Pollution and Harmful Algae Bloom Supervision:

Monitoring Chesapeake Bay Environmental Health

By: Nicholas Sabella, 114756671

Background and Motivation

Basic Information

The purpose of this tutorial is to investigate trends between nutrient pollution and the overall health of the Chesapeake Bay, as well as conducting exploratory analysis about monitoring these chemicals in regards to their regional prevalance and/or changes.

Nutrient Pollution

To begin to understand nutrient pollution, it is important to establish an the answer to the question: what is a nutrient? Simply put, a nutrient is something that provides energy for oragnisms to grow and carry out their biological functions.

Nitrogen and phosphorus are two of the most common nutrients, and thus can be found inside most gardens and almost every bag of fertilizer. However, as some readers who are gardening enthusiasts may know, when there are too many nutrients in an environment, this is a problem. In gardening it is known as a Fertilizer Burn, and when a body of water has too many nutrients, it is known as eutrophication.

Fertilizer Burn **Eutrophication**

This image of an algae-covered pond is not uncommon to most Maryland residents, as well as others across the country. Unlike where too much fertilizer causes damage to garden plants, when there is a sufficiently high concentration of nutrients in a body of water, Phytoplankton grow at an extreme rate, presenting as an explosion of algae growth. Such a sight is a sign of an unhealthy local ecosystem, and is known as a Harmful Algae Bloom (HAB); however, this is an issue which can extend to larger bodies of water, such as the Chesapeake Bay, which can impact all life inhabiting its watershed.

Harmful Algal Blooms and Water Health

Inituitively, one may know that the water in a body of water covered in algae is an unclean and unhealthy body of water, and one would not want to drink from it, or consume any organisms that came out of it. But why is this the case? When algae blooms become sufficiently massive, they block sunlight from reaching subaquatic plants, which in turn, means no photosynthesis and subsequent death. Moreover, they cause massive swings in oxygen concentration since they produce oxygen during the day via photosynthesis, and consume oxygen at night via cellular respiration, oftentimes causing hypoxic conditions at night that are not suitable to marine life. These areas are affectionately referred to as "Dead Zones."

Pfiesteria is the most common form of Harmful Algal Bloom, which casues other organisms in the ecosystem to die or become sick, and thus, is overall a detriment to the Chesapeake Bay Ecosystem and Watershed.

Pfiesteria: The Source of Harmful Algal Blooms

Toxic Algae

In recent years, there have been instances of shellfish poisoning due to a a type of Phytoplankton: dinoflagellates. There are a few species which have been specifically identified in Maryland as a cause of shellfish poisoning. However, the true cause of the toxicity is that the Phytoplankton can consume toxic cyanobacteria, which in turn, cause the Phytoplankton, and eventually the shellfish, to bioaccumulate large amounts of toxins, which can poison humans if they consume these affected shellfish.

Toxic Dinoflagellates and Their Food Chain

- Prorocentrum minimum
- Prorocentrum micans
- Prorocentrum gracile

Oxygen Concentration

Since one of the primary concerns regarding the impact of Harmful Algal Blooms is the effect on other marine life, and thus the ecosystem at large, by creating hypoxic dead zones. Therefore, measuring the oxygen concentration can be a useful indicator in determining the health of the surrounding ecosystem, and ensuring that it is not a dead zone.

Why track it? Can anything be done?

- Harmful algae blooms can be incredibly destructive to the environment, hurting both marine-life and humans alike, and thus tracking their progression/regression and being able to predict their future course gives ciritical insight into whether or not mitigation steps are necessary

    - This project is more focused on identification of trends, rather than policy insight into how to ensure that the extent of HAB's does not become too impactful, nor does it attempt to comment on evironmental health overall, since something like fertilizer may incredibly helpful for a forrest and its ecosystem, but the runoff of fertilizers could hurt the marine ecosystem miles downstream.
- NOAA closely monitors HAB growth for the above reason, however, there is currently no monitoring system for the Chesapeake Bay. Attached is a link to their monitering services for other regions. Monitoring Service
    - This specific tracking service could be financially lucrative for an effective model, as well as giving researchers the opportunity for new samples to be added to datasets and update the models in real time
    - **NOAA has an $116 million grant to investigate Harmful Algal Bloom Research** link
        - Previously awarded projects link

Setup and Functions

- Install necessary packages
- Create functions to clean & process data

Installing Tools

Here the packages that will be needed throughout the project are installed.

```
!pip install geojson
!pip install geopandas
!pip install descartes
!pip install mapclassify
import matplotlib.pyplot as plt
import folium
import json
import seaborn as sns
from geojson import dump
import numpy as np
import pandas as pd
import geopandas as gpd
import matplotlib.cm as cm
import matplotlib.colors as mcolors
import re
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
```

```
from sklearn.metrics import r2_score
from google.colab import drive
import os

drive.mount('/content/drive')
os.chdir('/content/drive/MyDrive/320Final')
```

Defining Functions

- create_map(path)
  - Creates Folium map from .geojson file
  - Used in the Advanced Reading Section
- read_by_region_and_year(path, regionType, year)
  - Opens a dataset, downloaded from Chesapeake Bay Water Quality DataHub, and cleans it
  - Processes the pollutant data based on the year and the provided regionType (FIPS, aka State County, in this project but explored additionally in Advanced Reading Section)
  - Can also process non-pollutant data from DataHub, such as with oxygen, and there are dozens of parameters to choose from
- make_choropleths(path, pollutant, mult)
  - Creates 20 choropleth maps from 2000 to 2020, one per year, to show the change in concetration over time
  - Separates the concentration geographically into something that can be understood, for what would otherwise be too many regions to directly model
  - Uses read_by_region_and_year(path, regionType, year) as a helper function
- read_algae_by_year(path, year)
  - Opens a dataset, downloaded from Chesapeake Bay Living Resources DataHub, and cleans it
  - Process the mean total algae cells of samples taken in that given year
- read_algae(path)
  - Uses read_algae_by_year(path, year) as a helper function to create DataFrame of mean total algae cells for every year
- read_total_pollutant(path)
  - Clean and process data obtained from MD Open Data Portal, for the total amount of pollutant/nutrient/oxygen each year.

```
def create_map(path):
  m = folium.Map(
    location=[38.90, -77.03],
    tiles="cartodbpositron",
    zoom_start=8,
  )

  with open(path) as f:
      data = json.load(f)


  folium.GeoJson(data=data, name="geojson").add_to(m)
  return m

def read_by_region_and_year(path, regionType, year):
  df = pd.read_csv(path, dtype={regionType:str})
  df['SampleDate'] = pd.to_datetime(df['SampleDate'])
  df['year'] = df['SampleDate'].dt.year
  mask = (df['SampleDate'] > (year + '-01-01')) & (df['SampleDate'] <= (year + '-12-31'))
  df = df.loc[mask]
  df = df.groupby([regionType, 'MeasureValue']).mean().reset_index()
  df.drop(df.iloc[:, 2:11], inplace = True, axis = 1)
  aggregation_functions = {'MeasureValue': 'mean', 'year' : 'mean'}
  df = df.groupby(df[regionType]).aggregate(aggregation_functions)
  df = df.reset_index()
  return df

def make_choropleths(path, pollutant, mult):
  years = ['2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009',
         '2010', '2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019', '202
  df_total = {}
  for x in years:
    df_new = read_by_region_and_year(path, 'FIPS', x)
    df_total[x] = df_new

  ntemp = np.arange(0,6) * mult
  normalize_colors = mcolors.Normalize(vmin=ntemp.min(), vmax=ntemp.max())
  color_scheme = cm.jet

  fig1 = plt.figure()
  fig1.set_figheight(50)
  fig1.set_figwidth(50)
  i = 1
  for y in df_total.keys():
    df_new = df_total[y]
    df_new = df_new.sort_values(["FIPS"], ascending=True)
    tempthing = pd.concat([df_new, A_sorted], axis=1)
    tempthing2 = gpd.GeoDataFrame(tempthing, geometry=tempthing['geometry'])
    ax = fig1.add_subplot(5, 4, i)
    i+=1
    tempthing2.plot(column="MeasureValue", ax=ax, cmap=color_scheme)
```

```
    tempthing3 = cm.ScalarMappable(norm=normalize_colors, cmap=color_scheme)
    tempthing3.set_array(ntemp)
    ax.set_title(pollutant + ' Concentration: ' + str(y), fontsize=38)
    cbar = plt.colorbar(tempthing3)
    plt.tick_params(labelsize=30)
    cbar.ax.tick_params(labelsize=30)
    cbar.set_label('Concentration of '+pollutant+' (mg/L)', rotation=270, fontsize=30, labelp

  fig1.tight_layout()
  plt.show()

def read_algae_by_year(path, year):
  dfplank= pd.read_csv(path, low_memory=False)
  dfplank['SampleDate'] = pd.to_datetime(dfplank['SampleDate'])
  dfplank['year'] = dfplank['SampleDate'].dt.year
  mask = (dfplank['SampleDate'] > (year + '-01-01')) & (dfplank['SampleDate'] <= (year + '-12
  dfplank = dfplank.loc[mask]
  dfplank = dfplank.groupby(['LatinName', 'ReportingValue']).mean().reset_index()
  dfplank.drop(dfplank.iloc[:, 2:6], inplace = True, axis = 1)
  aggregation_functions = {'ReportingValue': 'mean'}
  dfplank = dfplank.groupby(dfplank['year']).aggregate(aggregation_functions)
  dfplank = dfplank.reset_index()
  return dfplank

def read_algae(path):
  years = ['2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009',
    '2010', '2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018']
  df_total = {}
  for x in years:
    df_new = read_algae_by_year(path, x)
    df_new = df_new.drop(columns=['year'])
    df_total[x] = df_new
  for y in df_total.keys():
    df_total[y] = (df_total[y].to_numpy())[0]
  df_total = pd.DataFrame.from_dict(df_total)
  df_total = df_total.transpose()
  df_total = df_total.rename(columns={0:'ReportingValue'})
  return df_total

def read_total_pollutant(path):
  df = pd.read_csv(path, low_memory=False)
  df = df.drop(['Major Basin','Land-River Segment', 'FIPS', 'County', 'Tributary Basin', 'Sou
  years = {}
  for x in df.columns:
    years[x] = re.findall("\d{4}", x)[0]
  df.rename(columns=years, inplace=True)
  df = df.transpose()
  df['mean'] = df.mean(axis=1)
  df = df.drop(df.iloc[:, 0:4086], axis=1)
  return df
```

Modeling the data:

Chemical Prevalence

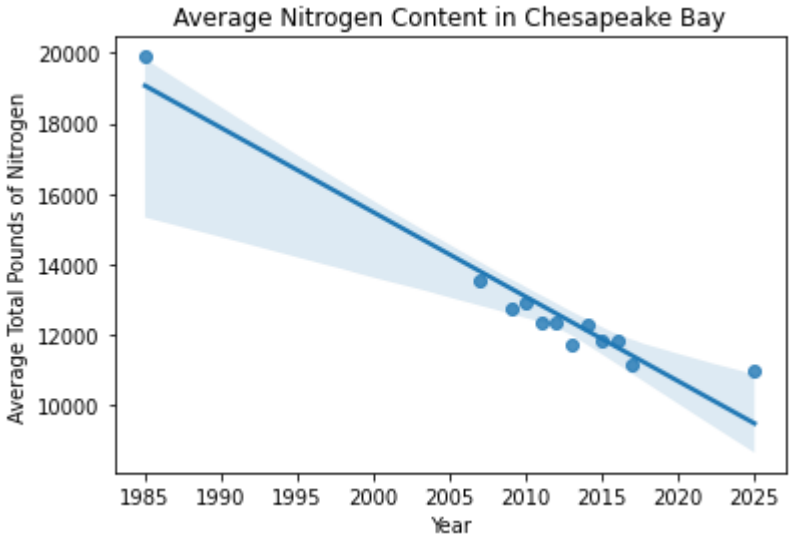
Here, linear regression models are applied to the total yearly nutrient weight. Note: due to the availability of these metrics, there is no recorded yearly data between 1986 and 2006, however significant Chesapeake Bay cleanup efforts transpired over this time period, so it is statistically significant to include the "historical" observation.

```
dfnitro= read_total_pollutant('./Chesapeake_Bay_Pollution_Loads_-_Nitrogen.csv')
dfphospho= read_total_pollutant('./Chesapeake_Bay_Pollution_Loads_-_Phosphorus.csv')
```

Nitrogen

```
fig, ax = plt.subplots(1,1)
```

```
ax = sns.regplot(x=dfnitro.index.to_numpy(dtype=int), y='mean', data=dfnitro)
ax.set_xlabel("Year");
ax.set_ylabel("Average Total Pounds of Nitrogen");
ax.set_title("Average Nitrogen Content in Chesapeake Bay");
```

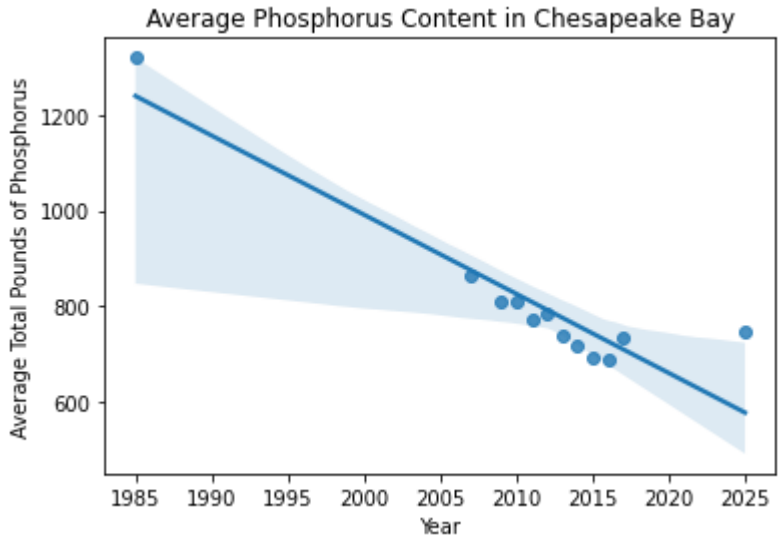


We observe that Nitrogen content is steadily decreasing at a nearly linear rate on a year-over-year basis. In fact, we observe that the predicted value from the dataset for 2025, is a conservative

estimate, assuming there will be less reduction over the next few years, and thus the value will fall in the upper range of predicted values.

Phosphorus

```
fig, ax = plt.subplots(1,1)
ax = sns.regplot(x=dfphospho.index.to_numpy(dtype=int), y='mean', data=dfphospho)
ax.set_xlabel("Year");
ax.set_ylabel("Average Total Pounds of Phosphorus");
ax.set_title("Average Phosphorus Content in Chesapeake Bay");
```



Similarly, we observe that Phosphorus content is steadily decreasing at a nearly linear rate on a year-over-year basis. In fact, we observe that the predicted value from the dataset for 2025, is a conservative estimate, assuming there will be less reduction over the next few years, and thus the value will fall in the upper range of predicted values.

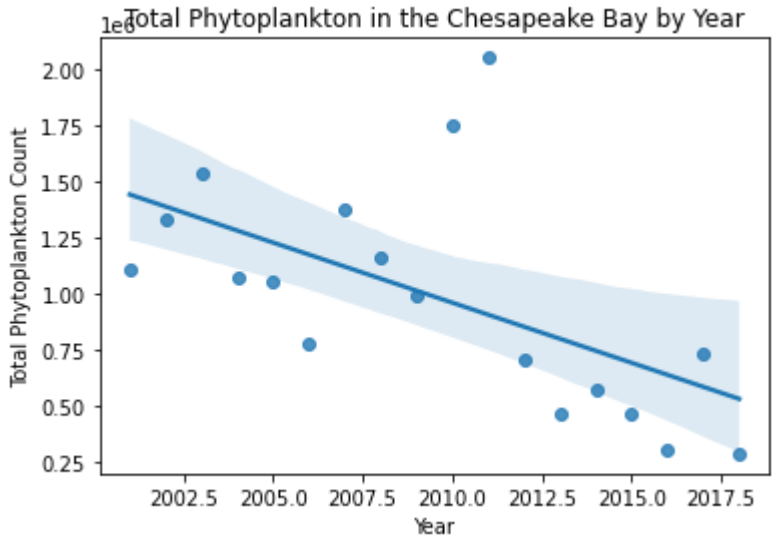
Algae Prevalence

Here, linear regression models are applied to total yearly phytoplankton count.

```
  dfalgae = read_algae('./PhytoplanktonCount.csv')
```

```
fig, ax = plt.subplots(1,1)
ax = sns.regplot(x=dfalgae.index.to_numpy(dtype=int), y="ReportingValue", data=dfalgae)
ax.set_xlabel("Year");
```

```
ax.set_ylabel("Total Phytoplankton Count");
ax.set_title("Total Phytoplankton in the Chesapeake Bay by Year");
```



In general, the phytoplankton count appears to be decreasing at a nearly linear rate, with two outlier years around 2010 and 2011. We expect some year to year variability in an observation such as phytoplankton, since a variety of factors beyond Eutrophication can affect phytoplankton growth, such as weather, temperature, and non-perrenial marine events/incidents.

Visualizing the data:

```
df66 = gpd.read_file('./Maryland_Physical_Boundaries_-_County_Boundaries_(Detailed).geojson')
df66 = df66.drop(labels=[3,23], axis=0)
A_sorted = df66.reset_index()
```

Displaying the Data

Nitrogen

```
print(make_choropleths('./WaterQualityWaterQualityFIPSNitrogen.csv', 'Nitrogen', 1))
```

Output is large, stored in Appendix for better clarity: See Appendix

Phosporous

```
make_choropleths('./WaterQualityWaterQualityFIPSPhosporous.csv', 'Phosphorus', .025)
```

Output is large, stored in Appendix for better clarity: See Appendix

Oxygen

```
make_choropleths('./WaterQualityWaterQualityFIPSOxygen.csv', 'Oxygen', 4)
```

Output is large, stored in Appendix for better clarity: See Appendix

Interpreting Visual Representation vs Modeling

When looking at the linear regressions, it is apparent that overall pollutant load is decreasing in nearly linear fashion over time for both Nitrogen and Phosphorus. Similarly, the overall Phytoplankton/Algae count has decreased in a linear fashion.

However, by examining the Choropleth graphs, we see this does not tell the full story. While the state as a whole has decreased in pollutant production, areas surrounding Baltimore City and on the inner regions of the Chesapeake Bay (particularly in the north near Baltimore for Nitrogen) have relatively high phosphorus and nitrogen content, and relatively low oxygen content.

Interestingly, we see that some of the regions of Maryland with the highest nitrogen, phosphorus, AND oxygen content are in Western and Northern Maryland, areas which would typically be considered much more rural. This makes sense that areas typically associated with agriculture would create nutrient run-off that would flow into the Chesapeake Bay, causing a buildup of concentration to occur inside the bay, rather than further up the watershed.

This resonates with but also challenges an intuitive understanding of water health, in that more urbanized areas along major bodies of water, typically have dirty water with ecological problems, whereas country rivers/lakes/streams are healthy and clean. The urbanized areas feel the downstream effects (literally) of agriculture, and thus have hypoxic, unhealthy water that is full of pollutants and nutrients. However, while water in more rural areas may appear clean and have thriving animal life due to the high oxygen content, there may be high concentrations of nutrients (which are toxic to humans in that form, as one could probably guess since humans can't eat plant fertilizer) hiding within the water as well, and in fact, the concentration is likely higher than it is in an urban area.

For greater exploration that will make this even more evident, specific dead zones are most common where major rivers pour into the Chesapeake Bay. Unfortunately, the map is too large to

render, so explore at your leisure.

Dead Zone Exploration

For the Advanced Reader:

Further Reading and Resources

This project provides a basic framework for demonstrating the power of data science to represent incomprehensibly large data sets into something that can be understood and analyzed by human minds. This project can be expanded upon in many different ways, to create even greater accuracy of insight, by analyzing more parameters and doing so in different geographic sub-units, or even using these in combination to create complex predictive models which can then be applied to the maps to show the distribution of a new predictive indicator.

While this project does so directly from datasets and a .geojson file of the FIPS regions, there are many pre-exisiting services which allow for easy drag & drop implementation for datasets to be visually displayed over a variety of files with borders and data with coordinate locations.

Map Editing Tools (can be found online):

- Easier Implementation
- Powerful Capabilities
- Arcgis has large database of maps in .json format
- Can work with layer control
- Can create a slider to display map changes over time

Additional Regional Breakdowns:

The following are different ecological groupings that could provide additional insight into sources of pollutants/nutrients, and where they concentrate/become problematic.

Available Groupings

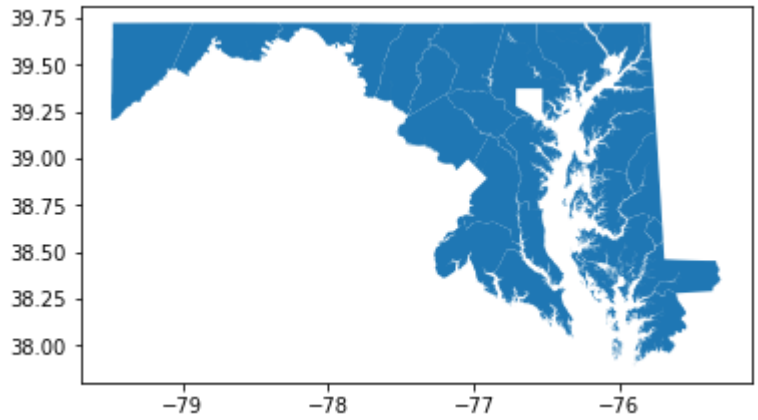
- FIPS (used in this project)
- HUC11 (federal watershed)
- HUC8 (watershed)
- HUC12 (subwatershed)

However, this is currently limited due to lack of available information sorted by grouping.

This can be worked around, however, by using the longitutde and latitude of each sample, and determining which Polygon region the coordinates fall inside. This is most effectively implemented using a .kml file, instead of a .geojson, but both can be accomplished.

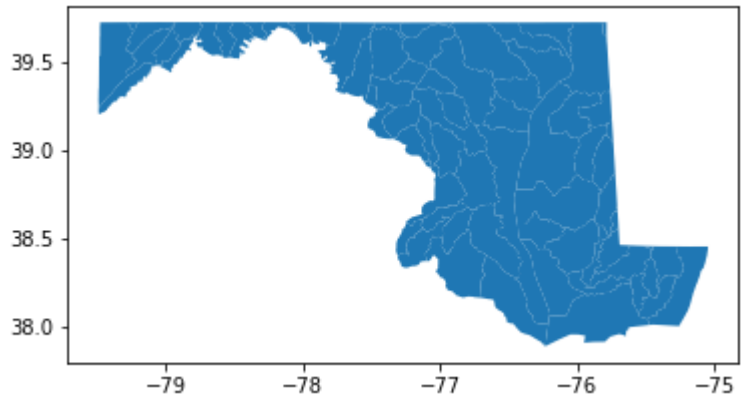
```
df66.plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f04862e49d0>
```



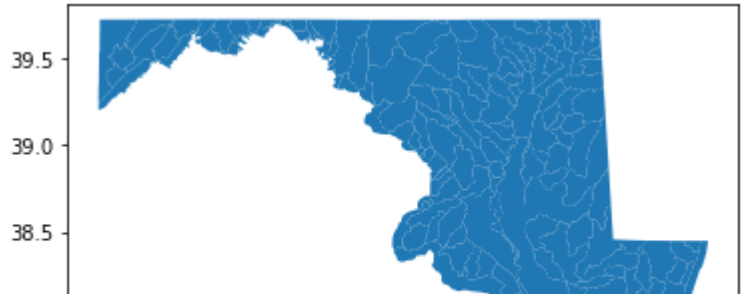

```
temp11digit = gpd.read_file('./Maryland_Watersheds_-_Federal_Watersheds_(HUC_11).geojson')
temp11digit.plot()
#create_map('./Maryland_Watersheds_-_Federal_Watersheds_(HUC_11).geojson')
#interactable map, doesn't work well for printing to pdf
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0495ea2610>
```



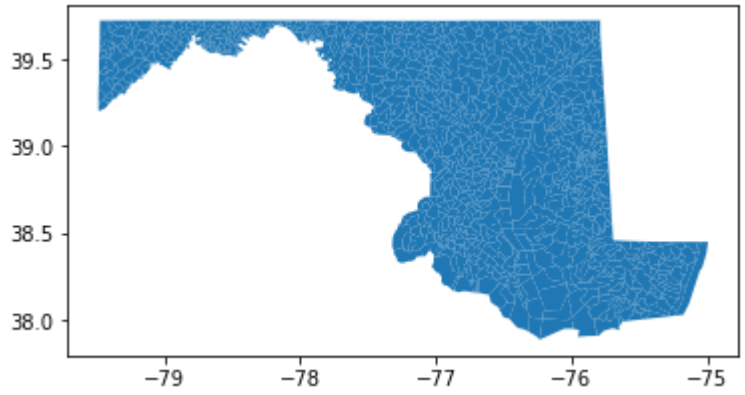

```
temp8digit = gpd.read_file('./Maryland_Watersheds_-_8_Digit_Watersheds.geojson')
temp8digit.plot()
#create_map('./Maryland_Watersheds_-_8_Digit_Watersheds.geojson')
#interactable map, doesn't work well for printing to pdf
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0495dd8750>
```




```
temp12digit = gpd.read_file('./Maryland_Watersheds_-_12_Digit_Watersheds.geojson')
temp12digit.plot()
#create_map('./Maryland_Watersheds_-_12_Digit_Watersheds.geojson')
#interactable map, doesn't work well for printing to pdf
```

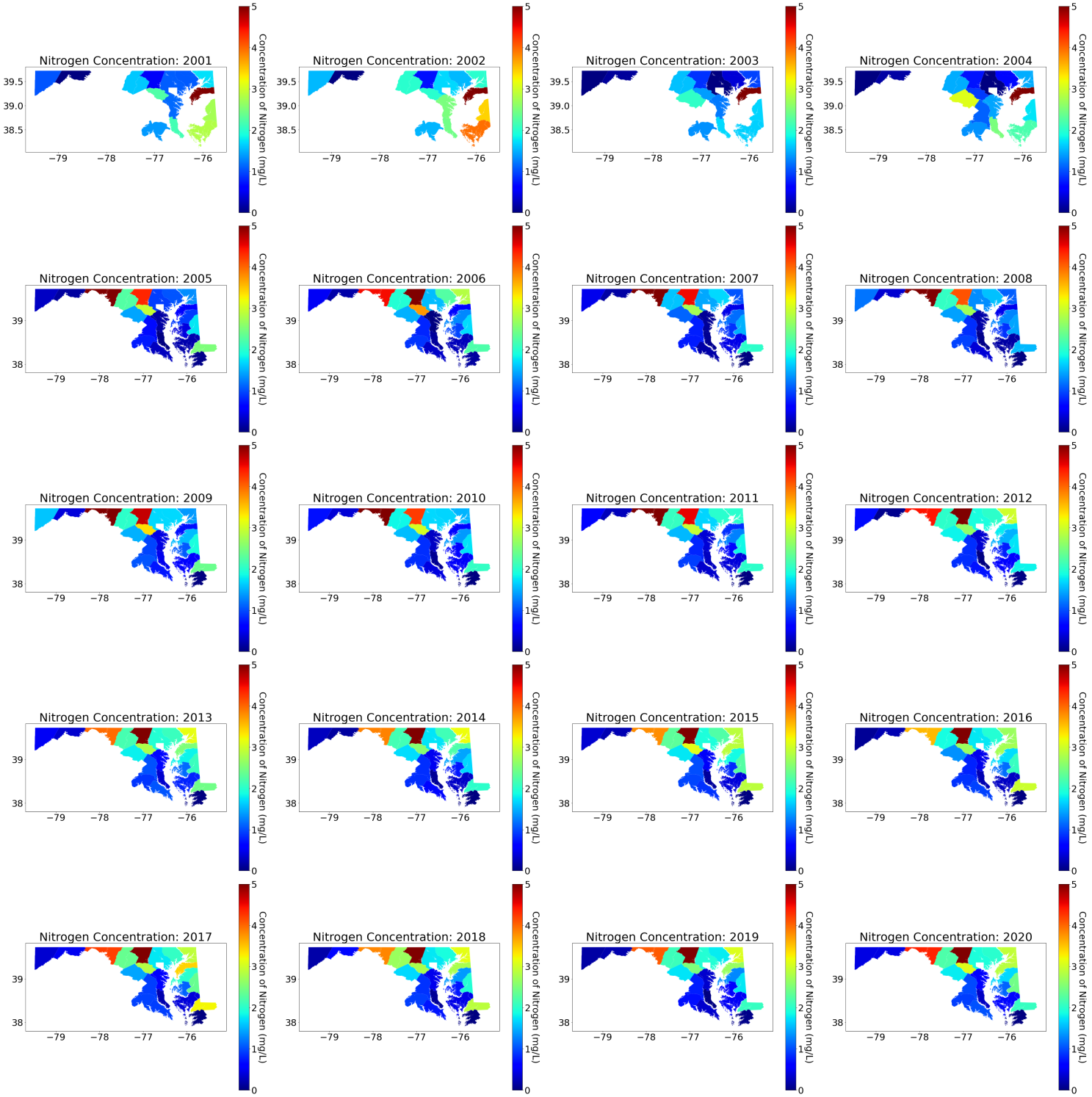
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0495bd3290>
```

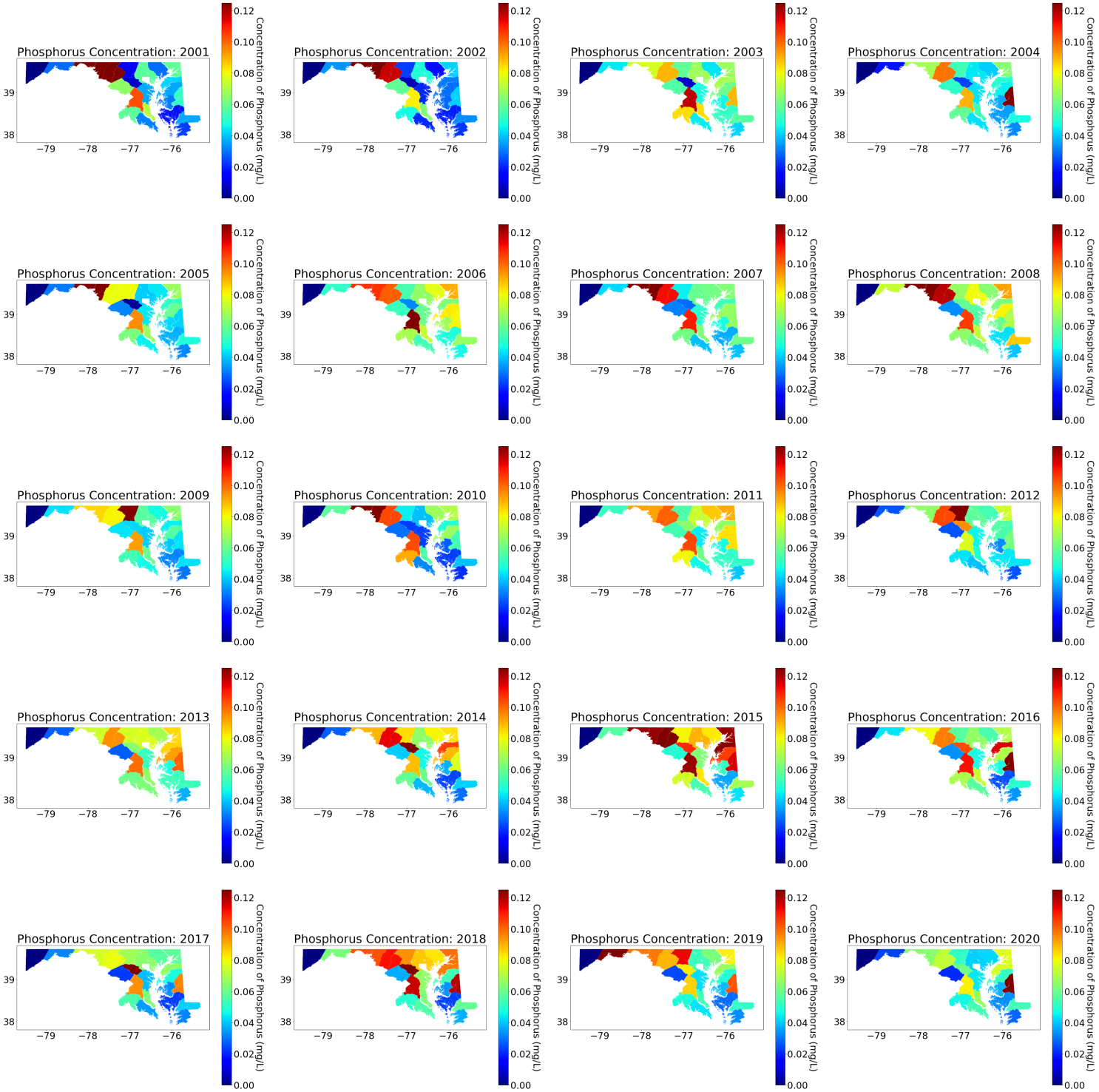


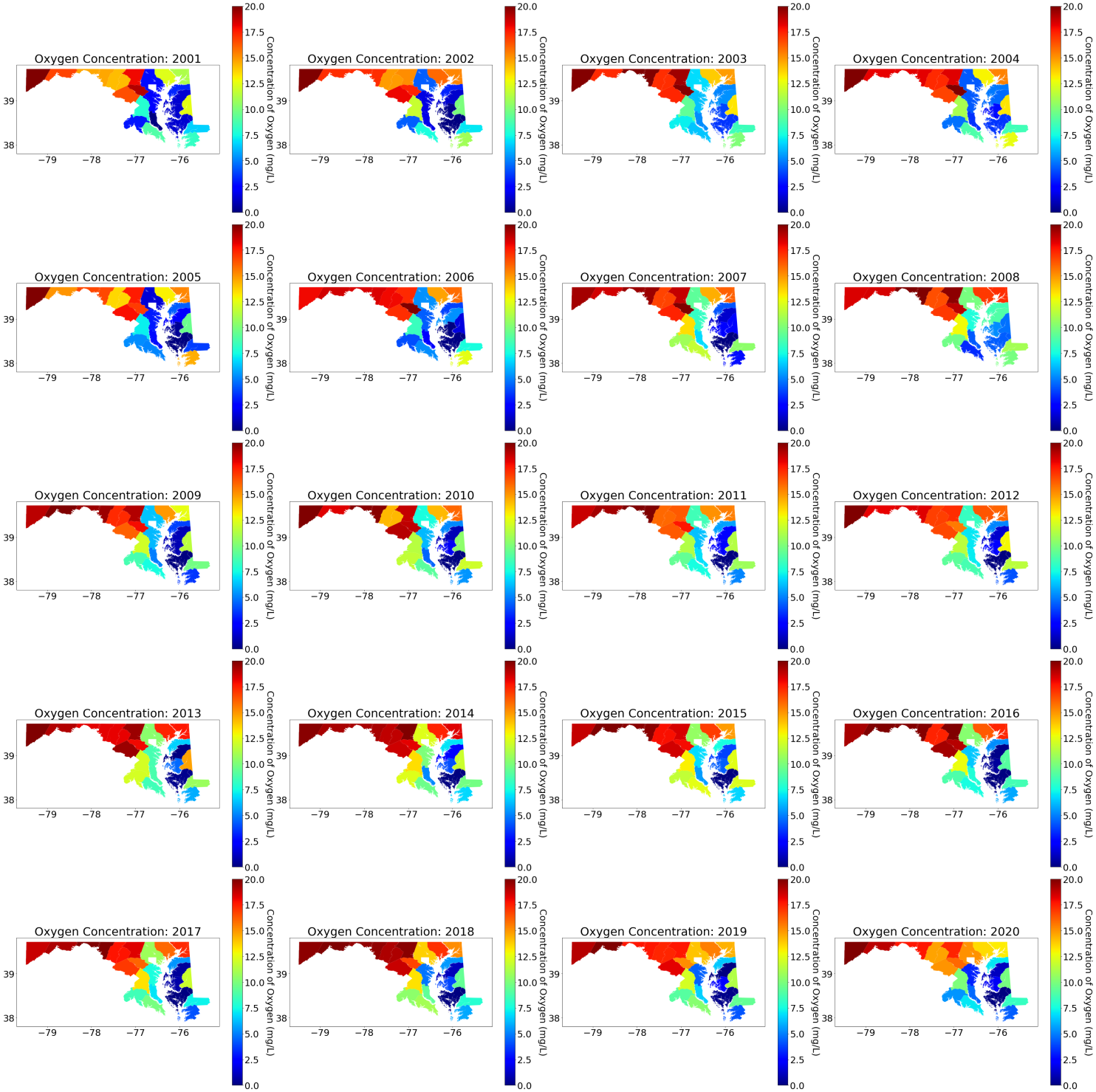
Conclusion

Throughout the course of this tutorial, we were able to observe a general trend of nearly linear decline of Nitrogen and Phosphorus in the Chesapeake Bay. We were also able to establish a general trend of decline in Phytoplankton in the Chesapeake Bay, which was unsurprising given the background information surrounding phytoplankton growth and the ambient nutrient concentration. However, by using data science to group the data and continue with our analysis, we were able to rectify common-sense understandings of the flow of nutrients that were harming the bay, as well as expand upon that understanding in ways previously not understood. Data science has the power to expand one's understanding and intuition, as well as create a useful product that can help others to understand a seemingly abstract and complex problem. So while it appears the Chesapeake Bay is on the right track health-wise, there is a stil a ways to go, and data science holds the key to actively monitoring whatever changes the future may hold. Maybe even, the reader of this tutorial will go on to create a monitoring service which receives part of that NOAA $116 million grant, helping the Chesapeake Bay, and their own bank account.

Appendix


Nitrogen Concentration: 2001
Nitrogen Concentration: 2002
Nitrogen Concentration: 2003
Nitrogen Concentration: 2004
Nitrogen Concentration: 2005
Nitrogen Concentration: 2006
Nitrogen Concentration: 2007
Nitrogen Concentration: 2008
Nitrogen Concentration: 2009
Nitrogen Concentration: 2010
Nitrogen Concentration: 2011
Nitrogen Concentration: 2012
Nitrogen Concentration: 2013
Nitrogen Concentration: 2014
Nitrogen Concentration: 2015
Nitrogen Concentration: 2016
Nitrogen Concentration: 2017
Nitrogen Concentration: 2018
Nitrogen Concentration: 2019
Nitrogen Concentration: 2020
Concentration of Nitrogen (mg/L)
0
1
2
3
4
5
−79
−78
−77
−76
38
38.5
39
39.0
39.5


Phosphorus Concentration: 2001
Phosphorus Concentration: 2002
Phosphorus Concentration: 2003
Phosphorus Concentration: 2004
Phosphorus Concentration: 2005
Phosphorus Concentration: 2006
Phosphorus Concentration: 2007
Phosphorus Concentration: 2008
Phosphorus Concentration: 2009
Phosphorus Concentration: 2010
Phosphorus Concentration: 2011
Phosphorus Concentration: 2012
Phosphorus Concentration: 2013
Phosphorus Concentration: 2014
Phosphorus Concentration: 2015
Phosphorus Concentration: 2016
Phosphorus Concentration: 2017
Phosphorus Concentration: 2018
Phosphorus Concentration: 2019
Phosphorus Concentration: 2020
39
38
−79
−78
−77
−76
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Concentration of Phosphorus (mg/L)


Oxygen Concentration: 2001
Oxygen Concentration: 2002
Oxygen Concentration: 2003
Oxygen Concentration: 2004
Oxygen Concentration: 2005
Oxygen Concentration: 2006
Oxygen Concentration: 2007
Oxygen Concentration: 2008
Oxygen Concentration: 2009
Oxygen Concentration: 2010
Oxygen Concentration: 2011
Oxygen Concentration: 2012
Oxygen Concentration: 2013
Oxygen Concentration: 2014
Oxygen Concentration: 2015
Oxygen Concentration: 2016
Oxygen Concentration: 2017
Oxygen Concentration: 2018
Oxygen Concentration: 2019
Oxygen Concentration: 2020
39
38
−79
−78
−77
−76
Concentration of Oxygen (mg/L)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0