A national, multi-decadal, water color and landsat dataset

Matthew Ross and lots of others!

14 August, 2018

Table of Contents

# Introduction

Since the beginning of the Landsat missions, the limnologists, oceanographers, and hydrologists have been interested in developing universal algorithms for extracting water quality information from remotely sensed images (Holyer 1978,Ritchie, Schiebe, and McHENRY (1976),Maul and Gordon (1975),Klemas, Borchardt, and Treasure (1973),Clarke, Ewing, and Lorenzen (1970)). Since these early efforts there has been almost fifty years of work with the basic goal of using spectral information to predict water quality parameters like total suspended sediment (TSS), Chlorophyll a (Chl.a), colored dissolved organic matter, and secchi disk depth (SDD). However, progress towards universal algorithms and unified approaches has been slow (**???**,Blondeau-Patissier et al. (2014),Gholizadeh, Melesse, and Reddi (2016)), with most papers published focusing on developing predictive methods as opposed to using predictions to interrogate process that control water quality dynamics [Topp2018]. Much of this slow evolution in methods and approaches comes from the inherent optical complexity of inland waters, where spectral signatures are the result of a complex mixture of inorganic sediment, organic sediment, algae, dissolved organic matter, and other constituents. Compared to oceanic remote sensing of water quality which benefits from robust, shared datasets of *in-situ* data paired with satellite overpass reflectance (Blondeau-Patissier et al. 2014,(**???**)), progress on inland water algorithms is further impeded by the lack of a shared overpass dataset. Such data could go a long way towards, if not the holy -grail of universal predictive algorithms, at least towards more unified approaches tested on a universal dataset. Here, we create and share the largest such overpass dataset ever assembled for inland waters by using Google Earth Engine (Gorelick et al. 2017) Landsat archive data from 1984-2018 with data from the Water Quality Portal (**???**) and phase one of the “lake multi-scaled geospatial and temporal database (LAGOSNE)”(**???**) for the conterminous USA and Alaska. Joining these datasets provides us with an unprecedented resource to model, predict, and understand the long-term and large-scale dynamics of variation in four key water quality constituents: TSS, SDD, Chl.a, and dissolved organic carbon (DOC). We also outline and share our approach, code, and intermediate data for bringing these three free datasets together; generating a high-graded analysis-ready dataset for remote sensors of water quality.

## Parameter description?

Wondering if this deserves it’s own section or just a reference to Topp2018. For now I’m assuming that this will be explicitly cast as sort of partB to that paper, and I am not writing out detailed explanations, but can easily add this section.

## Data source description

Combining *in-situ* data with Landsat reflectance information first requires a large repository of water quality samples, which increases the likelihood that a given sample happened to be taken on the same day as a Landsat overpass. For this paper, we focused on two databases of water quality. The first, the Water Quality Portal (WQP) has tens of millions of water observations in all types of inland surface waters, but there is no entity that harmonizes and cleans the data for quality (**???**). The second dataset we used, LAGOS, currently only covers lakes in the northeastern United States, with plans to expand and cover lakes across the entire USA (**???**). While LAGOS has less data than the WQP, a group of dedicated researchers has spent years combing through the data and ensuring data quality, making it a more analysis-ready dataset (**???**). These similar but contrasting datasets, one with more quantity (WQP) and the other with more quality assurances (LAGOS), ensures that our dataset covers the broadest possible number of waterbodies with the option of limiting analyses to only the highest quality subset.

### Water Quality Portal

The WQP is the largest dataset of water observations ever assembled with more than 290 million observations at 2.7 millions sites mostly in the USA, with data dating back more than a century (**???**). The WQP continuously gathers water quality information from more than 450 organizations including academic, government, NGO, tribal, and state datasets (**???**). These datastreams are gathered and distributed in a standardized format, making analysis across different collection methods more readily available. Yet, the diversity of data sources and variation in meta-data quality brings about some significant challenges to directly using the WQP as a analysis-ready dataset (**???**). Instead end-users of the data must carefully harmonize data across sampling methods, analytic approaches, and 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 more shallow less time-consuming but potentially more error-prone data quality check.

### LAGOS-NE

The LAGOS project was, in part, meant as a direct way to address some of the problems inherent to the WQP, with the explicit goal of building a publically available high-quality dataset for continental-scale lake analyses (**???**). In addition to pairing *in-situ* lake data with physical lake characteristics and local geologic setting, LAGOS researchers standardized key water quality measurements across the 87 water quality datasets that they gathered (**???**). In it’s current form, the LAGOS dataset covers only lakes in the northeast and midwest, two lake-rich regions of the USA. LAGOS provides an end-member dataset of the highest quality for matching *in-situ* data to Landsat overpasses.

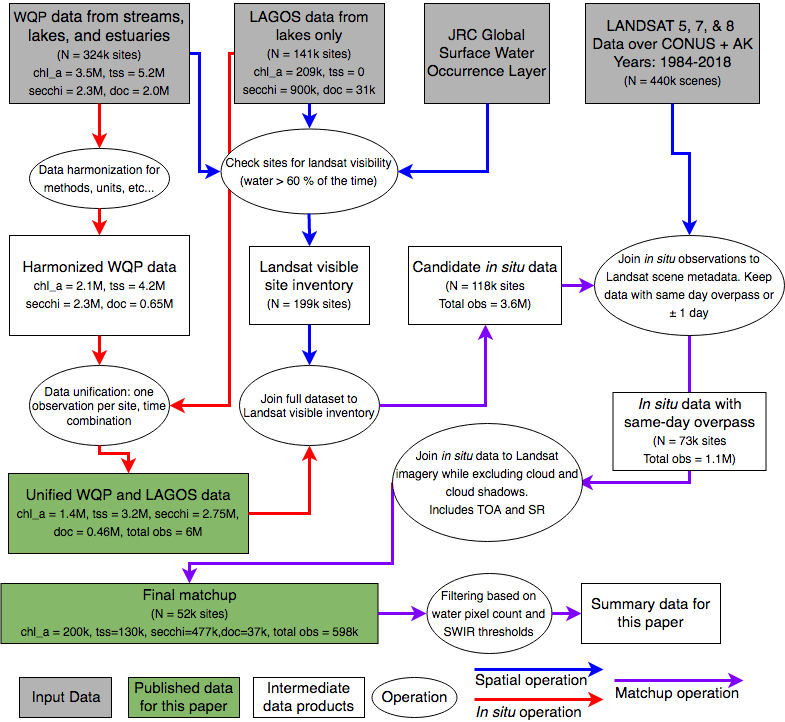
### Landsat

For this project, we join these two *in-situ* datasets with the Landsat data archive for Landsat missions 5, 7, and 8. The Landsat missions 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 (**???**). For this project we are only using the three most recent Landsat mission datasets: Landsat 5 with coverage from 1984-2012 and over 192745 available images; Landsat 7 which is still collecting data after launching in July of 1999 with 190976 images; and finally Landsat 8 which launched in November, 2013 still adding to its collection of 61255 images. Generally, these satellites complete a full imaging of the globe every sixteen days, except for the most polar regions (**???**,(**???**)), meaning that for most of the USA, a given spot will be imaged at least every sixteen days, and-when two missions are running at the same time- every eight days. All three satellites use different imagers to collect spectral information in the visible and infrared wavelengths.

# Data integration

## *In situ* data pull.

## WQP Parameters



Dataset generation

Landsat 5 has an onboard imager that collects seven bands of imagery centered on three visible wavelengths (blue, green, and red) and four infrared (near infrared, shortwave infrared 1,) ## Rivers, Lakes, and Estuaries/Deltas

## Water Quality Portal

### data pull and parameters therein

### Data harmonization approach and link to code output

## LAGOSNE

### Describe Lagos daasets

### In Situ data unification

## Joining landsat and water quality portal

### Google Earth Engine

### How we selected sites (pekel occurence)

### Diagram of joining procedures and counts of observations dropped

## Data quality flagging

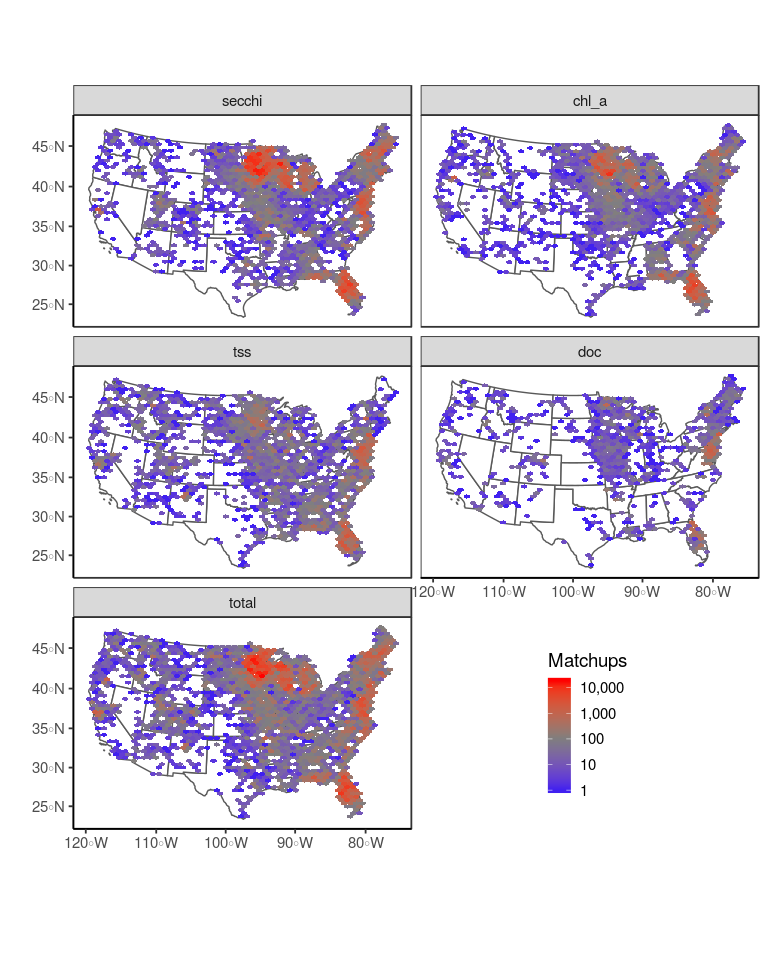
### Not sure what to put here or if we should have this section

# Results

For LAGOSNE data see [here](https://lagoslakes.org/the-lagos-database/)

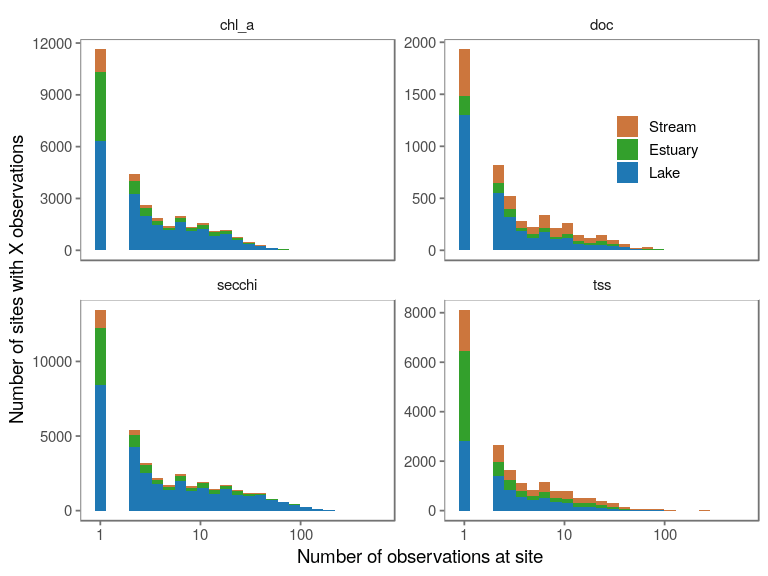
## Dataset description

### Map

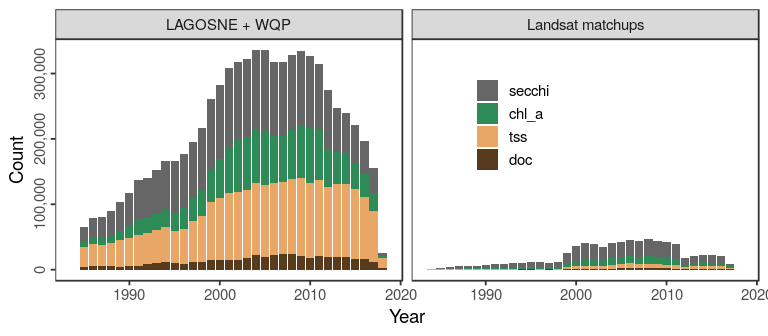


Distribution of observations across the conterminous USA. The data is split by observation type, where total represents an overpass for any of the four primary parameters

### Distribution of observations per site

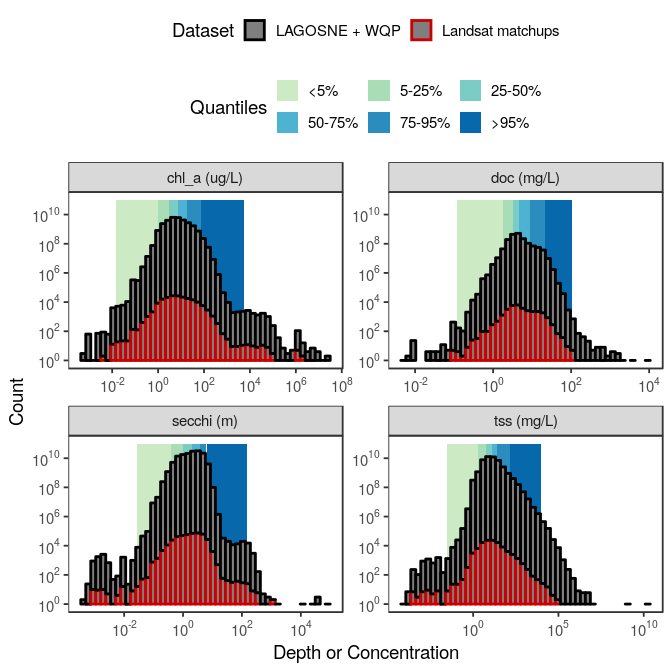


### Observations over time



## Variation captured by the datasets

### Distributions of in situ vs matchup datasets

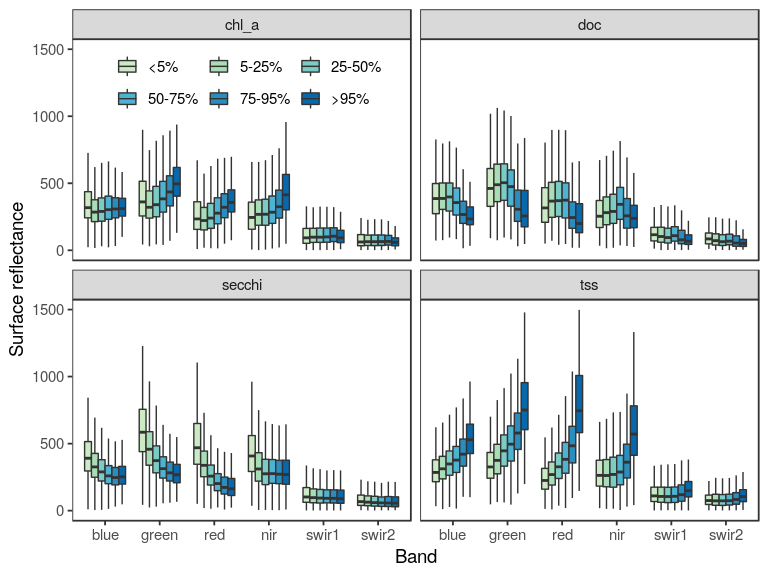


### Observations lost

For both DOC and TSS our matchup dataset is missing the long tail of data in the in situ dataset. What kind of sites were dropped to create this discrepancy? They are basically all streams.

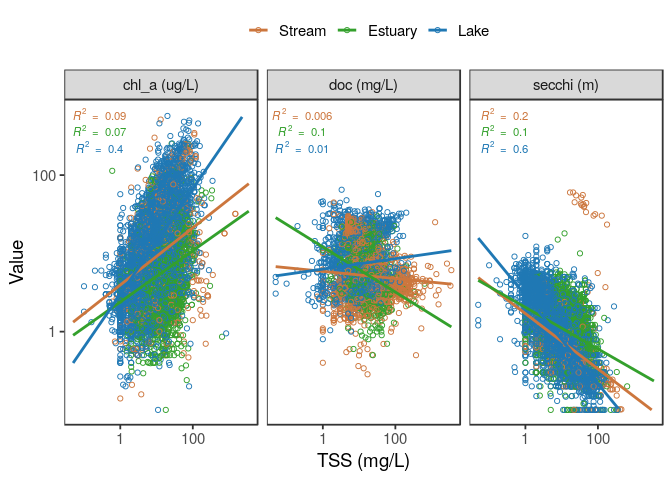
##   
## Estuary Lake Stream   
## 6 56 12881

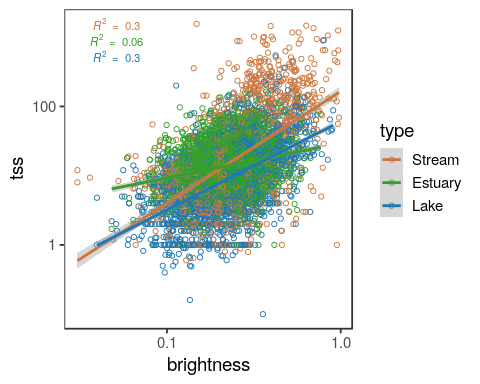
### Spectral variation



## Partitioning variation by sediment concentration and region

### TSS covarying with other constituents





Blondeau-Patissier, David, James F.R. Gower, Arnold G. Dekker, Stuart R. Phinn, and Vittorio E. Brando. 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. Elsevier Ltd: 23–144. doi:[10.1016/j.pocean.2013.12.008](https://doi.org/10.1016/j.pocean.2013.12.008).

Clarke, G. L., G. C. Ewing, and C. J. Lorenzen. 1970. “Spectra of Backscattered Light from the Sea Obtained from Aircraft as a Measure of Chlorophyll Concentration.” *Science* 167 (3921): 1119–21. doi:[10.1126/science.167.3921.1119](https://doi.org/10.1126/science.167.3921.1119).

Gholizadeh, Mohammad, Assefa Melesse, and Lakshmi Reddi. 2016. “A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques.” *Sensors* 16 (8): 1298. doi:[10.3390/s16081298](https://doi.org/10.3390/s16081298).

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. “Google Earth Engine: Planetary-scale geospatial analysis for everyone.” *Remote Sensing of Environment*. doi:[10.1016/j.rse.2017.06.031](https://doi.org/10.1016/j.rse.2017.06.031).

Holyer, Ronald J. 1978. “Toward universal multispectral suspended sediment algorithms.” *Remote Sensing of Environment* 7 (4): 323–38. doi:[10.1016/0034-4257(78)90023-8](https://doi.org/10.1016/0034-4257(78)90023-8).

Klemas, V., J. F. Borchardt, and W. M. Treasure. 1973. “Suspended sediment observations from ERTS-1.” *Remote Sensing of Environment* 2: 205–21. doi:[10.1016/0034-4257(71)90094-0](https://doi.org/10.1016/0034-4257(71)90094-0).

Maul, George A., and Howard R. Gordon. 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](https://doi.org/10.1016/0034-4257(75)90008-5).

Ritchie, JC, FR Schiebe, and JR McHENRY. 1976. “Remote sensing of suspended sediments in surface waters.” *American Society of* 42 (12): 1539–45. <https://trid.trb.org/view.aspx?id=66674>.