Sentinel-3 Chlorophyll Concentration Validation

by

Gaganjot Kaur
B.Tech., Guru Nanak Dev University, India, 2016

A Project Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Sentinel-3 Chlorophyll Concentration Validation

by

Gaganjot Kaur
B.Tech., Guru Nanak Dev University, India, 2016

Supervisory Committee

Dr. Daniela Damian, Co-Supervisor
(Department of Computer Science)

Dr. Yvonne Coady, Co-Supervisor
(Department of Computer Science)

Supervisory Committee

Dr. Daniela Damian, Co-Supervisor
(Department of Computer Science)

Dr. Yvonne Coady, Co-Supervisor
(Department of Computer Science)

ABSTRACT

Ocean health is very crucial for the balance of the ecosystem. Therefore, continuous monitoring of oceans is an important work being undertaken by remote sensing satellites. European Space Agency's (ESA) Sentinel 3 is deployed in the earth's orbit which keeps track of chlorophyll concentration in the oceans. This project is focused on validating the chlorophyll concentration data obtained by Sentinel 3. The data collected from the satellite is compared with the data directly retrieved from the ocean with the help of British Columbia (BC) ferries. The BC ferries are equipped with instruments and sensors that estimate the amount of chlorophyll. The area of study involves coastal British Columbia especially the southern Strait of Georgia. The goal of this project is to find how correlated the two datasets are. The project is extremely data-centric and involves extensive preprocessing and exploration followed by designing of efficient methodology for validation. The validation methodology is analyzed by statistical measures such as the Pearson Correlation. This project also sheds light on the assumptions and uncertainties involved in the data collection procedures which can affect the consistency and reliability of data.

Contents

Supervisory Committee **ii**

Abstract **iii**

Table of Contents **iv**

List of Figures **vi**

Acknowledgements **viii**

Dedication **ix**

1 Introduction **1**

1.1 Agenda 2

2 Problem Formulation & Related Work **4**

3 Data Exploration and Cleaning **7**

3.1 Data retrieved from BC ferries (Ferry Dataset) 7

3.2 Data retrieved from Sentinel 3 Satellite (Satellite Dataset) 9

4 Methodology Design and Implementation **14**

4.1 Methodology 14

4.1.1 Filtering ferry data with correspondence to satellite data availability 15

4.1.2 Filtering satellite data with correspondence to the location of the ferry 15

4.1.3 Oversampling and Interpolation 16

4.1.4 Pearsons Correlation Coefficient Calculation 17

4.2 Platform Used - Compute Canada 18

4.3 Software Dependencies 22

5 Experiments, Results and Analysis **23**

5.1 Experiments 23

5.2 Results and Analysis 24

5.3 Discussion 34

6 Conclusion and Future Work **37**

Bibliography **39**

List of Figures

Figure 3.1 Metadata from BC ferries dataset downloaded from ONC website 8
Figure 3.2 Example of features and data from BC ferries dataset 8
Figure 3.3 Example of processed features and data from BC ferries dataset 9
Figure 3.4 Data from different times of the day stored in multiple netcdf files 10
Figure 3.5 Metadata obtained from netcdf file for relevant attributes (latitude, longitude and logchl) 10
Figure 3.6 Netcdf format of 'longitude' 11
Figure 3.7 Netcdf format of 'latitude' 11
Figure 3.8 Netcdf format of 'logchl' 12
Figure 3.9 Example of processed netcdf file obtained from satellite dataset 13

Figure 4.1 Data from different times of the day stored in multiple netcdf files 16
Figure 4.2 Strengths of correlation with respect to the Pearson correlations ratio 17
Figure 4.3 Normal distribution of 200 Pearson's Correlation Coefficients . 18
Figure 4.4 Three layer inception on compute canada 19
Figure 4.5 Steps involved in building singularity container 20

Figure 5.1 Box plots of Pearson Correlation Coefficients obtained by varying the radius under consideration. 24
Figure 5.2 Choropleth visualization of the BC ferry locations considered for validation over a period of two months (July, August 2018 - Blue, Green Boat markers respectively). 25
Figure 5.3 Zoomed-in Choropleth Visualization for Mount Galiano and Mayne Island Area 27
Figure 5.4 Plot comparing the satelite chl-a values wrt mean ferry chl-a values for July 21. 28
Figure 5.5 Plot comparing the satelite chl-a values wrt mean ferry chl-a values for July 25. 28

Figure 5.6 Zoomed-in Chloropeth Visualization near the port of Tsawwassen near US-Canada Border 30
Figure 5.7 Plot comparing the satelite chl-a values wrt mean ferry chl-a values for July 01. 30
Figure 5.8 Zoomed-in Chloropeth Visualization near Active Pass 31
Figure 5.9 Plot comparing the satelite chl-a values wrt mean ferry chl-a values for August 13. 31
Figure 5.10Zoomed-in Chloropeth Visualization near Swanson Channel . . 32
Figure 5.11Plot comparing the satelite chl-a values wrt mean ferry chl-a values for August 6. 33
Figure 5.12Plot comparing the satelite chl-a values wrt mean ferry chl-a values for August 10. 33

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Yvonne Coady and Dr. Daniela Damian for trust, encouragement, guidance, and patience.

Dr. Derek Jacoby for mentoring, brainstorming the possible solutions, and constant support especially in troubleshooting the issues.

My family and friends especially my partner Ikagarjot Singh for motivating me and believing in me.

DEDICATION

This project is dedicated to everyone using science and technology to make this planet a better place to live.

Chapter 1

Introduction

Ocean life is very pivotal when it comes to maintaining the right balance in biological ecosystems which further affects various lives. The aquatic life is directly linked to marine health while the human lives are affected due to economic instabilities that may occur because of declining ocean health. That is why it is important to monitor the factors and parameters that can deteriorate the quality of water be it fresh water lakes or oceans. Along with in-situ sample collection and experimentation, the satellite imagery is playing a vital role in order to crack down the elements of ocean health around the globe.

One of the major elements which is responsible in degenerating the sanctity of water bodies is chlorophyll, also known as algae. Briefly, algae is a green pigment matter which tends to float on the surface of the water bodies. The algae blooms are the result of overabundance of nitrogen and phosphorus in the water which takes place due to dumping of human, agricultural or industrial waste into the water bodies. The green pigment of algae is responsible for holding photosynthesis properties due to which it absorbs the sunlight and prevents it from penetrating inside the water bodies [19]. This also reduces the amount of oxygen in the water bodies resulting in congesting the aquatic life residing under the surface, such a phenomenon is known as eutrophication [20]. To help reduce the adverse effects of chlorophyll/algae and any other substance that tends to contaminate the water bodies and threaten aquatic life, periodic monitoring is very crucial.

This kind of monitoring is conducted by satellites revolving around the earth. Various government as well private organizations such as NASA, ESA, JAXA, SpaceX etc. are contributing their resources for the surveillance of earths surface from outer

space. The physics of spectrum and optics play a major role in designing optimal algorithms for capturing scientific measurements and analyzing the geographical parameters, for instance, the sunlight getting reflected from ocean surfaces containing algae versus the reflectance from the clean ocean surfaces will carry different spectral information which will be helpful in determining the chlorophyll concentration.

However, there are some issues that are important to be handled when dealing with satellite imagery such as atmospheric correction, data loss during downlink data transmission, cloud coverage, haze, to name a few. Therefore, it is essential to check the reliability of the data provided through satellite imagery. The most obvious way of doing it is validating with the ground in-situ data collection.

This project is focused on validating the chlorophyll concentration in the coastal waters of southern Strait of Georgia. The satellite dataset is obtained from Ocean and Land Color Instrument (OLCI) push-broom sensors deployed on Sentinel 3 satellite, which analyze the geophysical attributes through the help of bio-optical algorithms to retrieve the spectral distribution of upwelling radiance [16]. The ground data is retrieved through the Ocean Network Canada website which uploads the results of in-situ experimentation [4]. The samples are being collected by BC ferries. Later on, satellite data is validated with the information provided by the data collected directly from the earth's surface. This is explained in more detail in upcoming chapters.

1.1 Agenda

This section provides the map of the project report to get a fair idea about the structure of the report

Chapter 1 contains an introductory base of the problem followed by an overview of the structure of the document itself.

Chapter 2 describes in detail the open problem which is to be tackled together with its context, related work, overall motivation and the research question.

Chapter 3 describes the process of data exploration and cleaning.

Chapter 4 is where the methodology and its implementation is fully described followed by the steps involved to setup the platform for running experiments.

Chapter 5 includes the results and in-depth analysis of the experiments performed followed by the discussion of the uncertainties and assumptions involved.

Chapter 6 concludes the problem statement and methodology followed by future work.

Chapter 2

Problem Formulation & Related Work

When the two datasets collected by different sources are brought together for comparison, it becomes very intriguing on the part of a data enthusiast to process both the datasets by understanding the assumptions made in the data collection procedures and overcoming the underlying differences that can hamper the reliability. This involves thorough data exploration and cleaning. This project focuses on statistical analysis of the two datasets considered for validation of chlorophyll (chl-a) concentration to understand the reliability of satellite data collection procedure and the atmospheric correction algorithms applied to enhance the quality of data.

The area of study in this project is southern Strait of Georgia. The algae blooms in Strait of Georgia has always been a topic of interest. In the Strait, the ocean water is mixed with fresh water coming from various rivers especially Fraser river which brings with it lots of suspended substances like sediments, silt, etc. This populates the ocean with different concentrations of suspended materials, affect the turbidity of the ocean and also impact other factors such as algal blooms. According to the research by Jennifer et al. in [13], the Strait of Georgia has the highest concentration of chl-a where average chl-a is 2 times larger than anywhere else in the Northeast Pacific region.

Various experiments have been conducted by number of organizations and institutions to monitor algal blooms. Monitoring is either done by direct contact of equipment with the surface of water or by remote sensing. Gower and King [11] monitor the pattern of seeding of algal blooms in spring season in Strait of Georgia

using MODIS satellite images. The authors also suggested that due to turbidity and continuous varying nature of coastal waters especially in Strait of Georgia, chloropyll flouroscence imaging has proved more efficient than radiometer imaging, thereby endorsing ESA's satellite equipped with Ocean and Land Color Instrument (OLCI) sensors. Carswell et al [5] validated various atmospheric correction algorithms to understand the impact of satellite (MODIS) monitoring.

Both of the ocean monitoring procedures (in-situ and remote sensing) have advantages and disadvantages. The advantage of direct or in-situ chl-a concentration measurements is that the numbers obtained tend to be more reliable since they are directly obtained from the surface of the ocean while the disadvantage is that in-situ monitoring cannot cover vast geographical areas. Remote sensing can rectify this drawback while the data collection procedure involved can introduce a lot of noise in the signals obtained and decline the degree of reliability. Therefore, validation of data collected by satellites is required.

In the process of validation of two datasets, there are various challenges which need attention. The differences in the data structure and format of recording measurements are the major challenges. Since, the sources and the format of the two datasets are entirely different, it is quite intriguing to find one particular format which can hold the two different datasets gracefully so that it is easy to compare. The other discrepancy in the given dataset is constant fluctuation in GPS coordinates due to which it is not possible to pick and choose one particular value of latitude and longitude and map them across. It has to be dealt by selecting a range of latitudes and longitudes rather than one specific value. The values of latitudes and longitudes in satellite data vary in the order of a few meters (1 -10) while in case of ground truth (ferry) data, it varies in the order of millimeters, which as a result poses a problem of choosing a specific range for validation.

After discussing the challenges involved in data exploration process, the goal of the project has narrowed down to the following research question :

> What is the maximum correlation between the chlorophyll concentration of the data obtained from BC Ferries and data obtained from satellites in a specific geographical area?

The answer to this question will be able to throw some light on the impact of ocean currents in changing the amount of chlorophyll concentration at a particular place. It will be extremely interesting to see the statistics with respect to region, time and date of the data collected and how effective the satellite surveillance is over a particular geographical region.

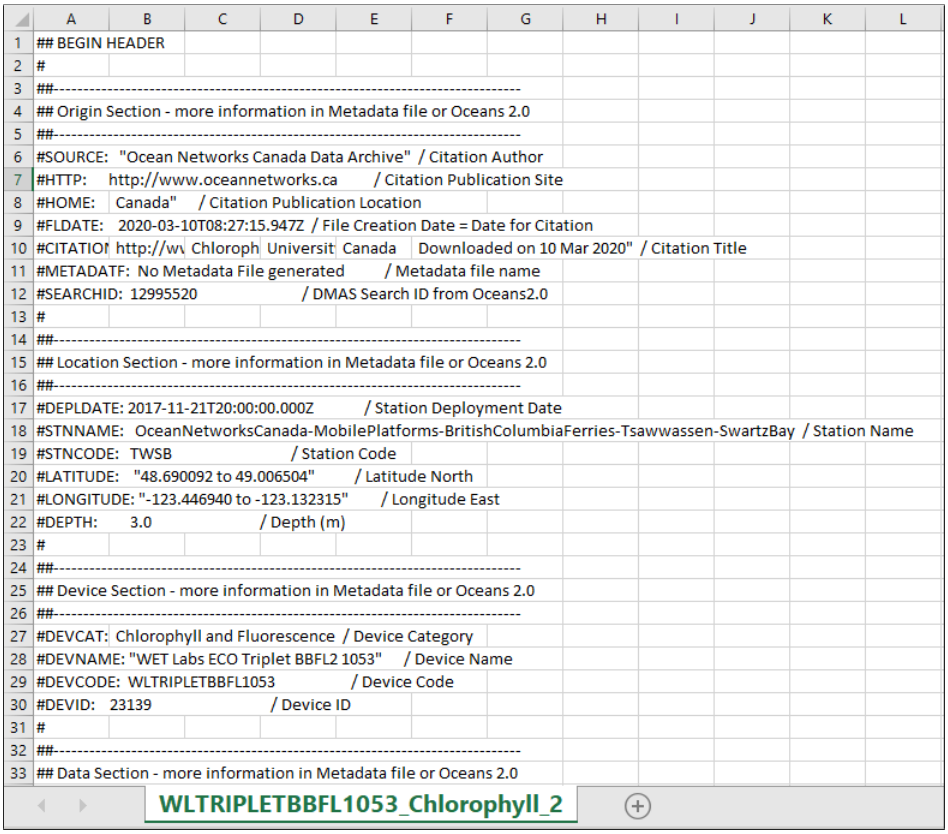
Chapter 3

Data Exploration and Cleaning

3.1 Data retrieved from BC ferries (Ferry Dataset)

To monitor the marine health in the Strait of Georgia, the BC ferries are equipped with sensors that can measure the concentration and distribution of phytoplankton also commonly referred to as algae. The area of study for this project is the southern Strait of Georgia. The BC ferry route in the southern Strait of Georgia includes the route between Tsawwassen and Swartz Bay. The fluorometers (Wetlab ECO Triplet) deployed on the ferry are responsible for estimating the amount of algae concentration on the surface or shallow depths of the ocean. The algae that resides in shallow depth or on the surface is also known as chlorophyll-a (chl-a). The fluorescence (emission of light by a substance that has absorbed light) of chl-a is used as a factor by fluorometers to estimate the amount of chl-a concentration. This data collection method is quite cost-efficient and productive. The initiative of data collection is taken by BC ferries and Ocean Networks Canada (ONC). The ONC website hosts the chlorophyll concentration data which can be downloaded in the form of CSV files.

The chl-a concentration data is collected every second of ferry operation hours. The BC ferry on this route normally operates for 14-15 hours a day and covers 8-9 round trips every day as estimated from the BC ferry schedule on the mentioned route [12]. This accounts for a plethora of data points. For this project, the dataset for two months of summer 2018, July and August were downloaded from the ONC website to serve as ground truth data in the process of validation. The dataset of July and August was automatically stored in the form of multiple CSV files. Each CSV file contains data from multiple days. The start of every CSV file has full-fledged



	A	B	C	D	E	F	G	H	I	J	K	L
1	## BEGIN HEADER											
2	#											
3	##----------											
4	## Origin Section - more information in Metadata file or Oceans 2.0											
5	##----------											
6	#SOURCE: "Ocean Networks Canada Data Archive" / Citation Author											
7	#HTTP: http://www.oceannetworks.ca / Citation Publication Site											
8	#HOME: Canada" / Citation Publication Location											
9	#FLDATE: 2020-03-10T08:27:15.947Z / File Creation Date = Date for Citation											
10	#CITATIOI http://w\ Chloroph Universit Canada Downloaded on 10 Mar 2020" / Citation Title											
11	#METADATF: No Metadata File generated / Metadata file name											
12	#SEARCHID: 12995520 / DMAS Search ID from Oceans2.0											
13	#											
14	##----------											
15	## Location Section - more information in Metadata file or Oceans 2.0											
16	##----------											
17	#DEPLDATE: 2017-11-21T20:00:00.000Z / Station Deployment Date											
18	#STNNAME: OceanNetworksCanada-MobilePlatforms-BritishColumbiaFerries-Tsawwassen-SwartzBay / Station Name											
19	#STNCODE: TWSB / Station Code											
20	#LATITUDE: "48.690092 to 49.006504" / Latitude North											
21	#LONGITUDE: "-123.446940 to -123.132315" / Longitude East											
22	#DEPTH: 3.0 / Depth (m)											
23	#											
24	##----------											
25	## Device Section - more information in Metadata file or Oceans 2.0											
26	##----------											
27	#DEVCAT: Chlorophyll and Fluorescence / Device Category											
28	#DEVNAME: "WET Labs ECO Triplet BBFL2 1053" / Device Name											
29	#DEVCODE: WLTRIPLETBBFL1053 / Device Code											
30	#DEVID: 23139 / Device ID											
31	#											
32	##----------											
33	## Data Section - more information in Metadata file or Oceans 2.0											

WLTRIPLETBBFL1053_Chlorophyll_2

Figure 3.1: Metadata from BC ferries dataset downloaded from ONC website

#"Time UTC (yyyy-mm-dd Thh:mm:ss.fff Z)"	"Chlorophyll (ug/l)"	"Chlorophyll QC Flag"	"Latitude (deg)"	"Latitude QC Flag"	"Longitude (deg)"	"Longitude QC Flag"	"Pitch (deg)"	"Pitch QC Flag"	"Roll (deg)"	"Roll QC Flag"	"True Heading (deg)"
2018-08-30T 05:45:10.974	3.8552	1	48.6901 0417	8	-123.41 24908	8	-6.5670 97097	8	3.6	8	311.354174 2

Figure 3.2: Example of features and data from BC ferries dataset

information about the data contained in that particular file, also known as metadata as described in the Figure 3.1. The metadata is then followed by the actual data. The actual data consists of rows and columns where each column contains information about the distinct parameter. The parameters include date/time, latitude, longitude, chl-a along with additional parameters like roll, pitch, degree, QC flags. Each row contains the record of an instance of data collection [Figure 3.2].

The ferry dataset needs to be processed to remove redundant and unnecessary elements such as metadata and additional parameters. As mentioned, each CSV file contains data from multiple days. To make the process of validation sorted, the data is distributed in CSV files in such a way that each CSV file contains data from one day only. This means we have 31 CSV files storing the chl-a concentration for July and another 31 CSV files storing data for August. The final data structure of the ferry dataset has all the unnecessary items eliminated and contains the relevant parameters only, which are date, time, latitude, longitude, and chl-a as shown in the Figure 3.3.

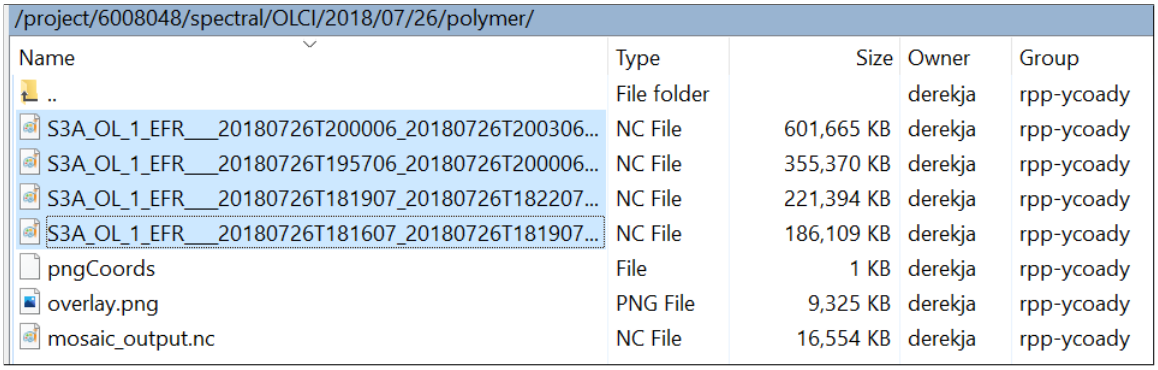
Date	Time	Latitude	Longitude	chl
2018-07-01	0:00:00	49.00648	-123.134	7.6616
2018-07-01	0:00:01	49.00648	-123.134	7.686
2018-07-01	0:00:02	49.00648	-123.134	7.6982
2018-07-01	0:00:03	49.00648	-123.134	7.7104
2018-07-01	0:00:04	49.00648	-123.134	7.7104
2018-07-01	0:00:05	49.00648	-123.134	7.7714
2018-07-01	0:00:06	49.00648	-123.134	7.7836
2018-07-01	0:00:08	49.00648	-123.134	7.8324
2018-07-01	0:00:09	49.00648	-123.134	7.8446
2018-07-01	0:00:10	49.00648	-123.134	7.8446

Figure 3.3: Example of processed features and data from BC ferries dataset

3.2 Data retrieved from Sentinel 3 Satellite (Satellite Dataset)

The dataset is collected by the European Space Agency's (ESA) satellite Sentinel 3. The satellite is deployed with Ocean and Land Color Instrument (OLCI) [16], which is a push-broom imaging spectrometer instrument used for capturing images of Earth at a spatial resolution of 300 meters. The dataset collected by push-broom sensors is affected by atmospheric disturbances such as reflectance, scattering of light, etc. which can contaminate the data captured. To appease the atmospheric effects, the atmospheric correction algorithms are used. The dataset used in the project was corrected by the 'Polymer' algorithm (by default). After atmospheric correction, the dataset was saved in netCDF file format as shown in figure 3.4. The figure shows

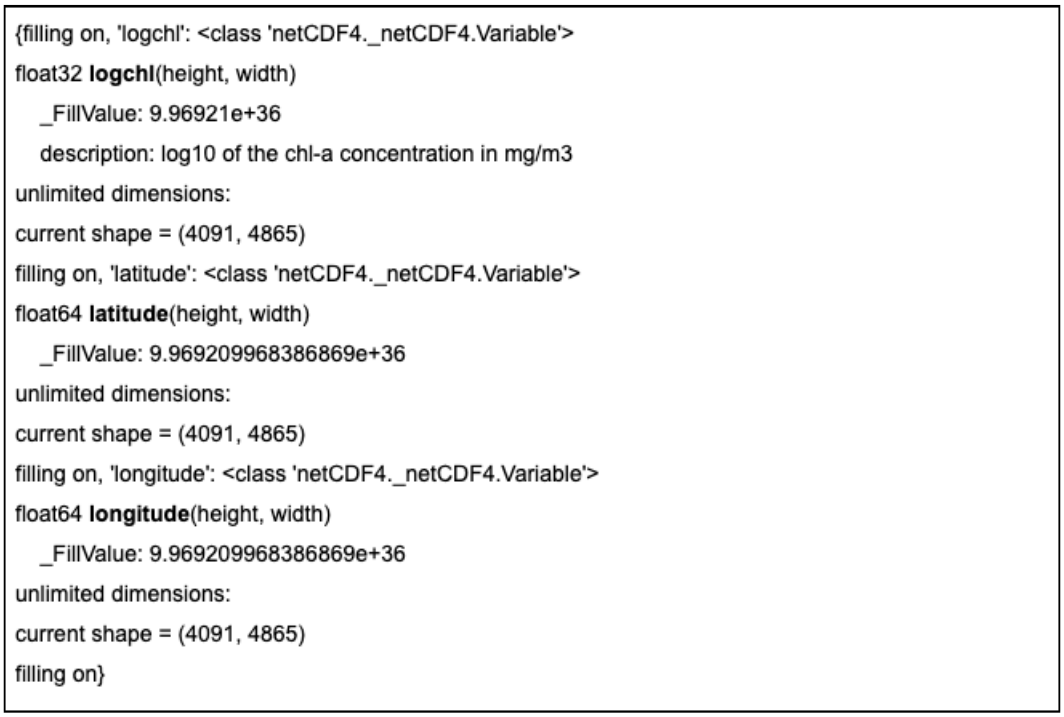
four netCDF files (highlighted in blue) representing satellite data collected on 26 July 2018. NetCDF data structure is exclusively used to store geospatial data to log the concentration of a substance at a particular coordinate on the surface of the earth. The metadata of the netCDF file is shown in Figure 3.5.



/project/6008048/spectral/OLCI/2018/07/26/polymer/

Name	Type	Size	Owner	Group
..	File folder		derekja	rpp-ycoady
S3A_OL_1_EFR___20180726T200006_20180726T200306...	NC File	601,665 KB	derekja	rpp-ycoady
S3A_OL_1_EFR___20180726T195706_20180726T200006...	NC File	355,370 KB	derekja	rpp-ycoady
S3A_OL_1_EFR___20180726T181907_20180726T182207...	NC File	221,394 KB	derekja	rpp-ycoady
S3A_OL_1_EFR___20180726T181607_20180726T181907...	NC File	186,109 KB	derekja	rpp-ycoady
pngCoords	File	1 KB	derekja	rpp-ycoady
overlay.png	PNG File	9,325 KB	derekja	rpp-ycoady
mosaic_output.nc	NC File	16,554 KB	derekja	rpp-ycoady

Figure 3.4: Data from different times of the day stored in multiple netcdf files



{filling on, 'logchl': <class 'netCDF4._netCDF4.Variable'>
float32 **logchl**(height, width)
_FillValue: 9.96921e+36
description: log10 of the chl-a concentration in mg/m3
unlimited dimensions:
current shape = (4091, 4865)
filling on, 'latitude': <class 'netCDF4._netCDF4.Variable'>
float64 **latitude**(height, width)
_FillValue: 9.969209968386869e+36
unlimited dimensions:
current shape = (4091, 4865)
filling on, 'longitude': <class 'netCDF4._netCDF4.Variable'>
float64 **longitude**(height, width)
_FillValue: 9.969209968386869e+36
unlimited dimensions:
current shape = (4091, 4865)
filling on}

Figure 3.5: Metadata obtained from netcdf file for relevant attributes (latitude, longitude and logchl)

For validating the satellite dataset, it is important to clean and pre-process it. The consequences of preprocessing is to reform the structure of the dataset so that it can

```
1 longitudes = nc.variables['longitude'][:]
2 longitudes
```

```
masked_array(
  data=[[-140.083383, -140.07938 , -140.075378, ..., -121.496959,
         -121.493381, -121.489804],
        [-140.083765, -140.079763, -140.075761, ..., -121.498366,
         -121.494789, -121.491212],
        [-140.084148, -140.080146, -140.076144, ..., -121.499774,
         -121.496197, -121.49262 ],
        ...,
        [-141.694965, -141.691704, -141.688442, ..., -126.332651,
         -126.329619, -126.326588],
        [-141.695371, -141.692109, -141.688848, ..., -126.333655,
         -126.330623, -126.327592],
        [-141.695777, -141.692515, -141.689254, ..., -126.334658,
         -126.331627, -126.328596]],
  mask=False,
  fill_value=1e+20)
```

Figure 3.6: Netcdf format of 'longitude'

```
1 latitudes = nc.variables['latitude'][:]
2 latitudes
```

```
masked_array(
  data=[[52.982373, 52.982151, 52.981929, ..., 50.405646, 50.404827,
         50.404009],
        [52.979813, 52.979591, 52.979369, ..., 50.403146, 50.402327,
         50.401509],
        [52.977254, 52.977031, 52.976809, ..., 50.400646, 50.399828,
         50.39901 ],
        ...,
        [42.505361, 42.505076, 42.504791, ..., 40.078929, 40.078227,
         40.077525],
        [42.502795, 42.50251 , 42.502224, ..., 40.076383, 40.075681,
         40.074979],
        [42.500229, 42.499944, 42.499658, ..., 40.073837, 40.073135,
         40.072433]],
  mask=False,
  fill_value=1e+20)
```

Figure 3.7: Netcdf format of 'latitude'

be easily accessed during the process of validation. NetCDF is a multi-dimensional data structure where latitude, longitude, and chl-a concentration (logarithmic scale) are stored in a grid format as shown in the figures 3.6, 3.7, 3.8. Therefore, pre-processing of the satellite dataset will transform it into a different data structure similar to that of a ferry dataset such that both the datasets are in concordance. The preprocessing was performed with the help of python libraries such as 'xarray' and pandas by reducing the dimensionality of the NetCDF files and storing them in a 2-dimensional dataframe, respectively.

```
1 chl = nc.variables['logchl'][:]
2 chl
```

```
masked_array(
  data=[[--, --, --, ..., --, --, --],
        [--, --, --, ..., --, --, --],
        [--, --, --, ..., --, --, --],
        ...,
        [--, --, --, ..., --, --, --],
        [--, --, --, ..., --, --, --],
        [--, --, --, ..., --, --, --]],
  mask=[[ True,  True,  True, ...,  True,  True,  True],
        [ True,  True,  True, ...,  True,  True,  True],
        [ True,  True,  True, ...,  True,  True,  True],
        ...,
        [ True,  True,  True, ...,  True,  True,  True],
        [ True,  True,  True, ...,  True,  True,  True],
        [ True,  True,  True, ...,  True,  True,  True]],
  fill_value=9.96921e+36,
  dtype=float32)
```

Figure 3.8: Netcdf format of 'logchl'

Not only the dimensionality of the data structure had to be changed, but other crucial changes were also required to make it relevant and good to use for validation. Sentinel-3 records data multiple times every day (saved in a netCDF file after going through atmospheric correction procedure). Every time OLCI push-broom sensors capture/record data for exactly three minutes and that recorded data can be of any part of the globe. Since only the southern Strait of Georgia is monitored in this project, we had to limit the satellite dataset to include only those data points which are present in the area of study or more specifically along the mentioned BC ferry route. The constraint was applied by filtering the data points in each file whose coordinates lie between Tsawwassen Terminal and Swartz Bay.

After simplifying the data structure and applying the constraints, the dataset was saved in CSV files. This process was repeated for every day of July and August 2018. After applying the constraints, it was observed that not every day of two months yielded data points as the satellite might not have hovered over the area under consideration (ferry route between Tsawwassen and Swartz Bay) on some of the days.

The cleaned dataset was saved in a similar hierarchy structure followed for the ferry dataset where each CSV file contained data for only one day. Every CSV file has parameters like date, start time (when push-broom sensor started capturing the data),

date	starttime	endtime	latitude	longitude	chl
2018-07-01	19:07:44	19:10:43	48.99936	-123.13	5.454649
2018-07-01	19:07:44	19:10:43	48.99936	-123.13	5.431354
2018-07-01	19:07:44	19:10:43	48.99847	-123.126	4.910563
2018-07-01	19:07:44	19:10:43	48.99758	-123.122	3.517093
2018-07-01	19:07:44	19:10:43	48.99669	-123.118	4.595998
2018-07-01	19:07:44	19:10:43	48.9995	-123.144	4.311829
2018-07-01	19:07:44	19:10:43	48.99861	-123.14	3.632063
2018-07-01	19:07:44	19:10:43	48.99773	-123.136	3.48322
2018-07-01	19:07:44	19:10:43	48.99684	-123.132	3.98394

Figure 3.9: Example of processed netcdf file obtained from satellite dataset

end time (when push-broom sensor stopped capturing the data), latitude, longitude, chl-a (anti-log of original chl-a values) as shown in the figure 3.9.

This chapter covered the process of exploration and cleaning of both the satellite data and ferry data in detail. The process of cleaning filters out unnecessary elements from the dataset and makes it ready for validation.

Chapter 4

Methodology Design and Implementation

The methodology cannot be implemented directly on the existing structures of datasets. Therefore, the datasets were processed and cleaned first as described in chapter 3. The methodology designed for validation was applied to the cleaned datasets to find out the maximum correlation. In the following sections, the methodology and its implementation is described.

4.1 Methodology

To exactly track down the chl-a concentration in both datasets, we have to match the spatial and temporal parameters of both datasets. This is achieved by following two steps executed in order:

1. Filtering the ferry dataset considering the temporal parameters of satellite dataset

2. Filtering the satellite dataset according to the spatial parameters of the *filtered* ferry dataset

After filtering, we can ensure that the geographical area covered by the data points in filtered satellite data and filtered ferry data roughly coincide with each other at the same date and time. The procedure of filtering is elaborated in the upcoming subsections, followed by the interpolation of datasets and finding Pearson Correlation.

4.1.1 Filtering ferry data with correspondence to satellite data availability

In the previous process of data cleaning and preprocessing, the ferry data was saved in CSV files where each CSV file contained the data for only one day. The ferry operates for 14-15 hours a day and the fluorometers deployed in the ferry collect the data every second while the satellite captures the data for exactly three minutes (180 seconds), multiple times a day. ESA states that OLCI sensors do not capture the same geographical region more than once on a single day [17] which means if the OLCI sensors have captured the southern Strait of Georgia once in a day, they are not going to capture it again on the same day. This ensures that we will only have one value for 'start-time' (the time when push-broom sensors started capturing the earth's surface) and one value for 'end-time' (the time when push-broom sensors stopped capturing the earth's surface) for a given day. We are only interested in the ferry data points collected between this start-time and end-time as that satisfies the temporal constraint.

Eventually, the ferry dataset is filtered which now contains 180 data points (for every day in which satellite is hovering over the BC ferry route in southern Strait of Georgia), that are gathered between the start-time and end-time of the push-broom sensors.

4.1.2 Filtering satellite data with correspondence to the location of the ferry

Time and geographical location are the two essential factors in the whole process of validation whose accuracy will yield reliable results. After filtering ferry data with respect to time, satellite data is filtered to extract relevant information with respect to the geographical location of the ferry such that only the satellite data points which are close to the geographical location of the ferry are taken into consideration.

This is achieved by drawing an imaginary circle whose center is the average location of the ferry (in the three-minute traversal) and radius is half of the euclidean geodesic distance between the ferry location at the start time and ferry location at the end time in the three-minute traversal. The approximate displacement of a ferry in the three minutes is 1.5 Km. The imaginary circle decides the vicinity of the ferry

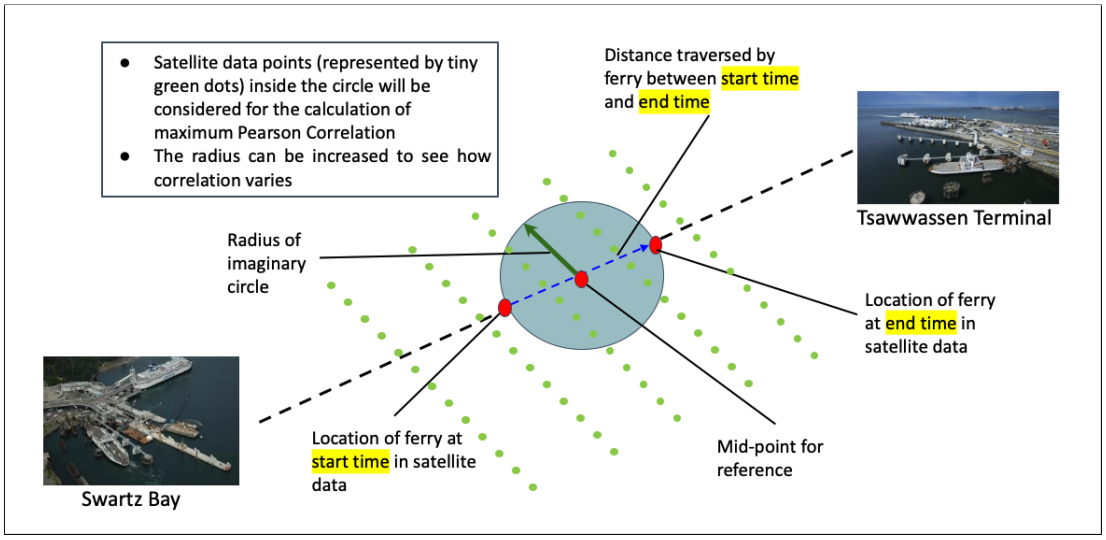


Figure 4.1: Data from different times of the day stored in multiple netcdf files

where the radius of the circle is a variable and can be changed to increase or decrease the area of the vicinity. Figure 4.1 provides a pictorial representation of this concept.

Drawing an imaginary circle is followed by filtering the satellite data in order to include the data points which lie inside or on the periphery of the imaginary circle. This is achieved by finding the geodesic distance to all the satellite data points from the average ferry location and sorting it in an ascending order to choose the closest ones that lie in the vicinity.

4.1.3 Oversampling and Interpolation

The next step is to find the Pearsons Correlation Coefficient of the two datasets obtained in order to know how divergent they are from each other. The concerning issue here is that to find correlation which involves finding covariance first, the length of the two datasets has to be the same. The filtered ferry dataset has approximately 180 data points whereas the filtered satellite dataset has fewer data points in comparison to ferry data, due to sparsely located satellite data points. In order to make the lengths of the two datasets equal, it is necessary to interpolate the data points in the satellite dataset on the basis of already existing points. Hence, to approximate the process of interpolation, random oversampling is done creating 200 random states with the help of the following code:

```
sat_sample = sat_filtered.sample(len(ferry_filtered),
                                 random_state=200,
                                 replace=True)
```

This process is followed by calculating Pearson's correlation coefficient of the interpolated satellite samples with respect to filtered ferry data.

4.1.4 Pearsons Correlation Coefficient Calculation

The Pearsons correlation coefficient is the covariance of the two datasets divided by the multiplication of the standard deviations of the two datasets. It gives a measure of how correlated two datasets are.

Absolute Value of r	Strength of Relationship
$r < 0.3$	None or very weak
$0.3 < r < 0.5$	Weak
$0.5 < r < 0.7$	Moderate
$r > 0.7$	Strong

Figure 4.2: Strengths of correlation with respect to the Pearson correlations ratio

The range of the correlation lies between -1 and 1 inclusive. If the range of the correlation is negative, it implies that both the datasets have reverse correlation or in other words, the datasets are following inversely proportional trends. The positive correlation means there is some correlation. As the correlation coefficient's value increases from 0 to 1, the strength of the correlation increases. For validation purposes, the two data sets should be in concordance directly rather than inversely. Figure 4.2 shows the degree of strength of the correlation with respect to the ratio r.

The average of the 200 correlations is taken and used for validation. The plot of 200 correlations gives normal distribution as shown in Figure 4.3 for one of the days. The average correlation is negative which states that for this day, time, or location, the algae concentration of the two datasets is very divergent. The data from the other days also gives similar distribution but with different means and standard deviations.

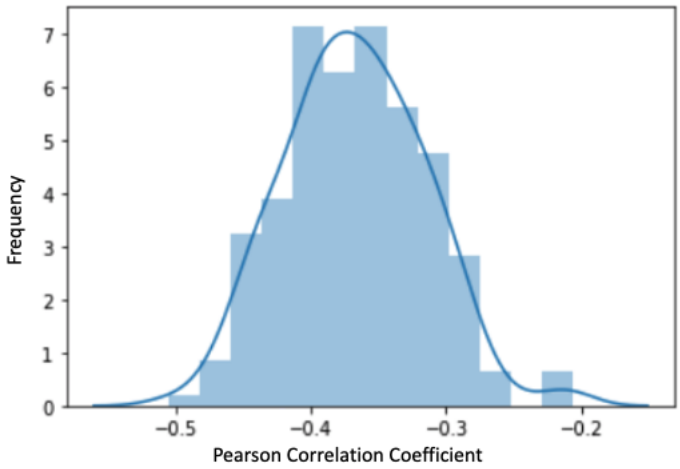


Figure 4.3: Normal distribution of 200 Pearson's Correlation Coefficients

4.2 Platform Used - Compute Canada

For the implementation of data exploration and data validation methodologies, working on local systems was not fruitful due to multifarious factors. Saving large netCDF files on local disk required a voluminous storage even for running an example on a small dataset. Not only this, reading these humongous files became computationally expensive on the local system which lead to memory errors after every few trials. Henceforth, it was necessary to make a transition to the cloud platform, that is when Compute Canada came into the picture. Compute Canada is a cloud computing platform that is free for researchers in Canada.

Compute Canada in general is divided into various clusters distributed across Canada. This project has been implemented on the Cedar cluster. The Cedar cluster is a cluster of CPUs and GPUs that has 94,528 CPU cores and 1352 GPU devices in total. The base cluster comprises 576 nodes with 32 cores each and 128 GB of memory [2]. There are two types of jobs that can be implemented on Compute Canada, batch jobs and interactive jobs. For debugging purposes, it is best to use interactive jobs which can be executed with 'salloc' command whereas to submit a job to be completed on its own, it is advisable to use batch jobs whose commands are written collectively in a bash script. To get the allocation of clusters quickly, the requested time for allocation must be less than or equal to three hours as Compute Canada follows the fair tree algorithm to give priority and to ensure a fair share of resources [3].

While working on Compute Canada, it is important to understand the hierarchy of the work environment. There is a three-layer inception while working on Compute Canada as described in the Figure 4.4.

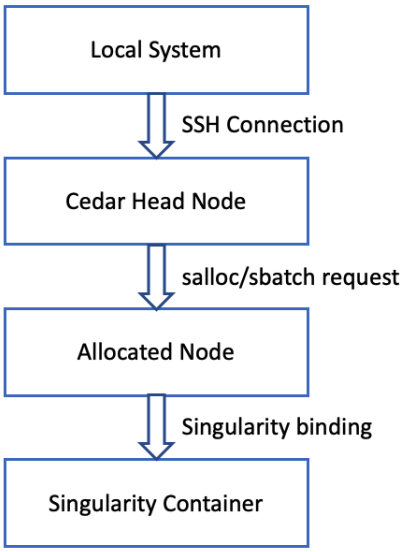


Figure 4.4: Three layer inception on compute canada

Initially, the user is at her/his local system followed by setting up a remote connection with the SSH protocol. SSH remote connection with the cedar cluster of Compute Canada lets the user enter into the Cedar cluster. Once in the cluster, the user will require a CPU/node to start working on, which can be allocated by salloc or sbatch command. Once the node is allocated, it can be considered as a personal computer where we need to install everything from scratch. It is not possible to install the requirements on the allocated node directly as it is the allocation is available tentatively for few hours only. In order to accomplish that, there is a requirement of containerization. In this project, singularity containers are used to fulfill the requirement. A singularity container contains an image of a system that has all the requirements installed to run the jobs successfully.

There are a bunch of actions that are essential to be taken before running the python scripts on the singularity container. The Figure 4.5 shows a flow of actions required to build a container.

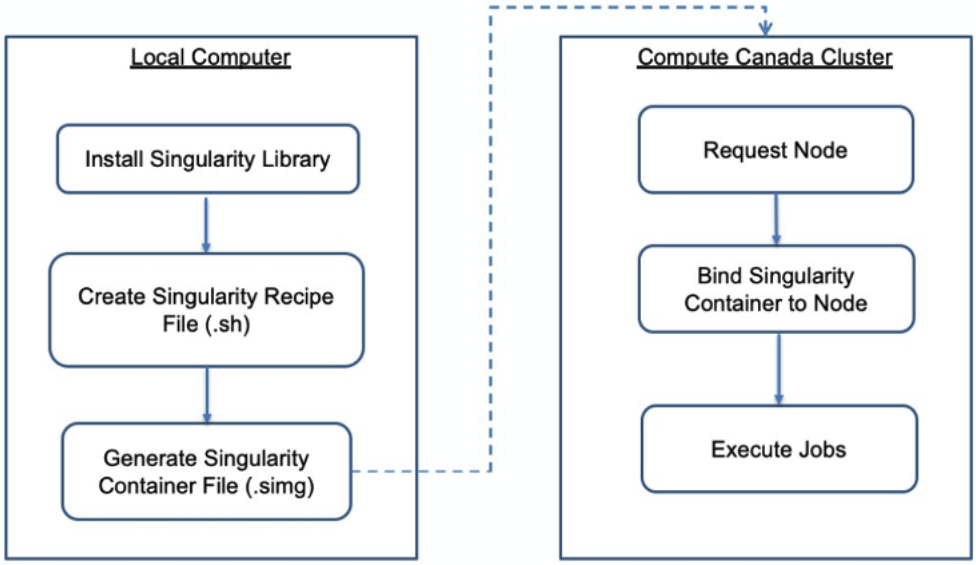


Figure 4.5: Steps involved in building singularity container

Singularity containers are only compatible with Linux. Compute Canada only allows the user to build the singularity container files on its system using the commands such as

```
sudo apt-get install -y singularity-container
```

The commands can be run directly on Compute Canada if the user is one of the admins of the Compute Canada cluster. For those who are not the admins need to build singularity on their Linux machine and transfer the resultant singularity container file on the Compute Canada platform [18].

In order to build the singularity container file on the local machine (Linux), the user has to first install singularity package. Building a singularity file is preceded by writing a singularity recipe. The recipe file has a proper format consisting of a header and a body. The header of the recipe file consists of the information on the type and version of the operating system on the container, to begin with. The body of singularity recipe is the accumulation of tiny scriptlets put together in a single shell script file containing environment variables, software packages, libraries and dependencies required for running the coding scripts. Here is the example of a tiny scriptlet for installing python3, python3 dependencies and useful python3 libraries.

```
# Header
BootStrap: debootstrap
OSVersion: xenial
MirrorURL: http://us.archive.ubuntu.com/ubuntu/
# Body
%post
# get and install python3 with various libraries
    sed -i 's/$/ universe/' /etc/apt/sources.list
    apt-get -y install software-properties-common build-essential
    apt-get -y update
    apt-get -y install python3 wget python3-pip
    add-apt-repository -y ppa:ubuntugis/ppa
    apt-get -y update
    pip3 install jupyter matplotlib numpy pandas scipy
    pip3 install netcdf4
    pip3 install xarray
```

After writing the singularity recipe file, the next step is to build a singularity container which will actually consist of all the defined packages, libraries, and dependencies. The file is generated by the following command

```
$sudo singularity build python_recipe_gagan.simg python_recipe.sh
```

The file named *'python_recipe_gagan.simg'* is then moved to Compute Canada platform. The container is executed over the node instance by the command :

```
$singularity shell --bind /project/rpp-ycoady/spectral:
/spectral /project/rpp-ycoady/spectral/gagan26
/python_recipe_gagan.simg
```

The above command is required to run every time when a new node is allocated in order to infuse a singularity container in it. Then after, we are ready to run jobs over Compute Canada.

4.3 Software Dependencies

The basic libraries which are used in every data analysis project are 'numpy', 'pandas', 'matplotlib', 'seaborn', 'glob', 'os' etc. There are two other libraries that are being majorly used in this project due to the type of the dataset involved. These are pythons 'netcdf4' [14] and 'xarray' [7] libraries. NetCDF is a set of interfaces for array-oriented data access and is used for storing multidimensional scientific variables such as pressure, humidity, temperature, chlorophyll, turbidity, etc.

The satellite images captured by push-broom sensors on-board are in the form of an image (.PNG) file. The geospatial image file is computationally and algorithmically expensive to analyze. In order to reduce the complexity of the process, there are special libraries to convert the information presented in the image file in the form of an array of numbers. The key components of geospatial images are latitude, longitude, and the attribute representing any event, object, or phenomena. In order to retain the simplicity, this is managed with the help of a python library called 'netcdf4' which stores the geospatial information in the form of masked arrays. 'Xarray' is another library that allows storing the grid structured data from 'netCDF' files into a 2-dimensional dataframe which can be further manipulated with the pandas library easily.

For visualization purposes, the choropleth maps are generated using a python library named folium [9] which have different plugins for various kinds of representations such as boat markers and circle markers. 'GeoPy' [6] is another python library used to calculate the euclidean distance between two coordinates on the earths surface.

Chapter 5

Experiments, Results and Analysis

After the design and implementation of the methodology, the next step is to test the methodology by performing multiple experiments and analyzing the results obtained to conclude the findings. This chapter provides a deeper insight to understand whether the data collection techniques used by BC ferries as well as Sentinel 3 satellite, are correct or there is something that needs to be changed/considered going forward for achieving more reliable output.

5.1 Experiments

To analyze the outcome of methodology, experiments were done on the two months of data (July and August 2018). The only variable in the whole experiment is the radius of the imaginary circle, as described in Figure 4.1. Four experiments were done on two months of data where the radius was incremented after every experiment. Increasing the radius will increase the area of the imaginary circle and hence, the number of satellite data points lying inside the imaginary circle will increase too, which means more satellite data points in the vicinity of the ferry will be considered for validation. (For more details about radius, refer to section 4.1.2)

Let us assume the radius of the imaginary circle is r where r is half of the distance traversed by BC ferry in 3 minutes [Figure 4.1]

Experiment 1 : radius = r

Experiment 2 : radius = 2 x r

Experiment 3 : radius = 3 x r

Experiment 4 : radius = 4 x r

All the experiments were conducted for every day of July and August. In the traversal of three minutes for each day, if the satellite hovers over the geographical region where the ferry is sailing, then the correlation is calculated for the data points gathered in those three minutes.

To reiterate, each day (from July and August 2018) will **either** have data points for just three minutes of the day when the satellite passed over the southern Strait of Georgia and captured data of that region in which BC ferry is sailing **or** have no data points at all due to the reason that satellite did not pass over the southern Strait of Georgia on that day. This implies that there will be just one correlation coefficient for a day for which the satellite crosses the southern Strait of Georgia or none otherwise.

5.2 Results and Analysis

All the correlation coefficients obtained for July 2018 and August 2018 are analyzed for each experiment and their distribution is shown in the box plots in Figure 5.1. A box plot is a graphical representation of the distribution of numerical data. The

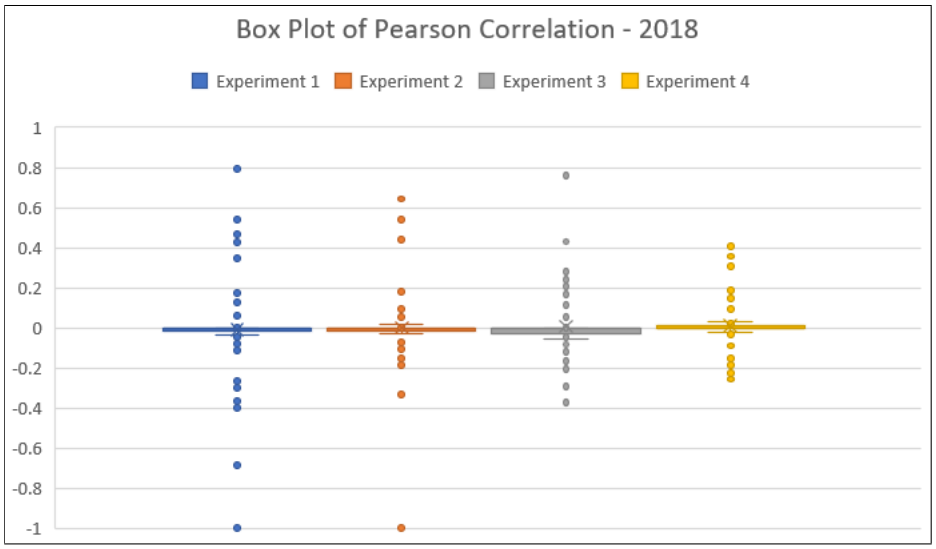


Figure 5.1: Box plots of Pearson Correlation Coefficients obtained by varying the radius under consideration.

box plot contains quartiles, whiskers, and outliers. The quartiles divide the number of data points into four more or less equal parts. The whiskers are lines indicating variability outside the first and last quartile whereas outliers refer to the data points which happen rarely.

Figure 5.1 provides a graphical analysis of the distribution of Pearson Correlation for two months of data. It contains four distinct box plots each showing distribution of one of the experiments. The plot does not show any significant difference between the results obtained from all the four experiments. The whiskers, quartiles, and means are mostly concentrated around 0 mg/m^3. In experiment 1, there are comparatively more instances that show moderate to strong correlation as there exist more outliers above 0.4 mg/m^3.

From the box plot analysis of four experiments each having a different radius factor, we can say that increasing or decreasing the radius of the imaginary circle has no significant impact on the correlation coefficients. Therefore, for further analysis and visualization, it is reasonable to choose the radius factor of 1 (pertaining to experiment 1).

The data is also visualized on the map to have an idea of how the correlation is changing with respect to geographical location. The geographical location impacts the correlation coefficients a lot. This might be due to the fact that at certain locations,

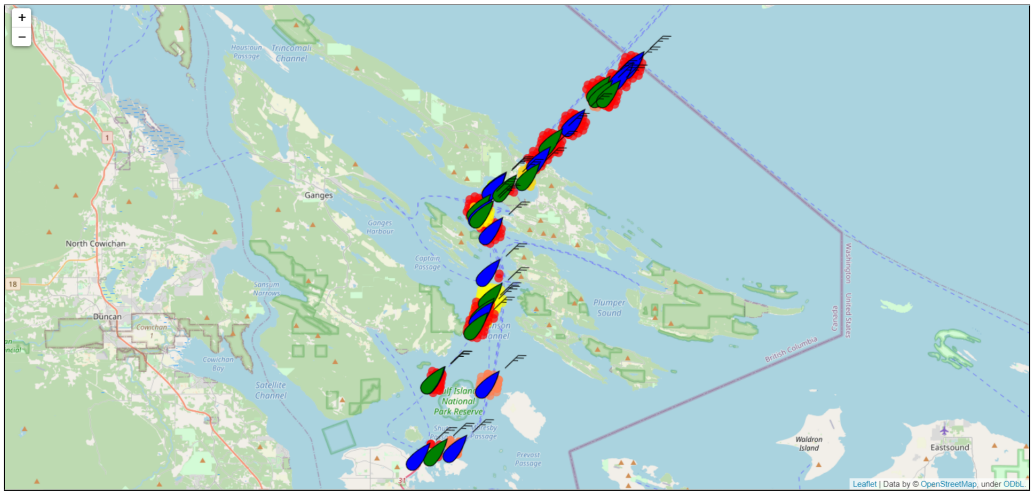


Figure 5.2: Choropleth visualization of the BC ferry locations considered for validation over a period of two months (July, August 2018 - Blue, Green Boat markers respectively).

the water currents are very rapid due to which the chl-a concentration changes very quickly resulting in random fluctuations in the chl-a concentration.

The visualization of the geospatial parameters on a map is known as choropleth visualization. The choropleth map in the Figure 5.2 is focused on the BC ferry route. The boat marker denotes the average location of the ferry (taken for reference) in three minutes. The blue and green boat markers are a symbol of data collected in July and August, respectively. The cluster around the boat marker is the visualization of an imaginary circle. Each small dot in the cluster represents a satellite data point which is considered for validation. Clusters are assigned colors according to the intensity of the Pearson Correlation. The red cluster signifies no correlation at all, the orange cluster signifies a very weak correlation, gold means slightly weak correlation, yellow refers to strong correlation and green signifies a very strong correlation. The intensity of correlation depends on the value of the correlation coefficient which can vary from -1 to 1 as described in Figure 4.2.

From the choropleth visualization in the Figure 5.2, the correlation coefficients varied a lot from one region to another. Therefore, it is intriguing as well as crucial to dive deeper into the statistics behind various geographical regions on the BC ferry route (on the southern Strait of Georgia). The following segment is a detailed analysis of chl-a concentration estimations in different regions of the southern Strait of Georgia.

1 Between Mount Galiano and Mayne Island

There are two instances when the location of ferry and area of surveillance of satellite coincide near Mayne Island, very near to Georgeson Bay. The two instances occur on 21 July 2018 and 25 July 2018. The instances show the correlation coefficient of 0.429 (weak correlation) on 21 July and 0.79 (very strong correlation) on 25 July. The zoomed-in choropleth visualization in Figure 5.3 shows the area near the Mayne Island and the two blue boat markers almost overlapping each other signifies that the instances carry data/information about chl-a concentration from identical regions occurring on the two mentioned dates. The green dots (signifying very strong correlation) and golden yellow dots (signifying weak correlation) represent satellite data points in the close vicinity of the BC ferry location (symbolized by blue boat markers).

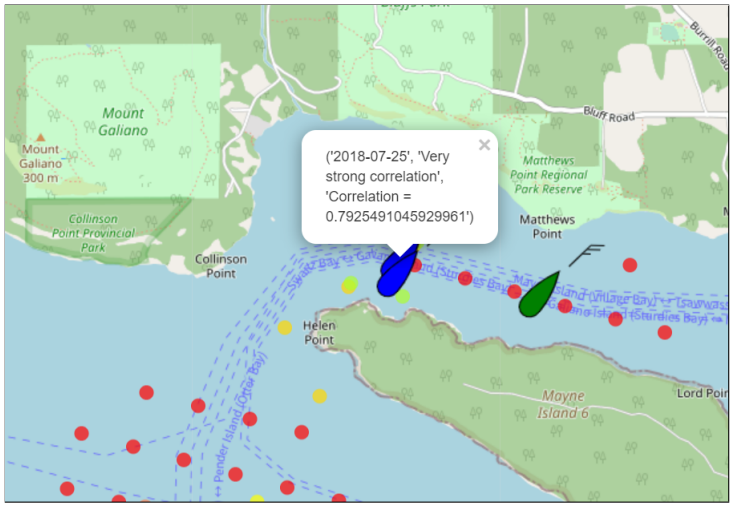


Figure 5.3: Zoomed-in Choropleth Visualization for Mount Galiano and Mayne Island Area

To further analyze the data in the map visualization, the chl-a concentration of satellite data points and their distance from the BC ferry reference location is fetched. Also, the ferry data in the three-minute window is fetched. The mean and standard deviation of chl-a concentration from ferry data is calculated and plotted. The chl-a concentration observed by satellite data points are also plotted on the same plot with respect to their distance (in kilometers) from the reference ferry location. The plot in figure 5.4 describes how the satellite chl-a varies as the distance from the reference location of the ferry increases.

For 25 July 2018, Figure 5.5 shows that upto a distance of approximately 1 Km from the location of the BC ferry, the satellite chl-a concentration is close to the average ferry chl-a concentration. The yellow dashed lines signify two standard deviations above and below the average ferry chl-a. The data from 25 July 2018 shows a very strong correlation when the radius factor is 1 (pertaining to experiment 1) and it starts decreasing in further experiments. This pattern is evident from the plot as the satellite chl-a concentration remains within the 3 standard deviations for upto 1 Km while it starts to deviate more as the distance from the reference ferry location increases. A similar statement is not true for the information collected on 21 July 2018 as the chl-a concentration fluctuates a lot around the mean ferry chl-a concentration in that region on 21 July, as shown in the Figure 5.4.

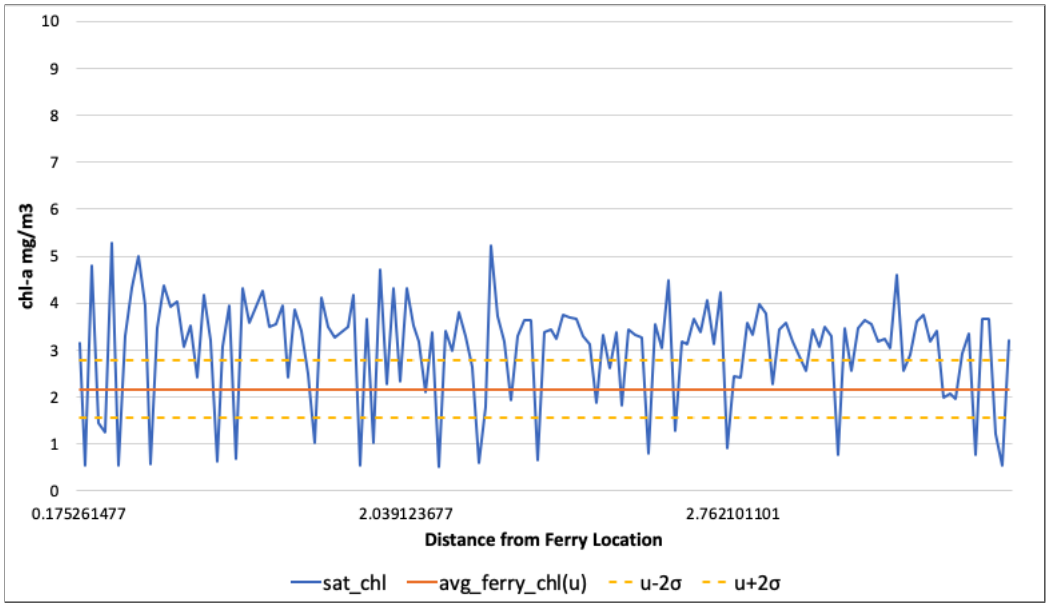


Figure 5.4: Plot comparing the satelite chl-a values wrt mean ferry chl-a values for July 21.

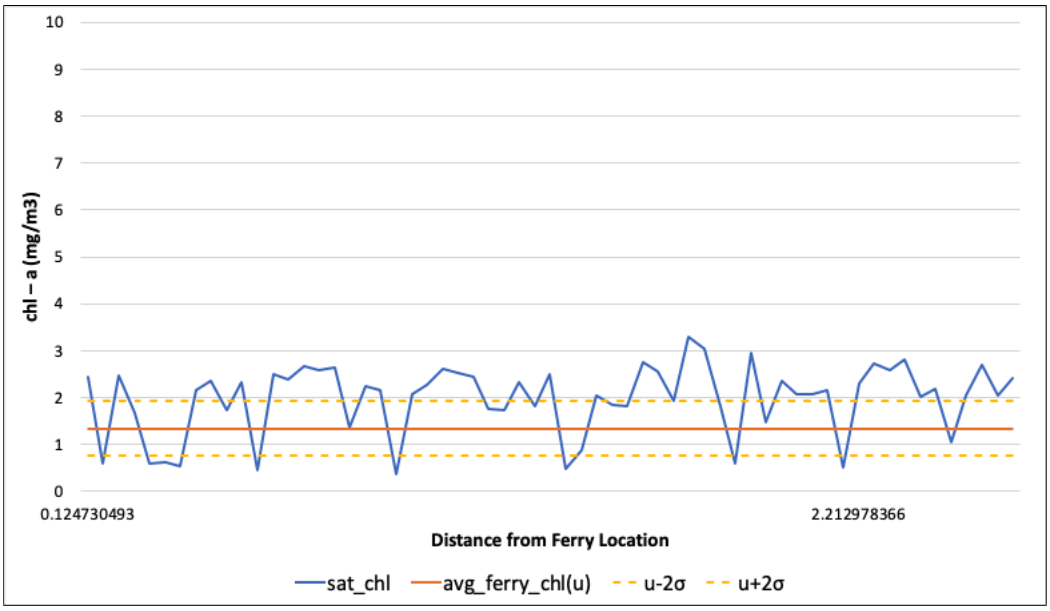


Figure 5.5: Plot comparing the satelite chl-a values wrt mean ferry chl-a values for July 25.

The map visualization in Figure 5.2 shows the location of the ferry just between Mount Galiano and Mayne Island. This area is known as Active Pass. The ferry location is just near the southern end of the Active Pass where the ocean currents are very rapid and the concentration of chlorophyll is subjected to

change rapidly. The wind speed in the Active Pass is also very high because it is a narrow passage surrounded by islands on both sides. Since the narrow opening of the active pass experiences high rapids, the chl-a concentration is ought to change instantaneously. Therefore, it was expected to get low chl-a correlation in this region especially, which suggests that information collected on 21 July 2018 is more relevant with the circumstances taking place in the Active Pass opening than that collected on 25 July 2018.

It is very important to check whether the satellite's OLCI sensors or atmospheric correction methods are overestimating before finding the correlation with the data obtained from BC ferries. In the satellite data obtained on 21 July 2018 collected from Active Pass near Mayne Island, there existed some extreme outliers that claimed the chl-a concentration to be 149 mg/m^3, which can never match the real scenario at that particular geographical region. This finding suggested that the data exploration process should involve removing of the outliers which can threaten the validity. After discovering the existence of the extreme outliers, the threshold was set such that the satellite data points having chl-a concentration lesser than 50 mg/m^3 are only involved for validation process. After executing the change and setting a new threshold for outliers, a slight improvement in the strength of the correlation was observed.

2 Near the port of Tsawwassen and near US-Canada Border

The map visualization (zoomed-in) in Figure 5.6 shows the satellite data points in red around the BC ferry location near the Tsawwassen Terminal, which signifies that there is no correlation of the ferry data points and the satellite data points in that area. There can be multiple reasons behind the poor correlation coefficients around that area which are discussed in the next section. The instances that witness the lack of correlation near Tsawwassen Terminal are collected on 1 July 2018, 28 July 2018, and 28 August 2018.

The graphical analysis of one of these dates (1 July 2018) is represented in the plot shown in Figure 5.7. We can see from the plot that the average chl-a of ferry data points (in three-minute duration) is around 2 mg/m^3 and the yellow dashed lines depicting chl-a value for two standard deviations above and below average chl-a are also closer to 2 mg/m^3 but the satellite chl-a concentration is

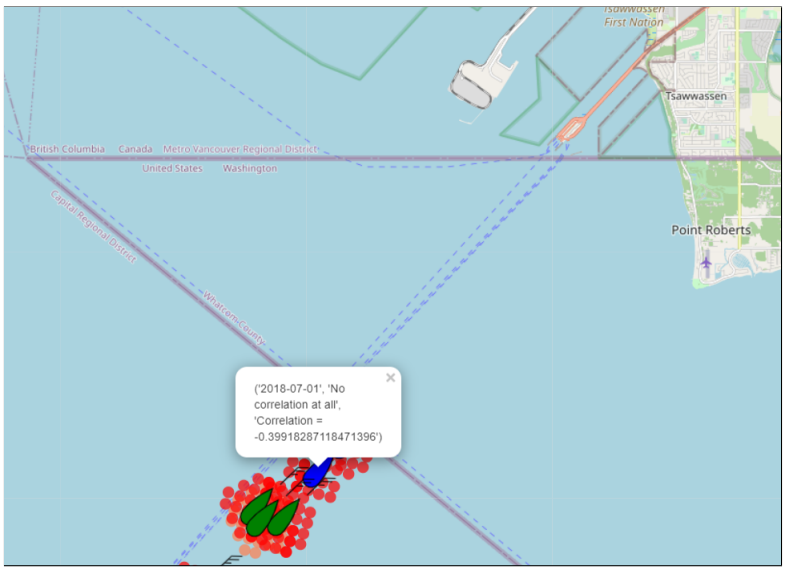


Figure 5.6: Zoomed-in Chloropeth Visualization near the port of Tsawwassen near US-Canada Border

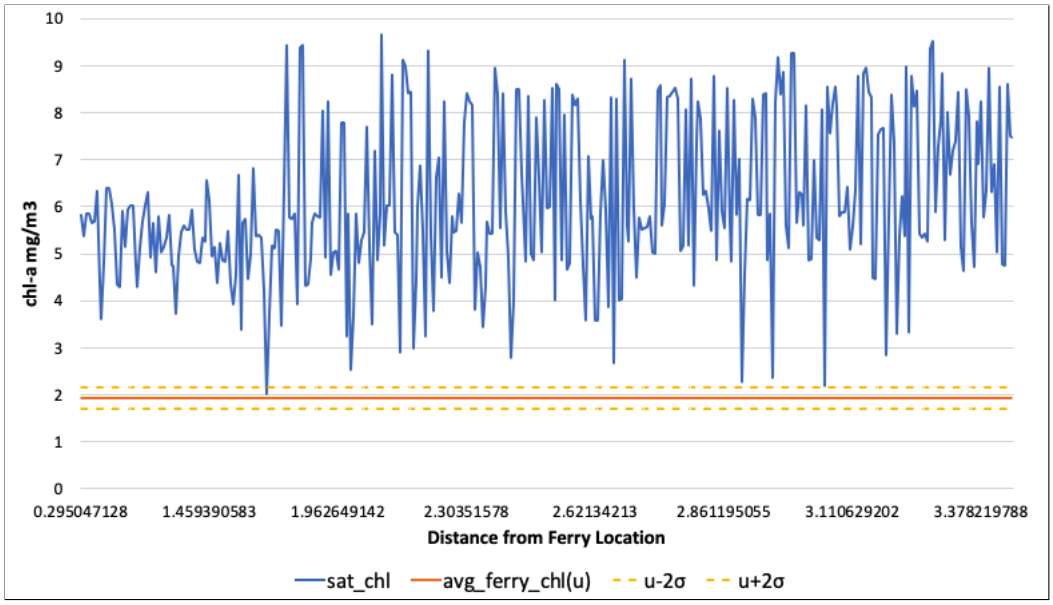


Figure 5.7: Plot comparing the satelite chl-a values wrt mean ferry chl-a values for July 01.

way higher than 2 mg/m^3 and deviates significantly as the distance from the reference ferry location increases. This behavior implied that either there is no correlation near Tsawwassen Terminal or the satellite overestimates the amount of chl-a concentration near Tsawwassen terminal.

3 Active Pass

The north opening of Active Pass witnesses a weak correlation of about 0.3. The zoomed-in map visualization of data points in Active Pass can be seen in Figure 5.8 where the cluster formed by satellite data points is golden yellow which signifies weak correlation.

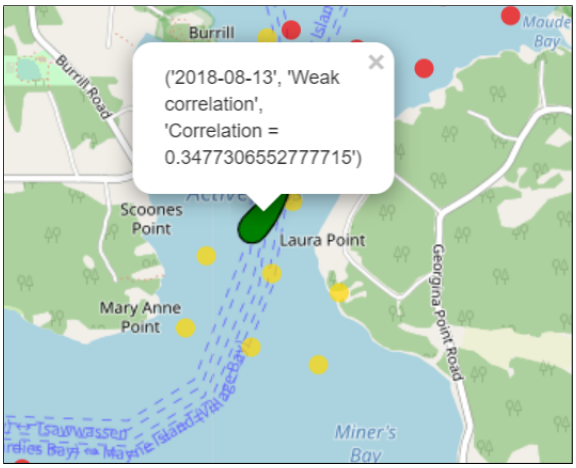


Figure 5.8: Zoomed-in Chloropeth Visualization near Active Pass

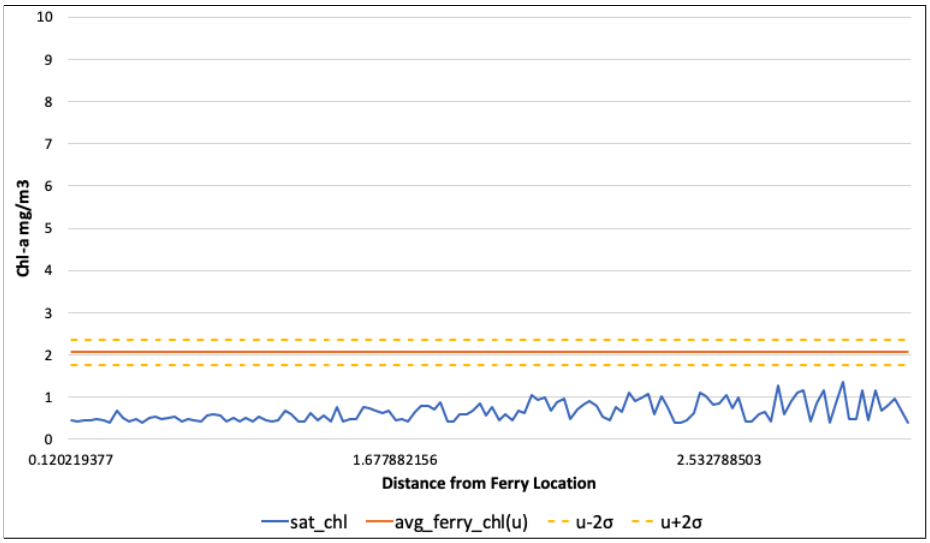


Figure 5.9: Plot comparing the satelite chl-a values wrt mean ferry chl-a values for August 13.

The graphical analysis in Figure 5.9 shows that the satellite chl-a concentration was much less than that of the average ferry chl-a concentration. It is dubious

to say whether the satellite sensors have underestimated the amount of chl-a over this region or the ferry chl-a has been overestimated or affected by high water currents around the opening.

4 Swanson Channel

Swanson Channel shows very confusing results. It is a very busy channel as there are quite many ferry routes that intersect on this channel. The zoomed-in version of choropleth visualization shows the presence of various instances when the location of BC ferry and area traversed by satellite surveillance coincided at the same time. The days that witness this are 6 August, 10 August, 18 July, and 14 August 2018 represented by the boat markers marked as 1,2, 3, and 4, respectively in the Figure 5.10. We can see different colors around each of these boat markers which are lying very close to each other. The yellow cluster around boat 1 (6 August) shows strong correlation while red cluster around the boat marked as 2 (10 August) shows no correlation at all. This is followed by boat marker 3 which is again showing weak correlation, indicated by golden yellow cluster around it and boat marker 4 which is lying very near shows no correlation with red cluster around it. The different strengths of Pearson Correlation on the same geographical region adds to the confusion.

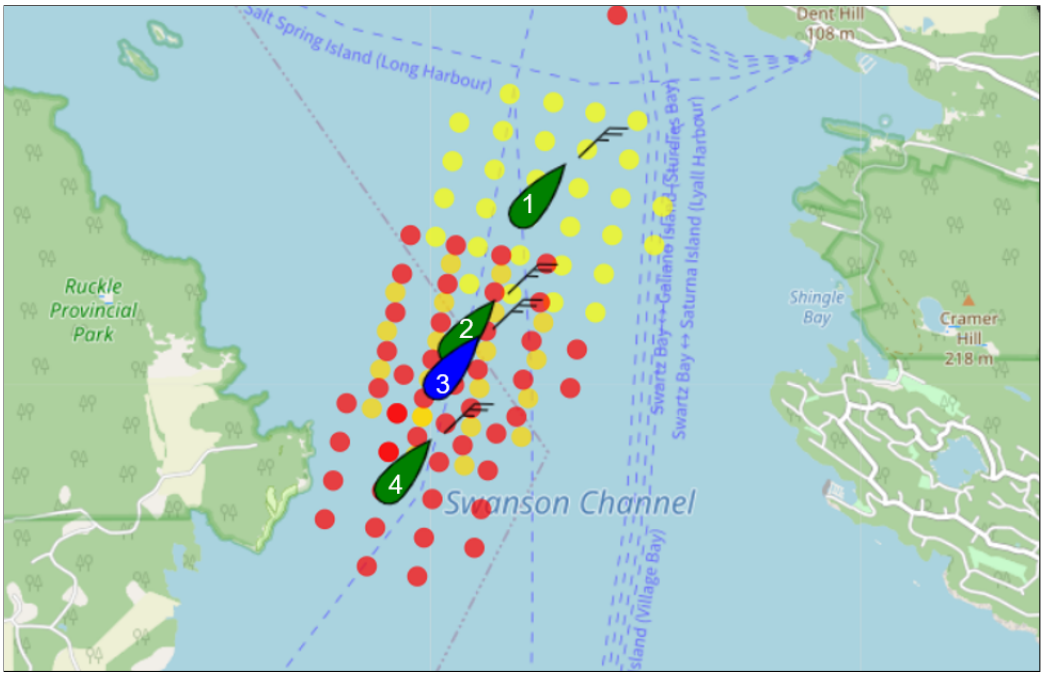


Figure 5.10: Zoomed-in Chloropeth Visualization near Swanson Channel

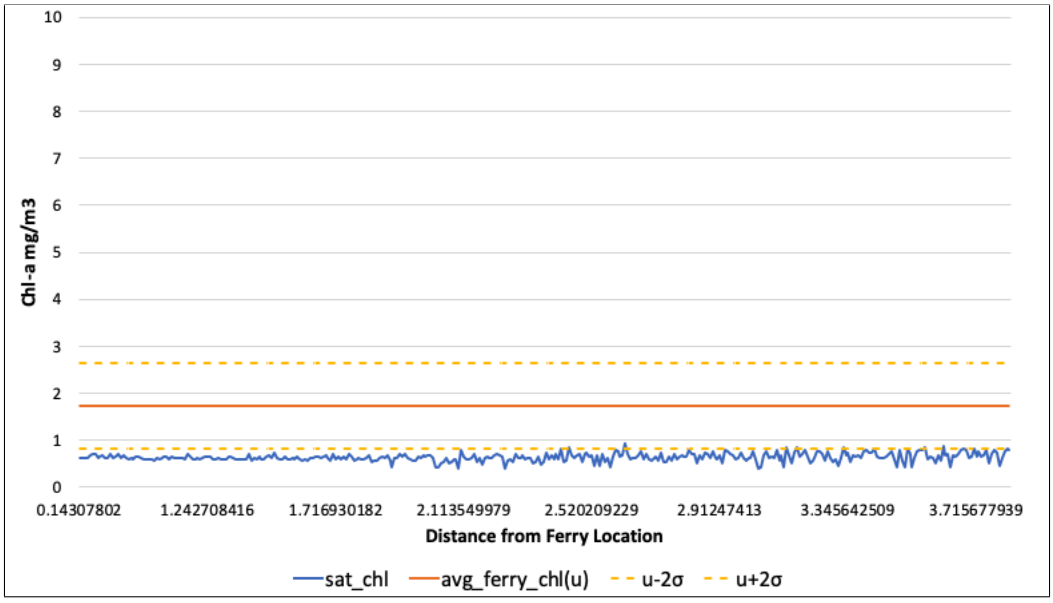


Figure 5.11: Plot comparing the satelite chl-a values wrt mean ferry chl-a values for August 6.

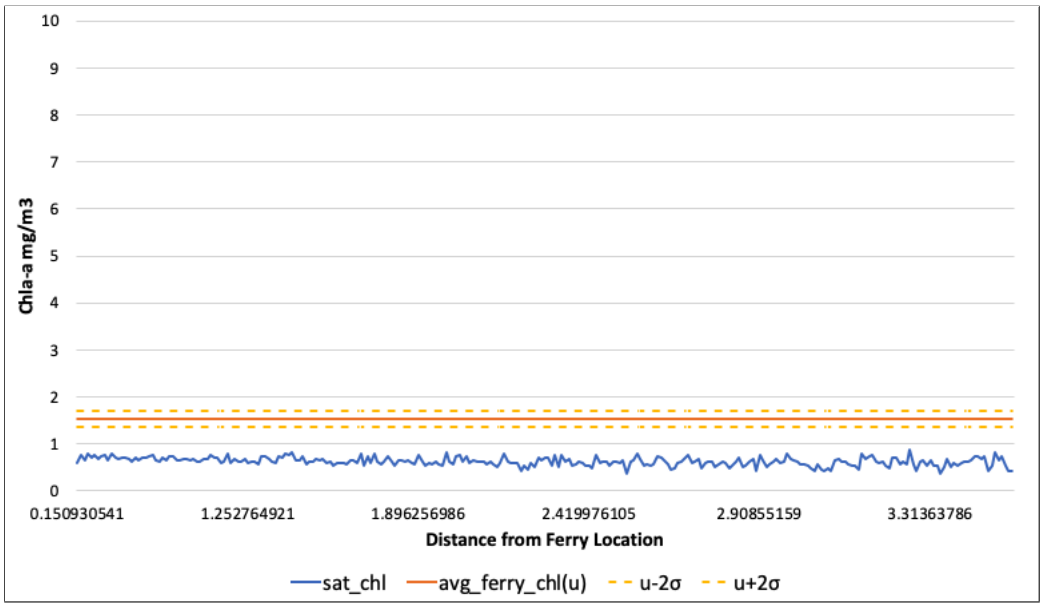


Figure 5.12: Plot comparing the satelite chl-a values wrt mean ferry chl-a values for August 10.

To get some insights behind the results displayed in the map visualization, a graph is plotted for the data collected on 6th August (Figure 5.11) as well as 10th August (Figure 5.12). In Figure 5.11, it is clear that, for data obtained on 6th August, the chl-a concentration recorded by ferry was having more standard

deviation and the chl-a concentration recorded by satellite is between second and third standard deviation from the mean ferry chl-a while it is not the case for data obtained on 10th August, in Figure 5.12, where the ferry chl-a concentration is not very deviated from its mean whereas the chl-a concentration recorded by satellite deviates a lot from the mean due to which there is no correlation.

5.3 Discussion

The graphs and numbers can only shed light on what is happening but cannot provide a holistic picture of the underlying reasons behind such uncertainties in the results. Therefore, there is a need to dive deeper into the understanding of ocean sciences and also question the efficiency of data collection methodology and equipment. There can be multiple factors acting together due to which we are facing uncertainties in the results. Ideally, the validation procedure was expected to provide a strong to a very strong correlation between the data collected by satellite and data collected by ferry, but most of the instances are showing no correlation. Some of them are showing very weak to weak correlation but just a few are resulting in strong correlation. Therefore, it is crucial to understand the uncertainties involved in the process of data collection and to digest the fact that the process of validation is not quite straightforward and linear, as multiple factors involved in a more pragmatic environment should be considered too.

The foremost uncertainty is the very nature of the coastal waters. The coastal waters are very productive and yielding towards the growth of many organisms, one of which is phytoplankton or algae. The southern Strait of Georgia has many channels where the river or freshwater is meeting the ocean. The rivers constitute silt and sediments which get mixed with oceans. The oceans have certain optical properties due to the presence of salt and minerals but when it gets contaminated by several other alien substances, its optical properties are affected and are influenced largely. Phytoplankton has a huge influence on the optical properties of the ocean as it reflects green light [1].

Gower and King [10] state the complications in dealing with coastal waters which represent a small fraction of oceans on the earth's surface. According to the authors, the coastal waters have very complex biological, physical, and optical properties due

to which there are a lot of variabilities throughout. The property of the variability of coastal waters makes them dynamic. The very fundamental conclusion drawn in the paper suggests that the presence of Colored Dissolved Organic Matter (CDOM) which emits blue light can confuse the algorithms present in sensors that perceive that the area has higher amounts of chlorophyll, hence overestimating chl-a concentration due to ambiguities in perception. Costa et al. [8], validates the ocean color images of Strait of Georgia obtained from satellite by comparing them to data obtained from BC ferries and concluded that in summer, the chl-a concentration estimated by satellite images show 12% of the variability. The authors used the methodology which is very similar to the one used in this project which considers temporal and spatial constraints. They observed less correlation and more variability, especially in the summer season which is similar to the observations derived in this project.

The fluorometers attached to BC ferries are responsible for estimating the chlorophyll concentration. The in-vivo fluorometers measure the chl-a from the intensity of light reflected from the chlorophyll/algae. The intensity of the reflectance is directly proportional to the intensity and angle of the incident light according to the famous Snell's law of reflection. Therefore, the amount of sunlight and the angle of sun rays also affects the performance of fluorometers. this implies that the fluorescence signals retrieved in the daylight and in the night are different. In the night, the signal/data obtained is less biased. The fluorometers used by BC ferries are Wetlabs' ECO triplet fluorometers. The accuracy and degree of reliability of data gathered by BC ferries also depend on how frequently the fluorometers are cleaned and re-calibrated. Some of the fluorometers from Wetlabs also come with the wipers attached to it which cleans the surface of the fluorometers but frequent calibration of the sensors is very important to maintain consistency in the data. Costa et al. [15] pointed out the issues of bio-fouling in fluorometers due to build-up surface growth and sediment present on measuring face of the fluorometers can affect the efficiency of sensors. It also explains how the signal attenuates over two weeks when cleaning and maintenance are not done efficiently.

The other main concern is the outliers produced by satellite data. Outliers in this scenario indicate very large chl-a concentration values which are not possible practically. The presence of outliers in certain areas means that either the satellite's OLCI sensors or the atmospheric correction algorithms are overestimating the chlorophyll concentration over certain geographical areas that need to be examined.

In this chapter, the results, analysis, and discussions over the analysis and uncertainties give a fair idea about the importance of validation of remote sensing data with the ground truth data. It has also shed some light on the loopholes that need to be fixed to further make the process of validation more efficient.

Chapter 6

Conclusion and Future Work

The main focus of the project was to validate the chlorophyll concentration estimated by remote sensing. The obtained data is validated by the chlorophyll concentration obtained directly from the oceans with the collaboration of BC ferries. The satellite Sentinel 3 which is equipped with OLCI sensors is monitoring ocean health. Ocean health is largely affected by chlorophyll/algae which is the basis of marine food webs.

The major challenge in the project was to align two entirely different datasets to make them compatible enough for effective comparison and validation. This is followed by thorough cleaning and preprocessing of both the datasets to filter out irrelevant information and look for outliers, if any. To validate the dataset, it is important to design an effective methodology that is sensitive to the temporal and spatial aspects of the data. Therefore, it is crucial to make sure that the time duration in which the data is collected by both the sources should align and the geographical area from where the two datasets are collected should also coincide. The design of the methodology revolves around coinciding with the two datasets considering time and space factors to produce reliable results. Whenever the time and space constraint is satisfied, an instance for validation is considered, signifying that the data points of BC ferry and satellite are overlapping at the same time and geographical location. The instance has its unique attributes which includes date, time, geographical location/coordinates. The data points involved in those instances are then statistically measured for validation. The statistical measure used for validation is the Pearson Correlation.

When it comes to playing with data, the process is not complete without visualizing it. The results and analysis are visualized in two different ways, namely choropleth

maps and box plots. Choropleth visualization shows all the data points or instances (in the two months of data), in which the time and space constraints are satisfied, on the map along with the correlation coefficients for each instance. Box plots visualize how the correlation coefficients of all the instances are distributed. There are also graphical representations of the statistical analysis performed on a few of the selected instances.

The overall results obtained indicate a very weak correlation between the satellite data and ferry data. There are some geographical areas such as near to Tsawwassen Terminal which shows no correlation while there are some areas such as Active Pass which show good correlation.

Various uncertainties play their role in the whole procedure, are discussed in section 5.3. The data collection techniques and equipment involved in the procedure also play an important role to ensure the consistency and reliability of the dataset. There are various environmental factors such as wind speed, traffic, rapid water currents, and geographical/natural structure of the passage which contributes towards the uncertainties. The coastal waters have always been an area of interest due to constant fluctuations in the chlorophyll concentrations.

The future work involves looking into the possible rectification of uncertainties such as re-calibration and cleaning of fluorometers deployed on BC ferries frequently and also changing the thresholds to avoid outliers in the satellite dataset. The current threshold is 50 mg/m^3 which can be reduced further to see if the results show further improvements. The presence of outliers also hints towards a significant work that needs to improve atmospheric correction algorithms to avoid overestimation.

Bibliography

[1] IOCCG (2000). Remote sensing of ocean colour in coastal, and other optically-complex, waters. sathyendranath, s. (ed.), reports of the international ocean-colour coordinating group, no. 3, ioccg, dartmouth, canada.

[2] Compute Canada. Cedar compute canada cc doc. URL https://docs.computecanada.ca/wiki/Cedar.

[3] Compute Canada. Job scheduling policies - cc doc. URL https://docs.computecanada.ca/wiki/Job_scheduling_policies/en.

[4] Ocean Networks Canada. Data search: Ocean networks canada - oceans 2.0. URL https://data.oceannetworks.ca/DataSearch.

[5] Tyson Carswell, M. Costa, Erika Young, Nicholas Komick, Jim Gower, and Ruston Sweeting. Evaluation of modis-aqua atmospheric correction and chlorophyll products of western north american coastal waters based on 13 years of data. *Remote Sensing*, 9, 10 2017.

[6] GeoPy Contributors. Geopy 2.0.0 documentation. URL https://geopy.readthedocs.io/en/stable/.

[7] Xarray Developers. Xarray: N-d labeled arrays and datasets in python. URL http://xarray.pydata.org/en/stable/.

[8] H Eckstrand, Trisalyn Nelson, M. Costa, and Nicholas Komick. Temporal and spatial analysis of chlorophyll-a measurements within the strait of georgia, british columbia: characterizing natural variability in relation to ocean colour images. 07 2020.

[9] Folium. Folium - folium 0.11.0 documentation. URL https://python-visualization.github.io/folium/.

[10] J. Gower and S. King. Use of satellite images of chlorophyll fluorescence to monitor the spring bloom in coastal waters. *International Journal of Remote Sensing*, 33(23):7469–7481, 2012.

[11] Jim Gower and Stephanie King. Satellite observations of seeding of the spring bloom in the strait of georgia, bc, canada. *International Journal of Remote Sensing*, 39(13):4390–4401, 2018.

[12] British Columbia Ferry Services Inc. Bc ferries schedules: Vancouver - victoria(tsawwassen-swartz bay). URL https://www.bcferries.com/schedules/mainland/tssw-current.php.

[13] Jennifer M. Jackson, Richard E. Thomson, Leslie N. Brown, Peter G. Willis, and Gary A. Borstad. Satellite chlorophyll off the british columbia coast, 19972010, Jul 2015.

[14] NetCDF. Netcdf 4 api documentation. URL https://unidata.github.io/netcdf4-python/netCDF4/index.html.

[15] A. Sastri, R. Fox, J. Krogh, and M. Costa. Real-time sea-surface measurements of coloured dissolved organic matter (cdom) in the strait of georgia, canada: Developing techniques to account for sensor fouling. https://meetings.pices.int/publications/presentations/PICES-2016/W7-Sastri.pdf.

[16] Sentinel. Olci instrument sentinel-3 olci technical guide sentinel online. URL https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-3-olci/olci-instrument.

[17] Sentinel. Orbit - sentinel-3 - mission - sentinel online. URL https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-3/satellite-description/orbit.

[18] Singularity. Installation - singularity container 2.6 documentation. URL https://sylabs.io/guides/2.6/user-guide/installation.html.

[19] National Oceanic US Department of Commerce and Atmospheric Administration. What is nutrient pollution?, Sep 2009.

[20] Wikipedia. Eutrophication. URL https://en.wikipedia.org/wiki/Eutrophication.