Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-$a$ and turbidity

Catherine Kuhn$^{a,*}$, Aline de Matos Valerio$^b$, Nick Ward$^{c,d}$, Luke Loken$^{e,f}$, Henrique Oliveira Sawakuchi$^{g,h}$, Milton Kampel$^b$, Jeffrey Richey$^i$, Philipp Stadler$^j$, John Crawford$^k$, Rob Striegl$^k$, Eric Vermote$^l$, Nima Pahlevan$^{l,m}$, David Butman$^{a,n}$

$^a$ Department of Forestry and Environmental Science, University of Washington, Seattle, WA, USA
$^b$ Divisão de Sensoriamento Remoto, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, São Paulo, Brazil
$^c$ Marine Sciences Laboratory, Pacific Northwest National Laboratory, Sequim, WA, USA
$^d$ School of Oceanography, University of Washington, Seattle, WA, USA
$^e$ U.S. Geological Survey, Wisconsin Water Science Center, Middleton, WI, USA
$^f$ Center for Limnology, University of Wisconsin-Madison, Madison, WI, USA
$^g$ Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, São Paulo, Brazil
$^h$ Department of Ecology and Environmental Science, Umeå University, KBC-huset, Linnéaus väg 6, 901 87 Umeå, Sweden
$^i$ School of Oceanography, University of Washington, Seattle, WA, USA
$^j$ TU Wien, Centre for Water Resource Systems, Karlplatz 13, A-1040 Vienna, Austria
$^k$ U.S. Geological Survey, Water Mission Area, USGS, Boulder, USA
$^l$ NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA
$^m$ Science Systems and Applications Inc., 10210 Greenbelt Road, Lanham, MD, USA
$^n$ Department of Civil and Environment Engineering, University of Washington, Seattle, WA, USA

ARTICLE INFO

Keywords:
Atmospheric correction
Ocean color
Landsat-8
Sentinel-2
Mississippi River
Amazon River
Columbia River
Turbid waters
Inland waters
Remote Sensing

ABSTRACT

Rivers and other freshwater systems play a crucial role in ecosystems, industry, transportation and agriculture. Despite the >40 years of inland water observations made possible by optical remote sensing, a standardized reflectance product for inland waters is yet forthcoming. The aim of this work is to compare the standard USGS land surface reflectance product to two Landsat-8 and Sentinel-2 aquatic remote sensing reflectance products over the Amazon, Columbia and Mississippi rivers. Landsat-8 reflectance products from all three routines are then evaluated for their comparative performance in retrieving chlorophyll-$a$ and turbidity in reference to shipborne, underway in situ validation measurements. The land surface product shows the best agreement (4% Mean Absolute Percent Difference) with field measurements of radiometry collected on the Amazon River and generates 36% higher reflectance values in the visible bands compared to aquatic methods (ACOLITE and SeaDAS) with larger differences between land and aquatic products observed in Sentinel-2 (0.01 sr$^{-1}$) compared to Landsat-8 (0.001 sr$^{-1}$). Choice of atmospheric correction routine can bias Landsat-8 retrievals of chlorophyll-$a$ and turbidity by as much as 59% and 35% respectively. Using a more restrictive time window for matching in situ and satellite imagery can reduce differences by 5–31% depending on correction technique. This work highlights the challenges of satellite retrievals over rivers and underscores the need for future optical and biogeochemical research aimed at improving our understanding of the absorbing and scattering properties of river water and their relationships to remote sensing reflectance.

1. Introduction

Rivers sustain terrestrial ecosystems and human communities (UN, 2009) yet are being transformed worldwide by anthropogenic pressures (Vörösmarty et al., 2010). Threats include harmful algal blooms, sediment loading, warming and eutrophication (Whitehead et al., 2009; Malmqvist et al., 2008). In terrestrial, ocean, coastal and lake ecosystems, satellites have been increasingly marshalled for ecological monitoring (Smith, 2003; Valero et al., 2017), yet rivers have received relatively little attention in the field of aquatic remote sensing, in part
due to their small spatial scale (< 100 km) and also because of their large dynamic range of optically significant constituents. Phytoplankton, chromophoric dissolved organic matter (CDOM), and non-algal particles can all be present and do not necessarily co-vary. This optical complexity, when combined with rapid changes in river flow and chemistry, results in a challenging observational environment (Hestir et al., 2015).

The dynamic nature of rivers necessitates the ability to evaluate ecosystem characteristics beyond point samplings to understand spatiotemporal variation and monitor long-term changes. For example, products retrieved from satellites such as chlorophyll-a (Chl-a), CDOM, and turbidity have been used for evaluating important processes/factors such as sediment and DOM transport (Griffin et al., 2018; Saraceno et al., 2009), total suspended matter (Shi et al., 2015) ecosystem productivity (Carr et al., 2006; Saba et al., 2011), and even greenhouse gas fluxes (Pay and McKinley, 2017) in the case of marine and lake settings, but these approaches are seldom applied to rivers.

Both atmospheric correction and bio-optical models are key processing steps to water color remote sensing (Ruddick et al., 2000). During atmospheric correction, remote sensing reflectance \( R_{\text{rs}}(\lambda); \text{sr}^{-1} \), defined as the ratio of water-leaving radiation below the water surface to downwelling irradiance above the water surface (Mobley, 1999), is retrieved from at-sensor measurements by correcting for surface effects and atmospheric influences. This process is paramount for the robust retrieval of chlorophyll-a, CDOM and sediment at regional and global scales (McCain et al., 2006) yet remains one of the largest sources of error and foremost challenges in aquatic remote sensing (Mobley et al., 2016; Mouw et al., 2015).

While sensor-specific atmospheric correction routines for land and ocean applications have existed for decades (Gordon and Wang, 1994), inland water techniques are still emerging. Atmospheric correction over water requires greater precision than over land because 70 to 90% of the top of the atmosphere signal over water is known to be from atmospheric effects and sun and sky glint from the water surface (Wang, 2010). In light of this challenge, the Joint National Aeronautics and Space Administration (NASA) and United States Geologic Survey (USGS) Landsat-8 Surface Reflectance Code (LaSRC), while primarily designed for terrestrial applications, has been modified to include a routine over surface waters (Vermote et al., 2016). At coarser resolutions, significant progress has been made in developing specialized corrections for coastal ocean applications. For example, NASA’s Ocean Color Biology Processing Group (OBPG) regularly generates atmospherically-corrected Level-2 products from ocean color sensors including the Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) (Barnes and Hu, 2016; O’Reilly et al., 1998; Wang and Shi, 2007) using the SeaWiFS Data Analysis System (SeaDAS).

The large pixel size (> 250 m) of standard ocean color products limits their use at smaller spatial scales relevant to inland waters (Mouw et al., 2015). While not designed for ocean color applications, moderate-resolution missions (10–100 m) such as the Operational Land Imager (OLI) on board NASA’s Landsat-8 (L8) and the Multispectral Instrument (MSI) on board the European Space (ESA) Agency’s Sentinel-2 (S2), abbreviated hereafter as L8 and S2, provide an improvement over prior generations of moderate-resolution sensors used for monitoring near-surface water constituents.

For example, Franz et al. (2015) demonstrated the atmospheric correction of L8 in SeaDAS over the Chesapeake Bay for the retrieval of Chl-a. Pahlevan et al. (2017b) used atmospherically-corrected S2 images to map total suspended sediment in moderately turbid coastal waters. Lymburner et al. (2016) used a land-based atmospheric correction of L8 over Australian lakes and produced reasonable estimates of surface reflectance, which were then successfully used to retrieve total suspended matter (TSM). L8-retrieved TSM values showed a strong correlation with in situ data, but the land-based atmospheric correction was thought to introduce a positive bias.

Surface reflectance over complex waters has also been estimated using a software platform called ACOLITE for both L8 (Vanhellemont and Ruddick, 2014) and S2 (Vanhellemont and Ruddick, 2016) for turbid coastal waters. Thus, the number of specialized atmospheric correction routines has increased substantially (Dörnhöfer and Oppelt, 2016) since the first ocean color correction was developed (Gordon, 1978) and performance comparisons over land are underway (Doxani et al., 2018). Despite this, few papers compare their merits (Hadjimitsis et al., 2004; Doxani et al., 2018; Martins et al., 2017) and standardized surface reflectance products are not yet available for inland water applications.

A second major barrier impeding our ability to evaluate water color retrievals is the limited number of concurrent, in situ, validation measurements (O’Reilly et al., 1998). Multi-parameter ship-borne sensing platforms are increasingly used in coastal and ocean settings to retrieve bio-optical properties (Werdell et al., 2013; Aiken and Hooker, 1997; Brewin et al., 2016; Matsuoka et al., 2016; Dall’Olmo et al., 2009; Slade et al., 2010; Fichot et al., 2015). These underway measurement systems allow for large-scale validation of satellite-based products. Underway measurements have expanded the possible scale of river observations, resulting in changes in our understanding of riverine carbon dynamics at the global scale (Sawakuchi et al., 2017). Conversely, in inland waters, especially rivers, the use of underway flow-through systems for calibration and validation of satellite products is still incipient.

In light of these advances, here we evaluate the potential of three atmospheric corrections to L8 and S2, validated with high-resolution underway measurements of river constituents, to develop space-based retrievals of Chl-a and turbidity in large rivers. Our primary objective is to examine the influence of three atmospheric correction routines (LaSRC, SeaDAS, ACOLITE) on estimating these parameters.

To achieve this goal, we begin by analyzing surface reflectance estimates, abbreviated hereafter as \( R_{\text{rs}} \), for readability, from three atmospheric correction routines and evaluating the differences in remote sensing reflectance spectra. We compare the reflectance spectra generated by this analysis to in-river field measurements as well as to satellite-derived spectra from the literature. We then examine the influence of these spectral differences on the performance of standard bio-optical algorithms for Chl-a and turbidity. Our goal is to examine how current remote sensing approaches may or may not suffice for estimating water constituents across a range of river conditions.

2. Site description

Continuous, underway data coincident to satellite overpasses were collected from the main stem of the Amazon, Columbia and Mississippi rivers (Table 1, Fig. 1). The rivers represent a natural gradient in water color from very clear to very turbid. The cruises are part of an ongoing effort (Stadler et al., 2018; Ward et al., 2018; Crawford et al., 2016, 2017) to characterize carbon cycling in major rivers using large-scale underway sampling transects.

3. Methods

The overarching methodological framework (Fig. 2d) for this study was to match satellite and in situ measurements for three optically diverse river systems ranging from the shallow and productive waters of the Mississippi (Fig. 2a) to the relatively clearer, deeper waters (Fig. 2b) of the Columbia river with the two cruises during Amazon High Water (HW) and Low Water ( LW) acting as a very turbid endpoint (Fig. 2c).

3.1. Underway river datasets for algorithm evaluation

Custom, flow-through systems delivered river water on board from an average depth of 0.2 m where it passed through a series of optical sensors configured to log simultaneously with a GPS unit as described in Crawford et al. (2017, 2016); Turner et al. (2016); Ward et al. (2018).
Optical parameters measured included turbidity (FNU) (ISO-7027 method, 860 ± 15 nm excitation, 90° scattering) and Chl-a fluorescence (mg m⁻³, excitation 470 nm ± 15 nm and emissions ± 685 20 nm), as measured by a Yellow Springs Instrument (YSI) EXO2 sonde. Fluorometers are well-suited for inline, large-scale mapping because of small sensor size and lower power requirements. Chlorophyll-a fluorescence (Chl-a) is commonly held to be a proxy for chlorophyll-a concentration. Known biases associated with fluorometric Chl-a include interference from other bio-optical components like non-algal particles as well as variability in phytoplankton physiology and species composition which can cause changes in the fluorescence to Chl-a ratio (Roesler et al., 2017; Mouw et al., 2013; Dierssen, 2010). Unfortunately, HPLC pigment data are not available so natural variations between Chl-a concentrations and fluorescence for these large rivers remain to be studied. Turbidity, in its formal optical definition, refers to the amount of attenuation and backscattering of light due to suspended solids and dissolved load. Data were logged per second at boat speeds ranging from 15 to 40 km h⁻¹ representing roughly a point every 4 to 11 m, or two to seven measurements per pixel depending on sensor. The Amazon cruise data were collected at one minute intervals; the Mississippi and Columbia datasets were converted from 1 s intervals to 1 min median bins (Dall’Olmo et al., 2009) to match the GPS unit logging interval.

### 3.2. In situ hyperspectral radiometry

While biogeochemical data exists for rivers worldwide, parallel radiometric measurements over rivers are very rare. As such, in situ reflectance measurements were only possible during the Amazon field campaigns. In situ radiometric data were collected at stationary sampling sites during the Amazon cruises (Valerio et al., 2017) to provide a more quantitative assessment of atmospheric correction techniques.

Above-water hyperspectral radiometry data were collected using a portable hyperspectral radiometer FieldSpec® (ASD Inc.) which collects radiance (L, μW m⁻² sr⁻¹) in the range of 350 to 1100 nm (bandwidth 1 nm) and a field-of-view of 25°. The acquisition geometry followed (Mobley, 1999) recommendations to avoid shadows and sun and sky glint contamination. Total water-leaving radiance (L(w)), sky radiance (L(sky)) and the radiance from a white Spectralon reference panel (L(g)) were consecutively measured 6 to 10 times in the same sequence using a fixed geometry, averaged and resampled to match satellite sensor bandwidths. During days with sparse clouds, the radiometer integration time was adjusted every time the sunlight condition changed and new measurements were made. The L(g) was used to estimate the downwelling irradiance (E(d)) (Eq. (1)):

\[ E_d(\lambda) = L_g(\lambda)f_c \]

where \( f_c \) is a correction factor estimated in laboratory by the ratio of a standard Spectralon reference that remains in the laboratory to the

### Table 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>River and region</th>
<th>Period</th>
<th>Distance</th>
<th>Variable</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-LW</td>
<td>Amazon Low Water, Lower Amazon, Brazil, South America</td>
<td>2016-11-04–2016-11-09</td>
<td>−526 km</td>
<td>Chl-a, Turb</td>
<td>3838</td>
</tr>
<tr>
<td>A-HW</td>
<td>Amazon High Water, Lower Amazon, Brazil, South America</td>
<td>2017-04-26–2017-05-02</td>
<td>−526 km</td>
<td>Chl-a, Turb</td>
<td>4299</td>
</tr>
<tr>
<td>LCR</td>
<td>Lower Columbia, USA, North America</td>
<td>2016-07-12–2016-07-18</td>
<td>−568 km</td>
<td>Chl-a, Turb</td>
<td>1436</td>
</tr>
<tr>
<td>UMR</td>
<td>Upper Mississippi, USA, North America</td>
<td>2015-08-01–2015-08-13</td>
<td>1385 km</td>
<td>Chl-a, Turb</td>
<td>4170</td>
</tr>
</tbody>
</table>

*a* Chlorophyll-a concentration (mg/m³).

*b* Turbidity (FNU).

**Fig. 1.** Field site locations and turbidity gradients. Map showing location of three large river basins: Amazon, Columbia and Mississippi (grey) and the field transects (red) (a). Overall 13,000 measurements of turbidity (FNU) were recorded during the four cruises, revealing spatial gradients in water clarity. (b) Upper Mississippi River transect August 2015 (c) Lower Amazon River cruise on November 2016 (d) Lower Columbia River Cruise in July 2016 and the (e) Lower Amazon River cruise on April 2017. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Spectranol panel used at the fieldwork. The remote sensing reflectance ($R_{rs}$) can be computed according to Eq. (2):

$$R_{rs} = \frac{L_u}{E_d} = \frac{L_u - p_{air-river} \cdot L_{sky}}{E_d}$$

(2)

where $L_u$ is the upwelling radiance that reaches the sensor and $p_{air-river}$ is a sun and sky glint correction coefficient at the air-sea interface. There are several methods in the literature to correct the optical signal from sky glint interference. The sun glint interference at this point should be minimum after redundant measurements made following viewing geometry proposed by Mobley (1999). The residual sun glint plus sky glint suggested by Ruddick et al. (2005) were corrected in the present study using the approach of Ruddick et al. (2006) for turbid to highly turbid waters. The $p_{air-river}$, is a function of wind and cloud cover (Eqs. (3a) and (3b)).

$$\frac{L_{sky}(\lambda = 750)}{E_d(\lambda = 750)} \geq 0.05 \rightarrow p_{air-river} = 0.0256$$

(3a)

or

$$\frac{L_{sky}(\lambda = 750)}{E_d(\lambda = 750)} < 0.05 \rightarrow p_{air-river} = 0.0256 + 0.00039W + 0.000034W^2$$

(3b)

where $W$ is the wind (m s$^{-1}$) measured concomitantly with the radiometric measurements. The residual glint and white offset correction was not performed for the spectra dataset. This correction usually is based on NIR spectrum (e.g. $R_{rs}$ (780) and $R_{rs}$ (780), Ruddick et al., 2005) and assumes that its shape is largely determined by pure water absorption. In very turbid waters with high NIR reflectance like the Amazon River (Fig. 1d–e), the NIR variability is not linear (Wang et al., 2012; Goyens et al., 2013). After $R_{rs}$ was calculated, the coefficient of variation ($CV = \text{standard deviation/mean} \cdot 100$) of the $R_{rs}$ spectra replicates was computed for each station. Only the spectrum (considering the interval of 400–840 nm) with $CV$ close or lower than 10% between the replicates was kept and averaged to get the final spectrum utilized in this study as a representative of $R_{rs}$ at each station.

### 3.3. Satellite data

L8, launched as a collaboration between the United States Geologic Survey (USGS) and National Aeronautics and Space Administration (NASA) on February 11, 2013, carries onboard the OLI pushbroom multispectral radiometer. While the 16-day revisit period is similar to previous Landsat missions, L8’s OLI sensor possesses several major enhancements including an additional band for coastal and aerosol applications (443 nm) and cirrus clouds (1374 nm). Designed to provide continuity with Landsat (Irons et al., 2012), the MultiSpectral Instrument (MSI) onboard ESA’s S2 has a 5-day revisit time and 13 spectral bands in the visible to near infrared (443 nm–2190 nm) (Drusch et al., 2012) and is also available through a variety of web providers including the USGS Earth Explorer site and Google Earth Engine (Gorelick et al., 2017). Compared to L8, the S2 sensor features a higher spatial resolution (10, 20 and 60 m), shorter revisit period (5 days) and three additional bands in the near-infrared (703, 740, and 783 nm) region. Both instruments (Table S1) are quantized at 12-bits, and have much higher signal to noise ratios compared to previous Landsat missions and less frequent saturation over highly reflective targets (Pahlevan et al., 2014; Roy et al., 2014). While existing ocean color missions like the MODerate Resolution Imaging Spectroradiometer (MODIS) also have high radiometric resolution (16 bits) and more frequent revisit times (1–2 days), the finer spatial (10 to 60 m) resolution of L8 and S2 is their major advantage over coarser ocean color sensors like Ocean and Land Color Instrument on board Sentinel-3 (300 m) and MODIS (250–1000 m).

Collection 1 Level 1 L8 and Level 1C S2 Top of the Atmosphere (TOA) data acquired during each cruise were identified by filtering to each region (Fig. 1) and cruise duration (Table 1) in Google Earth Engine (Gorelick et al., 2017). Over 140 total images were acquired over the rivers during the cruises. Of these, 121 were found unsuitable because of significant cloud cover (> 90%) or lack of overlap with the cruise transects in space or time (within ≤24 h). The final 19 TOA images (Table 2) were downloaded from the USGS Earth Explorer (https://earthexplorer.usgs.gov/) in January 2018 and defined in their native World Geodetic System 84 (WGS84) datum and Universal Transverse Mercator (UTM) projection for use as inputs to the atmospheric correction routines. Note that the fusion of these two sensors into continuous time series is possible but outside the scope of this study.

#### 3.4. Atmospheric correction techniques

In this study we test three atmospheric corrections using Level 1 TOA data as an input: the Landsat Surface Reflectance Code (LaSRC); the 12gen processor in SeaDAS; and a third method called ACOLITE (Vanhellemont and Ruddick, 2015). The first two are used by the USGS and NASA’s OBPG to create land and ocean products. The last method is an open-source software processor developed at the Royal Belgian Institute for Meteorology.
Institute of Natural Sciences (Vanhellemont and Ruddick, 2014, 2015, 2016). The overall differences between processors are described here and in Table S2.

The Landsat Surface Reflectance Code (LaSRC), was originally developed at NASA Goddard Flight Center for terrestrial applications. LaSRC uses the Second Simulation of the Satellite Signal in the Solar Spectrum (6SV) model and auxiliary data from MODIS climate grids to estimate aerosols, air temperature, water vapor (MOD09CMG), and ozone (MOD09CMA) (Vermote et al., 2016). LaSRC uses the coastal blue band (443–450 nm), where aerosols typically have a strong signal, in combination with the red band for retrieving aerosols (Roy et al., 2014). The resulting USGS Collection 1 L8 reflectance product is widely available, including from the USGS Earth Explorer (Woodcock et al., 2008) and cloud providers such as Google Earth Engine, Amazon Web Services and Planet Labs. For this application, L8 and S2 images were processed in their native resolution (see Table S1) using LaSRC (v 3.5.5) at NASA Goddard Space Flight Center’s Terrestrial Information Systems Laboratory.

NASA’s OBPG has, in parallel, developed an approach specifically tuned for water and distributed in SeaDAS (version 7.0). SeaDAS’s level 2 (L2gen, v9.1.0) processor was used to produce remote sensing reflectance products for L8 (Franz et al., 2015) and S2 (Pahlevan et al., 2017a). To accommodate complex coastal waters, ocean color processor now incorporates an iterative, NIR-based correction (Bailey et al., 2010), which has been shown to reduce negative SeaWiFS retrievals in combination with the blue (412–490 nm) by 40–100% for low to moderately turbid waters (Pahlevan et al., 2017c) more successfully than a combined NIR-SWIR method with an ε fixed per-scene to reduce pixel-level noise (Dogliotti et al., 2011). Clouds, land, glint and human infrastructure were masked using a SWIR (~1609 nm) threshold suggested by Vanhellemont and Ruddick (2015) and Wang and Shi (2006) in which Rayleigh-corrected reflectances above > 0.0215 (dimensionless) are considered to be non-water. Note that since the preparation of this manuscript, a new version of ACOLITE (Python 20180925.0) has been released that selects the atmospheric correction band from any part of the spectrum based on the resulting path radiance (Vanhellemont and Ruddick, 2018). This dark spectrum fitting method selects the “black pixel” in any wavelength, including the visible, and from whichever target, including building and tree shadows, that is the darkest. These changes are expected to alleviate some of the issues associated with using the SWIR band for atmospheric correction over water in the presence of adjacency effects.

3.5. Water color algorithms

In this study, standard satellite water color algorithms for inland waters were applied to atmospherically corrected data. Testing new approaches to bio-optical models in freshwater is an area of active research (see Supplementary Text 1 for discussion) outside the scope of this study. Here we selected standard, cross-platform approaches. Turbidity was estimated using a semi-empirical red band algorithm with L8’s 655 nm band and S2’s 665 nm band (Dogliotti et al., 2015; Nechad et al., 2009). Chl-a was estimated using the widely-tested OC3 algorithm. A complete description is given in Supplementary Text 3. Intended for concentrations > 0.2 mg m$^{-3}$, OC3 relates ratios of the maximum of the two blue bands (443 or 490 nm) and green bands (560 nm) to Chl-a with a fourth-order polynomial relationship (O’Reilly et al., 1998). While the red and NIR bands are in special cases used for estimating Chl-a over very turbid waters (Sun et al., 2014; Le et al., 2011; Dall’Olmo et al., 2009) to avoid overestimating chlorophyll-a, we wanted to evaluate the performance of standard approaches over the dynamic range of our river sites as pre-existing products are likely to be of the greatest use to the water management and limnology communities. Atmospherically corrected satellite data from the three approaches were used to produce spatially continuous estimates of Chl-a and turbidity over three river systems.

3.6. Satellite data to in situ matchup considerations

Outputs of the 19 images selected in this study (Table 2) from each of the three correction routines were then cloud-optimized, or tiled, for import into Google Cloud Storage (GCS). Satellite to in situ matchups were generated by importing all cloud-optimized imagery from GCS.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>River</th>
<th>Overpass date</th>
<th>Overpass time (UTC)</th>
<th>Time window (± hours)*</th>
<th>Path/row or tile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat-8</td>
<td>Mississippi</td>
<td>8/7/15</td>
<td>16:47:13, 16:46:50</td>
<td>3</td>
<td>025/032, 025/031</td>
</tr>
<tr>
<td></td>
<td>Amazon LW</td>
<td>11/6/16</td>
<td>13:34:59, 13:35:23</td>
<td>3</td>
<td>225/060, 225/059</td>
</tr>
<tr>
<td></td>
<td>Amazon HW</td>
<td>5/1/16</td>
<td>13:23:56, 13:34:20</td>
<td>3</td>
<td>225/060, 225/059</td>
</tr>
<tr>
<td>Sentinel-2</td>
<td>Mississippi</td>
<td>8/7/15</td>
<td>16:42:10</td>
<td>12</td>
<td>T15SYD, T16SBJ</td>
</tr>
<tr>
<td></td>
<td>Columbia</td>
<td>7/14/16</td>
<td>19:04:59, 19:01:37</td>
<td>12</td>
<td>T10TGR, T10TGS, T11TLM, T10TFR, T11TLL</td>
</tr>
<tr>
<td></td>
<td>Amazon LW</td>
<td>11/5/16</td>
<td>13:15:10</td>
<td>3</td>
<td>T22MDE, T22MEE, T22NEF, T22MFE</td>
</tr>
</tbody>
</table>

Table 2: Satellite overpasses from L8 and S2 coincident (± 24 h or fewer) of cruise activities using a more restrictive time window where possible. S2 images that share a time stamp have been repackaged by the USGS Earth Resources Observation and Science Center (EROS) into 100 km × 100 km granules, or tiles, excerpted from the same data set.

* Window of time in hours between closest coincident image acquisition and field measurements.
and field data points into Google Earth Engine (GEE) (Gorelick et al., 2017) for masking and sampling. To avoid the influence of clouds, all images were masked to clear, cloud-free pixels. L8 LaSRC Level 2 products allow users to specify both cloud and cloud shadow-free and water-only pixels using the pixel_qa band (60 m, where pixel_qa = 324), thus masking any pixels flagged with medium or high confidence as land, cloud, or ice (Foga et al., 2017). S2’s quality band (QA60, 60 m) was used to mask clouds (where QA60 = 0). However, this band only indicates the presence or absence of clouds without the additional flags for water presence/absence available from the L8 L2 product so the dilated shoreline mask described below was used to isolate water pixels.

Shoreline effects from mixed pixels and breaking surf at the water’s edge can contaminate dark targets and are a special concern for inland waters (Franz et al., 2015). To reduce this effect and minimize sub-pixel variability, field measurements collected within 3 pixels of the shoreline were discarded using a dilated shoreline mask. Shorelines were estimated using a land mask derived from the Surface Water Occurrence dataset (Pekel et al., 2016) where water occurrence > 90% is classified as water (0) and < 90% is classified as land (1). Each sampling point’s distance to the nearest non-zero land pixel (i.e. the shoreline) was then calculated using a fast distance transform function and points close to shore (< 3 L8 pixels) were excluded, resulting in a dataset representative of open water at a distance > 90 m from the shore. Data from both sensors was masked at 90 m for a conservative shoreline estimate.

Bailey and Werdell (2006) suggest match-ups between satellite and in situ data should ideally be restricted to field measurements collected within a 3-hour window of satellite overpass. However, other authors have demonstrated successful matchups using less restrictive windows of up to 3 days in lakes (Olmanon et al., 2008; Kloiber et al., 2002; Sriwongsitanon et al., 2011; Tebbetts et al., 2013) under stable hydrologic and atmospheric conditions. For S2 and L8, local overpass times were around 11 a.m. ± 0.5 h, which allowed for a time difference of < 3 h in most cases unless otherwise noted. The difference in acquisition versus in situ time is noted for each cruise and sensor (Table 2).

To generate matchups, we sampled a 3 x 3 pixel box centered on each validation point. We required a majority of pixels (n ≥ 5 pixels) inside each 3 x 3 box to be retrieved to ensure sample size homogeneity within each box. Medians, arithmetic means, standard deviations and counts were calculated for each 3 x 3 pixel box for Ramp, Chl-a and turbidity. A final filter included only points for which a valid reflectance was retrieved by all correction routines. This step ensured uncertainties were calculated using the same set of pixels from each technique.

To further constrain absolute accuracy, another set of filters as recommended by Bailey and Werdell (2006) was imposed for turbidity and Chl-a (Fig. S1). As such, negative values and pixels outside of one standard deviation were excluded to reduce the influence of extreme outliers and mixed pixels. As a result of these quality control steps, only high-quality chlorophyll-a and turbidity validation points were used, which restricted the final validation analysis to Columbia and Amazon High Water L8 acquisitions. The resulting data arrays were exported from GEE for statistical analysis and visualization using Python (version 2.7, Python Software Foundation, https://www.python.org/) and R (version 3.4.3, R Core Team, 2017, https://www.r-project.org/) in Jupyter Notebooks (Kluyver et al., 2016).

### 3.7 Evaluation functions

We evaluated algorithm performance following Bailey and Werdell (2006) to facilitate comparison to other studies. Statistics relating any satellite to in situ values included the median satellite to in situ ratio (R), the semi-interquartile range (SIQR) and the root mean square difference (RMSD). These metric, though widely used, should be deployed cautiously and in combination with more robust metrics that allow for non-Gaussian distributions and large dynamic ranges (Seegers et al., 2018). Therefore, we also calculated the median absolute percent difference (MAPD). These metrics are defined as:

\[
R_t = \frac{\text{median}(\frac{X_{\text{mod}}}{X_{\text{obs}}})}{X_{\text{obs}}} 
\]

\[
\text{SIQR} = \frac{Q_3 - Q_1}{2}
\]

\[
\text{RMSD} (\Psi) = \sqrt{\frac{\sum_{i=1}^{n} (X_{\text{mod}} - X_{\text{obs}})^2}{n}}
\]

\[
\text{MAPD} = \text{median} \left( 100 \times \frac{|X_{\text{mod}} - X_{\text{obs}}|}{X_{\text{obs}}} \right)
\]

where \(X_{\text{obs}}\) and \(X_{\text{mod}}\) are the in situ and satellite values for each sample point, \(n\) is the number of samples, and the \(Q_1\) and \(Q_3\) are the 25th and 75th quartiles respectively. The SIQR demonstrates the spread or uncertainty associated with the satellite-retrieved values and the median ratio captures the overall bias. The Slope (S) and Intercept (I) were estimated by applying a reduced major axis (RMA) type II regression model (R package lmfit2 (Legendre, 2014)) to accommodate errors in both the field and satellite measurements (Ricker, 1973).

\[
\text{Slope} = \frac{X_{\text{mod}} - 1}{X_{\text{obs}}}
\]

\[
\text{Intercept} = X_{\text{mod}} - S \times X_{\text{obs}}
\]

### 4. Results and discussion

#### 4.1 Underway data quality control

The compiled dataset resulted in > 13,000 measurements over 31 days and 3000 river kilometers. Measurements were collected over two different hydrologic conditions, Amazon low water (LW) and high water (HW) conditions in a tidally-influenced system and spanned a productivity gradient from the mesotrophic (0.1 < Chl-a ≤ 1 mg m\(^{-3}\)) Columbia to the eutrophic (Chl-a > 75 mg m\(^{-3}\)) Mississippi (Franz et al., 2005). To our knowledge, this is the first use of high resolution, underway data for evaluating estimates of river turbidity and Chl-a from L8. In situ values across these rivers revealed large gradients of turbidity (Figs. 1b-e, 2) at both small (e.g. mainstem versus tributary) and large (e.g. tropical versus temperate coniferous forest biomes) scales.

The dynamic range observed here (Fig. 3) falls within the range measured at 3400 marine stations (0.012-72.12 mg m\(^{-3}\)) as reported in NASA’s bio-Optical Marine Algorithm Data set (NOMAD) (Werdell and Bailey, 2005) as well as the range from the current largest Chl-a ocean color validation set (0–100 mg m\(^{-3}\)) described by Seegers et al. (2018). Therefore, our in situ dataset for three major inland rivers falls within the envelope used to develop satellite-based ocean color products, providing evidence that rivers should fall within the dynamic range used to develop current ocean and coastal water techniques. For the rivers (Mississippi and Columbia) where higher-frequency measurements were binned, the median and standard deviation of the 1-minute bins was 44.7 (4.7 μg/L) and 2.19 (0.37 μg/L) for Chl-a and was 26.3 FNU (2.09 FNU) and 1.8 (0.12) FNU for the Mississippi and Columbia River respectively. The relatively low standard deviation suggests spatial stability throughout the main stem.

Of the 13,744 field validation points acquired across all cruises, 6405 spatially coincided with a satellite overpass. Restricting to a 24-hour window reduced the dataset size by 90% (Fig. 4a). In order to ensure the same sample size for each correction technique, we included only pixels that passed all three corrections without masking (Fig. 4b). Pixels flagged under certain criteria are not processed and this varied by
routine.

For example, after masking with the QA band, LaSRC yielded 708 validation points, but ACOLITE and SeaDAS returned valid retrievals for 62% and 37% of those points respectively. This discrepancy is due to the fact that each routine uses a different set of flags, with SeaDAS (32 flags; Hooker et al., 2003) being more detailed in this case than ACOLITE (4 flags).

Less than 1% of samples were excluded due to low coverage (<5 pixels) inside the sampled 3 × 3 pixel window (Fig. 4c). The shoreline mask only excluded 8% of pixels (Fig. 4d), likely because boat surveys generally maintained a steady course in deeper navigation channels and because of the previous coverage filter which reduced sampling window variability.

4.2. Atmospheric correction

Here we will first evaluate the number of failed retrievals, as indicated by negative \( R_{rs} \) values, for each correction technique. We then discuss the differences between the land-based and aquatic correction techniques and quantitatively compare those results to field \( R_{rs} \) measurements. Finally, we evaluate bias, based on differences from field observations, in the resulting chlorophyll-a and turbidity retrievals introduced by each correction.

4.2.1. Negative \( R_{rs} \) retrievals

Analysis of L8 and S2 \( R_{rs} \) observations over rivers are still relatively rare. Fig. 5 shows the range of possible \( R_{rs} \) values in the green band across the entire river surface captured for the images (Table 2) before the quality control described in Section 3.6. This characterization shows the general performance of the atmospheric corrections before any uncertainties introduced by sampling the data at specific points. Negative retrievals can be an important diagnostic in determining the validity of \( R_{rs} \). Across the dataset, 1% of L8 and 2% of S2 pixels showed \( R_{rs} \leq 0 \text{ sr}^{-1} \) in the green bands (Fig. 5); more negative retrievals were observed in the blue bands (2% L8 and 10% S2), which are known to be sensitive to aerosol selection errors.

Negative retrievals can indicate a problem with atmospheric correction. Aquatic correction routines assume any SWIR reflectance to be from aerosols and subtract it as such from the rest of the spectra during the aerosol removal step. This assumption is based on the fact that even very shallow water absorbs longer wavelength light and thus should be dark (\( R_{rs} \sim 0 \text{ sr}^{-1} \)) in the SWIR bands (Wang and Shi, 2007; Shi and Wang, 2009). However, SWIR signals from strong adjacency effects over inland water pixels have been widely documented in most (although not all, see Pahlevan et al., 2019) cases. While SWIR signals over water could be caused by emergent macrophytes (Dogliotti et al., 2018), nearby land pixels contributing SWIR signals are more likely the cause (Richter et al., 2006; Bulgarelli and Zibordi, 2018). For example, 96% of negative retrievals were from the Columbia (Fig. 5a, b), whose shorelines feature steep and bright cliffs. The Amazon LW and Mississippi images had fewer failed returns (<1%) (Fig. 5c, d, e, f, g) although adjacency effects are present in the TOA spectra of all rivers in this study.

The negative retrieval rate was twice as high for S2 (5%) than for L8 (2%) in the visible to near-infrared bands (~400–865 nm). In addition to adjacency effects, another contributing factor could be S2's higher spatial resolution. Smaller pixels resolve smaller wave features (Kay et al., 2009), amplifying the confounding effects of sun glint. ACOLITE produced the most negative blue band retrievals (14%) relative to SeaDAS (8%) and LaSRC (6%).
4.2.2. Land-based versus aquatic corrections

After removal of failed retrievals, the resulting spectral plots derived from the scenes listed in Table 2 and processed as described in Section 3.6 show the top of the atmosphere (TOA) reflectance (unitless) and $R_{rs}$ (sr$^{-1}$) for the land (LaSRC) and aquatic (SeaDAS and ACOLITE) corrections for L8 (Fig. 6) and S2 (Fig. 7). The terrestrial technique produced higher $R_{rs}$ values than aquatic techniques by 36% MAPD, or 0.008 sr$^{-1}$ (Figs. 6, 7), across the L8 and S2 visible and near-infrared bands.

Clearer waters (Fig. 6e) showed larger departures between aquatic and terrestrial L8 techniques (0.004 sr$^{-1}$) in contrast to the turbid Amazon where differences narrowed to 0.001 sr$^{-1}$ (Fig. 6h) across bands. During the Amazon LW (Fig. 6g), aquatic and terrestrial methods overlapped, with notably high SeaDAS variability (IQR = 0.01 sr$^{-1}$) in the 430–512 nm range likely resulting from poor image quality. During the more turbid Amazon HW cruise, techniques converged in the NIR where differences between LaSRC and aquatic techniques closed to 0.0002 sr$^{-1}$.

Larger differences were observed in S2. Terrestrial and aquatic techniques diverged by 62%, yielding differences in $R_{rs}$ an order of magnitude higher for S2 (0.004 sr$^{-1}$) than for L8 (0.003 sr$^{-1}$). By processor, LaSRC L8 $R_{rs}$ was on average 0.0027 and 0.0031 sr$^{-1}$ greater than ACOLITE and SeaDAS respectively. For S2 the difference from LaSRC was larger for ACOLITE (0.02 sr$^{-1}$) but not SeaDAS (0.002 sr$^{-1}$).

Differences in $R_{rs}$ result from differences in processor assumptions and correction bands. LaSRC estimates aerosols over land using the blue and red bands and extrapolates over water. This assumption would not be appropriate over open ocean waters distant from shore but could be reasonable across the shorter spatial scales relevant to inland waters. In contrast, aquatic techniques assume water absorbs NIR-SWIR strongly and therefore any SWIR reflectance is assigned to aerosols. However, as noted in Section 4.2.1, NIR remote sensing reflectance values $> 0$ sr$^{-1}$ at the water surface were observed in all rivers (Figs. 6a–d, 7a–c), likely stemming from adjacency effects or sun and sky glint (Bulgarelli and Zibordi, 2018). Thus, while the SWIR-based black pixel assumption might be reasonable even over turbid waters, it becomes problematic for systems with strong adjacency effects such as inland waters. For waterbodies with nearby and/or bright shorelines, NIR/SWIR...
adjacency effects may be too strong for the NIR/SWIR atmospheric correction and consequently terrestrial or alternative methods may perform best.

However, considering the significant methodological differences in approaches, this convergence between aquatic and terrestrial techniques is encouraging and suggests choice of atmospheric correction may be less important when using L8 in highly reflective systems. The range of spectral shapes shown here also fall within that observed within an analysis conducted by Spyrakos et al. (2018) of over 250 inland and coastal water $R_{rs}$ spectra. The differences in spectra across optical gradients observed here are similar to those reported by Jackson et al. (2017) in a recent analysis of a large scale in situ $R_{rs}$ and Chl-$a$ dataset (OC-CCI v2.0, Valente et al., 2016).

Users must be warned, however, that small differences in spectra can lead to large differences in the absolute accuracy of satellite-retrieved Chl-$a$ and turbidity values, especially if the biases are spectrally-dependent. For example, band ratio models could be more impacted by differences in spectral shapes than magnitudes because of their dependence on relative contributions from each band. Here, processors resolved similar spectral shapes despite differences in magnitudes.

4.2.3. Aquatic corrections

While the terrestrial and aquatic methods produced spectra of differing magnitudes, the two aquatic corrections agree within 0.0006 sr$^{-1}$ across rivers and sensors. Differences were the smallest (0.001 sr$^{-1}$) in the blue and coastal blue bands for the Columbia (Fig. 6e) and for the green (0.0005 sr$^{-1}$) and NIR bands (0.00004 sr$^{-1}$) on the Amazon HW (Fig. 6h).

The difference between sensors was again pronounced, with a much larger gap between aquatic methods for S2 (0.002 sr$^{-1}$) than for L8 (0.0009 sr$^{-1}$). Zhang et al. (2018) reported S2 produced higher $R_{rs}$ values than L8 over land. These differences are likely related to environmental conditions or ground sampling resolution and not sensor specifications (Pahlevan et al., 2017a, 2017b). Despite difference between sensors, aquatic method disagreements are small relative to the differences between terrestrial and aquatic techniques. The similarity of these estimates likely stems from their shared use of the NIR-SWIR bands for atmospheric correction, in contrast to land-based methods which use different targets (dark dense vegetation), bands (blue and red) and assumptions.

Differences between aquatic methods showed spectral dependence. The S2 Mississippi spectra show a green peak (0.02 sr$^{-1}$) in the LaSRC spectrum but not in ACOLITE (Fig. 7e). The aquatic corrections also vary
in L8’s green band (560 nm) over the Columbia (Fig. 6e), where SeaDAS shows a peak (0.003 sr⁻¹) not observed in the ACOLITE spectra. A similar peak is observed in the S2 SeaDAS data (0.002 sr⁻¹) for the Columbia, despite only moderate in situ Chl-a (0.5–1 mg/m⁻³). Differences in NIR processing could drive these results. SeaDAS uses a modeled NIR value from iterative NIR processing while ACOLITE uses the NIR bands as given. While these spectral differences, as stated before, are likely to impact bio-optical models, the remote sensing reflectance estimated by these two aquatic techniques remain on the same order of magnitude.

4.2.4. Validation of remote sensing reflectance

Radiometric measurements for Rrs validation were collected during Amazon high and low water coincident to L8 acquisitions (Table 2, Fig. 8). Unfortunately, no coincident S2 overpasses occurred. In situ measurements show peaks (0.03–0.04 sr⁻¹) in the red and near-infrared characteristic of highly scattering waters (Fig. 8) and strong absorption in the shorter wavelengths due to organic matter. These spectra show agreement with the shape and range (0–0.02 sr⁻¹) measured over turbid Amazon floodplain lakes (Martins et al., 2017) and coastal waters (0–0.035 sr⁻¹) influenced by the Amazon (Froidelond et al., 2002).

In comparisons to in situ radiometry (Fig. 8), LaSRC had the best agreement with an average MAPD across the spectrum of 17% (low water) and 4% (high water) with a RMSD as low as 0.001 sr⁻¹ during high water. This is approaching the ± 5% uncertainty benchmark often cited as a target for clear water radiances and bottom-of-the-atmosphere reflectance (Drusch et al., 2012; Bailey and Werdell, 2006; Hooker et al., 1992). LaSRC also showed less uncertainty than the aquatic techniques evidenced by a narrower SIQR (Table 3).

Aquatic techniques for the scenes analyzed here showed reasonable spectra with the exception of SeaDAS LW (Fig. 8a, b). Median ratios were 0.82 < Rrs < 1.17 with the SIQR indicating uncertainties of < 0.30 except for SeaDAS (Table 3). No valid pixels coincident with field measurements were retrieved by ACOLITE during low water due to SWIR masking. During Amazon HW where all three corrections were available, ACOLITE MAPD (~8%) fell between LaSRC and SeaDAS. SeaDAS showed the least agreement with in situ measurements (MAPD of ~17–79%). These results fall within the range reported by Wei et al. (2018) for L8 as corrected by both SeaDAS and ACOLITE using the NIR/SWIR method over optically shallow and turbid waters, although ACOLITE shows higher performance here, possibly because this study used a fixed epsilon and a SWIR mask whereas Wei et al. (2018) used a per-pixel epsilon and did not specify use of a SWIR mask. Differences in these processing options could lead to higher MAPDs, especially because of the relatively low signal-to-noise ratios of the SWIR bands. RMSD ranged an order of magnitude (0.001–0.01 sr⁻¹) but across both cruises LaSRC had a lower RMSD than either aquatic technique. For example, LaSRC showed an order of magnitude lower RMSD than SeaDAS during LW and 50% and 75% lower RMSD than ACOLITE and SeaDAS during HW.

During high water, in situ Rrs was underestimated by the aquatic technique and overestimated by the terrestrial technique. During low water the bias varied spectrally. For example, satellite-retrieved Rrs was higher than in situ data in the blue and NIR during Amazon LW (Fig. 8a). Adiabatic effects could explain why the NIR signal observed from space is higher than in situ values (Bulgarelli and Zibordi, 2018). The presence of strong glint could also contribute additional NIR signal as observed by both field and satellite measurements.

Glint, a common issue for L8 and S2, could contribute additional scattered light to the satellite-observed signal and is currently not corrected for over turbid waters by most standard approaches. Glint-removal methods are predominantly developed for clear-water marine systems yet can profoundly influence Rrs estimates (Gilerson et al., 2018). For water retrievals, glint corrections can cause up to 43% MAPD in Rrs significantly impacting the resulting Chl-a retrievals (Garaba et al., 2015). Little research exists on this topic for inland waters, although Overstreet and Legleiter (2017) evaluated glint corrections over shallow rivers and Brando et al. (2016) evaluated glint effects in underway river samples. Harmel et al. (2018) developed a glint correction that reduced bias between in situ and satellite Rrs by 60%. However, neither in situ nor satellite observations over turbid waters are currently glint corrected in standard approaches, representing a major limitation.

In the opposite case, negative bias in the aquatic techniques was observed in comparison to field measurements during the Amazon high water cruise. The underestimation could be due to overcorrection resulting from adjacency effects or other factors such as cloud shadows observed in some areas of the scene and absorbing aerosols. Absorbing aerosols are not accounted for in current aerosol models (Gordon and
Wang, 1994) but are common in coastal zones. Absorbing aerosols are known to vary seasonally over the Amazon river mouth due to agricultural burning (Herman et al., 1997). These factors could all contribute to underestimation in the shorter wavelengths (440 and 480 nm) as observed in the high water cruise (Fig. 8b). It is also important to note the bidirectional reflectance effects, which are not considered or corrected for in these routines, could also add to discrepancies between field and satellite observations.

While the terrestrial approach, LaSRC, showed the closest match with field spectra, radiometry was not available from the Columbia and Mississippi so the relative difference between satellite and in situ \( R_a \) over less turbid waters remains unknown. The larger gap between techniques over the Columbia discussed in Section 4.2.2 suggests that comparisons to field radiometry over clearer inland waters should be a research priority.

### 4.3. Chlorophyll-a and turbidity

We examined the sensitivity and absolute accuracy of satellite-retrieved Chl-a and turbidity to atmospheric correction by comparing satellite and in situ measurements for cruises in which both aquatic techniques were available (see Fig. 8 for WRS Path/Row), which includes the L8 images acquired during the Columbia and Amazon HW cruises.

#### 4.3.1. Chlorophyll-a sensitivity to atmospheric correction

Overall, SeaDAS-derived Chl-a estimates were on average 2.7 times higher than ACOLITE and twice as high as field measurements (Fig. 9). For example, the 0.002 sr \(^{-1} \) and 0.001 sr \(^{-1} \) difference between SeaDAS and ACOLITE in the green and blue bands resulted in a difference of 6 mg m\(^{-3} \) Chl-a. Satellite-derived Chl-a estimates from SeaDAS and LaSRC overestimated Chl-a by ~4.7 mg m\(^{-3} \) and 0.6 mg m\(^{-3} \), with median ratios exceeding 1 (Fig. 9, Table 4).

The absolute percent difference for SeaDAS (MAPD = 59%) was higher than ACOLITE (MAPD = 30%) and LaSRC (MAPD = 32%), with SeaDAS-derived Chl-a in some cases being overestimated by an order of magnitude in the mesotrophic waters of the Columbia River. ACOLITE products underestimated Chl-a by ~1.29 mg m\(^{-3} \) across both rivers but approximated field values more closely than SeaDAS; for the Columbia River ACOLITE and in situ Chl-a distributions were not significantly different (p-value = 0.1).

As all three estimates were made using the same bio-optical algorithm (OC3), the major factor controlling these differences is the reflectance product used as an input and the underlying atmospheric correction processors used to generate those reflectance products. The consistent overestimation by SeaDAS directly results from the differences in blue and green bands (Sections 4.2.2-4.2.3) observed in the reflectance spectra. In this case, higher remote sensing reflectance accuracies over the Amazon resulted in satellite-retrieved Chl-a values close to field measurements.

The differences we observe here in these river systems are equivalent to those observed in other systems. For example, Dörnhöfer et al. (2018) also used a multisensor approach to estimate Chl-a over German lakes resulting in RMSD’s between 3.6 and 19.7 mg m\(^{-3} \) with errors varying between L8 and S2 sensors. Regardless of correction, differences from in situ measurements are lower than the factor of five commonly reported for empirically-based ocean Chl-a satellite retrievals (Dierssen and Karl, 2010) and fall below, with the exception of SeaDAS, the 40% error reported for the widely-used OC4v4 standard Chl-a product (LaLiberté et al., 2018), where errors were also largest in waters with < 0.5 mg m\(^{-3} \) Chl-a. It is important to note that the established OC3 coefficients were calibrated in reference to high performance liquid chromatography pigments (HPLC) whereas our underway validation data is derived from instantaneous fluorometers.

Where matchups were available at the exact moment of the satellite overpass, we conducted a sensitivity analysis to quantify changes in accuracies resulting from using more or less restrictive time windows. Boucher et al. (2016) showed using a time window of 2 instead of 5 days improved agreement between L8-retrieved and in situ Chl-a values in northeastern lakes. In this study, using a time window of ± 3 instead of 24 h reduced differences by 5% (ACOLITE) and 31% (SeaDAS) for Chl-a.

The blue green Chl-a algorithms tested here are designed for systems in which Chl-a is the dominant absorber. Consequently, in optically complex waters combinations of other components can result in a false Chl-a signal. For example, in terrestrially-influenced waters, non-algal particles and CDOM can also absorb in shorter wavelengths, changing the blue-green ratio and therefore leading to Chl-a overestimation. Non-algal particles and CDOM can also contribute fluorescence, resulting in an overestimation of fluorometric Chl-a (Roesler et al., 2017). For turbid systems like the Amazon, CDOM and suspended matter absorb strongly in shorter wavelengths, creating a green peak even in the absence of Chl-a. De Matos Valerio et al. (2018) reported CDOM absorption coefficients at 412 nm between 1.0 and 7.0 m\(^{-1} \) in non-turbid tributaries and the mainstem. The sensitivity of results to correction techniques, sensors and bio-optical model choice shown here indicates that satellite-retrieved Chl-a values should be interpreted with great caution in optically complex waters, especially because the relatively wide L8 bands are less than optimal for detecting aquatic signals in optically complex environments.

#### 4.3.2. Turbidity sensitivity to atmospheric correction

In contrast to Chl-a, for turbidity SeaDAS and ACOLITE show greater agreement with in situ measurements (Fig. 10). Differences between satellite-retrieved and measured values range from under-estimates of 13 FNU (SeaDAS) to overestimates of 12 FNU (LaSRC). Over the Columbia ACOLITE and SeaDAS-derived turbidity were statistically similar to in situ measurements (p-value = 0.02) (Fig. 10a).
LaSRC had a larger mean absolute difference to in situ measurements (35%) than ACOLITE (3%) or SeaDAS (13%) (Table 4) with larger differences observed at higher turbidities (Fig. 10b).

Over the Amazon, the range of satellite-retrieved turbidity (mean = 47–62 FNU) bracketed in situ values (mean = 50 FNU). Surprisingly, however, the $R_a$ spectra with the lowest MAPD across the spectral for Amazon HW did not result in the most accurate turbidity estimate. LaSRC over-predicted turbidity across both rivers which follows from high red band peaks (Fig. 6). Turbidity derived from in situ radiometry (636–673 nm) was also overestimated (70 FNU) in reference to direct measurements (50 FNU). Thus both satellite and in situ reflectance values overestimate turbidity. This could result from additional contributions in the red from surface glint that could be addressed in future studies through the use of in-water radiometry or could also potentially be remedied by revisiting the turbidity algorithm calibration coefficients or using longer wavelengths (Dogliotti et al., 2015; Novoa et al., 2017).

The turbidity algorithm used here is directly comparable across platforms and thus differences can be mainly attributable to the atmospheric correction and performance of the bio-optical algorithm across a wide dynamic range. Inconsistencies in radiometric calibration between the two sensors are expected to contribute to < 6% of the total uncertainty (Pahlevan et al., 2019). The wide range of turbidities observed across inland waters in this study suggests a blended approach that utilizes different algorithms at different reflectance ranges such as used for Chl-a (Hu et al., 2012) may be necessary to map turbidity across the entire dynamic range.

While turbidity is not a direct measure of a biogeochemical quantity (Boss et al., 2009), it’s relative ease of measurement, including through crowd-sourced smartphone apps (Leeuw and Boss, 2018) and its usefulness for water quality monitoring (Nechad et al., 2009; Dogliotti et al., 2015) makes it a practical candidate for satellite remote sensing of aquatic optical conditions. These results show that while absolute accuracies of retrievals still require improvement, bio-optical algorithms are broadly able to discriminate between rivers, provided rivers span a large optical gradient.

5. Summary & further work

This study highlights innovative underway field techniques used in combination with L8 and S2 satellite imagery to identify uncertainties in river remote sensing. We show that while all three corrections result in spectra on the same order of magnitude in most cases, the terrestrial atmospheric correction method produces a 36% (0.008 sr$^{-1}$) greater $R_a$ than aquatic techniques. The two aquatic approaches agree within 0.0006 sr$^{-1}$ but that varied by band, which in turn influences satellite-derived Chl-a and turbidity estimates. Where radiometric data were available, the standard land surface reflectance product had the best performance, achieving mean absolute differences as low as 4% relative to field measurements in turbid waters. When combined with bio-optical models, these $R_a$ estimates can be useful for examining broad spatial gradients of Chl-a and turbidity.

However, we strongly advise the cautious interpretation of these results because of uncertainties inherent to water color remote sensing that have yet to be resolved. Specifically, we advise future work on river remote sensing should occur in four major areas, especially as the application of remote sensing for water resource management applications is an increasing priority (McClain and Meister, 2012).

First, our findings show terrestrial correction techniques were able to resolve $R_a$ over the Amazon River comparably to aquatic methods, likely because adjacency effects undermine SWIR-based aquatic correction approaches and despite the fact that terrestrial techniques do not correct for glint. To test this interaction, radiometric measurements over a range of river conditions is needed in addition to research examining adjacency effects. For example, De Keukelaere et al. (2018) show improved reflectance results using an L8/S2 atmospheric correction framework (iCOR) designed to work over inland waters with an explicit adjacency correction (Sterckx et al., 2015). More radiometric validation data, especially over oligotrophic systems, is a major research need. Such measurements are required to constrain estimates of $R_a$ and to develop retrieval algorithms with sufficient absolute accuracy (Werdell et al., 2018).

Next, the assumptions and limitations of standard bio-optical retrieval methods need to be systematically re-evaluated in the context of rivers. While algorithms can resolve large differences in signals between systems, uncertainties in their quantitative, absolute accuracy must not be ignored (Pahlevan et al., 2016), especially because the wide bandwidths of the L8 and S2 sensors make them less than optimal for detecting aquatic signals in optically complex environments. A need exists to compare algorithm performance, especially across missions, to achieve product continuity (Mouw et al., 2015) and advance the field of inland water remote sensing. A first step is the development of sensor-agnostic, open source bio-optical models that can take outputs from a range of correction processors. This will accelerate testing over a wider range of sensors and water types.

Global measurements of river inherent optical properties are needed to inform a process-based understanding of the relationship between river biogeochemistry, optical conditions and remote sensing reflectance. More geographically diverse radiometric and inherent optical property (e.g. particle backscattering coefficients, absorption coefficients, diffuse attenuation) measurements are required. Fully understanding river productivity requires field measurements of parallel optical and biogeochemical properties. For example, although Chl-a is generally low in the Amazon River mainstem, evidence from dissolved oxygen stable isotopes suggest that primary production may still be occurring at up to 50% the rate of respiration (Gagne-Maynard et al., 2017). Establishing turbidity, CDOM and non-algal particle concentration thresholds for the use of blue/green chlorophyll-a algorithms could prevent the masking of chlorophyll-a by sediments and the over-estimation of chlorophyll-a caused by the presence of non-algal particles and CDOM. Spryakos et al. (2018) has made progress in classifying inland waters by their reflectance, which could be one approach to developing flexible bio-optical retrieval algorithms such as available for the open oceans. Future research is required to determine the dynamic range of river optical properties, their relationship to biogeochemistry, and their influence on remote sensing reflectance.

Finally, in addition to uncertainty from retrieval algorithms, a fourth major difficulty is establishing time benchmarks between field and satellite measurements in rivers, especially those influenced by short duration processes such as tides. We show constraining matchups...


