AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

TECHNICAL REPORTS: DATA

10.1029/2019WR024883

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Key Points:
- AquaSat contains ∼600,000 paired observations of water quality and Landsat reflectance, the largest such matchup data set
- Matchups capture diverse water bodies across the USA for 1984–2019; we see clear water quality/reflectance relationships
- AquaSat and open source code developed here will enable better development of models for remote sensing of water quality

Supporting Information:
- Supporting Information S1
- Supporting Information S2

Correspondence to:
M. R. V. Ross,
matt.ross@colostate.edu

Citation:
Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D. et al. (2019). AquaSat: A data set to enable remote sensing of water quality for inland waters. *Water Resources Research*, *55*, 10,012–10,025. https://doi.org/10.1029/2019WR024883

Received 28 FEB 2019
Accepted 20 SEP 2019
Accepted article online 11 OCT 2019
Published online 21 NOV 2019

AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters

Matthew R. V. Ross[1], Simon N. Topp[2], Alison P. Appling[3], Xiao Yang[2], Catherine Kuhn[4], David Butman[4], Marc Simard[5], and Tamlin M. Pavelsky[2]

[1]Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO, USA, [2]Department of Geological Sciences, University of North Carolina, Chapel Hill, NC, USA, [3]United States Geological Survey, Reston, VA, USA, [4]School of Environmental and Forest Sciences, University of Washington, Seattle, WA, USA, [5]NASA Jet Propulsion Laboratory, Pasadena, CA, USA

Abstract Satellite estimates of inland water quality have the potential to vastly expand our ability to observe and monitor the dynamics of large water bodies. For almost 50 years, we have been able to remotely sense key water quality constituents like total suspended sediment, dissolved organic carbon, chlorophyll a, and Secchi disk depth. Nonetheless, remote sensing of water quality is poorly integrated into inland water sciences, in part due to a lack of publicly available training data and a perception that remote estimates are unreliable. Remote sensing models of water quality can be improved by training and validation on larger data sets of coincident field and satellite observations, here called matchups. To facilitate model development and deeper integration of remote sensing into inland water science, we have built AquaSat, the largest such matchup data set ever assembled. AquaSat contains more than 600,000 matchups, covering 1984–2019, of ground-based total suspended sediment, dissolved organic carbon, chlorophyll a, and SDDSecchi disk depth measurements paired with spectral reflectance from Landsat 5, 7, and 8 collected within ±1 day of each other. To build AquaSat, we developed open source tools in R and Python and applied them to existing public data sets covering the contiguous United States, including the Water Quality Portal, LAGOS-NE, and the Landsat archive. In addition to publishing the data set, we are also publishing our full code architecture to facilitate expanding and improving AquaSat. We anticipate that this work will help make remote sensing of inland water accessible to more hydrologists, ecologists, and limnologists while facilitating novel data-driven approaches to monitoring and understanding critical water resources at large spatiotemporal scales.

1. Introduction

Production and effective dissemination of water quality data is a vital first step toward understanding natural and anthropogenic drivers of aquatic ecosystem change (Srebotnjak et al., 2012). Collecting such valuable data has historically been expensive and time consuming, and it has often proved difficult to maintain analysis-ready and open data sets. In many developed nations, however, data access and interoperability have been actively addressed over the last 10–20 years, leading to the publication and maintenance of large open-access data repositories of water quality measurements (Ballantine & Davies-Colley, 2014; Lack, 2000; Read et al., 2017; Soranno et al., 2017), but over a limited number of water bodies and with clear spatial and temporal biases (Stanley et al., 2019). Furthermore, access to robust historic water quality sampling data remains limited to a few economically developed countries (Sheffield et al., 2018).

With satellite remote sensing, we can augment in situ sampling efforts and provide water quality information in places with little or no data. Since the beginning of the Landsat missions, limnologists, oceanographers, and hydrologists have been interested in developing universal algorithms for extracting water quality information from remotely sensed images (Clarke et al., 1970; Holyer, 1978; Klemas et al., 1973; Maul & Gordon, 1975; Ritchie et al., 1976). From these early efforts, 50 years of work have used spectral information to estimate water quality parameters like total suspended solids (TSS), chlorophyll a (here abbreviated as Chl_a), colored dissolved organic matter (CDOM), and Secchi disk depth (SDD). However, progress toward universal algorithms and unified approaches has been slow for inland waters (including lakes, rivers, and estuaries; Blondeau-Patissier et al., 2014; Bukata, 2013; Gholizadeh et al., 2016; Palmer et al., 2015) especially at the global scale, despite some recent global efforts (Odermatt et al., 2018; Spyrakos et al., 2017) and ample regional efforts (Olmanson et al., 2008; Pavelsky & Smith, 2009; Torbick et al., 2013).

This slow progress for inland water remote sensing contrasts sharply with ocean remote sensing, which benefits from robust, open, and big data sets geared toward pairing both in situ and radiometric observations with satellite data, enabling rapid development of more universally effective algorithms and approaches (Blondeau-Patissier et al., 2014; Bukata, 2013). Ocean remote sensing also benefits from dedicated satellites designed specifically for ocean applications such as remote retrieval of Chl_a, but the spatial resolution of these sensors is too coarse to resolve most inland water bodies. As a result, inland water remote sensing has been limited to satellites built for terrestrial remote sensing (Hestir et al., 2015; Palmer et al., 2015). Methods development for rivers, lakes, and the near-shore environment is further challenged by the greater optical complexity of inland waters, where spectral signatures reflect a mixture of inorganic suspended sediment, organic suspended sediment, algae, dissolved organic matter, and other constituents (Mishra et al., 2017). We think some of these inherent challenges in inland water quality remote sensing can be met at a broad scale with a centralized, public remote sensing data set paired with in situ measurements of water quality (Palmer et al., 2015), building on similar work pursued by the Globolakes and LIMNADES projects (www.globolakes.ac.uk) (Spyrakos et al., 2017), and Diversity II (Odermatt et al., 2018).

In this data paper, we present AquaSat, a merged data set of in situ water quality measurements paired with same-day or ±1-day satellite reflectance, which we call "matchups." Here, matchups refer to reflectance data paired with direct measurements of water quality, which is a variation on the more typcial use of the term matchup, which refers to pairing satellite data with ground-truthed measurements of the exact response satellites are measuring, like pairing satellite reflectance with surface reflectance measurements made on the ground (Loew et al., 2017). Here, we include data for rivers, lakes, and estuaries in the continental United States and Alaska. This is the largest such matchup data set ever assembled for inland waters. To create AquaSat, we use the Landsat archive from 1984–2019, available in its entirety on the Google Earth Engine platform (Gorelick et al., 2017), in combination with data from the Water Quality Portal (WQP; Read et al., 2017) and the LAke multiscaled GeOSpatial and temporal database covering the northeastern United States (LAGOS-NE; Soranno et al., 2017). The WQP data we used covers all of the United STates. Joining these data sets provides us with an unprecedented resource to model, predict, and understand the long-term and large-scale dynamics of variation in TSS, SDD, Chl_a, and dissolved organic carbon (DOC) within inland waters. We also outline and share our approach, code, and intermediate data for bringing the WQP, LAGOS-NE, and Landsat data sets together.

2. Methods

2.1. Parameter Description

We focused on five common water quality parameters often targeted for remote sensing of water quality: TSS, DOC, CDOM, Chl_a, and SDD. These five parameters capture key ecological and physical factors that control water quality, and capabilities to remotely sense each of them have been demonstrated (Mishra et al., 2017)

TSS is a measure of the concentration of solids, both organic and inorganic, in a water column, measured in milligrams per liter. Waters with higher TSS generally scatter more sunlight at all visible and near-infrared wavelengths (Ritchie et al., 1976). Knowing TSS concentrations can provide insight into subsurface light conditions (Julian et al., 2008), erosion conditions (Syvitski & Kettner, 2011), and the hydrologic status of water bodies, where high TSS generally means sediment supply coupled with higher flow velocities (Pavelsky & Smith, 2009; Williams, 1989).

DOC, measured in milligrams per liter, is the broad description for the concentration of organic carbon dissolved in water and can provide insight into light conditions (Vähätalo et al., 2005), heterotrophic energy availability (Robbins et al., 2017), and terrestrial organic matter processing (Williamson et al., 2008). While DOC does not inherently alter the optical properties of water, its colored portion, CDOM, does affect optics and is often correlated with DOC concentration (Bricaud et al., 1981; Griffin et al., 2011). This correlation between CDOM and DOC can break down in places with low DOC concentrations (Griffin et al., 2018) or in areas with high photobleaching of DOC, which alters the DOC/CDOM fractionation (Cory et al., 2015; Spencer et al., 2009).

Chlorophyll a is a photosynthetically active pigment contained in all phytoplankton. Chlorophyll a can be used to detect algae blooms (Kutser, 2004), estimate primary productivity (Antoine et al., 1996), and understand algae dynamics (Richardson, 1996).

Finally, we gathered data on SDD (typically measured in meters), a long-standing method for estimating water clarity (Lee et al., 2018; Secchi, 1864). SDD is a simple measurement that integrates the optical properties of all water constituents and can provide information on the trophic status of water bodies (Carlson, 1977) or the algal status of a water body (Lorenzen, 1980).

2.2. Data Sources

Combining in situ data with the Landsat surface reflectance archive first requires a large repository of water quality samples in order to increase the probability of spatiotemporally colocated satellite and field samples. For this paper, we focused on the two largest databases of water quality in the United States: the WQP and LAGOS-NE. These data sets contrast in important ways: one has more data, emphasizing data quantity (WQP), and the other has more quality assurances (LAGOS-NE). Using both ensures sampling the largest possible number of water bodies, while retaining a harmonized, analysis-ready subset of the data.

2.2.1. WQP

The WQP, with mostly data from the United States, is the largest observation data set of water quality in the world. The WQP houses more than 290 million observations at 2.7 million sites dating back more than a century (Read et al., 2017). The WQP continuously gathers water quality information from more than 450 organizations including academic, government, NGO, tribal, and state data sets (Read et al., 2017). These data streams are then distributed in a standardized format, facilitating analysis across collection methods. While there is no entity that harmonizes the data across providers (Read et al., 2017), subsets of the data have been used in many publications analyzing water quality change in the United States (Booth et al., 2011; Oelsner et al., 2017; Sprague & Lorenz, 2009). As with many large data sets, the diversity of data sources and variation in metadata quality pose significant challenges to directly using the WQP as an analysis-ready data set (Sprague et al., 2017). Instead, end-users must carefully harmonize data across sampling methods, analytic approaches, and measurement units. The nature of harmonizing such large, distributed data generates a necessary trade-off between a deep, time-consuming exploration of data interoperability and a shallower, less time-consuming, but potentially more error-prone data quality check.

2.2.2. LAGOS-NE

The LAGOS project (which generated the data set LAGOS-NE) was, in part, designed to address some of the data harmonization issues inherent to the WQP, with the explicit goal of building a publicly available high-quality data set for continental-scale lake analyses (Soranno et al., 2015, 2017). In addition to pairing in situ lake data with physical lake characteristics and local geologic setting, LAGOS researchers harmonized key water quality measurements across the 87 water quality data sets that they gathered (Soranno et al., 2015, 2017). We used the LAGOSNE R package to access all data and used LAGOS-NE v.1.087.1 (Stachelek et al., 2017). The LAGOS-NE database includes only one sampling event per day per lake. In the few cases where more than one program sampled the same lake on a given day, the researchers selected the program that sampled at the deepest sampling location or that had the most water quality parameters measured. In addition, for all lakes in LAGOS-NE that had a water quality observation, the observation was assigned to the lake centroid regardless of sampling location within the lake because many water quality programs did not provide the position of the sampling location (Soranno et al., 2015). This approach is different from that used by the WQP, which often includes multiple sites and depths per water body and simultaneous observations. In its current form, the LAGOS-NE data set covers only lakes in the Northeast and Midwest, two lake-rich regions of the United States. LAGOS-NE (v1.087.1) provides a data set of the highest quality for matching in situ data to Landsat overpasses.

2.2.3. Landsat

For this project, we joined the in situ database (WQP and LAGOS-NE) with the Landsat Tier 1 products. The Landsat program started in July 1972, as the Earth Resources Observation Satellite with an explicit mission to provide solutions for some of Earth's pressing issues associated with industry and environmental change (Loveland & Dwyer, 2012). We only use the three most recent Landsat mission data sets with imagery over the United States: Landsat 5 (Thematic Mapper, 1984–2012; 192,745 available images), Landsat 7 (Enhanced Thematic Mapper +, 1999–present; 197,564 images), and Landsat 8 (Operational Land Imager, 2013–present; 69,030 images). We elected to exclude, Landsat 4 multispectral scanner data (MSS), because this sensor is not readily harmonized with the Thematic Mapper and Operational Land Imagers; however, this could be an additional data source in the future. Final data for this publication was queried on 2 May 2019. The total number of usable images is significantly lower because of cloud cover, which varies greatly by region and season. Furthermore, on 31 May 2003, the Landsat 7 scan line corrector failed, causing the Landsat 7 images after this date to have striped data gaps (Storey et al., 2005). We included all Landsat 7 data before and after this date, but did not fill gaps associated with the scan line error. The orbit repeat period of all three satellites is 16 days, though at high latitudes overlapping images result in shorter revisit times (Loveland & Dwyer, 2012; Wulder et al., 2016). In most of the United States, a given location will be imaged at least once every 16 days, and during periods of mission overlap, images are available on average at least every 8 days.

Landsat 5 and 7 have onboard imagers that collect seven bands of imagery centered on three visible wavelengths (blue, green, and red) and four infrared wavelengths (near infrared, shortwave infrared 1, shortwave infrared 2, and thermal band). Designed for continuity with previous sensors, Landsat 8 has bands in the same spectral regions and improved signal-to-noise ratios, with an additional ultrablue band (Barsi et al., 2014). Landsat 7 and 8 have panchromatic bands at 15-m resolution, while Landsat 5 does not. To keep matchup data in a standard format across time, we chose to use bands that were available and had the same spatial resolution in at least two of the Landsat missions (Table S1 in the supporting information), we also did not use any bands beyond the longest shortwave infrared.

Satellite image data need to be atmospherically corrected to account for differences between what the satellite can image from space and the actual reflectance on the surface of the Earth. When properly applied, atmospheric corrections can reduce the interference of absorbing and scattering aerosols, sun glint, and other processes that contribute to the signal observed at the satellite over water bodies, which can mask the optical information from the water body itself (Gordon, 1997). There are many options tailored for atmospheric correction over inland waters available for users on a scene-by-scene basis, For large-scale analysis, the United States Geological Survey (USGS) developed a surface reflectance product available in Google Earth Engine (Tier 1 collection), which uses a version of the 6-SV radiative transfer model called Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) for Landsat 5 and 7 (Ju et al., 2012) and the Landsat 8 Surface Reflectance Code (LaSRC) for Landsat 8 (Doxani et al., 2018; Vermote et al., 2016). While these surface reflectance products were developed for terrestrial remote sensing and not inland water observations, recent work by Kuhn et al. (2019) demonstrates that LaSRC performs well (within 4% difference of field radiometry) in estimating surface reflectance over the Amazon river. Also, the USGS product is the only standardized reflectance product that is globally available at the spatial scale required for inland water observation. Users may want to apply other atmospheric corrections, so while we only publish the surface reflectance data, our code can be used to work with top-of-atmosphere reflectance as well.

2.3. Data Integration

Building this data set required a flexible code architecture with a single workflow to download data from all three portals. Steps in the workflow included segmenting the data downloads into manageable pieces, conducting quality assurance checks, and joining data into the final data files (Figure 1). To avoid redundant data transfers and computations, we constructed a data pipeline that allowed us to only update each intermediate data product when needed—that is, when related sections of code were altered or when we wanted to bring in new source data. We implemented the pipeline using the R package *remake* (FitzJohn, 2018), which uses text files to declare the relationships among data and code files, then reruns only the code that must be rerun to keep the data up to date. The *remake* R package follows in the tradition of the make program for compiling computer software (Feldman, 1979). Although our project uses three different tools (R, Python, and Google Earth Engine), each tool is called directly from R—version 3.5.1 (R Foundation for Statistical Computing, 2018)—and RMarkdown files (Allaire et al., 2018), such that remake could be used to keep track of recent changes to code and data regardless of the tool. This data pipeline approach made our own analysis more efficient and should also increase efficiency for future researchers who may want to recreate the data set themselves or modify our specific approach.

2.3.1. Water Quality Data Download and Quality Control

We developed an automated method to retrieve our five water quality parameters from the WQP and LAGOS sites. For the WQP we used the dataRetrieval R package (Hirsch & De Cicco, 2015), which allows systematical downloading of WQP data. The WQP contains hundreds of parameter types (under the field "characteristicName" in the WQP), and we carefully selected those that best represented our target parameters based on our own expertise and previously published research using the same data sources, see

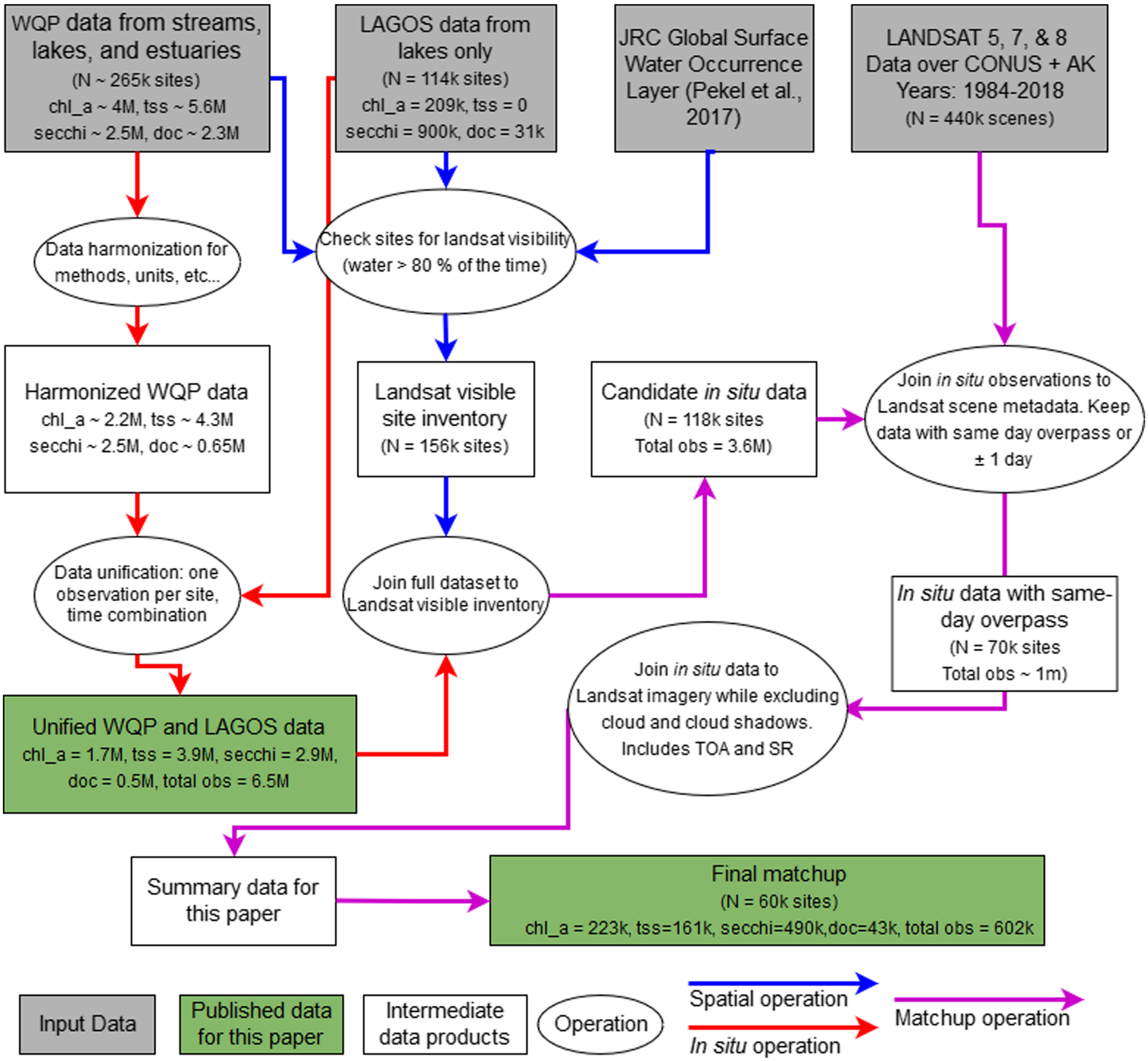
WQP data from streams, lakes, and estuaries
(N ~ 265k sites)
chl_a ~ 4M, tss ~ 5.6M
secchi ~ 2.5M, doc ~ 2.3M
LAGOS data from lakes only
(N = 114k sites)
chl_a = 209k, tss = 0
secchi = 900k, doc = 31k
JRC Global Surface Water Occurrence Layer (Pekel et al., 2017)
LANDSAT 5, 7, & 8
Data over CONUS + AK
Years: 1984-2018
(N = 440k scenes)
Data harmonization for methods, units, etc...
Check sites for landsat visibility (water > 80 % of the time)
Harmonized WQP data
chl_a ~ 2.2M, tss ~ 4.3M
secchi ~ 2.5M, doc ~ 0.65M
Landsat visible site inventory
(N = 156k sites)
Candidate in situ data
(N = 118k sites
Total obs = 3.6M)
Join in situ observations to Landsat scene metadata. Keep data with same day overpass or ± 1 day
Data unification: one observation per site, time combination
Join full dataset to Landsat visible inventory
In situ data with same-day overpass
(N = 70k sites
Total obs ~ 1m)
Join in situ data to Landsat imagery while excluding cloud and cloud shadows. Includes TOA and SR
Unified WQP and LAGOS data
chl_a = 1.7M, tss = 3.9M, secchi = 2.9M,
doc = 0.5M, total obs = 6.5M
Summary data for this paper
Final matchup
(N = 60k sites)
chl_a = 223k, tss=161k, secchi=490k,doc=43k, total obs = 602k
Input Data
Published data for this paper
Intermediate data products
Operation
Spatial operation
In situ operation
Matchup operation

Figure 1. Overview of data sources, steps taken to join data, and total observation counts. Note site counts are noted with "*N* =," while observation counts are associated with each parameter.

supporting information Table S2 for more information (Butman et al., 2016; Stets & Striegl, 2012). For all selected parameters, we downloaded data for all U.S. states except Hawaii for four water body categories: Lake, Reservoir, or Impoundment; Stream; Estuary; and Facility, where facility can indicate wastewater treatment facilities, including lakes and ponds. Finally, we restricted our queries to data sampled in water, excluding sediment and benthic samples.

Working with the LAGOS-NE data required many fewer decisions to combine parameters, since LAGOS researchers have already harmonized and combined parameters into simple categories that reflect our general parameter codes (Soranno et al., 2015, 2017). LAGOS-NE includes measurements of: DOC, Chl_a, and SDD, but no data on TSS or CDOM. As with the WQP, the data set can be simply loaded using an R package, LAGOSNE (Soranno et al., 2017).

Turning data from the WQP into an analysis-ready data set similar to LAGOS-NE requires a chain of decisions documented and justified in the supporting information. We have attempted to make these decisions both clear and justifiable, with the end goal of producing a high-quality data set. Figure 1 presents these data quality assurance procedures and shows how they reduce the number of observations at each step. The following are the most important decisions:

1. All observations were verified to have analytical methods related to parameter name; when this was not the case, samples were dropped. For example, if an observation was supposed to report TSS, but the analytical method was listed as "Nitrogen in Water," then that sample would be dropped. For TSS in particular, we assumed that the characteristicName "Suspended Sediment Concentration" reflected the same data as "TSS" despite some methodological differences in the data collection as documented by Gray and others (2000). However, we keep the original name so end-users of the data can filter based on method as they see fit.
2. We harmonized the data across exchangeable units such that TSS and DOC data are in milligrams per liter, Chl_a data is in micrograms per liter, and Secchi disk depth is in meters. We removed all observations with mismatched units (e.g., SDD in milligrams per liter).
3. Ideally, all observations would include sample depth information for accurate pairing of surface water data with reflectance. However, only 40% of the harmonized water quality data has depth data. For observations that did have sample depth data, we removed all observations deeper than 100 m (<1% of data), a depth where constituent concentration likely has little effect on radiation leaving the water body. For the samples with recorded depth, more than 88% of data were sampled within 2 m of the surface of the waterbody, suggesting most samples are near surface. Given this high proportion of surface samples in the WQP data, we made the decision to keep all data with no depth information, assuming the vast majority of it was collected near the surface. Because some users will not want to keep this depthless data, we have also included an additional data set so that end-users can filter out all depthless data as they wish.
4. We verified that both LAGOS-NE and WQP data have only one observation per site at a particular date and/or time. Some observations include date without timestamp; for our purposes we needed one observation per date if only date information was available and one per datetime if timestamps were recorded. Where the date, time, and observation value were the same for multiple observations, we converted duplicates to a single value. When the site and date or datetime were the same, but the parameter values were different, we averaged multiple observations to a single observation if the coefficient of variation (standard deviation/mean) was less than 10% and removed observations with too many simultaneous observations (five per date time combination) or too much variation with no metadata explaining the repeat observations.
5. TSS can be separated into subcategories by particle size (like sand, clay, and silt fractions) or by particle type (organic or inorganic), because many TSS observations (>400,000) included these fractioning data sets, we split them into two additional parameters of interest: fraction sand (p_sand) and total inorganic sediment (tis). We kept fraction sand instead of clay and/or silt, because there was limited data on clay and silt fractions.
6. Finally, we filtered in situ data to include only data with environmentally possible values by exploring the large data set itself and looking into the literature for reasonable values. The thresholds are as follows Chl_a >0.01 and <10,000 μg/L; TSS > 0.01 and <100,000 mg/L; SDD > 0.01 or 0 m and <100 m; and DOC > 0.01 and <500 mg/L. These thresholds are quite inclusive, and future users may need to further filter the data.

While other selection criteria could have been included to make all observations fully consistent, we avoided choices that removed the majority of the WQP data. For example, some analytical methods do not measure the exact same thing, such as measuring chlorophyll a with a fluorescence probe versus that with high performance liquid chromatography. If we elected to only keep perfectly exchangeable methods, a majority of the data would be lost. To allow user-defined selections, we included the methods attribute in an additional data set. Some decisions that resulted in retaining data included the following: not filtering data based on sampling method, not including temperature data as a filter for DOC and Chl_a samples, and including data that had unlabeled sample fraction metadata. While these decisions may preclude some types of analysis, our free and open source code allows future researchers to choose different data quality criteria and generate a strict or expanded data set to match user needs. All of the raw WQP data can be found here (https://figshare.com/articles/dataset/wqp_raw_zip/8139290). The harmonized and

LAGOS unified in situ data here (https://figshare.com/articles/dataset/Full_harmonized_in-situ_datasets/8139362), and the final matchup data set here (https://figshare.com/articles/dataset/AquaSat/8139383). All data for the project can be found here (https://figshare.com/account/collections/4506140).

2.3.2. Joining Landsat and In Situ Data Sets

Both the WQP and LAGOS-NE include sample latitude and longitude. Joining the in situ data to Landsat requires using this location data to determine which sites are visible by Landsat, gather spatially averaged reflectance, and match water quality observations to temporally coincident overpasses. As a sample location, LAGOS-NE and WQP use lake centroids and observation points, respectively, with different site identifiers. Therefore, the AquaSat database may contain duplicates (i.e., from same original water quality observation) which the user can choose. The first step in linking these data sets is to identify sites and water bodies that extend beyond a Landsat pixel (30 m) as to obtain reflectance measurement from the water surface alone. In addition, WQP and LAGOS-NE sites were preserved if they were within 200 m of at least 1 pixel with water occurrence of 80% in the Surface Water Occurrence data set (Pekel et al., 2016). All such sites were kept in the data set and were spatially joined to an inventory of Landsat WRS-2 paths and rows. Each site was related to its corresponding Landsat tile with sizes of about 5,000 × 5,000 pixels.

The resulting data set included the dates and times that each tile was observed by any of the three Landsat missions. We joined this data set to the in situ database by date. In cases with multiple in situ observations of the same site (in WQP or LAGOS-NE) on the same day, we kept only the observation closest in time to the Landsat overpass. In order to maximize the size of the data set, we retained all in situ data that fell within ±1 day of a Landsat overpass. This 1-day window is conservative compared to previous work in lakes (Olmanson et al., 2011; Torbick et al., 2013) and rivers (Griffin et al., 2011), but may result in mismatches between reflectance and water quality parameter values for estuaries and rivers with rapidly changing discharge where water quality values can vary on subhourly intervals (Rode et al., 2016). The timing difference between overpasses and in situ collection is preserved in the final data set, and users can specify minimum overpass timing if they so choose.

With this data set of Landsat-visible water quality sites matched to Landsat overpass times, we used Google Earth Engine to pair in situ observations with Landsat reflectance values near the observation site. To ensure the highest-quality reflectance data, we took several quality assurance steps. First, within a 200-m buffered zone we removed any pixel not classified as water at least 80% of the time in the Landsat archive (Pekel et al., 2016), which reduced the likelihood of using mixed pixels. Second, Landsat data include quality assessment bands for detection of water, clouds, aerosols, and other similar conditions. We used these bands to remove all pixels classified as cloud and cloud shadows, but we elected to keep all data classified as land, snow/ice, or water, since high sediment concentrations can lead to classification as land or ice. Third, because many of the samples in the WQP are taken from or near bridges narrower than 30 m, we also created a 30-m buffer around the TIGER road data set from the U.S. Census office (Ross et al., 2017). This step ensured that pixels within 30 m of any transport artery (road, traintracks, etc.) were removed, excluding mixed water/road pixels. After these quality assurance steps were taken, we calculated a spatial median of reflectance in each band from all remaining pixels in the 200-m buffer zone. To assess uncertainty and variance within these buffers, we also gathered the total pixel count and the standard deviation of reflectance values. The spatial medians include a median of the quality assessment band and standard deviation of the band. Together these values can be used to indicate if the quality band was uniformly a single value like water or was a mixture of land, ice, and/or water. These steps produced a "wide" (Wickham, 2014) data set matching in situ observation and Landsat reflectance values for all site and date combinations. Consistent with our philosophy of including as much data as possible, we did not set any specific thresholds for number of water pixels or qa bands but include these data for end-user filtering, though it is likely that most users will end up wanting to use sites only with at least 9 pixels per observation.

3. Results

The final data set structure is shown in Table 1. The quality assurance steps yielded more than 600,000 in situ samples that fell within ±1 day of a Landsat overpass. Matching in situ data to Landsat overpasses limits the data set to only 4–15% of the total in situ observations (Figure 1), with the biggest reduction in TSS observations and the smallest in SDD. This pattern stems from the fact that most TSS observations are made in streams too small to be visible from Landsat, while SDD observations are mostly in lakes, which are generally more visible. Given this reduction, we elected to remove CDOM from AquaSat because there were only

Table 1
Column names and data descriptions for AquaSat.

Column name	Content
system:index	Google Earth Engine system index
SiteID	Either the MonitoringLocationIdentifier from the WQP or lagoslakeid from LAGOS
blue	Median blue reflectance
blue_sd	Standard deviation of blue
date	Date without timestamp
date_unity	Date with timestamps when possible in local time zone
green	Median green reflectance
green_sd	Standard deviation of green
nir	Median nir reflectance
nir_sd	Standard deviation of nir
path	Landsat PATH
pixelCount	Number of water pixels that are averaged into each median and sd value
qa	The quality assessment band indicating clouds, land, and other classifications, see USGS surface reflectance summary for more info
qa_sd	Standard deviation of the quality band
red	Median red reflectance
red_sd	Standard deviation of red
row	Landsat ROW
sat	Landsat satellite (5,7,or 8)
swir1	Median shortwave infrared reflectance at 1,550–1,750 nm
swir1_sd	Standard deviation of shortwave infrared
swir2	Median of shortwave infrared reflectance at 2,000–2,350 nm
swir2_sd	Standard deviation of shortwave infrared
.geo	Geometry of Google Earth Engine spatial object
date_only	A T/F column marking whether or not there was timestamp available; T means no timestamp was available
chl_a	Chlorophyll a concentration in ug/L
doc	Dissolved organic carbon concentration in mg/L
p_sand	Percent sand in %
secchi	Secchi disk depth in m
tis	Total inorganic sediment in mg/L
tss	Total suspended sediment in mg/L
source	Source marking either WQP or LAGOS
lat	Latitude in WGS84
long	Longitude in WGS84
TZID	Site time zone ID
date_utc	Date with timestamp in UTC
clouds	Cloudiness score for the entire Landsat scene ranges from 0 (no clouds) to 100 (all clouds)
time	Timestamp
landsat_id	Landsat scene ID
timediff	Time difference between in situ water quality measurement and Landsat overpass
pwater	The median value for all water pixels in a 200-m radius from the sampling site, can be used to setup a stricter filter to keep only data from the center of a lake, river, or estuary
type	Type of water body, either lake, river, or estuary

WQP = Water Quality Portal; USGS = United States Geological Survey.

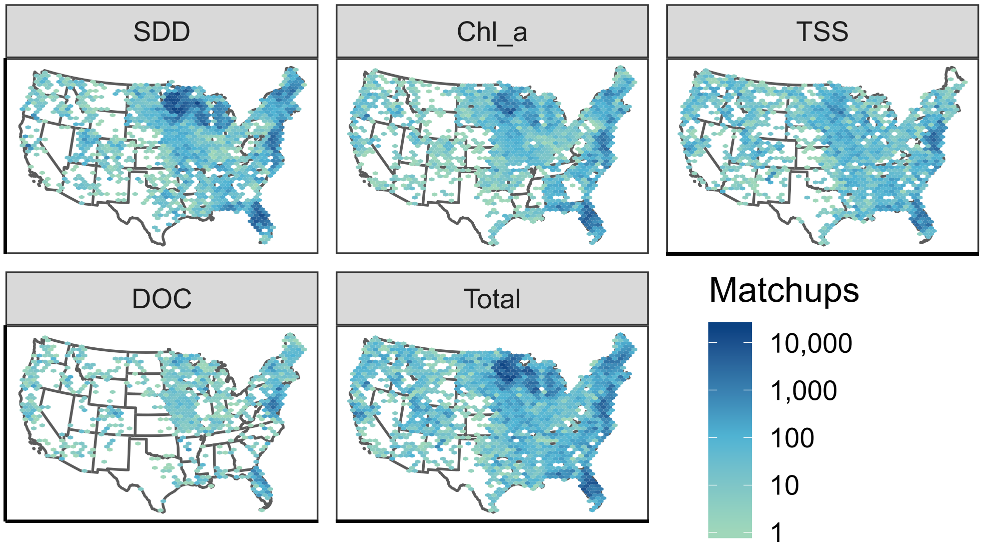

Figure 2. Distribution of observations across the conterminous United States, binning nearby observations, and integrating lake, river, and estuary data. The data are split by observation type, where total represents an overpass for any of the four primary parameters. Data for AK and HI not shown, though they are in the data set. SDD = Secchi Disk Depth; Chl_a = Chlorophyll a; TSS = Total Suspended Sediment; DOC = dissolved organic carbon.

2,761 in situ CDOM results in the entire WQP with too little information to harmonize between the many units and measurement approaches used to estimate CDOM. Given the challenges of data harmonization and the likelihood of only 100–200 overpasses for CDOM, we elected to drop it from the matchup analysis. The remaining data are well distributed across the parts of the United States with many lakes and rivers, including the Upper Midwest, Northeast, and Florida, with notable data concentrations near the Chesapeake Bay and along the U.S. East Coast in major estuarine environments (Figure 2). The western United States has notably less data available, which likely reflects lower concentrations of lakes and rivers in these states, the lack of LAGOS-NE data for these states, and, potentially, a bias in the completeness of the WQP toward certain states.

Lake-rich regions like Florida, Wisconsin, and Minnesota dominate the data set, with lakes contributing 70% of all data, mostly in the form of SDD observations. Figure 2 shows that DOC is the rarest observation in AquaSat, as it is in the in situ data (supporting information Figure S1). Half of all data comes from sites with only one or two matchups and less than 10% of sites have 25 or more observations. Given this limitation, reflectance-based water quality models borrowing information across many sites may be the most efficient way to use the database. Still, there are hundreds of sites for each parameter that have at least 50 matchups, which presents exciting opportunities for site-specific remote sensing research and possibly even assessments of water quality trends over time.

The temporal distribution of available data in our matchup data set generally reflects the availability of data in WQP and LAGOS-NE and the launch or retirement of Landsat missions (Figure S2). It also reflects the original WQP data (Read et al., 2017), as there are increasing data available in the in situ data sets from 1984 to 2012. The more recent decline in data availability may reflect a lag between agencies collecting data and submitting final data sets to the in situ databases and decreased funding (Myers & Ludtke, 2017) to monitoring organizations.

The data we captured in the matchup data set reflect the distribution of in situ data (Figure 3). This is especially true for Chl_a and SDD, where the overpass distribution shapes are nearly identical to the in situ distributions, just with fewer observations. The matchup data miss the largest values for both DOC and TSS, which occur almost entirely in small streams not visible to Landsat. Across parameter values (depths or concentrations), the matchup data span several orders of magnitude and capture environmentally meaningful variation in water quality. For each parameter, the data are approximately log-normally distributed, with the majority of the data occupying a relatively narrow range, within ~1–2 orders of magnitude of the median (Figure 3).

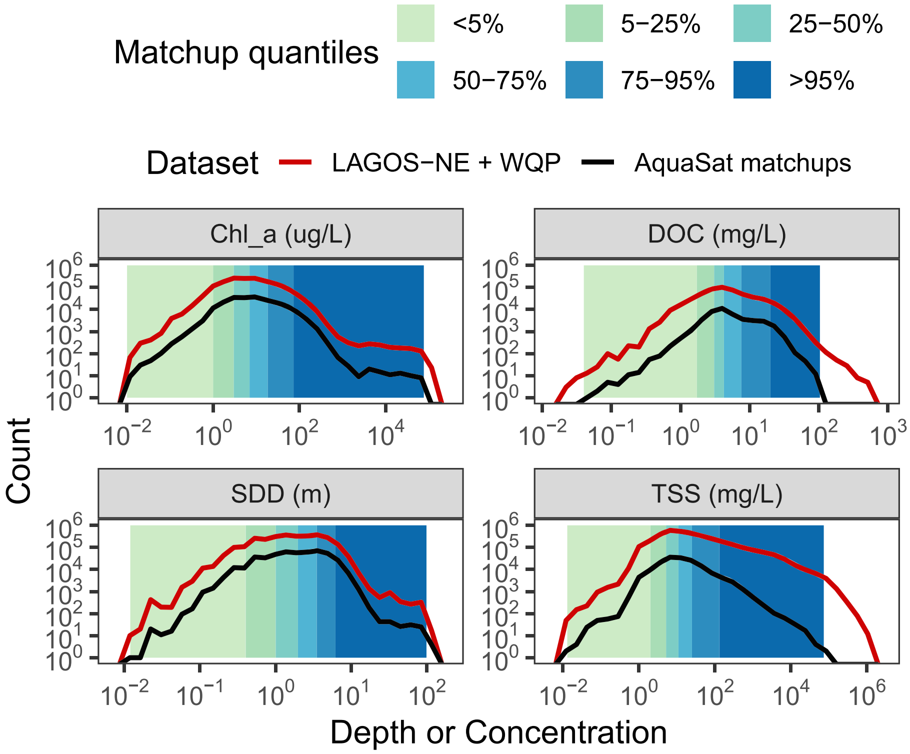

Figure 3. Data distributions for the in situ data (black) and the matchup data (red). Data quantiles are shown in the background as a color ramp from sage to blue, corresponding to the color scale in Figure 4. Quantiles were calculated for matchup data only. SDD = Secchi Disk Depth; Chl_a = Chlorophyll a; TSS = Total Suspended Sediment; DOC = dissolved organic carbon.

Based on decades of previous research, optically active constituents control absorption and scattering properties of water bodies, which in turn impacts their reflectance. While exploring these relationships at individual sites or regions is beyond the scope of this paper, we interrogate the data set to examine how variation in each water quality constituent maps to variation in reflectance in each spectral band. We divided AquaSat into the six quantiles shown in Figure 4 for each water quality parameter. Increasing concentrations of Chl_a, DOC, and TSS or increasing SDD control spectral variation, even when averaged across our three water body types (Estuary, Stream, and Lake) and averaged for the entire United States. Despite using such a heterogeneous data set, Figure 4 shows clear systematic variation in spectral response for each parameter as concentration or SDD increases.

4. Discussion

To our knowledge, the AquaSat data set is the largest matchup data set ever assembled for inland waters. Combining historical data sets of water quality and satellite reflectance maximizes the information we can gain from past data collections. Aquasat can inform our future approaches to in situ water quality monitoring by, for example, targeting sampling efforts near satellite overpass days. AquaSat, a data set essentially built on the overlapping of in situ water quality monitoring and Landsat imaging schedules, captures a broad range of variation in four major remotely observable water quality parameters across thousands of water bodies, and we anticipate it will unlock many avenues for remote water quality work. The four parameters in AquaSat (DOC, Chl_a, SDD, and TSS), within most of their range, have a significant impact on the observed surface reflectance, showing promising results for the value of these data to build predictive algorithms.

By publishing these data, we hope to contribute to the ongoing transition in the field from primarily developing methods to one where those methods are used to interrogate patterns in water quality and drivers of water quality change. For example, with an open, big data set, method comparisons for predictive models can be conducted, accelerating scientific discovery (Bukata, 2013). We built this data set to provide an easy way for nonexperts in remote sensing to begin using it in their research. Because the matchup data cover the

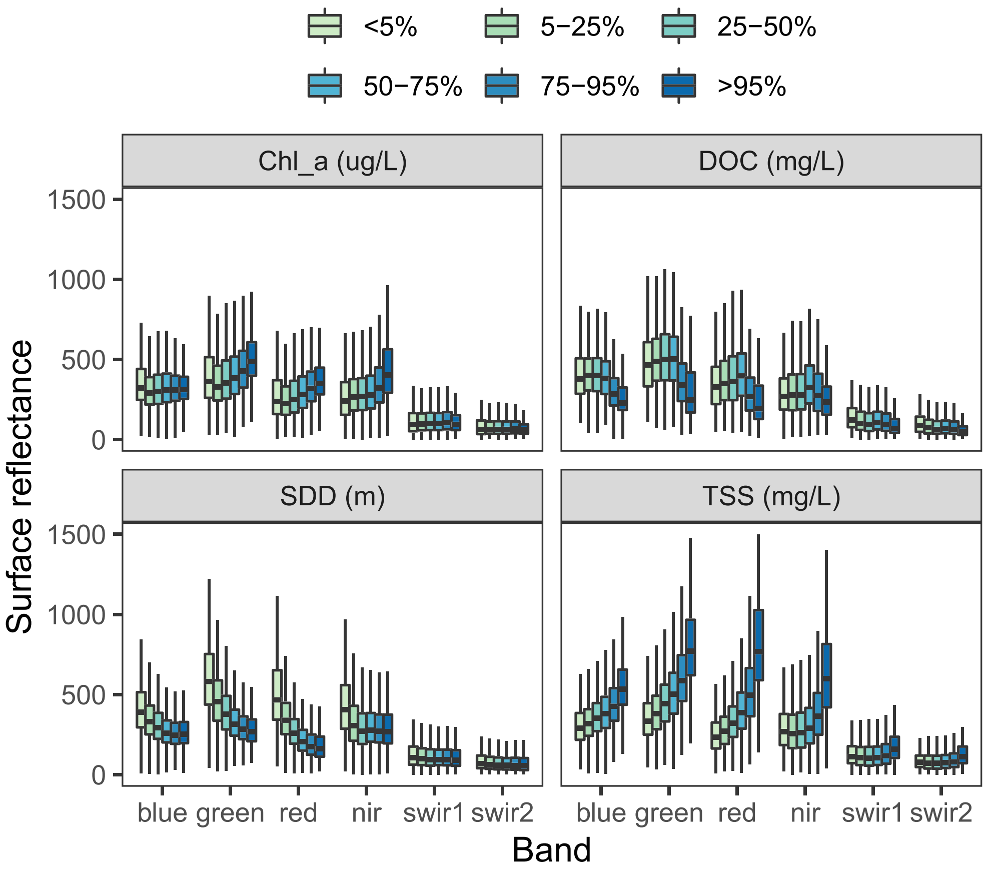

Figure 4. Shows spectral response, scaled and dimensionless, for each data quantile for each Landsat band. For Chl_a, DOC, and TSS, concentration increases moving from left to right for higher quantiles. For SDD, higher quantiles indicate increasing clarity or deeper SDD. The value range represented in each quantile is shown in Figure 3. SDD = Secchi Disk Depth; Chl_a = Chlorophyll a; TSS = Total Suspended Sediment; DOC = dissolved organic carbon.

United States, this work could range from classic approaches like building local water quality algorithms for detecting algae blooms in a single lake to more regional and national work predicting TSS in all the major rivers of the United States. We expect that as remote estimates of inland water quality become more common, they can be used as a complement to in situ data sets, vastly expanding our understanding of water quality trends and current status across the United States and world. We also anticipate that by enabling more work on remote sensing of water quality, we can fill in some of the spatial biases in water quality data that are inherent to the WQP and efforts like LAGOS (Stanley et al., 2019).

We built AquaSat to move toward continental- and global-scale remote sensing of water quality, but the data set comes with caveats and limitations. First and foremost, the WQP and LAGOS-NE have spatial biases in terms of which water bodies were sampled, which agencies fully report their data, and the completeness of records; these biases are carried over into AquaSat (Stanley et al., 2019). Second, our efforts to harmonize and unify the data in the WQP were performed with the explicit goal of including as much data as possible. Such inclusivity ensured a data set that allows future users the flexibility to set their own standards in line with the requirements of their individual needs, but it comes with intentionally limited quality. The LAGOS-NE data set, which was more extensively assured for quality, exemplifies a contrasting approach (Soranno et al., 2015, 2017). Lastly, for the remote sensing data, we used published surface reflectance estimates developed primarily with terrestrial remote sensing in mind, though recent work suggests these approaches may be as effective as custom approaches for aquatic remote sensing (Kuhn et al., 2019). For researchers who prefer their own atmospheric corrections, we also provide code for pulling top-of-atmosphere reflectance, which has no atmospheric correction applied. To enable future researchers to change some of these decisions, we also publish additional data sets with more raw information preserved, like analytical methods, so that other researchers can decide to filter data based on their own needs.

Our approach of pairing public in situ and satellite data can be expanded to any place with measurements of water quality. By publishing our code, we encourage use of our approach or code in other countries, moving toward truly global remote sensing of inland water quality. Additionally, there is ample previous work showing that remote sensing of water quality can be expanded to include constituents that are not optically active but are correlated with TSS or DOC, like mercury (Fichot et al., 2016; Telmer et al., 2006) or phosphorus (Kutser et al., 1995). Finally, this work can be adjusted to include other satellites with publicly available optical imagery (like Sentinel 2) or even private satellites with higher temporal and spatial resolution (like DigitalGlobe or PLANET). Our hope is that the content and philosophy of AquaSat help to accelerate progress in all of these areas.

Acknowledgments
All code used to generate this data set can be found at this site (https://github.com/GlobalHydrologyLab/AquaSat). The data for this paper come from the Landsat archive, LAGOSNE, and the Water Quality Portal. All of these data are free to download and the paper details extensively how we combined the data sets. Data generated by our work can be found on figshare (10.6084/m9.figshare.c.4506140). Part of this work was conducted at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. S. N. T. was funded by a NASA NESSF Grant (80NSSC18K1398) for work on this project. A. P. A. was supported by the Integration Information Dissemination Division of the U.S. Geological Survey. There are no conflicts of interest to report. We would also like to thank our reviewers for thorough, helpful, and encouraging reviews.

References

Allaire, J. J., Xie, Y., Mcphereson, J., Luraschi, J., Ushey, K., Atkins, A., et al. (2018). R markdown: Dynamic documents for R. R package version 1.11. https://rmarkdown.rstudio.com

Antoine, D., André, J.-M., & Morel, A. (1996). Oceanic primary production: 2. Estimation at global scale from satellite (Coastal Zone Color Scanner) chlorophyll. *Global Biogeochemical Cycles*, *10*(1), 57–69. http://doi.org/10.1029/95GB02832

Ballantine, D. J., & Davies-Colley, R. J. (2014). Water quality trends in New Zealand rivers: 1989–2009. *Environmental Monitoring and Assessment*, *186*(3), 1939–1950. http://doi.org/10.1007/s10661-013-3508-5

Barsi, J., Lee, K., Kvaran, G., Markham, B., Pedelty, J., Barsi, J. A., et al. (2014). The spectral response of the Landsat-8 operational land imager. *Remote Sensing*, *6*(10), 10,232–10,251. http://doi.org/10.3390/rs61010232

Blondeau-Patissier, D., Gower, JamesF. R., Dekker, A. G., Phinn, S. R., & Brando, V. E. (2014). A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Progress in Oceanography*, *123*, 23–144. http://doi.org/10.1016/j.pocean.2013.12.008

Booth, N. L., Everman, E. J., Kuo, I.-L., Sprague, L., & Murphy, L. (2011). A web-based decision support system for assessing regional water-quality conditions and management actions. *JAWRA Journal of the American Water Resources Association*, *47*(5), 1136–1150. http://doi.org/10.1111/j.1752-1688.2011.00573.x

Bricaud, A., Morel, A., & Prieur, L. (1981). Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domains. *Limnology and Oceanography*, *26*(1), 43–53. http://doi.org/10.4319/lo.1981.26.1.0043

Bukata, R. P. (2013). Retrospection and introspection on remote sensing of inland water quality: Like déjà vu all over again. *Journal of Great Lakes Research*, *39*, 2–5. https://doi.org/10.1016/j.jglr.2013.04.001

Butman, D., Stackpoole, S., Stets, E., McDonald, C. P., Clow, D. W., & Striegl, R. G. (2016). Aquatic carbon cycling in the conterminous United States and implications for terrestrial carbon accounting. *Proceedings of the National Academy of Sciences*, *113*(1), 58–63. https://doi.org/10.1073/pnas.1512651112

Carlson, R. E. (1977). A trophic state index for lakes. *Limnology and Oceanography*, *22*(2), 361–369. https://doi.org/10.4319/lo.1977.22.2.0361

Clarke, G. L., Ewing, G. C., & Lorenzen, C. J. (1970). Spectra of backscattered light from the sea obtained from aircraft as a measure of chlorophyll concentration. *Science*, *167*(3921), 1119–1121. https://doi.org/10.1126/science.167.3921.1119

Cory, R. M., Harrold, K. H., Neilson, B. T., & Kling, G. W. (2015). Controls on dissolved organic matter (DOM) degradation in a headwater stream: The influence of photochemical and hydrological conditions in determining light-limitation or substrate-limitation of photo-degradation. *Biogeosciences Discussions*, *12*(13), 9793–9838. https://doi.org/10.5194/bgd-12-9793-2015

Doxani, G., Vermote, E., Roger, J.-C., Gascon, F., Adriaensen, S., Frantz, D., et al. (2018). Atmospheric correction inter-comparison exercise. *Remote Sensing*, *10*(3), 352. https://doi.org/10.3390/rs10020352

Feldman, S. I. (1979). Make a program for maintaining computer programs. *Software: Practice and Experience*, *9*(4), 255–265. https://doi.org/10.1002/spe.4380090402

Fichot, C. G., Downing, B. D., Bergamaschi, B. A., Windham-Myers, L., Marvin-Dipasquale, M., Thompson, D. R., & Gierach, M. M. (2016). High-resolution remote sensing of water quality in the San Francisco Bay-Delta Estuary. *Environmental Science and Technology*, *50*(2), 573–583. https://doi.org/10.1021/acs.est.5b03518

FitzJohn, R. (2018). remake: Make-like build management. R package version 0.3.0. https://github.com/richfitz/remake

Gholizadeh, M., Melesse, A., & Reddi, L. (2016). A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors*, *16*(8), 1298. https://doi.org/10.3390/s16081298

Gordon, H. R. (1997). Atmospheric correction of ocean color imagery in the Earth Observing System era. *Journal of Geophysical Research*, *102*(D14), 17,081–17,106. https://doi.org/10.1029/96JD02443

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Griffin, C. G., Finlay, J. C., Brezonik, P. L., Olmanson, L., & Hozalski, R. M. (2018). Limitations on using CDOM as a proxy for DOC in temperate lakes. *Water Research*, *144*, 719–727. https://doi.org/10.1016/J.WATRES.2018.08.007

Griffin, C. G., Frey, K. E., Rogan, J., & Holmes, R. M. (2011). Spatial and interannual variability of dissolved organic matter in the Kolyma River, East Siberia, observed using satellite imagery. *Journal of Geophysical Research*, *116*, G03018. https://doi.org/10.1029/2010JG001634

Hestir, E. L., Brando, V. E., Bresciani, M., Giardino, C., Matta, E., Villa, P., & Dekker, A. G. (2015). Measuring freshwater aquatic ecosystems: The need for a hyperspectral global mapping satellite mission. *Remote Sensing of Environment*, *167*, 181–195. https://doi.org/10.1016/J.RSE.2015.05.023

Hirsch, R. M., & De Cicco, Laura (2015). User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data. *Techniques and Methods book 4*, *February*, 93. https://doi.org/10.3133/tm4A10

Holyer, R. J. (1978). Toward universal multispectral suspended sediment algorithms. *Remote Sensing of Environment*, *7*(4), 323–338. https://doi.org/10.1016/0034-4257(78)90023-8

Ju, J., Roy, D. P., Vermote, E., Masek, J., & Kovalskyy, V. (2012). Continental-scale validation of MODIS-based and LEDAPS Landsat ETM+ atmospheric correction methods. *Remote Sensing of Environment*, *122*, 175–184. https://doi.org/10.1016/J.RSE.2011.12.025

Julian, J. P., Doyle, M. W., Powers, S. M., Stanley, E. H., & Riggsbee, J. A. (2008). Optical water quality in rivers. *Water Resources Research*, *44*, W10411. https://doi.org/10.1029/2007WR006457

Klemas, V., Borchardt, J. F., & Treasure, W. M. (1973). Suspended sediment observations from ERTS-1. *Remote Sensing of Environment*, *2*, 205–221. https://doi.org/10.1016/0034-4257(71)90094-0

Kuhn, C., Valerio, AlinedeMatos, Ward, N., Loken, L., Sawakuchi, H., Kampel, M., et al. (2019). Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. *Remote Sensing of Environment*, *224*, 104–118. https://www.sciencedirect.com/science/article/pii/S0034425719300288

Kutser, T. (2004). Quantitative detection of chlorophyll in cyanobacterial blooms by satellite remote sensing. *Limnology and Oceanography*, *49*(6), 2179–2189. https://doi.org/10.4319/lo.2004.49.6.2179

Kutser, T., Arst, H., Miller, T., Käärmann, L., & Milius, A. (1995). Telespectrometrical estimation of water transparency, chlorophyll-a and total phosphorus concentration of Lake Peipsi. *International Journal of Remote Sensing*, *16*(16), 1–2. https://doi.org/10.1080/01431169508954609

Lack, T. (2000). Eurowaternet—A freshwater monitoring and reporting network for all European countries, *Transboundary water resources in the balkans* pp. 185–191). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4367-7_19

Lee, B. Z., Arnone, R., Boyce, D., Franz, B., Greb, S., Hu, C., et al. (2018). Global water clarity: Continuing a century-long monitoring. *Eos*, *99*(May), 1–10. https://doi.org/10.1029/2018EO097251

Loew, A., Bell, W., Brocca, L., Bulgin, C. E., Burdanowitz, J., Calbet, X., et al. (2017). Validation practices for satellite-based Earth observation data across communities. *Reviews of Geophysics*, *55*, 779–817. https://doi.org/10.1002/2017RG000562

Lorenzen, M. W. (1980). Use of chlorophyll-Secchi disk relationships. *Limnology and Oceanography*, *25*(2), 371–372. https://doi.org/10.4319/lo.1980.25.2.0371

Loveland, T. R., & Dwyer, J. L. (2012). Landsat: Building a strong future. *Remote Sensing of Environment*, *122*(October 2000), 22–29. https://doi.org/10.1016/j.rse.2011.09.022

Maul, G. A., & Gordon, H. R. (1975). On the use of the Earth resources technology satellite (LANDSAT-1) in optical oceanography. *Remote Sensing of Environment*, *4*(C), 95–128. https://doi.org/10.1016/0034-4257(75)90008-5

Mishra, D., Ogashawara, I., & Gitelson, A. (2017). *Bio-optical modeling and remote sensing of inland waters*: Elsevier. https://books.google.com/books?hl=en&lr=&id=jgNQCwAAQBAJ&oi=fnd&pg=PP1&dq=bio-optical+modelling+and+remote+sensing+of+inland+waters&ots=FoWel5zE_4&sig=52oX6rE9-IqeDYDcXgc8Mus0WYA .

Myers, D. N., & Ludtke, A. S. (2017). Progress and lessons learned from water-quality monitoring networks, *Chemistry and water* (vol. 33, pp. 23–120): Elsevier. https://doi.org/10.1016/B978-0-12-809330-6.00002-7

Odermatt, D., Danne, O., Philipson, P., & Brockmann, C. (2018). Diversity II water quality parameters from ENVISAT (2002–2012): A new global information source for lakes. *Earth System Science Data*, *10*(3), 1527–1549. https://doi.org/10.5194/essd-10-1527-2018

Oelsner, G. P., Sprague, L. A., Murphy, J. C., Zuellig, R. E., Johnson, H. M., Ryberg, K. R., et al. (2017). Water-quality trends in the nation's rivers and streams, 1972–2012. *USGS Scientific Investigations Report*, *5006*(October), 1972–2012. https://doi.org/10.3133/sir20175006

Olmanson, L. G., Bauer, M. E., & Brezonik, P. L. (2008). A 20-year Landsat water clarity census of Minnesota's 10,000 lakes. *Remote Sensing of Environment*, *112*(11), 4086–4097. https://doi.org/10.1016/j.rse.2007.12.013

Olmanson, L. G., Brezonik, P. L., & Bauer, M. E. (2011). Evaluation of medium to low resolution satellite imagery for regional lake water quality assessments. *Water Resources Research*, *47*, W09515. https://doi.org/10.1029/2011WR011005

Palmer, S. C. J., Kutser, T., & Hunter, P. D. (2015). Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sensing of Environment*, *157*, 1–8. https://doi.org/10.1016/j.rse.2014.09.021

Pavelsky, T. M., & Smith, L. C. (2009). Remote sensing of suspended sediment concentration, flow velocity, and lake recharge in the Peace-Athabasca Delta, Canada. *Water Resources Research*, *45*, W11417. https://doi.org/10.1029/2008WR007424

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

R Foundation for Statistical Computing (2018). R: A language and environment for statistical computing, Vienna, Austria. https://www.r-project.org

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., et al. (2017). Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research*, *53*, 1735–1745. https://doi.org/10.1002/2016WR019993

Richardson, L. L. (1996). Remote sensing of algal bloom dynamics. *BioScience*, *46*(7), 492–501. https://doi.org/10.2307/1312927

Ritchie, J., Schiebe, F., & McHenry, J. (1976). Remote sensing of suspended sediments in surface waters. *American Society of*, *42*(12), 1539–1545. https://trid.trb.org/view.aspx?id=66674

Robbins, C. J., King, R. S., Yeager, A. D., Walker, C. M., Back, J. A., Doyle, R. D., & Whigham, D. F. (2017). Low-level addition of dissolved organic carbon increases basal ecosystem function in a boreal headwater stream. *Ecosphere*, *8*(4), e01739. https://doi.org/10.1002/ecs2.1739

Rode, M., Wade, A. J., Cohen, M. J., Hensley, R. T., Bowes, M. J., Kirchner, J. W., et al. (2016). Sensors in the stream: The high-frequency wave of the present. *Environmental Science & Technology*, *50*(19), 10,297–10,307. https://doi.org/10.1021/acs.est.6b02155

Ross, W., Jarmin, R., Blumerman, L., Lamas, E., Ratcliffe, M. R., Hanks, G. F., & Doms, M. (2017). TIGER/Line Shapefiles 2017 Economic and Statistics Administration. https://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2017/TGRSHP 2017_TechDoc.pdf.

Secchi, P. A. (1864). Relazione delle esperienze fatte a bordo della pontificia pirocorvetta l'Immacolata concezione per determinare la trasparenza del mare; Memoria del P. A. Secchi. *Il Nuovo Cimento*, *20*(1), 205–238. https://doi.org/10.1007/BF02726911

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., & Verbist, K. (2018). Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. *Water Resources Research*, *54*, 9724–9758. https://doi.org/10.1029/2017WR022437

Soranno, P. A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., et al. (2017). LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. *GigaScience*, *6*(12), 1–22. https://doi.org/10.1093/gigascience/gix101

Soranno, P. A., Bissell, E. G., Cheruvelil, K. S., Christel, S. T., Collins, S. M., Fergus, C. E., et al. (2015). Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. *GigaScience*, *4*(1), 28. https://doi.org/10.1186/s13742-015-0067-4

Spencer, R. G. M., Stubbins, A., Hernes, P. J., Baker, A., Mopper, K., Aufdenkampe, A. K., et al. (2009). Photochemical degradation of dissolved organic matter and dissolved lignin phenols from the Congo River. *Journal of Geophysical Research*, *114*, G03010. https://doi.org/10.1029/2009JG000968

Sprague, L. A., & Lorenz, D. L. (2009). Regional nutrient trends in streams and rivers of the United States, 1993–2003. *Environmental Science and Technology*, *43*(10), 3430–3435. https://doi.org/10.1021/es803664x

Sprague, L. A., Oelsner, G. P., & Argue, D. M. (2017). Challenges with secondary use of multi-source water-quality data in the United States. *Water Research*, *110*, 252–261. https://doi.org/10.1016/j.watres.2016.12.024

Spyrakos, E., O'Donnell, R., Hunter, P. D., Miller, C., Scott, M., Simis, S. G. H., et al. (2017). Optical types of inland and coastal waters. *Limnology and Oceanography*, *63*, 846–870. https://doi.org/10.1002/lno.10674

Srebotnjak, T., Carr, G., de Sherbinin, A., & Rickwood, C. (2012). A global water quality index and hot-deck imputation of missing data. *Ecological Indicators*, *17*, 108–119. https://doi.org/10.1016/J.ECOLIND.2011.04.023

Stachelek, J., Oliver, S. K., & Masrour, F. (2017). LAGOS: R interface to the LAke multi-scaled GeOSpatial & temporal database. R package version 1.0.0.

Stanley, E. H., Collins, S. M., Lottig, N. R., Oliver, S. K., Webster, K. E., Cheruvelil, K. S., & Soranno, P. A. (2019). Biases in lake water quality sampling and implications for macroscale research. *Limnology and Oceanography*, *64*, 1572–1585. https://doi.org/10.1002/lno.11136

Stets, E., & Striegl, R. (2012). Carbon export by rivers draining the conterminous United States. *Inland Waters*, *2*(4), 177–184. https://doi.org/10.5268/IW-2.4.510

Storey, J., Scaramuzza, P., Schmidt, G., & Barsi, J. (2005). Landsat 7 scan line corrector-off gap filled product development. In Proceeding of *Pecora*, *16*, 23–27.

Syvitski, J. P. M., & Kettner, A. (2011). Sediment flux and the Anthropocene. *Philosophical transactions. Series A, mathematical, physical, and engineering sciences*, *369*(1938), 957–75. https://doi.org/10.1098/rsta.2010.0329

Telmer, K., Costa, M., Simões Angélica, R., Araujo, E. S., & Maurice, Y. (2006). The source and fate of sediment and mercury in the Tapajós River, Pará, Brazilian Amazon: Ground- and space-based evidence. *Journal of Environmental Management*, *81*(2), 101–113. https://doi.org/10.1016/j.jenvman.2005.09.027

Torbick, N., Hession, S., Hagen, S., Wiangwang, N., Becker, B., & Qi, J. (2013). Mapping inland lake water quality across the Lower Peninsula of Michigan using Landsat TM imagery. *International journal of remote sensing*, *34*(21), 7607–7624. https://doi.org/10.1080/01431161.2013.822602

Vähätalo, A. V., Wetzel, R. G., & Paerl, H. W. (2005). Light absorption by phytoplankton and chromophoric dissolved organic matter in the drainage basin and estuary of the Neuse River, North Carolina (U.S.A.) *Freshwater Biology*, *50*(3), 477–493. https://doi.org/10.1111/j.1365-2427.2004.01335.x

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, *185*, 46–56. https://doi.org/10.1016/J.RSE.2016.04.008

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Williams, G. P. (1989). Sediment concentration versus water discharge during single hydrologic events in rivers. *Journal of Hydrology*, *111*(1-4), 89–106. https://doi.org/10.1016/0022-1694(89)90254-0

Williamson, C. E., Dodds, W., Kratz, T. K., & Palmer, M. A. (2008). Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Frontiers in Ecology and the Environment*, *6*(5), 247–254. https://doi.org/10.1890/070140

Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S., Cohen, W. B., et al. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, *185*, 271–283. https://doi.org/10.1016/j.rse.2015.11.032