	Atmospheric correction is suspect
23						Spare	ALGICE	[Sea ice identified by nLw]
24						SEAICE	SEAICE	Pixel is over sea ice
25		X	X		X	NAVFAIL	NAVFAIL	Navigation failure
26						FILTER	FILTER	Insufficient data for smoothing filter

27						Spare	ALTCLD	[Cloud detected]
28						BOWTIEDEL	FOG	VIIRS deleted overlapping pixels [Fog]
29						HIPOL	FROMSWIR	High polarization [SWIR atm. corr. used]
30						PRODFAIL	PRODFAIL	Failure in any product
31						SPARE	OCEAN	[Pixel is over ocean]

* The L3 mask is used for generation of global composite data products.

‡ The "current" mask is that used throughout this study

† Includes additional flag(s) specific to C_a. Also used by Antoine et al. (2008), Mélin et al. (2007), Zibordi et al. (2009)

§ Includes additional flag for negative Rayleigh-corrected reflectance. Also used by Ahmed et al. (2013).

As such, this work follows two main objectives. First is to compare different validation methods of satellite *Rrs* data through the use of a large dataset covering a variety of water types ranging from estuarine, coastal, and oceanic in North America. The other is to evaluate these *Rrs* data products from VIIRS (both MSL12 and L2GEN processing) and MODISA (L2GEN processing only). Specifically, we present MODISA and VIIRS *Rrs* validation against the *in situ* dataset, assess typical data quality control methodologies, and provide environment-specific recommendations for future validation efforts, with the ultimate (and ongoing) goal of establishing high-quality, self- and cross-consistent environmental data records.

2. Methods

2.1. In situ data

Above-water reflectance data were collected between 2012-2017 using a handheld radiometer on 53 cruises in the Gulf of Mexico and waters off the southeast US coast (colloquially termed 'South Atlantic Bight', Fig. 3). Spectra were collected with the reflectance plaque radiance method described in the NASA ocean optics protocols (Mueller et al., 2003), using either a custom-built spectral radiometer (Spectrix; <3nm spectral resolution; ~350–800 nm) or a FieldSpec HandHeld 2-Pro Spectroradiometer (ASD). Specifically, at each station, multiple observations of upwelling radiance, diffuse downwelling irradiance (gray 10% diffuse reflector; Spectralon), and sky radiance were collected. During these data collections, senor zenith angle was constrained to 30 - 40° (from nadir for water measurements, and from zenith for sky measurements), while sensor azimuth was generally ~90°, up to 130° to avoid sun glint. For each radiance / irradiance parameter, obvious outlier scans were removed and the average of the remaining scans were used to calculate *Rrs*(λ) spectra (*N* = 432). The reflectance of the grey reference plaque was adjusted using solar zenith angles to reduce the biases introduced by the non-lambertian response of the reference plaque, while skylight and sunglint corrections were performed using optimization (Lee et al., 2010). Because the upwelling radiance below water is nearly isotropic for small angles, no BRDF correction was applied in the R_{rs} estimates. Error budgets for the R_{rs} dataset (*sensu* Zibordi, 2016) indicate uncertainties at 550 nm generally between 5% and 10%. Additionally, a round-robin comparison conducted with both above- and below-water R_{rs} measurements from ≥ 10 other groups during collection of many of the *in situ* data indicated between-sensor agreement within ~7% for wavelengths from 410 – 550 nm (Kovach and Ondrusek, 2018).

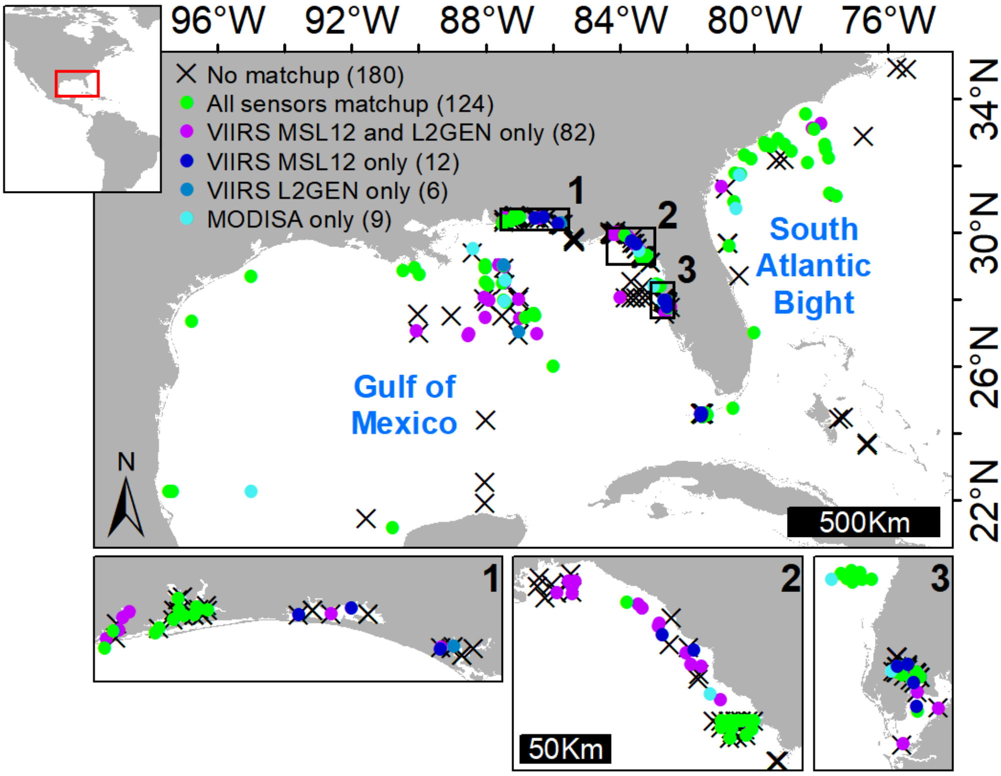

Figure 3: Map of sample locations, concentrated in the Gulf of Mexico and South Atlantic Bight, grouped according to matchup(s) with satellite data. These data were collected from 53 cruises of lengths from 1 to 35 days in 2012-2017. Enlargements shown for three regions (1. Florida Panhandle estuaries, 2. Florida Big Bend region, and 3. Tampa Bay) with high sample density (1-3 all have same spatial scale).

The *Rrs* quality assessment technique of Wei et al. (2016) was applied (using data subsampled to 9 wavelengths), yielding an *Rrs* quality score (hereafter termed 'QA_Wei') and water type for each spectra. For visualization of these water types, normalized *Rrs* (*nRrs*) spectra (dimensionless) were also calculated using these 9-band spectra by dividing each spectra by its root sum of squares (Wei et al., 2016). *Rrs* spectra with QA_Wei < 0.5 were further scrutinized to determine if any collection or processing characteristics (e.g., high solar zenith, unfavorable sea state, low scan repeatability, low signal-to-noise, etc.) warranted exclusion from the validation dataset. Note that while most spectra with QA_Wei < 0.5 were justifiably disqualified from further analyses, several seemingly high quality spectra showed very low QA_Wei (even QA_Wei = 0), but were not removed from the validation dataset for reasons explained below. All spectra were convolved to VIIRS and MODIS spectral bandwidths using the instrument- and band-specific relative spectral response functions. Note that while the VIIRS band centers (410, 443, 486, 551, and 671) differ slightly from associated MODIS band centers (412, 443, 488, 547, and 667), for this study we refer to the VIIRS band center names for both sensors, where appropriate.

2.2. Satellite data

MODISA and VIIRS granules covering the date and location of each *in situ Rrs* spectrum were downloaded at Level-2 from NASA GSFC archives (https://oceancolor.nasa.gsfc.gov) on 29 January 2018. These files conform to calibration 2018.0, for which atmospheric correction was performed with the iterative NIR approach (Bailey et al., 2010; Gordon and Wang, 1994; Mobley et al., 2016). VIIRS "science quality" data for these dates and locations were also acquired from NOAA CoastWatch (https://coastwatch.noaa.gov) on 21 February 2018. These data correspond to the April 2017 SDR and calibration update, with atmospheric correction performed using the NIR-SWIR procedure (Gordon and Wang, 1994; Jiang and Wang, 2014; Wang et al., 2017; Wang and Shi, 2007). Within this manuscript, VIIRS data from these two sources are termed 'VIIRS L2GEN' and 'VIIRS MSL12', respectively.

For each *in situ* spectrum and sensor, all same-day and collocated Level-2 satellite pixel(s) were identified. To account for overlapping scan lines and pixel enlargement at the scan edge, the "nearest" pixel was identified by first finding the scan line center which passed nearest to the sample location, then finding the geographically closest pixel within that scan line. Products including spectral *Rrs* (*nLw* for VIIRS MSL12) and Level-2 processing flags were extracted for each sample location and the surrounding 3x3 pixel box. For consistency, MSL12 *nLw*(λ) data were converted to *Rrs* [*Rrs*(λ)=*nLw*(λ)/*F0*(λ)] using spectral response integrated *F0* values (Thuillier et al., 2003). In practice, there are slight variations between the *F0* values used in the MSL12 and L2GEN processing routines, thus the *F0* values embedded in the Level-2 granules (NetCDF4 attributes) were used. Additionally, QA_Wei were calculated for all matchup spectra. Note that although VIIRS MSL12 L2 granules include a 'qa_score' product, the QA_Wei algorithm (as used in this manuscript) has been slightly updated since the MSL12 implementation (Menghua Wang, Jianwei Wei, personal communication). Although no Level-2 Processing Flags were applied to remove low-quality data at the time of data extraction, default processing precludes atmospheric correction (thus *Rrs* or *nLw* derivation) for any pixels identified as ATMFAIL, LAND, HILT, and CLDICE (termed "CLOUD" in MSL12 datasets; see Table 2 for a description of relevant L2 flags).

2.3. Statistical validation

Unbiased percent difference (UPD) and mean relative difference (MRD; also termed Relative Percent Difference, RPD, and Mean Percent Difference, MPD) were the primary measures used to assess satellite accuracy and bias, respectively, as:

$$UPD = \frac{100}{N} \times \sum_{i=1}^{N} \frac{|Y_i - X_i|}{0.5 \times (Y_i + X_i)}, \quad (1)$$

and

$$MRD = \frac{100}{N} \times \sum_{i=1}^{N} \frac{(Y_i - X_i)}{X_i}, \quad (2)$$

where X_i and Y_i are the *in situ* and satellite data, respectively, for matchup *i* of *N* total. Whereas most similar studies report *Rrs* accuracy as Mean Absolute Percent Difference (MAPD; or Average APD, AAPD), UPD was specifically selected in this study due to the uncertainties in both the satellite and *in situ* datasets (Hu and Le, 2014). For direct comparison to other published validation results, other statistical measures were also calculated, including Root Mean Squared Difference (RMSD), Mean Ratio (MR), MAPD, Mean Relative Bias (MRB), and coefficient of determination (R^2). Simple linear regression slope (β_1) and intercept (β_0) were calculated, as were β_0 and β_1 as determined from reduced major axis (RMA) regression (also termed 'Model II' regression), which accounts for error in the *in situ* data (Sokal and

Rohlf, 1995). For UPD, MRD, MAPD, and MR, margin of error for 95% confidence intervals were calculated as

$$ME_{95} = T_{(N-1)} * {}^{\sigma_{param}}/_{N} \tag{3}$$

where T is the critical t-value for a significance level (α) of 0.025 and N – 1 matchups, and σ_{param} is the standard deviation associated with a parameter (e.g., UPD). The 95% confidence intervals were also calculated for all regression coefficients. To reduce multiplicity, we did not perform pairwise t-tests to compare conditions, instead we considered groups 'statistically significant' only if their 95% confidence intervals did not overlap.

MRD and UPD were assessed according to a variety of exclusion criteria, including Level-2 Processing Flags, water type, QA_Wei, spatial homogeneity, and temporal difference between satellite and *in situ* measurements. Spatial homogeneity was assessed as the coefficient of variation (CV = standard deviation / mean) for the 3 x 3 pixel box with the matchup pixel in the center. In most analyses, satellite data were partitioned into 'Low C_a' and 'High C_a' categories according to the identified water type (Wei et al., 2016), with the former category encompassing water types 1-7 (exclusively offshore waters) and the latter being water types 8-23 (all collected nearshore). For these analyses, only categories with more than 10 matchups that met the conditions were considered. Additionally, for these analyses, implementation (or activation) of Level-2 processing flags is defined as excluding any data with ≥ 4 flag-identified pixels in the 3 x 3 pixel box surrounding the matchup pixel.

3. Results

3.1. In situ sample and matchup characteristics

After the quality control methods were applied, 413 *in situ Rrs* spectra remained for validation against satellite datasets. Temporal distribution of the *in situ* samples generally follows timing of cruise events, with winter months (January – March) and certain years (2014 and 2017) being underrepresented (Fig. 4a). Fortuitously, nearly 57% of *in situ* samples had a same-day matchup with at least one of the satellite datasets studied (N=233; lightly shaded bars in Fig. 4b-c). Over 65% of these matchups, however, were identified as low quality by at least one of the Level-2 Processing Flags considered in this study (i.e., excluding LAND, HILT, and CLDICE, see "current" mask in Table 2), leaving only 81 samples (20% of the original total) matching up with at least one satellite dataset.

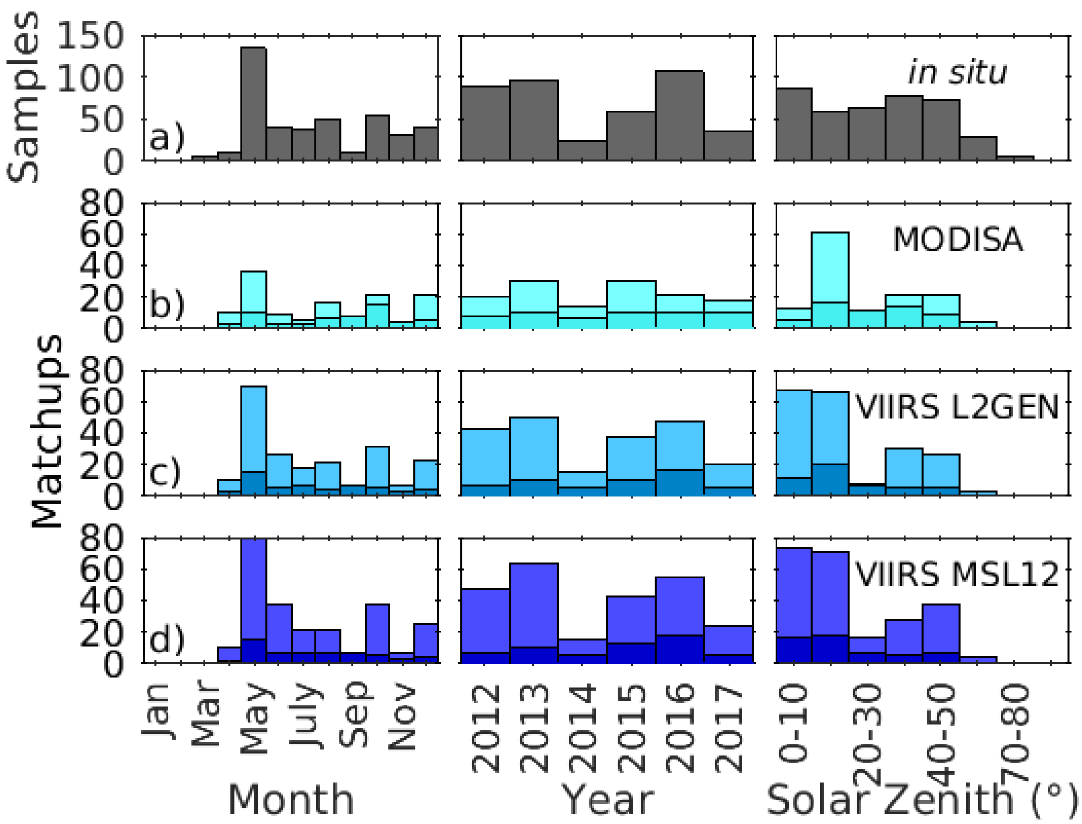

Figure 4: Distribution of (a) *in situ* samples, and (b-d) satellite / *in situ* matchups according to (left column) month, (middle column) year, and (right column) solar zenith angle. For b-d, lighter color shows any satellite / *in situ* matchups, while darker color excludes matchups identified by the "current" L2 flags (see Table 2). Solar Zenith angle histograms in (b-d) represent those for the satellite measurements, while data in (a) are correspond to the *in situ* measurements.

The 233 *in situ Rrs* that matched up with at least one satellite dataset were of overall high quality (mean QA_Wei = 0.9), and included all but two of the water types (9 and 14) described by Wei et al. (2016) (Fig.

5). Of particular note, six of these spectra (colored black in Fig. 5) had very low QA_Wei (mean = 0.2). These were all identified as water type 19, and were collected in Florida Big Bend coastal waters (2-5 m depth) with high chlorophyll concentrations (6-11 mg m^{-3}), extremely high CDOM absorption (a_g(443) = 4-18 m^{-1}), and low *Rrs*(551) (< 0.0005 sr^{-1}).

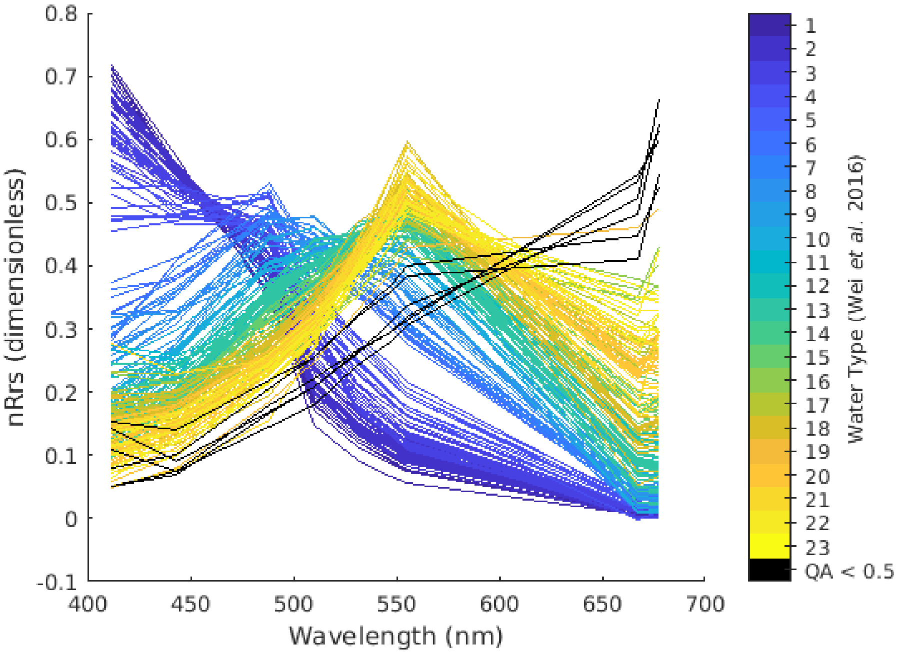

Figure 5: Normalized *Rrs* (*nRrs*) for *in situ* data with satellite matchups. Spectra are colored according to water type (see Figure 4 in Wei et al., 2016), with the exception of black spectra (all identified as water type 19) with low QA_Wei.

3.2. Level-2 Processing Flags

Level-2 Processing Flags are implemented by MSL12 and L2GEN to identify pixels with potentially low *Rrs* quality (e.g., optically complex atmosphere, adjacent to bright targets, bottom effects), which may indicate that the atmospheric correction routines are being applied to conditions outside their design bounds. Therefore, *Rrs* from flag-indicated pixels are likely to have larger uncertainties, thus increasing potential disagreement between satellite and *in situ* measurements. The default L2GEN and MSL12 processing routines both terminate atmospheric correction (thus do not produce *Rrs* or *nLw*) for any pixel identified as HILT, LAND, or CLDICE. Similarly, pixels flagged as ATMFAIL also do not produce *Rrs* (or *nLw*). The remaining flags, however, were individually activated (i.e., matchups were removed from analyses if ≥ 4 pixels were identified by the flag in the 3 x 3 pixel box) to assess impacts on both data quantity and quality (Fig. 6). UPD and MRD were also calculated for several flagging regimes, including "None" (no flags activated except HILT, LAND, CLDICE, and ATMFAIL), "All" (matchups removed if ≥ 4 pixels in the 3 x 3 box were indicated by any flag), "L3 Mask" (see Table 2), and the "Current" mask used for most of this work (Table 2). The latter is based off of the L3 Mask, but ignores the flags COCCOLITH, CHLFAIL, and ABSAER. This mask is thus largely a combination of the masks used by Bailey and Werdell (2006) and Hlaing et al. (2013), although it is slightly more stringent with inclusion of MODGLINT, MAXAERITER, and ATMWARN.

These analyses showed variability in both matchup statistics and data quantity resulting from masking by individual flags or specific masking regimes (Fig 6). With several exceptions, UPD and MRD were closest to 0 for flags (or masking regimes) which disqualified the most pixels. For example, the "all" masking regime (pixels excluded if identified by any flag) resulted in only a few data points with very high quality relative to other masking regimes for nearly all sensors and bands. Note, however, that data quality according to masking regime was not consistent by waveband or sensor. For example, activating the STRAYLIGHT flag resulted in the second lowest UPD values among individual flags for the 410 nm band for all sensors. This flag, however, had little impact on UPD (relative to other individual flags) for most other bands, and was worse (higher UPD and MRD) than other flags for the VIIRS MSL12 671 nm band.

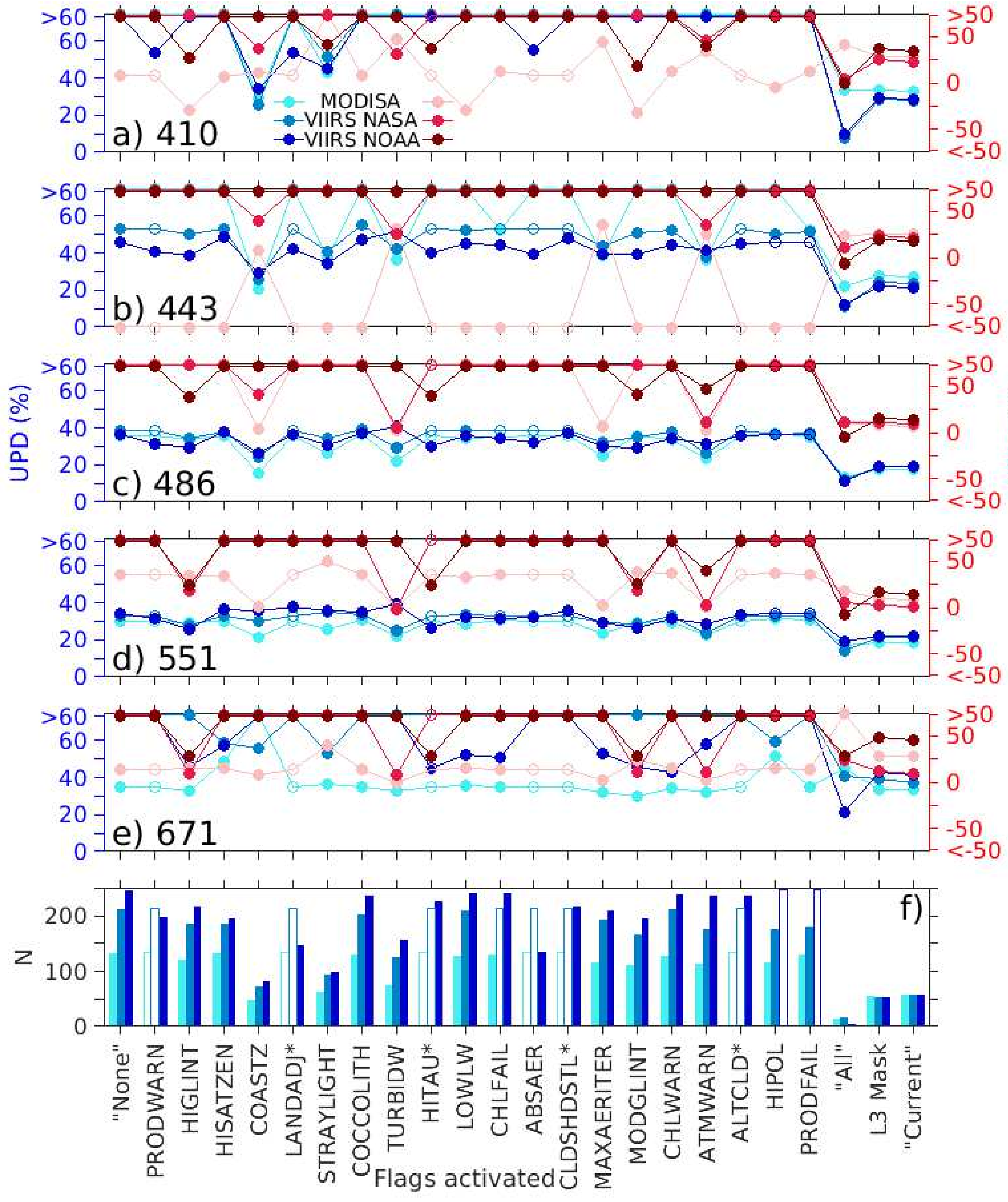

Figure 6: (a-e) UPD (blue shades, left axis) and MRD (red shades, right axis) of MODISA (cyan/pink), VIIRS L2GEN (blue/red), and VIIRS MSL12 (navy/maroon) matchups after masking by various individual flags or masking regimes. Results shown independently for (a) 410, (b) 443, (c) 486, (d) 551, and (e) 671 nm bands. The number of matchups remaining after masking (f) has the same color legend as the UPD data. Hollow data markers shown if there were no instances of flag activation in the dataset. * indicates flags used in MSL12 processing only.

3.3. Quality of satellite *Rrs*

The remaining spectra (i.e., those not masked by the "current" flagging regime) were assessed according to their QA_Wei scores and water types (Fig. 7). To emulate masking criteria as commonly used in validation exercises, all matchup pixels with QA_Wei ≥ various thresholds were used to calculate UPD and MRD. During this comparison, QA_Wei scores for neighboring pixels (i.e., the 3 x 3 pixel box) were not considered. Overall, both the low and high C_a datasets showed little variation in matchup statistics according to QA_Wei (most lines in Fig. 7 are relatively flat, with few significant differences between points). One exception is the most stringent QA_Wei threshold, whereby in datasets restricted to matchups with QA_Wei = 1, jumps in UPD relative to less stringent thresholds were observed (e.g., Low C_a, MODIS 412nm). In one instance (High C_a, VIIRS MSL12 443 nm band), this change was statistically significant (indicated by red circle in Fig. 7). Often this jump was in the positive direction, meaning that the reduction of data quantity was not coupled with improved data quality.

Irrespective of QA_Wei, for all sensors, Low C_a matchups (i.e., those identified as water types 1-7, which were exclusively offshore waters) generally showed improved (lower) UPD (Fig. 7) and reduced MRD (not shown) relative to higher C_a waters (water types 8-23, collected in nearshore waters). This effect was largest for the shorter wavebands, and reduced (or reversed) with increasing wavelength. Additionally, matchup statistics were considerably better for the 486nm and 551nm wavebands as compared to longer and shorter wavelengths.

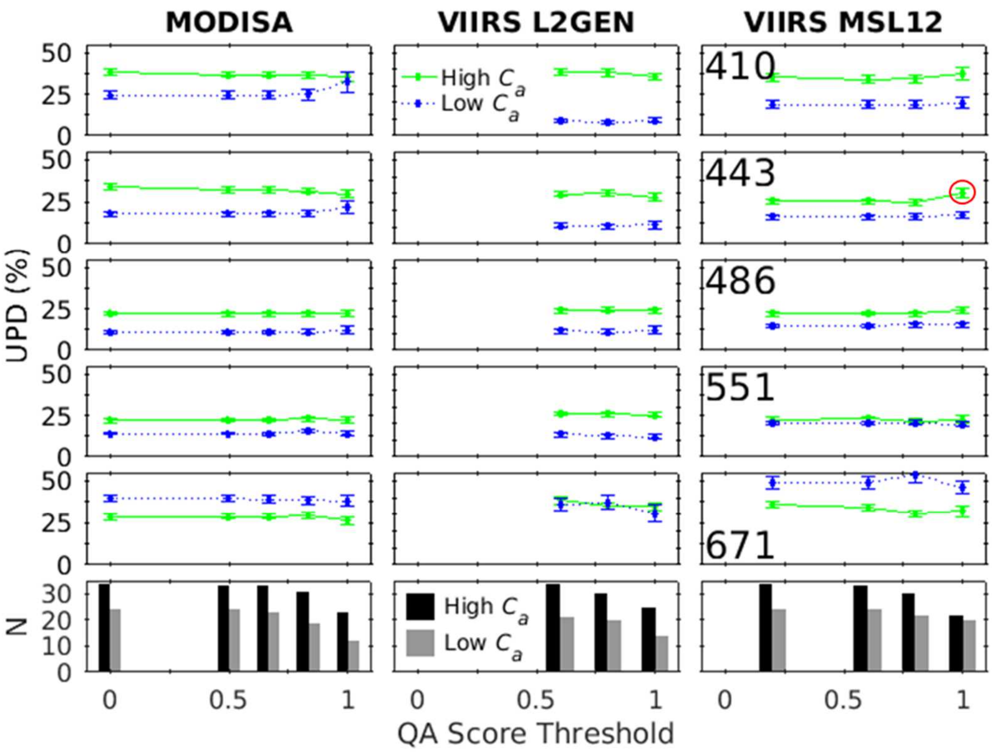

Figure 7: UPD (± 95% confidence intervals) and data quantity (bottom row) for matchup data according to various QA_Wei thresholds (Wei et al., 2016) – all pixels with QA_Wei ≥ the threshold are included in calculated UPD. Data from MODISA (left column), VIIRS L2GEN (center column), and VIIRS MSL12 (right column) are separated by waveband (from top to bottom row: 410, 443, 486, 551, and 671 nm), and partitioned into low C_a (blue dotted lines; water types 1-7) and high C_a (green solid lines; water types 8-23). Red circle indicates significant difference from preceding point (i.e., lower QA threshold).

3.4. Spatial homogeneity

Matchup data which remained after masking by the "current" L2 Flags masking regime were additionally partitioned according to spatial homogeneity, assessed as the CV of the 3x3 pixel box with the matchup location in the center (Fig. 8). MRD and UPD were calculated for all pixels with CV ≤ various thresholds. Note that CV calculations did not include flag-identified pixels (recall that matchups are discarded only if ≥ 4 of the 9 pixels in the 3 x 3 pixel box are flagged). As with the QA_Wei analysis (Section 3.3), this analysis was performed separately for the Low C_a (water types 1-7) and High C_a (water types 8-23) spectra. In most cases, little deviation in UPD or MRD (not shown) was observed for CV thresholds ≥ 0.2.

Results were variable for more stringent (i.e., lower) CV thresholds, with some sensors and bands showing improvement with decreasing CV (e.g., MODIS blue bands), while others show no change or even degradation of matchup statistics (e.g., VIIRS MSL12 blue for high C_a waters). Satellite data in the red bands have higher CV owing to the smaller magnitude of the reflectance data. Note that only a few matchups drive the significant differences between VIIRS L2GEN 442 and 551 nm data for the CV ≤ 0.4 threshold, as compared to the 0.6 threshold.

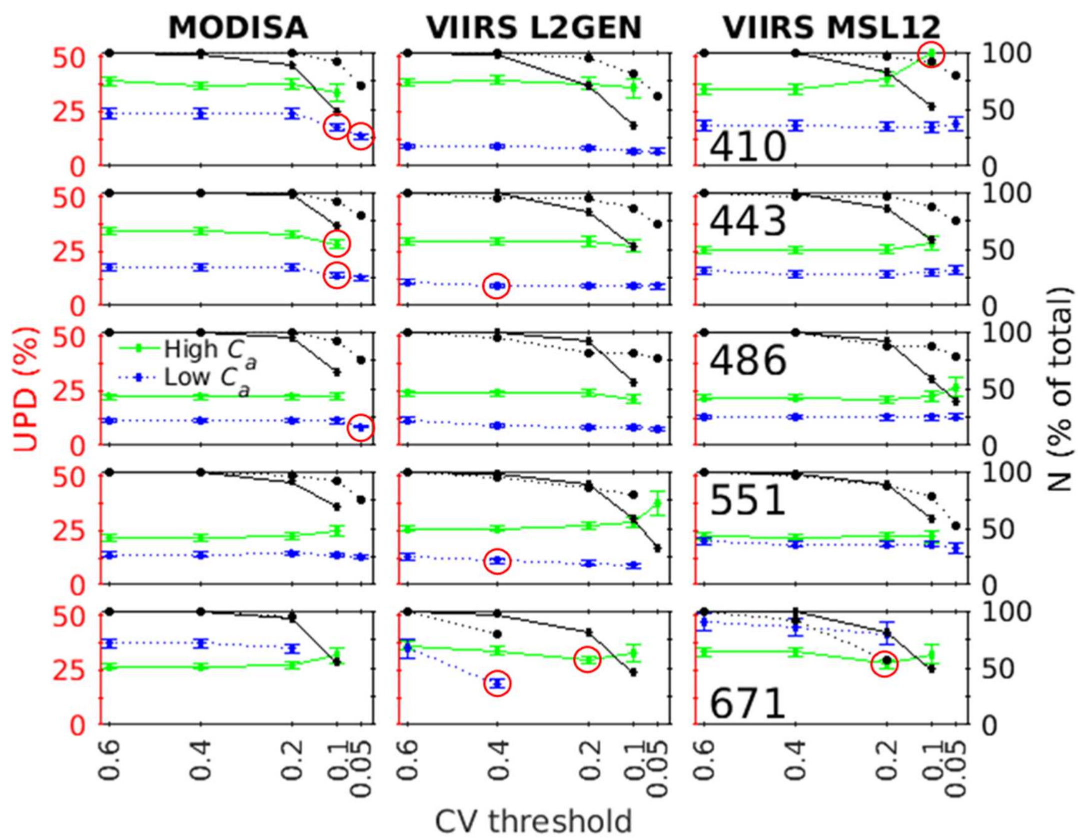

Figure 8: UPD (± 95% confidence intervals; left axes) and data quantity (as a percentage of the total N in each category; black; right axis) for matchup data according to various CV thresholds – UPD values represent all pixels with CV ≤ the CV threshold. Data from MODISA (left column), VIIRS L2GEN (center column), and VIIRS MSL12 (right column) are separated by waveband (from top to bottom row: 410, 443, 486, 551, and 671 nm), and partitioned into low C_a (blue dotted lines; water types 1-7) and high C_a (green solid lines; water types 8-23). Data partitions with N <

10 are excluded. Red circles indicate significant difference from preceding point (i.e., higher CV threshold).

3.5. Temporal concordance

Finally, matchups were assessed according to the temporal gap between the satellite and *in situ* measurement times, again using separate partitions for Low C_a (water types 1-7) and High C_a (water types 8-23) spectra (Fig. 9). Specifically, UPD and MRD were calculated for all pixels (those which were not excluded by "current" L2 Flags masking regime) for which the temporal gap between the satellite and *in situ* data was ≤ various thresholds (1 to 6 hours in 1 hour increments). Although most trends were not statistically significant, for VIIRS data (both L2GEN and MSL12), Low C_a waters showed a general upward trend with tightening temporal difference thresholds, while high C_a waters showed the opposite effect. MODIS data were more variable, especially for high C_a waters, for which increases in UPD associated with tighter temporal overlap criteria were observed for the 410 and 443 nm bands, while a decrease was seen for the 551 nm band. Data quantity was lacking (N < 10) for the Low C_a condition for VIIRS and the High C_a and Low C_a conditions for MODISA, precluding further interpretation of these trends.

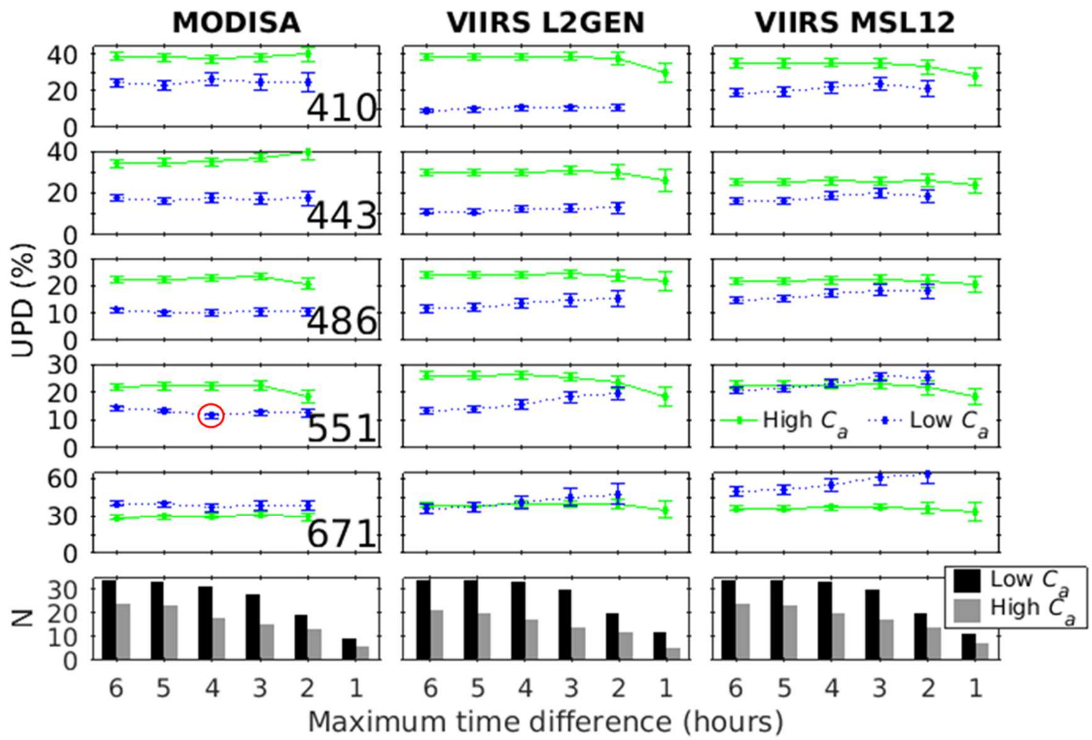

Figure 9: UPD (± 95% confidence intervals) and data quantity (bottom row) for matchup data according to various thresholds of temporal difference between satellite and *in situ* measurements. UPD and MRD values represent all pixels with time difference ≤ the maximum threshold. Data from MODISA (left column), VIIRS L2GEN (center column), and VIIRS MSL12 (right column) are separated by waveband (from top row: 410, 443, 486, 551, and 671 nm), and partitioned into low C_a (blue dotted lines; water types 1-7) and high C_a (green solid lines; water types 8-23). Data partitions with N < 10 are excluded. Unlike Figures 6-8, axis limits are not the same for all wavebands. Red circle indicates significant difference from preceding point (i.e., longer threshold for temporal difference between measurements).

3.6. Overall matchup statistics

The analyses above highlight some examples of improvement (although variable by band) in matchup statistics through the application of various L2 Flags or masking regimes. However, none of the other methods to cull low quality data individually showed widespread (across sensors and bands) effectiveness at improving the statistical relationships. As such, we compared scatterplot and matchup statistics for three QA schemes: (1) masking using the minimal L2 Flags ("none" mask; i.e., all matchups are allowed), (2) implementation of the "current" L2 Flags mask, and (3) implementation of both the "current" L2 Flags mask and thresholds for CV and temporal overlap of 0.2 and 2 h, respectively (Zibordi et al., 2009). These results are presented in scatterplots (Fig. 10), as well as tabular form for the latter two datasets (Tables 3-5). Again, CV calculations do not incorporate flag-indicated pixels, while pixels with ≥ 4 flagged pixels in the 3 x 3 pixel box are excluded.

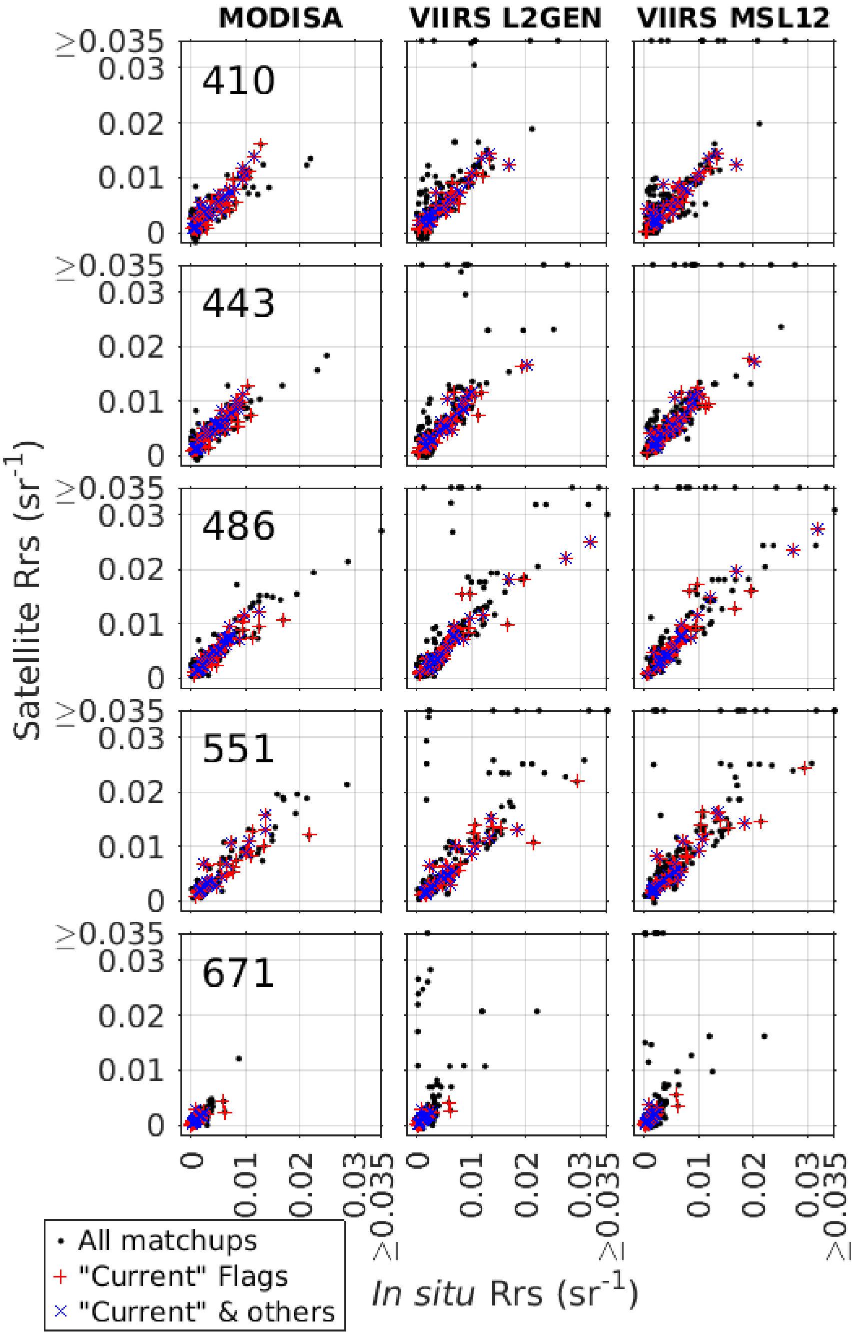

Figure 10: Scatterplots showing *in situ* / satellite *Rrs* (sr^{-1}) matchups for three QA schemes: all matchups (L2 flags regime “None”, black dots), “Current” Flags activated (red ‘+’), and “Current” Flags activated, CV < 0.2, and temporal overlap < 2h (blue ‘x’). Data shown separately for MODISA (left column), VIIRS L2GEN (middle column), and VIIRS MSL12 (right column), and by waveband (from top row: 410, 443, 486, 551, and 671 nm).

Overall, Fig. 10 and Tables 3-5 show a general concordance between satellite and *in situ* data, with the exception of obvious outliers that were exclusively restricted to the most lenient flagging scheme. Nevertheless, variable performance was seen in matchup statistics between these three QA masking schemes and by satellite dataset. The QA masking schemes also had substantial impacts on data quantity and dynamic range, with increasingly stringent masking schemes generally culling at least half or more of the data, particularly affecting higher *Rrs* values (as measured both *in situ* and by satellite). Interestingly, the different metrics used occasionally disagreed on the “best” performing QA scheme (Tables 3-5). For example, looking at MODISA *Rrs*(412) matchups (Table 3), UPD identified the most restrictive mask as better performing, however MRD and MR found that the dataset masked only by the “Current” L2 Flags outperformed the other masking scheme.

Table 3: Matchup statistics for MODISA data according to two QA schemes.

	“Current” L2 Flags applied					“Current” Mask, CV < 0.2, +/- 2 h				
Band (nm)	412	443	488	547	667	412	443	488	547	667
UPD (%)	33 (1)	27 (1)	18 (1)	19 (1)	33 (1)	29 (2)	26 (2)	14 (1)	16 (1)	29 (2)
MRD (%)	27 (2)	24 (2)	7 (1)	8 (1)	27 (2)	39 (6)	34 (3)	13 (2)	13 (2)	29 (3)
RMSD	0.0014	0.0012	0.0013	0.0018	0.0007	0.0014	0.0011	0.0009	0.0012	0.0004
MAPD (%)	43 (2)	35 (2)	19 (1)	21 (1)	42 (2)	45 (5)	37 (3)	17 (1)	19 (2)	38 (3)
MR	0.98 (0.02)	0.92 (0.01)	0.99 (0.01)	0.99 (0.01)	0.93 (0.01)	0.84 (0.02)	0.82 (0.01)	0.91 (0.01)	0.93 (0.01)	0.86 (0.02)
R^2	0.87	0.85	0.84	0.82	0.71	0.89	0.91	0.93	0.91	0.67
β_0 (*10^4)	2.9 (5.8)	6.4 (5.7)	9.5 (5.9)	9.8 (5.9)	3.5 (1.6)	6.8 (8.2)	7.1 (6.2)	4.4 (5.8)	3.2 (6.7)	2 (2.4)
β_1	1.03 (0.1)	0.94 (0.11)	0.82 (0.1)	0.76 (0.09)	0.61 (0.11)	1.01 (0.15)	1 (0.12)	0.99 (0.11)	1 (0.13)	0.87 (0.24)
RMA β_0 (*10^4)	-0.2 (4.8)	2.8 (5)	5.6 (5.3)	6.1 (4.6)	2.3 (1.1)	4.1 (7.1)	5.1 (5.5)	2.7 (5.3)	1.1 (5.4)	0.6 (2)
RMA β_1	1.11 (0.1)	1.02 (0.1)	0.89 (0.09)	0.84 (0.09)	0.73 (0.1)	1.08 (0.14)	1.05 (0.11)	1.02 (0.1)	1.06 (0.12)	1.07 (0.22)
Max (sr^{-1})	0.013	0.011	0.017	0.022	0.006	0.012	0.009	0.012	0.014	0.002
N	58	58	58	58	58	27	30	29	30	29

**Numbers in parentheses indicate 95% confidence intervals (± ME) for listed statistics, underlined values indicate significant improvement.

Table 4: Matchup statistics for VIIRS L2GEN data according to two QA schemes.

	“Current” L2 Flags applied					“Current” Mask, CV < 0.2, +/- 2 h				
Band (nm)	410	443	486	551	671	410	443	486	551	671

UPD (%)	27 (1)	23 (1)	19 (1)	21 (1)	37 (1)	25 (2)	20 (2)	17 (1)	21 (2)	27 (2)
MRD (%)	21 (2)	20 (2)	9 (1)	1 (1)	10 (3)	26 (5)	20 (3)	7 (2)	0 (3)	7 (5)
RMSD	0.0013	0.0015	0.0021	0.0024	0.0007	0.0015	0.0015	0.0019	0.0018	0.0005
MAPD (%)	37 (2)	30 (2)	22 (1)	23 (1)	43 (2)	36 (5)	26 (3)	19 (2)	23 (3)	31 (4)
MR	0.97 (0.01)	0.93 (0.01)	0.98 (0.01)	1.08 (0.01)	1.26 (0.03)	0.92 (0.02)	0.89 (0.02)	0.99 (0.02)	1.09 (0.02)	1.06 (0.03)
R^2	0.9	0.88	0.87	0.84	0.7	0.87	0.87	0.95	0.84	0.6
β_0 ($*10^4$)	6.4 (5.4)	10 (6.3)	12.9 (7.8)	10.3 (7.3)	2.9 (1.7)	11.3 (9.4)	13.5 (9.5)	12.2 (7.4)	6.6 (10.5)	3.2 (3.9)
β_1	0.9 (0.08)	0.87 (0.09)	0.82 (0.09)	0.75 (0.09)	0.61 (0.11)	0.86 (0.14)	0.85 (0.14)	0.8 (0.07)	0.82 (0.15)	0.71 (0.28)
RMA β_0 ($*10^4$)	3.8 (4.6)	6.7 (5.3)	8.9 (6.1)	6.3 (5.6)	1.7 (1.2)	8 (8.1)	10.1 (8.3)	10.8 (5.4)	2.6 (9)	0.7 (3.8)
RMA β_1	0.95 (0.08)	0.93 (0.09)	0.88 (0.08)	0.82 (0.08)	0.73 (0.1)	0.91 (0.13)	0.91 (0.13)	0.82 (0.07)	0.89 (0.14)	0.92 (0.24)
Max (sr^{-1})	0.017	0.02	0.032	0.029	0.006	0.017	0.02	0.032	0.019	0.003
N	55	55	55	55	55	26	26	27	26	21

**Numbers in parentheses indicate 95% confidence intervals (± ME) for listed statistics, underlined values indicate significant improvement.

Table 5: Matchup statistics for VIIRS MSL12 data according to two QA schemes.

	"Current" L2 Flags applied					"Current" Mask, CV < 0.2, +/- 2 h				
Band (nm)	410	443	486	551	671	410	443	486	551	671
UPD (%)	28 (1)	22 (1)	19 (1)	22 (1)	41 (1)	31 (3)	22 (2)	18 (1)	22 (2)	31 (3)
MRD (%)	34 (3)	18 (2)	13 (1)	14 (1)	45 (3)	49 (9)	23 (4)	13 (3)	15 (4)	44 (7)
RMSD	0.0015	0.0015	0.0021	0.0021	0.0006	0.0019	0.0015	0.0017	0.0019	0.0007
MAPD (%)	46 (3)	28 (2)	23 (1)	26 (1)	61 (3)	58 (9)	31 (4)	22 (2)	28 (3)	48 (6)
MR	0.95 (0.02)	0.94 (0.01)	0.94 (0.01)	0.96 (0.01)	1.01 (0.03)	0.86 (0.02)	0.91 (0.02)	0.94 (0.02)	0.96 (0.02)	0.81 (0.03)
R^2	0.85	0.87	0.87	0.86	0.71	0.82	0.86	0.95	0.83	0.57
β_0 ($*10^4$)	11.1 (6.2)	10.3 (6.2)	12.7 (7.7)	11.1 (7.6)	3.4 (2)	16.1 (10.3)	11.9 (9.1)	11.7 (8.2)	7.3 (11.8)	2.8 (4.8)
β_1	0.88 (0.1)	0.88 (0.09)	0.87 (0.09)	0.86 (0.09)	0.78 (0.13)	0.83 (0.15)	0.88 (0.14)	0.88 (0.08)	0.93 (0.17)	0.97 (0.4)
RMA β_0 ($*10^4$)	7.5 (5.2)	6.8 (5.3)	8.7 (6.1)	7.1 (5.8)	2 (1.4)	11.7 (8.8)	8.3 (7.9)	10.1 (6)	2.5 (9.9)	-0.3 (4.4)
RMA β_1	0.96 (0.09)	0.95 (0.09)	0.93 (0.08)	0.93 (0.09)	0.92 (0.12)	0.91 (0.14)	0.95 (0.13)	0.9 (0.08)	1.02 (0.16)	1.29 (0.34)
Max (sr^{-1})	0.017	0.02	0.032	0.029	0.006	0.017	0.02	0.032	0.019	0.002
N	58	58	58	58	58	28	29	28	27	22

**Numbers in parentheses indicate 95% confidence intervals (± ME) for listed statistics, underlined values indicate significant improvement.

4. Discussion

4.1. Overall performance

These analyses, in the aggregate, show reliable performance of both the MODISA and VIIRS instruments as well as the most recent calibration efforts (and associated atmospheric correction routines) and reprocessing efforts of both NOAA (April 2017 SDR) and NASA (2018.0). For all three datasets studied, UPD for the green band *Rrs* hovers around 20%, only slightly higher than the ~15 % MAPD reported by numerous other studies (Table 1). When matchups identified as "Low C_a" (water types 1-7) were analyzed independently, results showed UPD and MRD very close to those previously reported MAPD of

~15%. Most datasets showed slight positive bias relative to *in situ* data for all wavebands. For MODISA, this contrasts with some previous assessments (Antoine et al., 2008; Maritorena et al., 2010; Mélin et al., 2007; Zibordi et al., 2009), but agrees with more recent findings (Hlaing et al., 2013). Note, however, that changes in bias may result directly from changes to instrument calibration coefficients, which can vary by processing and calibration versions (see Table 1 for versions used in previous validation efforts). Also, because the purpose of this study was to show the effects of QA procedures on uncertainties estimates, no attempt was made to separate the uncertainties from the satellite and *in situ* sources. The final uncertainty estimates thus inherently contain those from *in situ* measurements.

Aggregate results were also variable between the MSL12- and L2GEN-based VIIRS processing schemes (Tables 4-5), with neither proving consistently more accurate (even when considering only common pixels, results not shown). The MSL12-based VIIRS processing resulted in slightly more matchup points than L2GEN-based VIIRS processing when identical flagging schemes were used (Tables 4-5). Moreover, response to QA procedures were occasionally variable between these two datasets, especially for individual L2 flags (Fig. 6). We also note an apparent preference in the literature for the L2GEN (SeaDAS) processing for VIIRS data. This is perhaps due to familiarity within the ocean color community to the SeaDAS software package (MSL12 is much newer) or to availability of the SeaDAS software for custom processing and application to other sensors.

The difference in performance between the “All Matchups” and “Current” L2 Flags masking regimes (Fig. 10) also highlights the effectiveness of the L2 flags as a QA method. For MODISA data, the default L2 Flags (LAND, HILT, CLDICE, ATMFAIL) masking scheme performed well, with few obvious outliers in any band (Fig. 10), and decent matchup statistics for the green and red bands. For the VIIRS datasets, however, outliers after simple default L2 Flag masking were much more prominent (Fig. 10). Activating the “Current” L2 mask for VIIRS data removed most of these outliers and generally reduced (i.e., improved) the UPD by approximately half (much more in some cases), with even larger improvement in MRD. This impact was not as drastic for MODISA data, especially for the 671 nm band.

4.2. QA Methods

As with any satellite ocean color investigation or algorithm development study, validation analyses inherently include a compromise between data quantity and quality. For both the *in situ* and satellite datasets, the approach is generally to include the largest number of matchups with the largest dynamic range (thereby maximizing statistical power) without compromising from the highest quality data available.

L2 flags are typically the first tool used to cull satellite measurements of potentially reduced quality. Generally, this is performed with little (if any) assessment on their impacts to both the quality and quantity of the matchup dataset as a whole. In this study, we found variability in the impact of individual L2 Flags by wavelength, both in terms of data quantity and quality (Fig. 6). In particular, flags for conditions associated with coastal waters (e.g., COASTZ, LANDADJ, TURBIDW, and ABSAER) often identified the largest number of pixels. Activating these flags caused improvement in matchup statistics for the blue bands, but the effects were much more muted for other bands (even substantially diminishing statistics for the red band). Another consequence of activating these flags, however, is a large restriction in the dynamic range of the validation dataset, as most nearshore and optically complex waters are identified and removed by these flags. The STRAYLIGHT flag also caused a large reduction in the quantity of data, and resulted in matchup statistic trends similar to those of the “coastal” flags mentioned above. The STRAYLIGHT flag is implemented as a 5x7 pixel box from any HILT pixel, which includes land targets. As a consequence, the STRAYLIGHT flag masks many estuarine matchups (see Fig.

3). Nevertheless, we included the STRAYLIGTH flag in our “Current” mask due to precedent in the literature (Table 2) as well as our determination that the improvements in matchup statistics outweighed the negative impacts on data quantity and dynamic range. Note that while Feng and Hu (2016) suggested that the STRAYLIGHT flag could be implemented as a 3x3 pixel box without sacrificing data quality in open ocean waters, it is not clear if this finding holds for nearshore waters, and assessment of such was beyond the scope of the present study.

As noted in Table 2, there is no real consensus on which particular flags should be applied when performing validation of satellite data. Indeed, most studies do not even list the specific flags used for this purpose. Nevertheless, the general assumption is that removing more flag-identified pixels will improve validation results. The analysis of UPD and MRD changes resulting from the removal of data identified by individual flags (Fig. 6) challenges this assumption. For example, activation of certain flags (e.g., TURBIDW) often decreased performance relative to the unmasked (“no” flags) dataset. Furthermore, although the “All” flags mask produced the best (or close to the best) matchup statistics for most bands and sensors, MODIS red band matchups remaining after application of this mask were actually worse than the “no” flags dataset (Fig. 6). For the 486 and 551 bands, activation of “All” flags showed no substantial improvement in UPD or MRD relative to the “Current” or “L3 Mask” flagging schemes, especially when considering that 63-87% of the data were disqualified. Nevertheless, it should be remembered that these L2 Flags represent globally optimized QA procedures as implemented by the processing agencies (NASA and NOAA). Use (or exclusion) of these flags for validation purposes should be done with caution – researchers need to consider if the masking scheme is justifiable.

Beyond L2 Flags, three additional different QA schemes were individually assessed for impacts on both data quantity and quality (Fig. 7-9). With some specific exceptions, none of these methods demonstrated widespread applicability for improvement in matchup statistics. Given the widespread use of these methods in culling data (Table 1), this result is somewhat surprising, but not unprecedented (Barnes and Hu, 2015; Mélin et al., 2007). Note that the matchup statistics for these three QA schemes were calculated after implementation of the "current" L2 flags mask, so these findings might not hold true for solo implementation. Indeed, the average QA_Wei values for the satellite datasets masked with "no" flags is quite low (0.73 - 0.78) relative to those after excluding pixels identified by the "Current" mask (0.90 - 0.92), meaning Fig. 7 would show much more substantial trends when calculated using the "no" flags data.

Similar to the L2 Flags analyses, comparison of various QA_Wei thresholds (Fig. 7) highlights an issue with unsupervised exclusion of data points meeting (or failing to meet) certain criteria. Specifically, both the *in situ* and satellite datasets included spectra with extremely low QA_Wei (even QA_Wei = 0), many of which were collected in "dark" or "black" coastal waters where *Rrs*(551) < 0.0005 sr^{-1}. Water samples associated with *in situ* spectra show high chlorophyll concentrations (6-11 mg m^{-3}) and CDOM absorption (a_g(443) = 4-18 m^{-1}). Indeed, it seems that such waters are not represented in any of the QA_Wei water types, indicating the need for revision of that metric to either include an additional water type or relax the boundaries of an existing water type (likely 19) to include such conditions. In either case, it is often difficult to obtain valid satellite *Rrs* in such waters due to low signal:noise and atmospheric correction uncertainties.

It is also important to highlight that although none of the QA schemes (beyond L2 Flags) resulted in widespread improvement in matchup statistics (Figs. 7-9), scatterplots (Fig. 10) do show a few individual outliers which are included in the dataset with only "Current" flags applied, but removed from the dataset with additional CV and temporal difference thresholds. In Figure 10, these show as red '+'

without overlying blue 'x.' This is especially apparent in the 486 and 551 bands (Fig. 10) for all sensors, and is indicated mostly via improvements in RMSD, R^2, and β_1 (Tables 3-5). These outliers are largely coastal, and thus have somewhat smaller impacts on other metrics (i.e., UPD, MRD, MAPD, and MR) due to the larger denominator. Thus, we note that (1) the choice of metric is important, with various metrics showing differences depending on the data quantity and dynamic range; while (2) multiple QA schemes implemented in concert may show improvements in matchup statistics that are not apparent in solo implementations.

4.3. Limitations and recommendations

To our knowledge, the findings reported here represent the first attempt to extensively document effects of QA exclusion methods on satellite / *in situ Rrs* validation statistics. We have largely refrained from pairwise comparisons for each of the studied groupings, primarily because limited data quantity does not support such rigorous analysis for the multitude of QA options and thresholds assessed. Even in the absence of such statistics, the number of data points excluded by each incrementally tightening QA threshold is extremely important. For instance, a small quantity of matchups in highly heterogeneous environments (in time or space) may lead one to the conclusion that time difference between measurements or CV have little impact. Thus, we have refrained from drawing conclusions from changes in UPD resulting from only a few data points. Likewise, because different applications may have different requirements on uncertainties, it is impractical to define which matchup criteria lead to uncertainties meeting various requirements. This is especially true when considering that even the highest-quality MODIS reflectance data from ocean gyres can show reflectance uncertainties higher than the traditional requirements of 5% for blue bands in waters with C_a > 0.1 mg m^{-3} (Hu et al., 2013). For more productive waters, reflectance uncertainties can be substantially higher (Moore et al., 2014).

Although we tested implementation of several QA schemes (and combinations thereof) beyond those shown here, the results generally showed limited (and variable) impacts similar to those presented here. This is especially true across wavebands, as QA approaches that appear to provide maximum statistical benefit for blue bands often diminish results for green and red bands. This presents a challenge for identifying best-practice recommendations for future studies involving satellite / *in situ* matchups. We are similarly hesitant to unequivocally state that the results found here will generalize to other datasets. Additionally, we recognize that different datasets and / or objectives may be best suited by disparate QA approaches.

On the other hand, it is also not our goal to advocate an “anything goes” approach to removing low quality data, as some level of standardization is important towards attaining comparable results across studies. It is also especially important to emphasize that decisions with respect to the specific flagging scheme and QA procedures need to be made with consideration of the real impacts to the dataset (e.g., reduction in data quantity, decrease in dynamic range, or exclusion of data from a specific environment or with an otherwise common attribute). Without this consideration, researchers can artificially improve matchup statistics by selectively implementing QA procedures that remove undesirable data.

Therefore, we argue that the process detailed in this work (or a simplified version) can be applied as an important component to validation works going forward, allowing investigators to make informed determinations of the QA techniques and thresholds which most effectively remove low quality data while maximizing retained data quantity and retaining robustness of the dataset. While not necessary to test impacts of each individual L2 flag, quantifying the effects of a few flag combinations may lead to significant improvements (or degradations) in results. Of course, the final selection of flags must be made with consideration of the reason why a particular flag should be excluded. For example, the

COASTZ flag uses a static bathymetry to identify pixels shallower than 30m. While excluding pixels indicated by COASTZ would likely improve matchup results in many cases, this is alone is not a justifiable culling method for validation activities.

With some modification, QA_Wei may be another effective method to identify low-quality data, although it is likely duplicative with L2 Flags. Fig. 8 provides some evidence that CV thresholds can be effective in offshore waters (low C_a), which concords with their stated purpose. However, it is clear that some of the more stringent data culling thresholds may actually degrade statistical performance. In most cases, for coastal waters, reducing the temporal gap between satellite and *in situ* measurements improved performance (which comports with intuition), while smaller disimprovements in performance were noted with tightening temporal gaps for offshore waters. Where possible, matchups should be extracted at Level-2 to avoid issues related to homogeneity assessment at scan edges. As for the particular statistical metrics, given the uncertainties associated with *in situ Rrs* data (Hooker et al., 2002; Hooker and Maritorena, 2000), we recommend use of UPD and RMA regression (as opposed to the more widely used MAPD and simple linear regression). Although it is difficult to statistically compare disparate metrics (e.g., UPD vs MAPD), with a few exceptions, UPD and RMA coefficients were improved as compared to their more commonly used analogs.

Finally, the statistical measures (UPD, MPD, etc.) presented here represent those from point matchups after applying various QA techniques, and they do not represent uncertainties in satellite global products after spatial and temporal binning. The spatial homogeneity test and temporal matchup windows, in addition to other QA criteria, are intended to serve as the best effort to minimize the impact of differences between *in situ* measurements (point sample) and satellite measurements (integrated ≥1 km^2 pixel). These criteria are not and should not be used when generating global products. Additionally, uncertainties in the global products are expected to reduce significantly as data at pixel-resolution are binned in space and/or time (Qi et al., 2017). The intention of this study is therefore to provide a comparison and recommendation on the QA criteria when validating satellite-derived *Rrs* data products rather than detailing the various uncertainty sources in satellite data products at various spatial and temporal scales. For the latter, readers are referred to a recent community effort led by the International Ocean Colour Coordination Group (IOCCG, Mélin and Doerffer, 2015).

5. Conclusions

In this paper, we quantify the statistical performance of commonly used satellite reflectance datasets against a collection of high-quality *in situ* data and critically assess some standards used in validation exercises. The overall strong validation statistics reflect positively on the calibration efforts and atmospheric correction schemes developed by both NOAA and NASA. The variability in results according to QA regimes leads us to recommend that future studies include some consideration of the impacts of methods used to discard low quality data, followed by clear presentation of the methods used in generation of the final results. These moderate changes will hopefully lead to larger datasets with wider dynamic range being used in validation studies, with documentation allowing fair tracking of satellite ocean color data over time (and across processing versions), towards the ultimate goal of ensuring high quality and consistent environmental data records across multiple satellites.

6. Acknowledgments

This work was supported by NASA (NNX16AQ71G) and by the Joint Polar Satellite System (JPSS) funding for the NOAA ocean color calibration and validation (Cal/Val) project (NA15OAR4320064). The authors wish to thank both NASA and NOAA for providing the processed satellite data used in this study. The *in situ* data were collected mainly by the USF Optical Oceanography Lab and Florida Fish and Wildlife Research Institute from targeted cruise surveys or cruises of opportunity, including DEEPEND cruises

(DP01-DP05). Some of the *in situ* data used in this study were collected during the NOAA dedicated VIIRS ocean color calibration and validation (Cal/Val) cruises (NF-14-09, NF-15-13, NF-16-08) and the GOMECC-3 cruise, supported by the Joint Polar Satellite System (JPSS) program, NOAA Office of Marine and Aviation Operations, and GOMECC. We thank many individuals on these cruises for their assistance in collecting field data. Gulf of Mexico Research Initiative (GoMRI) data are publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org.

7. References

Ahmed, S., Gilerson, A., Hlaing, S., Weidemann, A., Arnone, R., Wang, M., 2013. Evaluation of ocean color data processing schemes for VIIRS sensor using in-situ data of coastal AERONET-OC sites. Proc. SPIE - Int. Soc. Opt. Eng. 8888. https://doi.org/10.1117/12.2028821

Antoine, D., d'Ortenzio, F., Hooker, S.B., Bécu, G., Gentili, B., Tailliez, D., Scott, A.J., 2008. Assessment of uncertainty in the ocean reflectance determined by three satellite ocean color sensors (MERIS, SeaWiFS and MODIS-A) at an offshore site in the Mediterranean Sea (BOUSSOLE project). J. Geophys. Res. 113, C07013. https://doi.org/10.1029/2007JC004472

Bailey, S.W., Franz, B.A., Werdell, P.J., 2010. Estimation of near-infrared water-leaving reflectance for satellite ocean color data processing. Opt. Express 18, 7521–7527. https://doi.org/10.1364/OE.18.007521

Bailey, S.W., Werdell, P.J., 2006. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sens. Environ. 102, 12–23. https://doi.org/10.1016/j.rse.2006.01.015

Barnes, B.B., Hu, C., 2016. Dependence of satellite ocean color data products on viewing angles: A comparison between SeaWiFS, MODIS, and VIIRS. Remote Sens. Environ. 175, 120–129. https://doi.org/10.1016/j.rse.2015.12.048

Barnes, B.B., Hu, C., 2015. Cross-sensor continuity of satellite-derived water clarity in the Gulf of Mexico: Insights into temporal aliasing and implications for long-term water clarity assessment. IEEE Trans. Geosci. Remote Sens. 53, 1761–1772. https://doi.org/10.1109/TGRS.2014.2348713

Barnes, B.B., Hu, C., Schaeffer, B.A., Lee, Z., Palandro, D.A., Lehrter, J.C., 2013. MODIS-derived spatiotemporal water clarity patterns in optically shallow Florida Keys waters: A new approach to remove bottom contamination. Remote Sens. Environ. 134. https://doi.org/10.1016/j.rse.2013.03.016

Blackwell, S.M., Moline, M.A., Schaffner, A., Garrison, T., Chang, G., 2008. Sub-kilometer length scales in coastal waters. Cont. Shelf Res. 28, 215–226. https://doi.org/10.1016/j.csr.2007.07.009

Brando, V.E., Lovell, J.L., King, E.A., Boadle, D., Scott, R., Schroeder, T., 2016. The potential of autonomous ship-borne hyperspectral radiometers for the validation of ocean color radiometry data. Remote Sens. 8. https://doi.org/10.3390/rs8020150

Brown, C., Huot, Y., Werdell, P., Gentili, B., Claustre, H., 2008. The origin and global distribution of second order variability in satellite ocean color and its potential applications to algorithm development. Remote Sens. Environ. 112, 4186–4203. https://doi.org/10.1016/j.rse.2008.06.008

Cao, C., Xiong, X., Wolfe, R., DeLuccia, F., Liu, Q., Blonski, S., Lin, G., Nishihama, M., Pogorzala, D., Oudrari, H., Hillger, D., 2013. Visible Infrared Imaging Radiometer Suite (VIIRS) Sensor Data Record (SDR) User's Guide, Version 1.2. NOAA Technical Report NESDIS 142. Washington, D.C.

Feng, L., Hu, C., 2016. Cloud adjacency effects on top-of-atmosphere radiance and ocean color data products: A statistical assessment. Remote Sens. Environ. 174, 301–313. https://doi.org/10.1016/j.rse.2015.12.020

Franz, B.A., Bailey, S.W., Werdell, P.J., McClain, C.R., 2007. Sensor-independent approach to the vicarious calibration of satellite ocean color radiometry. Appl. Opt. 46, 5068–82.

Garaba, S.P., Zielinski, O., 2013. Methods in reducing surface reflected glint for shipborne above-water remote sensing. J. Eur. Opt. Soc. 8. https://doi.org/10.2971/jeos.2013.13058

Gordon, H.R., Clark, D.K., 1981. Clear water radiances for atmospheric correction of coastal zone color scanner imagery. Appl. Opt. 20, 4175–4180. https://doi.org/10.1364/AO.20.004175

Gordon, H.R., Du, T., Zhang, T., 1997. Remote sensing of ocean color and aerosol properties: resolving the issue of aerosol absorption. Appl. Opt. 36, 8670. https://doi.org/10.1364/AO.36.008670

Gordon, H.R., Wang, M., 1994. Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm. Appl. Opt. 33, 443–452. https://doi.org/10.1364/AO.33.000443

Harding, L.W., Magnuson, A., Mallonee, M.E., 2005. SeaWiFS retrievals of chlorophyll in Chesapeake Bay and the mid-Atlantic bight. Estuar. Coast. Shelf Sci. 62, 75–94. https://doi.org/10.1016/j.ecss.2004.08.011

Hlaing, S., Gilerson, A., Foster, R., Wang, M., Arnone, R., Ahmed, S., 2014. Radiometric calibration of ocean color satellite sensors using AERONET-OC data. Opt. Express 22, 23385. https://doi.org/10.1364/OE.22.023385

Hlaing, S., Harmel, T., Gilerson, A., Foster, R., Weidemann, A., Arnone, R., Wang, M., Ahmed, S., 2013. Evaluation of the VIIRS ocean color monitoring performance in coastal regions. Remote Sens.

Environ. 139, 398–414. https://doi.org/10.1016/j.rse.2013.08.013

Hooker, S.B., Esaias, W.E., 1993. An overview of the SeaWiFS Project. Eos, Trans. Am. Geophys. Union. https://doi.org/10.1029/93EO00945

Hooker, S.B., Esaias, W.E., Feldman, G.C., Gregg, W.W., McClain, C.R., 1992. An overview of SeaWiFS and ocean color. NASA Tech. Memo., vol. 104566. Greenbelt, MD.

Hooker, S.B., Lazin, G., Zibordi, G., McLean, S., 2002. An Evaluation of Above- and In-Water Methods for Determining Water-Leaving Radiances. J. Atmos. Ocean. Technol. 19, 486–515.

Hooker, S.B., Maritorena, S., 2000. An evaluation of oceanographic radiometers and deployment methodologies. J. Atmos. Ocean. Technol. 17, 811–830. https://doi.org/10.1175/1520-0426(2000)017<0811:AEOORA>2.0.CO;2

Hu, C., Barnes, B.B., Qi, L., Corcoran, A.A., 2015. A harmful algal bloom of Karenia brevis in the northeastern Gulf of Mexico as revealed by MODIS and VIIRS: A comparison. Sensors (Switzerland) 15. https://doi.org/10.3390/s150202873

Hu, C., Carder, K.L., Muller-Karger, F.E., 2001. How precise are SeaWiFS ocean color estimates? Implications of digitization-noise errors. Remote Sens. Environ. 76, 239–249. https://doi.org/10.1016/S0034-4257(00)00206-6

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

Hu, C., Le, C., 2014. Ocean Color Continuity From VIIRS Measurements Over Tampa Bay. IEEE Geosci. Remote Sens. Lett. 11, 945–949. https://doi.org/10.1109/LGRS.2013.2282599

Jiang, L., Wang, M., 2014. Improved near-infrared ocean reflectance correction algorithm for satellite ocean color data processing. Opt. Express 22, 21657. https://doi.org/10.1364/OE.22.021657

Kovach, C., Ondrusek, M., 2018. Uncertainties associated with ocean color satellite data, in: 2018 Ocean Sciences Meeting, Portland, OR. IS44A-2677.

Le, C., Hu, C., 2013. A hybrid approach to estimate chromophoric dissolved organic matter in turbid estuaries from satellite measurements : A case study for Tampa. Opt. Express 21, 18849–18871. https://doi.org/10.1364/OE.21.018849

Le, C., Hu, C., Cannizzaro, J., English, D., Muller-Karger, F., Lee, Z., 2013a. Evaluation of chlorophyll-a remote sensing algorithms for an optically complex estuary. Remote Sens. Environ. 129, 75–89. https://doi.org/10.1016/j.rse.2012.11.001

Le, C., Hu, C., English, D., Cannizzaro, J., Chen, Z., Feng, L., Boler, R., Kovach, C., 2013b. Towards a long-term chlorophyll-a data record in a turbid estuary using MODIS observations. Prog. Oceanogr. 109, 90–103. https://doi.org/10.1016/j.pocean.2012.10.002

Lee, Z., Ahn, Y.-H., Mobley, C., Arnone, R., 2010. Removal of surface-reflected light for the measurement of remote-sensing reflectance from an above-surface platform. Opt. Express 18, 26313. https://doi.org/10.1364/OE.18.026313

Li, R.R., Lewis, M.D., Gould, R.W., Lawson, A., Amin, R., Gallegos, S.C., Ladner, S., 2015. Inter-comparison between viirs and MODIS radiances and ocean color data products over the Chesapeake Bay. Remote Sens. 7, 2193–2207. https://doi.org/10.3390/rs70202193

Maritorena, S., D'Andon, O.H.F., Mangin, A., Siegel, D.A., 2010. Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. Remote Sens. Environ. 114, 1791–1804. https://doi.org/10.1016/j.rse.2010.04.002

Meister, G., Franz, B., 2014. Corrections to the MODIS Aqua Calibration Derived From MODIS Aqua Ocean Color Products. IEEE Trans. Geosci. Remote Sens. 52, 6534–6541.

Meister, G., Franz, B.A., Kwiatkowska, E.J., McClain, C.R., 2012. Corrections to the Calibration of MODIS Aqua Ocean Color Bands Derived From SeaWiFS Data. IEEE Trans. Geosci. Remote Sens. 50, 310–319. https://doi.org/10.1109/TGRS.2011.2160552

Mélin, F., Doerffer, R., 2015. Uncertainties in Ocean Colour Remote Sensing [WWW Document]. Rep. IOCCG. URL http://www.ioccg.org/Meetings/IOCCG20/Uncertainties-report_draft_20150225_1.pdf (accessed 9.11.18).

Mélin, F., Zibordi, G., Berthon, J.-F., 2007. Assessment of satellite ocean color products at a coastal site. Remote Sens. Environ. 110, 192–215. https://doi.org/10.1016/j.rse.2007.02.026

Mobley, C.D., Werdell, J., Franz, B., Ahmad, Z., Bailey, S., 2016. Atmospheric Correction for Satellite Ocean Color Radiometry.

Moore, T.S., Campbell, J.W., Feng, H., 2014. Characterizing the uncertainties in spectral remote sensing reflectance for SeaWiFS and MODIS-Aqua based on global in situ matchup data sets. Remote Sens. Environ. 159, 14–27. https://doi.org/10.1016/j.rse.2014.11.025

Mueller, J.L., Davis, C.O., Arnone, R.A., Frouin, R., Carder, K.L., Lee, Z., Steward, R.G., Hooker, S.B., Mobley, C.D., Mclean, S., 2003. Above-water radiance and remote sensing reflectance measurement and analysis protocols, in: Mueller, J.L., Fargion, G.S., Mcclain, C. (Eds.), Ocean Optics Protocols for Satellite Ocean Color Sensor Validation, Revision 4, Volume III: Radiometric Measurements and Data Analysis Protocols. pp. 21–31.

Patt, F.S., Barnes, R.A., Eplee, R.E., Franz, B.A., Robinson, W.D., Feldman, G.C., Bailey, S.W., Gales, J., Werdell, P.J., Wang, M., Frouin, R., Stumpf, R.P., Arnone, R.A., Gould, R. W., J., Martinolich, P.M., Ransibrahmanakul, V., O'Reilly, J.E., Yoder, J.A., 2003. Algorithm Updates for the Fourth SeaWiFS Data Reprocessing, NASA Tech Memo 2003-206892, Volume 22, in: Hooker, S.B., Firestone, E.R. (Eds.), SeaWiFS Postlaunch Technical Report Series.

Qi, L., Lee, Z., Hu, C., Wang, M., 2017. Requirement of minimal signal-to-noise ratios of ocean color sensors and uncertainties of ocean color products. J. Geophys. Res. Ocean. 122, 2595–2611. https://doi.org/10.1002/2016JC012558

Ruddick, K.G., Ovidio, F., Rijkeboer, M., 2000. Atmospheric correction of SeaWiFS imagery for turbid coastal and inland waters. Appl. Opt. 39, 897–912.

Salama, M.S., Su, Z., 2011. Resolving the subscale spatial variability of apparent and inherent optical properties in ocean color match-up sites. IEEE Trans. Geosci. Remote Sens. 49, 2612–2622. https://doi.org/10.1109/TGRS.2011.2104966

Siegel, D.A., Wang, M., Maritorena, S., Robinson, W., 2000. Atmospheric correction of satellite ocean color imagery: the black pixel assumption. Appl. Opt. 39, 3582–3591. https://doi.org/10.1364/AO.39.003582

Sokal, R.R., Rohlf, F.J., 1995. Biometry, Biometry Third edition.

Stumpf, R.P., Arnone, R.A., Gould, R.W., Martinolich, P.M., Ransibrahmanakul, V., 2003. A partially coupled ocean-atmosphere model for retrieval of water-leaving radiance from SeaWiFS in coastal waters, in: Hooker, S.B., Firestone, E.R. (Eds.), SeaWiFS Postlaunch Technical Report Series, Volume 22, Algorithm Updates for the Fourth SeaWiFS Data Reprocessing. pp. 51–59.

Thuillier, G., Hersé, M., Labs, D., Foujols, T., Peetermans, W., Gillotay, D., Simon, P.C., Mandel, H., 2003. The solar spectral irradiance from 200 to 2400 nm as measured by the SOLSPEC spectrometer from the ATLAS and EURECA missions. Sol. Phys. 214, 1–22. https://doi.org/10.1023/A:1024048429145

Toole, D.A., Siegel, D.A., Menzies, D.W., Neumann, M.J., Smith, R.C., 2000. Remote-sensing reflectance determinations in the coastal ocean environment: impact of instrumental characteristics and environmental variability. Appl. Opt. 39, 456. https://doi.org/10.1364/AO.39.000456

Uprety, S., Cao, C., Xiong, X., Blonski, S., Wu, A., Shao, X., 2013. Radiometric intercomparison between suomi-NPP VIIRS and aqua MODIS reflective solar bands using simultaneous nadir overpass in the low latitudes. J. Atmos. Ocean. Technol. 30, 2720–2736. https://doi.org/10.1175/JTECH-D-13-00071.1

Vandermeulen, R.A., Arnone, R., Ladner, S., Martinolich, P., 2015. Enhanced satellite remote sensing of coastal waters using spatially improved bio-optical products from SNPP-VIIRS. Remote Sens. Environ. 165, 53–63. https://doi.org/10.1016/j.rse.2015.04.026

Wang, M., Jiang, L., Liu, X., Son, S., Sun, J., Shi, W., Tan, L., Mikelsons, K., Wang, X., Lance, V., 2016. VIIRS ocean color products: A progress update. Int. Geosci. Remote Sens. Symp. 2016–Novem, 5848–5851. https://doi.org/10.1109/IGARSS.2016.7730528

Wang, M., Liu, X., Jiang, L., Son, S., 2017. Visible Infrared Imaging Radiometer Suite (VIIRS) Ocean Color Produts Algorithm Theoretical Basis Document Version 1.0.

Wang, M., Liu, X., Jiang, L., Son, S., Sun, J., Shi, W., Tan, L., Naik, P., Mikelsons, K., Wang, X., Lance, V., 2015. VIIRS ocean color research and applications. Int. Geosci. Remote Sens. Symp. 2015–Novem, 2911–2914. https://doi.org/10.1109/IGARSS.2015.7326424

Wang, M., Liu, X., Jiang, L., Son, S., Sun, J., Shi, W., Tan, L., Naik, P., Mikelsons, K., Wang, X., Lance, V., 2014. Evaluation of VIIRS ocean color products 92610E. https://doi.org/10.1117/12.2069251

Wang, M., Liu, X., Tan, L., Jiang, L., Son, S., Shi, W., Rausch, K., Voss, K., 2013. Impacts of VIIRS SDR performance on ocean color products. J. Geophys. Res. Atmos. 118, 10347–10360.

https://doi.org/10.1002/jgrd.50793

Wang, M., Shi, W., 2007. The NIR-SWIR combined atmospheric correction approach for MODIS ocean color data processing. J. Geophys. Res. 15, 15722–15733. https://doi.org/10.1029/2004JD004950

Wang, M., Shi, W., Jiang, L., 2012. Atmospheric correction using near-infrared bands for satellite ocean color data processing in the turbid western Pacific region. Opt. Express 20, 741. https://doi.org/10.1364/OE.20.000741

Weeks, S., Werdell, P., Schaffelke, B., Canto, M., Lee, Z., Wilding, J., Feldman, G., 2012. Satellite-Derived Photic Depth on the Great Barrier Reef: Spatio-Temporal Patterns of Water Clarity. Remote Sens. 4, 3781–3795. https://doi.org/10.3390/rs4123781

Wei, J., Lee, Z., Shang, S., 2016. A system to measure the data quality of spectral remote-sensing reflectance of aquatic environments. J. Geophys. Res. Ocean. 121, 8189–8207. https://doi.org/10.1002/2016JC012126

Werdell, P.J., Bailey, S.W., Franz, B. a., Harding Jr., L.W., Feldman, G.C., McClain, C.R., 2009. Regional and seasonal variability of chlorophyll-a in Chesapeake Bay as observed by SeaWiFS and MODIS-Aqua. Remote Sens. Environ. 113, 1319–1330. https://doi.org/10.1016/j.rse.2009.02.012

Zibordi, G., 2016. Experimental evaluation of theoretical sea surface reflectance factors relevant to above- water radiometry 24, 838–850. https://doi.org/10.1364/OE.24.00A446

Zibordi, G., Berthon, J.-F., Mélin, F., D’Alimonte, D., Kaitala, S., 2009. 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. https://doi.org/10.1016/j.rse.2009.07.013