Adil Adetunji and Jason Louie

Log Book

December 16th, 2023:

- Identified project idea and problem
  - Identified problem:
    - Thousands of lakes that are difficult to access, making diagnostics on their health very difficult)
- Began research on the problem and background information
  - How cyanobacteria blooms affect lake health

December 27th, 2023:

- Continued research on the problem and background information.
  - Background information:
    - Factors that affect an cyanobacteria bloom
  - Problem:
    - Algae blooms affect lake health.
    - Chlorophyll A levels indicate the scale of the algae bloom

January 5th, 2024:

- Pondered possible solutions to problem and then researched them
- Decided to use AI to predict Chlorophyll A levels given a dataset
- Researched different AI models

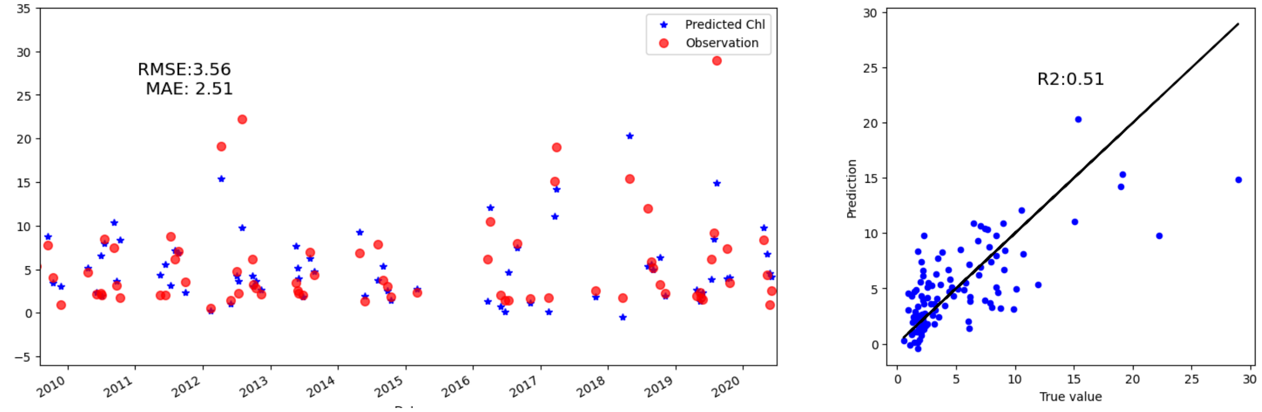
January 16th, 2024:

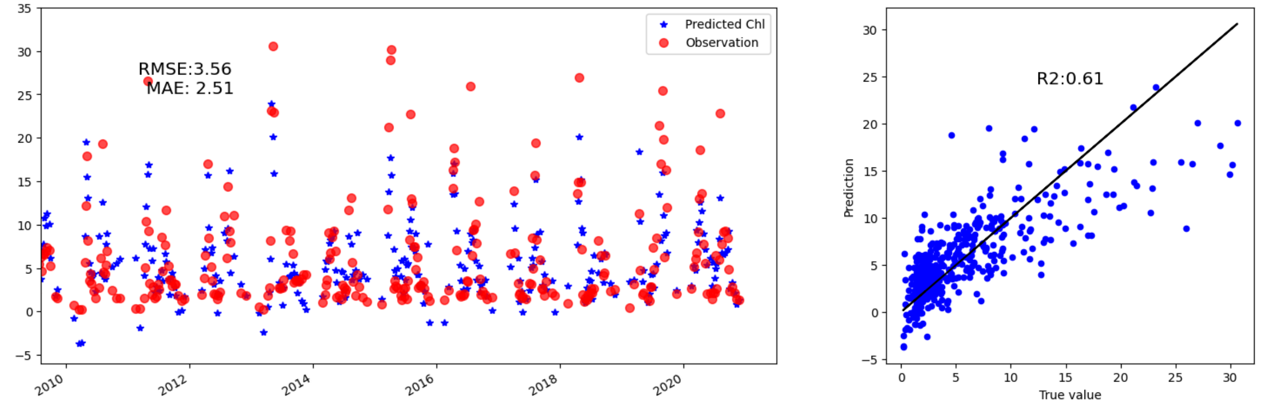
- Determined our solution to the problem:
  - Create our own code that analyzes data from a given lake and predicts the Chlorophyll A levels
  - Data used was from Erken Lake in Sweden
    - Good data because it had been observed daily for over 20 years.
  - Created using Linear Regression model
    - Researched what a Linear Regression model is, as well as inaccuracy values such as RMSE, MAE, R2
      - RMSE: Root Mean Squared Error
        - Performance indicator
        - The lower the better (theoretical best case is a 0)
      - MAE: Mean Absolute Error
        - Average size of mistakes in predicted values
        - The lower the better
      - R^2: R-Squared

-
    -
        - The closer the value is to 1, the better the results.
        - 0 means randomness in results, 1 means the results are accurate.
- Determined method, materials, and hypothesis for our project
- We then began development on an AI model that would predict Chlorophyll A levels in Chestermere Lake

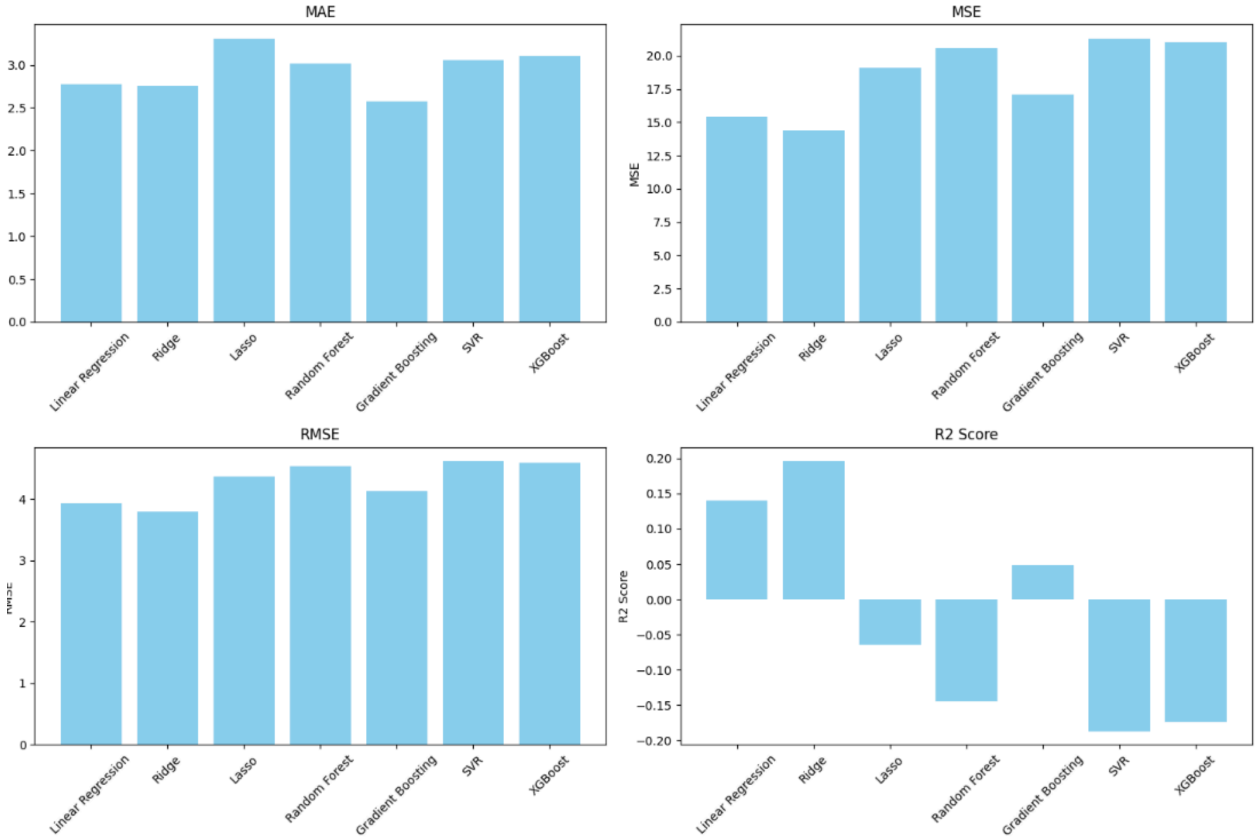
January 29th, 2024:

- Finished the AI model using data from Lake Erken and got these results
    - Used libraries:
        - pandas
            - dataframe library
        - numpy
            - mathematical operations on arrays library
        - sklearn
            - machine learning library
        - sklearn LinearRegression
            - used Linear Regression library from sklearn in our model
        - matplotlib
            - visualization library used to visualize our data
    - The AI model outputted these graphs (first is training data, the second is testing data):
        - Training data and testing data were split in order to prevent overfitting
            - Overfitting is when the model fits the training data 100% and cannot accurately predict with other data





- Finished AI model on Chestermere Lake
    - Due to lack of data on Chestermere Lake, we could not create the graphs like we did for Erken Lake
    - Instead, we analyzed the effectiveness of different models



-
    - And what each model saw as an import variable:
-
- Model Research
    - Linear Regression
        - covered before
    - Ridge Regression
        - Useful in solving problems that have less than one hundred thousand samples or when number of parameters exceed samples
        - Almost identical to linear regression, except that bias is introduced, essentially meaning that starting with a slightly worse data fit, better long term predictions can be made.
    - Lasso Regression
        - Used to estimate relationships between variables and make predictions
        - LASSO stands for Least Absolute Shrinkage and Selection Operator
        - A regularization technique that uses shrinkage, meaning that data values are shrunk towards a central point as they mean.

  - Random Forest Regression
    - An algorithm that uses multiple decision trees (that are separate from each other) to predict a value.
  - Gradient Boosting
    - Uses a tree structure. Prediction starts at the “roots” of the tree, then it branches out. There is a target “goal leaf” that is the prediction result.
  - SVR
    - Different from traditional linear regression because it finds a hyperplane that best first the data points in a continuous space.
  - XGBoost
    - A decision tree style machine learning algorithm, and it allows for parallel tree boating (meaning that multiple trees can be used together).

February 1st, 2024:

- Started work on slides
  - Formatted research
  - Added analysis part

February 4th, 2024:

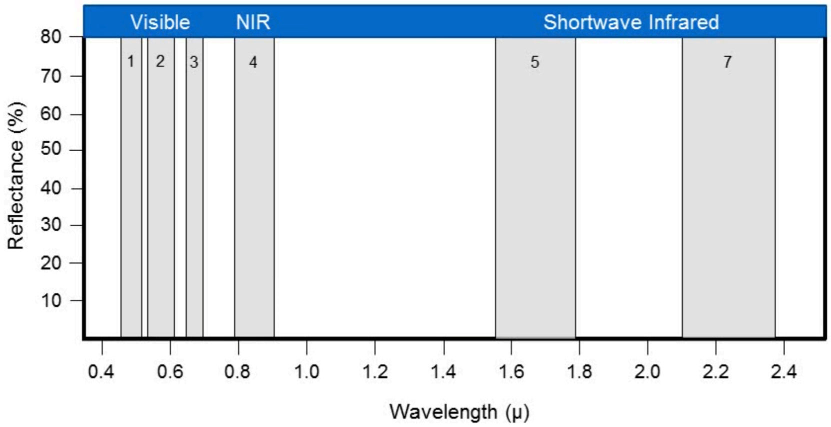
- Completed conclusions
  - Determined effectivity
  - Our learnings and understandings
  - Identified future plans for this project.
    - Utilize satellite imagery and data to add assist in our model

February 5th, 2024:

- Finalized Google Slides presentation
- Created trifold

February 20th, 2024:

- https://www.youtube.com/watch?v=zesEJouvNU4
  - Took notes on this video
  - Taught us how band combinations work
    - Different objects have different levels of reflection of light
- Looked at different ways cyanobacterial blooms are treated



-
- Surface of Earth reflects the light from these bands differently, and thus different satellites have different sensors to detect different wavelengths.

February 25th, 2024:

- https://www.researchgate.net/publication/312202874_A_survival_guide_to_Landsat_preprocessing
  - Took notes on this article, which is an article explaining what preprocessing is and why it's necessary for satellite imagery.

Introduction

- Landsat data has become highly utilized in monitoring Earth in the past decade due to its free and global coverage.
- Preprocessing of Landsat data is necessary for ecological analyses, but there is a possibility of introducing errors and is challenging for non-remote sensing scientists due to the complexity and variety of techniques involved.
- This paper aims to clarify the preparation of Landsat imagery for ecological applications

Overview and synthesis of the Landsat missions

- Landsat missions were initiated by NASA (National Aeronautics and Space Administration) in 1972 and managed by the U.S. Geological Survey since the 1990s. These missions have led to a continuous stream of satellite imagery, but the difference between satellites and sensors present a challenge for experts and non-experts.
- Landsat satellites can be organized into three groups (based on sensor and platform characteristics):
  - Landsat 1-3 with Multispectral Scanner (MSS)
  - Landsat 4-7 with Thematic Mapper (TM) or Enhanced Thematic Mapper (ETM+)

  - Landsat 8 with Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS)
- Data collected by Landsat have multispectral bands and are in a 30 m
  - Serves various ecological purposes such as vegetation mapping, water quality assessment, wildfire ecology, and land cover classification

Data coverage and dissemination

- Landsat satellites capture data during orbit, which are organized into images based on scene location and date.
  - Scene is the extent (footprint) that exists on the ground.
  - Image is the collection of spatially arranged measurements (i.e. bands) captured in the scene at a single time.
- Scenes has an assigned location defined by the Worldwide Reference System (WRS), with Landsat data distributed across two systems, WRS1 for Landsat 1–3 and WRS2 for Landsat 4–8 due to differing orbit altitudes.
- The USGS manages and disseminates Landsat data, focusing on Level-1 and Climate Data Records (CDR) products, which undergo systematic processing to ensure standardized quality.
- The USGS is also working on Analysis Ready Data (ARD) project aimed at providing standardized products to users, potentially reducing the need for preprocessing, but understanding the motivations of preprocessing remains valuable in choosing between Level-1 or ARD.

Preprocessing

- Landsat images undergo preprocessing to correct for sensor, solar, atmospheric, and topographic distortions.
  - Important for accurate analysis in ecological applications.
- Preprocessing is important to minimize errors introduced during data acquisition and ensures data is in appropriate units for analysis, such as digital numbers (DN), radiance, and reflectance.
- Ecological studies often require additional preprocessing beyond standard Level-1 or Climate Data Records (CDR) products, with a general workflow comprising geometric, absolute, and relative corrections.
- Implementation of preprocessing steps can be done using various software packages, both free (e.g. R packages, QGIS) and proprietary (e.g. ENVI, ArcGIS), depending on analysts' familiarity and analysis requirements.

Geometric correction

- Georeferencing and orthorectifying are crucial components of geometric correction for Landsat imagery
    - Georeferencing: alignment of imagery to its correct geographic location
    - Orthorectifying: correction for the effects of relief and view direction on pixel location
- Landsat Level-1 products undergo systematic processing called “terrain correction” with precision registration suitable for pixel-level analysis, classified into tiers based on quality and processing level.
- Registration ensures alignment between data layers. Critical for preprocessing Landsat imagery, especially in change detection analyses.

Absolute radiometric correction

- Absolute radiometric correction involves preprocessing steps to obtain “true” and comparable values for sensor, solar, atmospheric, and topographic effects, facilitating comparisons across time, space, or sensors, and is contrasted with relative radiometric correction.

Solar correction

- The solar correction preprocessing step adjusts at-sensor radiance to top-of-atmosphere (TOA) reflectance, considering solar influences such as solar irradiance, Earth-Sun distance, and solar elevation angle, crucial for accurate inter-image comparisons.

February 29th, 2024:

- Found two research papers regarding the use of LANDSAT-8 data to analyze cyanobacterial blooms within the lake

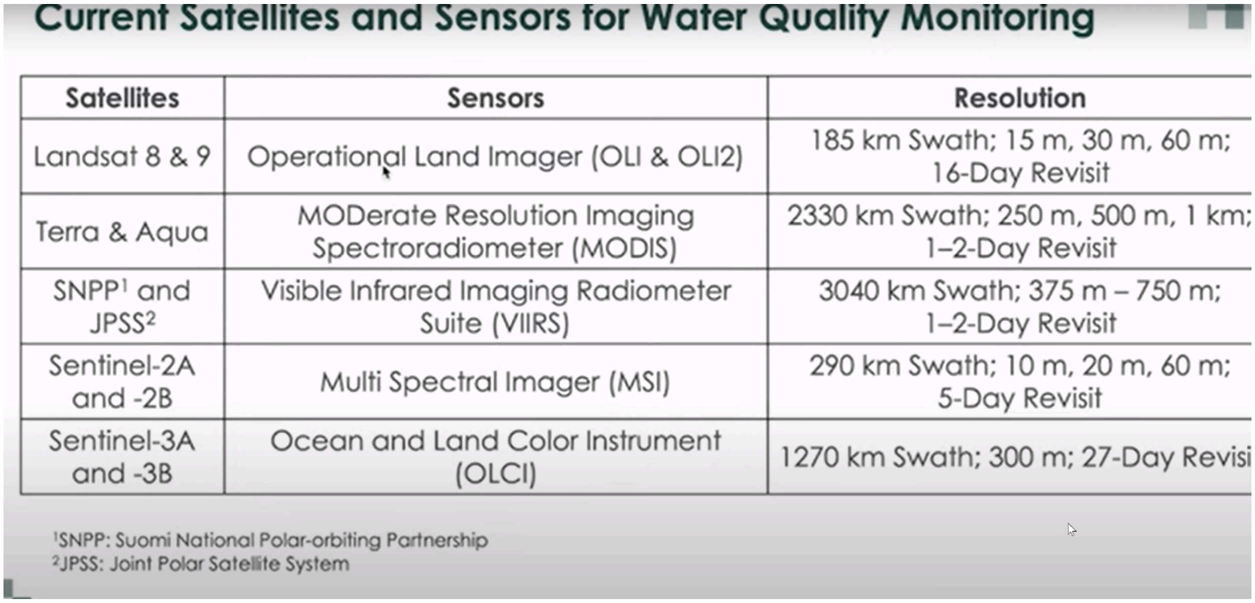
March 3rd, 2024:

- Consulted with Zainab Akhtar, a geospatial analyst with the Qatar Computing Research Institute
    - recommended we utilize Google Earth Engine to collect satellite data
    - recommended how we use the data from Google Earth Engine
        - use the data from Google Earth Engine and add it to our in-situ observations to provide more correlations and increase accuracy

March 6th, 2024:

https://www.youtube.com/watch?v=UaDybXuIW1c

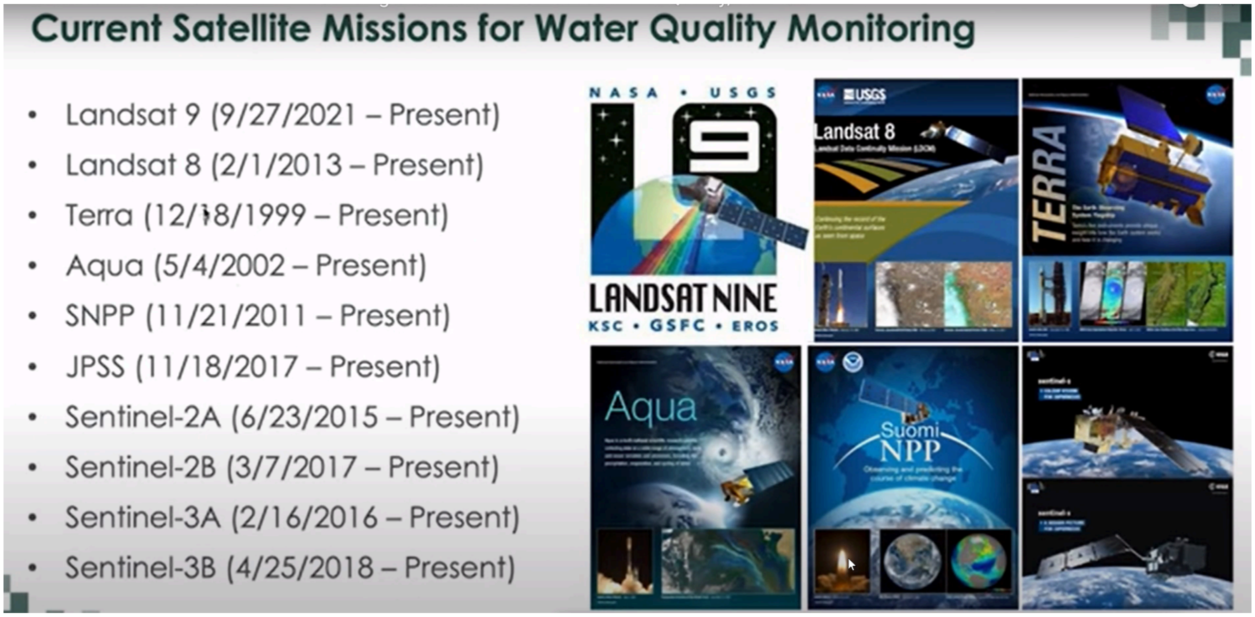
- Took notes and followed the process on this video, which is NASA’s course of Google Earth Engine remote sensing.
- Key screenshots:



Current Satellites and Sensors for Water Quality Monitoring

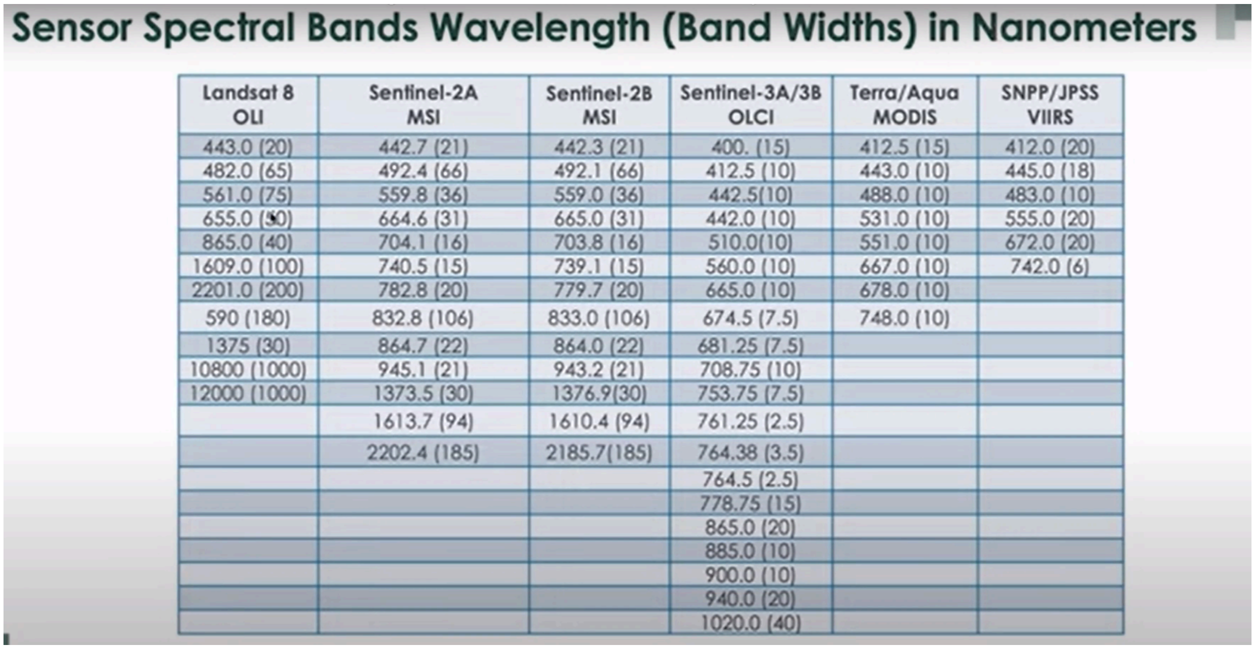
Satellites	Sensors	Resolution
Landsat 8 & 9	Operational Land Imager (OLI & OLI2)	185 km Swath; 15 m, 30 m, 60 m; 16-Day Revisit
Terra & Aqua	MODerate Resolution Imaging Spectroradiometer (MODIS)	2330 km Swath; 250 m, 500 m, 1 km; 1–2-Day Revisit
SNPP[1] and JPSS[2]	Visible Infrared Imaging Radiometer Suite (VIIRS)	3040 km Swath; 375 m – 750 m; 1–2-Day Revisit
Sentinel-2A and -2B	Multi Spectral Imager (MSI)	290 km Swath; 10 m, 20 m, 60 m; 5-Day Revisit
Sentinel-3A and -3B	Ocean and Land Color Instrument (OLCI)	1270 km Swath; 300 m; 27-Day Revisi

[1] SNPP: Suomi National Polar-orbiting Partnership
[2] JPSS: Joint Polar Satellite System



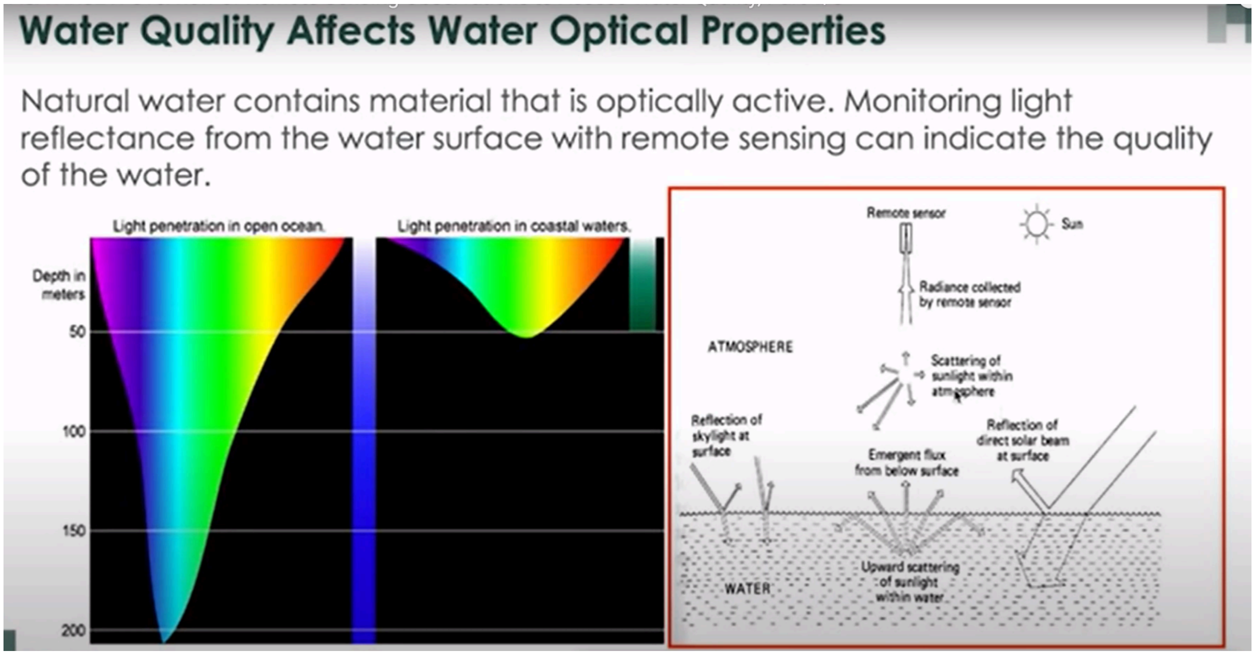
Current Satellite Missions for Water Quality Monitoring

- Landsat 9 (9/27/2021 – Present)
- Landsat 8 (2/1/2013 – Present)
- Terra (12/18/1999 – Present)
- Aqua (5/4/2002 – Present)
- SNPP (11/21/2011 – Present)
- JPSS (11/18/2017 – Present)
- Sentinel-2A (6/23/2015 – Present)
- Sentinel-2B (3/7/2017 – Present)
- Sentinel-3A (2/16/2016 – Present)
- Sentinel-3B (4/25/2018 – Present)



Sensor Spectral Bands Wavelength (Band Widths) in Nanometers

Landsat 8 OLI	Sentinel-2A MSI	Sentinel-2B MSI	Sentinel-3A/3B OLCI	Terra/Aqua MODIS	SNPP/JPSS VIIRS
443.0 (20)	442.7 (21)	442.3 (21)	400. (15)	412.5 (15)	412.0 (20)
482.0 (65)	492.4 (66)	492.1 (66)	412.5 (10)	443.0 (10)	445.0 (18)
561.0 (75)	559.8 (36)	559.0 (36)	442.5(10)	488.0 (10)	483.0 (10)
655.0 (30)	664.6 (31)	665.0 (31)	442.0 (10)	531.0 (10)	555.0 (20)
865.0 (40)	704.1 (16)	703.8 (16)	510.0(10)	551.0 (10)	672.0 (20)
1609.0 (100)	740.5 (15)	739.1 (15)	560.0 (10)	667.0 (10)	742.0 (6)
2201.0 (200)	782.8 (20)	779.7 (20)	665.0 (10)	678.0 (10)	
590 (180)	832.8 (106)	833.0 (106)	674.5 (7.5)	748.0 (10)	
1375 (30)	864.7 (22)	864.0 (22)	681.25 (7.5)		
10800 (1000)	945.1 (21)	943.2 (21)	708.75 (10)		
12000 (1000)	1373.5 (30)	1376.9(30)	753.75 (7.5)		
	1613.7 (94)	1610.4 (94)	761.25 (2.5)		
	2202.4 (185)	2185.7(185)	764.38 (3.5)		
			764.5 (2.5)		
			778.75 (15)		
			865.0 (20)		
			885.0 (10)		
			900.0 (10)		
			940.0 (20)		
			1020.0 (40)		



Water Quality Affects Water Optical Properties

Natural water contains material that is optically active. Monitoring light reflectance from the water surface with remote sensing can indicate the quality of the water.

Light penetration in open ocean.
Light penetration in coastal waters.
Depth in meters
50
100
150
200
Remote sensor
Sun
Radiance collected by remote sensor
ATMOSPHERE
Scattering of sunlight within atmosphere
Reflection of skylight at surface
Emergent flux from below surface
Reflection of direct solar beam at surface
Upward scattering of sunlight within water
WATER

March 9th, 2024:

- Consulted with Gijs van den Dool, a senior geospatial data scientist and independent researcher
    - He told us the importance of utilizing short-wave infrared (SWIR) wavelengths as well as near-infrared (NIR) and red band wavelengths in our research

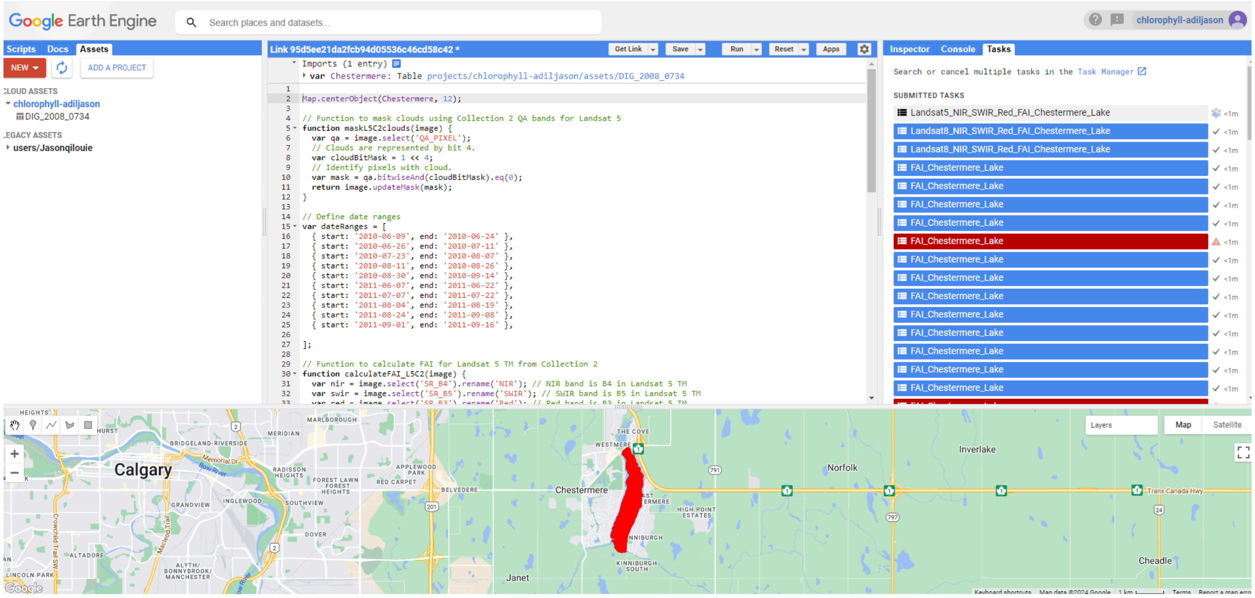
March 11th, 2024:

- Adil: problem and method were completed

- Researched health effects of cyanobacterial toxins
- Researched what “dead zones” were

March 13th, 2024:

- Adil: Finished the conclusion, analysis
- Jason: finished the satellite scraping data as well as the improved machine learning model



```
Imports (1 entry)
  var Chestermere: Table projects/chlorophyll-adiljason/assets/DIG_2008_0734

Map.centerObject(Chestermere, 12);

// Function to mask clouds using Collection 2 QA bands for Landsat 5
function maskL5C2clouds(image) {
  var qa = image.select('QA_PIXEL');
  // Clouds are represented by bit 4.
  var cloudBitMask = 1 << 4;
  // Identify pixels with cloud.
  var mask = qa.bitwiseAnd(cloudBitMask).eq(0);
  return image.updateMask(mask);
}

// Define date ranges
var dateRanges = [
  { start: '2010-06-09', end: '2010-06-24' },
  { start: '2010-06-26', end: '2010-07-11' },
  { start: '2010-07-23', end: '2010-08-07' },
  { start: '2010-08-11', end: '2010-08-26' },
  { start: '2010-08-30', end: '2010-09-14' },
  { start: '2011-06-07', end: '2011-06-22' },
  { start: '2011-07-07', end: '2011-07-22' },
  { start: '2011-08-04', end: '2011-08-19' },
  { start: '2011-08-24', end: '2011-09-08' },
  { start: '2011-09-01', end: '2011-09-16' },

];

// Function to calculate FAI for Landsat 5 TM from Collection 2
function calculateFAI_L5C2(image) {
  var nir = image.select('SR_B4').rename('NIR'); // NIR band is B4 in Landsat 5 TM
  var swir = image.select('SR_B5').rename('SWIR'); // SWIR band is B5 in Landsat 5 TM
  var red = image.select('SR_B3').rename('Red'); // Red band is B3 in Landsat 5 TM
```

```
Search or cancel multiple tasks in the Task Manager

SUBMITTED TASKS
Landsat5_NIR_SWIR_Red_FAI_Chestermere_Lake    <1m
Landsat8_NIR_SWIR_Red_FAI_Chestermere_Lake  ✓ <1m
Landsat8_NIR_SWIR_Red_FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ⚠ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
FAI_Chestermere_Lake  ✓ <1m
```

  - Snapshot of Google Earth Engine

March 14, 2024:

- Presentation was finished and the video was recorded
  - https://www.youtube.com/watch?v=jdZNgJVpvVw