Final Project - Time Series Analysis

Omer Eyal
315044636

February 19, 2025

1 Introduction

The Tapajós National Forest, located in the state of Pará, Brazil, spans an area of 549,066.87 hectares. It is bordered by the Tapajós River, the Cupari River, and the BR-163 Santarém–Cuiabá road. The region receives an average annual rainfall of approximately 1,800 mm, with temperatures ranging from 21 to 31 °C and a mean temperature of 25.8 °C. The forest is classified as an Evergreen Broadleaf Forest with a closed canopy at 40 m and below, and it is considered primary or old-growth.

The BR-Sa1 measurement site (Latitude: -2.8567, Longitude: -54.9589) within the forest is part of the Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA), which aims to improve understanding of the regional carbon balance. The dataset is obtained from the FLUXNET network, which integrates regional networks of Earth system scientists. Carbon, water, and energy fluxes between the biosphere and atmosphere are measured using the tower-based Eddy Covariance technique [1],[2].

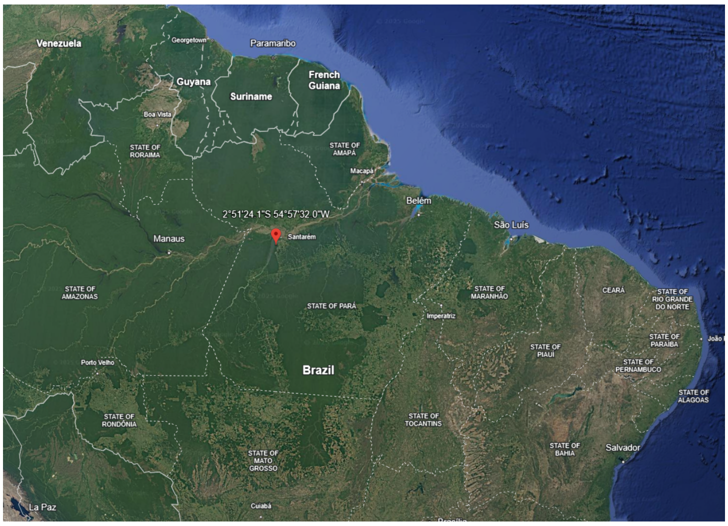


Figure 1: The BR-Sa1 site location

Among the many variables measured at the site, this project focuses primarily on hourly temperature, vapor pressure deficit (VPD), and net ecosystem exchange (NEE) of carbon dioxide. The primary objectives of this analysis are to examine temporal trends and periodic patterns in the data and to investigate potential correlations between these variables. Additionally, since the dataset includes flagged gap-filled values provided by FLUXNET, I plan to apply alternative gap-filling techniques and compare their performance.

2 Exploratory Analysis

2.1 Outliers identification

All variables in this project underwent an outlier identification process using the interquartile range (IQR) sliding window method. Outliers were filtered using a window size of 50, which corresponds to 50 hours when the sample interval is one hour. A threshold (k) of 1.5 times the IQR was applied from the first and third quartiles to identify and remove outliers. Figure 2 illustrates an example for temperature, comparing the original time series to the filtered one.

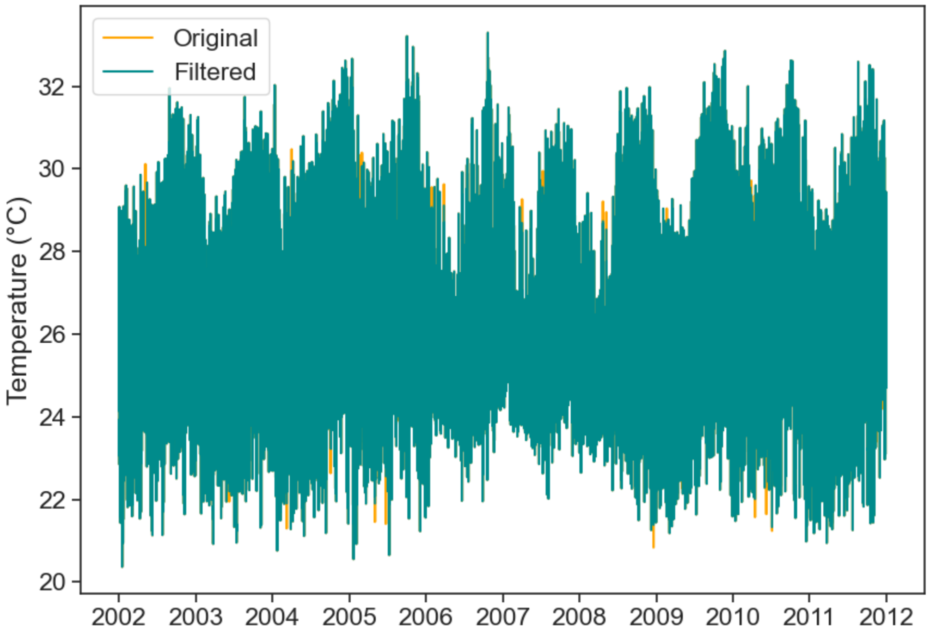
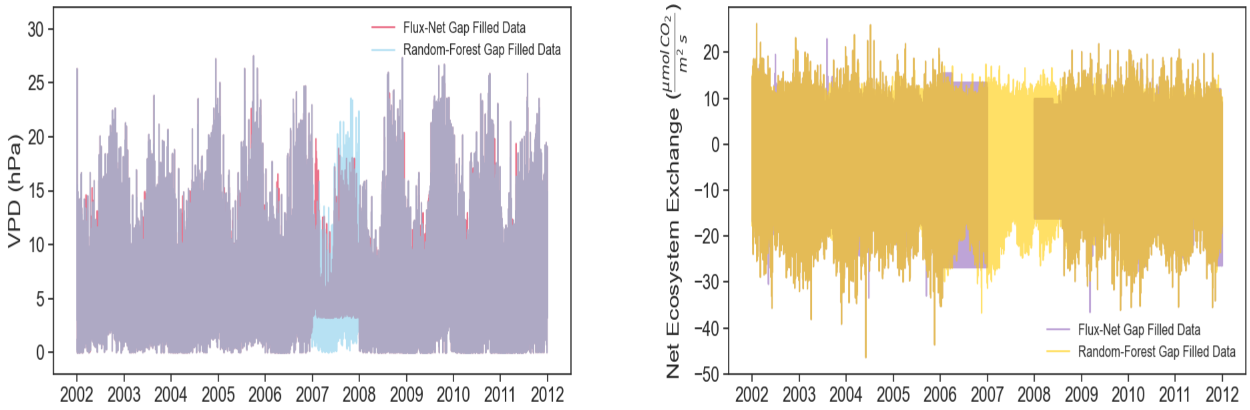


Figure 2: Example of outliers identification for the temperature series

2.2 Gap-filling methods comparison

After identifying outliers, I compared the FLUXNET gap-filling method with the Random Forest model for VPD and NEE. First, I filtered the raw data to include only actual measurements using the flag columns. Then, I applied the Random Forest method for gap filling and compared the results with the FLUXNET gap-filled data.

For temperature, I used the FLUXNET gap-filled data alongside the Random Forest gap-filled data, after removing outliers. Figure 3 compares both methods, showing that the Random Forest method more closely resembles the actual data than the FLUXNET method. Therefore, for further analyses, I used the Random Forest gap-filled data.



(a)

(b)

Figure 3: Comparison of the Random Forest and FLUXNET gap-filling methods for (a) VPD and (b) NEE.

2.3 Visualize the processed data

Figure 4 presents the processed data after outlier identification and gap-filling for the selected variables, including precipitation. While precipitation is not included in further analyses, it is shown here to provide additional context about the tropical climate. The figure already reveals the annual cycles of temperature and VPD, as well as continuous rainfall throughout the year, with distinct peak events observed during the measured time range.

To observe long-term trends, I resampled the data to compute the monthly averages for the selected variables, as shown in Figure 5. Overall, the data exhibit a relatively stable behavior over the years.

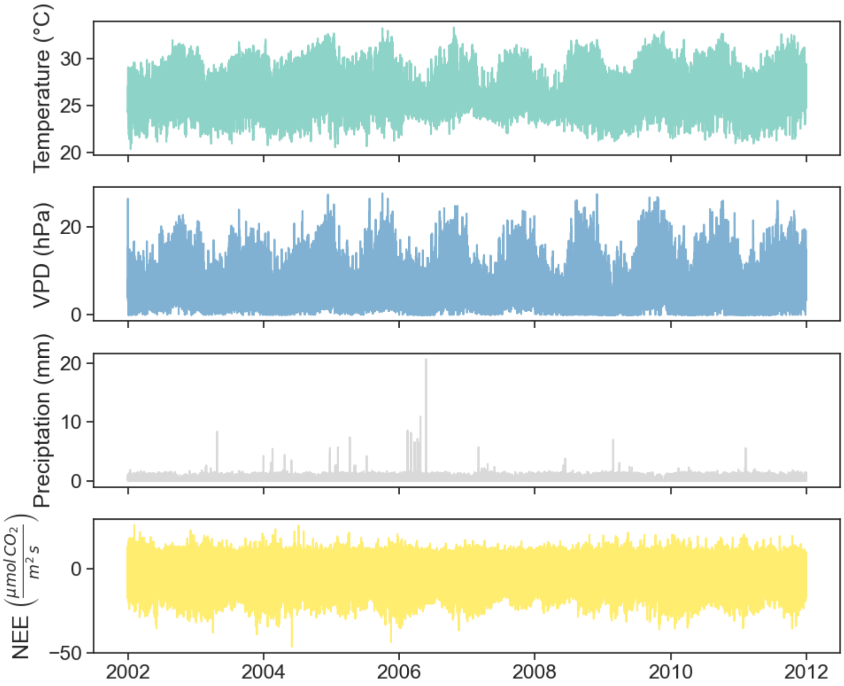


Figure 4: Processed data for the selected variables after outlier removal and gap-filling.

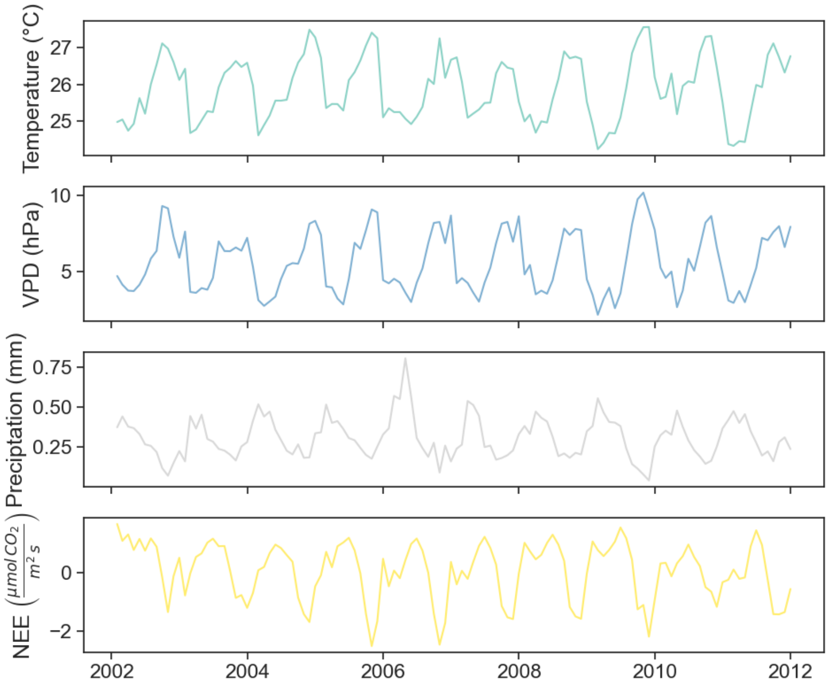


Figure 5: Monthly average data of the chosen variables

3 Detailed Analysis

3.1 Seasonal Decomposition

Each variable underwent seasonal decomposition into three components: the yearly trend $T(t)$, the seasonal component $S(t)$, and the residual component $e(t)$, assuming that these components sum to reconstruct the original observed time series $Y(t)$:

$$Y(t) = T(t) + S(t) + e(t) \tag{1}$$

Figure 6 shows the seasonal decomposition of the temperature time series. The seasonal component exhibits minor fluctuations, mostly ranging between -1°C and 1°C, indicating the absence of a strong seasonal pattern. Additionally, the trend component does not display a clear directional trend, fluctuating within a narrow range. A similar pattern is observed in the other variables.

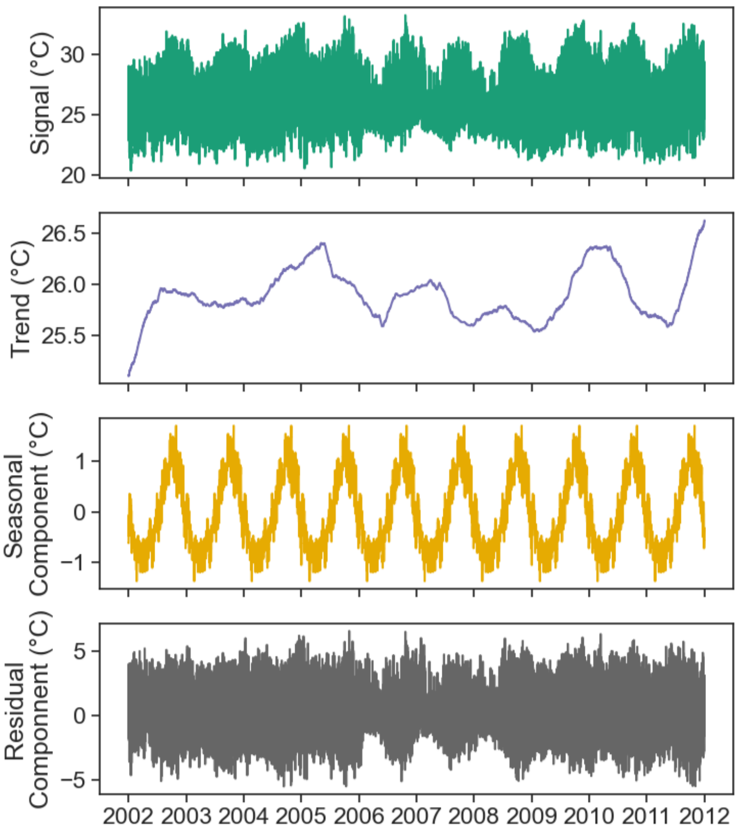


Figure 6: Seasonal decomposition of the temperature series

3.2 Frequencies

For analyzing the time series cycles, I applied the Fourier transform $F(k)$ to identify the dominant frequencies in the series $f(t)$:

$$F(k) = \int_{-\infty}^{\infty} f(t)e^{-2\pi ikt}dt \tag{2}$$

where k represents the frequencies. In practice, the Fourier transform is computed using the discrete Fourier transform (DFT), where the integral is replaced by a summation over a finite number of points (corresponding to the length of the dataset).

Figure 7 shows the absolute value of the power spectrum of the Fourier transform for NEE. The strongest frequencies correspond to yearly, daily, and half-daily cycles. A similar pattern is observed for VPD and temperature, suggesting that these variables may follow annual and diurnal cycles.

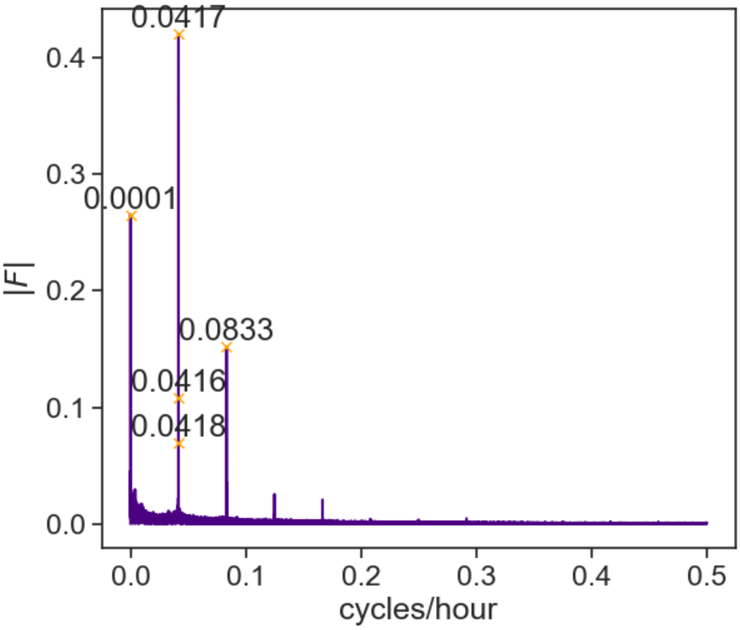


Figure 7: The absolute value of the Fourier transform power spectrum for the NEE series

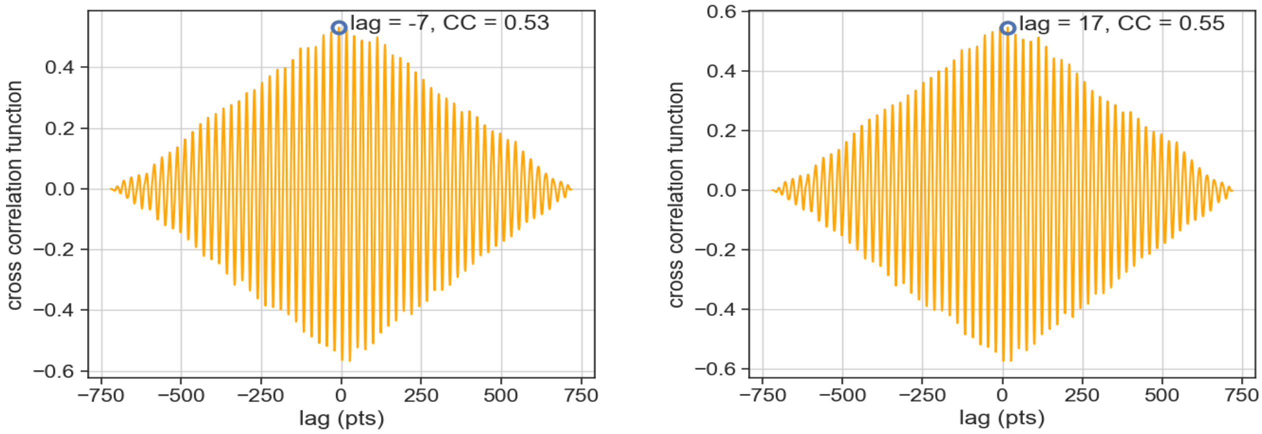
3.3 Cross-Correlation

By applying the cross-correlation function, I examined the relationship between VPD and temperature with NEE. Given the Fourier transform results, which suggest that all variables exhibit a strong daily cycle, they share the same cycle length.

Additionally, I analyzed the correlation between the residual components of the variables to isolate the relationship while removing the influence of trends and seasonal effects.

Figure 8 presents the cross-correlation function at different lags between temperature and NEE for both the original time series and their residual components. The maximum cross-correlation values for both cases are similar (0.53 and 0.55), occurring at lags of -7 and +17. Since the data follows an hourly resolution, these lags correspond to the same point within a 24-hour cycle (e.g., if the reference hour is 00:00, then a lag of -7 corresponds to 17:00 on the previous day, while +17 corresponds to 17:00 on the same day).

A similar pattern is observed for VPD and NEE, with a maximum cross-correlation of 0.50 and 0.51 at a lag of -7, further reinforcing the presence of daily periodicity.



(a)

(b)

Figure 8: Comparison of the cross-correlation function at different lags between temperature and NEE: (a) for the original time series and (b) for their residual components.

3.4 Stationarity

The Augmented Dickey–Fuller (ADF) test was used to assess whether the NEE signal is stationary. To perform the test, I extracted a one-month segment from the NEE time series (January 2003) and applied the ADF test. The results indicate that the series is **stationary**, with a p-value of 8.85×10^{-17}.

4 Conclusions and Pitfalls

Given the tropical location (Latitude: -2.8567), it is unsurprising that the seasonal component does not play a major role in the data. Even when analyzing the monthly averages, the fluctuations appear minor, with no noticeable trend. The seasonal decomposition further confirms this, showing a relatively small seasonal component and a subtle trend. However, a deeper investigation into the trend could still provide valuable insights.

When examining the periodic patterns of the variables, a strong daily cycle is evident in temperature, VPD, and NEE. While this is expected for temperature and VPD, it is particularly interesting for NEE. This daily cycle makes sense, as NEE is influenced by photosynthesis, which follow a diurnal pattern.

Regarding the correlation between variables, the similarity in correlation values between temperature-NEE and VPD-NEE is expected, given that temperature and VPD are strongly correlated (correlation of 0.94 with a lag of 0). The cross-correlation lag of -7 or +17, as previously explained, essentially represents the same point in a daily cycle. This lag can be interpreted as follows: while temperature and VPD peak around midday, NEE peaks at night, when photosynthesis is minimal and respiration dominates, leading to a positive NEE. I examined whether cross-correlation between the residual components would yield higher correlations by removing seasonal and trend effects. However, the improvement was only slight, which makes sense given the minimal influence of the seasonal component and the trend previously discussed.

As for the comparison between gap-filling methods, the Random Forest approach appears to perform better than the FLUXNET method, assuming there were no major changes during the gap periods. Since FLUXNET is a large network, site-specific optimizations may not always be ensured, making the Random Forest method a more reliable choice for this dataset.

This dataset is very large, containing numerous variables, making it challenging to select the most relevant ones. Eventually, I decided to focus on temperature and VPD, considering that these are the primary driving forces and among the most frequently measured variables influencing NEE.

A possible next step could involve incorporating data from multiple sites and adding additional variables, such as radiation, respiration, and photosynthesis, to gain deeper insights into the system's dynamics.

5 Code Availability

The complete Jupyter Notebook, containing the full code and additional analyses, is available here.

6 References

1 Tóta, Julio, et al. ”Amazon rain forest subcanopy flow and the carbon budget: Santarém LBA-ECO site.” Journal of Geophysical Research: Biogeosciences 113.G1
https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2007JG000597

2 FLUXNET - https://fluxnet.org/about/