Analysis of Dissolved Organic Content and Water-Color Trends within Inland Lakes Using Satellite Remote Sensing

REU Scholar: Aisha Malik[1],

IUSE Scholar: Fahmeda Khanom[2],Touheda Khanom[2]

Mentor: Dr. M. Azarderakhsh[3]

[1]Computer Science, Hunter College, CUNY
[2] Computer Engineering, New York City College of Technology, CUNY
[3] Construction Management and Civil Engineering, New York City College of Technology, CUNY

ABSTRACT

In the 21st century, climate change stands as a formidable threat to New York State's natural resources, including over 3000 lakes and their watersheds. The lakes within the Adirondack Park have benefited from strict land use laws due to the Clean Air Act to control the anthropogenic impacts on the lakes' health. In the late 1970s, the Adirondack Lake Survey (ALS) took extensive field data including water chemistry to better understand the health of more than 30 represented lakes to monitor the recovery from acidification. The impact of climate change raises concerns about the lake ecosystem and there is a significant data gap in monitoring beyond the selected lakes.

Here, we seek to explore spatial and temporal patterns of water chemistry data, leveraging historic and long-term monitoring programs and integrating satellite remote sensing (RS) imagery to inform future research on watercolor change as an indicator of water quality beyond the selected sampling lakes through a future Survey of Climate change and Adirondack Lake Ecosystems (SCALE) Pilot Program. A dataset with sampling since 1985 is compiled for 50 lakes and Dissolved Organic Content (DOC) is selected for comparison with several empirical algorithms using Landsat-5 and Landsat-8 surface reflectance (SR) observations.

While many studies develop predictive relationships between RS-SR and water parameters, these relationships are often limited to a specific geographic region and have little applicability in other areas. The preliminary data analysis of several algorithms does not show a strong correlation for the represented ALTM lakes. However, they exhibit consistent long-term trends using Landsat-5 SR data suggesting lake color change in several of the sampled lakes.

A machine-learning (ML) approach can learn the complex relation between the inputs and data types. Several ML techniques are employed to estimate DOC using different input features including SR. This analysis is performed on an openly accessible Python script on the Google Earth Engine Platform for processing cloud-based publicly available data and will allow the determination of relationships over various lakes and studying the impact of climate change within the larger Adirondack region.

Keywords: Dissolved Organic Content (DOC), Landsat-5, Landsat-8, Multispectral Instrument, Colored Dissolved Organic Matter (CDOM), Machine Learning algorithms, Random Forest Regression, Lake water quality, Climate Change, Remote Sensing, Google Earth Engine (GEE)

1. INTRODUCTION

Dissolved Organic Content (DOC) and its colored subset, Colored Dissolved Organic Matter (CDOM), are key determinants that shape the dynamics of lake ecosystems. CDOM has two primary sources: (1) allochthonous, originating from the decomposition of woody plants on land, and (2) autochthonous, arising from the breakdown of algae and aquatic vegetation within the water. Within this environment, CDOM, characterized by its ability to absorb visible light, gives water coloration. CDOM contributes to ecological and biogeochemical processes that are important for the functioning of lake ecosystems (Brezonik et al., 2015).

The relationship between CDOM and watercolor holds major significance in assessing water quality. Variations in watercolor, which are tied to shifts in CDOM concentration levels, offer insight into alterations in Dissolved Organic Content (DOC). However, this correlation cannot be proven as increases in DOC do not always indicate increases in CDOM (Brezonik et al., 2015).

To observe the relationships between CDOM, DOC, and watercolor, the integration of remote sensing technologies is a powerful tool for monitoring their spatial and temporal dynamics within lake ecosystems (Koll-Egyed et al., 2021; Olmanson et al., 2020). Merging historical data with modern remote sensing capabilities can allow for greater comprehension of these dynamics.

In this study, remote sensing techniques and technologies are utilized for the investigation and monitoring of lake water bodies. Data is acquired from two satellites, Landsat-5 and Landsat-8. Through the implementation of an empirical formula and an algorithm, values of CDOM are retrieved from remote sensing data. A goal of this research is to compare in-situ DOC data with CDOM estimated values obtained from Landsat-5 and Landsat-8 satellite observations to create a dataset spanning from 1985 to the present. This study evaluates shifts in watercolor as a marker for water quality beyond the selected sampling lakes, through the future Survey of Climate Change and Adirondack Lake Ecosystems (SCALE) Pilot Program.

To achieve this goal, Google Earth Engine Python API is employed for the development of an algorithm. Additionally, machine learning algorithms are integrated for the predictive extrapolation of in-situ DOC measurements. This data was collected from the Adirondack Long Term Monitoring Program (ALTM), conducted by the Adirondacks Lake Survey Corporation (ALSC). This effort is aimed at enhancing the understanding of over 50 represented lakes within the Adirondack region and their recovery from acidification.

2. BACKGROUND

2.1. Correlation Between CDOM and DOC

In studies that use remote sensing for CDOM measurement, a common assumption made is that the findings can be used to estimate DOC concentrations. This assumption, however, is not always true. Brezonik et al. (2015) conducted field measurements in the summer of 2013, depicting a strong correlation between DOC and CDOM, with a large R^2 value of 0.925, indicating a possible direct relationship. However, concerns arose regarding the reliability of using CDOM values, as indicators for predicting DOC concentrations.

To make accurate conclusions regarding CDOM and DOC correlations, it is crucial to consider outside drivers. Brezonik et al. (2015) proposed that factors such as human activities can influence DOC levels and, as a result, impact the CDOM-DOC relationship. For instance, In situations where water

bodies heavily impacted by human activities, resulting in DOC with low color, are excluded from analysis, a stronger CDOM-DOC relationship can emerge.

Studies conducted across various regions have consistently shown strong positive correlations between CDOM and DOC. Huang et al. (2023) observed a strong correlation (R^2 = 0.9) between CDOM absorption and DOC concentration in the Pearl River Estuary (PRE) in China. Similarly, Sun et al. (2021) conducted a study in Southern Canada and identified a significant positive correlation between CDOM and DOC levels in lakes within the region. These findings support the idea that CDOM can serve as an indicator for DOC in water bodies, highlighting the significance of CDOM measurements in evaluating water quality.

2.2. CDOM Estimation Methods and Algorithms

There are a variety of strategies used to extract CDOM measurements from remote sensing data, including empirical, semi-analytical, matrix inversion, and optimization methods (Brezonik et al., 2015). However, Zhu et al. (2014) cautions that these methods can sometimes generate unreliable results; some algorithms may lead to underestimations in high CDOM conditions, while others may overestimate in low CDOM conditions. The unique characteristics of CDOM contribute to this challenge of obtaining accurate CDOM measurements from satellite imagery. CDOM can absorb light without causing large amounts of scattering or reflection, resulting in weak light signals in areas abundant in CDOM. By interacting with minerals, algae, and other suspended particles in water bodies, the likelihood of light scattering increases, further reducing the ability to obtain accurate CDOM measurements. As CDOM displays a gradual, quasi-exponential decline in absorption as the wavelength increases, and as there is an absence of distinct wavelength bands directly associated with CDOM in the visible spectrum, retrieving precise CDOM measurements in inland water bodies is made greatly difficult (Brezonik et al., 2015).

Researchers have developed algorithms and models specifically for CDOM estimation to address the challenges of extracting accurate CDOM measurements. Kutser et al. (2005) developed a band-ratio type algorithm that leveraged ALI band 2 and band 3 to estimate CDOM content in lakes. Similarly, Olmanson et al. (2016) found strong models, using the widely used green/red band ratio model, with R^2 values of 0.79 for Landsat 7 and 0.81 for CDOM estimation. Several studies, including Kim et al. (2022) and Huang et al. (2023), have also employed band ratio CDOM retrieval algorithms. These band selection and feature extraction techniques attempt to derive the best CDOM estimation results from remote sensing data.

Stepwise regression is also a significant technique for extracting CDOM information from satellite data. This approach involves finding the most suitable band or band ratio model, where parameters such as the natural log (ln) transformed CDOM act as the target variable, while different bands and band ratios act as independent variables. The model produced good results, with R^2 values ranging from 0.84 to 0.86 for CDOM estimation using Landsat 8, Sentinel-2, and Sentinel-3 bands (Brezonik et al., 2015). Similarly, Olmanson et al. (2016) employed stepwise regression to identify the best Secchi Depth model, incorporating bands and band ratios, even considering new ultra-blue and narrower NIR bands. This strategy aimed to improve CDOM retrieval from satellite imagery.

Empirical models have also been used for CDOM estimation. Martins et al. (2018) successfully retrieved CDOM information from TM/Landsat-5 spectral band configurations using empirical algorithms. The study's best-adjusted model achieved an R^2 value of 0.91, signifying the potential of empirical approaches for CDOM estimation. Semi-analytical algorithms such as the Quasi-Analytical Algorithm (QAA) and Generalized Inherent Optical Property (GIOP) algorithm have also been implemented for CDOM concentration estimation (Aurin et al., 2018).

2.3. Remote Sensing of CDOM Seasonal and Temporal Variation

Temporal fluctuations in CDOM levels have been observed in various studies. Brezonik et al. (2015) revealed significant temporal variations in CDOM concentrations within Florida lakes over a multi-year sampling period. While some lakes exhibited stable CDOM levels for long periods of time, others displayed large changes over shorter intervals. Brezonik et al. (2015) explained that large variations in CDOM are likely driven by factors such as rainfall and runoff. However, the complexity of hydrological systems made it challenging to establish direct relationships between rainfall amounts and CDOM levels. Brezonik et al. (2015) cautioned that unless temporal stability in CDOM levels is confirmed, data collection should be performed close to the image acquisition date, ideally within 1–2 months for lakes and even more quickly for rivers and streams, especially if there are expected changes in precipitation conditions.

Olmanson et al. (2020) conducted an analysis of CDOM levels to reveal trends in CDOM concentrations in different ecoregions (or biomes) across the state of Minnesota, USA, using Landsat and Sentinel satellite imagery. Lakes located in agricultural ecoregions exhibited a tendency towards decreasing CDOM levels between 2015 and 2016, likely attributed to dilution by rainfall. In contrast, lakes located in ecoregions characterized by high forest and wetland cover experienced increases in CDOM, indicating CDOM runoff from forested wetlands (Olmanson et al., 2020).

2.4. Machine Learning Approaches for CDOM Estimation and Prediction

Various machine learning algorithms and methods have been implemented in previous studies for CDOM estimation and prediction. Koll-Egyed et al. (2021) employed a Random Forest model, utilizing median band values from multiple summer seasons, which yielded an adjusted R^2 of 0.70. The models constrained within a 30-day time window exhibited lower performance, indicating the importance of using extensive sets of images to improve model accuracy when analyzing larger regions. Kim et al. (2022) found that the Random Forest model outperformed other models in CDOM estimation, achieving an R^2 value of 0.85, indicating the effectiveness of the Random Forest algorithm in capturing CDOM variability.

A series of studies focusing on CDOM estimation and prediction implemented a large scope of machine-learning techniques that resulted in varying performances. Huang et al. (2023) employed six machine-learning techniques for CDOM absorption estimation, with XGBoost emerging as the top-performing algorithm with an R^2 value of 0.9. Machine learning models Gaussian Process Regression (GPR) and Support Vector Regression (SVR) were used by Keller et al. (2018) for CDOM estimation, achieving R^2 values ranging from 0.899 to 0.946, with multiple variables used as input. Sun et al. (2021) employed machine learning techniques including Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting Decision Tree (XGBoost), Convolutional Neural Network (CNN), K-nearest Neighbor Regression (KNN), and Multi-Layer Perceptron (MLP) to estimate CDOM absorption coefficients. Results showed that Gaussian Process Regression (GPR) exhibited higher stability and estimation accuracy (R^2 = 0.74) compared to other models. Ruescas et al. (2018) investigated machine learning methods applied to Sentinel-2 and Sentinel-3 simulated reflectance data for CDOM retrieval. The Random Forest Regression (RFR) model performed well, with R^2 = 0.992. Additionally, the regularized linear regression (RLR) method exhibited improved results when considering all reflectance bands and ratios. Aurin et al. (2018) utilized machine learning models, including Random Forest tree-bagger (RFTB) and Multivariate Linear Regression (MLR), for CDOM retrievals. In this study, however, the empirical approaches slightly overestimated CDOM in waters with low CDOM and underestimated CDOM in waters with high CDOM, suggesting that these methods are not very responsive which is evident in their generated low R^2 values.

While machine learning methods show potential in estimating and predicting CDOM concentrations, it is important to recognize their limitations. Firstly, the accuracy of some machine learning models can differ based on the quantity and quality of the input data. Secondly, careful data processing may be required since certain methods might be sensitive and reactive towards outliers or “noise” in the data. Thus, the choice of a suitable machine learning model should align with the dataset's unique elements and research goals.

3. METHODS & MATERIALS

3.1. Study Sites

Data is collected from Adirondack Park in upstate New York (Fig. 1). Adirondack Park is a significant study area due to its unique characteristics and ecological diversity. The park encompasses over six million acres, making it one of the largest protected areas in the United States. Adirondack Park is known for its extensive network of more than 3,000 lakes and ponds, in addition to thousands of miles of rivers and streams. The park's geography ranges from mountains to lowland regions, providing a variety of lake ecosystems (*More About the Adirondack Park*, n.d.). This variability in landscape contributes to diverse lake sizes and CDOM content, making it an ideal location for the investigation of DOC, CDOM, and watercolor trends.

To assess CDOM, DOC, and watercolor trends over this region, satellite remote sensing imagery and field data are acquired over different periods of time for 50 sampled lakes across the Adirondacks. The data collected from Adirondack Park contributed to the development of the empirical algorithm.

Figure 1. Location of Adirondack Park in New York and locations of major lakes in the region. Map by Department of Environmental Conservation (DEC) (*Adirondack Park Campground Map - NYS Dept. of Environmental Conservation*, n.d.).

3.2. In-situ Measurements

3.2.1. ALTM Data

In-situ measurements were carried out in 50 lakes (and 50 monitoring stations) in Adirondack Park since 1985 by the Adirondack Long-Term Monitoring Program (ALTM) (Table 1). This ALTM program is supported by organizations including the New York State Energy Research Development Authority (NYSERDA), the New York State Department of Environmental Conservation (NYSDEC), and the United States Environmental Protection Agency (USEPA) (Corporation, n.d.). ATLM data provides DOC measurements from 1982 to 2021.

3.2.2. ALAP Data

The Adirondack Lake Assessment Program (ALAP) was launched by the Protect the Adirondacks (PROTECT) and the Paul Smith's College Adirondack Watershed Institute (AWI) organizations in 1998. The program began with nine sample lakes and grew to consistently monitoring approximately over 150 lakes across the Adirondack Park region. In-situ measurements of CDOM, DOC, Chl-a, and other water chemistry parameters are available from 1995 to 2022 for the sampled lakes. Additional parameters including lake transparency, trophic state, and maximum depth are also reported in the reports by ALAP.

Table 1.
Table of 50 ALTM and ~150 ALAP sampled lakes with lake coordinates, classification, lake depth, surface area, and trophic state.

Name	LAT	LON	Surface Area (ha)	Lake Depth	Classification	Trophic State
Alford Pond	44.26168	-74.03662	15.5	0.6	-	Mesotrophic
Amber Lake	44.40034	-74.61754	45	2.1	-	Eutrophic
Arbutus Pond	43.98787	-74.24170	48	-	MTD	Oligotrophic
Augur Lake	44.46067	-73.49256	146	6.4	-	Mesotrophic
Austin Pond	43.67642	-73.96434	9	-	-	Mesotrophic
Avalanche Lake	44.13287	-73.96680	4.4	7	MTD	-
Barnes Lake	43.56611	-75.22690	-	-	-	-
Bartlett Pond	44.10615	-73.51104	40	6.1	-	Oligotrophic
Big Cherry Patch Pond	44.29080	-73.94441	7	4.6	-	Mesotrophic
Big Hope Pond	44.51369	-74.12680	-	-	-	-
Big Moose Lake	43.83221	-74.84650	512.5	21.3	ThinTD	Oligotrophic
Black Pond	44.30749	-74.38150	10.1	13.4	ThickTD	Oligotrophic
Blue Mountain Lake	43.86476	-74.46158	500	30.5	-	Oligotrophic
Bone Pond	44.36123	-74.30440	5.6	-	-	Oligotrophic
Brandreth Lake	43.91803	-74.70441	362	54	-	Oligotrophic
Brook Trout Lake	43.60298	-74.66310	-	-	-	-
Butternut Pond	44.43070	-73.49566	65.8	6.3	-	Mesotrophic
Canada Lake	43.16811	-74.52288	294	45.7	-	Oligotrophic
Carry Pond	43.68229	-74.48870	2.8	4.6	MS	-
Cascade Lake	43.79039	-74.80240	40.4	6.1	MTD	-
Catlin Lake	44.02284	-74.26950	261	-	-	Oligotrophic
Chapel Pond	44.13971	-73.74798	8	23.8	-	Oligotrophic
Chase's Lake	43.75775	-75.30333	47	-	-	Mesotrophic
Chazy Lake	44.75867	-73.81291	746.6	21.9	-	Oligotrophic
Clear Pond	44.48658	-74.16071	42.1	16.8	ThickTD	Oligotrophic
Connery Pond	44.31184	-73.93404	34.3	15.2	-	Oligotrophic
Constable Pond	43.83290	-74.79820	-	-	-	-
Copperas Pond	44.31398	-74.37626	10	5.8	-	Oligotrophic
Cranberry Lake	44.16540	-74.80336	2750	11.6	-	Mesotrophic
Dart Lake	43.79548	-74.86410	51.8	17.7	ThinTD	-
Deer Lake	44.03999	-74.25246	38	-	-	Oligotrophic
Eagle Lake	43.84747	-74.47866	64	9.4	-	Oligotrophic
East Caroga Lake	43.12495	-74.48064	94	12.4	-	Oligotrophic
East Pine Pond	44.33900	-74.41904	27.1	10.1	-	Oligotrophic
Echo Lake	44.29718	-73.96365	7	1	-	Mesotrophic
Eighth Lake	43.77880	-74.70198	-	-	-	-
Eli Pond	43.74821	-73.82918	9	-	-	Mesotrophic
Fawn Lake	43.48825	-74.45545	-	-	-	-
Fern Lake	44.48869	-73.71850	172.4	-	-	Mesotrophic
Fifth Lake	43.74807	-74.79209	-	-	-	-
Fish Creek East	44.30404	-74.35174	35	5.3	-	Mesotrophic
Fish Creek Pond	44.30344	-74.37264	85.7	15.6	-	Mesotrophic

Fish Creek West	44.29867	-74.35952	30.3	9.3	-	Mesotrophic
Floodwood Pond	44.33385	-74.40375	94.4	9.5	-	Mesotrophic
Follensby Clear Pond	44.31910	-74.34692	200.4	18.3	-	Oligotrophic
Fourth Lake	43.75851	-74.84068	-	-	-	-
Frank Pond	43.86185	-74.16343	10	-	-	Oligotrophic
Franklin Falls Reservoir	44.41941	-73.98986	181.6	6.1	-	Mesotrophic
Friends Lake	43.62624	-73.84389	-	-	-	Oligotrophic
G Lake	43.41421	-74.63250	32.2	9.8	ThinTD	-
Garnet Lake	43.51868	-74.02228	133	-	-	Oligotrophic
Gordon Pond	44.34215	-74.34085	2.1	-	-	Oligotrophic
Grass Pond	43.69207	-75.06170	5.3	5.2	MTD	-
Grass Pond (3)	44.65688	-74.49560	4.4	-	-	-
Green Lake	43.17572	-74.50664	-	-	-	Mesotrophic
Green Pond	44.33968	-74.33714	26.1	18.3	-	Mesotrophic
Gull Pond	44.21073	-74.52914	117	23.2	-	Oligotrophic
Heart Lake	44.18229	-73.96920	10.7	16.8	MTD	-
Heavens Lake	44.13068	-74.42652	-	-	-	-
Hewitt Lake	44.21073	-74.52914	166	16.5	-	Oligotrophic
Hidden Lake	43.38357	-73.76335	8	7.6	-	Mesotrophic
Highlands Forge Lake	44.41010	-73.44475	50.3	-	-	Mesotrophic
Hoel Pond	44.35078	-74.35509	187.1	24.2	-	Oligotrophic
Holcomb Pond	44.29140	-73.92192	11.8	0.6	-	Oligotrophic
Horseshoe Pond	44.32109	-74.35737	35.6	7.9	-	Mesotrophic
Indian Lake	43.61745	-74.75570	33.2	10.7	ThinTD	-
Indian Lake- Franklin County	43.68557	-74.33000	134	4.9	-	Mesotrophic
Indian Lake- Hamilton County	44.71626	-74.13451	2155	25.9	-	Mesotrophic
Irving Pond	43.16633	-74.47055	-	-	-	Mesotrophic
Jabe Pond	43.70337	-73.53962	59.3	22.9	-	Mesotrophic
Jockeybush Lake	43.30330	-74.59360	17.3	11.3	ThinTD	-
Jordan Lake	44.37662	-74.60806	72	9.8	-	Mesotrophic
Kiwassa Lake	44.29565	-74.15686	114.3	13.7	-	Oligotrophic
Lake Abanakee	43.76577	-74.25421	208	-	-	Mesotrophic
Lake Adirondack	43.78842	-74.26167	78	5.8	-	Mesotrophic
Lake Algonquin	43.39669	-74.29397	-	-	-	-
Lake Alice	44.86874	-73.48638	27.9	1.6	-	Oligotrophic
Lake Alice 2	44.87488	-73.47838	3.3	1.6	-	Mesotrophic
Lake Clear	44.36860	-74.25256	448.5	8.5	-	Oligotrophic
Lake Colby	44.34184	-74.15384	125.6	14.3	-	Mesotrophic
Lake Colden	44.12266	-73.97950	15.4	7.3	ThinTD	-
Lake Durant	43.84444	-74.41476	-	-	-	Oligotrophic
Lake Eaton	43.97758	-74.46556	-	-	-	-
Lake Everest	44.38940	-73.81698	18.1	3.4	-	Mesotrophic
Lake Flower	44.31448	-74.12648	131.1	3.7	-	Mesotrophic
Lake Kushaqua	44.52077	-74.11230	153.9	27.4	-	Oligotrophic
Lake Madeleine	44.12526	-74.45926	-	-	-	-
Lake of the Pines	43.68131	-72.27588	3	-	-	Eutrophic
Lake Placid	44.33687	-73.96647	-	-	-	-
Lake Pleasant	43.47345	-74.38938	-	-	-	-
Lake Rondaxe	43.76697	-74.90140	90.5	10.1	ThinTD	-
Lake Roxanne	44.89606	-73.80877	80.7	2.4	-	Eutrophic
Lake Titus	44.72493	-74.28219	177	9.1	-	Mesotrophic
Lens Lake	43.39690	-74.01374	22	-	-	Mesotrophic
Limekiln Lake	43.71197	-74.79870	186.9	21.9	MTD	-
Little Clear Pond	44.66237	-74.50040	1.9	14	MS	-
Little Echo Pond	44.30849	-74.35560	-	-	-	-
Little Green Pond	44.35737	-74.30006	29.8	12.2	-	Oligotrophic
Little Hope Pond*	44.51731	-74.12630	2.8	6.2	MS	-
Little Jabe Pond	43.71395	-73.53718	3.8	6.7	-	Oligotrophic
Little Long Lake	43.56026	-75.15073	64	10.7	-	Oligotrophic
Little Polliwog Pond	44.32691	-74.36355	8.2	1.8	-	Mesotrophic
Little Rainbow Pond	44.35935	-74.32546	5	2.1	-	Mesotrophic
Little Simon Pond*	44.15531	-74.44470	58.1	32	MTD	-
Little Square Pond	44.32029	-74.38776	41.3	8.8	-	Mesotrophic
Long Lake	44.07008	-74.33023	1685	13.7	-	Oligotrophic
Long Pond	43.83689	-74.48090	1.7	4	MTD	-
Long Pond- Essex County	44.38100	-73.45366	120.3	-	-	Mesotrophic

Long Pond- Franklin County	44.35965	-74.39301	139.3	15.2	-	Mesotrophic
Loon Hollow Pond	43.96292	-75.04310	48.9	-	-	-
Loon Lake- Franklin County	44.56291	-74.07676	143.9	16.5	-	Oligotrophic
Loon Lake- Warren County	44.55092	-74.06294	212	-	-	Mesotrophic
Lost Pond*	43.64652	-74.55700	68	1.2	ThinTD	-
Lower Ausable Lake	44.10553	-73.83346	58.2	6.1	-	Mesotrophic
Lower Beaver Pond	43.85127	-74.15225	-	-	-	Mesotrophic
Lower Cascade Lake	44.22822	-73.87178	11.1	12.5	-	Oligotrophic
Lower Chateaugay Lake	44.81902	-74.01850	234	7.6	-	Mesotrophic
Lower Saranac Lake	44.31541	-74.17953	870.5	18.3	-	Oligotrophic
Lower St. Regis Lake	44.42703	-74.24805	-	-	-	-
McCauley Pond	44.35295	-74.20341	32.8	3.6	-	Mesotrophic
Middle Branch Lake	43.70161	-75.09830	17	5.2	ThinTD	-
Middle Pond*	44.33935	-74.37910	24.3	3.3	Cl	-
Middle Saranac Lake	44.25937	-74.26723	572.6	6.1	-	Mesotrophic
Middle Settlement Lake	43.68510	-75.09740	15.8	11	ThinTD	-
Mink Pond	43.84333	-74.12747	56	-	-	Oligotrophic
Mirror Lake	44.28914	-73.98218	50.5	18.3	-	Oligotrophic
Moody Pond	44.32908	-74.11812	10.8	5.2	-	Mesotrophic
Moose Pond	44.37197	-74.06274	66	21.3	-	Oligotrophic
Morehouse Lake	43.36335	-74.67830	-	-	-	-
Moss Lake	43.78483	-74.85050	45.7	15.2	MTD	Oligotrophic
Mountain View Lake	44.70098	-74.13091	95	2.7	-	Mesotrophic
Nate Pond*	43.85752	-74.09040	8.3	6.4	MTD	-
North Lake	43.53937	-74.92710	176.8	17.7	ThinTD	-
Oseetah Lake	44.28152	-74.13294	306	-	-	Mesotrophic
Osgood Pond	44.45107	-74.22851	108	-	-	Mesotrophic
Otter Lake	43.18837	-74.49940	-	-	-	-
Otter Pond	44.36804	-74.59479	4	2.4	-	Mesotrophic
Owen Pond	44.32156	-73.90110	7.6	9.4	ThickTD	-
Owl Pond	44.26948	-74.15314	7	2.6	-	Mesotrophic
Oxbow Lake	43.44403	-74.47968	-	-	-	-
Paradox Lake	43.89136	-73.67191	377	18.5	-	Oligotrophic
Penfield Pond	43.91783	-73.53866	72.4	2.5	-	Mesotrophic
Pine Lake	43.19733	-74.51269	67	-	-	Oligotrophic
Pine Pond	44.26473	-74.14378	20.3	19.8	-	Oligotrophic
Pleasant Lake	43.68881	-75.28767	6	10.4	-	Oligotrophic
Polliwog Pond	44.33395	-74.35368	86.5	24.4	-	Oligotrophic
Putnam Pond	43.83596	-73.58016	114.6	10.4	-	Mesotrophic
Queer Lake	43.81053	-74.79950	54.5	21.3	ThinTD	-
Ragged Lake	44.71862	-74.06978	-	-	-	Mesotrophic
Rainbow Lake	44.48441	-74.15711	149.6	17.7	-	Mesotrophic
Raquette Lake	43.85232	-74.65121	2183	29	-	Oligotrophic
Raquette Lake Reservoir*	43.79301	-74.65190	1.5	3	MTD	-
Rat Pond	44.35455	-74.31243	13.7	8.8	-	Mesotrophic
Rich Lake	43.97461	-74.21492	154	19.8	-	Oligotrophic
Rollins Pond	44.31266	-74.41678	183.5	23.5	-	Mesotrophic
Round Pond	43.35228	-73.67690	15	14.3	-	Oligotrophic
Rush Pond	43.34945	-73.70331	12.2	4	-	Mesotrophic
Sacandaga Lake	43.48406	-74.42376	-	-	-	-
Sagamore Lake*	43.76758	-74.61940	68	22.9	MTD	-
Schroon Lake	43.83391	-73.75291	1722	46.4	-	Oligotrophic
Seventh Lake	43.74453	-74.74319	-	-	-	-
Silver Lake	44.50575	-73.87669	324.9	-	-	Oligotrophic
Simon Pond	44.19285	-74.44115	287	-	-	Oligotrophic
Slang Pond	44.36377	-74.37969	21.2	7	-	Mesotrophic
Sochia Pond	44.35283	-74.29430	-	-	-	-
South Lake (East Branch)	43.51165	-74.88820	-	-	-	-
Spitfire Lake	44.41798	-74.26963	-	-	-	-
Split Rock Pond	43.86486	-74.15348	-	-	-	Oligotrophic
Spy Lake	43.39657	-74.51803	-	-	-	-
Squash Pond	43.82649	-74.88700	-	-	-	-
Squaw Lake	43.63276	-74.73860	36.4	6.7	ThinTD	-
Star Lake	44.15829	-75.04935	83	19.8	-	Oligotrophic
Stony Creek Pond	44.21459	-74.31106	76	12.5	-	Oligotrophic
Sunday Pond	44.34610	-74.30070	-	-	-	-
Taylor Pond	44.48425	-73.86349	358	13.4	-	Oligotrophic

Thirteenth Lake	43.70516	-74.12690	128	14.9	-	Oligotrophic
Tripp Lake	43.59667	-73.79535	19	-	-	Oligotrophic
Trout Lake	43.54478	-73.69983	104.6	22.9	-	Oligotrophic
Trout Pond	44.41972	-73.57318	13.5	2.1	-	Mesotrophic
Tupper Lake	44.16784	-74.54009	2447	25.9	-	Oligotrophic
Turtle Pond	44.36014	-74.36126	28.7	10	-	Oligotrophic
Twitchell Lake	43.85140	-74.88244	58	10.4	-	Oligotrophic
Union Falls Reservoir	44.49194	-73.93538	660	7.6	-	Mesotrophic
Upper Ausable Lake	44.07723	-73.87371	60.5	14.6	-	Oligotrophic
Upper Cascade Lake	44.22337	-73.87948	10.6	19.2	-	Oligotrophic
Upper Chateaugay Lake	44.73259	-73.96976	1038	21.9	-	Mesotrophic
Upper Saranac Lake	44.25341	-74.32595	1970.6	26	-	Mesotrophic
Upper St. Regis Lake	44.40839	-74.28513	-	-	-	-
West Caroga Lake	43.13715	-74.49546	129	21.3	-	Oligotrophic
West Pond	43.81102	-74.87940	-	-	-	-
Whey Pond	44.30769	-74.39285	47.4	6.1	-	Mesotrophic
White Lake	43.54389	-75.15221	97	22.9	-	Oligotrophic
Willis Lake	43.36919	-74.24300	-	-	-	-
Willys Lake	43.96932	-74.95510	24.3	13.7	ThinTD	-
Windfall Pond	43.80529	-74.82890	-	-	-	-
Windover Lake	43.63260	-74.01294	38	3.1	-	Mesotrophic
Wolf Lake	44.02058	-74.21967	59	-	-	Oligotrophic
Woods Lake	43.87003	-74.95230	24.7	10.1	ThinTD	-
Zack Pond	43.93347	-74.18654	-	-	-	Oligotrophic

Notes. Blue font indicates lakes from ALTM data, black font indicates lakes from ALAP data, and orange font indicates lakes represented in both in-situ datasets. Classification - Medium Till Drainage (MTD), Thin Till Drainage (ThinTD), Thick Till Drainage (ThickTD)...

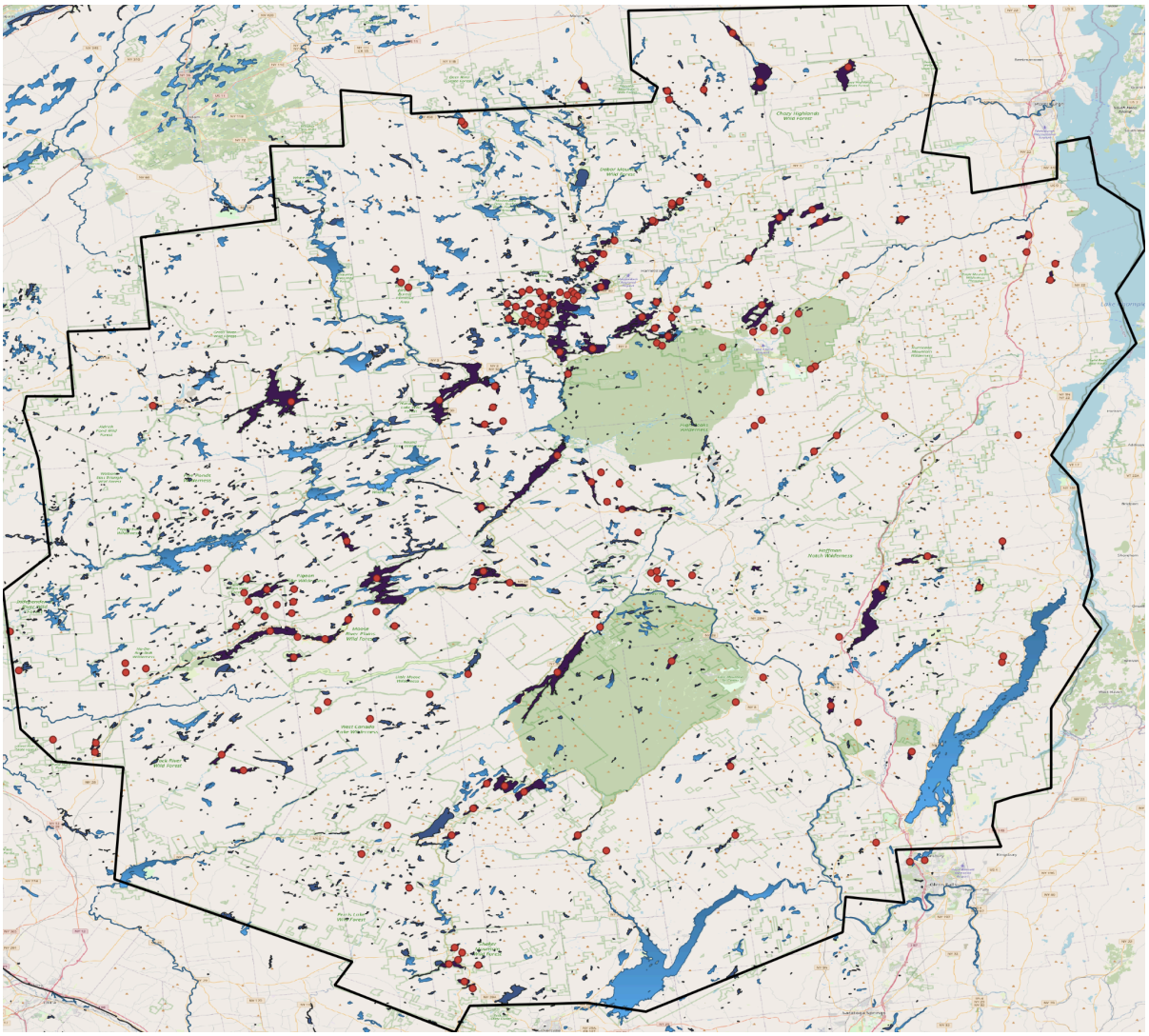

Figure 2. Location of ALAP and ALTM sampled lakes across Adirondack Park. Highlighted in a darker shade are the sampled lakes, represented by red markers.

3.3. Remote Sensing Instruments

Landsat-5 has captured data from 1984 to 2013 and provides multispectral imagery with a spatial resolution of 30 meters and temporal resolution of 8 days across 7 spectral bands (*USGS Landsat 5 Level 2, Collection 2, Tier 1 | Earth Engine Data Catalog | Google for Developers*, n.d.).

Table 2.
Landsat-5 Thematic Mapper (TM) band information

Name	Description	Wavelength	Resolution
B1	Blue	0.45 - 0.52 µm	30 meters
B2	Green	0.52 -0.60 µm	30 meters
B3	Red	0.63 - 0.69 µm	30 meters

B4	Near infrared	0.77 - 0.90 µm	30 meters
B5	Shortwave infrared 1	1.55 - 1.75 µm	30 meters
B6	Shortwave infrared 2	2.08 - 2.35 µm	30 meters

Notes. Bands of focus are B1-B4

Landsat-8 has acquired data from 2013 to 2023. Similar to Landsat-5, Landsat-8 provides a 30-meter multispectral spatial resolution and has a temporal resolution of 8 days (*USGS Landsat 8 Collection 2 Tier 1 Raw Scenes | Earth Engine Data Catalog | Google for Developers*, n.d.).

Table 3.
Landsat-8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) band information

Name	Description	Wavelength	Resolution
B1	Coastal Aerosol	0.43 - 0.45 µm	30 meters
B2	Blue	0.45 - 0.51 µm	30 meters
B3	Green	0.53 - 0.59 µm	30 meters
B4	Red	0.64 - 0.67 µm	30 meters
B5	Near Infrared	0.85 - 0.88 µm	30 meters
B6	Shortwave infrared 1	1.57 - 1.65 µm	30 meters
B7	Shortwave infrared 2	2.11 - 2.29 µm	30 meters
B8	Band 8 - Panchromatic	0.52 - 0.90 µm	15 meters
B9	Cirrus	1.36 - 1.38 µm	30 meters
B10	Thermal infrared 1	10.60 - 11.19 µm	30 meters
B11	Thermal infrared 2	11.50 - 12.51 µm	30 meters

Notes. Bands of interest are B1-B4

The surface reflectance data collected by these satellites span several decades, providing a 40-year record. This data is used to assess the capabilities of empirical algorithms and machine learning models.

3.4. Atmospheric Correction and CDOM Estimation Algorithms

To accurately retrieve surface reflectance data for water quality parameters from Landsat-5 and Sentinel-2 imagery, an essential step is an atmospheric correction. The presence of atmospheric constituents causing light scattering or absorption can distort the radiance signals and alter the accuracy of the derived CDOM information.

Google Earth Engine (GEE) is used to obtain USGS Landsat 5 TM Collection 2 Tier 1 Raw Scenes data. Landsat 5 surface reflectance data is generated using a pre-implemented atmospheric correction algorithm, specifically the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm. USGS Landsat 8 OLI Collection 2 data is also pre-processed through the Land Surface Reflectance Code (LaSRC) algorithm.

3.4.1. Small Scale Atmospheric Correction Algorithm Development Process for Landsat-5 and 8 Data

The Landsat Quality Assessment (QA) band is used to generate masks that remove low-quality pixels and clouds. In the algorithm, a variable is generated by applying bitwise operations on the QA band that selects pixels where no bits are set to 1, indicating good pixel quality. The mask then filters out undesired pixels. Another mask is generated by selecting pixels where the Radiometric Saturation (QA_RADSAT) is equal to zero. The QA_RADSAT mask identifies and removes image pixels that are saturated. The cloud mask is defined by identifying pixels affected by clouds using specific bits in the QA band. The cloud mask variable is generated by combining bitwise operations on the QA band with the bits corresponding to cloud presence (bit 5, bit 7, and bit 3). The pixels in the satellite image marked as clouds in the QA band are masked out, leaving only clear pixels for processing.

The Landsat surface reflectance bands are converted from Digital Numbers (DN) to reflectance values using conversion coefficients. The formula applied to the optical bands is: $(DN \times 0.0000275) - 0.2$. This conversion is made by scaling the DN values and then subtracting an offset to retain reflectance values. The thermal band is also corrected for radiometric calibration using coefficients. The generated masks are then applied to the image to retain only the pixels of interest.

The atmospheric-corrected images from Landsat-5 and Landsat-8 are processed through a CDOM estimation function to derive CDOM values for the water bodies.

3.5. CDOM Data Collection Methods: Lake Polygons

This study uses two methods of data collection to enhance the accuracy of CDOM estimation: (1) Creation of lake polygon over lake monitoring station coordinates, (2) Creation of lake polygon over Lake Center Coordinates (reference Table 1 for lake coordinates). Lake polygons define the boundaries of each sampled lake, and they consider potential variations in CDOM concentrations within the water body.

3.5.1. Lake Polygon over Lake Monitoring Station

Lake polygons are generated to encompass the geographical coordinates of monitoring stations located within each lake. This method gives insight into CDOM at specific monitoring points, providing localized information on CDOM behavior. A monitoring station is a designated location within a lake where regular observations and measurements are carried out. This method targets specific areas within the lake that have been consistently monitored, allowing for tracking changes in CDOM concentrations over time.

A challenge was encountered in accurately positioning certain lake monitoring station coordinates within the lakes. To manually adjust the station coordinates, Geographic Information System (GIS) tools, including QGIS, were used. Before, certain station coordinates for some of the 50 ALTM sampled lakes were incorrectly located over land areas outside the lakes, resulting in inaccurate CDOM measurements (Fig. 2). The relocation of these points within the lake boundaries increases the reliability of the CDOM data collected from the lake polygon over the station of each lake.

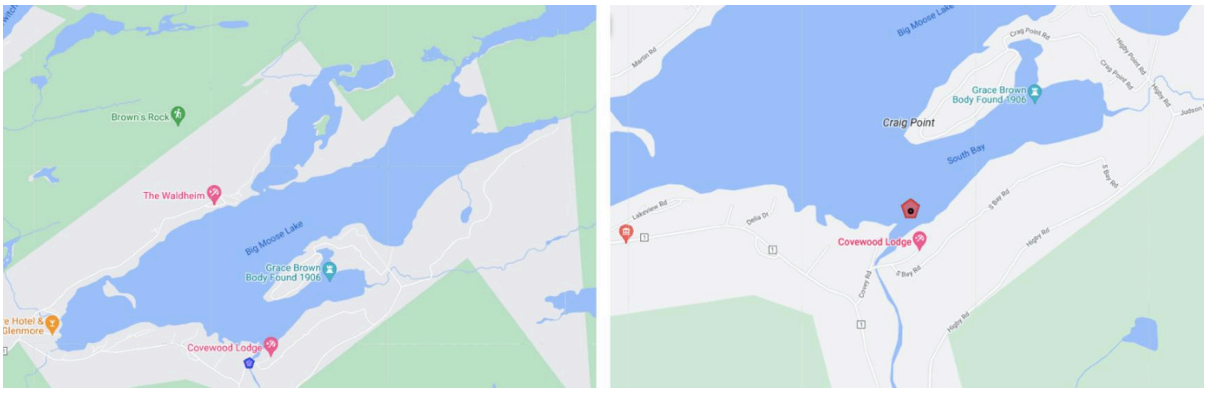

Figure 2. Map of Big Moose Lake (43.83221°N, 74.8465°W) and its lake polygon before and after manual relocation of station coordinates using GIS. For a pixel size of 20 meters (for Sentinel-2) and pixel size of 30 meters (for Landsat-5), a buffer distance of 3 times this pixel size is calculated in meters. This buffer distance is then converted to degrees for latitude and longitude adjustments, with an approximate conversion factor of 111320.0 meters per degree.

3.5.2. Lake Polygon over Lake Center Coordinates

Lake polygons are also generated to encompass the geographical coordinates of the center of each lake. This method provides a broader perspective of CDOM distribution by focusing on the central region of the lake. Examining CDOM behavior in these regions provides insight into how CDOM concentrations can differ across the lake's central and outer regions. This methodology adds to the station-centered approach, providing a holistic understanding of CDOM's spatial distribution within the entire lake ecosystem.

3.6. Machine Learning Models

Several machine learning bagging and boosting models are implemented to predict Dissolved Organic Content (DOC) levels using various input features derived from surface reflectance data.

Bagging methods create multiple models that each learn from different parts of the training data. In this study, RandomForest, SVR, and MLPRegressor machine learning bagging models are used to reduce overfitting and capture complex relationships in the data. Boosting methods build strong predictive models by focusing on challenging instances. In this study, AdaBoost, XGBoost, and GradientBoosting machine learning boosting models are used to improve model performance, adaptability, and generalization.

To evaluate the performance of these models, a data split of 70% for training and 30% for testing is utilized. This data split ensures that the models are trained on a large portion of the data while being evaluated on new data to assess generalization. To find the best-performing model, the impact of lake classification, atmospheric correction algorithms, and lake water depths on the models' performance are investigated. The assessment of model performance includes the use of metrics such as the r-squared error (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Different time windows of 1, 3, and 7 days are examined to assess how variations in observation periods impact the model's ability to predict DOC concentrations accurately.

Table 3.
CDOM predictive models.

Model	Equation form	Coefficients	R^2	RMSE	Lake Type (size, depth, etc)	CDOM	R^2	RMSE
Kuster et al. (2005)	$\ln(_{a440}) = a_1 + a_2\ln(B3/B4)$ *Landsat 8*	1.582, – 1.507	0.84	0.47	Size: 0.009 - 126.7 k^2	N/A		
Brezonik et al. (2005)	$\ln(_{a440}) = a_1 + a_2$ (B2) + a_3(B2/B5) *Landsat 8*	2.304, – 255.88, – 0.2542	0.53	1.32	Depth: 1.8 - 39.0 meters	N/A	0.00	15.99
Brezonik et al. (2015)	$\ln(_{a440}) = a_1 + a_2\ln$(B2:B6) *Sentinel-2*	1.872, – 0.830	0.86	0.44	N/A	High		
" "	$\ln(_{a440}) = a_1 + a_2\ln$(B1/B5) *Landsat 8*	1.168, – 0.957	0.72	0.62	N/A	High		
" "	$\ln(_{a440}) = a_1 + a_2\ln$(B2/B5) *Landsat 8*	1.441, – 0.841	0.70	0.63	N/A	High		
Olmanson et al. (2016)	$\ln(_{a440}) = a_1 + a_2$(B1/B3) + a_3 (B2/B3) *Landsat 7*	14.62, 2.77, – 14.21	0.83	0.43	N/A	Low		
" "	$\ln(_{a440}) = a_1 + a_2$(B1/B4) + a_3 (B2/B4) *Landsat 8*	29.2, 109.3, – 143.2	0.83	0.44	N/A	Low	0.06	8.38
Martins et al. (2018)	$a_{CDOM}(485) = a_1 + a_2$ (B4/B1) *Landsat 5*	– 0.5986, 5.5510	0.91	0.78	N/A*	High*	0.00	204787.59
Olmanson et al. (2020)	$\ln(_{a440}) = a_1(R_{rs}$(B4)/ R_{rs}(B3)) + $a_2(R_{rs}$(B5)/ R_{rs}(B3)) + a_3 *Landsat 8*	0.42, 1.79, 6.07	0.85	0.49	Size: ≥ 4 ha	Low	0.02	0.72
Koll-Egyed et al. (2021)	$\ln(CDOM(_{a440})) = a_1 - a_2$ (ln(B3/B4)) - a_3 (ln(B2) *Landsat 8*	3.65, 2.91, 0.41	0.47	0.73	Size: ≥ 10 ha	Low	2.98	0.67
Zhang et al. (2022)	$X = a_1R_{rs}$(B1) + $_{a2}R_{rs}$(B2) + a_3R_{rs}(B3) + a_4R_{rs}(B4) *Landsat 8*	11.053, 6.950, 51.125, 34.457	0.79	N/A	Depth: 33 meters*	N/A*		
Huang et al. (2023)	$ag(290) = a_1$ (B3/B2)+ a_2 *Landsat 8*	9.302, – 8.801	0.55	1.36	Depth: 30 meters*	N/A		
This study:	CDOM = a_1 - a_2(B2/B3) + a_3 (B3/B4) *Sentinel-2 & Landsat 8*	20.3, 10, 2.4						

Notes. The published equations by Kuster et al. (2005) and Brezonik et al. (2005) are reformatted by Brezonik et al. (2015) to a measure of CDOM (a440) using bands of Landsat 8 OLI 1-5. R^2 and RMSE values on the last two columns are the tested values in this study using their respective equations, which are produced with merged Landsat-5 and 8 data over lake centroids. Information marked with an asterisk indicates the respective study is focused on a single inland water body.

4. RESULTS & DISCUSSION

4.1. Time Series Plots & DOC vs. CDOM Scatter Plots

Big Moose Lake, one of the lakes included in the 50 sampled lakes of the Adirondack Long-Term Monitoring (ALTM) program, was chosen as a focal point to explore the relationship between CDOM and DOC concentrations at a localized level within a smaller, more specific context.

4.1.1. Landsat-5 DOC vs. CDOM Time Series

The time series scatter plot, using data from Landsat-5 and 8, did not reveal a correlation between CDOM and DOC concentrations within Big Moose Lake. The plot demonstrates that while DOC exhibits a clear and consistent linear increase over time, the behavior of CDOM is notably different. CDOM's data over time lacks a clear pattern, suggesting that its concentration does not follow a linear trajectory like DOC (Fig. 3). While DOC and CDOM frequently show correlations and similar trends, it is important to note that an upward linear trend in DOC over time does not necessarily always imply a similar behavior in CDOM data.

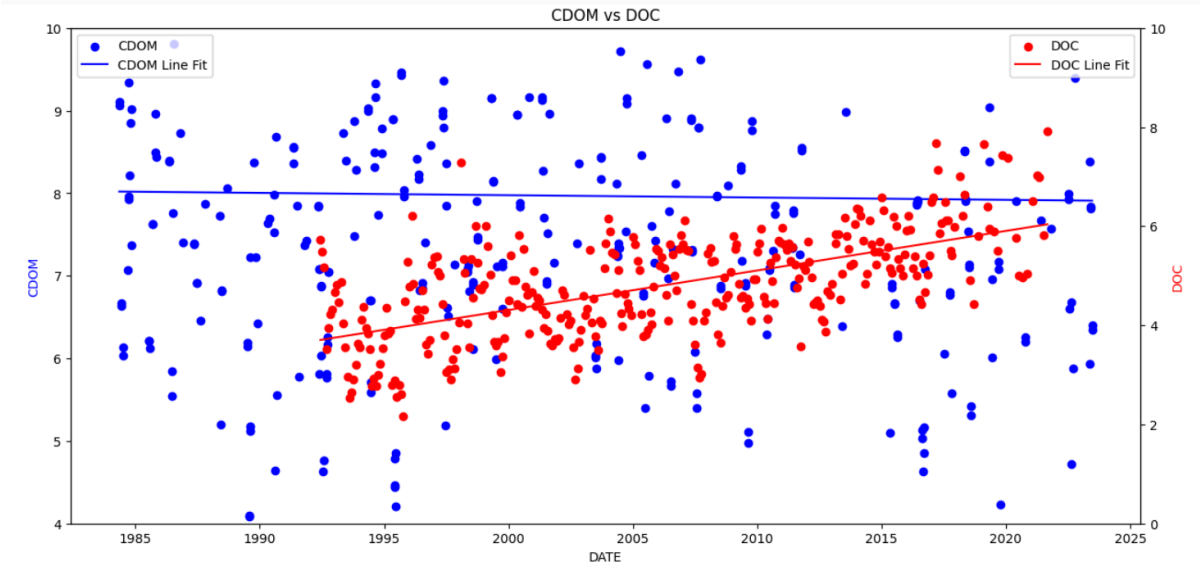

Figure 3. Landsat-5 time series of CDOM vs. DOC from 1985-2013 over Big Moose Lake. Landsat-8 is included in this time series to provide a larger time scale from 1985-2023.

The time series scatter plot using only Landsat-5 data, however, did reveal a correlation between CDOM and DOC concentrations within Big Moose Lake. The plot illustrates that both CDOM and DOC follow a positive linear trend. However, because Landsat-5 has data up until 2011 for CDOM, the line stops at that specific year, while the in-situ DOC data continues until 2023. The line of fit for CDOM depicts an R^2 of 0.03 while the line of fit for DOC depicts an R^2 of 0.39, indicating that the correlation of CDOM from Landsat 5 is not enough to determine its relationship with DOC's increasing linear pattern over time.

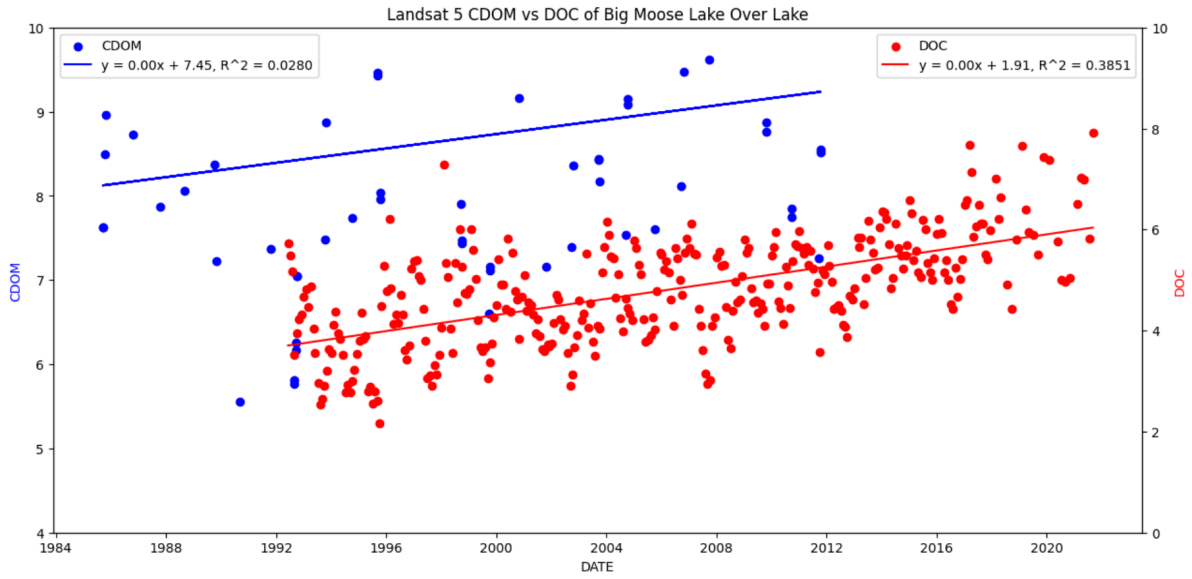

Figure 4. Landsat-5 time series of CDOM vs. DOC from 1985-2013 over Big Moose Lake. Landsat-5 data is limited to before 2011.

4.1.2. Landsat-8 DOC vs. CDOM Time Series

4.1.3. Landsat-5 Merged DOC vs. CDOM Scatter Plots

The merging process involves combining information from the in-situ DOC data and the CDOM estimated data from satellite surface reflectance measurement using the empirical formula. By specifying a 7-day tolerance, this method matches and merges rows from the datasets that fall within a close time range, integrating related data points within the Landsat-5 data collection timeframe. A 7-day time window is selected for Landsat-5 as it helps identify trends and patterns that might be overlooked in shorter intervals while preserving the overall trends visible in longer timeframes. This time window allows for the consideration of environmental conditions, such as weather patterns, seasonal transitions, and potential anthropogenic influences, which could affect DOC and CDOM concentrations.

In the scatter plot comparing DOC and CDOM concentrations, a lack of correlation suggests that changes in CDOM do not consistently correspond to changes in DOC concentration (Fig 5). No correlation between CDOM and DOC, however, could indicate the influence of other factors such as local sources of CDOM (e.g., terrestrial runoff) or interactions between CDOM and inorganic particles in the water.

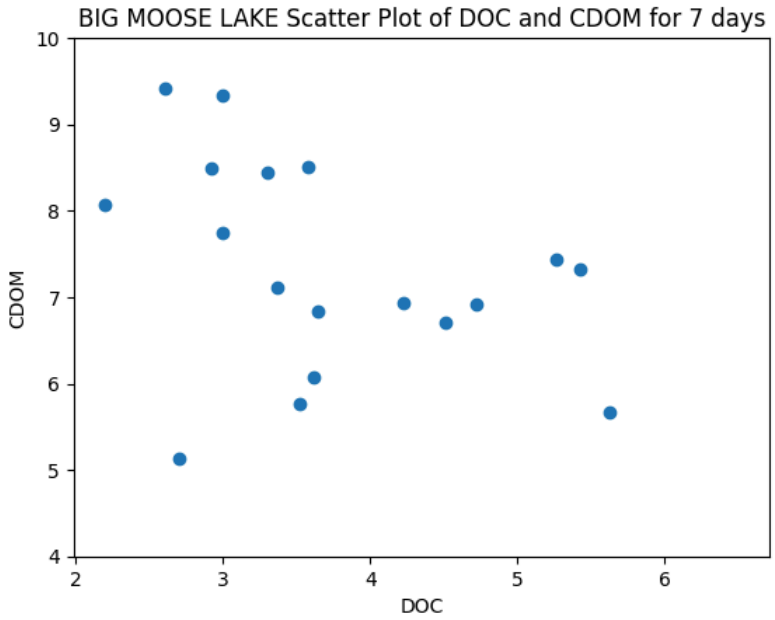

Figure 6. Landsat-5 scatterplot of merged DOC vs. CDOM over Big Moose Lake

4.1.4. Landsat-8 Merged DOC vs. CDOM Scatter Plots

4.2. DOC vs. CDOM Scatterplot for All Sampled Lakes

Despite the limited outcomes observed for Big Moose Lake, scatterplots for Landsat-5 (Fig. 7) and Sentinel-2 (Fig. 8) depicting the relationship between DOC and CDOM were also generated for all 50 sampled ALTM lakes. This broader analysis was conducted to determine whether any correlations could be discovered at a generalized level.

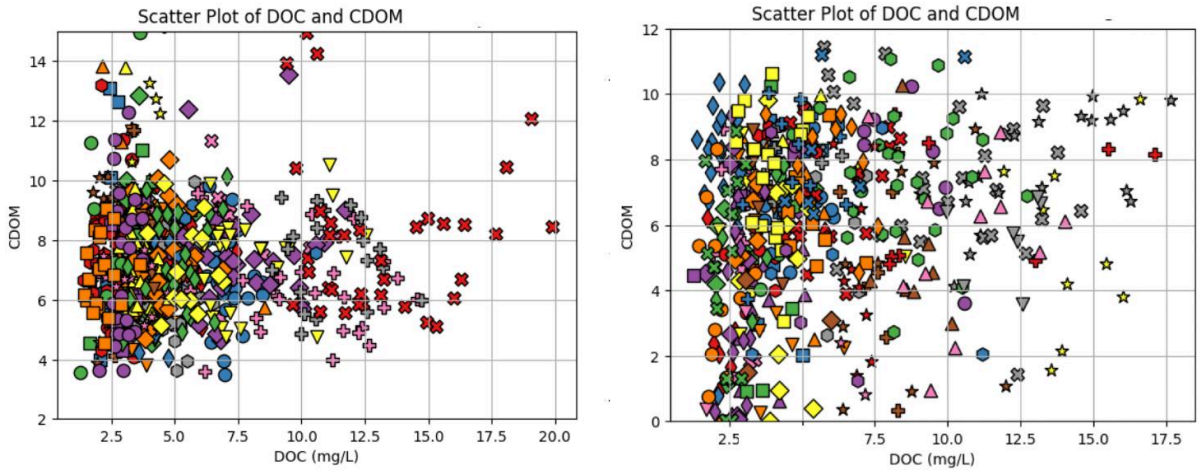

Figure 8. Landsat-5 merged DOC vs. CDOM for all 50 sampled ALTM lakes. Left plot is generated using lake center coordinates and the plot on the right is generated using station coordinates.

However, the results obtained from these scatterplots do not reveal any significant correlations either. The lack of clear patterns in the relationship between DOC and CDOM concentrations across the sampled ALTM lakes suggests that the traditional analysis methods might not be sufficient. As a result, a transition towards more machine learning models is necessary. These models have the potential to uncover underlying patterns and interactions that might be overlooked by traditional statistical methods, offering an understanding of the relationships between DOC and CDOM concentrations in lake ecosystems.

4.3. Surface Reflectance Data

4.3.1. Landsat-5 Surface Reflectance Data for Big Moose Lake

Analyzing the reflectance values captured in the red, green, blue, and Near Infrared (NIR) bands over time helps reveal patterns that highlight the seasonal alterations in DOC and CDOM concentration levels. The goal of this analysis is to determine whether there are consistent patterns of reflectance that correlate with shifts in CDOM and DOC concentrations.

The graph illustrates the variations in reflectance values for each band, highlighting how these values evolve across different seasons (Fig. 9). The Near-Infrared band generates higher reflectance values in comparison to the red, green, and blue bands. This behavior in the Near-Infrared band's reflectance is due to its sensitivity to CDOM and DOC. Because CDOM and DOC absorb light in the visible spectrum, lower reflectance in the red, green, and blue bands is observed. On the contrary, the Near-Infrared band produces higher reflectance due to its lower absorption and longer wavelength.

When analyzing the Landsat-5 reflectance values collected in the red, green, and blue bands over time, consistent patterns across different seasons and years are observed, depicting similar trends and variations. The red, green, and blue bands are sensitive to factors including water clarity, sediment content, and the presence of suspended particles in lake ecosystems. These factors also can influence the scattering and absorption of light across the visible spectrum and can inaccurately represent DOC and CDOM concentrations. Because of this, the patterns in the reflectance values for these bands often mirror each other.

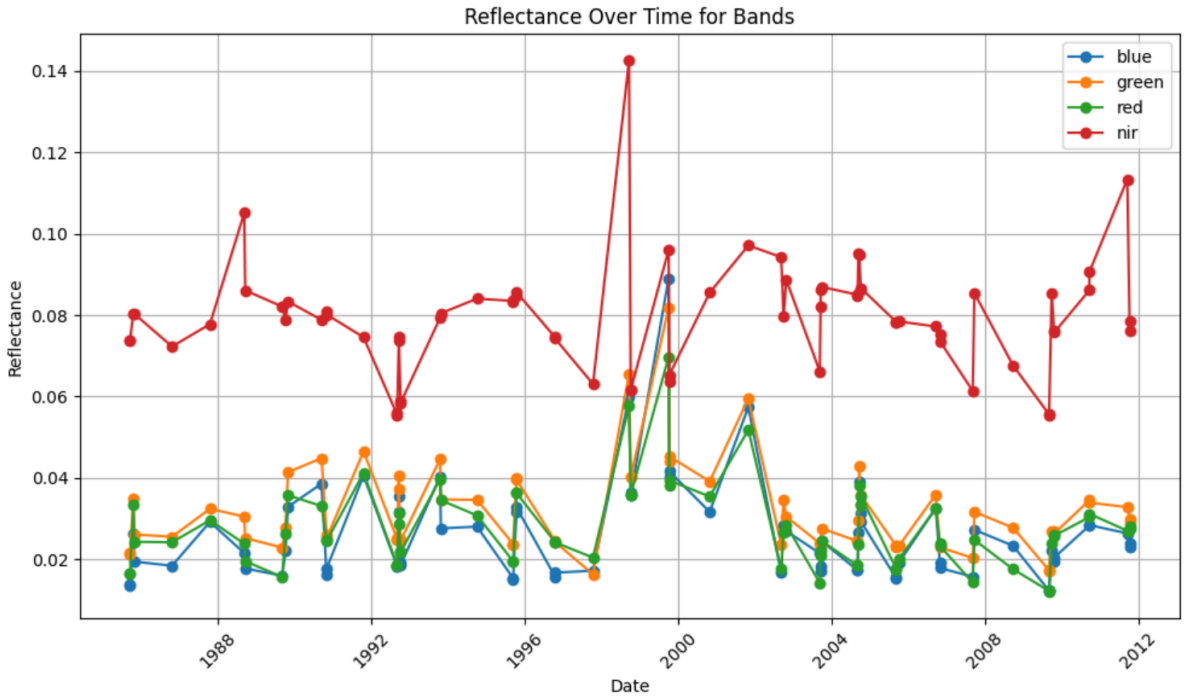

Figure 10. Landsat-5 Surface Reflectance values over Big Moose Lake for bands Red, Green, Blue, and Near-Infrared over years 1985-2011

4.3.2. Landsat-8 Surface Reflectance Data for Big Moose Lake

4.4. Machine Learning Models

4.4.1. Landsat-5 Machine Learning Data

In addition to the spectral bands, lake classification parameters, lake depth mean and lake label were used as input features to produce the best results. For Landsat-5, bagging models RandomForest and LightGBM produced the highest results with an R^2 of 0.72, suggesting a strong correlation (Table 2).

The scatterplot depicting the relationship between measured DOC and predicted DOC values provides a visual representation of the RandomForest models' predictive accuracy. The strong, positive correlation observed in this scatterplot indicates that the machine learning model is capable of approximating actual DOC concentrations and has the ability to generalize to unseen data points (Fig. 11).

Table 2.
Landsat-5 Machine Learning model error results for Surface Reflectance (using 1-day time window)

Model	R^2	MSE	RMSE	MAE
RandomForest	0.72	3.35	1.83	1.24
AdaBoost	0.59	4.96	2.23	1.82
XGBoost	0.69	3.69	1.92	1.33
GradientBoosting	0.70	3.53	1.88	1.35
LightGBM	0.72	3.31	1.82	1.28
SVR	-0.16	13.92	3.73	2.63
MLPRegressor	0.04	11.52	3.39	2.50

Notes. Error table for Figure 12, Parameters: X = [['blue', 'green', 'red', 'nir', 'LAKE_DEPTH_MEAN', 'LAKE_LABEL']] # Input (features); y = ['DOC_MG_L'] # Target variable

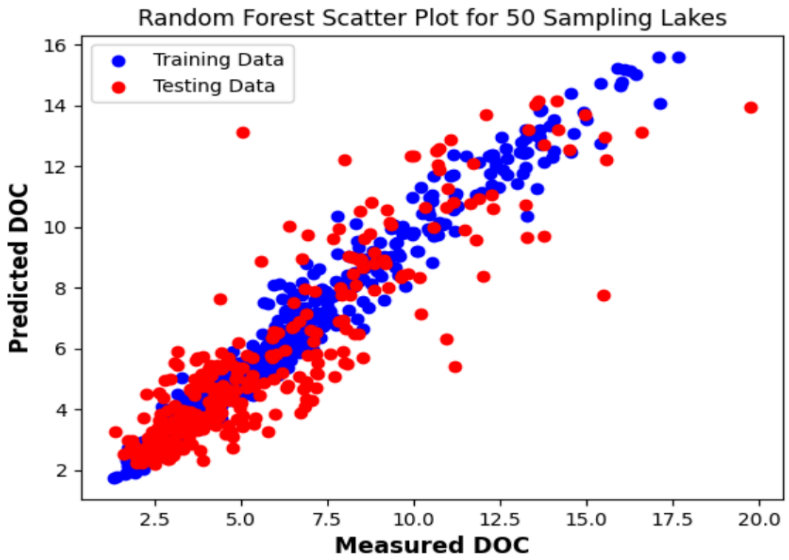

Figure 12. Landsat-5 Random Forest machine learning model scatterplot with 1-day window of training and testing data from all 50 sampled ALTM lakes

Top of Atmosphere (TOA) reflectance values were also examined as predictors for DOC beyond the regular Surface Reflectance values. It was observed that the RandomForest and LightGBM bagging models continued to provide large R^2 values. These models achieved R^2 values of 0.80, an improvement that indicates an even stronger correlation between the predicted and measured DOC values (Table 3). The scatterplot for TOA depicts the strong, positive correlation between the measured DOC concentrations and the DOC values predicted by the RandomForest and LightGBM models (Fig. 12).

Table 3.
Landsat-5 Machine Learning model error results for Top of Atmosphere (using 1-day time window)

Model	R^2	MSE	RMSE	MAE
RandomForest	0.80	2.21	1.49	1.01
AdaBoost	0.58	4.59	2.14	1.80
XGBoost	0.80	2.13	1.46	1.02
GradientBoosting	0.81	2.10	1.45	1.02
LightGBM	0.82	1.96	1.40	0.96
SVR	0.13	9.48	3.08	1.99
MLPRegressor	0.30	7.62	2.76	2.06

Notes. Error table for Figure 13, Parameters: X = [['blue', 'green', 'red', 'nir', 'LAKE_DEPTH_MEAN', 'LAKE_LABEL']] #Input (features); y = ['DOC_MG_L'] # Target variable

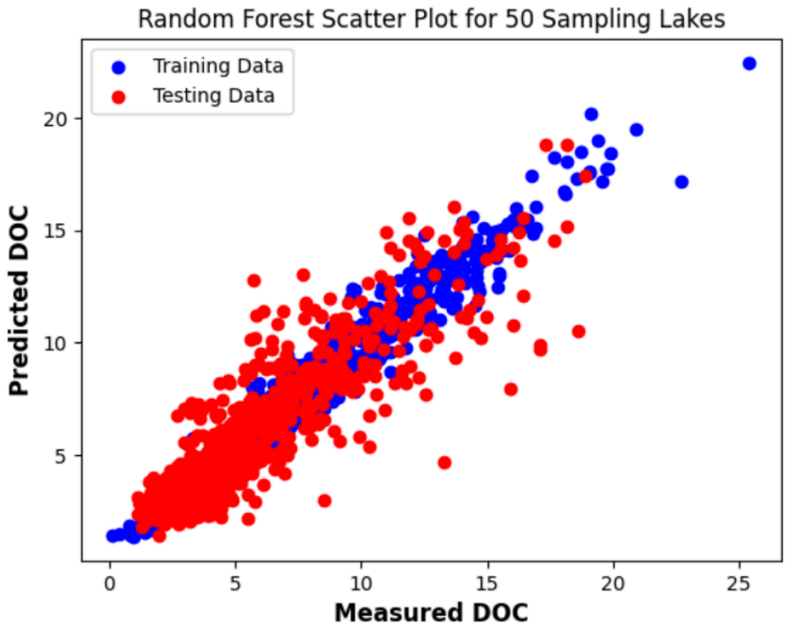

Figure 13. Landsat-5 Random Forest machine learning model scatterplot with a 1-day window of training and testing data from all 50 sampled ALTM lakes

4.4.2. Landsat-8 Machine Learning Data

5. CONCLUSION

Although the findings from both Landsat-5 and Landsat-8 reflectance data analysis do not depict consistent strong correlations between CDOM and DOC, machine learning evidently showcases its ability to uncover distinct, underlying patterns of DOC in lake ecosystems. In this study, Machine learning models demonstrated better performance for CDOM estimation compared to the empirical model. Leveraging machine learning algorithms and satellite-based CDOM estimation demonstrates its ability to reveal strong correlations between CDOM and DOC to successfully assess water color as an indicator of water quality in inland water bodies. Overall, the utilization of satellites and machine learning algorithms has great potential in revealing long-term water quality trends, understanding lake regional dynamics, and potentially aiding the identification of representative sampling lakes within the Adirondack region for future surveys.

6. FUTURE WORK

A data gap remains beyond the studied lakes. However, the scalability of machine learning models offers an opportunity for extrapolation to other lakes, allowing for the understanding of climate dynamics and outside impacts on Dissolved Organic Content (DOC) and Colored Dissolved Organic Matter (CDOM) within the broader Adirondack area. Future work will revolve around the development of CDOM maps across diverse lakes. These maps, designed to visually depict the spatial distribution of CDOM, will enhance the understanding of its variability across the lake ecosystems. These efforts are set to improve the understanding of how watercolor indicates water quality using CDOM and DOC.

7. ACKNOWLEDGEMENT

This summer research project was supported by NY Department of Environmental Conservation Grant #DEC01-C01714GG-3350000 (NYS-DEC), NSF GRANT# AGS-2023174 (NSF-IUSE), and NSF Grant AGS-1950629 (REU). Special appreciation to the New York State SCALE (Survey of Climate Change and Adirondack Lake Ecosystems) project for sharing the in-situ data. The statements contained within the poster are not the opinions of the funding agency or the U.S. government but reflect the author's opinions.

6. REFERENCES

Adirondack Park Campground Map—NYS Dept. Of Environmental Conservation. (n.d.). Retrieved August 24, 2023, from https://www.dec.ny.gov/outdoor/33154.html

Aurin, D., Mannino, A., & Lary, D. J. (2018). Remote Sensing of CDOM, CDOM Spectral Slope, and Dissolved Organic Carbon in the Global Ocean. *Applied Sciences*, *8*(12), 2687–2687. https://doi.org/10.3390/app8122687

Brezonik, P. L., Olmanson, L. G., Finlay, J. C., & Bauer, M. E. (2015). Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters. *Remote Sensing of Environment*, *157*, 199–215.

Brezonik, P. L., Menken, K., & Bauer, M. E. (2005). Landsat-based Remote Sensing of Lake Water Quality Characteristics, Including Chlorophyll and Colored Dissolved Organic Matter (CDOM). *Lake and Reservoir Management*, *21*(4), 373–382.

Corporation, A. L. S. (n.d.). *Adirondack Lakes Survey Corporation—ALTM*. Retrieved August 24, 2023, from http://www.adirondacklakessurvey.org/

Huang, Y., Pan, J., & Devlin, A. T. (2023). Enhanced Estimate of Chromophoric Dissolved Organic Matter Using Machine Learning Algorithms from Landsat-8 OLI Data in the Pearl River Estuary. *Remote Sensing (Basel, Switzerland)*, *15*(8), 1963-. https://doi.org/10.3390/rs15081963

Keller, S., Maier, P. M., Riese, F. M., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake, A., & Hinz, S. (2018). Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophylla, Diatoms, Green Algae and Turbidity. *International Journal of Environmental Research and Public Health*, *15*(9), 1881-.

Kim, J., Jang, W., Kim, J. H., Lee, J., Cho, K. H., Lee, Y.-G., Chon, K., Park, S., Pyo, J., Park, Y., & Kim, S. (2022). Application of airborne hyperspectral imagery to retrieve spatiotemporal CDOM distribution using machine learning in a reservoir. *International Journal of Applied Earth Observation and Geoinformation*, *114*, 103053-.

Koll-Egyed, T., Cardille, J. A., & Deutsch, E. (2021). Multiple Images Improve Lake CDOM Estimation: Building Better Landsat 8 Empirical Algorithms across Southern Canada. *Remote Sensing (Basel, Switzerland)*, *13*(18), 3615-.

Kutser, T., Pierson, D. C., Kallio, K. Y., Reinart, A., & Sobek, S. (2005). Mapping lake CDOM by satellite remote sensing. *Remote Sensing of Environment*, *94*(4), 535–540.

Martins, S., Chokmani, K., Alcântara, E., Ogashawara, I., & El-Alem, A. (2018). Mapping the coloured dissolved organic matter absorption coefficient in a eutrophic reservoir using remotely sensed images. *Inland Waters*, *8*(4), 488–504. https://doi.org/10.1080/20442041.2018.1482153

More About the Adirondack Park. (n.d.). Retrieved August 24, 2023, from https://apa.ny.gov/about_park/more_park.html

Olmanson, L. G., Brezonik, P. L., Finlay, J. C., & Bauer, M. E. (2016). Comparison of Landsat 8 and Landsat 7 for regional measurements of CDOM and water clarity in lakes. *Remote Sensing of Environment*, *185*, 119–128.

Olmanson, L. G., Page, B. P., Finlay, J. C., Brezonik, P. L., Bauer, M. E., Griffin, C. G., & Hozalski, R. M. (2020). Regional measurements and spatial/temporal analysis of CDOM in 10,000+ optically variable Minnesota lakes using Landsat 8 imagery. *The Science of the Total Environment*, *724*, 138141–138141.

Ruescas, A., Hieronymi, M., Mateo-Garcia, G., Koponen, S., Kallio, K., & Camps-Valls, G. (2018). Machine Learning Regression Approaches for Colored Dissolved Organic Matter (CDOM) Retrieval with S2-MSI and S3-OLCI Simulated Data. *Remote Sensing (Basel, Switzerland)*, *10*(5), 786-. https://doi.org/10.3390/rs10050786

Sentinel-2 MSI: MultiSpectral Instrument, Level-2A | Earth Engine Data Catalog. (n.d.). Google for Developers. Retrieved August 24, 2023, from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR

Sun, X., Zhang, Y., Zhang, Y., Shi, K., Zhou, Y., & Li, N. (2021). Machine Learning Algorithms for Chromophoric Dissolved Organic Matter (CDOM) Estimation Based on Landsat 8 Images. *Remote Sensing (Basel, Switzerland)*, *13*(18), 3560-. https://doi.org/10.3390/rs13183560

USGS Landsat 5 Level 2, Collection 2, Tier 1 | Earth Engine Data Catalog | Google for Developers. (n.d.). Retrieved August 24, 2023, from

https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LT05_C02_T1_L2

Zhang, W., Wang, S., Zhang, B., Zhang, F., Shen, Q., Wu, Y., Mei, Y., Qiu, R., & Li, J. (2022). Analysis of the water color transitional change in Qinghai Lake during the past 35 years observed from Landsat and MODIS. *Journal of Hydrology. Regional Studies*, *42*, 101154-.

Zhu, W., Yu, Q., Tian, Y. Q., Becker, B. L., Zheng, T., & Carrick, H. J. (2014). An assessment of remote sensing algorithms for colored dissolved organic matter in complex freshwater environments. *Remote Sensing of Environment*, *140*, 766–778.

USGS Landsat 8 Collection 2 Tier 1 Raw Scenes | Earth Engine Data Catalog | Google for Developers. (n.d.). Retrieved September 19, 2023, from https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1