Gideon Szamet

5-May-2025

Final Machine Learning Project

Phase 0: Dataset Origin and Selection Rationale

For this project, I used daily ecosystem flux data from the **FLUXNET2015** database. FLUXNET is a global network of eddy covariance sites that monitor exchanges of carbon, water, and energy between terrestrial ecosystems and the atmosphere.

I decided to combine datasets from six different FLUXNET sites:

- **US-Var** (Vaira Ranch – grassland ecosystem)
- **US-Ton** (Tonzi Ranch – oak savanna)
- **US-MMS** (Morgan Monroe State Forest – deciduous broadleaf forest)
- **US-Kon** (Konza Prairie – native tallgrass prairie)
- **US-UMB** (University of Michigan Biological Station – mixed forest)
- **US-Ha1** (Harvard Forest EMS – deciduous forest)

These sites span diverse ecosystems including forests, grasslands, and savannas across the continental U.S. The reasoning behind this multi-site selection was to include a broad spectrum of ecological variability, which can improve the generalizability and ecological realism of any predictive models developed.

Each dataset contained similar daily variables such as radiation, temperature, humidity, and carbon fluxes. I added a site column to track origin and joined them into one cohesive dataset for unified preprocessing and modeling.

Phase 1: Data Exploration and Cleaning

1. Initial EDA ("Understand Your Data")

The dataset includes daily ecosystem measurements from eddy covariance towers. Each row corresponds to one day's data at one site. Variables include meteorological conditions, energy fluxes, and derived productivity metrics.

Data Types:

- Numerical variables: All predictors and the target are float-type
- Categorical variable: site (added during preprocessing)

Dataset Dimensions:

- Rows: 20,895
- Columns: 13 (10 predictors + 1 target + site + timestamp)

Descriptive statistics:

All numerical features (target and predictors) -

- **GPP_NT_VUT_REF** - Gross Primary Productivity (gC/m²/day):
  range = -9.90 to 22.31, mean = 3.18, variance = 17.17
- **TA_F** - Air Temperature (°C):
  range = -17.52 to 36.61, mean = 15.23, variance = 73.53
- **VPD_F** - Vapor Pressure Deficit (hPa):
  range = 0.00 to 46.51, mean = 9.57, variance = 67.69
- **PPFD_IN** - Photosynthetic Photon Flux Density (µmol/m²/s):
  amount of photosynthetically active light
  range = 0.00 to 997.77, mean = 431.04, variance = 49167.52
- **PA_F** - Air Pressure - barometric pressure at the surface (kPa):
  range = 94.15 to 101.31, mean = 98.76, variance = 1.03 –

- **P_F** – Precipitation – daily rainfall or snowfall (mm/day):
  range = 0.00 to 139.19, mean = 1.89, variance = 40.46
- **WS_F** - Wind Speed– near-surface wind speed (m/s):
  range = 0.39 to 11.67, mean = 2.43, variance = 1.67
- **LE_F_MDS** - Latent Heat Flux - proxy for evapotranspiration rate (W/m²):
  range = -22.89 to 252.65, mean = 33.46, variance = 1346.81 –
- **H_F_MDS** - Sensible Heat Flux - convective air heat exchange (W/m²):
  range = -78.47 to 175.42, mean = 44.04, variance = 1858.57
- **G_F_MDS** - Ground Heat Flux – heat exchange with the soil (W/m²):
  range = -63.50 to 32.01, mean = 0.84, variance = 56.50
- **NETRAD** - Net Radiation – net energy available at the surface (W/m²):
  range = -133.91 to 374.74, mean = 106.66, variance = 6432.61

Correlation Analysis:

Correlation of the target with all other features:

Feature	Corr. with GPP
LE_F_MDS	**+0.87**
NETRAD	**+0.36**
RADIATION_BALANCE	**+0.36**
TA_F	+0.32
PPFD_IN	+0.29
G_F_MDS	+0.22
WS_F	+0.10
P_F	-0.01
TEMP_X_VPD	-0.04
VPD_F	**-0.07**
H_F_MDS	**-0.20**
PA_F	**-0.21**

During correlation analysis, I identified an extremely strong correlation (r = 0.99) between SW_IN_F (shortwave radiation) and PPFD_IN (photosynthetically active radiation) (figure 1). These variables both represent incoming solar energy, but PPFD_IN is specific to the light spectrum most relevant for photosynthesis.

To avoid multicollinearity and redundancy, I decided to drop SW_IN_F and retain PPFD_IN, as it is ecologically more meaningful for predicting Gross Primary Productivity.

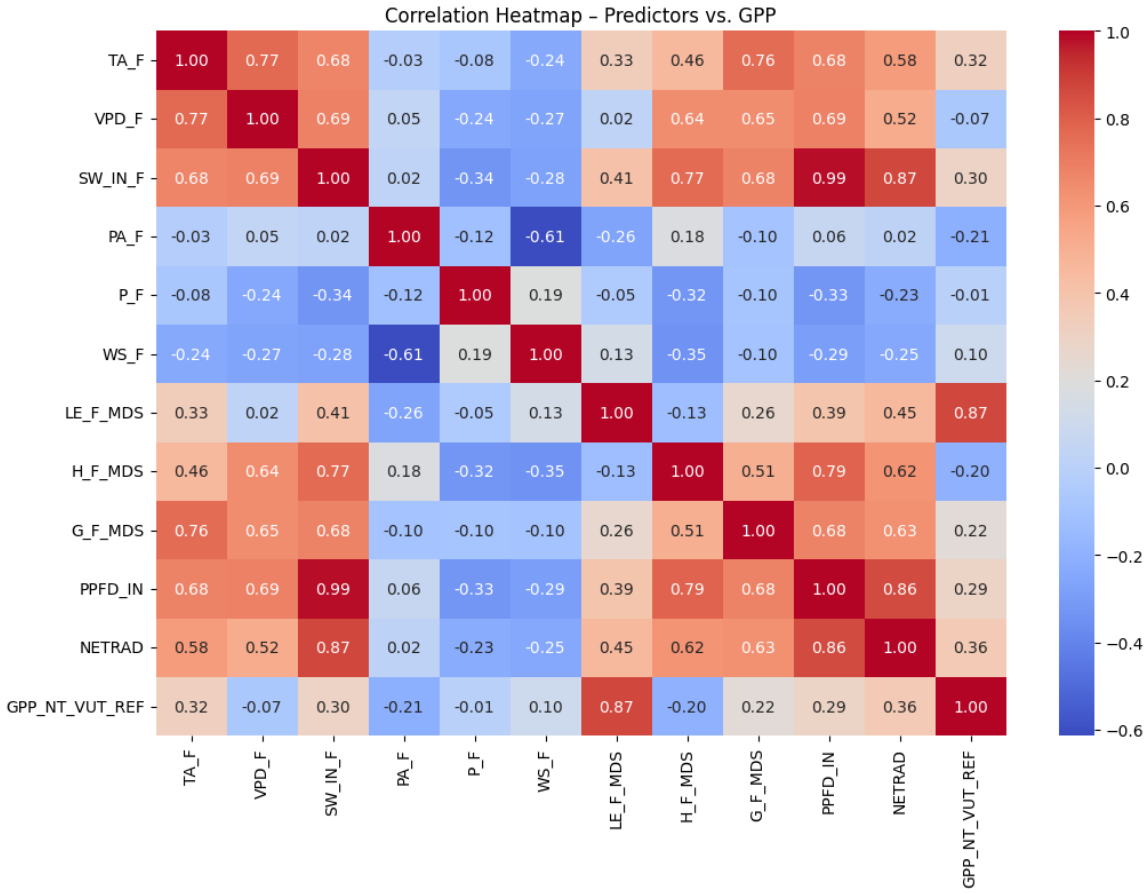

Figure 1: Heatmap of correlations between predictors and GPP.

Missing Values:

All missing values were originally coded as -9999. These were first converted to NaN for proper detection and handling.

2. Clean and Prepare

To prepare the dataset for modeling, I standardized all numerical predictor variables using StandardScaler, ensuring that each feature had zero mean and unit variance. This step is crucial for distance-based algorithms and gradient descent optimization in later modeling phases. I retained the original, unscaled dataset alongside the standardized version (df_scaled) for comparison and flexibility.

No categorical encoding was needed, as the only categorical feature site was used for exploratory plots but not included as a model predictor.

3. Handle Missing Data

After converting -9999 placeholders to NaN, I assessed missing data per feature. Fortunately, none of the selected variables for modeling contained missing values at this stage. However, I did introduce missing values deliberately when engineering lag and rolling features (e.g., 7-day rolling variance), which were handled by row-wise removal of records containing NaN. I chose complete case analysis for this phase (dropping rows with missing engineered features) because:

- The original dataset was large (20K+ rows), so dropping ~100 rows had negligible impact.
- Temporal integrity was preserved per site.
- These missing values were not structural but a result of rolling-window computation (thus MCAR).

4. Visual Exploration and Pattern Detection

I plotted histograms of all standardized predictors and GPP to evaluate distribution shapes and skewness. Most features were unimodal but skewed, with precipitation showing a clear zero-inflated distribution (figure 2). I also created boxplots grouped by site to assess variation across ecosystems (figure 3), and a box plot of the target (GPP) across sites (figure 4).

These visualizations highlighted that:

- GPP and radiation-related variables are highest in grasslands (e.g., US-Var) and lowest in forests (e.g., US-MMS).
- Wind speed, Air Pressure and Sensible Heat Flux showed substantial variability across sites.
- Outliers existed but were not extreme enough to warrant transformation.

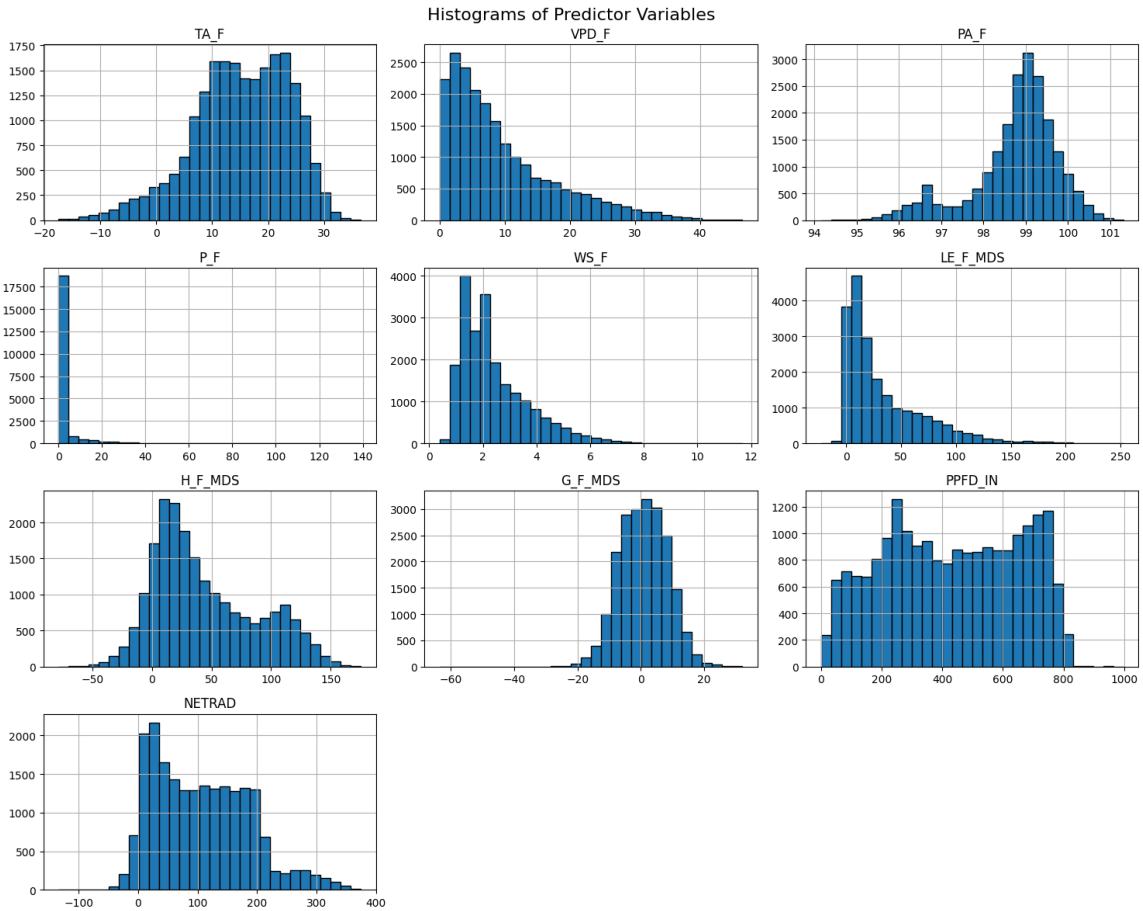

Figure 2: Histograms of standardized predictor variables

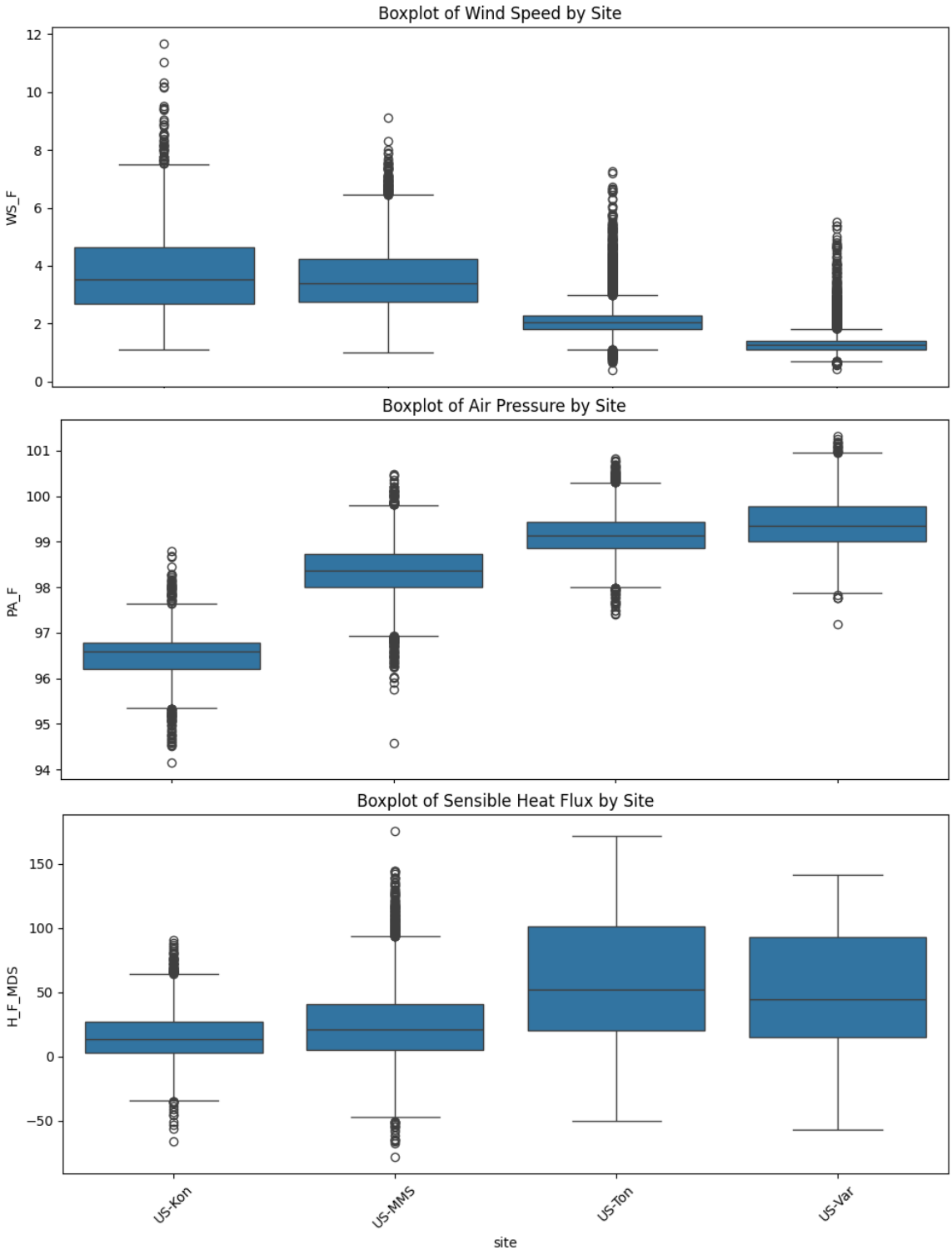

Figure 3: Boxplots of standardized predictors grouped by site

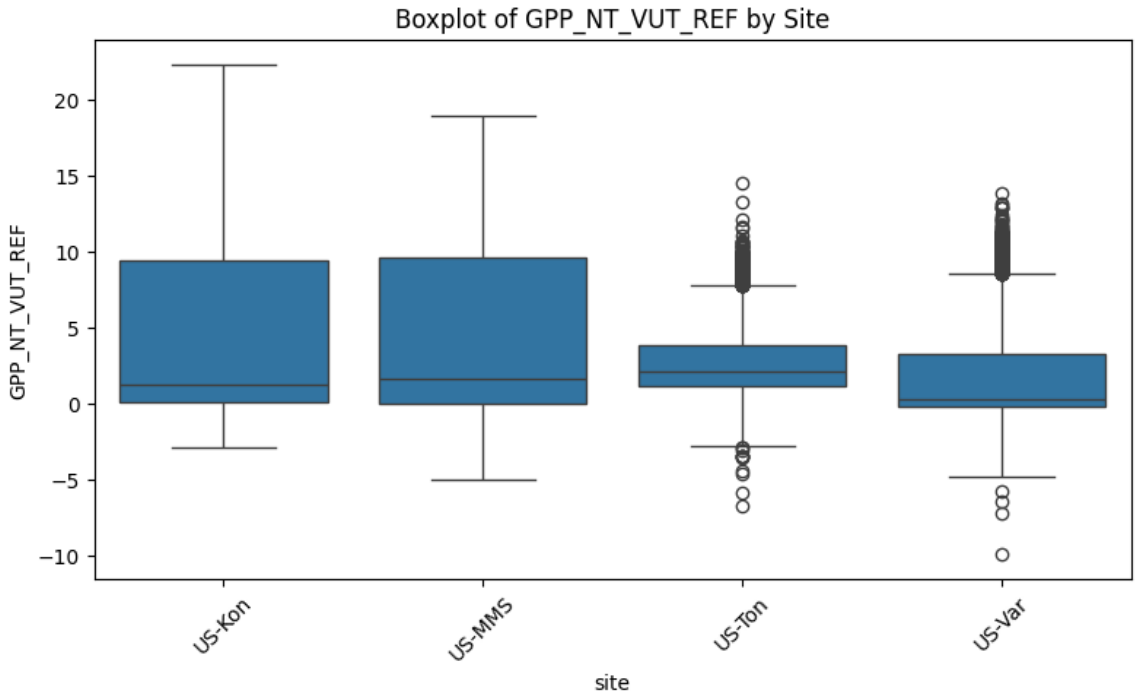

Figure 4: Boxplot of GPP across sites

5. Feature Engineering

I engineered several new features based on ecological knowledge and temporal dynamics:

- TEMP_X_VPD: interaction between air temperature and vapor pressure deficit
- RADIATION_BALANCE: sum of net radiation and ground flux (surface energy balance proxy)
- GPP_LAG1: previous day's GPP (to capture temporal dependence)
- GPP_DIFF: daily change in GPP (first derivative)
- GPP_VAR7: 7-day rolling variance in GPP (temporal stability indicator)

I re-evaluated the correlation between all predictors and GPP after adding the engineered features. Notably:

- GPP_LAG1 had the strongest correlation with GPP (**r = 0.95**)
- GPP_VAR7 also showed substantial correlation (**r = 0.48**)

- Interaction and energy balance features slightly improved explanatory power

The updated dataset was then standardized again using StandardScaler, and visualized via histograms (figure 5).

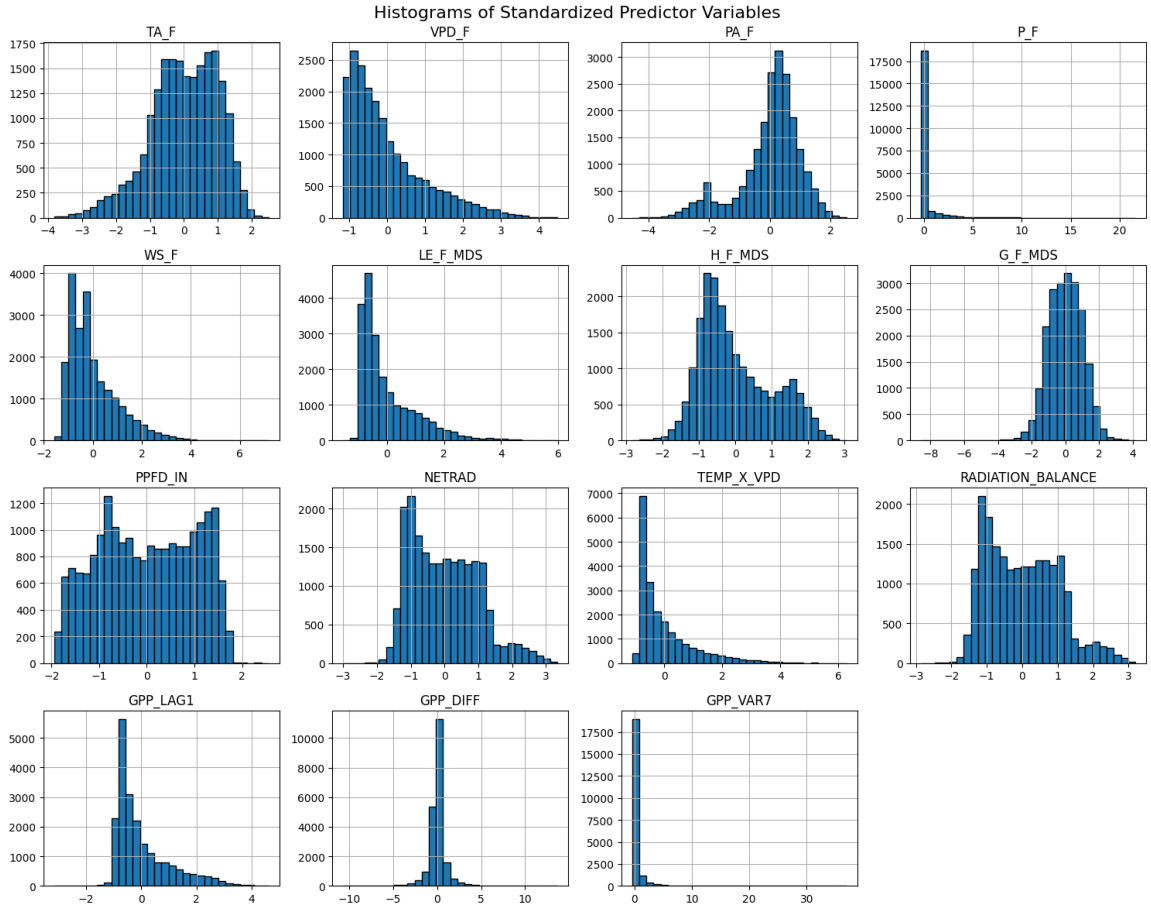

Figure 5: Histograms of standardized engineered predictors.

Phase 2: Hypothesis Formulation

1. Frame a Business-Oriented Hypothesis

I hypothesize that:

Gross Primary Productivity (GPP) can be accurately predicted from daily measurements of ecosystem energy fluxes, atmospheric conditions, and recent temporal trends across diverse terrestrial biomes.

This hypothesis is grounded in prior ecological knowledge and the findings from Phase 1. The high correlation between GPP and both latent energy flux (LE_F_MDS) and recent GPP dynamics (GPP_LAG1, GPP_VAR7) supports this premise. By integrating meteorological, biophysical, and temporal variables, I aim to model how ecosystems convert atmospheric carbon into biomass under different environmental conditions.

However, while it is tempting to use highly predictive temporal variables such as GPP_LAG1, GPP_DIFF, and GPP_VAR7, it is important to recognize that they rely on prior knowledge of the target variable. Such features can only be used for forecasting where previous GPP measurements are available. For estimating GPP in new or unmonitored ecosystems, models based on these variables are not applicable.

Therefore, I decided to develop two modeling pipelines:

- One that includes past-GPP variables (useful for forecasting or gap-filling)
- One that excludes them (to simulate use in novel ecosystems based solely on environmental variables)

2. Important Trends, Relationships, and Features identification

Objective 1: Assess Interaction Effects Between Features and the Target

To examine interactions, I fit an OLS regression model with key predictors (TA_F, VPD_F, LE_F_MDS) and their pairwise interaction terms. The model achieved an R^2 of **0.816**, and all interaction terms were highly significant ($p < 0.001$). This indicates that not only do these variables affect GPP independently, but their combinations (e.g., temperature × humidity, or temperature × energy flux) significantly modulate ecosystem productivity.

Objective 2: Explore Correlations and Clusters

First, I plotted the target var of each site on a time axis to understand the importance of temporal changes on the GPP, there the seasonality is seen clearly, emphasizing the importance and centrality of time in the year on the GPP (figure 7). This observation had led me to another feature engineering of creating 2 predicting temporal variable – month (numerical) and season (categorial) - which was one-hot-key encoded.

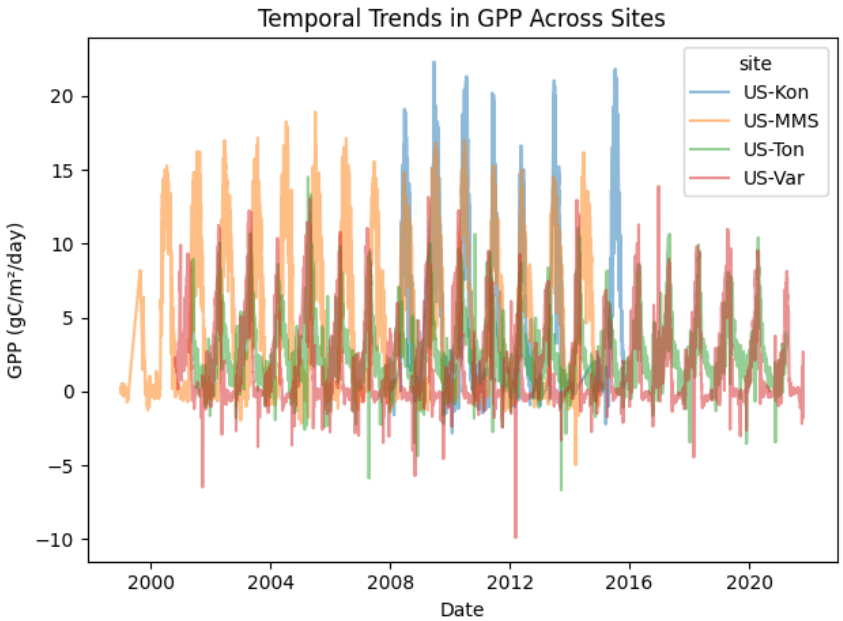

Figure 6: Temporal fluctuations of GPP across sites

To uncover hidden structure in the feature space, I applied Principal Component Analysis (PCA) to reduce dimensionality and then used K-Means clustering (k=3) to look for latent groupings. However, the resulting cluster visualization did not reveal meaningful clusters. The clusters appear arbitrarily distributed in the PCA space and do not align with site-specific patterns or any identifiable environmental regimes (figure 7). This outcome suggests that while the features are predictive of GPP, they do not form coherent unsupervised groupings across ecosystems — possibly due to the continuous gradients in climate and energy fluxes rather than categorical biome types.

To uncover seasonality patterns across years I applied PCA projection colored by season, which revealed a distinct season clustering along PC1 axis (figure 8). A principal component coefficients test reveals that the most dominant predictor composing PC1 is the variable 'Month' with a coefficient of 0.98 (figure 9).

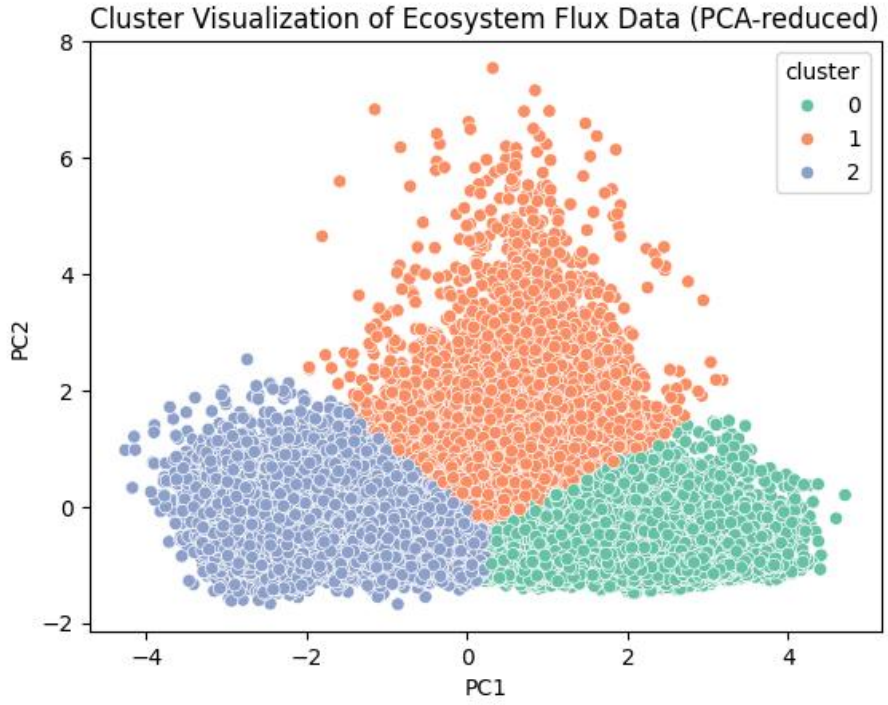

Figure 7: PCA Analysis on continuous variable, excluding temporal variables, facilitating k-means (k=3) for cluster detection

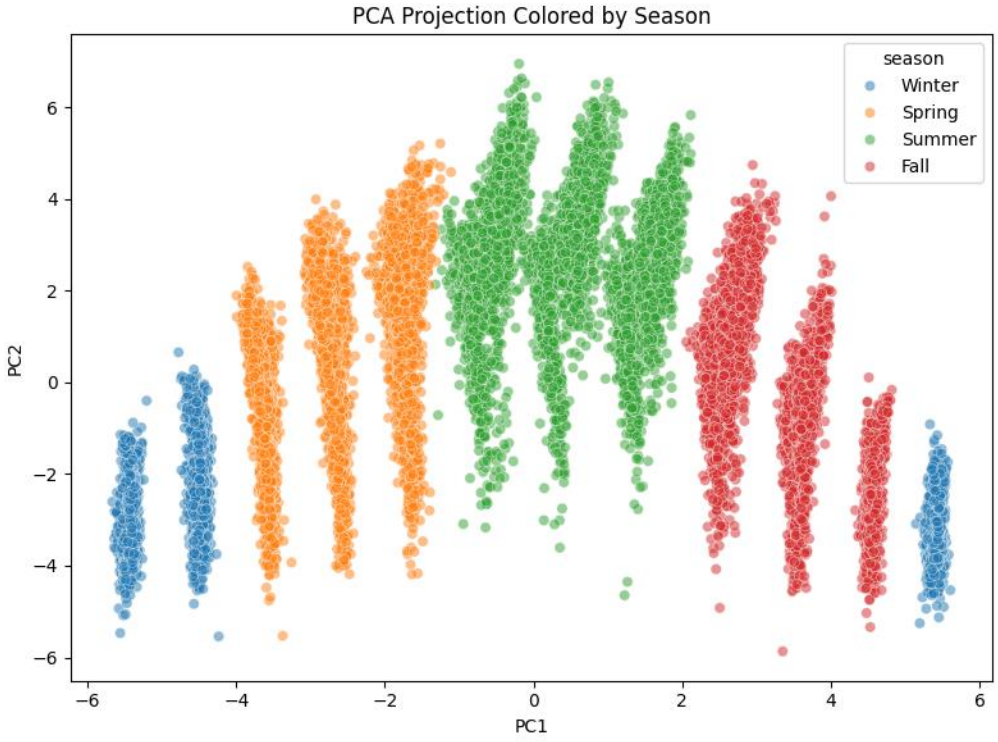

Figure 8: PCA projection based on all predictors (continues and categorial), colored by season category. Clusters formation clearly on PC1 axis

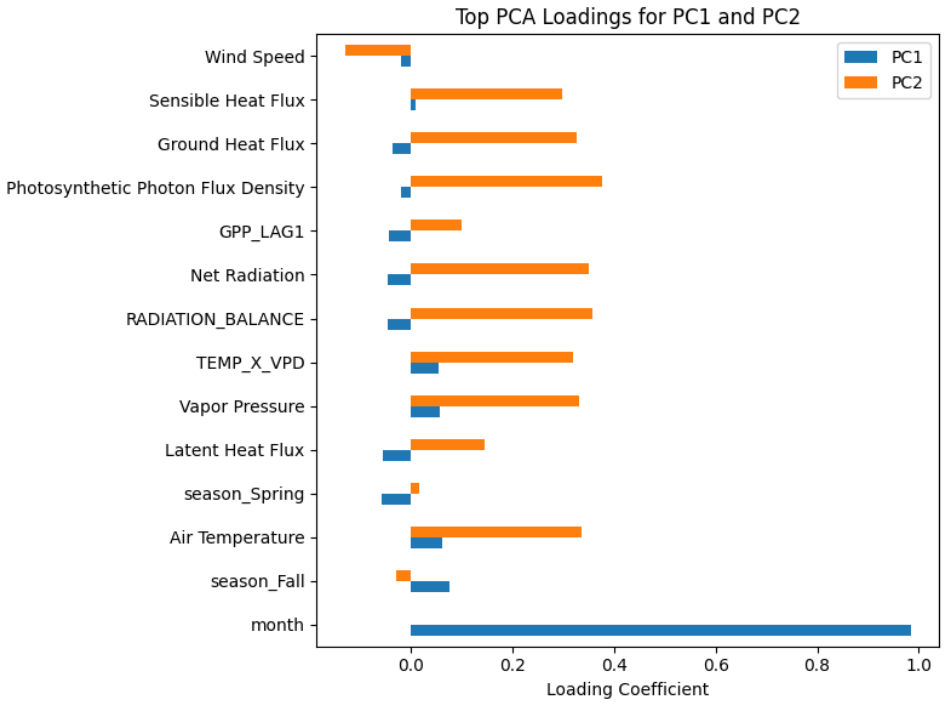

Figure 9: principal component coefficients for PC1 and PC2 of PCA projection colored by season

Phase 3: Supervised Learning

Here I am aiming to test whether Gross Primary Productivity (GPP) can be accurately predicted from environmental variables and temporal patterns. Based on the business-oriented hypothesis formulated earlier, I constructed supervised learning models to quantify the relationship between daily GPP and its potential drivers.

To support both forecasting and generalization use cases, I trained models under two scenarios:

1. Using all available features, including prior GPP values (for short-term forecasting),
2. Using only meteorological and flux variables, excluding past GPP (for prediction in unmonitored ecosystems).

I selected a combination of linear and non-linear models to compare both interpretability and predictive power.

- Linear models (OLS, Ridge, Lasso) serve as strong baselines and provide insight into feature weights.
- Tree-based models (Random Forest, XGBoost) are well-suited for non-linear, high-dimensional data like mine and provide robustness and built-in feature selection.
- SVR was added as a margin-based learner that can generalize well, though it may be slower to train.
- KNN is included to assess how well local similarity captures variation in GPP.

Based on my data (20K+ rows, non-linear relationships, feature interactions, and no strong outliers), I expect Random Forest Regressor or XGBoost Regressor to perform best. These models handle non-linearity, interaction effects, and unscaled features well, and they've already shown high feature importance alignment with ecological expectations.

Ordinary Least Squares (OLS) Regression

The OLS model that includes GPP history performs almost perfectly, achieving an average R^2 of 1.0000 ± 0.0000 and RMSE of 0.0000 ± 0.0000 in cross-validation. The fitted equation is:

$$GPP_{NT-VUT-REF} = 3.1767 + 4.1459 \cdot GPP_{LAG1} + 1.3165 \cdot GPP_{DIFF}$$

This performance reflects the strong autoregressive nature of GPP across consecutive days. The model selected only two predictors—GPP_LAG1 and GPP_DIFF—to reconstruct current GPP, ignoring all environmental variables. While effective for short-term forecasting, this model can only be applied in ecosystems where recent GPP observations are already available and is therefore not suitable for broader ecological applications.

In contrast, the OLS model trained without GPP history variables still achieves strong performance, with an average R^2 of 0.8172 ± 0.0073 and RMSE of 1.7669 ± 0.0093. Its fitted equation includes a diverse set of environmental predictors:

$$\begin{aligned} GPP_{NT-VUT-REF} &= 3.3757 + 1.2483 \cdot TA\_F - 3.0815 \cdot VPD\_F + 0.1767 \cdot PA\_F \\ &\quad - 0.2355 \cdot P\_F + 0.0535 \cdot WS\_F + 2.6809 \cdot LE\_F\_MDS - 0.9474 \\ &\quad \cdot H\_F\_MDS - 0.2264 \cdot G\_F\_MDS + 0.9301 \cdot PPFD\_IN - 0.0507 \\ &\quad \cdot NETRAD + 1.6382 \cdot TEMP\_X\_VPD - 0.0677 \\ &\quad \cdot RADIATION\_BALANCE - 0.2773 \cdot season\_Fall - 0.1123 \\ &\quad \cdot season\_Spring + 0.9109 \cdot season\_Summer - 0.5213 \\ &\quad \cdot season\_Winter - 0.0344 \cdot month \end{aligned}$$

This model demonstrates that meteorological and seasonal variables alone can explain a substantial portion of GPP variability, making it more applicable in broader, real-world ecological contexts where GPP is not continuously measured.

General Comment: Since GPP_LAG1 and GPP_DIFF fully reconstruct the target with a simple OLS, there's little value in further modeling that version. From here the investigation will continue with the GPP-independent dataset only, ensures focus on generalizable predictors.

Ridge Regression

I trained a Ridge regression model using the GPP-independent dataset and performed 5-fold cross-validation with a grid of alpha values ranging from 0.01 to 100. The optimal regularization strength was alpha = 0.01.

The model achieved an average RMSE of 2.26 and an R^2 of 0.48, both notably lower than the OLS model (R^2 = 0.82). This reduction in performance suggests that Ridge's regularization, while useful for mitigating overfitting in complex models, penalizes coefficients too strongly in this low-noise, well-preprocessed dataset.

The resulting regression equation retained a similar structure to the OLS model, with latent heat flux (LE_F_MDS), temperature (TA_F), and photosynthetically active radiation (PPFD_IN) remaining strong positive contributors to GPP, while vapor pressure deficit (VPD_F) and sensible heat flux (H_F_MDS) had strong negative effects. Seasonal terms also contributed to the prediction, with summer increasing GPP and winter decreasing it.

Given its weaker predictive performance and the fact that regularization appears to dampen meaningful ecological signals, I do not consider Ridge regression a good choice for this problem.

Lasso Regression

The Lasso model selected an alpha of 0.0001, resulting in an RMSE of 1.76 and an R^2 of 0.82 on the training set. This performance is slightly better than the ordinary least squares (OLS) model without past GPP, suggesting that Lasso's

regularization helped refine the model by emphasizing the most impactful features.

The model retained nearly all predictors, assigning substantial weights to key variables such as latent heat flux (proxy for evapotranspiration), vapor pressure deficit (humidity stress), and the interaction between air temperature and vapor pressure. Seasonal categories and air temperature also contributed meaningfully to the prediction.

While Lasso regression imposes a penalty to encourage simpler models, in this case it did not eliminate many predictors. As a result, the final model is slightly more complex than the standard OLS but achieves marginally improved predictive accuracy.

Random Forest Regression

To capture nonlinear relationships and potential feature interactions, I trained a Random Forest Regressor on the dataset without prior GPP values. The model was evaluated using 5-fold cross-validation.

It achieved an average R^2 of 0.899 and an RMSE of 1.31, significantly outperforming the linear models. This performance indicates that the Random Forest model can explain nearly 90% of the variance in Gross Primary Productivity using only environmental and seasonal predictors. These results highlight the presence of nonlinear dynamics and complex interactions in the data that linear models like OLS and Lasso could not fully capture.

Given its strong performance and generalizability without relying on past GPP observations, Random Forest emerges as a promising candidate for robust GPP prediction across diverse ecosystems.

XGBoost (Extreme Gradient Boosting)

The final model tested using the feature set that excluded prior GPP observations. It achieved a mean RMSE of 1.31 and an R^2 of 0.899 in 5-fold

cross-validation. This performance is effectively tied with the Random Forest model, and substantially better than all linear models tested.

XGBoost is known for its ability to capture complex non-linear interactions between features through an ensemble of shallow trees. Its performance confirms that ecosystem productivity can be reliably estimated from environmental measurements alone, even in the absence of prior GPP data. Compared to linear models like OLS, Ridge, or Lasso, XGBoost captures richer relationships in the data, which aligns with ecological intuition—plant productivity is driven by multiple interacting factors including temperature, vapor pressure deficit, light availability, and seasonal trends.

The model revealed clear dominance of Latent Heat Flux and seasonality (particularly Summer) in predicting Gross Primary Productivity (GPP). These two features alone accounted for 62.8% of the total predictive importance, emphasizing the role of evapotranspiration and seasonal vegetation activity in ecosystem carbon uptake. Notably, the feature Radiation Balance, which I engineered as the sum of Net Radiation and Ground Heat Flux, emerged as one of the top predictors. This highlights the value of domain-informed feature engineering, as it captures the net energy available for photosynthesis and water fluxes. Other moderately important predictors included Sensible Heat Flux and Air Pressure, both consistent with known ecological drivers. Interestingly, more conventional variables such as Air Temperature, Vapor Pressure Deficit, and Photosynthetic Photon Flux Density played relatively minor roles, suggesting that the model relies more on integrated energy and moisture dynamics than on temperature alone.

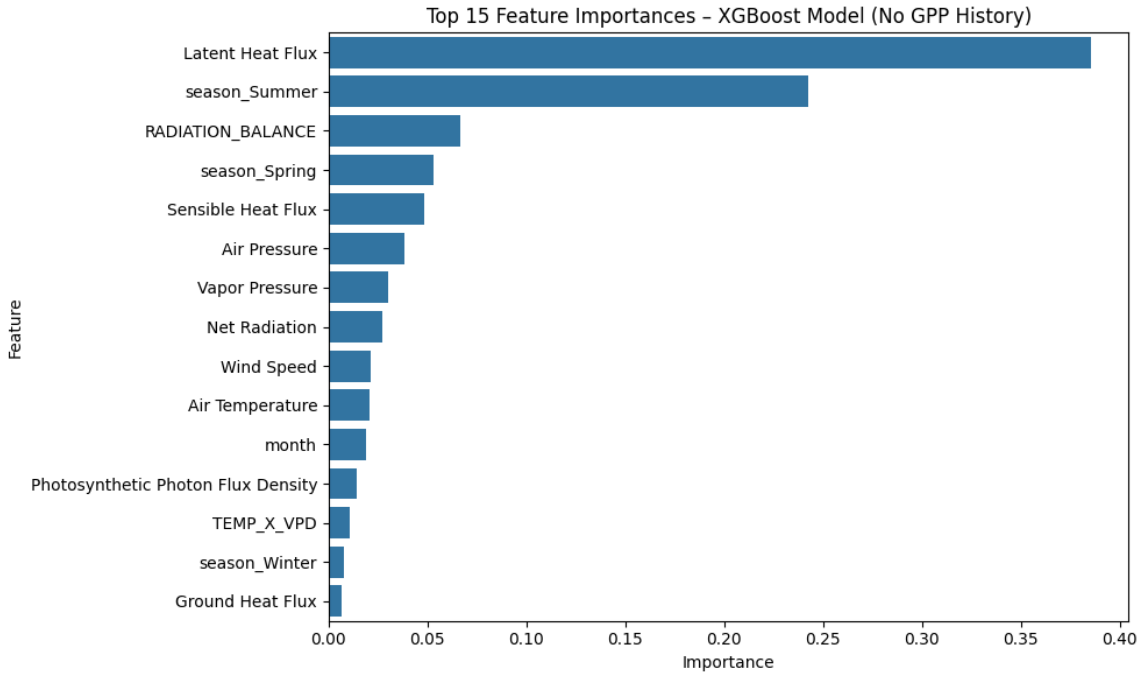

Figure 10: XGBoost model features importance

Support Vector Regression (SVR)

Support Vector Regression was tested using an RBF kernel, with hyperparameter tuning over C and epsilon. The best configuration was C=10, epsilon=0.1, and kernel='rbf'. The model achieved an average RMSE of 1.72 and an R^2 of 0.73 in cross-validation. While SVR performed reasonably well, its performance was lower than that of Random Forest and XGBoost, and its training time was significantly longer — over 13 minutes for tuning and evaluation. This makes it less practical for large datasets, despite its ability to capture non-linear relationships.

Evaluation

Since this is a regression task with a continuous target variable (Gross Primary Productivity), I used Root Mean Squared Error (RMSE) and the coefficient of

determination (R^2) to evaluate model performance. RMSE provides an interpretable estimate of the average prediction error in the units of the target ($gC/m^2/day$), while R^2 reflects the proportion of variance in GPP that can be explained by the predictors.

I compared the performance of five supervised learning models on the test set using the GPP-independent feature set:

- OLS Regression: RMSE = 1.78, R^2 = 0.82
- Ridge Regression: RMSE = 1.78, R^2 = 0.82
- Lasso Regression: RMSE = 1.78, R^2 = 0.82
- Random Forest: RMSE = 1.26, R^2 = 0.91
- XGBoost: RMSE = 1.26, R^2 = 0.91

Linear models (OLS, Ridge, and Lasso) provided reasonable performance, capturing over 80% of the variance in test data. Notably, Lasso slightly outperformed the others, suggesting that regularization may help reduce overfitting. However, both Random Forest and XGBoost significantly outperformed linear models, reducing RMSE by nearly 30% and achieving over 90% explained variance.

I also examined residuals to assess model fit. The residuals from both Random Forest and XGBoost models were approximately symmetric, normally distributed, and homoscedastic, which suggests a stable model across the target range. These diagnostics reinforce the reliability of the tree-based methods for this prediction task.

Based on both cross-validation and test performance, I selected the **XGBoost** model as the final model for downstream application. It matched the performance of the Random Forest model in both RMSE and R^2 but executed significantly faster — training in approximately 5 seconds compared to over 5

minutes for Random Forest. This efficiency makes XGBoost more practical for iterative workflows, hyperparameter tuning, and future scalability, while still offering strong predictive power and robust feature importance metrics.

Although these results are satisfactory, further improvements could come from:

- Incorporating additional temporal features (e.g., moving averages of meteorological variables)
- Testing site-specific models rather than pooled ones
- Experimenting with hybrid models that combine linear trends and tree ensembles

Overall, the modeling phase confirms that environmental and temporal features alone can predict daily ecosystem productivity with high precision, even in the absence of previous GPP observations.

Phase 4: Unsupervised Learning

Clustering Model Selection and Application

Before applying clustering algorithms, I hypothesized that K-Means would be the most effective for this dataset. The data was already standardized and likely exhibits continuous gradients (e.g., environmental variables), which K-Means is well-suited to detect. However, due to the ecological nature of the data and the possibility of non-spherical clusters, I also tested Agglomerative Clustering and DBSCAN.

Features Used for Clustering

I selected all environmental variables excluding the target (GPP_NT_VUT_REF) and lag-based features. These included:

- Air Temperature
- Vapor Pressure
- Air Pressure
- Precipitation
- Wind Speed
- Latent Heat Flux
- Sensible Heat Flux
- Ground Heat Flux
- Net Radiation
- Photosynthetic Photon Flux Density
- TEMP_X_VPD
- RADIATION_BALANCE
- month (numerical)
- Seasonal dummy variables (Spring, Summer, Winter)

These features capture both abiotic drivers and seasonal cycles that likely structure ecosystem processes.

Clustering Algorithms to be Applied

- K-Means: with k=3 to k=6
- Agglomerative Clustering: With Ward linkage, 3–6 clusters
- DBSCAN: Multiple values of eps (0.5–1.5) and min_samples (5–20)

K-Means Clustering Analysis

To explore latent groupings in the data, I applied K-Means clustering with k ranging from 3 to 6. The clustering was performed on PCA-reduced data to capture the most informative variance in a two-dimensional space.

The silhouette scores for each value of k were as follows:

- **k = 3**: 0.367
- k = 4: 0.334
- k = 5: 0.332
- k = 6: 0.283

The highest silhouette score was achieved with **k = 3**, suggesting that three clusters offer the most cohesive separation in the feature space. The PCA visualization of the clusters shows distinct but overlapping regions, indicating moderate cluster separation (figure 11). This structure likely reflects ecological gradients (e.g., energy flux, temperature) rather than discrete ecosystem types. These clusters may represent different functional regimes or environmental states across the diverse FLUXNET sites included in the dataset.

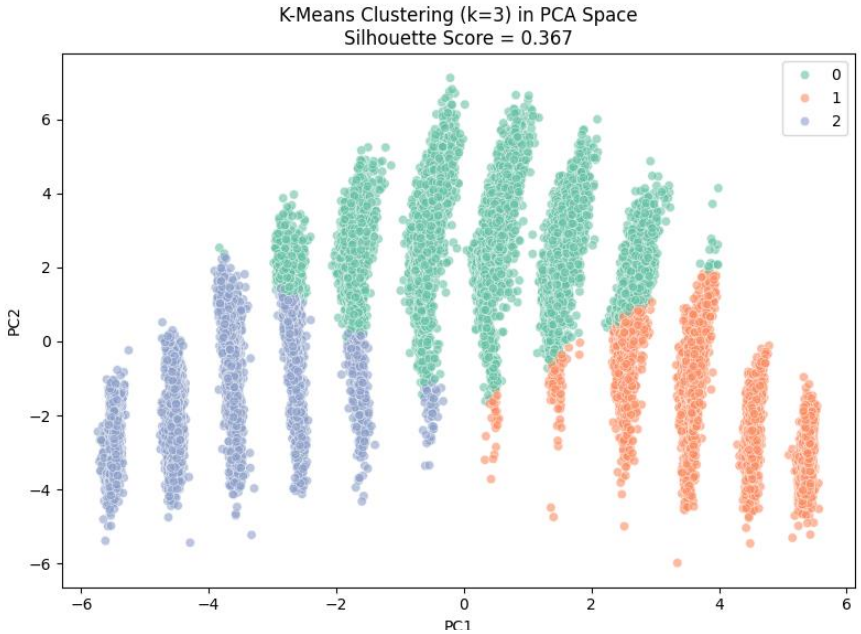

Figure 11: K-Means Clustering (k=3) in PCA space

Agglomerative Clustering Analysis

I applied Agglomerative Clustering using Ward linkage with the number of clusters ranging from 3 to 6. The silhouette scores were consistently lower than those obtained using K-Means, indicating weaker cohesion within clusters:

- k = 3: 0.204 (figure 12)
- k = 4: 0.178
- k = 5: 0.191
- k = 6: 0.174

These results suggest that Agglomerative Clustering was less effective at capturing structure in the feature space. This may reflect the continuous nature of the environmental gradients and the lack of discrete hierarchical relationships between ecosystem states in the data.

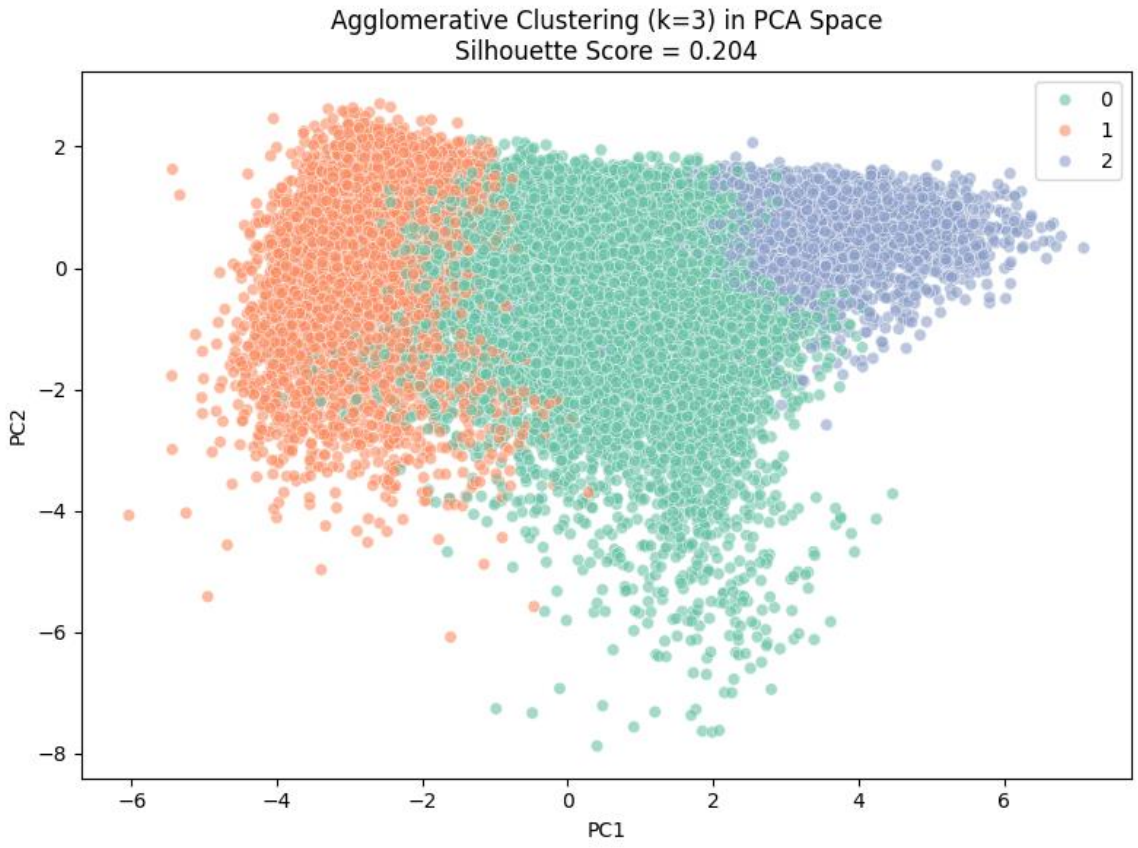

Figure 12: Agglomerative Clustering (k=3) in PCA space

DBSCAN Clustering Analysis

To explore non-spherical and density-based groupings, I applied DBSCAN with varying eps (0.5 to 1.5) and min_samples (5 to 20). This method can identify clusters of varying shape and is robust to noise, making it suitable for complex ecological datasets.

The results showed a wide range in the number of clusters and noise points:

- At low eps values (e.g., 0.5), the algorithm produced many small clusters and a high proportion of noise points (e.g., 109 clusters and 13,269 noise points with min_samples=5), with poor silhouette scores.
- As eps increased, the number of clusters decreased, and silhouette scores gradually improved.
- The best silhouette score (0.381) was achieved with eps=0.5 and min_samples=20, though it still left over 18,000 points as noise.
- At eps=1.5, DBSCAN produced fewer clusters but often collapsed into a single cluster, making silhouette evaluation invalid.
- The best DBSCAN configuration found with eps=1.5 and min_samples=5, produced 3 clusters, with 214 noise points and Silhouette Score: 0.339 (figure 13)

Overall, while DBSCAN was able to form a few cohesive groupings under certain parameter combinations, it struggled to identify stable and interpretable clusters across the dataset. Compared to K-Means (silhouette up to 0.367) and Agglomerative Clustering (up to 0.204), DBSCAN offered moderate structure but with limited practical separation and high sensitivity to parameter choice.

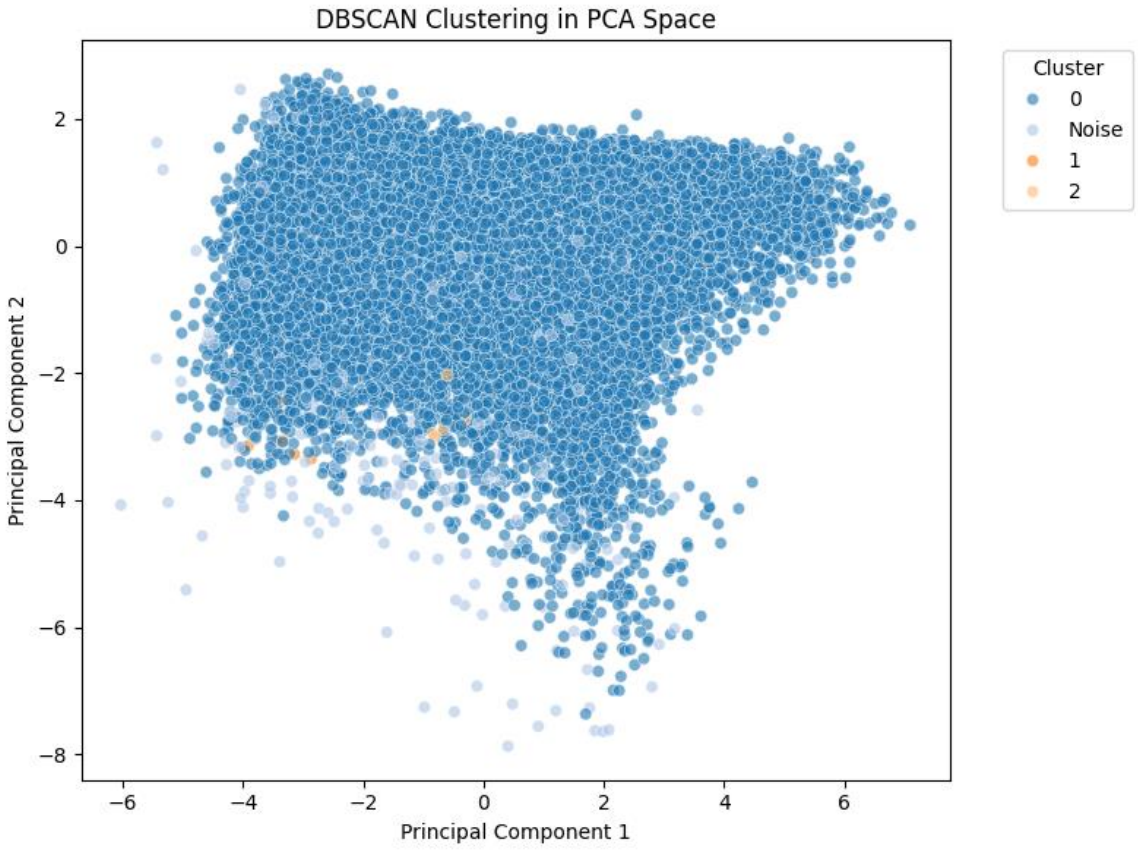

Figure 13: DBSCAN clustering results projected into PCA space (eps = 1.5, min_samples = 5). The clustering produced 3 compact clusters with minimal noise (214 points) and a silhouette score of 0.339. Unlike K-Means, DBSCAN can identify arbitrarily shaped clusters

Interpreting Clustering Results in Business Context

Despite extensive experimentation with multiple unsupervised learning algorithms—including K-Means, Agglomerative Clustering, and DBSCAN—the results revealed **only weak and ambiguous clustering structure** in the dataset. Even the best-performing model (K-Means with $k = 3$) yielded a modest silhouette score (0.367), indicating limited internal cohesion and substantial cluster overlap. DBSCAN, though capable of identifying noise points and irregular clusters, was similarly unable to isolate ecologically meaningful groupings, often collapsing into one or two clusters or fragmenting into many small ones.

These findings suggest that the environmental and flux variables in this dataset vary along continuous gradients, rather than forming discrete ecosystem regimes that unsupervised algorithms can readily detect. Consequently, clustering did not reveal clear subgroups that could be translated into actionable insights or targeted strategies.

While unsupervised learning did not produce useful segmentation in this case, the process was still informative: it confirmed the absence of strong latent structure and emphasized the importance of using target-based predictive modeling (as done in the supervised phase) to understand ecosystem function. Future work could revisit clustering using more domain-specific transformations or functional trait variables to uncover hidden ecological modes, or explore semi-supervised approaches that blend partial label information with clustering objectives.