MC02 TECHNICAL REPORT

Course: CPEN106 - Big Data Analytics 2

Laboratory Title: Data Mining and Data Visualization for Water Quality Prediction in Taal Lake

Instructor: JOVEN R. RAMOS

Date Performed:
Date Submitted:

Group Members: Dizon, Rockwell E
Lineses, Dann B.
Mojica, Warreon Dave A.
Pangilinan, Patrick James A.
Papa, Mark Jamir C.
Vidad, Ranjo B.

Objectives

- Apply data mining techniques to extract useful insights from environmental datasets.
- Develop predictive models for water quality using machine learning.
- Compare ensemble learning techniques such as CNN, LSTM, and Hybrid CNN LSTM.
- Visualize trends and patterns in water quality parameters.
- Interpret the impact of environmental and volcanic activity on water quality.
- Predict Water Quality Index (WQI) and Water Pollutant Levels with actionable insights for environmental monitoring and intervention.

Materials & Tools

2.1. Software Requirements

- Python (Jupyter Notebook/Google Colab)
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, TensorFlow/Keras (for deep learning), Streamlit (for dashboarding)

2.2. Hardware Requirements

- Computer with at least 8GB RAM and quad-core processor

2.3. Dataset

- Taal Lake Water Quality Reports from 2013-2023 given by the Bureau of Fisheries and Aquatic Resources (BFAR)
    - Weather Factors:
        - Weather Condition
        - Wind Direction
        - Air Temperature

    - Water Quality Parameters:
        - Water Temperature (Surface, Middle, Bottom)
        - pH
        - Dissolved Oxygen
        - Nitrogen (Nitrite or Nitrate)
        - Ammonia
        - Phosphate
    - Volcano Activity:
        - Sulfide
        - Carbon Dioxide

Procedure

Phase 1: Data Preparation & Preprocessing

Step 1: Import necessary libraries

```
from google.colab import files
import io
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.impute import KNNImputer
from sklearn.preprocessing import
MinMaxScaler
```

Step 2: Load dataset into a Pandas Dataframe

```
uploaded = files.upload()
df =
pd.read_csv(io.BytesIO(uploaded['Spreadmeat_W
QxVAxMF.csv']))
```

Step 3: Clean missing values (dropna() or fillna())

```
# Calculate the number of null values in each
row
null_counts_per_row = df.isnull().sum(axis=1)

# Remove rows with no datas from the df
df_new = df[null_counts_per_row < 10]

# Impute the remaining null values
imputer = KNNImputer(n_neighbors=5)
float_cols =
df_new.select_dtypes(include=['float']).colum
ns
```

```
df_new[float_cols] =
imputer.fit_transform(df_new[float_cols])
```

Step 4: Remove duplicates with pandas' "drop_duplicates()" function

```
# Remove duplicates
df.drop_duplicates(inplace=True)
```

Step 5: Convert date-time formats using pandas' "to_datetime()" function

```
# Change the date format into
'YEAR-MONTH-DAY'
df_new['Month'] = df_new['Month'].astype(str)
df_new['Month'] =
df_new['Month'].str.capitalize()
df_new['Month'] =
pd.to_datetime(df_new['Month'],
format='%B').dt.month

# Create Date column
df_new['Date'] =
pd.to_datetime(df_new[['Year',
'Month']].assign(DAY=1))

# Rearrange the dataframe
cols = list(df_new.columns)
col_to_move = cols.pop()
cols.insert(1, col_to_move)
df_new = df_new[cols]

# Remove Month and Year columns
df_new = df_new.drop(columns=['Month',
'Year'])
```

Step 6: Normalize values with sklearn's MinMaxScaler

```
# Create a MinMaxScaler object
scaler = MinMaxScaler()

# Select the columns you want to normalize
numeric_cols =
df_new.select_dtypes(include=['number']).colu
mns

# Fit the scaler to the selected columns and
transform them
df_new[numeric_cols] =
scaler.fit_transform(df_new[numeric_cols])
```

Phase 2: Exploratory Data Analysis (EDA)

Step 7: Generate summary statistics with the “describe()” function

```
df_new.describe()
```

Step 8: Compute correlation matrix with “corr()” function

```
# Remove non-numeric temporarily
numeric_df                                              =
df.select_dtypes(include=['number'])

# Generate Correlation matrix
correlation_matrix = numeric_df.corr()

print(correlation_matrix)
```

Step 9: Visualize trends among the parameters:

A. *Line Charts for Time-Series Data*

```
# Set 'Date' as the index if it's not
already

if     not     isinstance(df_new.index,
pd.DatetimeIndex):

    df_new = df_new.set_index('Date')

# Unique color list (extend if you have
more variables)

colors   =   ['blue',   'green',   'red',
'purple', 'orange', 'black', 'magenta',
'brown', 'violet', 'gray']

# Get numeric columns

numeric_columns                         =
df_new.select_dtypes(include=np.number).
columns

# Safety check: extend colors if needed

if len(numeric_columns) > len(colors):

    from itertools import cycle

    color_cycle = cycle(colors)

    assigned_colors = [next(color_cycle)
for _ in range(len(numeric_columns))]

else:

                   assigned_colors      =
colors[:len(numeric_columns)]
```

```
# Plot each column with a unique color

for       column,       color       in
zip(numeric_columns, assigned_colors):

    plt.Figure(figsize=(10, 6))

                  plt.plot(df_new.index,
df_new[column], color=color)

    plt.xlabel("Date")

    plt.ylabel(column)

       plt.title(f"Time  Series  Plot  of
{column}")

    plt.grid(True)

    # Format the x-axis for yearly ticks

plt.gca().xaxis.set_major_locator(mdates
.YearLocator())

plt.gca().xaxis.set_major_formatter(mdat
es.DateFormatter('%Y'))

    plt.xticks(rotation=45)

    plt.tight_layout()

    plt.show()
```

B. Heatmaps for Feature Correlation

```
# Create a heatmap for visualization
plt.Figure(figsize=(12,  10))   # Adjust
Figure size if needed
sns.heatmap(correlation_matrix,
annot=True,  cmap='coolwarm',  fmt=".2f",
linewidths=.5)
plt.title('Correlation Matrix Heatmap')
plt.show()
```

C. Scatter Plots for Parameter Relationships

a. Effects of Environmental Factors in the Water Quality

```
# Surface Temperature
plt.Figure(figsize=(10, 6))
plt.scatter(df_new['Air
Temperature   '],   df_new['Surface
Temp'],  label='Air  Temperature  vs
Surface Temperature', alpha=0.7)
plt.xlabel("Air Temperature")
plt.ylabel("Surface Temperature")
plt.title("Effect of Environmental
Factors on Water Quality")
plt.legend()
plt.grid(True)
plt.show()

# Dissolved Oxygen
plt.Figure(figsize=(10, 6))
plt.scatter(df_new['Air
Temperature '], df_new['Dissolved
Oxygen'], label='Air Temperature
vs Dissolved Oxygen', alpha=0.7)
plt.xlabel("Air Temperature")
plt.ylabel("Dissolved Oxygen")
plt.title("Effect of Environmental
Factors on Water Quality")
plt.legend()
plt.grid(True)
plt.show()



plt.Figure(figsize=(12, 8))

plt.subplot(2, 2, 1)
plt.scatter(df_new['Wind
Direction'], df_new['Ammonia'],
label='Wind Direction vs Ammonia',
alpha=0.7)
plt.xlabel("Wind Direction")
plt.ylabel("Ammonia")
plt.title("Wind Direction vs
Ammonia")
plt.legend()
plt.grid(True)

plt.subplot(2, 2, 2)
plt.scatter(df_new['Wind
Direction'], df_new['Nitrate'],
label='Wind Direction vs Nitrate',
alpha=0.7)
plt.xlabel("Wind Direction")
plt.ylabel("Nitrate")
plt.title("Wind Direction vs
Nitrate")
plt.legend()
plt.grid(True)
plt.subplot(2, 2, 3)
plt.scatter(df_new['Wind
Direction'], df_new['Phosphate'],
label='Wind Direction vs
Phosphate', alpha=0.7)
plt.xlabel("Wind Direction")
plt.ylabel("Phosphate")
plt.title("Wind Direction vs
Phosphate")
plt.legend()
plt.grid(True)

plt.subplot(2, 2, 4)
plt.scatter(df_new['Wind
Direction'], df_new['Nitrite'],
label='Wind Direction vs Nitrite',
alpha=0.7)
plt.xlabel("Wind Direction")
plt.ylabel("Nitrite")
plt.title("Wind Direction vs
Nitrite")
plt.legend()
plt.grid(True)

plt.tight_layout()
plt.show()



weather_params = ['pH', 'Dissolved
Oxygen', 'Nitrate', 'Nitrite',
'Ammonia', 'Phosphate']
for param in weather_params:
    plt.Figure(figsize=(10, 6))
          sns.scatterplot(x='Weather
Condition', y=param, data=df_new)
                plt.xlabel('Weather
Condition')
    plt.ylabel(param)
      plt.title(f'Weather Condition
vs. {param}')
                  plt.legend([],[],
frameon=False)     # remove the
legend
     plt.xticks(rotation=90)  # <--
rotate labels vertically
    plt.tight_layout()
    plt.show()
```

b. Effects of Volcanic Activity to the Water Quality

```
## CO2 may decrease pH = more acidic = bad
plt.Figure(figsize=(12, 8))
plt.scatter(df_new['Carbon Dioxide'], df_new['pH'], label='Carbon Dioxide vs pH', alpha=0.7)
plt.xlabel("Carbon Dioxide")
plt.ylabel("pH")
plt.title("Carbon Dioxide vs pH")
plt.legend()
plt.grid(True)
plt.show()

## CO2 may decrease Dissolved Oxygen = bad
plt.Figure(figsize=(12, 8))
plt.scatter(df_new['Carbon Dioxide'], df_new['Dissolved Oxygen'], label='Carbon Dioxide vs Dissolved Oxygen', alpha=0.7)
plt.xlabel("Carbon Dioxide")
plt.ylabel("Dissolved Oxygen")
plt.title("Carbon Dioxide vs Dissolved Oxygen")
plt.legend()
plt.grid(True)
plt.show()

## Sulfide may decrease DO = bad
plt.Figure(figsize=(12, 8))
plt.scatter(df_new['Sulfide'], df_new['Dissolved Oxygen'], label='Sulfide vs Dissolved Oxygen', alpha=0.7)
plt.xlabel("Sulfide")
plt.ylabel("Dissolved Oxygen")
plt.title("Sulfide vs Dissolved Oxygen")
plt.legend()
plt.grid(True)
plt.show()

## Sulfide may decrease pH = bad
plt.Figure(figsize=(12, 8))
plt.scatter(df_new['Sulfide'], df_new['pH'], label='Sulfide vs pH', alpha=0.7)
plt.xlabel("Sulfide")
plt.ylabel("pH")
plt.title("Sulfide vs pH")
plt.legend()
plt.grid(True)
plt.show()
```

Phase 3: Predictive Modeling using Ensemble Learning

Step 10: Create CNN, LSTM, CNN-LSTM Models for Predicting pH/DO

View Appendix A for Code

Step 11: Evaluate the Models with MAE & RMSE

View Appendix A for Code

Step 12: Perform Time-Based Prediction (Monthly & Yearly)

View Appendix A for Code

Step 13: Calculate WQI for Predicted Values

View Appendix A for Code

Phase 4: Data Visualization

Step 14: Compare Actual vs Predicted values using Seaborn

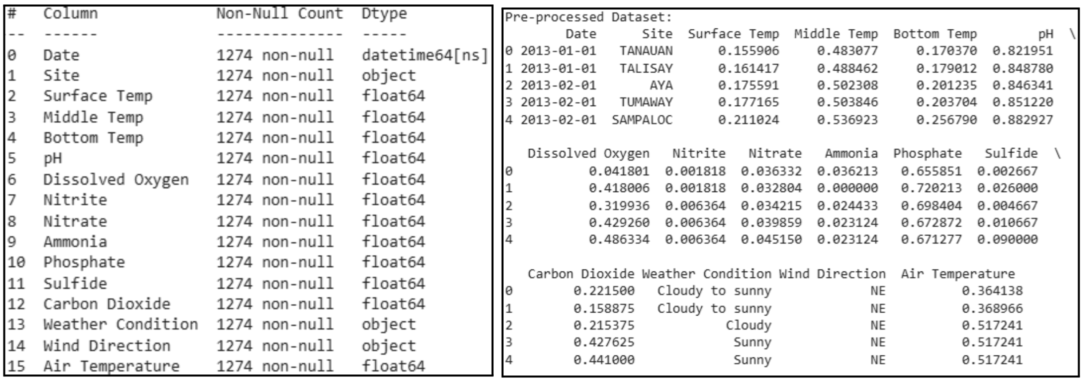
View Appendix A for Code

Step 15: Create a Streamlit Dashboard to Showcase Results

View Appendix B for Website

Data and Results

The dataset for this activity was provided by the Bureau of Fisheries and Aquatic Resources (BFAR) and covers the years 2013 to 2023. It includes monthly measurements of water quality, including dissolved oxygen, pH, nitrate, nitrite, ammonia, phosphate, and water temperature (at the surface, middle, and bottom). Additionally, it includes environmental data such as air temperature, weather condition, wind direction, as well as volcanic activity indicators (sulfide and carbon dioxide levels) to fully understand the factors that affect the water quality in Taal Lake. As seen in Figure 1, the dataset had a total of 1274 samples (after preprocessing) and a total of 16 parameters.

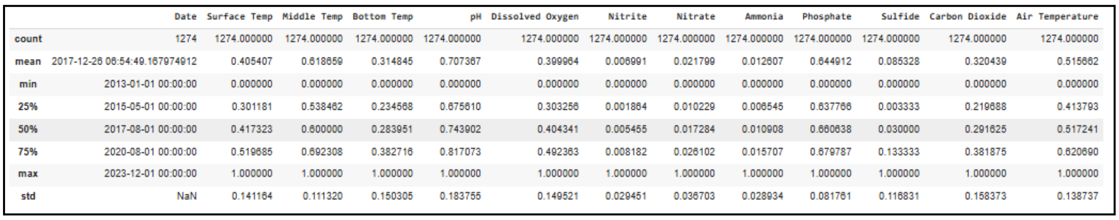


```
#   Column             Non-Null Count  Dtype
--  ------             --------------  -----
0   Date               1274 non-null   datetime64[ns]
1   Site               1274 non-null   object
2   Surface Temp       1274 non-null   float64
3   Middle Temp        1274 non-null   float64
4   Bottom Temp        1274 non-null   float64
5   pH                 1274 non-null   float64
6   Dissolved Oxygen   1274 non-null   float64
7   Nitrite            1274 non-null   float64
8   Nitrate            1274 non-null   float64
9   Ammonia            1274 non-null   float64
10  Phosphate          1274 non-null   float64
11  Sulfide            1274 non-null   float64
12  Carbon Dioxide     1274 non-null   float64
13  Weather Condition  1274 non-null   object
14  Wind Direction     1274 non-null   object
15  Air Temperature    1274 non-null   float64
```

```
Pre-processed Dataset:
        Date      Site  Surface Temp  Middle Temp  Bottom Temp        pH  \
0 2013-01-01   TANAUAN      0.155906     0.483077     0.170370  0.821951
1 2013-01-01   TALISAY      0.161417     0.488462     0.179012  0.848780
2 2013-02-01       AYA      0.175591     0.502308     0.201235  0.846341
3 2013-02-01   TUMAWAY      0.177165     0.503846     0.203704  0.851220
4 2013-02-01  SAMPALOC      0.211024     0.536923     0.256790  0.882927

   Dissolved Oxygen   Nitrite   Nitrate   Ammonia  Phosphate   Sulfide  \
0          0.041801  0.001818  0.036332  0.036213   0.655851  0.002667
1          0.418006  0.001818  0.032804  0.000000   0.720213  0.026000
2          0.319936  0.006364  0.034215  0.024433   0.698404  0.004667
3          0.429260  0.006364  0.039859  0.023124   0.672872  0.010667
4          0.486334  0.006364  0.045150  0.023124   0.671277  0.090000

   Carbon Dioxide Weather Condition Wind Direction  Air Temperature
0        0.221500   Cloudy to sunny             NE         0.364138
1        0.158875   Cloudy to sunny             NE         0.368966
2        0.215375            Cloudy             NE         0.517241
3        0.427625             Sunny             NE         0.517241
4        0.441000             Sunny             NE         0.517241
```

Figure 1. Taal Lake Water Quality Dataset after Preprocessing

To start the Exploratory Data Analysis (EDA) process, the preprocessed dataset was first described which shows the summarized statistics of each parameter that will be used in the data analysis and data modelling processes. This includes the important statistics such as the count, mean, minimum and maximum value, standard deviation and the percentage in which values lie.

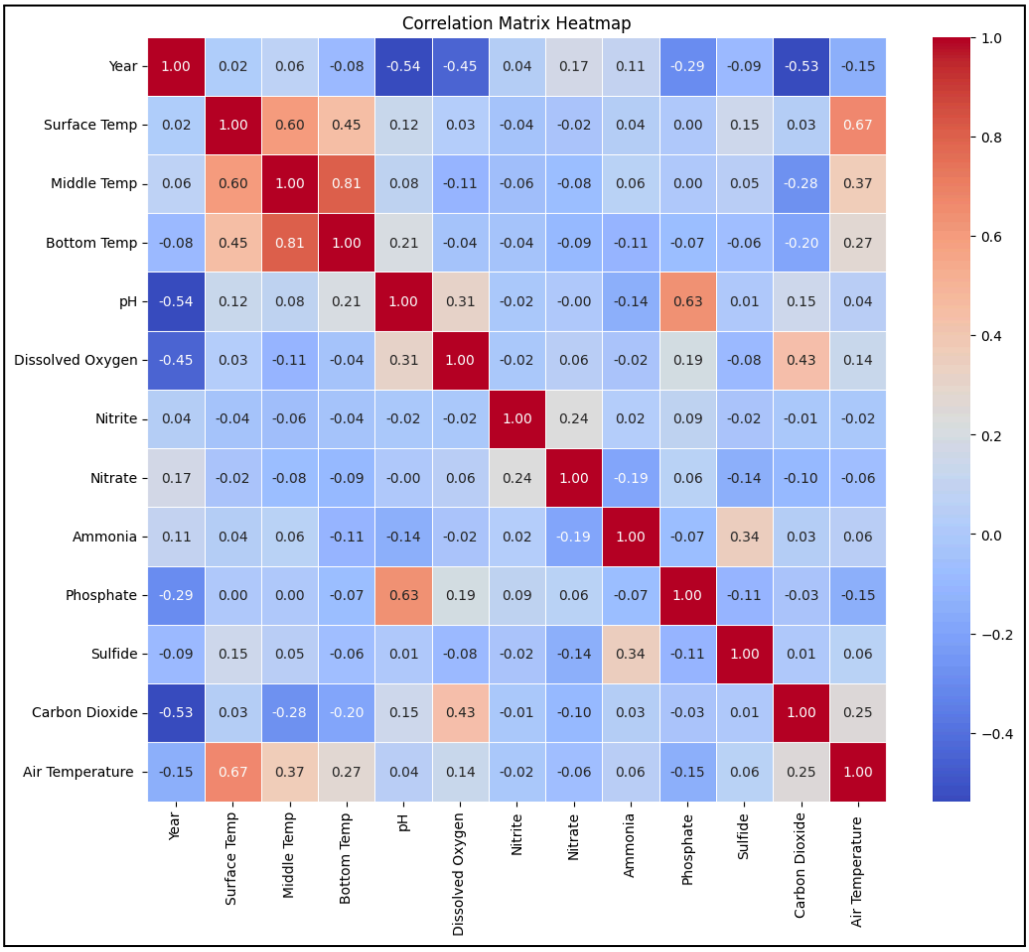


	Date	Surface Temp	Middle Temp	Bottom Temp	pH	Dissolved Oxygen	Nitrite	Nitrate	Ammonia	Phosphate	Sulfide	Carbon Dioxide	Air Temperature
count	1274	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000	1274.000000
mean	2017-12-26 06:54:49.167974912	0.405407	0.618659	0.314845	0.707367	0.399964	0.006991	0.021799	0.012607	0.644912	0.085328	0.320439	0.515662
min	2013-01-01 00:00:00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2015-05-01 00:00:00	0.301181	0.538462	0.234568	0.675610	0.303256	0.001864	0.010229	0.006545	0.637766	0.003333	0.219688	0.413793
50%	2017-08-01 00:00:00	0.417323	0.600000	0.283951	0.743902	0.404341	0.005455	0.017284	0.010908	0.660638	0.030000	0.291625	0.517241
75%	2020-08-01 00:00:00	0.519685	0.692308	0.382716	0.817073	0.492363	0.008182	0.026102	0.015707	0.679787	0.133333	0.381875	0.620690
max	2023-12-01 00:00:00	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
std	NaN	0.141164	0.111320	0.150305	0.183755	0.149521	0.029451	0.036703	0.028934	0.081761	0.116831	0.158373	0.138737

Figure 2. Summary statistics of the preprocessed dataset

Impact of Environmental and Volcanic Activity on Water Quality

With the inclusion of the other parameters to account for the volcanic activity and environmental factors, as shown in Figure 3, the parameters have relatively low negative correlation. However, some parameters show a high positive and negative correlation between them. This includes the positive correlations between the Water Temperatures, Ammonia and Phosphate, and Surface and Air temperature. This shows that the water temperatures are highly correlated with the air temperature affecting the temperature of the surface water and moderately affecting the middle and bottom temperatures. The positive rise between the Ammonia and Phosphate might be due to pollution or biological processes. Some parameters also show a negative correlation such as the Year in correlation with pH, Dissolved Oxygen, Phosphate and Carbon Dioxide. This suggests that pH, Dissolved Oxygen, Phosphate and Carbon Dioxide, overtime, decreases which might be due to acidification, warmer water temperatures that may be due to climate change, lesser biological activity and seasonal changes, respectively. pH and Dissolved Oxygen, Sulfide and Ammonia, and Carbon Dioxide and Dissolved Oxygen also shows a moderately positive correlation. Carbon Dioxide also has a moderately negative correlation with the middle and bottom water temperature.



	Year	Surface Temp	Middle Temp	Bottom Temp	pH	Dissolved Oxygen	Nitrite	Nitrate	Ammonia	Phosphate	Sulfide	Carbon Dioxide	Air Temperature
Year	1.00	0.02	0.06	-0.08	-0.54	-0.45	0.04	0.17	0.11	-0.29	-0.09	-0.53	-0.15
Surface Temp	0.02	1.00	0.60	0.45	0.12	0.03	-0.04	-0.02	0.04	0.00	0.15	0.03	0.67
Middle Temp	0.06	0.60	1.00	0.81	0.08	-0.11	-0.06	-0.08	0.06	0.00	0.05	-0.28	0.37
Bottom Temp	-0.08	0.45	0.81	1.00	0.21	-0.04	-0.04	-0.09	-0.11	-0.07	-0.06	-0.20	0.27
pH	-0.54	0.12	0.08	0.21	1.00	0.31	-0.02	-0.00	-0.14	0.63	0.01	0.15	0.04
Dissolved Oxygen	-0.45	0.03	-0.11	-0.04	0.31	1.00	-0.02	0.06	-0.02	0.19	-0.08	0.43	0.14
Nitrite	0.04	-0.04	-0.06	-0.04	-0.02	-0.02	1.00	0.24	0.02	0.09	-0.02	-0.01	-0.02
Nitrate	0.17	-0.02	-0.08	-0.09	-0.00	0.06	0.24	1.00	-0.19	0.06	-0.14	-0.10	-0.06
Ammonia	0.11	0.04	0.06	-0.11	-0.14	-0.02	0.02	-0.19	1.00	-0.07	0.34	0.03	0.06
Phosphate	-0.29	0.00	0.00	-0.07	0.63	0.19	0.09	0.06	-0.07	1.00	-0.11	-0.03	-0.15
Sulfide	-0.09	0.15	0.05	-0.06	0.01	-0.08	-0.02	-0.14	0.34	-0.11	1.00	0.01	0.06
Carbon Dioxide	-0.53	0.03	-0.28	-0.20	0.15	0.43	-0.01	-0.10	0.03	-0.03	0.01	1.00	0.25
Air Temperature	-0.15	0.67	0.37	0.27	0.04	0.14	-0.02	-0.06	0.06	-0.15	0.06	0.25	1.00

Figure 3. Correlation Heatmap of Taal Lake Water Quality Dataset

The time series line plots shown in Figures 4 to 8 show the behavior of each parameter overtime. Majority of the plots show a decline during the early and later months of the year while it increases during the middle of the year. This includes the surface, middle and bottom water temperatures, carbon dioxide and air temperature. pH shows a relatively similar value with a sudden drop towards the end of the graph, similar to phosphate. Dissolved oxygen shows a fluctuating graph as well as the carbon dioxide. While nitrite, nitrate and ammonia shows a very low valued graph which means it has a relatively low value while having a couple of peaks here and there due to events of contamination, pollution or environmental/natural events. Sulfide also shows a low graph with a lot of low to moderate spikes and having small amounts of big spikes towards the middle of the year around 2016 to 2018.

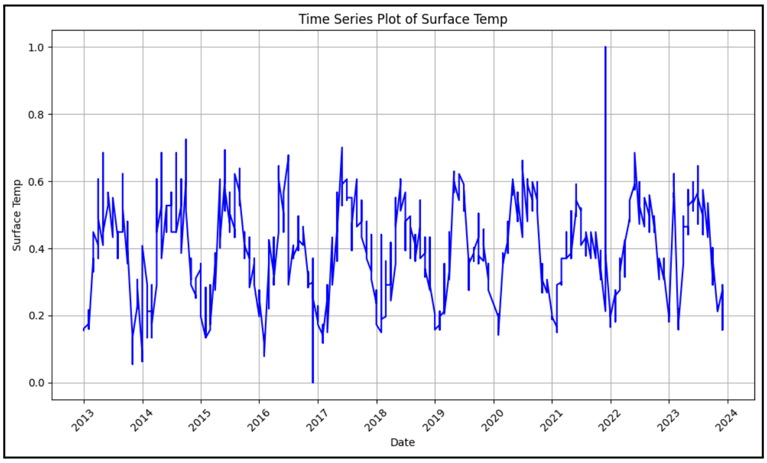


Figure 4. Time Series Plot for the Surface Temperature

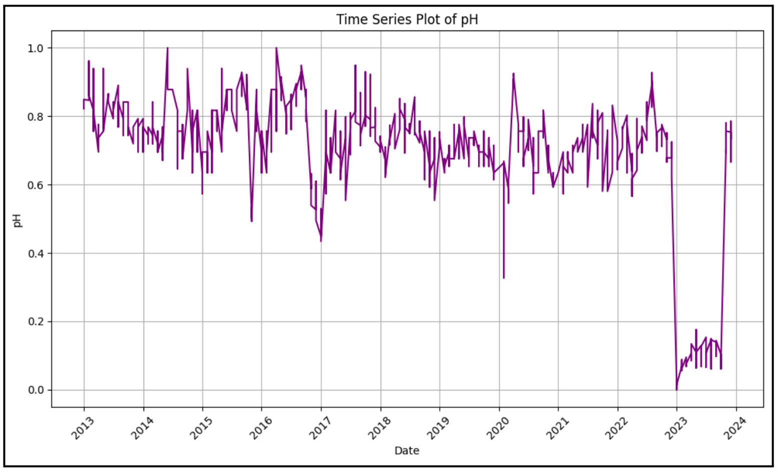


Figure 5. Time Series Plot for the pH

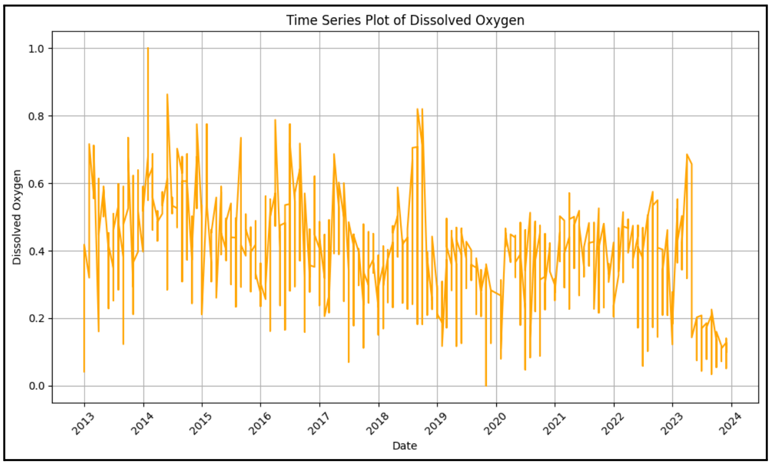


Figure 6. Time Series Plot for the Dissolved Oxygen

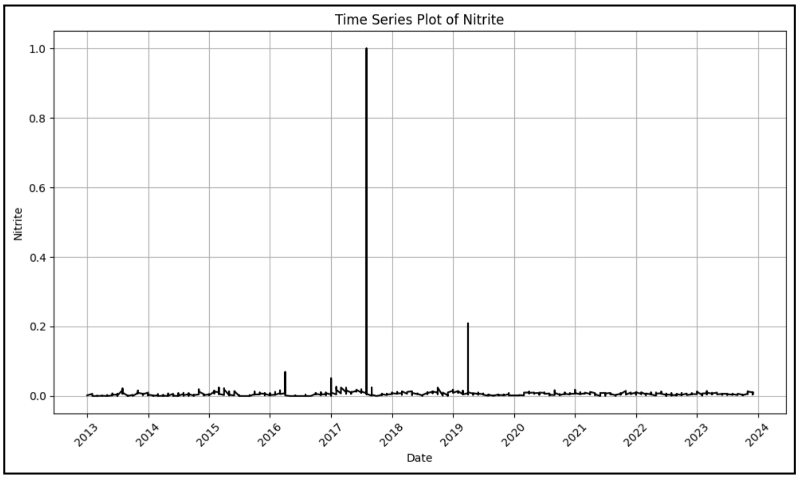


Figure 7. Time Series Plot for the Nitrite

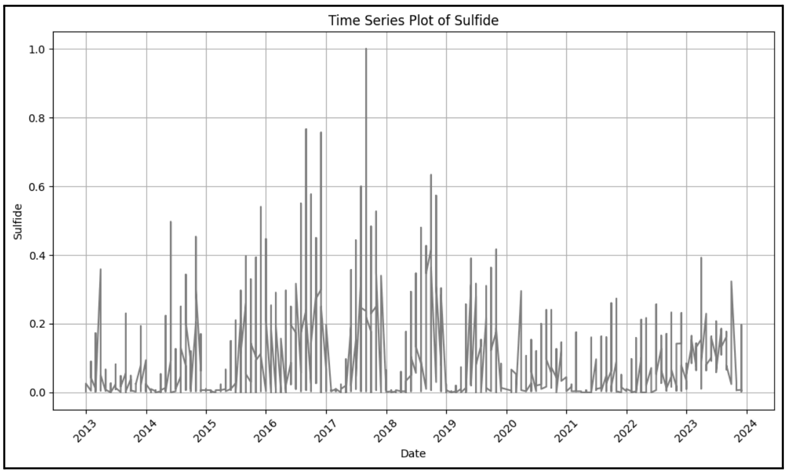


Figure 8. Time Series Plot for the Sulfide

The effects of environmental factors and volcanic activity to the water quality of Taal Lake was also taken into account and analyzed. Scatter plots were used to see if there was any correlation between the parameters. In environmental factors, in Figure 9, the surface water temperature shows a positive trend towards air temperature which means that surface water temperature increases as the air temperature increases. Dissolved oxygen and air temperature also shows a lower positive trend than with surface water temperature.

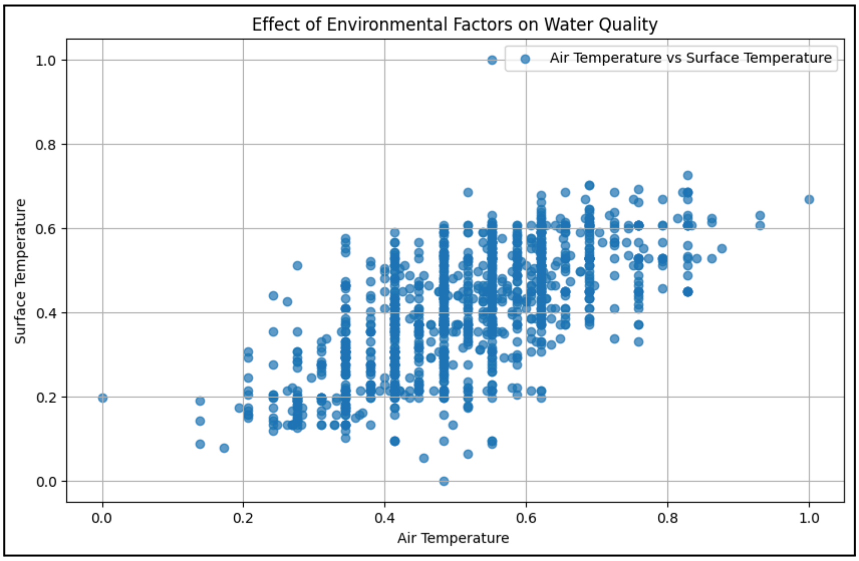


Figure 9. Correlation between Surface water temperature and Air Temperature

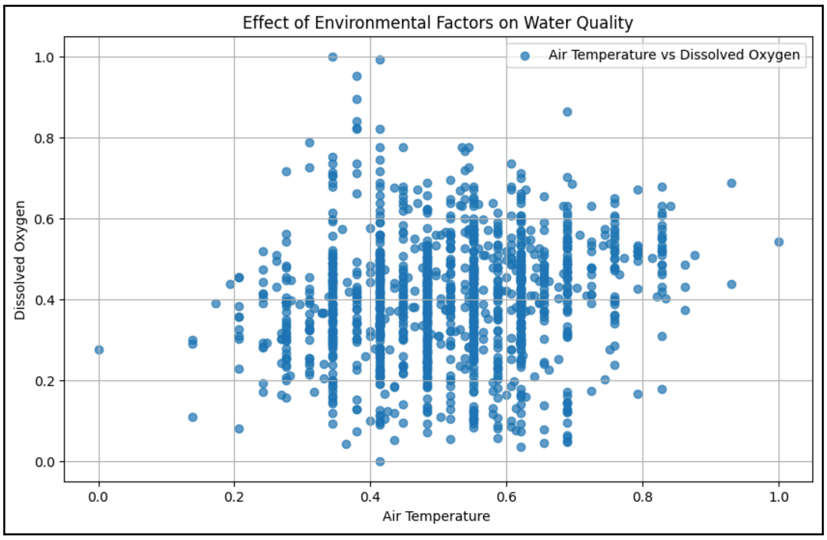


Figure 10. Correlation between Surface water temperature and Air Temperature

Scatter plots were also used to show the effect of wind direction to the water quality of the Taal Lake. As shown in the graphs, the wind direction does not have much effect on the ammonia. The graph in nitrate and nitrite looks similar to ammonia but it may be attributed to its lack of data. The wind direction had the most effect in phosphate and dissolved oxygen as seen in Figure 11. Each wind direction tends to have a great effect on the phosphate while in the dissolved oxygen, the result varies with North Eastern and South Western winds having the majority of the data and showing the sensitivity of the dissolved oxygen to the wind direction.

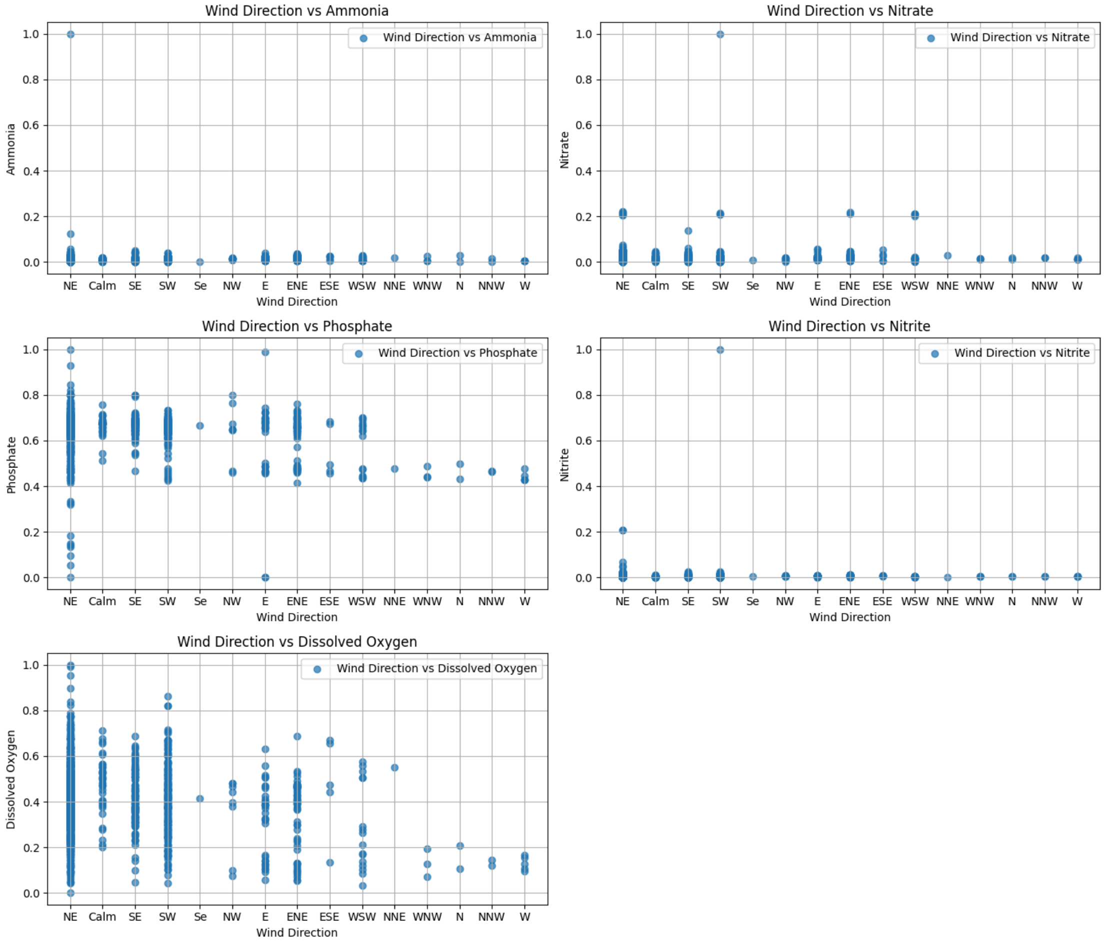


Figure 11. Correlation of Wind Direction to the Water Quality of Taal Lake

As per the weather condition's effect to the water quality of Taal Lake, the weather condition does not have much effect on ammonia, nitrite and nitrate while having the most effect in pH, dissolved oxygen and phosphate wherein the values vary the most. pH and phosphate tend to have higher values especially on days with cloudy, fair, partly cloudy, or sunny weather while having some days with normal values. While dissolved oxygen shows the most variation, especially on days with relatively cloudy, sunny or fair days.

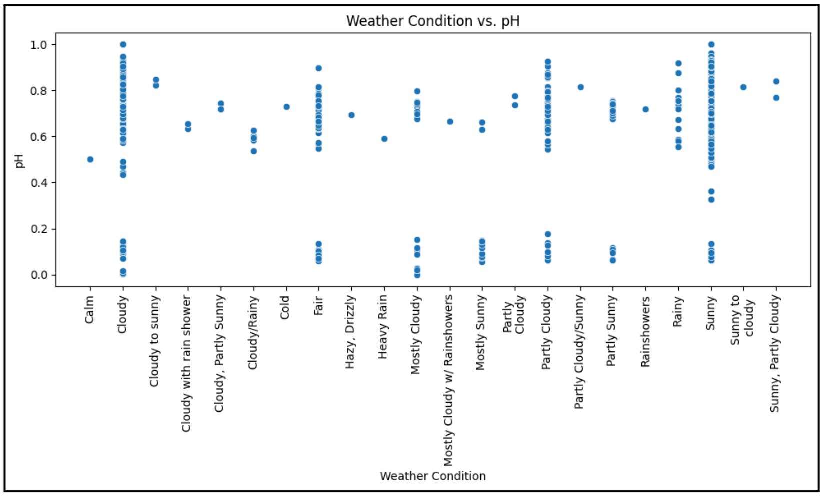


Figure 12. Correlation of Weather Condition to pH

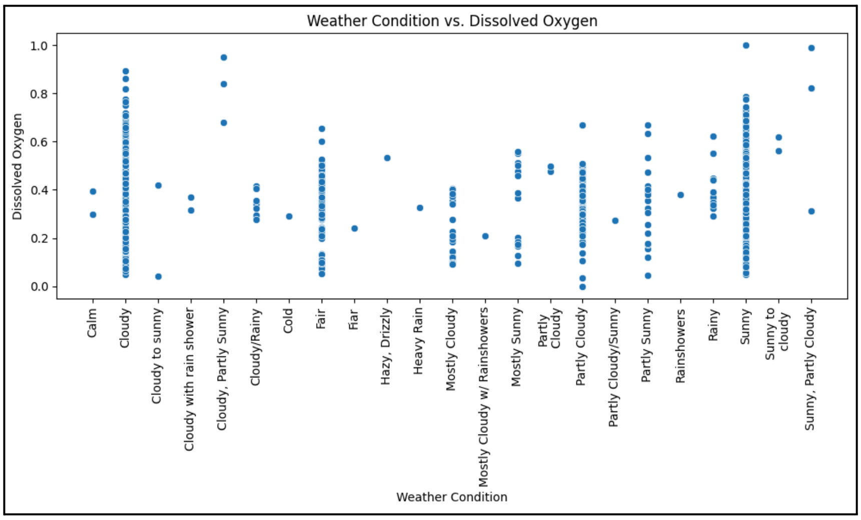


Figure 13. Correlation of Weather Condition to the Dissolved Oxygen

To measure the volcanic activity's effect on the water quality, the correlation between the carbon dioxide and sulfide to the pH and dissolved oxygen were visualized. It is shown in Figure 14 that in the correlation between carbon dioxide and pH, the value of pH tends to rise when the value of the carbon dioxide is low showing a negative correlation.

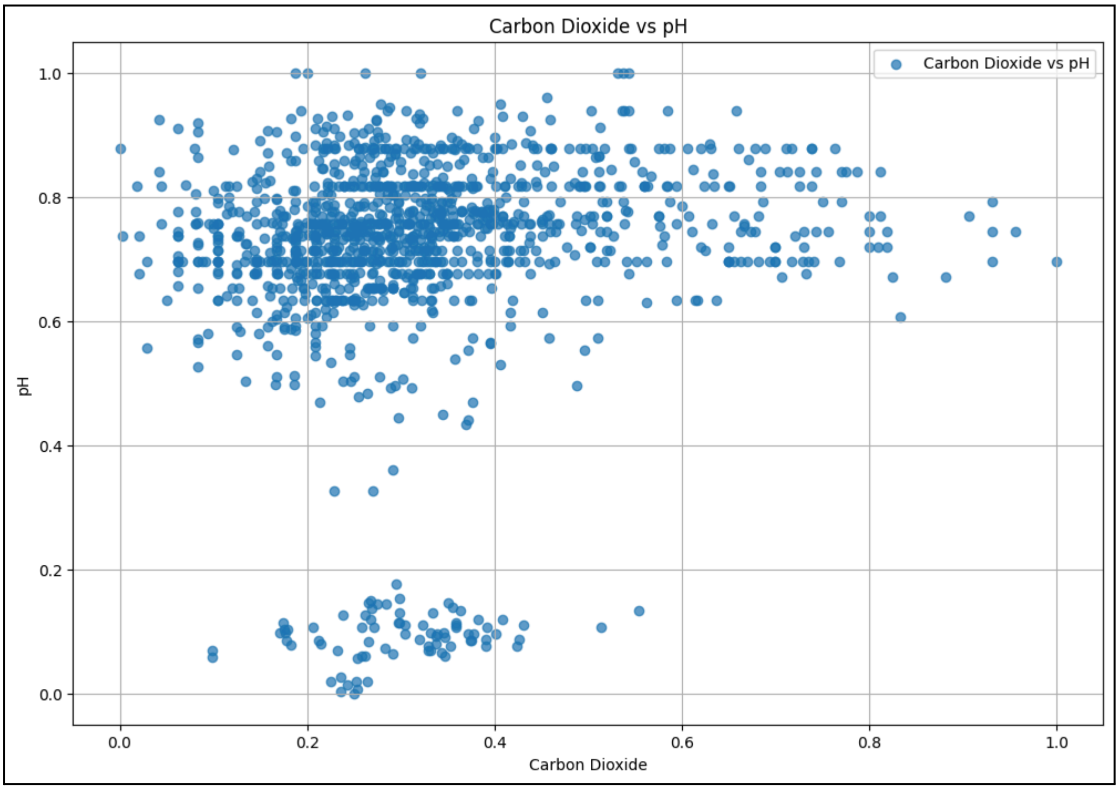


Figure 14. Correlation of Carbon Dioxide to pH

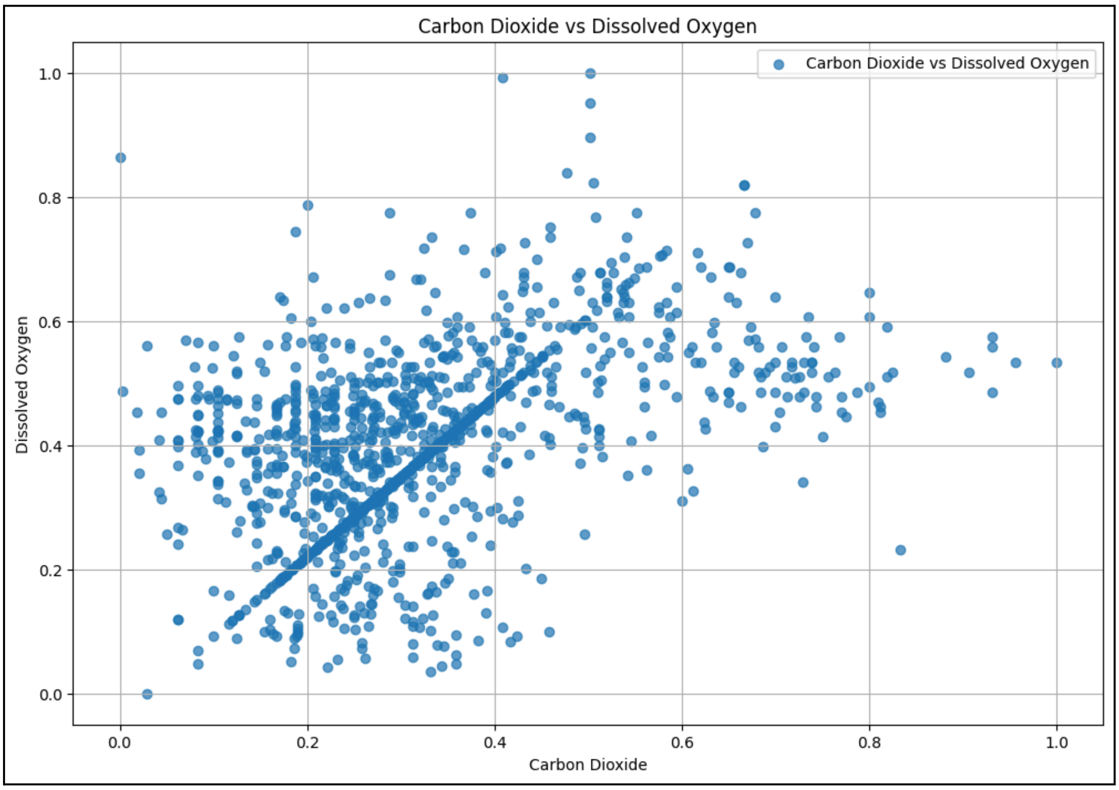


Figure 15. Correlation of Carbon Dioxide to Dissolved Oxygen

On the other hand, sulfide has little to no effect towards the values of dissolved oxygen as Figure 13 shows that, especially on lower levels of sulfide, the value of dissolved oxygen varies ranging from low to high which indicates an inconsistent relationship between the two. Unlike dissolved oxygen, sulfide, although little, has more effect on pH as even at lower values of sulfide, pH tends to have a higher value. This might mean that there are other factors to be observed when measuring the water quality of Taal Lake than Carbon Dioxide and Sulfide.

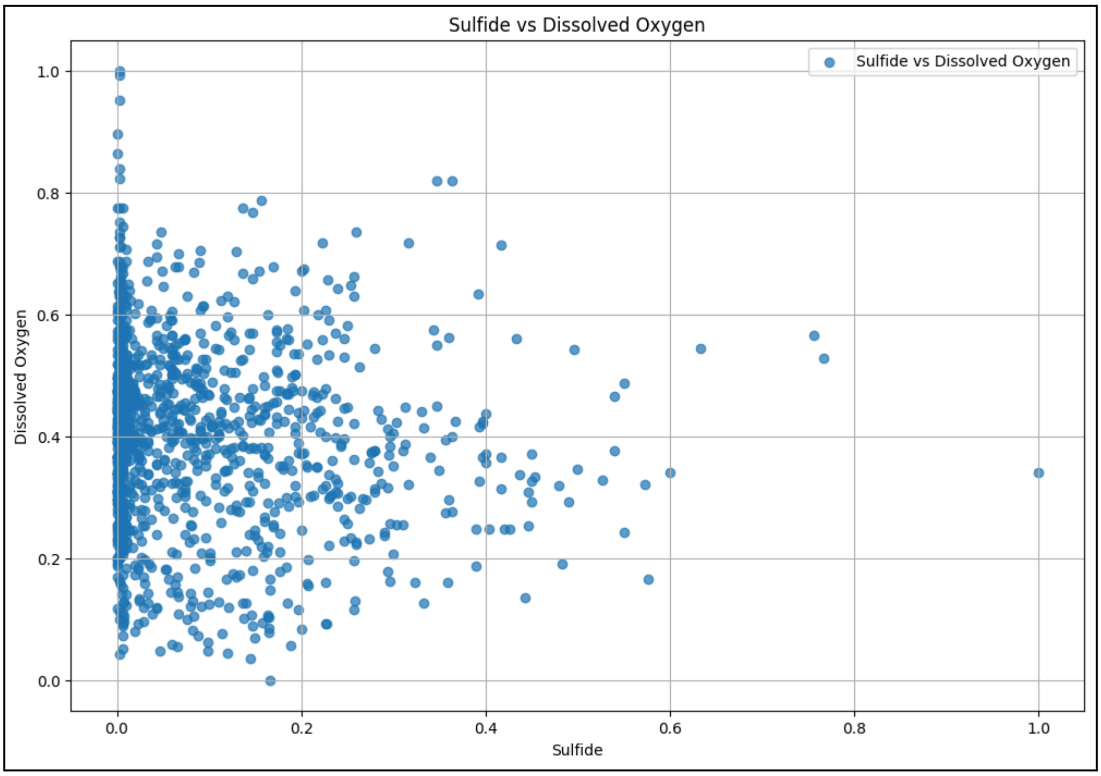


Figure 16. Correlation of Sulfide to Dissolved Oxygen

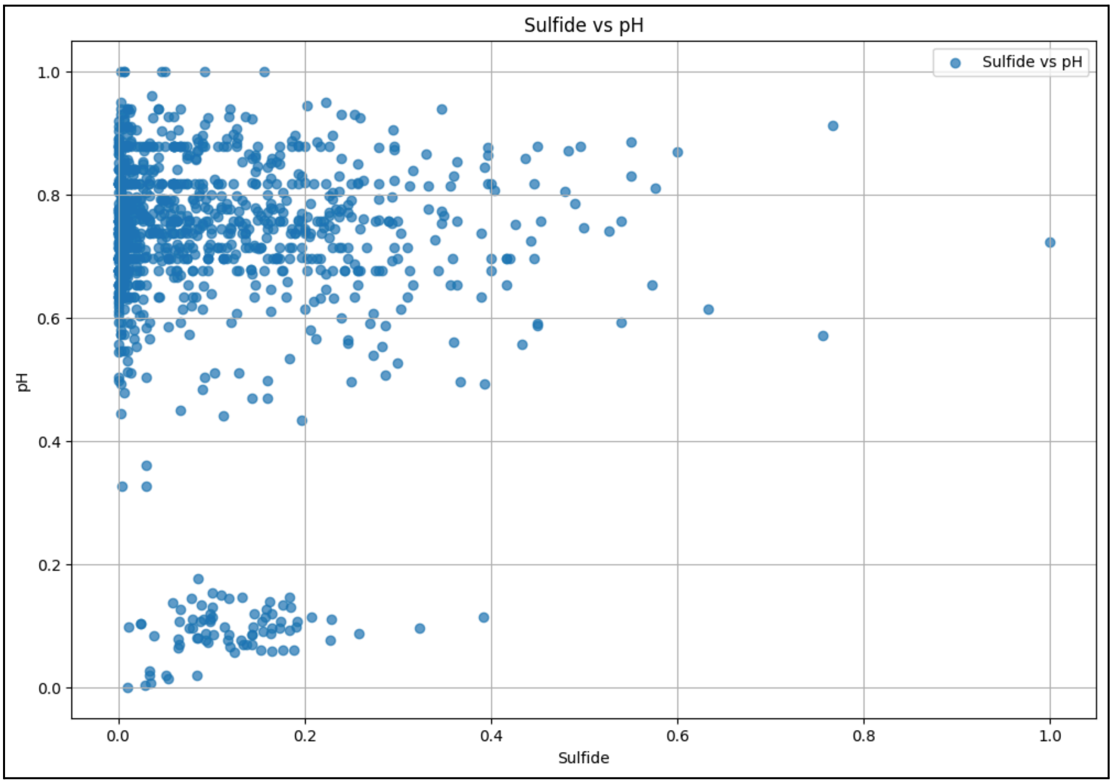


Figure 17. Correlation of Sulfide to pH

Evaluation and Comparison of CNN, LSTM and HYBRID CNN-LSTM Models

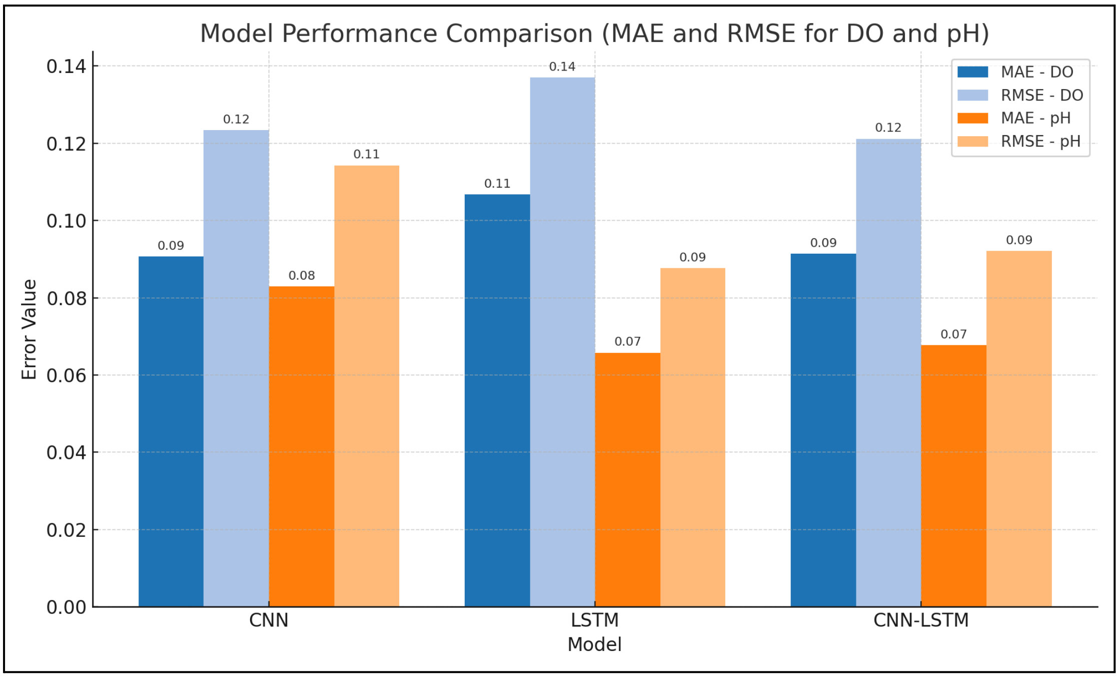


Figure 18. Comparison of CNN, LSTM, and Hybrid CNN-LSTM Models Based on MAE and RMSE for DO and pH Prediction

Convolutional Neural Network (CNN)

The CNN model, where convolutional layers captured spatial patterns, performed good, specifically when we also consider the environmental factors rather than focusing only on the water parameters. Without incorporating environmental inputs, the CNN_DO model had an MAE of 0.1118 and an RMSE of 0.1444, but the CNN_pH model performed very slightly better (MAE = 0.1080, RMSE = 0.1640). When environmental conditions were incorporated, performance greatly improved. The MAE of the CNN_DO model fell to 0.0907 and RMSE to 0.1234, and the CNN_pH model to an MAE of 0.0829 and RMSE of 0.1143. These results show CNN's strength in utilizing more contextual information to improve prediction accuracy.

Long Short-Term Memory(LSTM)

The LSTM model, which is suitable for sequential time-series data, performed optimally in pH prediction when water parameters alone were used. The LSTM_pH model performed an excellent MAE of 0.0658 and RMSE of 0.0877, better than CNN and CNN-LSTM on that particular task. But its performance was affected when environmental features were added LSTM_pH model's MAE increased to 0.1149, and RMSE to 0.1643, suggesting that the added features added noise or complexity the LSTM could not model well. Correspondingly, in case of DO prediction, LSTM performance degraded somewhat on the inclusion of environmental features (MAE escalated from 0.1068 to 0.1108). This points toward the probability that LSTM without proper tuning can be insufficient when dealing with mixed-type inputs.

HYBRID CNN-LSTM

The dual CNN-LSTM model, mixing the spatial potential of CNN and temporal modeling power of LSTM, tended to increase when environmental elements were incorporated. Without environmental variables, the CNN-LSTM_DO model achieved an MAE of 0.1220 and RMSE of 0.1547, and the CNN-LSTM_pH model was slightly better (MAE = 0.0998, RMSE = 0.1625). But with the incorporation of environmental information, the CNN-LSTM_DO model was much better at MAE = 0.0914 and RMSE = 0.1212, and the CNN-LSTM_pH model at MAE = 0.0677, RMSE = 0.0921 practically equivalent or superior to CNN in certain respects. This indicates that the hybrid model is advantaged by the extra features and can potentially describe more complex dependencies if it is provided with spatial and temporal information.

Comparison

Overall, the three models CNN, LSTM, and hybrid CNN-LSTM showed differences in strength and weakness when it comes to modeling water quality parameters. Most typically, the hybrid CNN-LSTM model works the best, particularly when water parameters and environmental factors are taken into consideration. Its ability to maximize on CNN's spatial feature learning and LSTM's temporal modeling improved its ability to model complex, high-dimensional data with more flexibility and most importantly with more accuracy. CNN alone also works notably well, particularly with environmental data being used, suggesting it is incredibly skilled at learning patterns from structured, spatially correlated input features. LSTM works incredibly well alone in simple, time-series-only scenarios (particularly pH prediction) but does not work with larger, more complex data. This suggests LSTM is best suited to sequence data but may not generalize as well in mixed-input scenarios without additional fine-tuning. Finally, the hybrid model has the most stable performance, so it is the better option when data complexity and predictive accuracy are concerns.

Water Quality Index and Water Pollutant Levels Prediction

In predicting the values of each parameter—including the non-pollutants—in different time scales using the CNN-LSTM model, a range of future values were calculated. For weekly based prediction, values for the next four weeks were calculated, values for the next 12 months for monthly based prediction, and values for the next five years for yearly-based prediction. Focusing on the performance of the model in predicting water pollutant levels, particularly the Nitrite, Nitrate, Ammonia, and Phosphate, the model showed promising results in predicting values based on different time granularity as shown in their RMSE, MAE, and R^2 results as shown in the Tables below.

Time	Nitrite	Nitrate	Ammonia	Phosphate
Weekly	0.0054	0.0199	0.0030	0.0164
Monthly	0.0035	0.010	0.0115	0.0462
Yearly	0.0020	0.0378	0.0044	0.0442

Table 1. RMSE Results for Water Pollutant Levels Prediction with CNN-LSTM

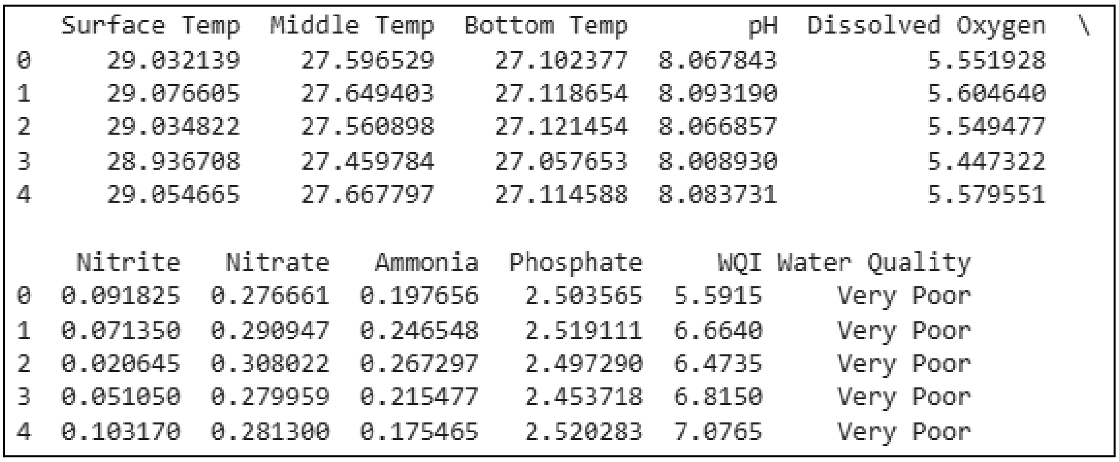
Time	Nitrite	Nitrate	Ammonia	Phosphate
Weekly	0.0053	0.0164	0.0029	0.0152
Monthly	0.0030	0.0086	0.0090	0.0417
Yearly	0.0015	0.0347	0.0039	0.0339

Table 2. MAE Results for Water Pollutant Levels Prediction with CNN-LSTM

Time	Nitrite	Nitrate	Ammonia	Phosphate
Weekly	-229.664	-19.085	-152.525	-36.733
Monthly	-0.5216	-1.4553	-0.1974	-0.3404
Yearly	-0.0463	-0.0004	-1.7214	-0.6863

Table 3. R^2 Results for Water Pollutant Levels Prediction with CNN-LSTM

The results for predicting the water pollutant levels across different time granularities provide key insights about the CNN-LSTM Hybrid model's accuracy. Overall, RMSE results indicate varying prediction performance that depends on the pollutant and the time scale. Nitrite and Ammonia predictions resulted in consistent low RMSE values across all time frames. Nitrate and Phosphate resulted in higher RMSEs, particularly at the yearly-based prediction. However, the increase in RMSE at longer time frames may be due to data aggregation.



```
   Surface Temp  Middle Temp  Bottom Temp        pH  Dissolved Oxygen  \
0     29.032139    27.596529    27.102377  8.067843          5.551928
1     29.076605    27.649403    27.118654  8.093190          5.604640
2     29.034822    27.560898    27.121454  8.066857          5.549477
3     28.936708    27.459784    27.057653  8.008930          5.447322
4     29.054665    27.667797    27.114588  8.083731          5.579551

    Nitrite   Nitrate   Ammonia  Phosphate     WQI Water Quality
0  0.091825  0.276661  0.197656   2.503565  5.5915     Very Poor
1  0.071350  0.290947  0.246548   2.519111  6.6640     Very Poor
2  0.020645  0.308022  0.267297   2.497290  6.4735     Very Poor
3  0.051050  0.279959  0.215477   2.453718  6.8150     Very Poor
4  0.103170  0.281300  0.175465   2.520283  7.0765     Very Poor
```

Figure 19. WQI Calculation Results

Based on the Water Quality Index (WQI) calculated using a weighted combination of key parameters such as pH, dissolved oxygen, nitrite, nitrate, ammonia, phosphate, and temperature at the surface, middle, and bottom of the lake, the majority of the samples in the dataset were classified as having "Very Poor" water quality. This classification indicates that the overall health of the water body is critically low, likely due to a combination of low dissolved oxygen levels and elevated nutrient concentrations such as phosphate and ammonia, which are known contributors to water pollution and eutrophication.

Conclusion and Recommendations

In conclusion, this lab activity offered practical experience in data mining, machine learning, and data visualization to evaluate the environmental factors affecting water quality. From our analysis, we noticed some clear patterns—like how hotter air makes the surface water warmer, and how more carbon dioxide usually means lower pH levels. Using this information, we created prediction models with CNN, LSTM, and a mix of both (CNN-LSTM). The LSTM model worked well when we only used time data to predict pH, but it didn't do as well when we added other environmental data. CNN did better with more environmental info. In the end, the combined CNN-LSTM model gave the best and most reliable results for predicting both dissolved oxygen and pH. Lastly, the results of the WQI calculation showed that the water around the sites near Taal Lake is very poor.

To improve the water quality of Taal Lake, we recommend regular monitoring of the lake especially for pH, oxygen levels, temperature, and contaminants that could originate from volcanic activity. Preserving or introducing native fish species that feed on algae can also support ecological balance. In addition, installing aerators can improve oxygen levels in the water, helping to suppress the growth of harmful algae. Finally, raising awareness among the local community about the importance of lake conservation and encouraging active participation are important to improve the quality of the water.

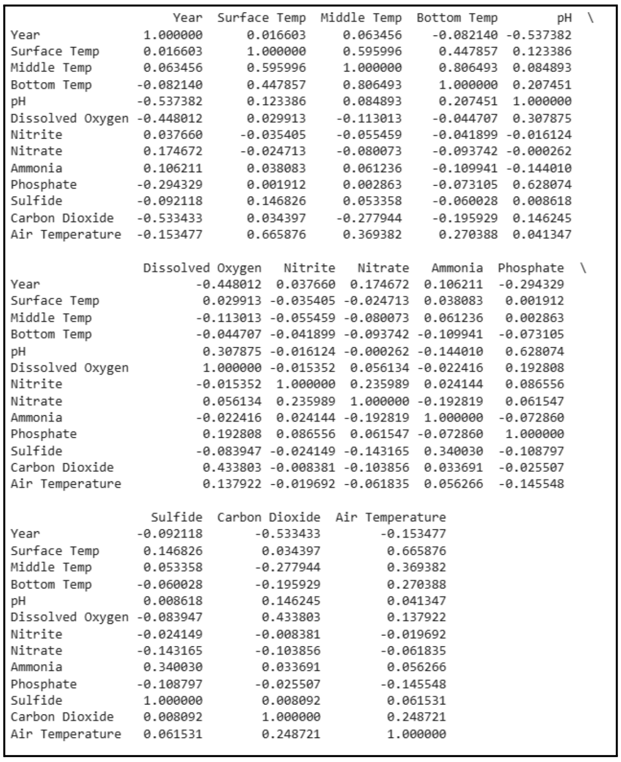
To improve the project, we can get more environmental and weather data from PHIVOLCS and PAGASA to make predictions more accurate. We should aim for daily water quality predictions for each parameter at every location. Adding data from the year 2024 will help keep the analysis up to date and improve trend detection. Fine-tuning the machine learning model settings can also boost performance and reduce errors. Lastly, using newer models like Transformers or AutoML can help improve prediction by handling sequences better and automating the optimization process.

APPENDICES

Appendix A. Google Colab Link for MC02
https://colab.research.google.com/drive/13T3xTy8WtG3WZLI94KjlOJ78gy7esQny

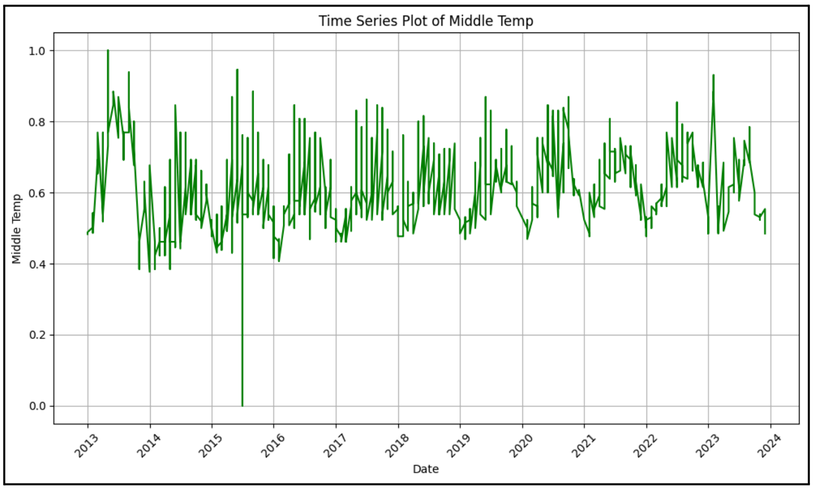
Appendix B. Streamlit Website Link
https://mc02-cpen106-dlmppv.streamlit.app/?embed_options=light_theme

Appendix C. Correlation Matrix of the Parameters

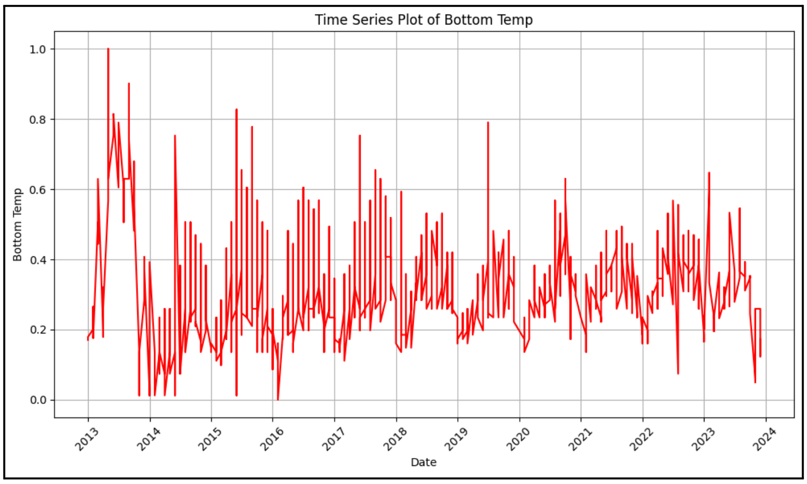


```
                      Year  Surface Temp  Middle Temp  Bottom Temp        pH  \
Year              1.000000      0.016603     0.063456    -0.082140 -0.537382   
Surface Temp      0.016603      1.000000     0.595996     0.447857  0.123386   
Middle Temp       0.063456      0.595996     1.000000     0.806493  0.084893   
Bottom Temp      -0.082140      0.447857     0.806493     1.000000  0.207451   
pH               -0.537382      0.123386     0.084893     0.207451  1.000000   
Dissolved Oxygen -0.448012      0.029913    -0.113013    -0.044707  0.307875   
Nitrite           0.037660     -0.035405    -0.055459    -0.041899 -0.016124   
Nitrate           0.174672     -0.024713    -0.080073    -0.093742 -0.000262   
Ammonia           0.106211      0.038083     0.061236    -0.109941 -0.144010   
Phosphate        -0.294329      0.001912     0.002863    -0.073105  0.628074   
Sulfide          -0.092118      0.146826     0.053358    -0.060028  0.008618   
Carbon Dioxide   -0.533433      0.034397    -0.277944    -0.195929  0.146245   
Air Temperature  -0.153477      0.665876     0.369382     0.270388  0.041347   

                  Dissolved Oxygen   Nitrite   Nitrate   Ammonia  Phosphate  \
Year                     -0.448012  0.037660  0.174672  0.106211  -0.294329   
Surface Temp              0.029913 -0.035405 -0.024713  0.038083   0.001912   
Middle Temp              -0.113013 -0.055459 -0.080073  0.061236   0.002863   
Bottom Temp              -0.044707 -0.041899 -0.093742 -0.109941  -0.073105   
pH                        0.307875 -0.016124 -0.000262 -0.144010   0.628074   
Dissolved Oxygen          1.000000 -0.015352  0.056134 -0.022416   0.192808   
Nitrite                  -0.015352  1.000000  0.235989  0.024144   0.086556   
Nitrate                   0.056134  0.235989  1.000000 -0.192819   0.061547   
Ammonia                  -0.022416  0.024144 -0.192819  1.000000  -0.072860   
Phosphate                 0.192808  0.086556  0.061547 -0.072860   1.000000   
Sulfide                  -0.083947 -0.024149 -0.143165  0.340030  -0.108797   
Carbon Dioxide            0.433803 -0.008381 -0.103856  0.033691  -0.025507   
Air Temperature           0.137922 -0.019692 -0.061835  0.056266  -0.145548   

                   Sulfide  Carbon Dioxide  Air Temperature  
Year             -0.092118       -0.533433        -0.153477  
Surface Temp      0.146826        0.034397         0.665876  
Middle Temp       0.053358       -0.277944         0.369382  
Bottom Temp      -0.060028       -0.195929         0.270388  
pH                0.008618        0.146245         0.041347  
Dissolved Oxygen -0.083947        0.433803         0.137922  
Nitrite          -0.024149       -0.008381        -0.019692  
Nitrate          -0.143165       -0.103856        -0.061835  
Ammonia           0.340030        0.033691         0.056266  
Phosphate        -0.108797       -0.025507        -0.145548  
Sulfide           1.000000        0.008092         0.061531  
Carbon Dioxide    0.008092        1.000000         0.248721  
Air Temperature   0.061531        0.248721         1.000000  
```

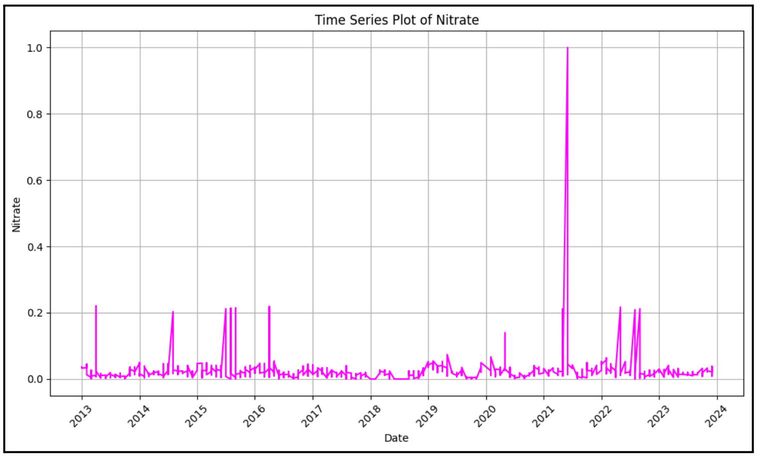
Appendix D. Time Series Plot for the Middle Temperature



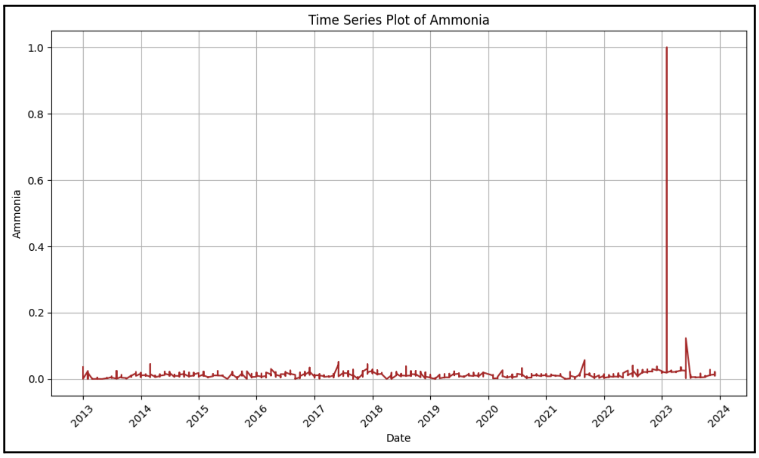
Appendix E. Time Series Plot for the Bottom Temperature



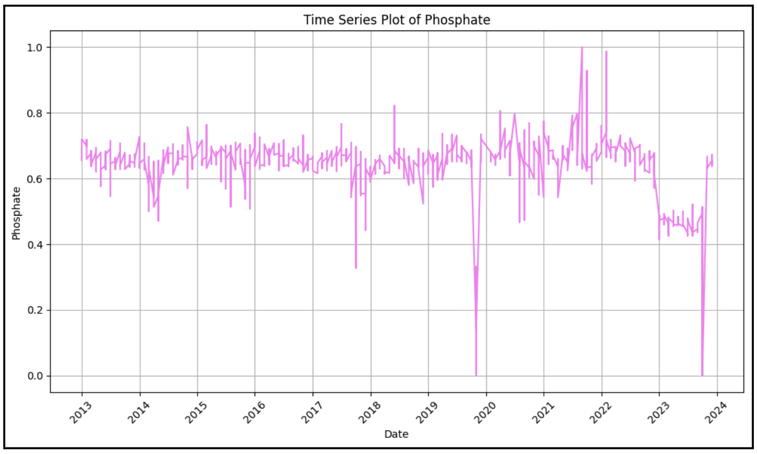
Appendix F. Time Series Plot for the Nitrate



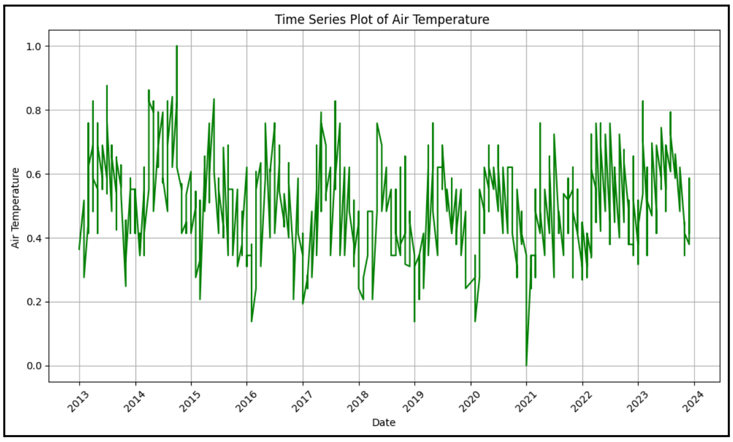
Appendix G. Time Series Plot for the Ammonia



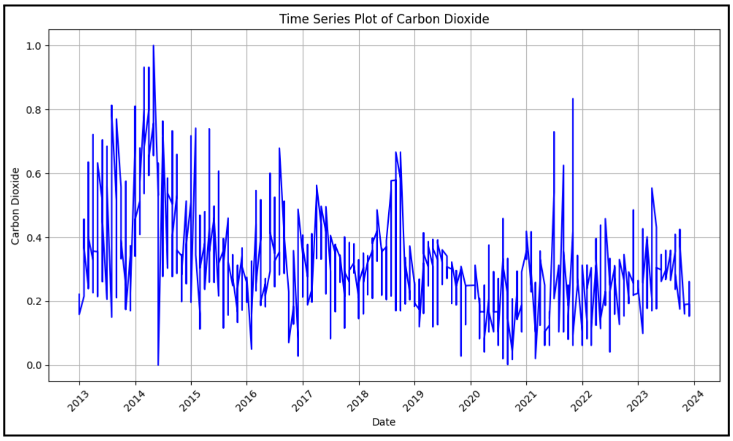
Appendix H. Time Series Plot for the Phosphate



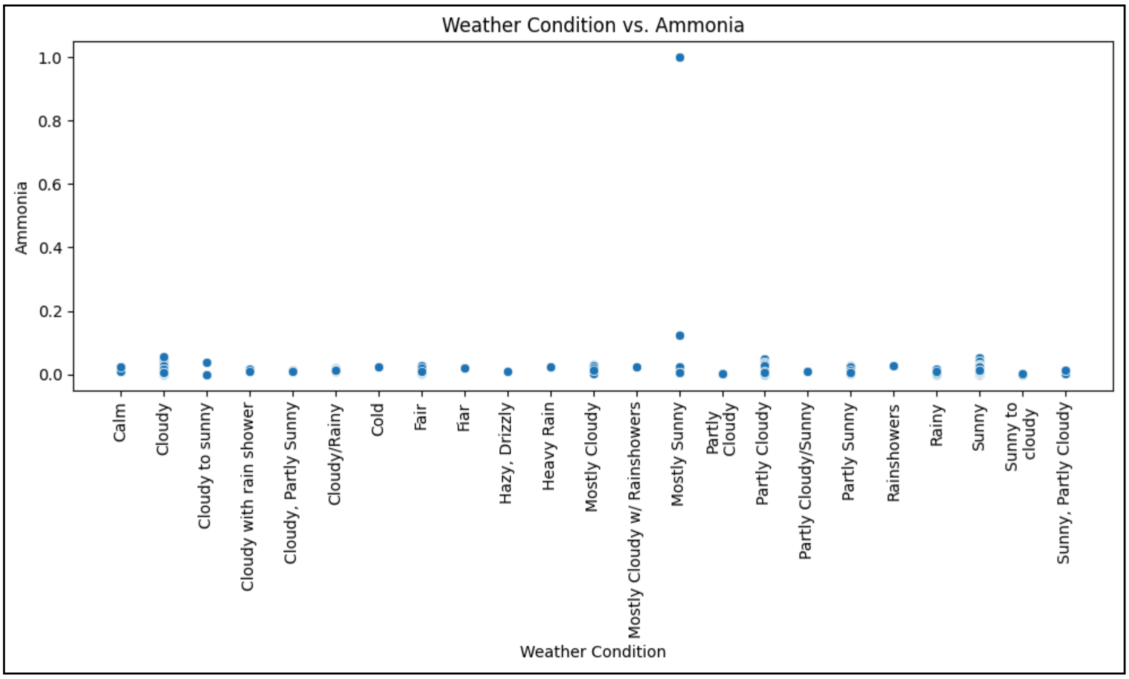
Appendix I. Time Series Plot for the Air Temperature



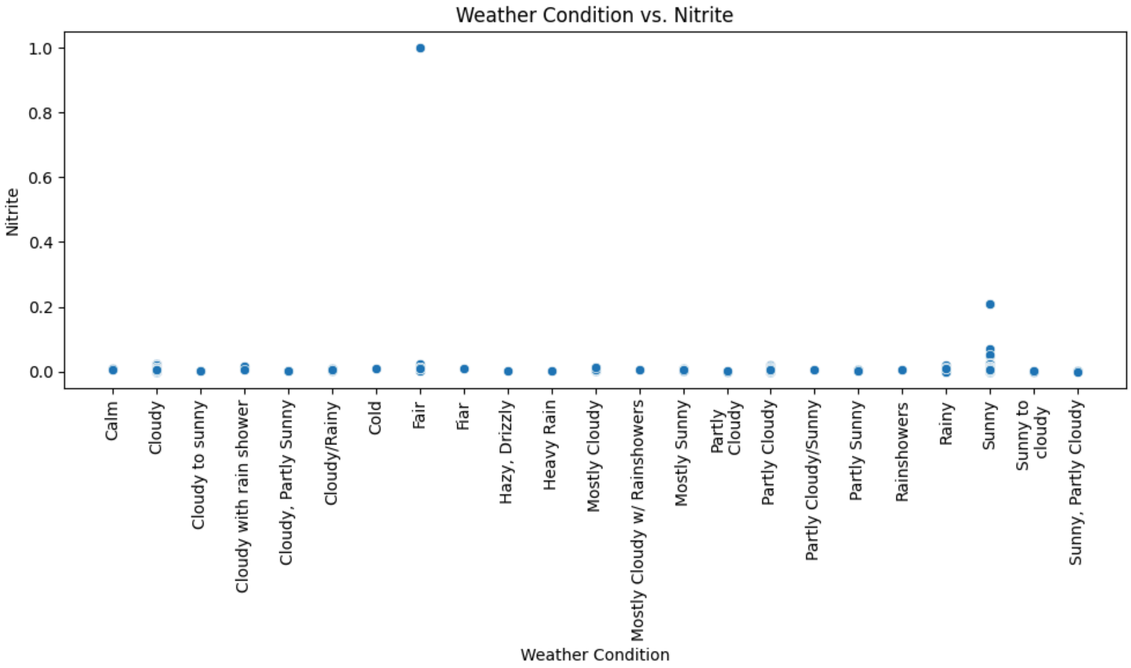
Appendix J. Time Series Plot for the Carbon Dioxide



Appendix K. Correlation between the Weather Condition and Ammonia



Appendix K. Correlation between the Weather Condition and Nitrite


Weather Condition vs. Nitrite
Nitrite
Calm
Cloudy
Cloudy to sunny
Cloudy with rain shower
Cloudy, Partly Sunny
Cloudy/Rainy
Cold
Fair
Fiar
Hazy, Drizzly
Heavy Rain
Mostly Cloudy
Mostly Cloudy w/ Rainshowers
Mostly Sunny
Partly Cloudy
Partly Cloudy
Partly Cloudy/Sunny
Partly Sunny
Rainshowers
Rainy
Sunny
Sunny to cloudy
Sunny, Partly Cloudy
Weather Condition