Course: Time Series Analysis for Environmental Sciences

Name: Yehuda Yungstein

ID: 205487143

This report provides an overview of the project documented solely through code, as detailed in the attached codebook. While the report is brief and condensed, I suggest consulting the code notebook for further elaboration, interactive graphs, and additional information. The report's structure mirrors that of the code notebook, making it easier to navigate between the two documents.

1. Introduction

This study delves into the dynamics of the Hula Lake ecosystem in Israel (N 33° 6.200040 E 35° 36.380760), employing a comprehensive analysis of meteorological and remote sensing data. The overarching goal is to elucidate the environmental impacts of water and vegetation factors within this unique lacustrine environment, with a particular emphasis on describing the changes between the lake and the vegetation affecting the energy balance and carbon emissions.

Wetlands contain a disproportionate amount of the earth's total soil carbon; holding between 20 and 30% of the estimated 1,500 Pg of global soil carbon despite occupying 5–8% of its land surface [Nahlik and Fennessy., 2016], underscoring the necessity to isolate and monitor the various factors within these ecosystems. This project aims to achieve an initial understanding of the differences in the annual and daily ranges of various parameters between the water station and the vegetation station, paving the way for future forecasting and linking remote sensing data to physical parameters.

1.1 Data Acquisition

The dataset comprises in-situ measurements obtained from two meteorological stations: one situated on the lake's water surface and the other on a vegetated island within the lake (Fig 1)**.** These stations recorded a atmospheric and hydrological parameters, including:

- Temperature
- Relative humidity
- Wind speed and direction
- Energy balance components
- Water fluxes
- Carbon dioxide levels
- Evaporation
- Precipitation

Additionally, remotely sensed data from the Sentinel-2 satellite was acquired, spanning the period from 2017 to 2024, with a temporal resolution of 5 days and a spatial resolution of 10 meters.

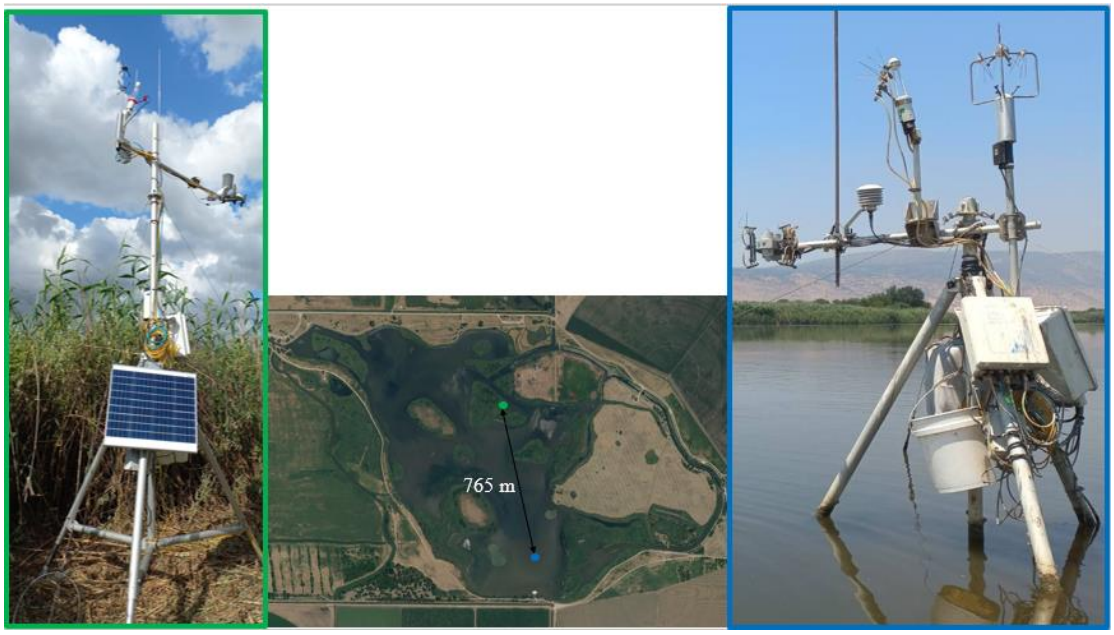


Figure 1 - Map of the Hula Lake ecosystem, depicting the locations of the water and vegetation meteorological stations.

1.2 Analysis Approach

The analysis is divided into two distinct phases: near-term and long-term.

Near-Term Analysis

The near-term analysis focuses on the high-frequency meteorological data collected from October 23, 2023, to February 24, 2024, encompassing approximately 9 months. The dataset consists of 12,000 rows and 25 columns, with measurements sampled at a half-hourly temporal resolution.

The analysis will proceed as follows:

1. **Data Preprocessing**: Implement cleaning procedures to handle noise and missing data (date formatting, outliers, filling gaps, etc.).
2. **Daily Trend Comparison**: Compare daily trends in recorded parameters between the two stations to discern potential micro-environmental differences.
3. **Cross-Correlation Analysis**: Employ cross-correlation analysis on the daily trends to quantify time differences.

Long-Term Analysis

For the long-term analysis, remotely sensed data from the Sentinel-2 satellite will be utilized, resulting in approximately 300 samples of average NDVI (Normalized Difference Vegetation Index) values for the lake water and vegetated island.

The analysis will involve the following steps:

1. **Time Series Generation**: Generate time series of average NDVI values for the lake water and vegetated island.
2. **Smoothing and Resampling**: Perform smoothing and resampling procedures on the time series data.
3. **Decomposition**: Decompose the time series into trend and cyclical components.
4. **Trend Observation**: Observe and compare annual trends in vegetation dynamics between the lake and the island.
5. **Stationarity Test**: Conduct stationarity tests on the time series data.
6. **ACF and PACF**: Analyze the autocorrelation and partial autocorrelation functions of the time series data.
7. **ARIMA Modeling**: Implement ARIMA modeling as a forecasting tool.

Throughout the analysis, I will adhere to rigorous scientific standards, employing appropriate statistical techniques and data visualization methods to effectively communicate our findings.

2. Near-Term Analysis

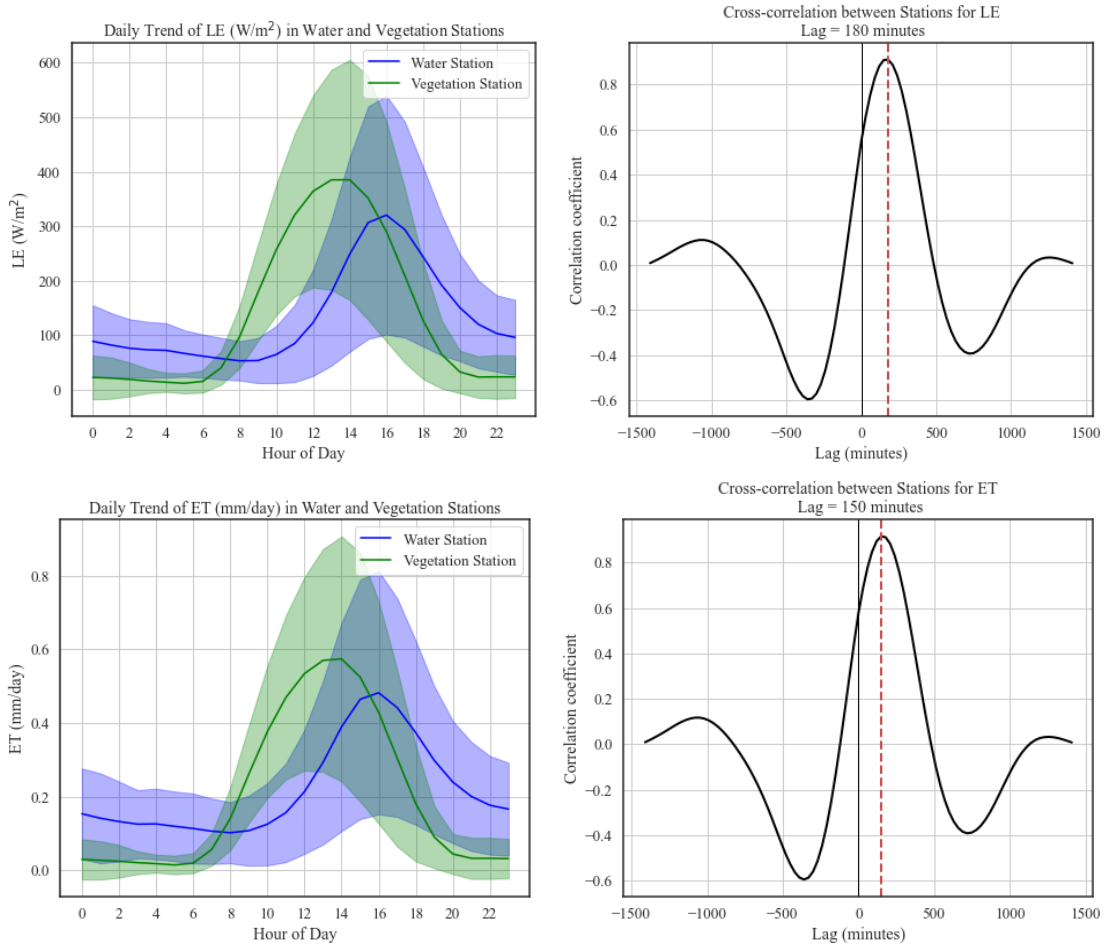
2.1 Data Cleaning
2.2 Filling Gaps
2.3 Smoothing the Data
2.4 Add Soil / Water Heat Flux Based on Energy Balance Equation
2.5 Plot the Daily Trend

1. Plot the daily trends of the parameters for both stations.

2. Calculate Cross-Correlation function between the daily trends of the parameters for both stations.

3. Compare the time lag for the maximum correlation.

LE and ET Daily Trend



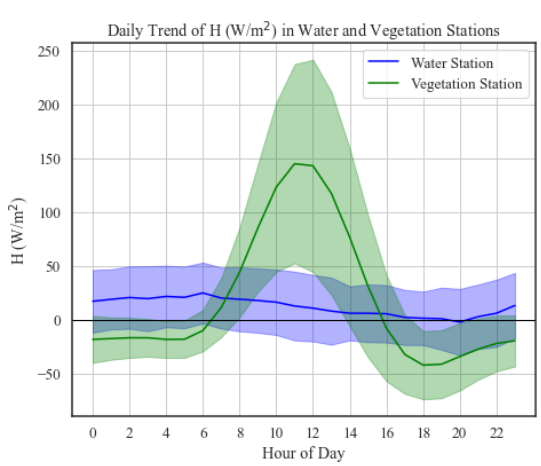
LE and **ET** represent the latent heat flux and evapotranspiration, respectively.

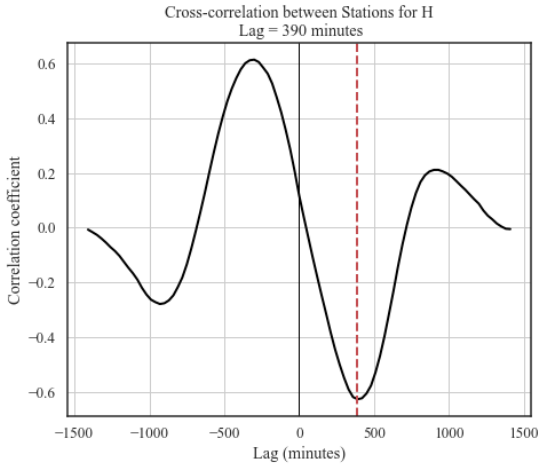
As we can observe, the daily trends of these parameters exhibit a strong correlation, with a time lag of approximately 2.5 hours. This lag indicates that the **latent heat flux** and **evapotranspiration** values at the **water station** are slightly delayed compared to the **vegetation station**. This can be explained by the following factors:

1. **Water heat capacity**: Water has a higher heat capacity than vegetation, meaning it takes more energy to change the temperature of water compared to vegetation and evaporate water from the surface.
2. **Stomatal sensitivity**: The stomata of the vegetation are more sensitive to environmental conditions, such as radiation, leading to a faster response in evapotranspiration.

Interestingly, the **amplitude** of the latent heat flux and evapotranspiration values is **equivalent between the two stations**, suggesting that the vegetation has **no water stress** during the observed period. This makes sense, as the data is from an island in the middle of the lake, where water availability is likely not a limiting factor.

H Daily Trend





Sensible heat flux (H) represents the transfer of heat from the surface to the air layer. Similar to the previous observations, there is a **time lag** between the **water station** and the **vegetation station**, although in this case, the **difference in amplitude** appears to be more significant.

While the **daily average** of *H* is relatively similar between the two stations (17 and 11 W/m² for vegetation and water station, respectively), the **standard deviation** at the **vegetation station** is **~3 times higher**.

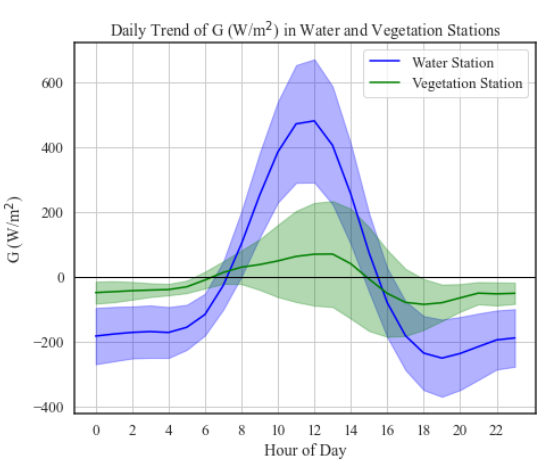
This trend is clearly observed in the graph:

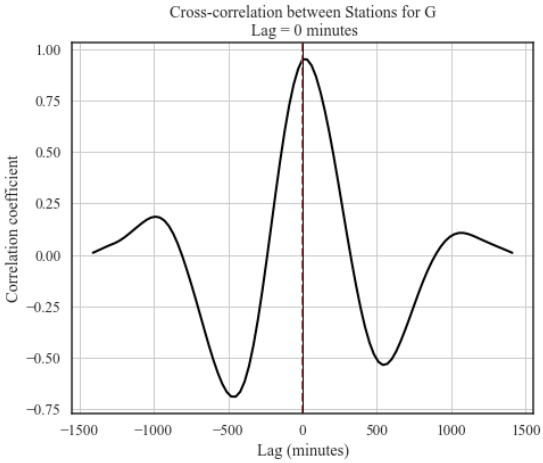
- The *H* values at the **water station** remain **relatively constant**.
- In contrast, the **vegetation station** exhibits a **clear daily trend**, with an **increase during the day** and a **decrease during the night**, even reaching **negative values** (indicating a negative temperature difference between the air and the ground).

These phenomena can be explained by the physical differences between the two environments:

1. **Water heat capacity**: Water has a higher heat capacity, which leads to a more **gradual and muted response** to changes in environmental conditions.
2. **Vegetation heat capacity**: Vegetation has a lower heat capacity, allowing it to **respond more quickly** to climatic and environmental changes.

G (Soil / Water Heat Flux) Daily Trend





- The **water station** exhibits a relatively **stable and muted daily trend** of *G*, with **smaller fluctuations**.
- The **vegetation station** shows a **more pronounced daily trend**, with a **clear increase in *G* during the day** and a **decrease at night**, even **reaching negative values**.

Amplitude Difference:

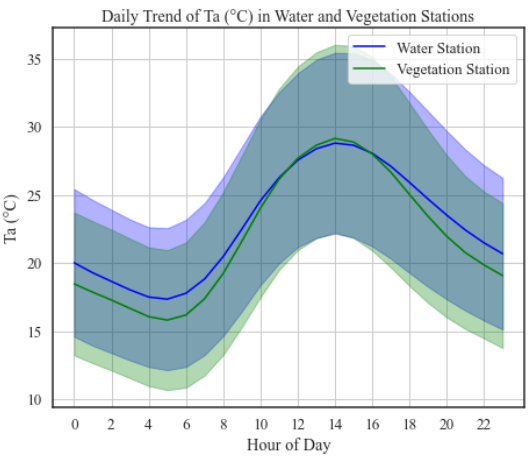
- The **amplitude** (range) of *G* is **larger at the vegetation station** compared to the water station.

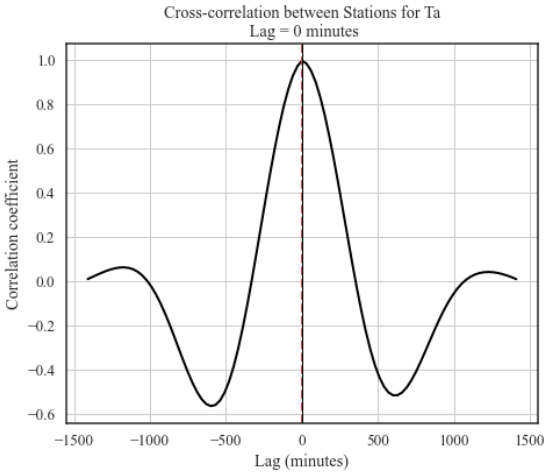
Time Lag:

- There is a **time lag of half an hour** between the peak values of *G* at the water station and the vegetation station.

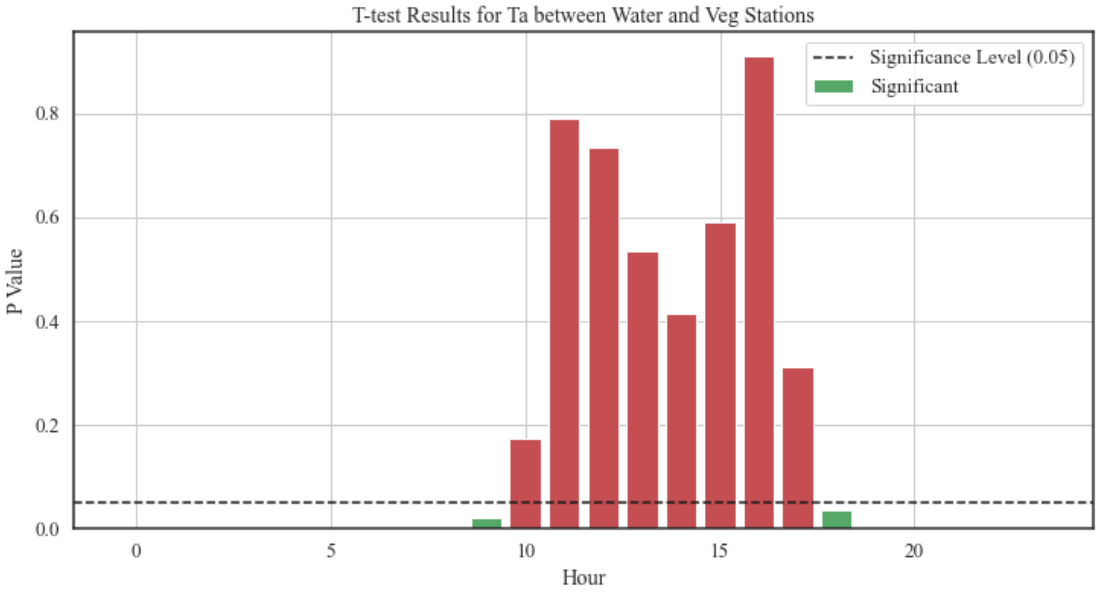
Here, also the time lag is observed, with the vegetation station experiencing changes in *G* slightly earlier than the water station. This can be attributed to the faster response of vegetation to environmental conditions, leading to a more pronounced daily cycle in *G*.

Temperature and RH Daily Trend

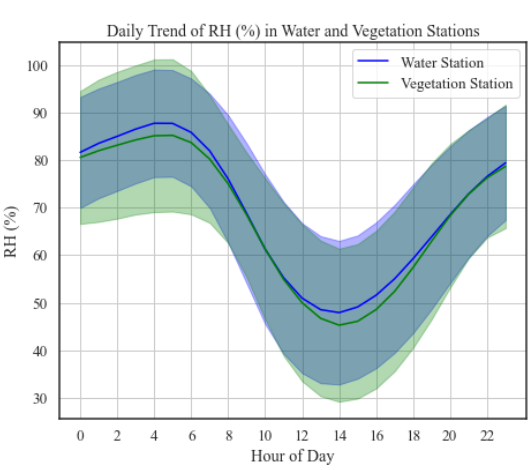


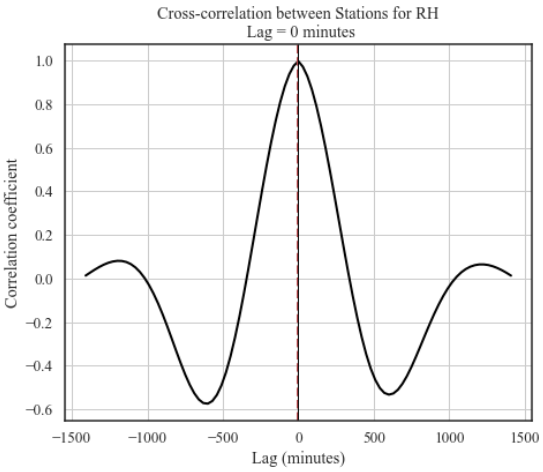


T-test of the mean temperature of each hour between the water station and vegetation station:



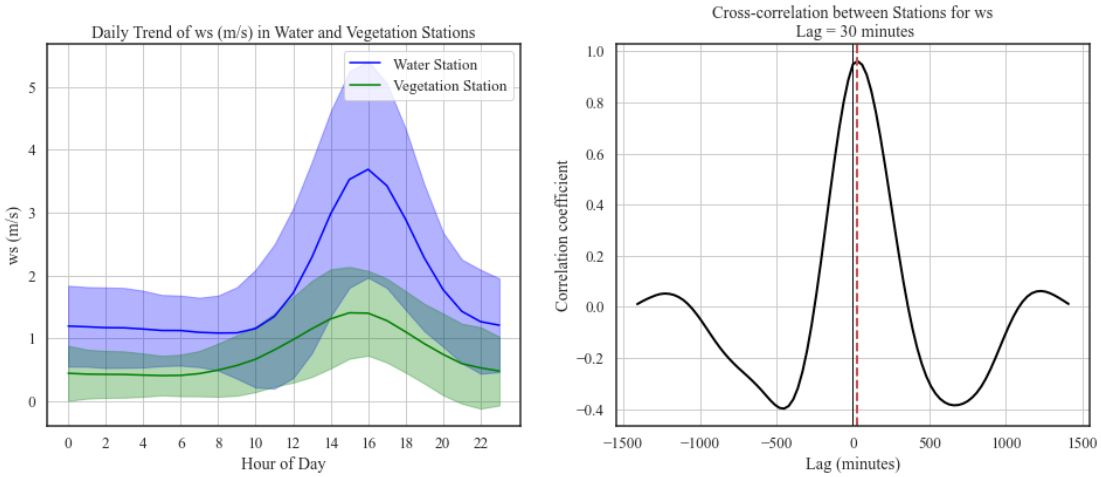
- The **2 stations exhibit a similar daily trend in temperature** , with a peak around midday and a minimum during the night.
- There is no time lag between the 2 stations, as the cross-correlation plot shows a strong positive correlation at lag 0.
- From the t-test, we can see that **the difference in temperature between the 2 stations is statistically significant (p-value < 0.05) during the night until morning (10:00 AM)**, with the vegetation station having lower temperatures compared to the water station. This difference can be attributed to the heat capacity of water, which leads to more stable temperatures compared to the vegetation station.





As expected, the RH exhibits an inverse relationship with temperature, with higher RH values observed during the night and early morning when temperatures are lower. The daily trend of RH is relatively consistent between the two stations, with both showing a peak in RH during the night and a decrease during the day. The time lag between the two stations is 0, indicating that the RH values are synchronized.

Wind Speed (w_s) Daily Trend



1. The **vegetation station** exhibits a relatively **smooth and muted daily trend** of w_s, with **smaller fluctuations**.
2. The **water station** shows a **more pronounced daily trend**, with a **clear increase in** w_s **during the day** and a **decrease at night**.
3. **Amplitude Difference:**

   The **amplitude** (range) of w_s is **larger at the water station** compared to the vegetation station.
4. **Time Lag:**

   There is a **time lag of approximately 30 minutes** between the peak values of w_s at the water station and the vegetation station.

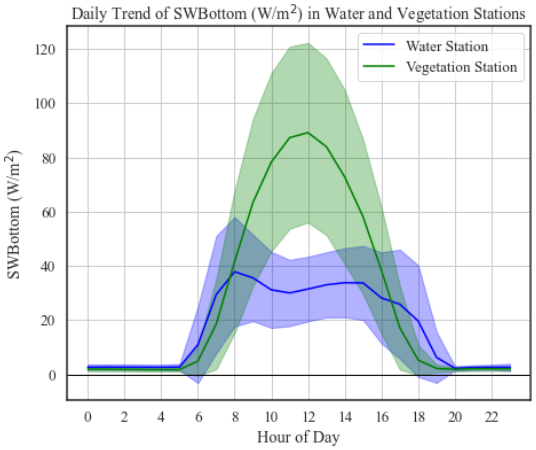
The differences in the behavior of w_s between the water and vegetation stations can be explained by:

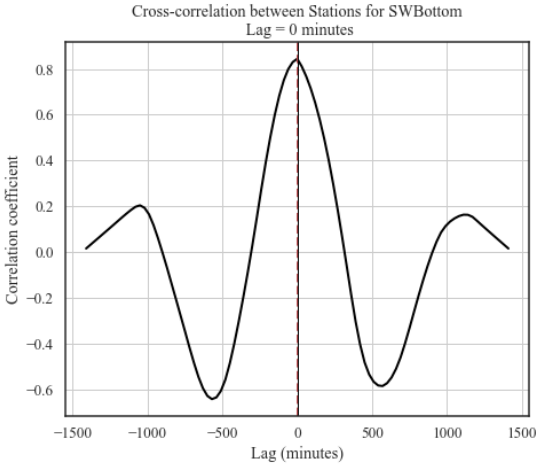
Surface roughness: The **vegetation station** has a **higher surface roughness** compared to the **water station**, which can lead to more turbulence and fluctuations in wind speed.

Wind sheltering: The presence of vegetation can **shield the wind**, leading to lower wind speeds compared to the open water surface.

SWBottom Daily Trend

SWBottom is the shortwave radiation reflected from the surface towards the sky.





1. **Daily Trend of *SWBottom*:**

   The **vegetation station** exhibits a **more pronounced daily trend** in *SWBottom*, with **higher amplitude** compared to the **water station**.

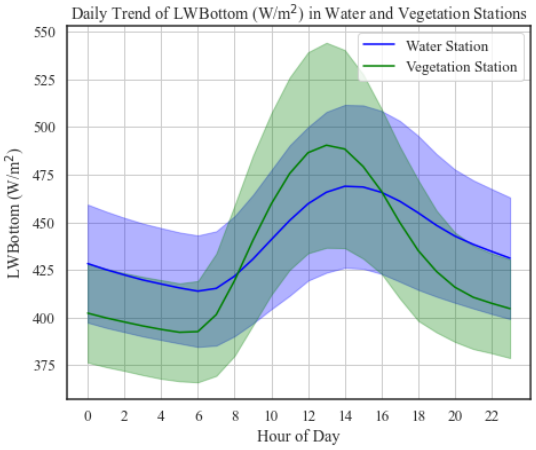
   The **water station** shows an increase in *SWBottom* during the morning, constant values during the day, and a decrease in the evening.

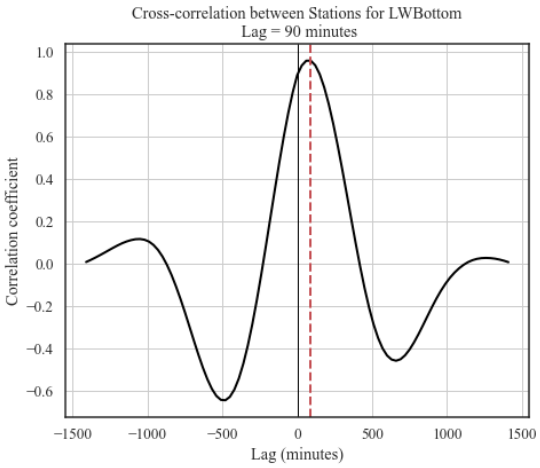
2. **Amplitude Difference:**

   The **amplitude** of *SWBottom* is **higher at the vegetation station** compared to the water station.

The differences in the behavior of *SWBottom* between the water and vegetation stations can be explained by the contrasting physical properties of the two environments. The water absorbs more solar radiation compared to vegetation, leading to a more stable and consistent trend in *SWBottom* at the water station and less reflection towards the sky.

LWBottom Daily Trend





1. **Daily Trend of LWBottom (longwave radiation from the bottom up):**

The **water station** exhibits a relatively **smooth and muted daily trend** of LWBottom, with **smaller fluctuations**.

The **vegetation station** shows a **more pronounced daily trend**, with a **clear increase in LWBottom during the day** and a **decrease at night**.

2. **Amplitude Difference:**

   The **amplitude** (range) of LWBottom is **larger at the vegetation station** compared to the water station.

3. **Time Lag:**

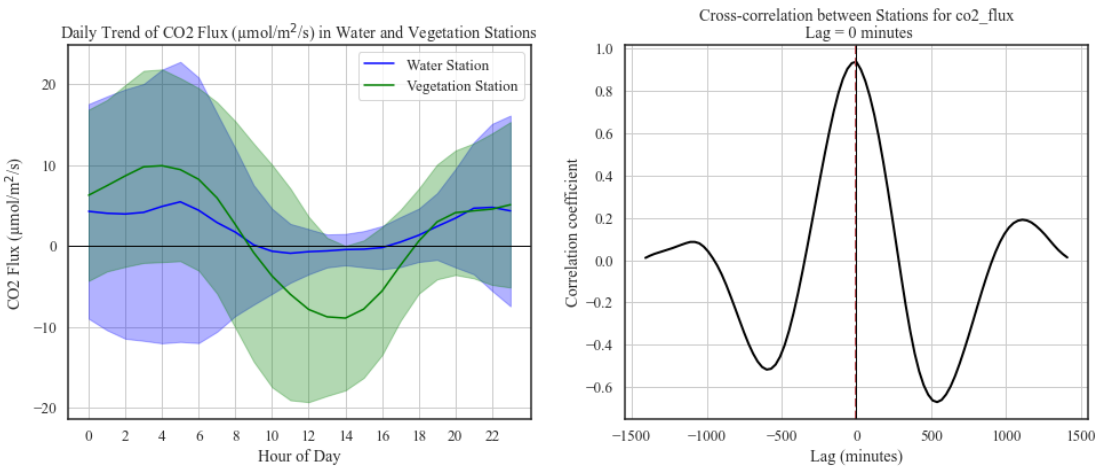
   There is a **time lag of approximately 90 minutes** between the peak values of LWBottom at the water station and the vegetation station.

The differences in the behavior of LWBottom between the water and vegetation stations can be explained by the following factors:

Thermal properties: The **water station** exhibits a more stable trend in LWBottom due to the higher heat capacity of water, which leads to slower changes in longwave radiation.

Vegetation response: The **vegetation station** shows a more dynamic trend in LWBottom, reflecting the rapid changes in longwave radiation associated with vegetation properties and environmental conditions.

CO_2 Flux Daily Trend



The daily trend graphs reveal some interesting insights into the carbon flux dynamics between the water and vegetation stations:

Vegetation Station:

The daily trend clearly shows a pattern of **carbon fixation during the day** (negative values) due to photosynthesis, and **carbon emission at night** (positive values) due to respiration.

However, the **overall carbon sequestration should be significantly higher than the emission**, given the assumption that wetlands are important carbon sinks.

The observed trend requires further research to understand the underlying factors. **Dividing the data into specific periods and examining it in more detail** could provide additional insights.

Evaluating the **microbial content of the soil** can also help in understanding the carbon dynamics at the vegetation station, as the soil microbiome can significantly influence the carbon dynamics.

Water Station:

The situation is more complex for the water station, as it involves a mix of carbon sources and sinks, such as animals, secretions, plants, and algae.

The **more muted daily trend** observed at the water station suggests a **more balanced carbon budget**, with both fixation and emission processes occurring.

Determining the **net carbon flux** and understanding the **relative contributions of the different components** (e.g., aquatic biota, water-air exchange) would require a more in-depth analysis.

CO2 flux is central to my research focus, and a **more comprehensive investigation would be valuable**. **Dividing the data into specific periods**, examining **seasonal variations**, and incorporating **additional factors** (e.g., soil microbiome, aquatic biota) could provide a more nuanced understanding of the carbon dynamics at the two stations.

Exploring the **underlying mechanisms and drivers** of the observed trends, as well as **quantifying the net carbon sequestration or emission**, would be important next steps (but beyond the scope of this analysis).

3. Long-Term Analysis

1. **Read the NDVI data of Sentinel-2 for both station**s:

The data contains the NDVI values for the water station and the vegetation station over the period from 2017 to 2024. Every point in each time series represents the average NDVI value over a 5-day period (temporal resolution) and the mean NDVI of a polygon of the area of the station (spatial resolution).

The data downloaded from the Sentinel-2 satellite, which provides relatively high-resolution optical images of the Earth's surface.

NDVI (normalized difference vegetation index) is a widely-used metric for quantifying the health and density of vegetation using sensor data. It is calculated from spectrometric data at two specific bands: red and near-infrared by the following formula:

$$NDVI = \frac{NIR + RED}{NIR - RED}$$

2. **Explore the distribution of NDVI values for both stations.**

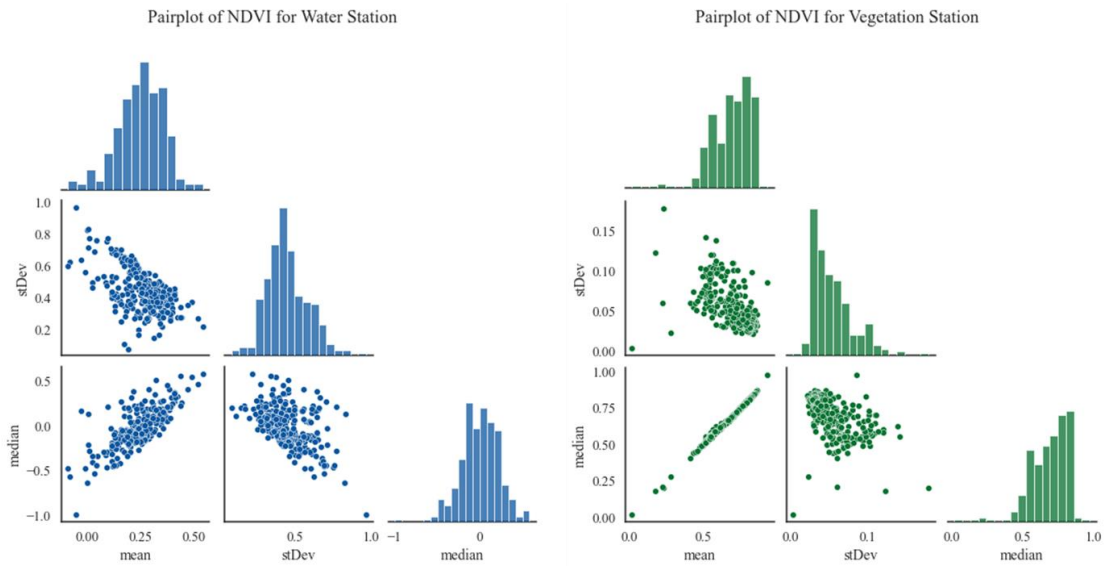


Figure 3. Pair plots of the mean, std, and median of NDVI ROI values of water station (blue) and vegetation station (green)

- As expected, the **NDVI values** for the **vegetation station** are **higher** on average compared to the **water station**, indicating **denser and vegetation** on the island.
- The NDVI values of water bodies are typically between -1 to 0. Here, the **water station** shows **NDVI values close to 0** (even positive values), which could be due to the presence of **algae, aquatic plants, or floating vegetation**.
- The correlation of mean with median is **higher for the vegetation station** (1 and 0.6 respectively), indicating a **more homogeneous distribution** of NDVI values in the vegetation area compared to the water area. This can be also shown by the distribution of standard deviation, which is higher for the water station.
- The distribution of the median NDVI values of the water station is much wider than the distribution of the mean NDVI values, indicating that the NDVI values are **skewed**. This could be due to the presence of **outliers** or **extreme values**.

3. **Smoothing the NDVI data** using rolling mean with a window size of 7.
4. **Resample the NDVI to weekly** and interpolate using linear interpolation for missing values.

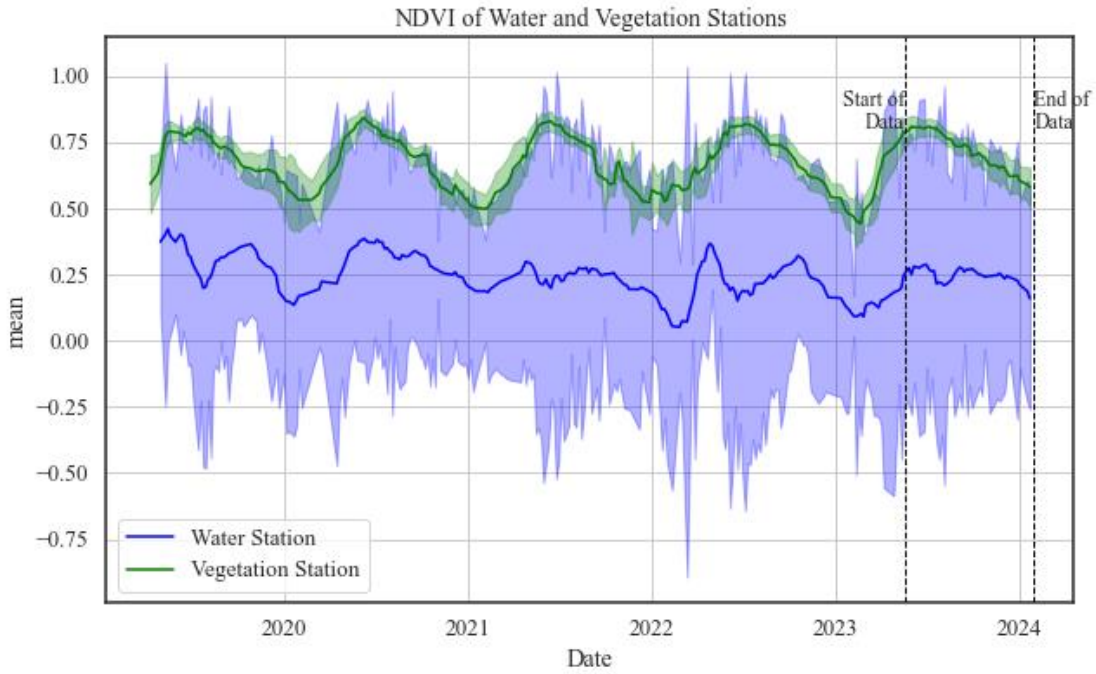
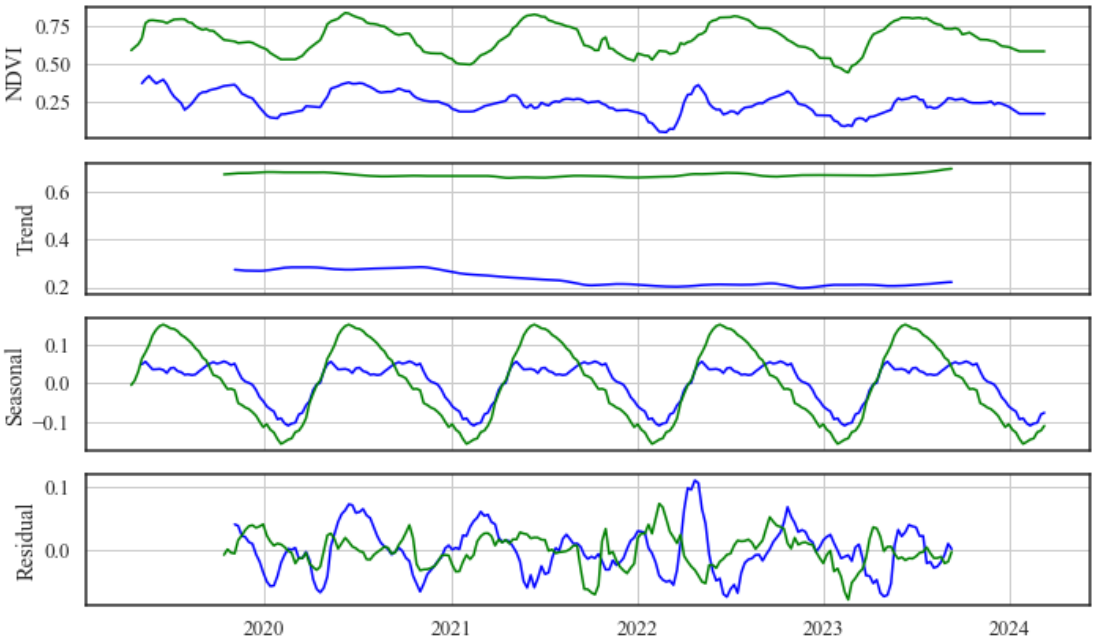


Figure 4. NDVI mean±std time series after smoothing and resampling the data. The black dashed lines represent the time of the measurements so far.

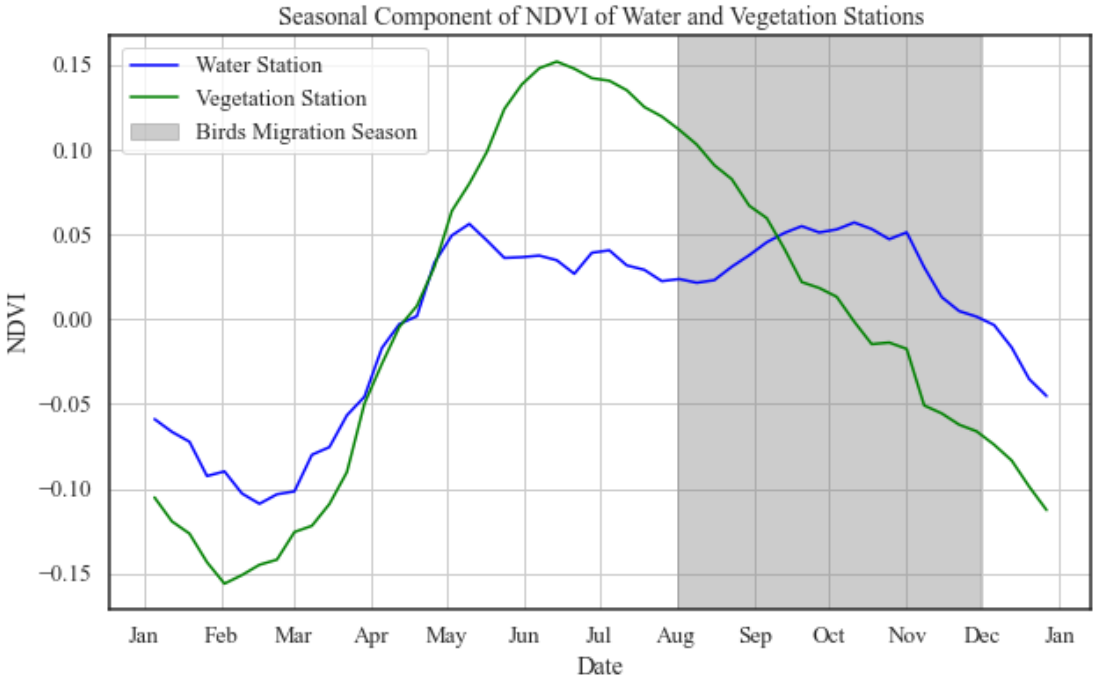
5. **Decompose** the NDVI time series into trend, seasonal, and residual components in order to observe the yearly and long-term dynamics of vegetation at the two stations.



From the decomposition of the NDVI time series, we can observe the following trends:

- There is no significant trend in the NDVI values for the vegetation station, indicating **relatively stable vegetation dynamics** over the years. Water station presents a slight decrease in the trend between 2020 to 2021.

- The seasonal component shows a clear **annual cycle** for vegetation, with **higher NDVI values** during the **spring and summer months** and **lower values** during the **fall and winter months**. This pattern is consistent with the **seasonal growth and dormancy** of vegetation.
- The seasonal component of the water station is interesting, it seems that there are 2 peaks in the NDVI values, one in the spring before the peak of vegetation, and one between August to December. Let's investigate this further by plotting the seasonal component of the NDVI values for the water station of one year in addition to the Bird migration season, which occurs in the fall between this months:



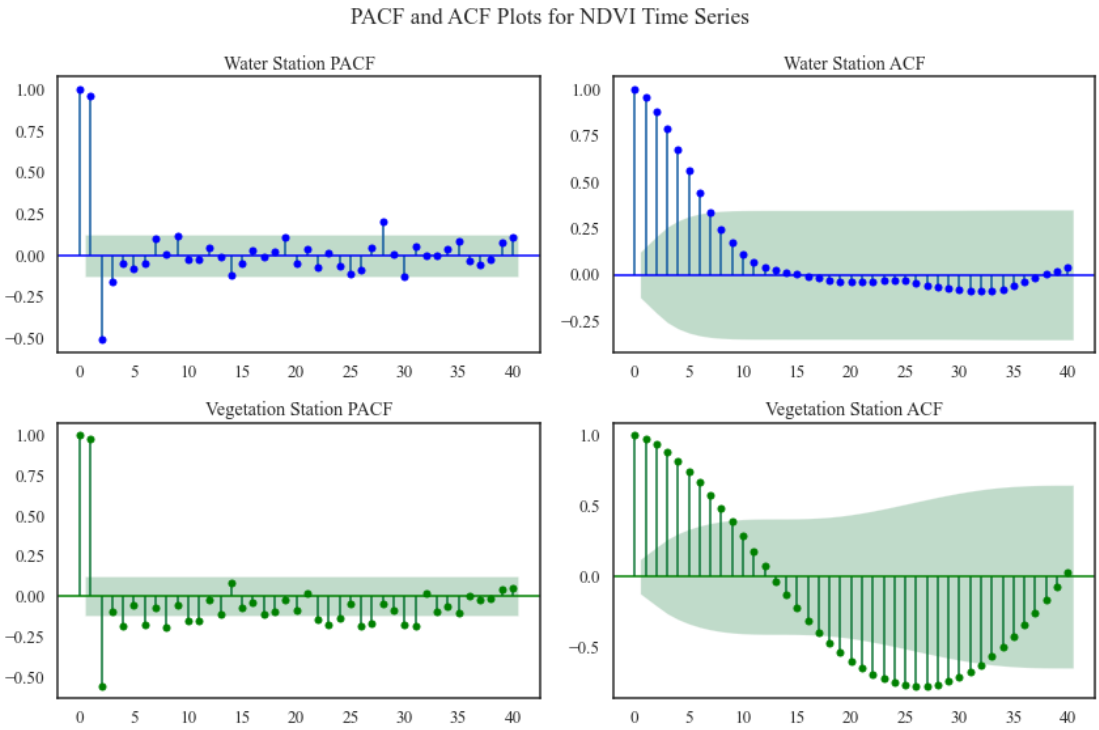
The observed data suggests that the **second peak in NDVI coincides with the bird migration season**. This observation leads to the following hypothesis:
If it were not for the bird migration, we would expect a **decrease in NDVI values** after the initial increase in the spring. However, the persistence of the peak at the vegetation station compared to the water station may be explained by the **specific vegetation type** in the area.
If this hypothesis is correct, it implies that the **presence of migratory birds increases the amount of vegetation in the lake during the migration period**. This could potentially be due to the **enrichment of the water with nutrients** from the birds' activities, such as excrement or foraging.
Further research is needed to:

1. **Confirm the correlation between the NDVI peak and the bird migration season**.
2. **Investigate the mechanisms by which the birds' presence leads to increased vegetation**, such as nutrient input or other ecological interactions.
3. **Explore the differences in vegetation type and phenology between the water and vegetation stations** that may contribute to the observed patterns.

6. **Stationarity Test**: Perform the Augmented Dickey-Fuller test to check the stationarity of the NDVI time series for both stations:

   The results of the Augmented Dickey-Fuller test indicate that the NDVI time series for both the water and vegetation stations are **stationary**. This suggests that the time series data does not exhibit a trend or systematic pattern over time, as we see in the decomposition analysis.

7. **PACF and ACF**: Plot the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) of the NDVI time series for both stations:



1. Water Station:

- **PACF:** The PACF plot for the water station shows significant spikes at lags 1, 2, and 3, indicating the potential presence of an autoregressive (AR) component in the time series (order 2).
- **ACF:** The ACF plot for the water station exhibits a gradual decay, suggesting the possible presence of a moving average (MA) component in the time series.

2. Vegetation Station:

- **PACF:** The PACF plot for the vegetation station shows a more complex pattern, with multiple significant spikes at various lags. This suggests the potential need for an order 2.

- **ACF:** The ACF plot for the vegetation station also exhibits a gradual decay, similar to the water station, implying the presence of a moving average (MA) component.

8. **ARIMA Model**: Fit an ARIMA model to the NDVI time series for both stations and make predictions for the last 2 years.

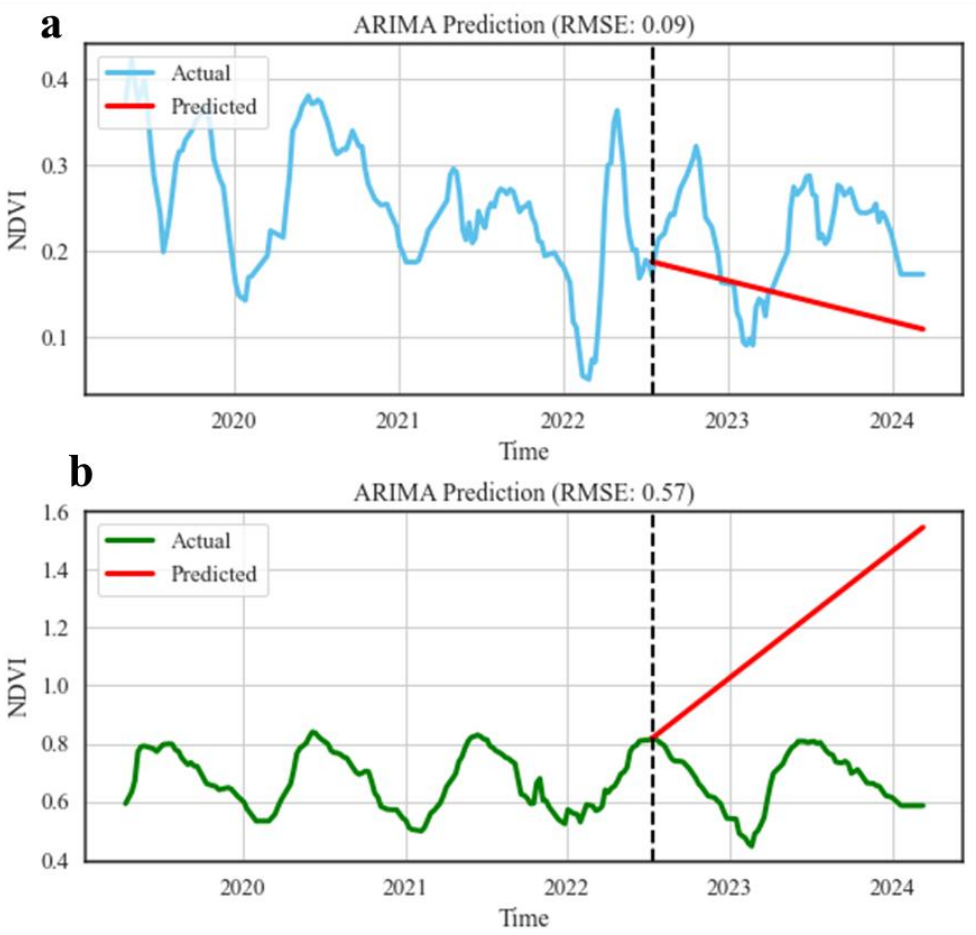
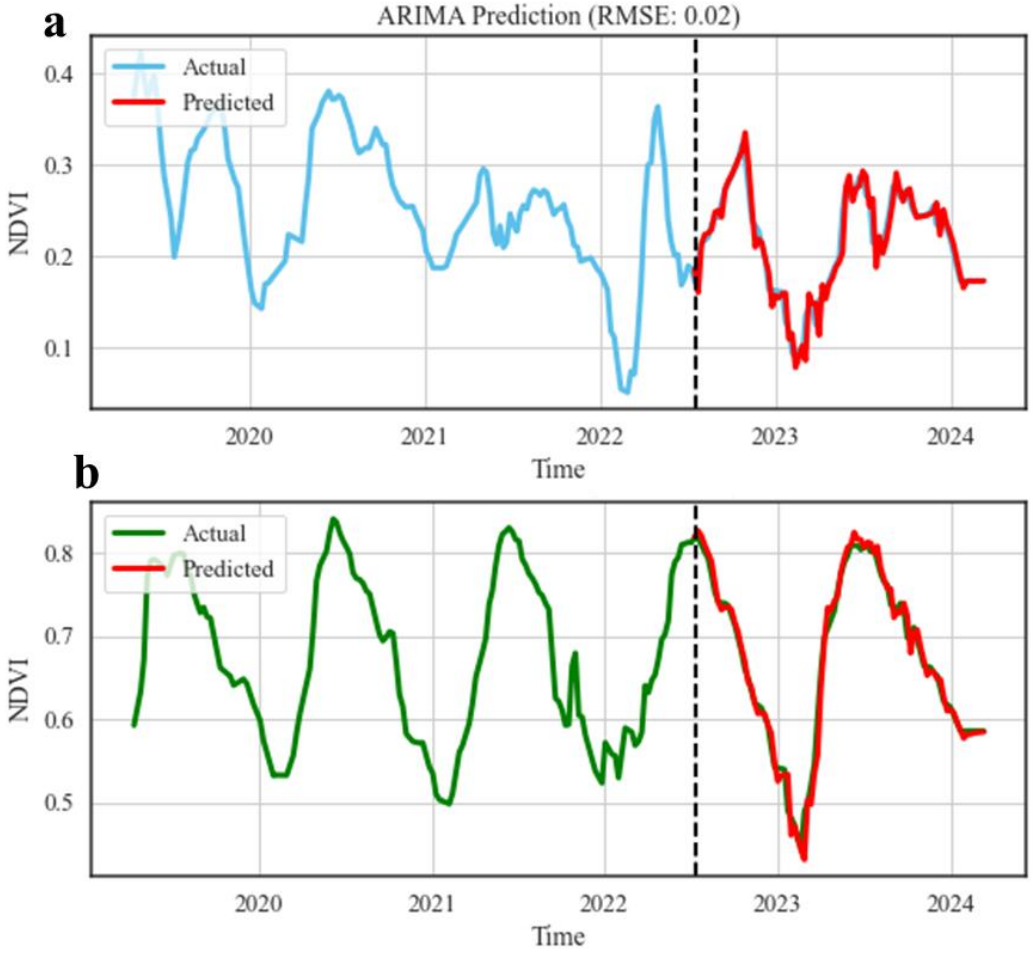


Figure 5. ARIMA Model Forecast for the NDVI Time Series of the Water Station (blue) and Vegetation Station (green). The red line is the predicted values for the last 2 years.

As it can be seen from the plots, the ARIMA model fails to capture the complex dynamics of the NDVI time series for both stations. The predictions are relatively flat and do not reflect the observed patterns in the data. Let's try to predict every sample in different step. This of course is not the best way to predict the NDVI values, but it can let the model predict the values in a more accurate way:



Now the results are much better. Of course, predict on value is much easier than predict on 300 values, but the model can predict the values in a more accurate way.