

CHE507 Assignment 01

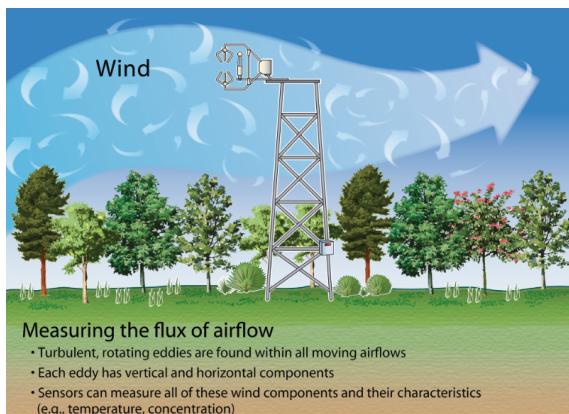
Aiqi Zhang, Engineering Science (Engineering Physics) 2T6, 1009253521

October 13, 2025

1 Introduction

The dataset for analysis is the measurement results for an eddy-covariance flux tower. An eddy covariance flux tower is an instrumented platform used to measure the exchange of gases, energy, and momentum between the land surface and the atmosphere. It typically consists of a 3D sonic anemometer that records wind speed and direction at high frequency, along with gas analyzers that continuously measure concentrations of CO_2 , H_2O , and CH_4 [1]. By capturing rapid, simultaneous fluctuations in vertical wind velocity and gas concentration, the system computes fluxes such as net ecosystem exchange of carbon. Data are logged at high temporal resolution (e.g., 10–20 Hz) and later processed to produce half-hourly or hourly averages. Complementary sensors often record meteorological and radiation variables, providing context for interpreting ecosystem-atmosphere interactions. A sketch of the tower can be found in Figure 1a

The dataset used in this analysis is from the Turkey Point (CA-TPD) site, covering the period from 2012-01-01 to 2017-12-31 [2]. CA-TPD is face into a deciduous broadleaf forest (DBF) ecosystem, characterized by woody vegetation with more than 60% canopy cover and an average height exceeding 2m, shown in Figure 1b.



(a) Sketch of eddy-covariance flux tower [3].



(b) Flux tower at CA-TPD site [2].

Figure 1: Overview of eddy-covariance setup and the CA-TPD flux tower.

The flux tower measurements can be affected by several sources of noise and disturbance. Small errors may come from sensor drift, calibration issues, or misalignment of instruments. Weather conditions such as rain, frost, or dust can also interfere with readings. In addition, temporary power or data interruptions may cause missing values or unstable signals. To reduce these effects, the data are carefully filtered and quality-controlled during post-processing to ensure reliable results.

2 Exploratory Data Analysis

The data product is relatively complex, containing 93 columns and a total of 105,216 rows of time-series measurements recorded at 30-minute intervals. It includes meteorological variables, radiation measurements, soil state indicators, and carbon fluxes, with each variable accompanied by the corresponding uncertainty estimates or quality flags. For the purpose of exploratory data analysis, we list here only some of the most commonly used variables, which are the ones we will be used in our analysis with abbreviation in Table 1. Table 2 list out the basic statistics of all these time series, and Table 3 present a key distribution measurement for each time series.

Table 1: Variable Abbreviations and Descriptions

| Abbreviation | Description |
|--------------------|--|
| TA | Air Temperature ($^{\circ}\text{C}$) |
| VPD | Vapor Pressure Deficit (kPa) |
| PA | Air Pressure (kPa) |
| P | Precipitation (mm) |
| RH | Relative Humidity (%) |
| TS ₁ | Soil Temperature at 5 cm depth ($^{\circ}\text{C}$) |
| PPFD _{in} | Incoming Photosynthetic Photon Flux Density ($\mu\text{mol m}^{-2} \text{s}^{-1}$) |
| CO ₂ | CO ₂ Concentration (ppm) |
| GPP _{DT} | Gross Primary Productivity (daytime, $\mu\text{mol m}^{-2} \text{s}^{-1}$) |

Table 2: Basic Numeric Statistics (mean, median, min, max)

| Symbol | Unit | Mean | Median | Min / Max |
|--------------------|--------------------------------------|---------|---------|-------------------|
| TA | $^{\circ}\text{C}$ | 9.881 | 10.390 | -25.490 / 33.680 |
| VPD | kPa | 3.997 | 2.560 | 0.000 / 34.964 |
| PA | kPa | 101.246 | 101.300 | 96.300 / 103.900 |
| P | mm | 0.049 | 0.000 | 0.000 / 27.400 |
| RH | % | 74.071 | 74.860 | 10.910 / 100.000 |
| TS ₁ | $^{\circ}\text{C}$ | 9.635 | 9.720 | -3.584 / 24.805 |
| PPFD _{in} | $\mu\text{mol m}^{-2} \text{s}^{-1}$ | 282.767 | 10.047 | 0.000 / 2024.239 |
| CO ₂ | ppm | 399.520 | 403.731 | 310.844 / 501.603 |
| GPP _{DT} | $\mu\text{mol m}^{-2} \text{s}^{-1}$ | 3.486 | 0.003 | -0.000 / 42.019 |

Table 3: Distribution Summary (q05, q25, q50, q75, q95)

| Symbol | Unit | q05 | q25 | q50 | q75 | q95 |
|--------------------|--------------------------------------|---------|---------|---------|---------|----------|
| TA | $^{\circ}\text{C}$ | -8.090 | 1.821 | 10.390 | 18.910 | 24.920 |
| VPD | kPa | 0.192 | 1.095 | 2.560 | 5.551 | 12.542 |
| PA | kPa | 99.700 | 100.800 | 101.300 | 101.800 | 102.500 |
| P | mm | 0.000 | 0.000 | 0.000 | 0.000 | 0.200 |
| RH | % | 45.890 | 62.650 | 74.860 | 87.300 | 98.500 |
| TS ₁ | $^{\circ}\text{C}$ | -0.170 | 1.923 | 9.720 | 16.560 | 20.435 |
| PPFD _{in} | $\mu\text{mol m}^{-2} \text{s}^{-1}$ | 0.000 | 0.000 | 10.047 | 420.986 | 1344.440 |
| CO ₂ | ppm | 359.012 | 390.696 | 403.731 | 411.443 | 422.391 |
| GPP _{DT} | $\mu\text{mol m}^{-2} \text{s}^{-1}$ | 0.000 | 0.000 | 0.003 | 2.154 | 21.168 |

Daily averaged time series (original dataset has per half an hour data), and monthly averaged box plot over 6 years for all variables mentioned above are plotted from Figure 2 to 10.

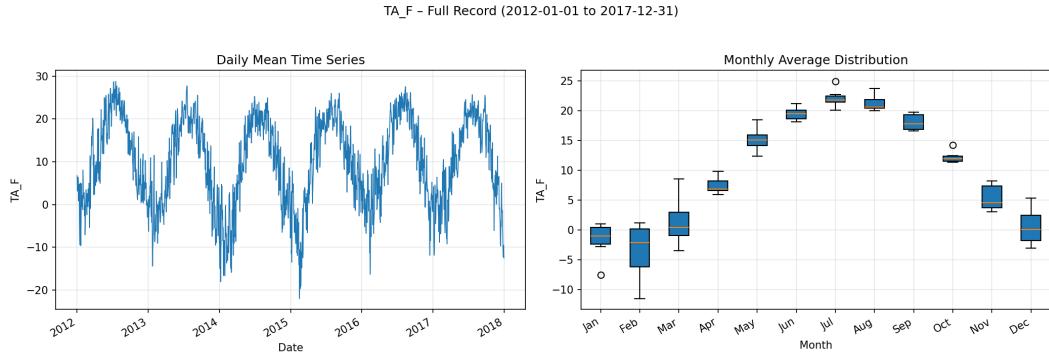


Figure 2: Air Temperature (TA) timeseries boxplot.

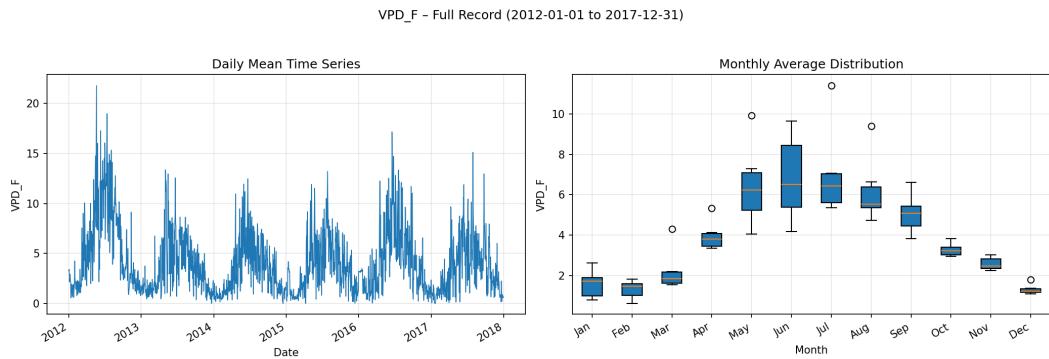


Figure 3: Vapor Pressure Deficit (VPD) timeseries boxplot.

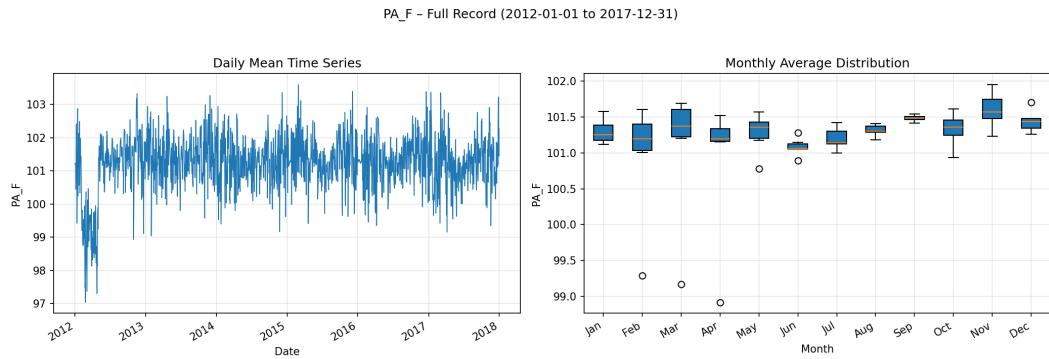


Figure 4: Air Pressure (PA) timeseries boxplot.

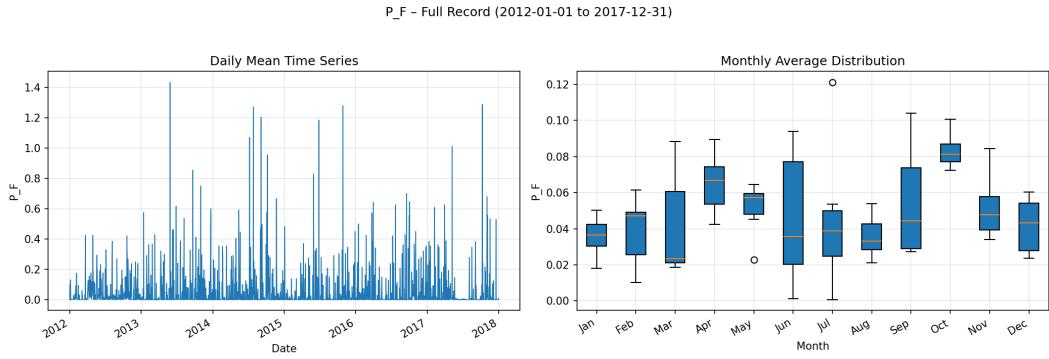


Figure 5: Precipitation (P) timeseries boxplot.

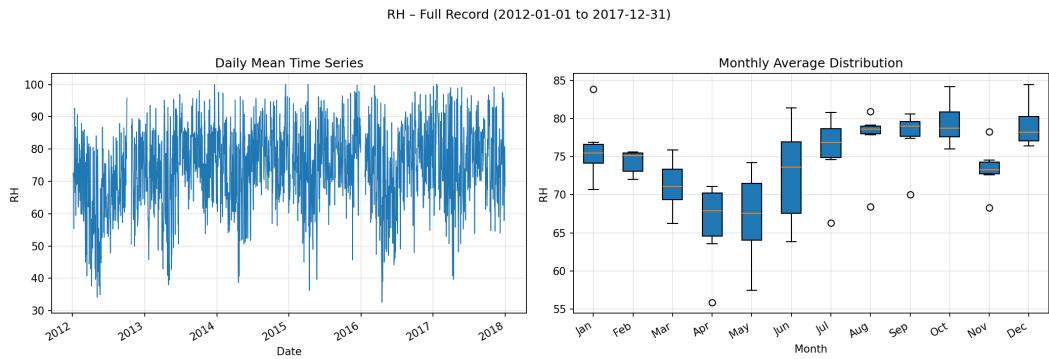


Figure 6: Relative Humidity (RH) timeseries boxplot.

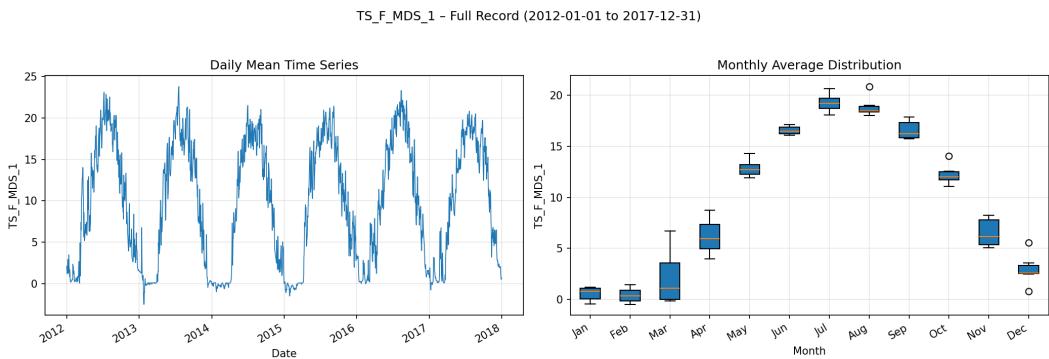


Figure 7: Soil Temperature at 5 cm depth (TS₁) timeseries boxplot.

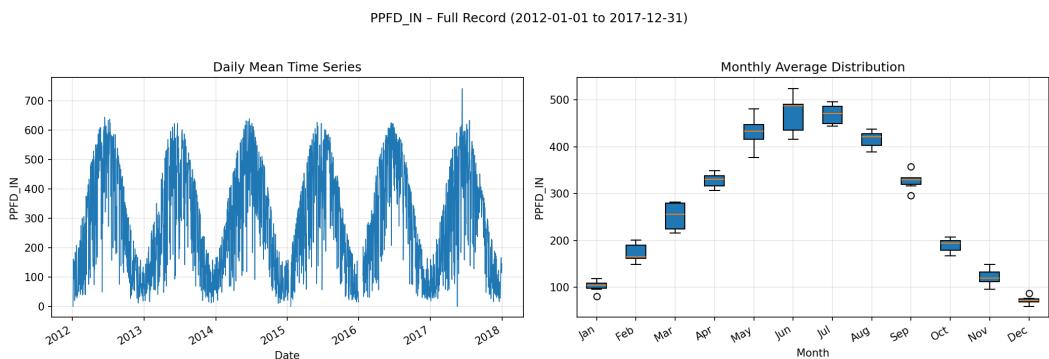


Figure 8: Incoming PAR (PPFD_{in}) timeseries boxplot.

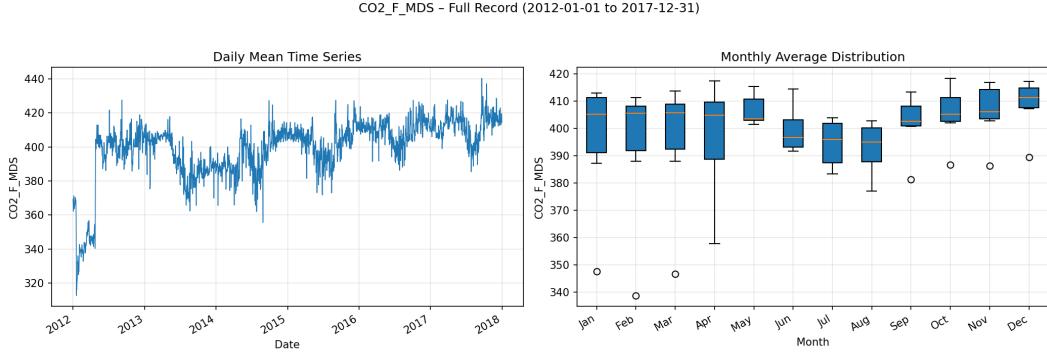


Figure 9: CO₂ concentration timeseries boxplot.

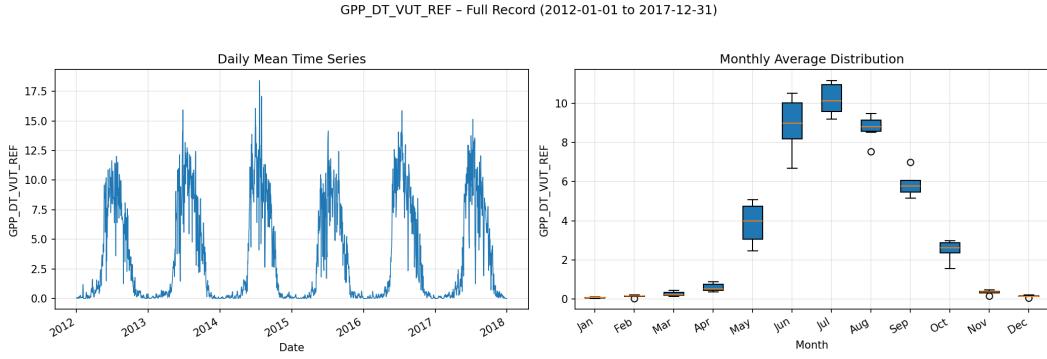


Figure 10: Gross Primary Productivity (GPP_{DT}) timeseries boxplot.

As environmental variables are closely interconnected, correlations can often be observed among their time series. For instance, Figure 11 shows a regression between air temperature and soil temperature, revealing a moderately strong relationship with $R^2 = 0.85$. However, soil temperature rarely falls below 0°C, whereas air temperature can reach much lower values. Moreover, soil temperature responds more slowly to changes compared to air temperature, which aligns with physical expectations: convection occurs readily in the atmosphere but not within the soil, making heat transfer through the soil more gradual and limited.

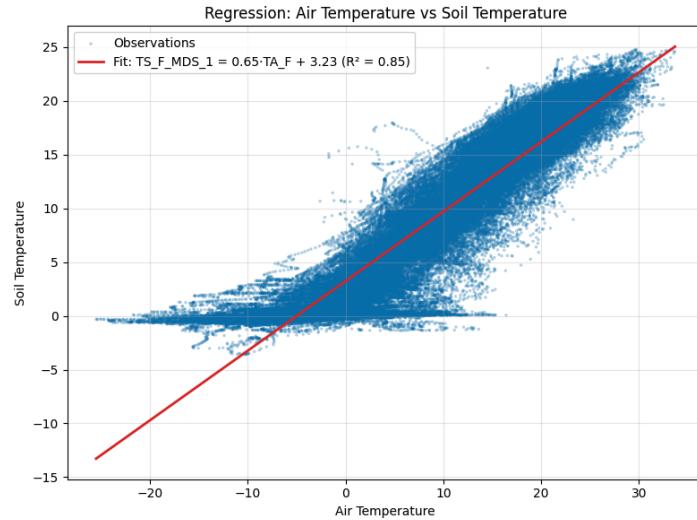


Figure 11: Soil temperature regression with air temperature.

3 Stationary & Preprocessing

For the stationary and preprocessing analysis, we use two variables—precipitation (P) (Figure 5) and gross primary productivity (GPP)(Figure 10), which represents the amount of carbon fixed by vegetation, to perform the stationary check.

3.1 GPP

Table 4: ADF and KPSS stationarity check for GPP time series

| Metric | Raw | First difference | Seasonal difference |
|------------------------|-----------------------|------------------|---------------------|
| ADF Statistic | -7.5505 | -59.3480 | -68.9512 |
| ADF 5% Critical Value | -2.8616 | -2.8616 | -2.8616 |
| ADF p -value | 3.2×10^{-11} | 0.0000 | 0.0000 |
| KPSS Statistic | 0.5132 | 0.0005 | 0.0052 |
| KPSS 5% Critical Value | 0.4630 | 0.4630 | 0.4630 |
| KPSS p -value | 0.0387 | 0.10 | 0.10 |

The raw GPP (Gross Primary Production) time series exhibits partial non-stationarity characteristics by Table 4. The ADF test yielded a strongly negative test statistic (-7.55 , $p \approx 3.2 \times 10^{-11}$), indicating statistical stationarity. However, the complementary KPSS test produced a statistic of 0.513 ($p = 0.039$), exceeding the 5% critical threshold and thus rejecting stationarity.

This disagreement between ADF and KPSS implies that, despite the absence of a stochastic trend, the series likely retains deterministic components such as seasonal variation or low-frequency drift.

After applying first differencing, both ADF ($p = 0.000$) and KPSS ($p = 0.10$) results consistently indicated stationarity, confirming the removal of the long-term trend. Further applying seasonal differencing (lag = 365) reinforced this result, with both tests showing strongly stationary behavior.

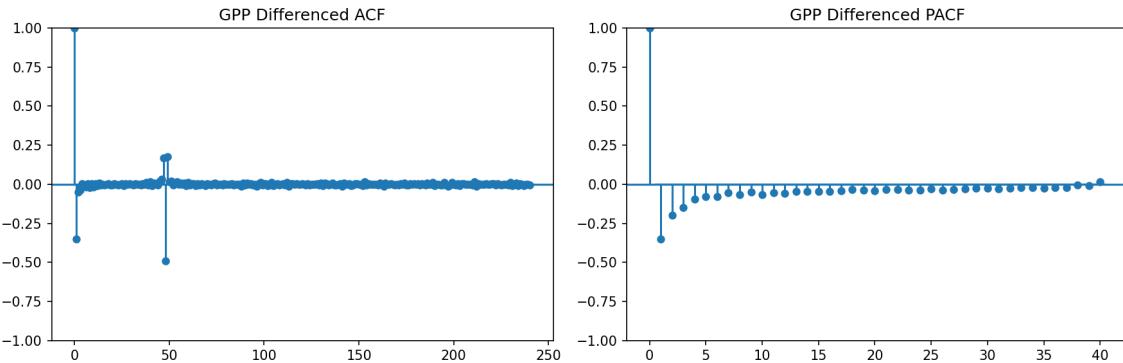


Figure 12: ACF and PACF results for GPP.

Figure 12 presents the autocorrelation (ACF) and partial autocorrelation (PACF) plots of the first-differenced GPP series. In the ACF plot, the correlation at lag 0 equals 1 by definition, while the sharp drop to a small negative value at lag 1 and the rapid decay thereafter indicate

that serial dependence has largely been removed by differencing. A minor secondary spike near lag 50 suggests weak residual periodicity, possibly related to phenological and other indicators.

The PACF plot shows a strong negative correlation at lag 1, implying that the differenced GPP series is primarily influenced by its most recent value and consistent with the ACF plot. Therefore, the differenced GPP series can be regarded as stationary and suitable for subsequent ARIMA modeling.

3.2 Precipitation

Table 5: ADF and KPSS stationarity check for Precipitation (P) time series

| Metric | Raw | First difference | Seasonal difference |
|------------------------|----------|------------------|---------------------|
| ADF Statistic | -49.6962 | -62.5579 | -71.2751 |
| ADF 5% Critical Value | -2.8616 | -2.8616 | -2.8616 |
| ADF <i>p</i> -value | 0.0000 | 0.0000 | 0.0000 |
| KPSS Statistic | 0.0669 | 0.0049 | 0.0034 |
| KPSS 5% Critical Value | 0.4630 | 0.4630 | 0.4630 |
| KPSS <i>p</i> -value | 0.10 | 0.10 | 0.10 |

Table 5 for raw precipitation (P) time series shows substantial temporal variability, reflecting both short-term fluctuations and pronounced seasonal cycles, as expected, raining is generally a random behavior of the nature compare to seasonal cycles of carbon cycles.

The ADF test yielded a highly negative statistic ($-49.70, p = 0.000$), indicating statistical stationarity. Similarly, the KPSS test returned a low statistic ($0.0669, p = 0.10$), failing to reject the stationarity. Although both tests formally indicate stationarity, the wide variability in monthly precipitation and the pronounced annual cycle shown in Figure 5 suggest that the series is not stationary in practice, as its mean and variance change systematically with the seasons and shifts in rainfall regimes.

After applying first differencing, both ADF ($p = 0.000$) and KPSS ($p = 0.10$) results confirmed stationarity, and additional seasonal differencing (lag = 365) further stabilized the mean and variance. Therefore, the differenced precipitation series is considered stationary and suitable for time-series modeling and correlation analysis with GPP.

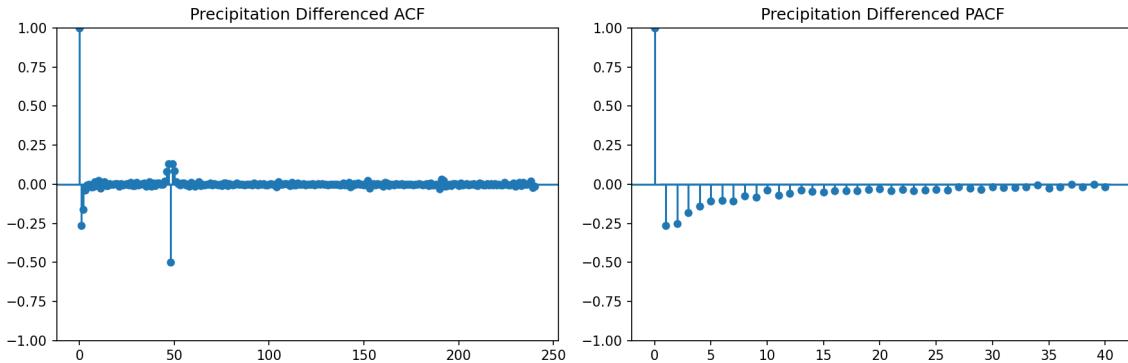


Figure 13: ACF and PACF results for Precipitation.

By observation of Figure 13 and Figure 12, the ACF and PACF patterns of the differenced

precipitation and GPP series appear really similar, which can be explained by physical factors.

From a climatic perspective, precipitation and GPP are closely coupled through shared environmental drivers such as radiation, temperature, and moisture availability. Periods of high rainfall generally enhance plant productivity, while dry periods suppress photosynthetic activity, leading to parallel temporal dynamics in both variables. After applying first differencing, the long-term and seasonal components were effectively removed, leaving short-term fluctuations that reflect common meteorological variability.

4 Model Identification, Fitting, Residue Diagnostics, and Evaluation

4.1 ARIMAX

4.1.1 Model Construction

With first pre-processing the variable, including calculating a daily mean for half-hour measurement, and also proceed first + seasonal differencing. The ARIMAX model was constructed to capture both short-term autocorrelation and the influence of meteorological variables on daily GPP. A simple ARIMA(1,0,1) kernel was selected to represent temporal persistence, while key environmental drivers (PPFD_IN, TA_F, RH) were included as exogenous regressors.

Classical SARIMA seasonality was not used because daily flux data often contain irregular gaps and non-stationary seasonal amplitudes, which make fixed differencing at a 365-day period unstable. Instead, a single Fourier harmonic ($K = 1$) was introduced to represent the smooth annual cycle of GPP. This sine–cosine pair provides a continuous and flexible seasonal baseline without imposing rigid integer-period differencing, allowing the model to capture gradual phenological transitions across years.

Table 6: Estimated ARIMAX model coefficients and significance levels - base model.

| Variable | Coef. | Std. Err. | z | P> z | [0.025 | 0.975] |
|--------------------|---------|-----------|--------|-------|--------|--------|
| TA_F | 0.0054 | 0.017 | 0.322 | 0.747 | -0.027 | 0.038 |
| VPD_F | 0.1298 | 0.028 | 4.689 | 0.000 | 0.076 | 0.184 |
| PA_F | -0.0008 | 0.015 | -0.056 | 0.956 | -0.030 | 0.028 |
| P_F | -0.9272 | 0.126 | -7.380 | 0.000 | -1.173 | -0.681 |
| RH | 0.0384 | 0.006 | 6.209 | 0.000 | 0.026 | 0.051 |
| TS_F_MDS_1 | 0.0320 | 0.036 | 0.887 | 0.375 | -0.039 | 0.103 |
| PPFD_IN | 0.0073 | 0.000 | 32.018 | 0.000 | 0.007 | 0.008 |
| CO2_F_MDS | -0.0055 | 0.004 | -1.563 | 0.118 | -0.012 | 0.001 |
| sin_1 | -0.7799 | 0.256 | -3.048 | 0.002 | -1.281 | -0.278 |
| cos_1 | -2.5592 | 0.408 | -6.265 | 0.000 | -3.360 | -1.759 |
| ar.L1 | 0.8259 | 0.015 | 56.739 | 0.000 | 0.797 | 0.854 |
| ma.L1 | -0.0103 | 0.018 | -0.563 | 0.573 | -0.046 | 0.026 |
| sigma ² | 1.0696 | 0.023 | 46.366 | 0.000 | 1.024 | 1.115 |

Table 6 summarizes the estimated coefficients from the ARIMAX model, with all variable introduced above as input variables.

The fitted ARIMA model generally demonstrates strong explanatory power for daily GPP dynamics, with most meteorological variables showing statistically significant effects. Among the predictors, PPFD_IN (photosynthetic photon flux density) exhibits the highest magnitude and significance ($p < 0.001$), confirming that light availability is the dominant control on

carbon uptake. Vapor pressure deficit also shows a significant positive relationship ($p < 0.001$), indicating enhanced GPP under moderate atmospheric dryness, consistent with expected stomatal responses. Precipitation has a strong negative coefficient ($p < 0.001$), suggesting that wet periods correspond to reduced photosynthetic activity, likely due to cloud cover and low radiation. Relative humidity is positively associated with GPP, reflecting the role of moist atmospheric conditions in maintaining photosynthetic efficiency. The Fourier sine and cosine terms are both significant ($p < 0.01$), capturing the smooth annual cycle of phenological change, while the autoregressive parameter ($\text{ar.L1} = 0.83$) confirms strong temporal persistence in GPP.

Other variables, such as air pressure, soil temperature, and CO₂ concentration, are not statistically significant, implying minor or redundant contributions once primary meteorological drivers are included. So the final model cut those variable and brought up a cleaner version for the updated ARIMA model in table 7. **The AIC value also drops from 4411.219 from the previous full variable input to 4403.957. Although it might not be significant, but it's still a valid drop.**

Table 7: Final ARIMA model coefficients and significance.

| Variable | Coef. | Std. Err. | z | P> z | [0.025 | 0.975] |
|--------------------|---------|-----------|--------|-------|--------|--------|
| VPD_F | 0.1469 | 0.021 | 6.856 | 0.000 | 0.105 | 0.189 |
| PA_F | -0.0208 | 0.005 | -4.169 | 0.000 | -0.031 | -0.011 |
| P_F | -0.8411 | 0.122 | -6.919 | 0.000 | -1.079 | -0.603 |
| RH | 0.0405 | 0.005 | 8.230 | 0.000 | 0.031 | 0.050 |
| PPFD_IN | 0.0072 | 0.000 | 32.748 | 0.000 | 0.007 | 0.008 |
| sin_1 | -0.8705 | 0.257 | -3.389 | 0.001 | -1.374 | -0.367 |
| cos_1 | -3.0722 | 0.320 | -9.592 | 0.000 | -3.700 | -2.444 |
| ar.L1 | 0.8263 | 0.014 | 57.098 | 0.000 | 0.798 | 0.855 |
| ma.L1 | 0.0042 | 0.018 | 0.232 | 0.817 | -0.031 | 0.040 |
| sigma ² | 1.0691 | 0.023 | 46.413 | 0.000 | 1.024 | 1.114 |

4.1.2 Residual Diagnostics

Table 8: Residual Diagnostic Tests for Training Data (SARIMAX Model)

| Test | Statistic / Value | Interpretation |
|-------------------|---|---|
| Zero-mean check | Mean(resid) = -0.000524 | Residual mean is close to zero, indicating no systematic bias in the model. |
| Constant variance | Var(resid) = 1.071 | Residual variance appears stable, suggesting approximately constant variance over time. |
| Ljung–Box test | $p(7) = 1.9 \times 10^{-7}$, $p(14) = 8.3 \times 10^{-16}$, $p(21) = 3.8 \times 10^{-15}$ | Very low p-values indicate significant autocorrelation in the residuals; they are not completely white noise. |

The diagnostic statistics of the ARIMAX(1,0,1) model indicate that the residuals are approximately unbiased, with a mean close to zero (-0.0005), suggesting that the model effectively captures the systematic component of the data. The residual variance is 1.07, remaining relatively stable across observations, which implies that the assumption of constant variance holds reasonably well. However, the Ljung–Box test yields very small p-values at multiple lags, indicating that significant autocorrelation persists in the residuals.

Overall, while the model successfully captures the mean behavior of the series, the presence of autocorrelation suggests that further refinement or inclusion of additional explanatory variables may improve the model’s performance.

4.1.3 Forecast & Validation

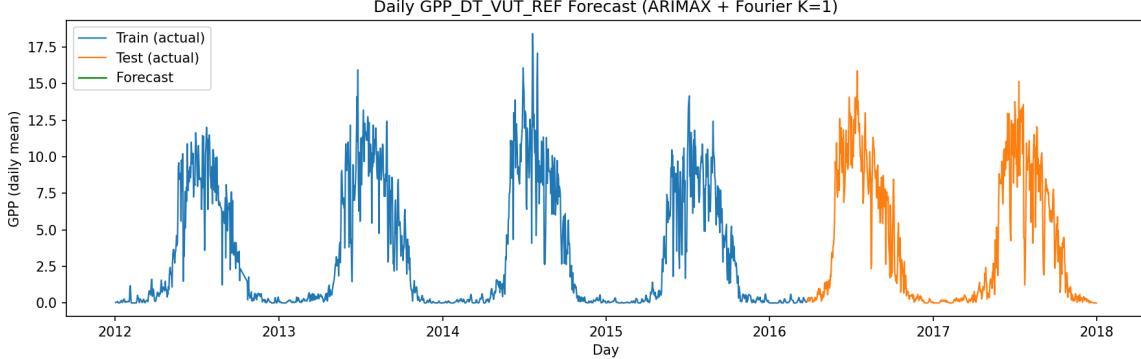


Figure 14: Forecast result with 70% train/test split.

Table 9: ARIMA Model Performance Metrics for Training and Test Sets

| Dataset | MAE | RMSE |
|--------------|--------|--------|
| Training set | 0.6986 | 1.0346 |
| Test set | 0.7124 | 1.1021 |

The model demonstrates consistent performance across the training and test datasets refer to Figure 14 and Table 9. The close alignment between training and test errors indicates that the model generalizes well and does not suffer from significant overfitting. However, the marginally higher test RMSE suggests that some short-term variability or unseen dynamics in the test set may not be fully captured by the model.

4.2 Machine Learning Model - ANN

4.2.1 Model Construction

The ANN model adopts a feed-forward architecture with two hidden layers (32, 16) to balance model capacity and generalization. A ReLU activation and Adam optimizer were chosen for stable and efficient convergence on non-linear flux–meteorology relationships. Daily aggregation and feature standardization reduce noise and scale imbalance, while Fourier terms capture seasonality. A time-series split is used to preserve temporal order during cross-validation, and early stopping prevents overfitting. Residual diagnostics and performance metrics further ensure the model’s reliability on unseen data.

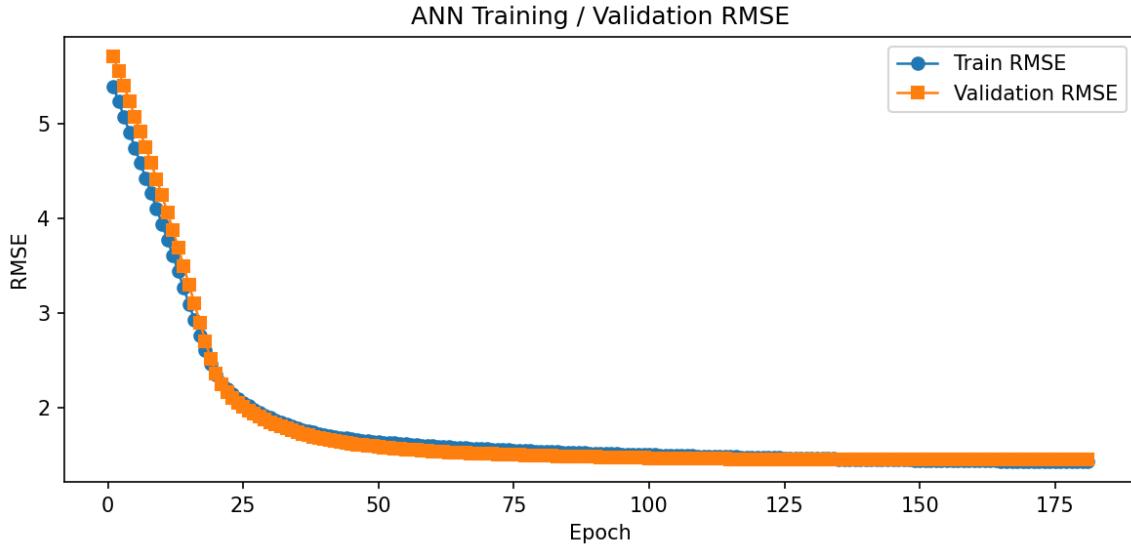


Figure 15: Training and Validation curve for ANN.

Figure 15 shows the RMSE evolution of the ANN during training. Both the training and validation errors decrease rapidly within the first 25 epochs and gradually level off, indicating efficient learning and stable convergence without overfitting. The close alignment of the two curves suggests that the model generalizes well to unseen data. The entire training process completes in approximately three seconds, demonstrating that the network is computationally lightweight and well-suited for daily GPP prediction tasks.

Table 10: Time Series Split Cross-Validation Metrics for the ANN Model

| Fold | Train MAE | Train RMSE | Val MAE | Val RMSE | Val R^2 |
|-------------|-----------|------------|---------|----------|-----------|
| 1 | 0.6844 | 1.0735 | 1.1689 | 1.8438 | 0.805 |
| 2 | 0.7133 | 1.1650 | 1.1616 | 1.9171 | 0.824 |
| 3 | 0.8092 | 1.3785 | 1.0537 | 1.5584 | 0.833 |
| 4 | 0.8642 | 1.4220 | 1.0875 | 1.6888 | 0.855 |
| 5 | 0.8045 | 1.3479 | 0.8981 | 1.4638 | 0.881 |
| Mean | 0.7751 | 1.2774 | 1.0740 | 1.6944 | 0.840 |
| Std. | 0.0742 | 0.1503 | 0.1098 | 0.1895 | 0.030 |

The feed-forward ANN shows stable performance across the five time-based folds. Training errors are small (average RMSE ≈ 1.28), while validation errors are slightly higher (average RMSE ≈ 1.69), meaning the model fits the data well and still performs reasonably on new periods it has not seen. The validation R^2 values (0.81–0.88) indicate that the model explains about 80–88% of the variation in unseen data, suggesting good predictive ability across different seasons. Since the errors stay fairly consistent across folds, the model appears reliable over time.

Overall, the ANN captures the main patterns between GPP and the environmental factors, though some remaining structure in the residuals suggests that adding time-related or seasonal information could make the predictions even more accurate.

4.2.2 Residual Diagnostics

Table 11: Residual Diagnostic Tests for Hold-out Data (ANN Model)

| Test | Statistic / Value | Interpretation |
|-------------------|--|--|
| Zero-mean check | Mean(resid) = -0.0935 | Slight negative bias; residuals are close to zero on average but indicate a small underprediction. |
| Constant variance | Var(resid) = 2.7606 | Variance is finite but not constant over time. |
| Ljung–Box test | $p(7) = 3.94 \times 10^{-177}$, $p(14) = 7.32 \times 10^{-219}$, $p(21) = 9.99 \times 10^{-254}$ | Extremely small p values indicate strong serial autocorrelation; residuals are not white noise. |

Despite reasonable hold-out errors (MAE = 1.116, RMSE = 1.663), the diagnostics reveal serial correlation, heteroskedasticity, and non-normal heavy tails in the residuals. This suggests the ANN captures much of the mean structure but leaves temporally correlated and variance-changing effects unexplained. Potential remedies include adding autoregressive/lagged features (e.g., past GPP/drivers), richer seasonal/covariate terms, variance-stabilizing transforms, or hybrid models that explicitly model residual dynamics.

4.2.3 Forecast

As the evaluation of model performance has been provided above, Figure 16 presents the time series prediction.

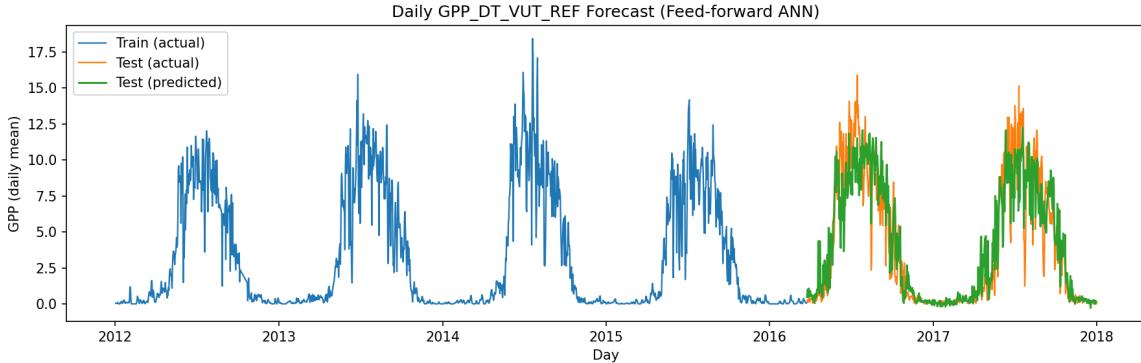


Figure 16: Time series prediction for ANN.

5 Feature Engineering

In the field, a general semi-empirical model for GPP prediction would be Light Use Efficiency (LUE) model. LUE model describes gross primary productivity (GPP) as the product of absorbed light and the efficiency with which plants convert that light into chemical energy. In simple form, it assumes that GPP increases with incoming radiation but is constrained by environmental stressors such as temperature, humidity, and vapor pressure deficit (VPD). Based on this idea, a proxy variable was created as $\text{LUE_proxy} = \text{PPFD_IN} \times \text{RH} / (\text{VPD_F} + 10^{-3})$, representing the amount of light adjusted by atmospheric moisture conditions. This variable captures the intuition that photosynthetic efficiency is higher under bright and humid conditions and declines when the air is dry.

The additional feature was incorporated into the previous ANN model, which generally requires less computation time and offers greater flexibility compared to the ARIMA model. The training and testing accuracies after adding the feature are summarized in Table 12.

Table 12: Comparison of Model Performance Before and After Adding the LUE-based Feature

| Model Version | MAE | | RMSE | | MSE | |
|--------------------|--------|--------|--------|--------|--------|--------|
| | Train | Test | Train | Test | Train | Test |
| Before LUE feature | 0.8202 | 1.1160 | 1.3945 | 1.6629 | 1.9446 | 2.7651 |
| After LUE feature | 0.9962 | 1.2807 | 1.5941 | 1.8224 | 2.5412 | 3.3211 |

The addition of the LUE-based feature caused a small increase in both training and testing errors. This suggests that, although the feature reflects meaningful environmental patterns, it may overlap with information already captured by other variables such as humidity and VPD. In future versions, adjusting how this feature is scaled or combined with others could help the model make better use of it.

6 Process Insights

6.1 Significance of Input Variables

Both the statistical and neural network models consistently identify PPFD_IN (incoming light) as the dominant driver of GPP, which aligns with the fundamental role of radiation as the primary energy source for photosynthesis. Relative humidity (RH) and vapor pressure deficit (VPD_F) also exert strong influences, highlighting the importance of atmospheric moisture in regulating stomatal conductance and photosynthetic efficiency. Precipitation (P_F) exhibits a negative correlation with GPP, likely reflecting the reduction in light availability during cloudy or rainy periods.

By contrast, variables such as air pressure (PA_F) and shallow soil temperature (TS_F_MDS_1) contribute less once the dominant factors of light and humidity are accounted for. Overall, these findings are physically consistent with ecosystem functioning: light governs potential carbon uptake, while moisture-related variables control the efficiency with which vegetation utilizes that energy for photosynthesis.

6.2 Control Loop

In principle, the developed models could be integrated into a feedback or control framework aimed at predicting and managing ecosystem productivity or greenhouse operations. Because both the ARIMAX and ANN models respond predictably to key meteorological inputs, they could be used to forecast short-term changes in GPP under varying light, humidity, or temperature conditions. However, the residual autocorrelation and variance fluctuations observed in the diagnostics indicate that not all temporal dependencies are fully represented.

For operational control applications, further refinement—such as the inclusion of lagged input features, adaptive retraining, or state-space extensions—would improve model stability and responsiveness. With these improvements, the framework could potentially support near real-time monitoring and decision-making in environmental or agricultural systems.

6.3 Process Dynamics

Both modeling approaches successfully capture essential dynamics of the carbon flux system. The autoregressive structure in the ARIMAX model and the temporal learning capacity of the ANN reveal strong persistence in GPP, meaning that present fluxes depend heavily on conditions from preceding days. The inclusion of Fourier terms allows both models to represent the smooth annual cycle associated with plant phenology, while the nonlinear activation functions within the ANN enable the representation of processes such as photosynthetic saturation under high light intensity.

Some short-term residual correlations remain, suggesting that processes such as soil moisture response, canopy acclimation, or delayed physiological feedback introduce time lags that are not yet fully modeled. Nevertheless, the combined effects of persistence, nonlinearity, and seasonality captured in both approaches provide a realistic and physically meaningful representation of ecosystem-scale carbon dynamics.

7 Reflection

Block diagram

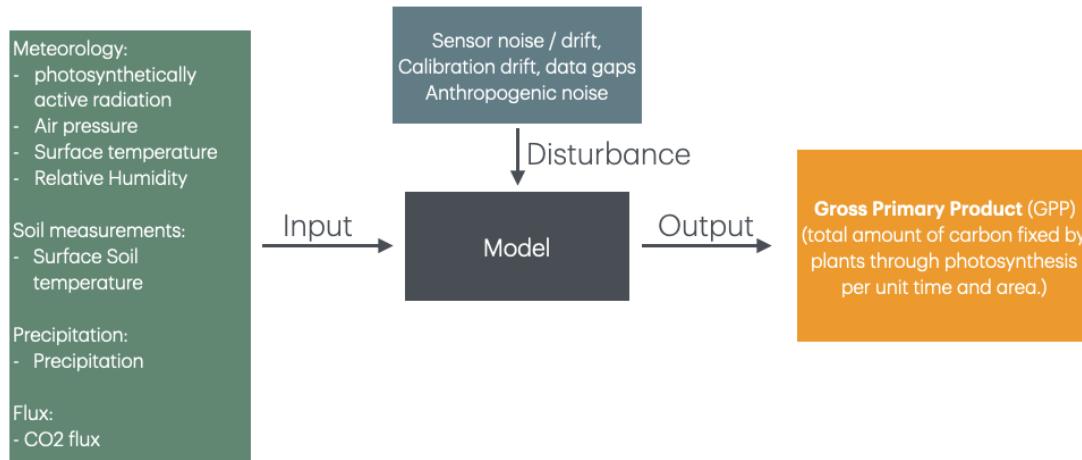


Figure 17: Block diagram (As this is not a chem-eng project, I'm not sure about the block diagram and it seems to be a over-complicating diagram to me not sure if this is true).

Proposed Additional Experiment or Measurement

Because this dataset represents one component of my broader thesis work in atmospheric physics, a logical next step would be to incorporate additional measurements that describe vegetation phenology and canopy dynamics. Satellite-based indices such as the Normalized Difference Vegetation Index (NDVI) or Sun-Induced Chlorophyll Fluorescence (SIF) could provide complementary information on the physiological state of vegetation, including chlorophyll content, canopy structure, and stress responses, which are not fully captured by meteorological variables alone.

Integrating these remote-sensing indicators with flux tower data would strengthen the connection between surface fluxes and ecosystem function. Such combined datasets could help explain residual variability in GPP associated with phenological transitions—such as leaf-out,

senescence, or drought stress—that are currently represented only as unexplained noise in the models. By linking atmospheric and canopy-level information, the predictive accuracy and physical interpretability of the model could be substantially improved, leading to a more complete representation of carbon exchange processes across temporal scales.

Reflection on Model Choice for Real-time Control

For real-time prediction and control applications, the neural network (ANN) model would be the preferred approach over the ARIMA framework. While ARIMA offers higher interpretability and transparency through its linear structure, it is limited in capturing nonlinear interactions among environmental drivers. In contrast, the ANN model flexibly learns complex, multidimensional relationships—such as the joint influence of light, temperature, and humidity—making it better suited for dynamic and coupled atmospheric processes.

Another practical advantage of the ANN model is computational efficiency: training requires only a few seconds, allowing for rapid retraining and deployment in real-time monitoring systems. Its nonlinear design also enables it to adapt to sudden changes in atmospheric conditions, such as shifts in radiation, humidity, or wind patterns, which can strongly affect short-term photosynthetic activity. By continuously updating with new data, the model can maintain high predictive performance even under changing environmental conditions.

Although the neural network is less interpretable than ARIMA, its higher accuracy and adaptability make it more effective for real-time applications. In atmospheric modeling, where rapid responses to variability are critical, flexibility often outweighs simplicity. With proper validation and periodic recalibration, the ANN framework can serve as a reliable and responsive tool for forecasting ecosystem behavior and supporting environmental decision-making.

Interpretability and Model Transparency

It is acknowledged that neural networks present challenges in interpretability and are more susceptible to overfitting compared with traditional time-series models like ARIMA. Their architecture involves multiple interconnected layers and nonlinear transformations that collectively capture complex dependencies within the data but are difficult to express in explicit analytical form. Nevertheless, the fundamental principle remains straightforward: through iterative learning, the network adjusts its parameters to minimize prediction error, gradually uncovering how environmental factors interact to drive carbon flux variability.

Although the internal mechanisms may appear opaque, the patterns learned by the network are grounded in physical processes observed in the data. When combined with systematic validation, regular retraining, and appropriate feature design, these models can serve as robust and insightful tools for understanding and predicting atmospheric–biospheric interactions. In this sense, the neural network functions less as a “black box” and more as a data-driven extension of process understanding, capable of revealing relationships that complement traditional physical modeling approaches.

Code Availability

All code and processed data used in this analysis are publicly available at the following repository: https://github.com/aiqiz/CHE_507_code_repo.

References

- [1] D. D. Baldocchi, “How eddy covariance flux measurements have contributed to our understanding of global change biology,” *Global Change Biology*, vol. 26, no. 1, pp. 242–260, 2020. DOI: <https://doi.org/10.1111/gcb.14807>.
- [2] M. A. Arain, *Ameriflux base ca-tpd ontario - turkey point mature deciduous*, version 3-5, Dataset, AmeriFlux AMP, 2025. DOI: [10.17190/AMF/1246152](https://doi.org/10.17190/AMF/1246152).
- [3] AmeriFlux Management Project, *About ameriflux data*, <https://ameriflux.lbl.gov/data/aboutdata/>, Lawrence Berkeley National Laboratory, 2025. Accessed: Oct. 11, 2025.