AquaSat: a dataset 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] **Tamlin Pavelsky** [2]

[1]Department of Ecosystem Science and Sustainability, Colorado State University,
[2]Department of Geological Sciences, University of North Carolina
[3]United States Geological Survey
[4]School of Environmental and Forest Sciences, University of Washington
[5]NASA Jet Propulsion Laboratory

Key Points:

- AquaSat contains >550,000 paired observations of water quality and Landsat reflectance, the largest such matchup dataset
- Matchups capture diverse waterbodies across the USA for 1984-2018; 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

Corresponding author: Matthew R.V. Ross, `matt.ross@colostate.edu`

Abstract

Satellite predictions 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 (TSS), Dissolved Organic Carbon (DOC), Chlorophyll a, and Secchi Disk Depth (SDD). 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 datasets 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 dataset ever assembled. AquaSat contains more than 550,000 matchups, covering 1984-2018, of ground-based TSS, DOC, Chlorophyll a, and SDD 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 datasets covering the contiguous United States, including the Water Quality Portal, LAGOS, and the Landsat archive. In addition to publishing the dataset, 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 towards understanding natural and anthropogenic drivers of aquatic ecosystem change (Srebotnjak, Carr, de Sherbinin, & Rickwood, 2012). Collecting such valuable data has historically been expensive and time-consuming, and it has often proved difficult to maintain analysis-ready and open datasets. 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. 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, Ewing, & Lorenzen, 1970; Holyer, 1978; Klemas, Borchardt, & Treasure, 1973; Maul & Gordon, 1975; Ritchie, Schiebe, & McHenry, 1976). From these early efforts, fifty years of work have used spectral information to predict 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 towards universal algorithms and unified approaches has been slow (Blondeau-Patissier, Gower, Dekker, Phinn, & Brando, 2014; Bukata, 2013; Gholizadeh, Melesse, & Reddi, 2016; Palmer, Kutser, & Hunter, 2015) especially at the global scale. Further, most papers published to date have focused on developing predictive methods as opposed to using predictions to interrogate processes that control water quality dynamics (Topp et al., *in review*).

This slow progress contrasts sharply with ocean remote sensing, which benefits from robust, open and big datasets geared towards 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 waterbodies. As a result, inland water remote sensing has been limited to satellites built for terrestrial remote sensing (Palmer et al., 2015). Methods development for inland waters 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. We think some of these inherent challenges in inland water quality remote sensing can be met at a broad scale because with a centralized, public remote sensing dataset paired with *in-situ* measurements of water quality (Palmer et al., 2015).

In this data paper, we present AquaSat, a merged dataset of *in-situ* water quality measurements paired with same-day or ± 1-day satellite reflectance (which we call "matchups"). This is the largest such matchup dataset ever assembled for inland waters. To create AquaSat, we use the Landsat archive from 1984-2018, 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 multi-scaled GeOSpatial and temporal database covering the northeastern United States (LAGOS-NE, Soranno et al., 2017). The WQP data we used covers all of the USA. Joining these datasets 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). We also outline and share our approach, code, and intermediate data for bringing the WQP, LAGOS-NE, and Landsat datasets 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 and CDOM, Chlorophyll a and Chl_a. These four parameters capture key ecological and physical factors that control water quality, and capabilities to remotely sense each of them have been demonstrated (Topp et al., *in review*).

TSS is a measure of the concentration of solids, both organic and inorganic, in a water column, measured in mg/L. 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, Doyle, Powers, Stanley, & Riggsbee, 2008), erosion conditions (Syvitski & Kettner, 2011), and the hydrologic status of waterbodies, where high TSS generally means sediment supply coupled with higher flow velocities (Pavelsky & Smith, 2009; Williams, 1989).

DOC, measured in mg/L, is the broad description for the concentration of organic carbon dissolved in water, and can provide insight into light conditions (Vähätalo, Wetzel, & Paerl, 2005), heterotrophic energy availability (Robbins et al., 2017), and terrestrial organic matter processing (Williamson, Dodds, Kratz, & Palmer, 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, Morel, & Prieur, 1981; Griffin, Frey, Rogan, & Holmes, 2011) (Bricaud et al., 1981; Griffin et al., 2011). This correlation between CDOM and DOC can break down in places with low DOC concentrations (Griffin, Finlay, Brezonik, Olmanson, & Hozalski, 2018), or in areas with high photobleaching of DOC, which alters the DOC/CDOM fractionation (Cory, Harrold, Neilson, & Kling, 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, André, & Morel, 1996), and understand algae dynamics (Richardson, 1996).

Finally, we gathered data on Secchi disk depth (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 waterbodies (Carlson, 1977) or the algal status of a waterbody (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 co-located 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 datasets 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 waterbodies, while retaining a harmonized, analysis-ready subset of the data.

2.2.1 Water Quality Portal

The WQP, with mostly data from the USA, is the largest observation dataset 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 datasets (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 USA (Booth, Everman, Kuo, Sprague, & Murphy, 2011; Oelsner et al., 2017; Sprague & Lorenz, 2009). As with many large datasets, the diversity of data sources and variation in meta-data quality pose significant challenges to directly using the WQP as an analysis-ready dataset (Sprague, Oelsner, & Argue, 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 dataset 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 dataset for continental-scale lake analyses (Soranno et al., 2017, 2015). 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 datasets that they gathered (Soranno et al., 2017, 2015). Because LAGOS researchers combined data from many different sources, they chose to identify all data for a single lake with the lake centroid. If two different organizations were measuring Secchi disk depth at the north and south end of a lake, the LAGOS dataset would combine all of these measurements into a single time series, located at the lake centroid. For same day observations, the deepest observation would have been kept (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 dataset covers only lakes in the Northeast and Midwest, two lake-rich regions of the USA. LAGOS-NE (v1.087.1) provides a dataset 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). For this project we are only using the three most recent Landsat mission datasets: Landsat 5 (Thematic Mapper, 1984-2012, 192745 available images), Landsat 7 (Enhanced Thematic Mapper +, 1999-present, 197564 images), and Landsat 8 (Operational Land Imager, 2013-present, 69030 images). The total number of usable images is significantly lower because of cloud cover, which varies greatly by region and season. Furthermore, on May 31, 2003, the Landsat 7 scan line corrector failed, causing the Landsat 7 images after this date to have striped data gaps (Storey, Scaramuzza, Schmidt, & Barsi, 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 sixteen days, though at high latitudes overlapping images result in shorter revisit times (Loveland & Dwyer, 2012; Wulder et al., 2016). In most of the USA, a given location will be imaged at least once every sixteen days, and during periods of mission overlap, images are available on average at least every eight 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 ultra-blue band (Barsi et al., 2014). Landsat 7 and 8 have panchromatic bands at 15m 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 (SI Table 1).

Satellite image data needs 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 waterbodies, which can mask the optical information from the waterbody 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 USGS developed a surface reflectance product available in Google Earth Engine which uses a version of the 6SV radiative transfer model called Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) for Landsat 5 and 7 (Ju, Roy, Vermote, Masek, & Kovalskyy, 2012) and the Landsat 8 Surface Reflectance Code (LaSRC) for Landsat 8 (Doxani et al., 2018; Vermote, Justice, Claverie, & Franch, 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 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 here the surface

reflectance data, our code can be used to work with top-of-atmosphere reflectance as well.

2.3 Data integration

Building this dataset 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 – i.e., 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 this 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 dataset 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 SI table 2 for more information (Butman et al., 2016; Stets & Striegl, 2012). For all selected parameters, we downloaded data for all US 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., 2017, 2015). LAGOS-NE includes measurements of: DOC, Chl_a, and SDD, but no data on TSS or CDOM. As with the WQP, the dataset can be simply loaded using an R package called LAGOSNE (Soranno et al., 2017).

Turning data from the WQP into an analysis-ready dataset similar to LAGOS-NE requires a chain of decisions documented and justified in the supplemental html file (SI 3). We have attempted to make these decisions both clear and justifiable, with the end goal of producing a high-quality dataset. 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

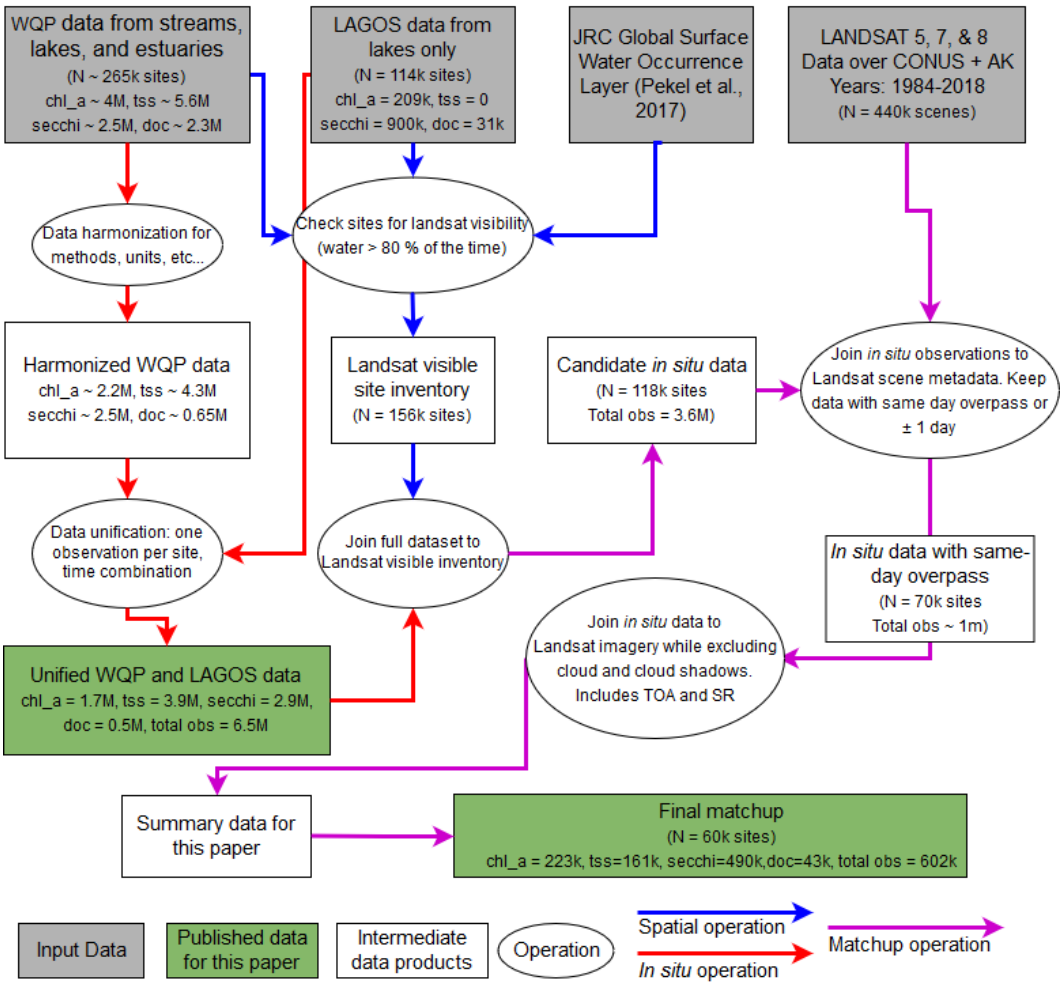

Figure 1. Overview of data sources, steps taken to join data, and total observation counts

"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 mg/L, Chl_a data is in mug/L, and Secchi disk depth is in meters. We removed all observations with mismatched units (e.g. SDD in mg/L).
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 meters (< 1% of data), a depth where constituent concentration likely has little effect on radiation leaving the waterbody. For the samples with recorded depth, more than 97% of data were sampled within 20m of the surface of the waterbody, suggesting most samples are near surface.
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 (5 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, silt fractions) or by particle type (organic or inorganic), because many TSS observations (> 400,000) included these fractioning datasets, 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 elected to keep all non-negative values not excluded by the preceding steps, requiring users to conduct their own assessment of how well the data reflect their expert knowledge of the system and parameter.

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 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 dataset. Some decisions that resulted in retaining data included: 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 dataset to match user needs.

3 Results

The quality assurance steps yielded almost 600,000 *in-situ* samples that fell within ± 1 day of a Landsat overpass. Matching *in-situ* data to Landsat overpasses limits the dataset 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 visible. Given this reduction, we elected to remove CDOM from AquaSat because there were only 2761 *in-situ* CDOM results in the entire WQP, which would have resulted in less than 100 total matchups. The remaining data are well distributed across the parts of the USA 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 towards certain states.

Lake-heavy regions like Florida, Wisconsin, and Minnesota dominate the dataset. They contribute 71% of the 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 (SI Fig 1). 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

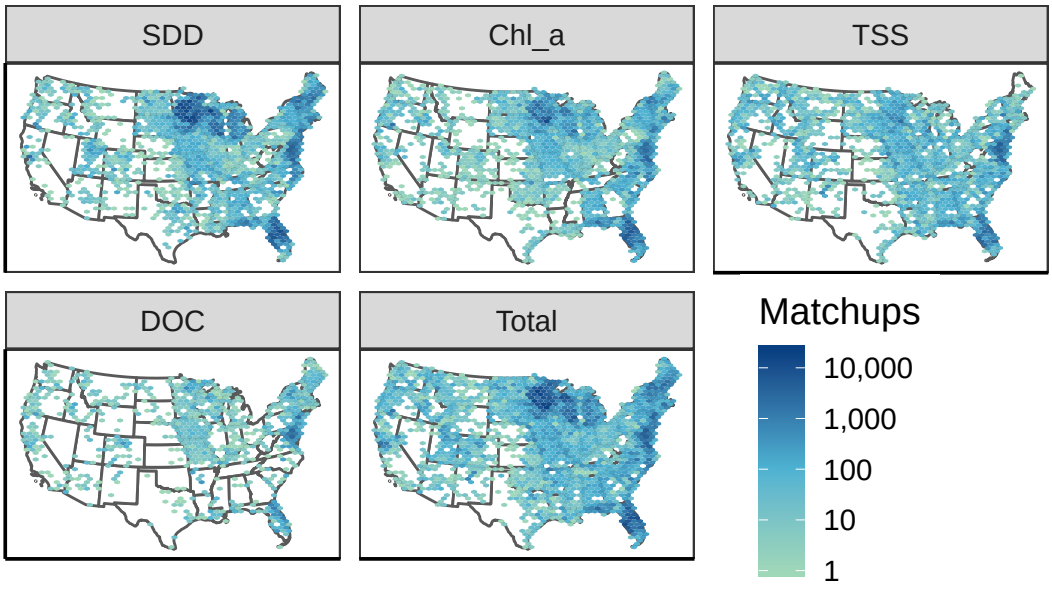

Figure 2. Distribution of observations across the conterminous USA, binned by local region. The data is split by observation type, where total represents an overpass for any of the four primary parameters. Data for AK and HI not shown, though they are in the dataset.

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 dataset generally reflects the availability of data in WQP and LAGOS-NE and the launch or retirement of Landsat missions (SI Fig 1). It also reflects the original WQP data (Read et al., 2017), as there is increasing data available in the *in-situ* datasets from 1984 to 2012. The more recent decline in data availability may reflect a lag between agencies collecting data and submitting final datasets to the *in-situ* databases and decreased funding (Myers & Ludtke, 2017) to monitoring organizations.

The data we captured in the matchup dataset reflects the distribution of *in-situ* data (Fig. 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 misses 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 spans several orders of magnitude and captures environmentally meaningful variation in water quality. For each parameter, the data is approximately log-normally distributed, with the majority of the data

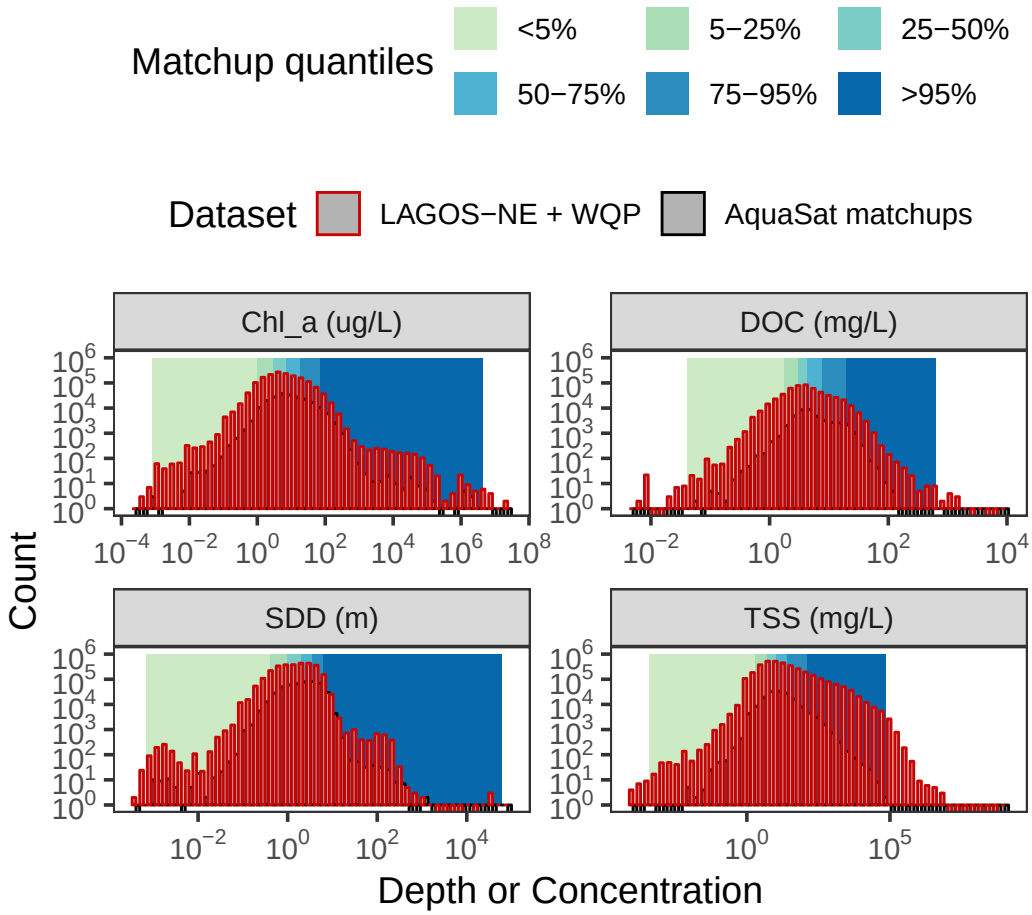

Figure 3. Data distributions for the *in-situ* data (gray) and the matchup data (red). Data quantiles are shown in the background as a color ramp from sage to blue, corresponding to the color scale in Figure 4. Quantiles were calculated for matchup data only.

occupying a relatively narrow range, within ˜1-2 orders of magnitude of the median (Fig. 3).

Based on decades of previous research (Topp et al, in review WRR), 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 dataset 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 waterbody types (Estuary, Stream, and Lake) and averaged for the entire USA. Despite using such a heterogeneous dataset, figure 4 shows clear systematic variation in spectral response for each parameter as concentration or SDD increases.

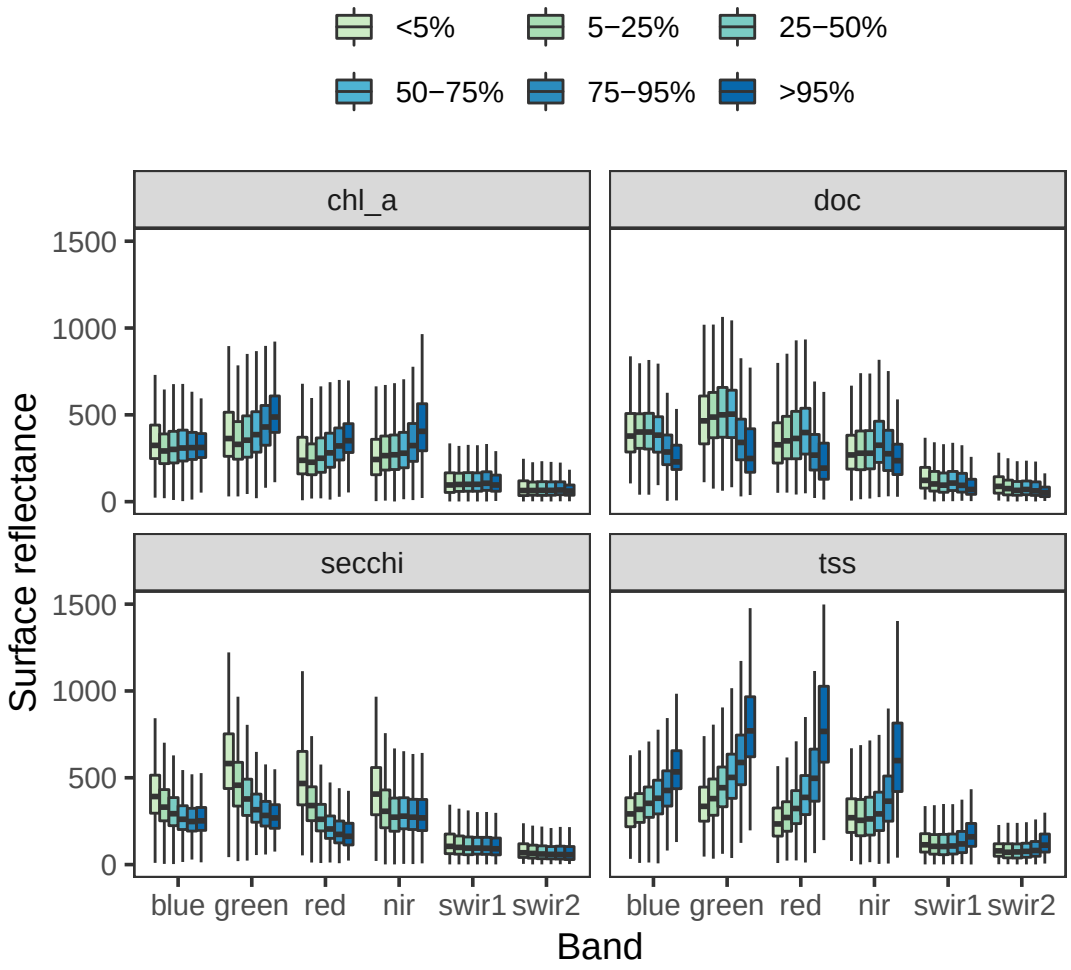

Figure 4. Shows spectral response, scaled and dimensionless, for each data quantile for each Landsat band. For Chl_a, DOC, and TSS, concentration increases moving from left to right for higher quantiles. For SDD, higher quantiles indicate increasing clarity or deeper SDD. The value range represented in each quantile are shown in figure 3.

4 Discussion

To our knowledge, the AquaSat dataset is the largest matchup dataset ever assembled for inland waters. Combining historical datasets 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 dataset 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 waterbodies, and we anticipate it will unlock many avenues for remote water quality work. The four parameters in AquaSat (DOC, Chl_a, SDD, TSS), within most of their range, have a significant impact on the observed surface reflectance, showing promising results for the value of this data to build predictive algorithms.

By publishing this 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 (Topp et al, in review WRR). For example, with an open, big dataset, method comparisons for predictive models can be conducted, accelerating scientific discovery (Bukata, 2013). We built this dataset to provide an easy way for non-experts in remote sensing to begin using it into their research. Because the matchup data covers the 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 USA. We expect that as remote predictions of inland water quality become more common, they can be used as a complement to *in-situ* datasets, vastly expanding our understanding of water quality trends and current status across the USA 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.

We built AquaSat to move towards continental-and global-scale remote sensing of water quality, but the dataset 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. 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 dataset 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 dataset, which was more extensively assured for quality, exemplifies a contrasting approach (Soranno et al., 2017, 2015). 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 (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.

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 towards 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, Costa, Simões Angélica, Araujo, & Maurice, 2006) or phosphorus (Kutser, Arst, Miller, Käärmann, & Milius, 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

The acknowledgments must list:

All code used to generate this dataset can be found at: https://github.com/GlobalHydrologyLab/watersat. The data for this paper come from the Landsat archive, LAGOSNE, and the Water Quality Portal. All of this data is free to download and the paper details extensively how we combined the datasets. Data generated by our work can be found at CUAHSI, Zenodo, and NASA archives. Part of this work was conducted at the Jzet Propulsion Laboratory, California Institute of Technology, under a contract with the National

Aeronautics and Space Administration. SNT was funded by a NASA NESSF grant (#80NSSC18K1398) for work on this grant. APA was supported by the Integration Information Dissemination Division of the U.S. Geological Survey. There are no conflicts of interest to report.

References

Allaire, J., Xie, Y., Mcphereson, J., Luraschi, J., Ushey, K., Atkins, A., ... Iannone, R. (2018). *rmarkdown: Dynamic Documents for R. R package version 1.11.* Retrieved from `https://rmarkdown.rstudio.com`

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

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

Barsi, J., Lee, K., Kvaran, G., Markham, B., Pedelty, J., Barsi, J. A., ... Pedelty, J. A. (2014, oct). The Spectral Response of the Landsat-8 Operational Land Imager. *Remote Sensing*, *6*(10), 10232–10251. Retrieved from `http://www.mdpi.com/2072-4292/6/10/10232` doi: 10.3390/rs61010232

Blondeau-Patissier, D., Gower, J. F., 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. Retrieved from `http://dx.doi.org/10.1016/j.pocean.2013.12.008` doi: 10.1016/j.pocean.2013.12.008

Booth, N. L., Everman, E. J., Kuo, I.-L., Sprague, L., & Murphy, L. (2011, oct). A Web-Based Decision Support System for Assessing Regional Water-Quality Conditions and Management Actions1. *JAWRA Journal of the American Water Resources Association*, *47*(5), 1136–1150. Retrieved from `http://doi.wiley.com/10.1111/j.1752-1688.2011.00573.x` doi: 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. Retrieved from `http://doi.wiley.com/10.4319/lo.1981.26.1.0043` doi: 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. Retrieved from `http://dx.doi.org/10.1016/j.jglr.2013.04.001` doi: 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. Retrieved from `http://www.pnas.org/lookup/doi/10.1073/pnas.1512651112http://www.pnas.org/content/113/1/58{\%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26699473{\%}5Cnhttp://www.pnas.org/content/113/1/58.full.pdf{\%}5Cnhttp://www.pnas.org/content/113/1/58.short{\%}5Cnhttp://www.pnas.or` doi: 10.1073/pnas.1512651112

Carlson, R. E. (1977, mar). A trophic state index for lakes1. *Limnology and Oceanography*, *22*(2), 361–369. Retrieved from `http://doi.wiley.com/10.4319/lo.1977.22.2.0361` doi: 10.4319/lo.1977.22.2.0361

Clarke, G. L., Ewing, G. C., & Lorenzen, C. J. (1970). Spectra of Backscat-

tered Light from the Sea Obtained from Aircraft as a Measure of Chlorophyll Concentration. *Science*, *167*(3921), 1119–1121. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.167.3921.1119{\%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/17829405 doi: 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. doi: 10.5194/bgd-12-9793-2015

Doxani, G., Vermote, E., Roger, J.-C., Gascon, F., Adriaensen, S., Frantz, D., ... Vanhellemont, Q. (2018, feb). Atmospheric Correction Inter-Comparison Exercise. *Remote Sensing*, *10*(3), 352. Retrieved from http://www.mdpi.com/2072-4292/10/2/352 doi: 10.3390/rs10020352

Feldman, S. I. (1979, apr). Make — a program for maintaining computer programs. *Software: Practice and Experience*, *9*(4), 255–265. Retrieved from http://doi.wiley.com/10.1002/spe.4380090402 doi: 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. doi: 10.1021/acs.est.5b03518

FitzJohn, R. (2018). *remake: Make-like build management. R package version 0.3.0.* Retrieved from 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. Retrieved from http://www.mdpi.com/1424-8220/16/8/1298 doi: 10.3390/s16081298

Gordon, H. R. (1997, jul). Atmospheric correction of ocean color imagery in the Earth Observing System era. *Journal of Geophysical Research: Atmospheres*, *102*(D14), 17081–17106. Retrieved from http://doi.wiley.com/10.1029/96JD02443 doi: 10.1029/96JD02443

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017, dec). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0034425717302900 doi: 10.1016/j.rse.2017.06.031

Griffin, C. G., Finlay, J. C., Brezonik, P. L., Olmanson, L., & Hozalski, R. M. (2018, nov). Limitations on using CDOM as a proxy for DOC in temperate lakes. *Water Research*, *144*, 719–727. Retrieved from https://www.sciencedirect.com/science/article/pii/S0043135418306286 doi: 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: Biogeosciences*, *116*(3), 1–12. doi: 10.1029/2010JG001634

Hirsch, R. M., & De Cicco, L. (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. Retrieved from http://pubs.usgs.gov/tm/04/a10/ doi: http://dx.doi.org/10.3133/tm4A10

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

Ju, J., Roy, D. P., Vermote, E., Masek, J., & Kovalskyy, V. (2012, jul). Continental-scale validation of MODIS-based and LEDAPS Landsat ETM+ atmospheric

correction methods. *Remote Sensing of Environment*, *122*, 175–184. Retrieved from https://www.sciencedirect.com/science/article/pii/S003442571200051X doi: 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*(10), 1–19. doi: 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. doi: 10.1016/0034-4257(71)90094-0

Kuhn, C., Valerio, A. d. M., Ward, N., Loken, L., Sawakuchi, H., Kampel, M., ... Butman, D. (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*, *Accepted*.

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

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

Lack, T. (2000). Eurowaternet- A Freshwater Monitoring and Reporting Network for All European Countries. In *Transboundary water resources in the balkans* (pp. 185–191). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-011-4367-7{_}19 doi: 10.1007/978-94-011-4367-7_19

Lee, B. Z., Arnone, R., Boyce, D., Franz, B., Greb, S., Hu, C., ... Wernand, M. (2018, may). Global Water Clarity : Continuing a Century-Long Monitoring. *Eos*, *99*(May), 1–10. doi: 10.1029/2018EO097251

Lorenzen, M. W. (1980, mar). Use of chlorophyll-Secchi disk relationships. *Limnology and Oceanography*, *25*(2), 371–372. Retrieved from http://doi.wiley.com/10.4319/lo.1980.25.2.0371 doi: 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. Retrieved from http://dx.doi.org/10.1016/j.rse.2011.09.022 doi: 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. doi: 10.1016/0034-4257(75)90008-5

Myers, D., & Ludtke, A. (2017). Progress and Lessons Learned from Water-Quality Monitoring Networks. In *Chemistry and water* (Vol. 33, pp. 23–120). Elsevier. Retrieved from http://dx.doi.org/10.1016/B978-0-12-809330-6.00002-7https://linkinghub.elsevier.com/retrieve/pii/B9780128093306000027 doi: 10.1016/B978-0-12-809330-6.00002-7

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

Palmer, S. C., Kutser, T., & Hunter, P. D. (2015, feb). Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sensing of Environment*, *157*, 1–8. Retrieved from https://www.sciencedirect.com/science/article/pii/S0034425714003666 doi: 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*(11). doi: 10.1029/2008WR007424

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

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., ... Winslow, L. A. (2017, feb). Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research*, *53*(2), 1735–1745. Retrieved from http://doi.wiley.com/10.1002/2016WR019993 doi: 10.1002/2016WR019993

Richardson, L. L. (1996, aug). Remote Sensing of Algal Bloom Dynamics. *BioScience*, *46*(7), 492–501. Retrieved from http://academic.oup.com/bioscience/article/46/7/492/322759/Remote-Sensing-of-Algal-Bloom-DynamicsNew-research doi: 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. Retrieved from 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, apr). Low-level addition of dissolved organic carbon increases basal ecosystem function in a boreal headwater stream. *Ecosphere*, *8*(4), e01739. Retrieved from http://doi.wiley.com/10.1002/ecs2.1739 doi: 10.1002/ecs2.1739

Secchi, P. (1864, dec). 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. Retrieved from http://link.springer.com/10.1007/BF02726911 doi: 10.1007/BF02726911

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

Soranno, P. A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., ... Yuan, S. (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. doi: 10.1093/gigascience/gix101

Soranno, P. A., Bissell, E. G., Cheruvelil, K. S., Christel, S. T., Collins, S. M., Fergus, C. E., ... Webster, K. E. (2015, dec). Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. *GigaScience*, *4*(1), 28. Retrieved from https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0067-4 doi: 10.1186/s13742-015-0067-4

Spencer, R. G. M., Stubbins, A., Hernes, P. J., Baker, A., Mopper, K., Aufdenkampe, A. K., ... Six, J. (2009). Photochemical degradation of dissolved organic matter and dissolved lignin phenols from the Congo River. *Journal of Geophysical Research: Biogeosciences*, *114*(3), 1–12. doi: 10.1029/2009JG000968

Sprague, L. A., & Lorenz, D. L. (2009, may). Regional nutrient trends in streams and rivers of the United States, 1993-2003. *Environmental Science and Technology*, *43*(10), 3430–3435. Retrieved from http://pubs.acs.org/doi/abs/10.1021/es803664x doi: 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. Retrieved from http://dx.doi.org/10.1016/j.watres.2016.12.024 doi: 10.1016/j.watres.2016.12.024

Srebotnjak, T., Carr, G., de Sherbinin, A., & Rickwood, C. (2012, jun). A global Water Quality Index and hot-deck imputation of missing data. *Ecological*

Indicators, *17*, 108–119. Retrieved from https://www.sciencedirect.com/science/article/pii/S1470160X1100104X doi: 10.1016/J.ECOLIND.2011.04.023

Stets, E., & Striegl, R. (2012, oct). Carbon export by rivers draining the conterminous United States. *Inland Waters*, *2*(4), 177–184. Retrieved from http://www.tandfonline.com/doi/full/10.5268/IW-2.4.510 doi: 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. *PECORA 16 Conference Proceedings, Sioux Falls, South Dakota*, 23–27. Retrieved from http://www.asprs.org/a/publications/proceedings/pecora16/Storey{_}J.pdf

Syvitski, J. P. M., & Kettner, A. (2011, mar). Sediment flux and the Anthropocene. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *369*(1938), 957–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21282156 doi: 10.1098/rsta.2010.0329

Telmer, K., Costa, M., Simões Angélica, R., Araujo, E. S., & Maurice, Y. (2006, oct). 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. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0301479706001009 doi: 10.1016/j.jenvman.2005.09.027

Vähätalo, A. V., Wetzel, R. G., & Paerl, H. W. (2005, mar). 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. Retrieved from http://doi.wiley.com/10.1111/j.1365-2427.2004.01335.x doi: 10.1111/j.1365-2427.2004.01335.x

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016, nov). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, *185*, 46–56. Retrieved from https://www.sciencedirect.com/science/article/pii/S0034425716301572 doi: 10.1016/J.RSE.2016.04.008

Williams, G. P. (1989, jan). Sediment concentration versus water discharge during single hydrologic events in rivers. *Journal of Hydrology*, *111*(1-4), 89–106. Retrieved from https://www.sciencedirect.com/science/article/pii/0022169489902540 doi: 10.1016/0022-1694(89)90254-0

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

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