DATA SCIENCE FOR SUSTAINABLE AGRICULTURE

Name: - BHAVANA PARUPALLI (113536974)

Semester: - SPRING – 2023

Course: - DSA5900 Professional Practicum (4 credit hours)

Faculty Supervisor: - Dr. David Ebert

Company & Sponsor – Data Institute for Societal Challenges (DISC) – Dr. Ebert

Mail: - parupallibhavana123@ou.edu

Credits – United States Department of Agriculture (USDA)

Table of Contents

1. INTRODUCTION

Agriculture is one of the primary sectors of the global economy, and sustainable agricultural practices are crucial for ensuring food security, environmental conservation, and economic growth. However, climate change and environmental degradation pose significant challenges to sustainable agriculture. To address this challenge, data science techniques can be used to study the dynamics and present insights that can help farmers and policymakers make more informed decisions and reduce the environmental impact of agriculture. This project focuses on studying the effects of 1) methane turbulent flux, 2) ecosystem respiration, and 3) MODIS EVI (greenness of grassland) from available environmental variables.

Methane is a potent greenhouse gas that is emitted during agricultural activities such as livestock management and rice cultivation. The prediction of methane emissions from available environmental variables such as solar radiation, air temperature, soil temperature, relative humidity, and vapor pressure deficit can aid in developing sustainable agricultural practices that reduce greenhouse gas emissions.

Ecosystem respiration is a critical process that plays a crucial role in carbon cycling in terrestrial ecosystems. The prediction of ecosystem respiration from environmental variables can aid in the understanding of the carbon balance in agricultural systems and help develop management practices that reduce carbon losses.

MODIS EVI is an indicator of the greenness of grassland and can provide valuable information about vegetation productivity. The prediction of MODIS EVI from environmental variables such as solar radiation, air temperature, soil temperature, relative humidity, and vapor pressure deficit can aid in understanding the dynamics of grassland productivity and provide insights into sustainable management practices..

This project will provide an overview of the data science techniques used for prediction and an analysis of the results. The datasets for this study are provided by the United States Department of Agriculture (USDA)[1].

2. OBJECTIVES

The main objective of this practicum is to study the effect of environmental variables on methane turbulent flux, ecosystem respiration and MODIS EVI (greenness of grassland) to predict their values.

2.1. Technical Project Objectives

1. Understand the hypothesis and background of the study. Do required reading on agriculture and weather-related terminology within the project's scope.

[1] https://www.ars.usda.gov/

2. Analyze the available data to process critical information.
3. Visualizing the patterns and trends in the data to gain insights and improve model performance.
4. Evaluating and comparing the performance of various machine learning algorithms for predicting the values of methane turbulent flux, ecosystem respiration and MODIS EVI.

2.2. Individual Learning Objectives

1. Become more familiar with working on Agriculture and weather data.
2. Expand my knowledge of sustainable agriculture and environmental issues by reading relevant literature.
3. Acquire experience by working with real-world data to know how it affects farmers' day-to-day work life.
4. Carry out a complete modeling process, starting with the problem understanding through data analysis, model training, and testing, and ending with the model's validation.
5. Finally, to get experience with a professional team working towards making farmer's life better using technology, in my case data science.

3. DATA

3.1. DATA INGESTION

The data for this study was provided by the United States Department of Agriculture (USDA) and consists of three datasets:

1. Methane turbulent flux vs environment variables.

2. Ecosystem respiration (ER) vs environment variables.

3. MODIS EVI (greenness of grassland) vs Daily weather data.

Each dataset contains information on various environmental variables such as solar radiation, air temperature, soil temperature, relative humidity, and vapor pressure deficit.

The methane turbulent flux dataset includes measurements of methane emissions from grassland ecosystems. The dataset contains daily measurements of methane flux from 2020 to 2021, along with corresponding environmental variables.

The ecosystem respiration dataset includes measurements of ER from grassland ecosystems. The dataset contains daily measurements of ecosystem respiration for every 30-minute interval from 2019 to 2022, along with corresponding environmental variables.

The MODIS EVI dataset includes measurements of vegetation greenness from grassland ecosystems. The dataset contains 8- day composite satellite-derived vegetation greenness data for a native tallgrass prairie pasture (MODIS EVI) from 2000 to 2022, along with corresponding daily weather data.

3.2. DATA EXPLORATION AND CLEANING

In this section, we will explore the provided datasets to gain insights and understanding of the data. We will analyze the data to identify any missing values, outliers, and correlations between variables. All datasets were in csv format and were loaded using Python's panda's library. And all variables considered for analysis are numeric variables.

3.2.1. Methane Turbulent Flux Vs Environmental variables

The methane turbulent flux dataset contains measurements of methane flux in umol m-2 s-1 from a grassland site. The dataset has 13 variables including environmental variables such as solar radiation, relative humidity, air temperature, soil temperature and vapor pressure deficit. The dataset consists of 11759 observations collected in 2020 and 2021. Figure 3.2.1a shows the first few rows of dataset.

	Date_Time	Date	Time	Year	DoY	month	Hour	Rg_f	VPD_f	rH_f	Tair_f	Tsoil_f	Methane Turbulent Flux
0	5/6/2020 0:00	5/6/2020	0:00	2020	126	5	24.0	1.3	399.4	65.6	18.3	16.0	0.003
1	5/6/2020 0:30	5/6/2020	0:30	2020	127	5	0.5	1.3	587.9	64.9	18.4	15.6	0.023
2	5/6/2020 1:00	5/6/2020	1:00	2020	127	5	1.0	0.6	534.4	65.5	18.1	15.5	0.008
3	5/6/2020 1:30	5/6/2020	1:30	2020	127	5	1.5	0.5	483.3	68.0	17.5	14.8	0.005
4	5/6/2020 2:00	5/6/2020	2:00	2020	127	5	2.0	1.1	560.6	66.3	17.6	14.7	0.001

FIG 3.2.1a: Sample data of Methane turbulent flux

During the data exploration phase, no missing values were found in the dataset. And figure 3.2.1b illustrates how the data is distributed for environmental variables and methane in 2020 and 2021. For solar radiation (Rg_f), relative humidity (rH_f) and soil temperature (Tsoil_f), we can observe that the overall shape and distribution of the tips are similar for both time periods, but there are some outliers in the case soil temperature for 2020. For vapor pressure deficit (VPD_f), we see the largest difference in the shape of the distribution for both years. For methane turbulent flux, there are more outliers in the year 2021. We can also observe that in the year 2021 for both air temperature and soil temperature the data distribution is same. For better understanding of distribution, fig 3.2.1c shows the data variation in the year 2021 for both soil temperature and air temperature.

So, after interpreting these visualizations, only 2020 data has been considered for further analysis. Since 2021 data is abnormal and contain errors. So, the USDA team mentioned to proceed with 2020 data.

The final size of the dataset is 5328 records. After removing outliers, it is 5271 records.

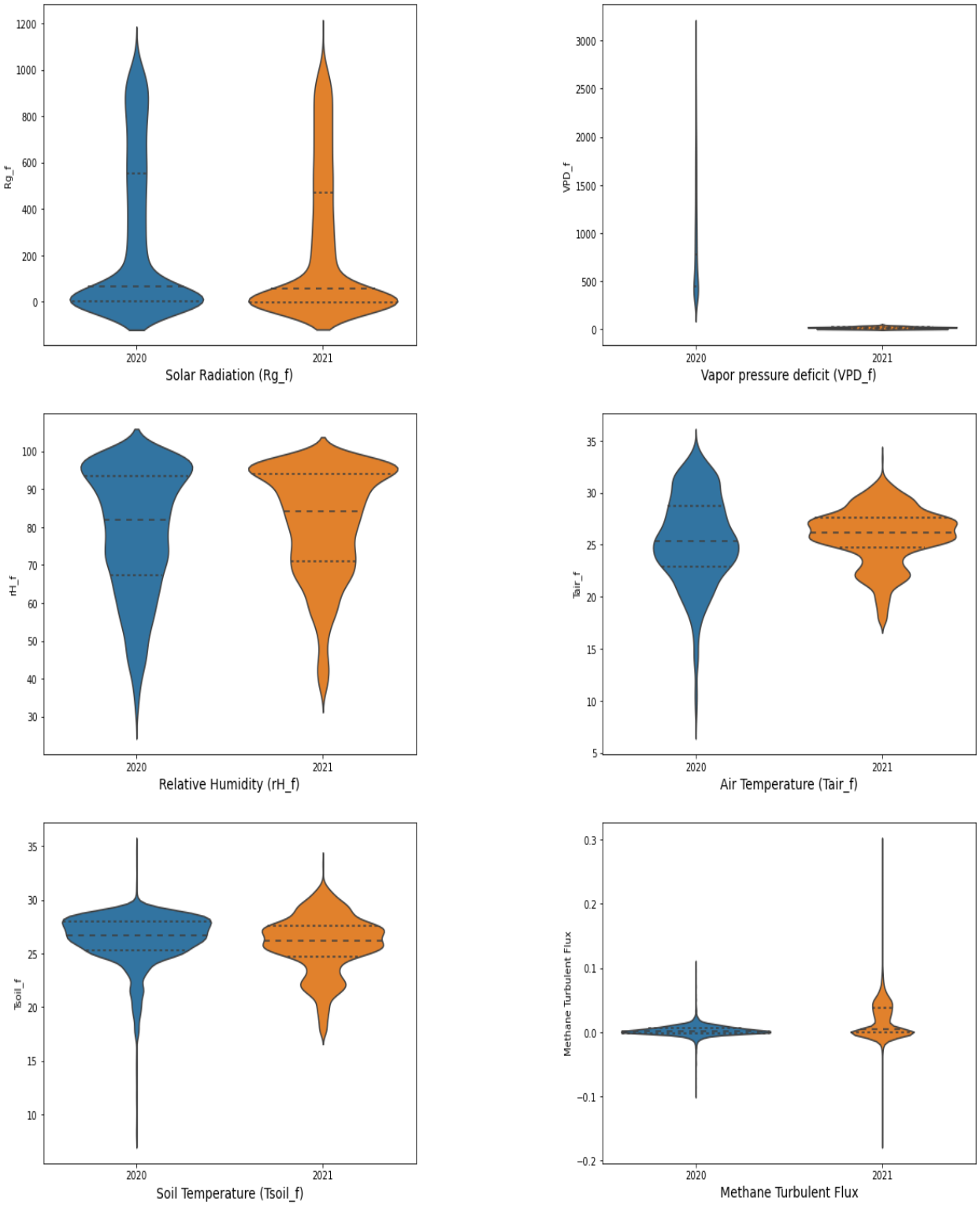


FIG 3.2.1b: Violin plot describing distribution of data for methane and environment variables.

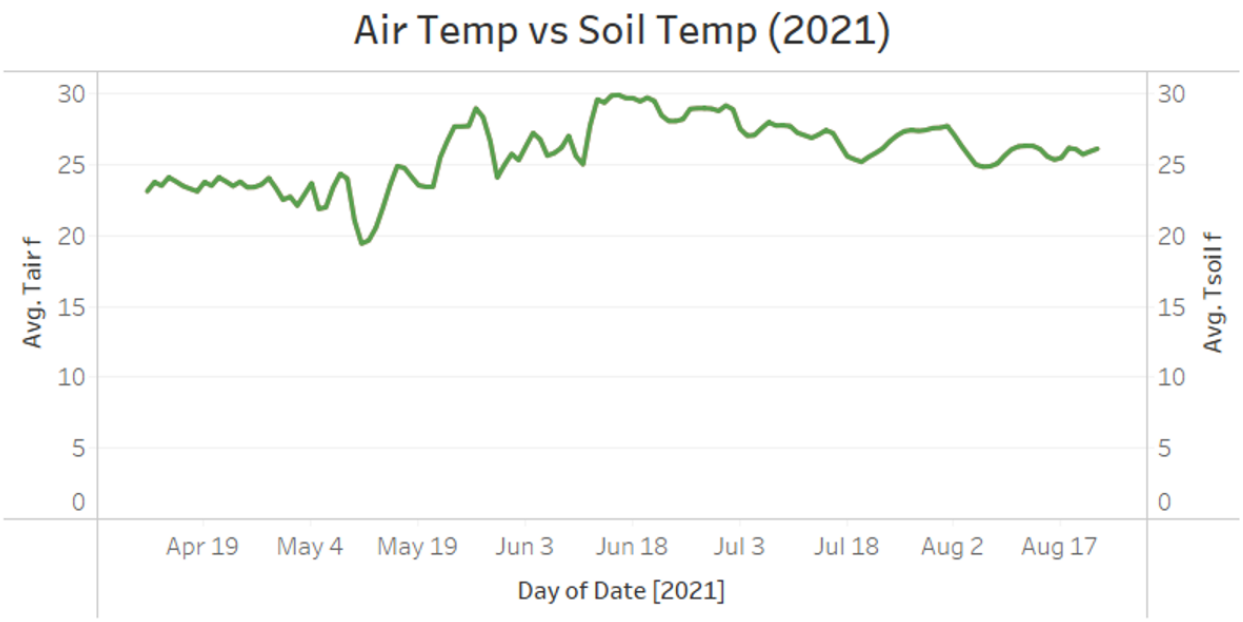


Fig 3.2.1c: Data varying in the year 2021 for soil and air temperature.

3.2.2. Ecosystem respiration (ER) vs Environment variables

The ER dataset contains measurements of ER value and environmental variables with 65807 records from 2019 to 2022.Environmental variables includes vapor pressure deficit (VPD in hPa), Turbulence (Ustar in ms-1), soil water content (SWC in m3 m-3). Figure 3.2.2a shows the first few rows of the dataset.

	Year	DoY	Hour	ER	Rg	Tair	Tsoil	rH	VPD	Ustar	SWC	Date
0	2019	91	0.5	1.16305	0.0	2.57	9.73	73.91	NaN	0.05	0.31	2019-04-01
1	2019	91	1.0	0.84226	0.0	2.56	9.39	73.33	NaN	0.03	0.31	2019-04-01
2	2019	91	1.5	1.84811	0.0	1.88	9.12	76.41	NaN	0.02	0.31	2019-04-01
3	2019	91	2.0	1.12012	0.0	1.78	8.83	74.55	NaN	0.03	0.31	2019-04-01
4	2019	91	2.5	1.25492	0.0	1.84	8.57	68.79	NaN	0.03	0.31	2019-04-01

Fig 3.2.2a: Sample data of Ecosystem Respiration dataset.

In the dataset, it is observed some of the attributes had missing values (refer fig 3.2.2b). The presence of missing values in the dataset is a common problem that needs to be addressed before building any predictive models. The missing values may arise due to various reasons such as errors during data collection, data processing, or data entry. In our dataset, we observed missing values in the attributes 'ER', 'Rg', 'Tsoil', 'Tair', 'rH', 'VPD', 'Ustar' and 'SWC'. To remove these missing values, we explored different strategies such as removing rows with a missing value in ‘ER’ attribute and removing rows with maximum Nan values in a row and imputing the missing

values with appropriate techniques such as mean and median. For normally distributed attribute replaced missing value with mean and skewed distributed with median.

Next, outliers were identified using box plots, pair plots and removed from the dataset. After cleaning, data size reduced to 29586 records.

```
ER_data.isna().sum()
```

```
Year         0
DoY          0
Hour         0
ER       33519
Rg       21473
Tair     20715
Tsoil    21500
rH       20718
VPD      29981
Ustar    27786
SWC      19899
Date         0
dtype: int64
```

Fig 3.2.2b: Missing values in each attribute in ER dataset.

Exploratory Data Analysis:

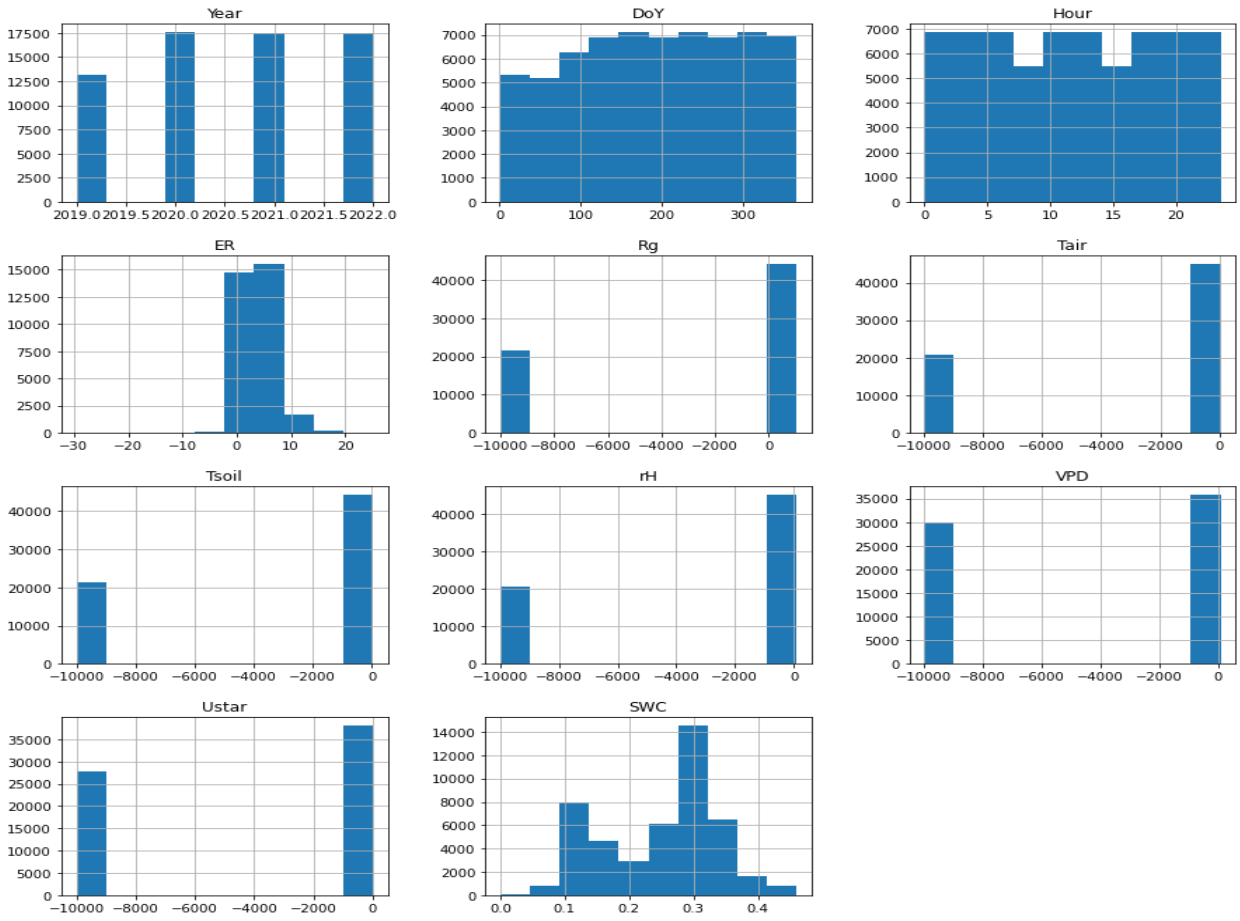


Fig 3.2.2c: Data distribution of every attribute in ER dataset.

From figure 3.2.2c, Shows the distribution of data in histogram for every attribute. Here, you can observe missing values (‘-9999’ value) which is far from most of the data distribution in attributes like ‘Rg’, ’Tair’, ‘Tsoil’, ‘rH’, ‘VPD’, and ‘Ustar’.

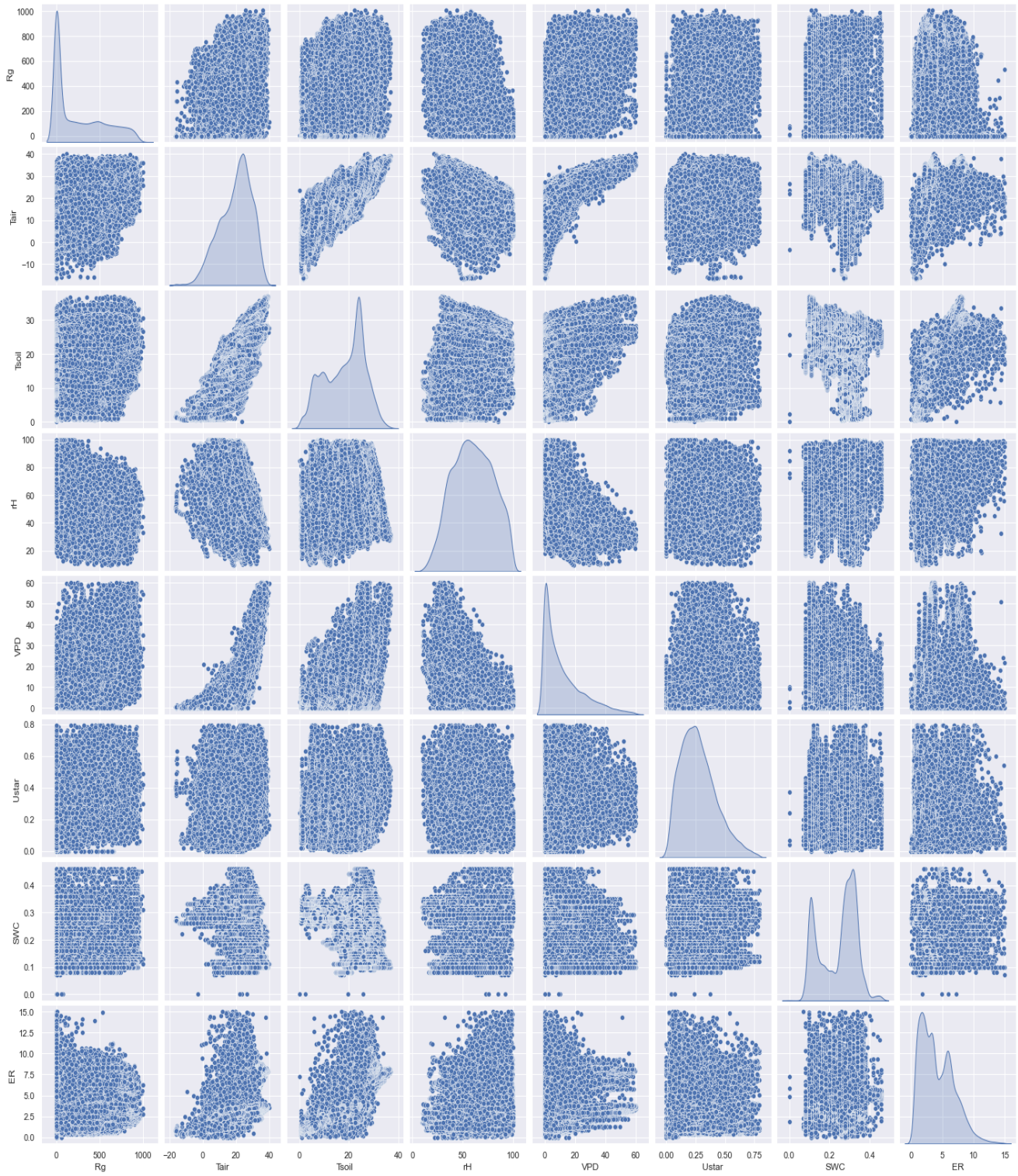


Fig 3.2.2d: Pair plot to understand the distribution between two variables.

Figure 3.2.d illustrates the pairs plot which builds on two basic figures, the density distribution, and the scatter plot. The density distribution on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship between two variables. From this plot, we found outliers. For example, in 'SWC' attribute and other attributes relation plots few data points are away from the group of datapoints which are identified as outliers.

Once the data is cleaned, the data was standardized to ensure that all variables have a comparable scale. This was done by centering and scaling the data such that each variable has zero mean and unit variance. Standardization facilitates the modeling process and improves the performance of some machine learning algorithms.

3.2.3. MODIS EVI (greenness of grassland) vs Daily weather data.

Firstly, MODIS (Moderate Resolution Imaging Spectroradiometer) is an instrument that measures the health and productivity of vegetation on a global scale. It is a remote sensing instrument that is mounted on two NASE Earth observing System (EOS) satellites – Terra and Aqua. The Enhanced Vegetation Index (EVI) is a vegetation index derived from the MODIS sensor data. The EVI is a measure of the greenness and vigor of vegetation, considering the influence of atmospheric conditions, such as aerosols and clouds.

Data Inspection:

The MODIS EVI dataset aimed to predict the MODIS EVI values based on weather data. In the MODIS EVI dataset (Fig 3.2.3a) weather data is collected on daily basis which consists of 8,402 observations along with 19 variables which is 2 decades worth of data. Whereas MODIS EVI data is of 8-day composite data which consists of 1046 rows and 2 columns.

From the Fig 3.2.3a the abbreviations of the variables are as follows: **TMAX** (maximum air temperature(F)), **TMIN** (minimum air temperature(F)), **TAVG** average air temperature(F)), **HAVG** (average relative humidity(%)), **VDEF** (average daily vapor deficit(mb)), **HDEG** (heating degree-days (65F standard)), **CDEG** (cooling degree-days (65F standard)), **WCMN** (minimum wind chill index temperature(F)), **WSPD** (average wind speed (mph)), **ATOT** (solar radiation (MJ/m2)), **RAIN** (daily rainfall (inch)), **SAVG** (average soil temperature 10cm under sod(F)), **BAVG** (average soil temperature 10cm under bare soil(F)), **TRO5** (soil moisture calibrated Delta-T at 5cm), **TR25** (soil moisture calibrated Delta-T at 25cm), **TR60** (soil moisture calibrated Delta-T at 60cm)

	YEAR	MONTH	DAY	TMAX	TMIN	TAVG	HAVG	VDEF	HDEG	CDEG	WCMN	WSPD	ATOT	RAIN	SAVG	BAVG	TR05	TR25	TR60
0	2000	1	1	69.04	38.58	53.69	60.74	6.76	11.19	0.0	31.63	15.82	11.45	0.00	45.16	46.20	1.8101	1.5840	1.5508
1	2000	1	2	60.58	31.12	48.24	58.04	5.47	19.15	0.0	25.76	6.21	12.16	0.00	47.04	48.35	1.8401	1.5840	1.5464
2	2000	1	3	55.06	26.92	40.14	86.71	1.39	24.01	0.0	12.03	13.38	8.88	0.11	45.64	45.87	1.8145	1.5792	1.5420
3	2000	1	4	44.77	17.64	29.75	64.11	2.28	33.80	0.0	6.09	11.32	12.77	0.01	41.85	39.53	1.7812	1.5782	1.5392
4	2000	1	5	53.69	26.63	38.42	48.98	4.76	24.84	0.0	15.09	15.63	12.38	0.00	40.03	38.82	1.7900	1.5799	1.5392
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8397	2022	12	28	70.36	39.86	53.52	45.95	8.38	9.89	0.0	29.75	22.30	8.33	0.00	40.68	34.44	1.5763	1.4643	1.4874
8398	2022	12	29	66.43	39.02	52.80	50.40	7.70	12.27	0.0	32.12	13.69	8.42	0.00	44.00	42.76	1.5718	1.4613	1.4876
8399	2022	12	30	54.61	24.61	38.87	76.54	2.48	25.39	0.0	16.57	6.11	11.48	0.00	42.53	41.94	1.5754	1.4620	1.4864
8400	2022	12	31	58.57	36.79	45.07	75.99	2.97	17.32	0.0	29.39	9.58	9.04	0.00	43.04	42.62	1.5773	1.4635	1.4880
8401	2023	1	1	71.74	37.75	50.95	69.97	5.39	10.26	0.0	34.38	5.69	9.25	0.00	44.06	44.35	1.5755	1.4606	1.4868

8402 rows × 19 columns

Fig 3.2.3a: Weather data

	Date	MODIS_EVI
0	2/26/2000	0.170
1	3/5/2000	0.183
2	3/13/2000	0.189
3	3/21/2000	0.205
4	3/29/2000	0.192
...	...	...
1041	10/16/2022	0.173
1042	10/24/2022	0.173
1043	11/1/2022	NaN
1044	11/9/2022	0.177
1045	11/17/2022	NaN

1046 rows × 2 columns

Fig 3.2.3b: MODIS EVI data

Data Cleaning:

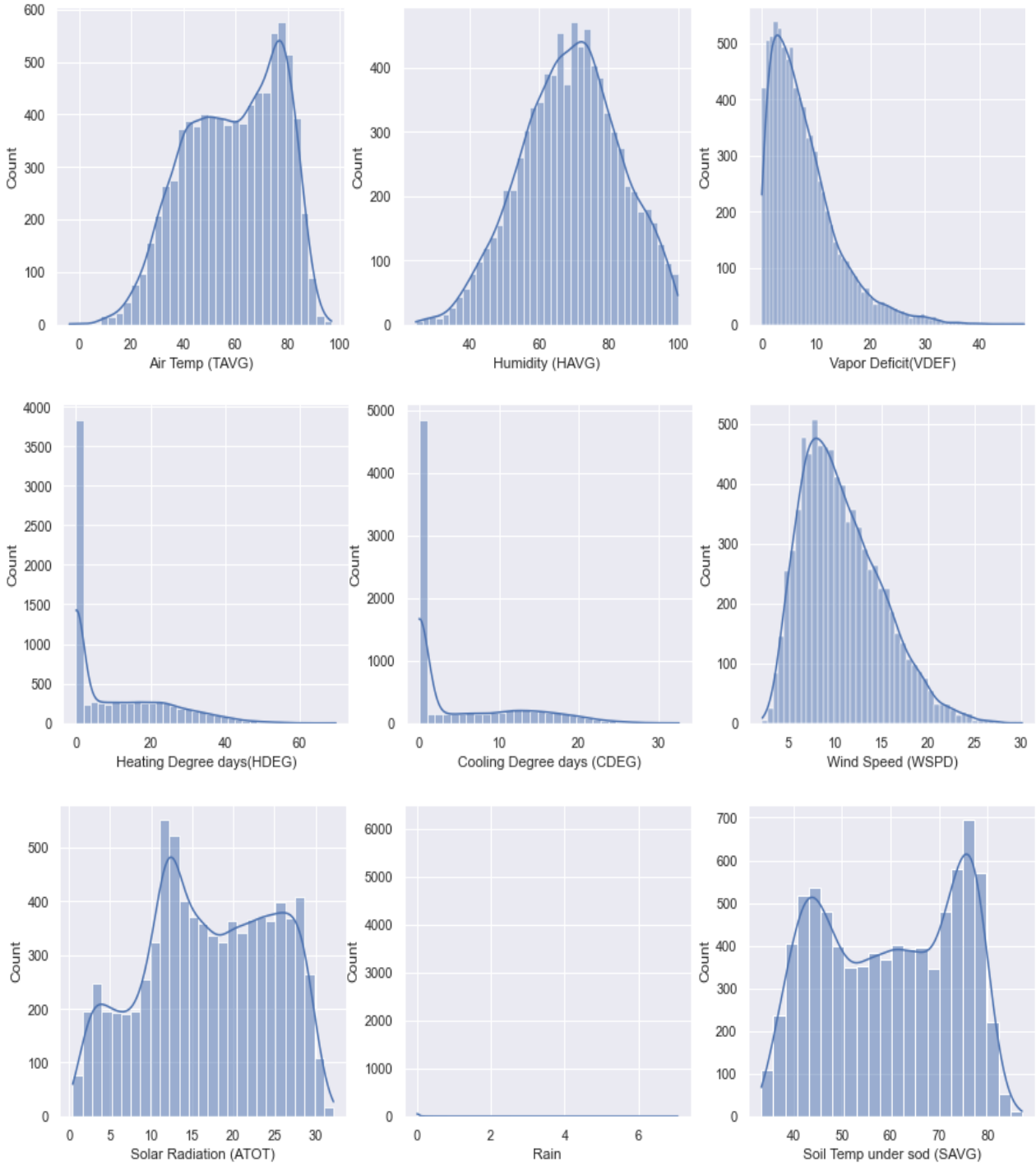
To begin with the YEAR, MONTH and DAY columns in the weather data are merged using a pandas datetime object in python. Then we dropped all the 3 columns YEAR, MONTH and DAY.

The total number of missing values in each column of weather data is listed in the Fig 3.2.3c

```
TMAX      79
TMIN      79
TAVG      79
HAVG      59
VDEF      87
HDEG      79
CDEG      79
WCMN    4060
WSPD     283
ATOT     297
RAIN      24
SAVG     110
BAVG     479
TR05     233
TR25     186
TR60     110
```

Fig 3.2.3c: Total number of missing values in each column.

As half of the values are missing in WCMN (minimum wind chill index temp(F)) that column is dropped. Along with the WCMN, the values of TAVG are like TMAX and TMIN. Therefore, only TMAX and TMIN are dropped. Moreover, Fig 3.2.3d illustrates how weather data is distributed.



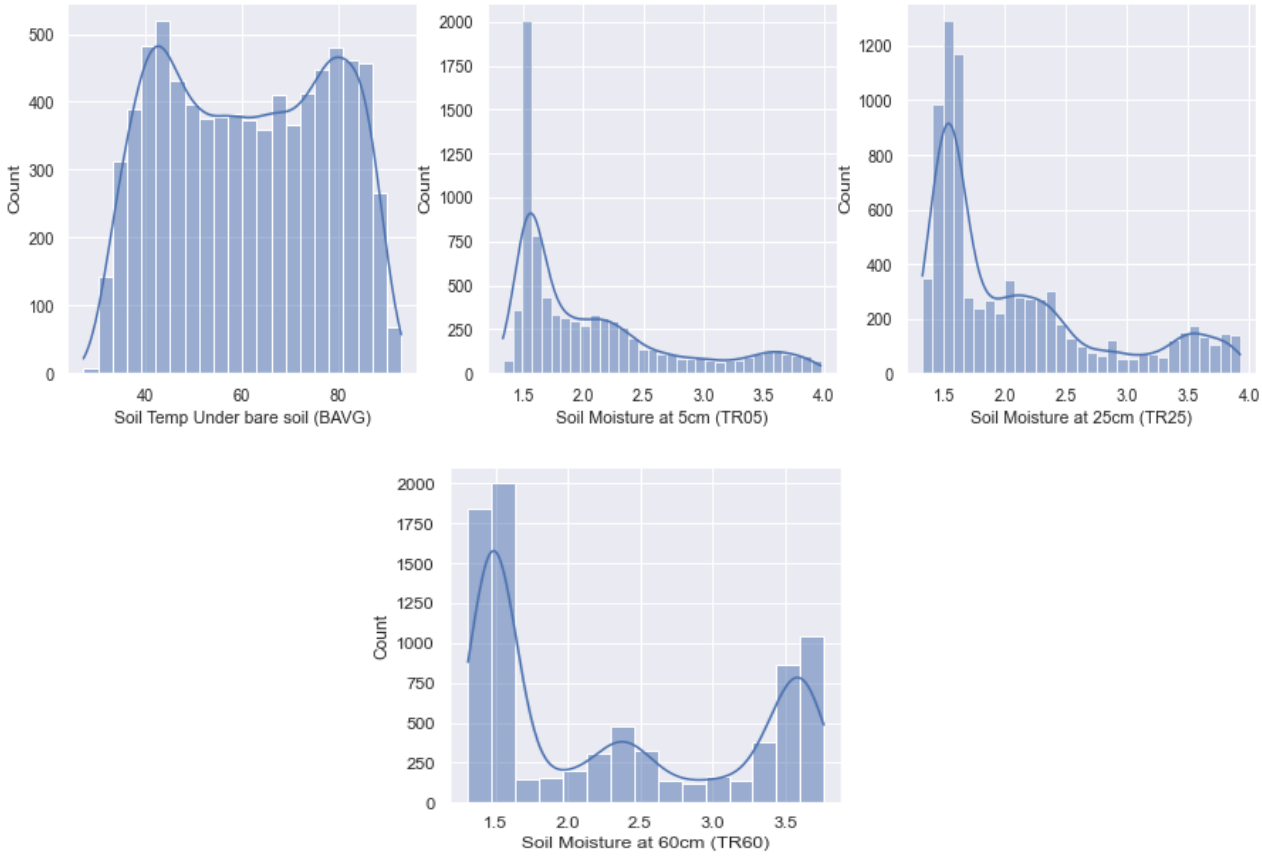


Fig 3.2.3d: Visualizations of raw weather data.

Missing value imputation:

From the above Fig 3.2.3d, the data is skewed for some variables such as VDEF, HDEG, CDEG, RAIN, TR05, TR25. There are several or large numbers of data points that act as outliers. Outlier's data points will have significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model. Hence, median value is imputed in place of missing values. For all the other variables whose data is normally distributed mean bases imputation is implemented in place of missing values.

Data integration:

The MODIS EVI data is an 8-day composite satellite-derived vegetation greenness data, whereas the weather data is collected daily. So, transformed weather data to a similar scale that is aggregating into 8-day averages for all variables but sum for rainfall. Finally, combining the MODIS EVI data with the weather data and dropped the records where MODIS EVI values are missing. The records have been dropped because the MODIS EVI data is satellite derived data and if we try to impute some other values in place of missing values, then there might be inaccurate predictions. Fig 3.2.3e represents the final data set after data pre-processing. The total number of rows and columns after data preparation are 862 and 15 respectively.

	DATE	Tavg_mean	Havg_mean	Vdef_mean	Hdeg_mean	Cdeg_mean	Wspd_mean	Atot_mean	Rain_sum	Savg_mean	Bavg_mean	Tr05_mean
0	2/26/2000	48.71250	66.61750	5.25500	16.66875	0.00000	12.51125	14.828610	0.78	49.21875	49.43125	1.671738
1	3/5/2000	50.03750	72.90625	4.11375	14.25750	0.00000	13.49750	17.770000	0.71	51.34375	51.30125	1.562450
2	3/13/2000	47.88250	81.58250	3.14125	15.82625	0.00000	10.42750	12.791250	0.95	50.05750	49.97750	1.594888
3	3/21/2000	57.95000	75.16625	5.15500	6.97000	0.00000	10.11750	16.667500	2.39	56.53125	56.36875	1.507513
4	3/29/2000	50.39000	75.42750	4.47000	15.15250	0.31125	11.22500	15.853750	1.31	53.35625	52.09125	1.512400
...	...	...	...	...	...	...	...	...	...	...	...	...
857	9/22/2022	70.36750	50.75500	15.16375	0.00000	5.49625	8.51375	18.623750	0.02	68.46875	75.99375	3.573925
858	9/30/2022	65.58125	48.60625	13.80000	1.37250	1.70125	6.01875	16.581250	0.05	65.18625	72.14250	3.616913
859	10/16/2022	59.23625	48.31250	11.17750	8.67125	3.12500	11.68000	14.632500	0.00	58.64625	61.48875	2.277975
860	10/24/2022	54.55625	70.89750	5.19500	9.21625	0.00000	9.23250	16.958877	2.35	57.97875	56.63875	1.569400
861	11/9/2022	41.83750	73.89750	2.92875	23.40375	0.93500	11.42000	16.958877	0.66	53.01625	49.11750	1.538763

862 rows × 15 columns

Tr25_mean	Tr60_mean	MODIS_EVI
1.570663	1.491400	0.170
1.557400	1.478662	0.183
1.547525	1.471650	0.189
1.531750	1.465475	0.205
1.518875	1.462813	0.192
...	...	...
3.598238	3.652275	0.270
3.629125	3.671975	0.237
3.617062	3.687700	0.173
2.167937	3.687425	0.173
1.451175	3.677050	0.177

Fig 3.2.3e: Final dataset after pre-processing.

4. METHODOLOGY

4.1. Methane turbulent flux vs environment variables.

(I) Techniques

Supervised Learning:

In Methane dataset we are trying to predict the Methane turbulent flux which is a continuous time-series value. This falls into regression techniques under supervised learning. As this is time-series data some of the algorithms we selected are autoregressive integrated moving average (ARIMA) and prophet. Time-series models can be used to forecast future values of a variable based on historical data.

(ii) Procedure

Model training: After the data is preprocessed model is selected, it can be trained on the prepared data. This involves splitting the data into training and validation sets, using the training set to train the model and the validation set to assess its performance. 80% of data is used for training and 20% of data is used for validation.

Evaluation metrics:

Mean Square Error (MSE) - It is a measure of the average squared distance between the predicted and actual values, and it gives more weight to larger errors.

Mean Absolute Error (MAE) - It is a measure of the average absolute distance between the predicted and actual values, and it gives equal weight to all errors.

Root Mean Square Error (RMSE) - RMSE is the square root of the average of the squared differences between the predicted and actual values of the target variable.

R-Squared - R-squared is a statistical measure that represents the proportion of the variance in the target variable that is explained by the independent variables in the model. It is a measure of how well the model fits the data, and it ranges from 0 to 1. A higher R-squared value indicates a better fit of the model to the data.

4.2. Ecosystem respiration (ER) vs environment variables.

(I) Techniques

Supervised Learning:
The variable we are attempting to forecast in the ER dataset is ER value, which is a continuous value. When we want to predict continuous values, we typically use regression techniques in supervised learning. Regression techniques are used to predict a continuous target variable based on one or more input variables. Some of the regression techniques we have selected to predict the ER value are KNN regressor, Support Vector Regressor (SVR), Random Forest and Decision Tree regressor.

(II) Procedure

Model training: After preprocessing, the data is divided into training and testing datasets. The ER dataset uses 80% of the data for training and the remaining 20% for assessing the model's performance on unobserved data. Multiple regression algorithms are used to train the training dataset and then the trained model is evaluated on test dataset. Before tuning the accuracy of Random Forest Regressor was 63%.

Model tuning: Here are some hyperparameters that are used to tune the Random Forest Regressor using GridSearchCV approach. After implementing the hyperparameter tuning the accuracy of the Random Forest Regressor has increased to 80%.

'max_depth': [10, 20, 30, 50],
'n_estimators': [100, 500, 1000,1500],

'bootstrap': [True, False],
'min_samples_split': [2, 5, 10, 12]

Evaluation metrics:

MSE, MAE, RMSE, and R-Squared are metrics used to evaluate the performance of the model.

4.3. MODIS EVI (greenness of grassland) vs Daily weather data.

(I) Techniques

Supervised Learning:

In the MODIS EVI dataset, the variable we are trying to predict is MODIS EVI which is a continuous value. Therefore, supervised learning techniques are used when we have labeled data and we want to predict the output label for new, unseen input data. Regression is a type of supervised learning technique used for predicting continuous output values based on a set of input features. Various regression models like Random Forest, Linear Regression, Support Vector Regressor (SVR), Decision Tree Regressor and KNN regressor are chosen to find the best fit line or curve that can capture the relationship between the dependent variable and independent variables and use this relationship to predict the values of the dependent variable for new or unseen data.

(ii) Procedure

Model training: Once the data is pre-processed, it is split into training and testing datasets. For the MODIS EVI dataset 80% of data is used for training and the remaining 20% for evaluating the performance of the model on unseen data. Then the selected regression model is trained on the training dataset. This involves fitting the model to the training data and adjusting the model parameters to minimize prediction error. The trained model is then evaluated on the testing dataset to determine its performance in making predictions on new data. Random Forest regressor is very good at predicting compared to other regression models with 80% accuracy.

Model tuning: The hyperparameters in Random Forest Regressor can be tuned to improve the performance of the model. To tune the hyperparameters in Random Forest Regressor, we used Grid Search technique. GridSearchCV involves specifying a grid of hyperparameter values and evaluating the model's performance for each combination of values. Cross-validation is typically used to evaluate the performance of the model during the tuning process. Here are some of the hyperparameters that are used to tune Random Forest Regressor:

'n_estimators': [50, 100, 200, 300, 400, 500, 600, 700, 800],
'max_depth': [None, 5, 10, 20, 30, 40, 50, 60],
'min_samples_split': [2, 5, 10, 15, 20],
'min_samples_leaf': [1, 2, 4, 6, 8]

Evaluation metrics:

MSE, MAE, RMSE, and R-Squared are metrics used to evaluate the performance of the model.

5. RESULTS AND ANALYSIS

5.1. Methane turbulent flux vs environment variables.

Summary Statistics of Methane dataset

	Year	DoY	month	Hour	Rg_f	VPD_f	rH_f	Tair_f	Tsoil_f	Methane Turbulent Flux
count	5328.0	5328.000000	5328.000000	5328.00000	5328.000000	5328.000000	5328.000000	5328.000000	5328.000000	5328.000000
mean	2020.0	181.979167	6.477102	12.25000	274.700544	968.943994	78.830800	25.474343	26.269482	0.003023
std	0.0	32.044965	1.072832	6.92735	336.379765	606.040249	16.433958	4.313989	2.492104	0.009422
min	2020.0	126.000000	5.000000	0.50000	0.000000	300.700000	30.100000	7.900000	7.800000	-0.098000
25%	2020.0	154.000000	6.000000	6.37500	0.900000	443.500000	67.400000	22.900000	25.300000	-0.001000
50%	2020.0	182.000000	6.000000	12.25000	65.450000	778.850000	82.000000	25.400000	26.700000	0.002000
75%	2020.0	210.000000	7.000000	18.12500	551.000000	1384.825000	93.500000	28.700000	28.000000	0.007000
max	2020.0	237.000000	8.000000	24.00000	1064.500000	2995.900000	100.000000	34.600000	34.900000	0.108000

Correlation Plot

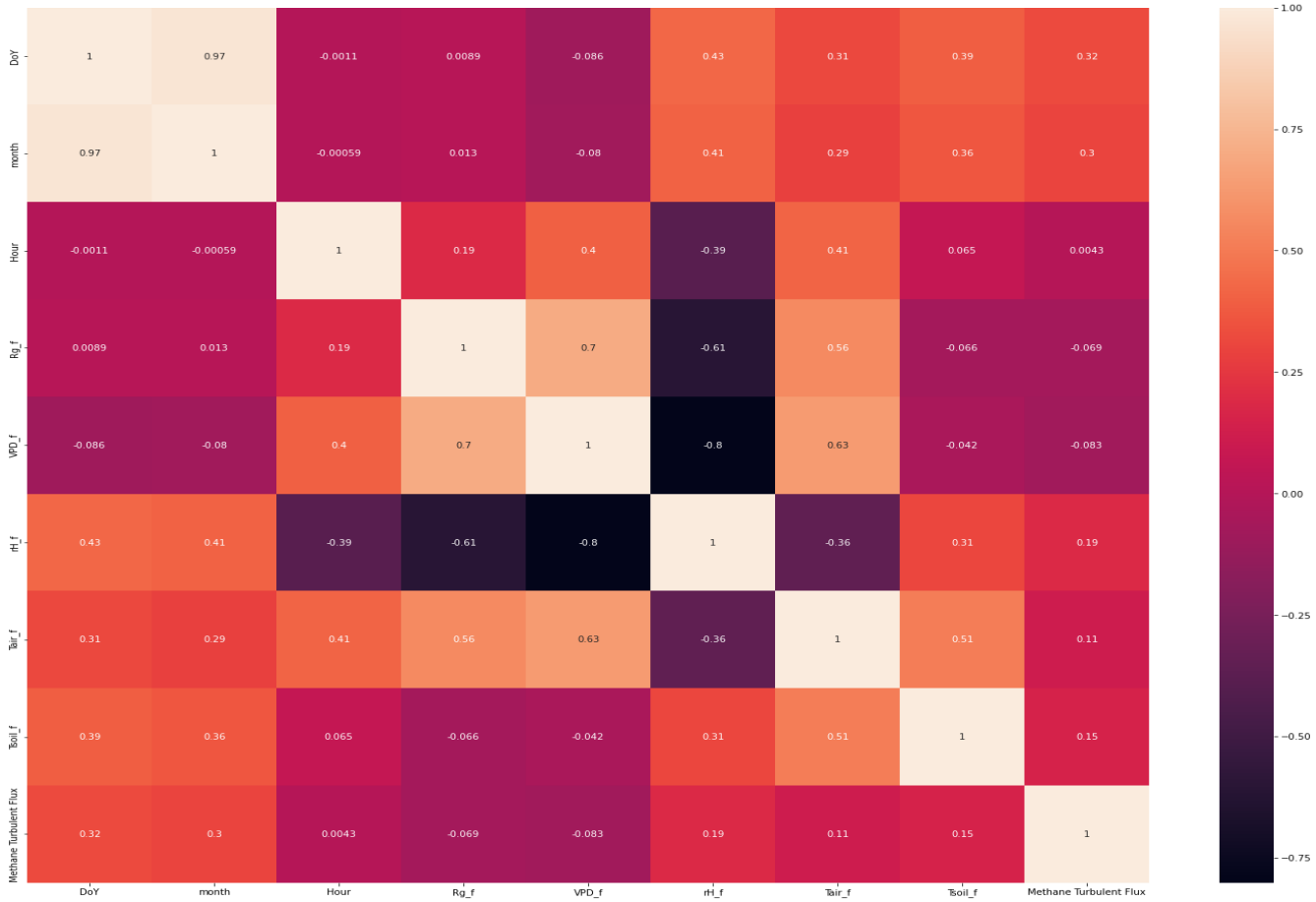


Fig 5.3.1a: Heat map of Methane dataset.

In Fig 5.3.1a shows that environment variables do not have much correlation with the Methane Turbulent Flux.

Model Results

From the following table we can see that MSE is very low. However, the Root Mean Squared Error (RMSE) is relatively high, which suggests that the errors may be larger relative to the scale of the data. The negative R-squared value indicates that models are performing worse than the mean of the data, which suggests that the model is not explaining much of the variation in the data.
A low MSE and high RMSE with a negative R-squared value indicate that the model is making accurate predictions on average but is not capturing the overall trends or patterns in the data.

MODEL	MSE	RMSE	R-SQUARED
Auto regressive Integrated Moving Average (ARIMA)	0.00006	0.007	-0.15
Prophet	0.00009	0.007	-0.74

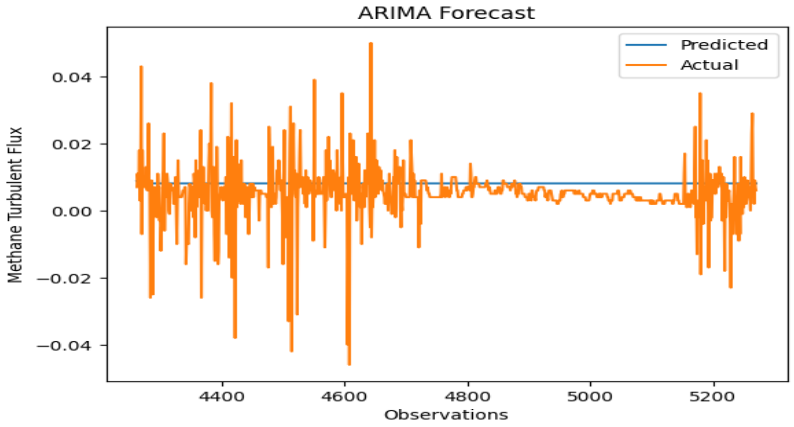


Fig 5.3.1b ARIMA forecast for actual vs predicted labels.

Fig 5.3.1b shows how well the ARIMA model fits the time series data over the entire range of data. The predicted values are plotted against the actual values of the time series. From the Fig 5.3.1b we can see that predicted values of ARIMA model do not closely follow the actual data. As the predicted values deviate significantly from the actual data, it indicates that the model is not capturing all the important information in the data.

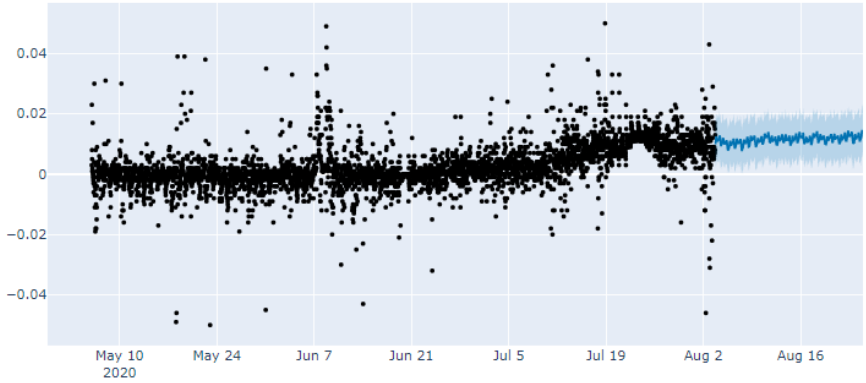


Fig 5.3.1c Prophet forecast for actual vs predicted labels.

Fig 5.3.1c typically shows the actual time series data as a black line, and the forecasted values as a blue line. The uncertainty around the forecasted values is represented by shaded regions, which show the upper and lower bounds of the prediction intervals. From Fig 5.3.1c, the prophet plot does not fit actual data well.

5.2. Ecosystem respiration (ER) vs Environment variables.

Summary Statistics of pre-processed ER dataset.

	Rg	Tair	Tsoil	rH	VPD	Ustar	SWC	ER
count	29586.000000	29586.000000	29586.000000	29586.000000	29586.000000	29586.000000	29586.000000	29586.000000
mean	256.852298	19.632038	18.371029	59.768618	12.071701	0.272822	0.243969	4.166194
std	292.266593	9.496737	7.892328	19.566660	12.986571	0.152384	0.088659	2.619989
min	0.000000	-16.660000	0.710000	9.600000	0.000000	0.000000	0.070000	0.000000
25%	0.000000	13.110000	11.700000	44.850000	1.730000	0.160000	0.160000	1.982900
50%	126.925000	21.010000	19.740000	59.768618	7.550000	0.250000	0.270000	3.600000
75%	485.960000	26.687500	24.390000	74.990000	18.557500	0.360000	0.310000	5.994375
max	1006.340000	40.270000	36.950000	100.000000	59.990000	0.790000	0.460000	14.980000

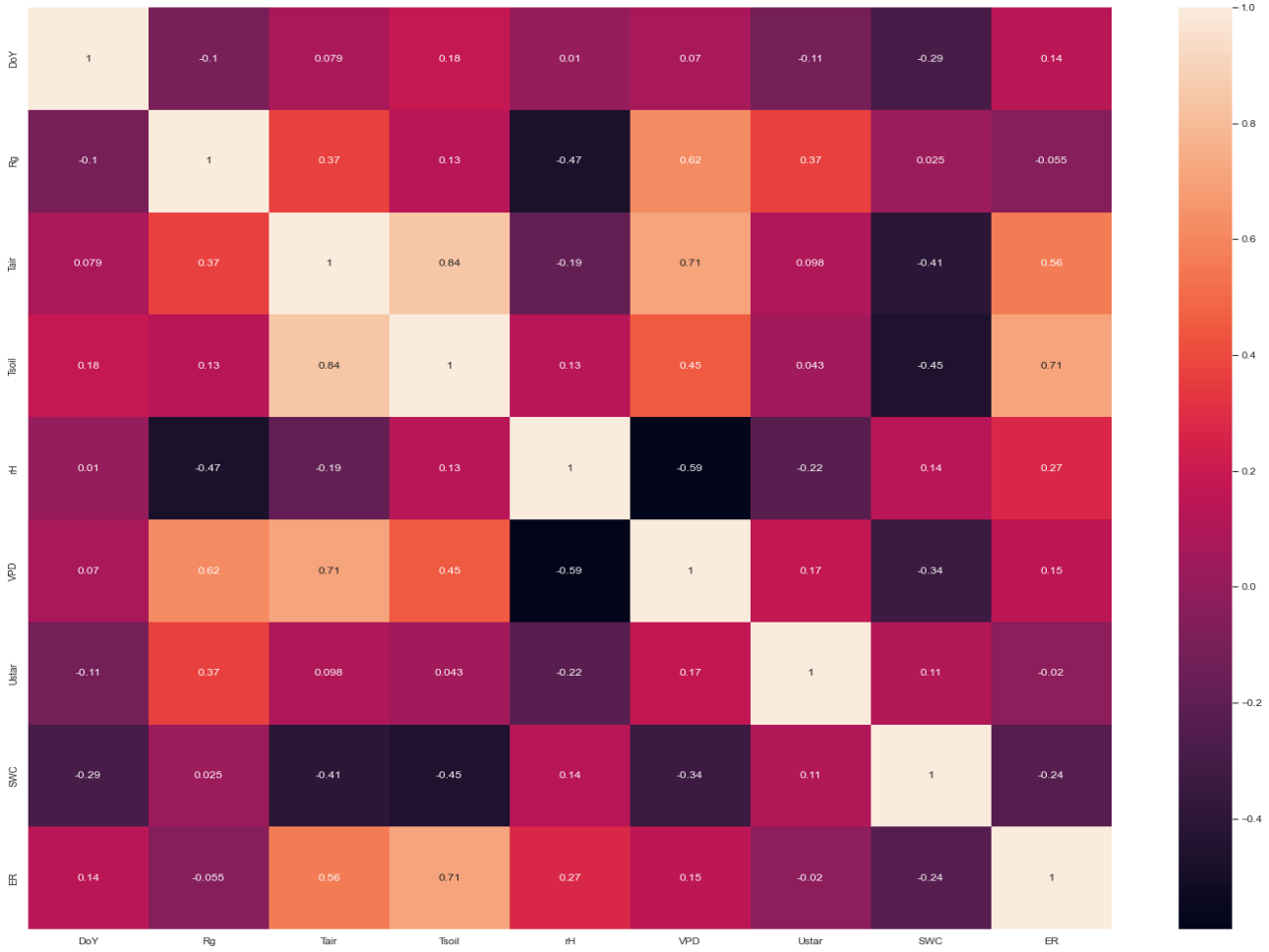


Fig 5.2a: Heat map of pre-processed ER dataset.

Figure 5.2a is the heatmap to visualize the correlation between input variables and target variable (in our case 'ER') in the dataset. We can see that the soil temperature and ER have positive correlation, whereas other variables have less correlation. Here, 'Rg', 'Ustar' and 'SWC' have weak negative correlation with target variable 'ER'.

Model Results:

Table 5.2a: Model Evaluating Results

MODEL	MSE	RMSE	R-SQUARED
Random Forest Regressor	1.388	1.178	0.63
KNN	1.646	1.283	0.763
Support Vector Regressor (SVR)	2.129	1.459	0.694
Decision Trees	2.627	1.620	0.623
Random Forest Regressor (with tuning)	1.387	1.177	0.80

In Table 5.2a, performance of each model on the ER dataset is shown. Here, Random Forest Regressor after hyper-parameter tuning got the satisfied results compared to other models. Best parameters are {max_depth=30, min_samples_split=5, n_estimators=1500}. A RMSE of 1.17 means that on average, the model's predictions are off by approximately 1.17 units from the actual values. A R2 score of 0.80 indicates that 80% of the variability in the target variable can be explained by the model's input features.

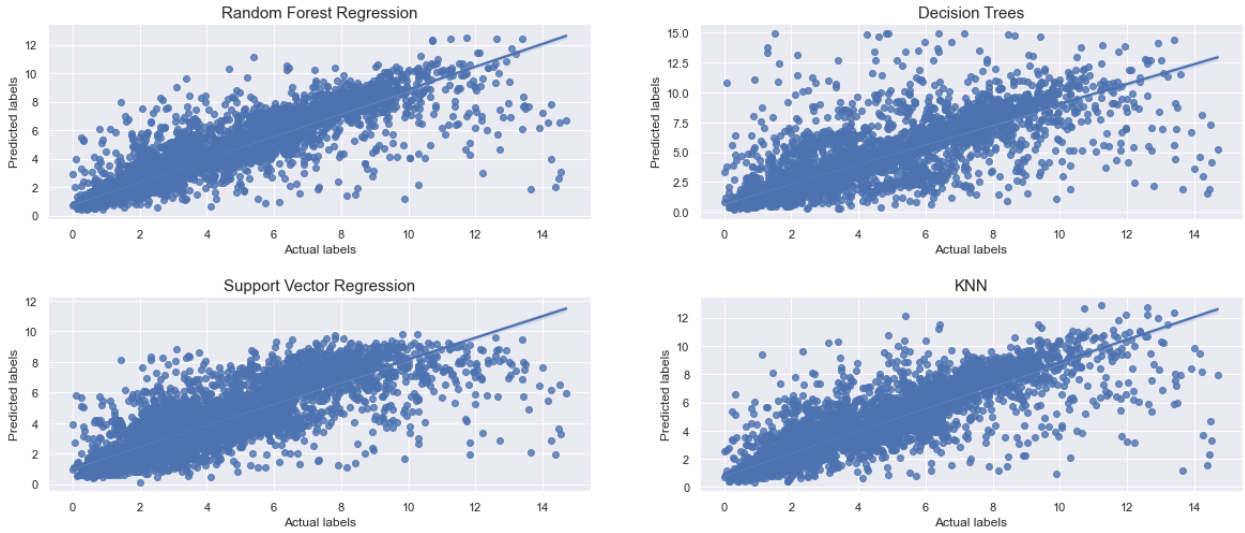


Fig 5.2b plots representing actual vs predicted labels for each model.

Figure 5.2b shows the actual vs predicted plot for ER data, the x-axis represents the actual values of ER, while the y-axis represents the predicted values of ER by the model. The subplot shows how well the model can predict the actual values of ER. Ideally, the points on the plot should fall

along a diagonal line, indicating that the predicted values are close to the actual values. If the points on the plot are scattered around the diagonal line, it indicates that the model is not able to accurately predict the actual values. The distance between the points and the diagonal line represents the magnitude of the prediction error. A smaller distance indicates a lower prediction error and better model performance. Here in the first subplot, we can see that the random forest regressor model is able to predict the actual values of ER with reasonable accuracy, as most points fall close to the diagonal line. However, there are some points that are scattered away from the diagonal line, indicating that the model has some level of prediction error.

5.3. MODIS EVI (greenness of grassland) vs Daily weather data.

Summary Statistics of pre-processed MODIS EVI dataset

	Tavg_mean	Havg_mean	Vdef_mean	Hdeg_mean	Cdeg_mean	Wspd_mean	Atot_mean	Rain_sum	Savg_mean	Bavg_mean	Tr05_mean
count	862.000000	862.000000	862.000000	862.000000	862.000000	862.000000	862.000000	862.000000	862.000000	862.000000	862.000000
mean	61.552654	68.212281	8.426435	8.801964	5.642329	10.749356	17.873294	0.719733	61.769991	63.597185	2.142651
std	15.886086	9.789248	5.568940	10.531065	6.724976	2.281442	5.875566	1.116422	12.966244	15.447726	0.653967
min	24.102000	32.486250	0.532500	0.000000	0.000000	4.872500	4.595000	0.000000	35.877500	32.848750	1.344362
25%	48.945937	62.416875	4.559688	0.000000	0.000000	9.121368	12.708332	0.010000	50.334063	51.008437	1.609341
50%	62.916250	69.105000	7.032500	3.706875	1.987500	10.665500	18.132985	0.290000	63.413125	63.714375	1.921284
75%	76.077500	75.039250	10.709375	15.726875	11.458750	12.283521	22.683125	0.927500	73.729062	77.290937	2.482747
max	92.728750	93.836250	40.068750	40.466000	27.852500	19.828750	29.428750	8.570000	85.607500	91.628750	3.949313

Tr25_mean	Tr60_mean	MODIS_EVI
862.000000	862.000000	862.000000
2.125261	2.320384	0.324304
0.719747	0.874108	0.136696
1.327800	1.313862	0.096000
1.560119	1.491003	0.201000
1.864194	2.105338	0.300000
2.449122	3.348525	0.440000
3.916762	3.760875	0.654000

Correlation Plot

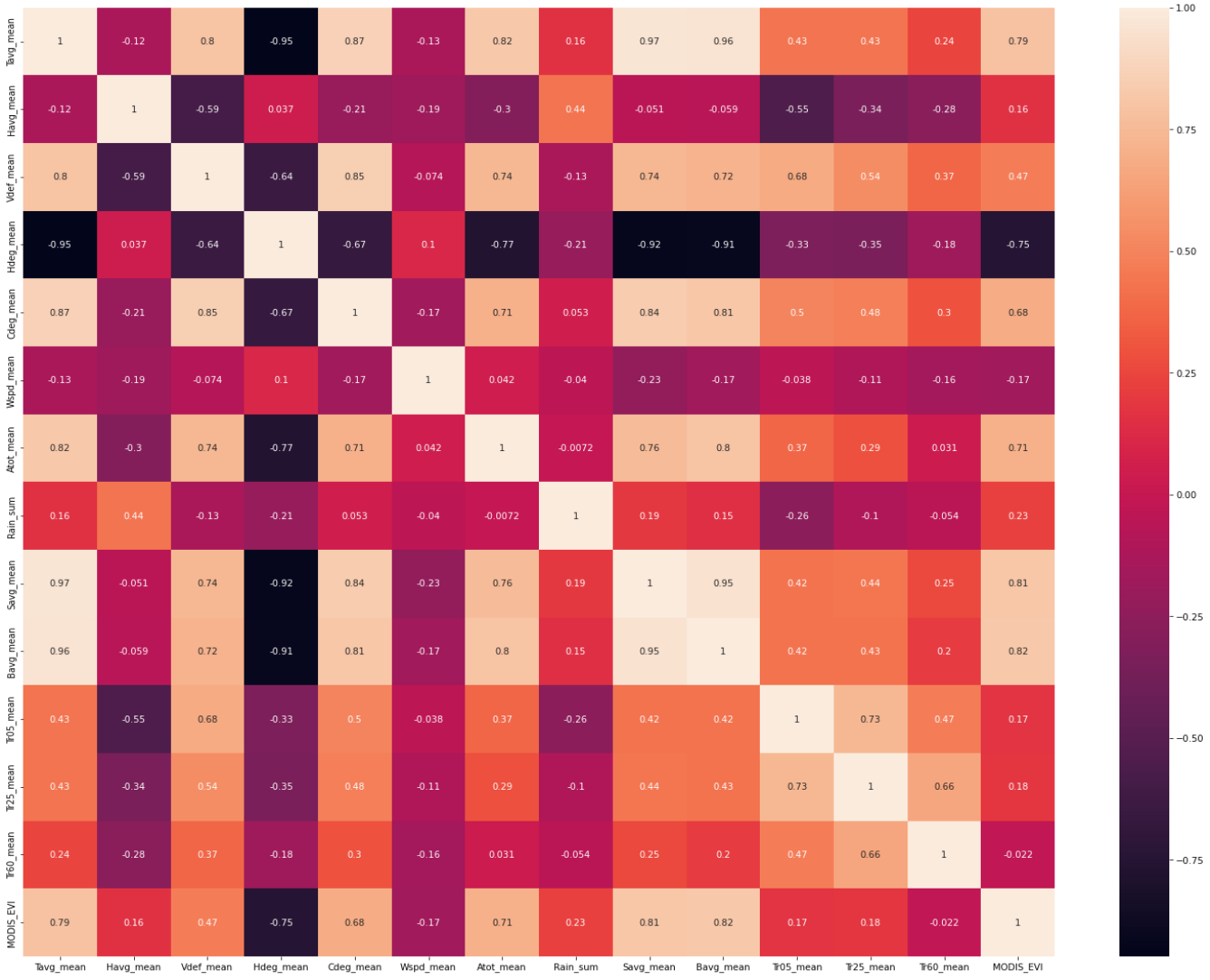


Fig 5.3.3a: Heat map of preprocessed MODIS EVI dataset.

The heatmap from Fig 5.3.3a provides insights about the relationship between two variables. It can be seen how each variable is correlated with all the other attributes. Here, only 3 variables (Hdeg_mean, Wspd_mean, and Tr60_mean) are negatively correlated. Whereas all the other attributes have a positive correlation with MODIS EVI.

Model Results

The following table explains the performance of each model. Moreover, Random Forest Regressor is performing great when compared to all the other models. If a model has less MAE, MSE, and RMSE but more R-squared, it means that the model is performing well in terms of its predictive accuracy and its ability to explain the variance in the dependent variable. In addition to this, low MAE, MSE, RMSE also indicates that the model is making more accurate predictions, while a higher R-squared indicates that the model is a good fit for the data.

MODEL	MSE	MAE	RMSE	R-SQUARED
Random Forest Regression	0.002	0.04	0.05	0.83
KNN	0.003	0.04	0.05	0.81
Linear Regression	0.003	0.04	0.06	0.78
Support Vector Regressor (SVR)	0.003	0.04	0.06	0.77
Decision Trees	0.005	0.57	0.07	0.68

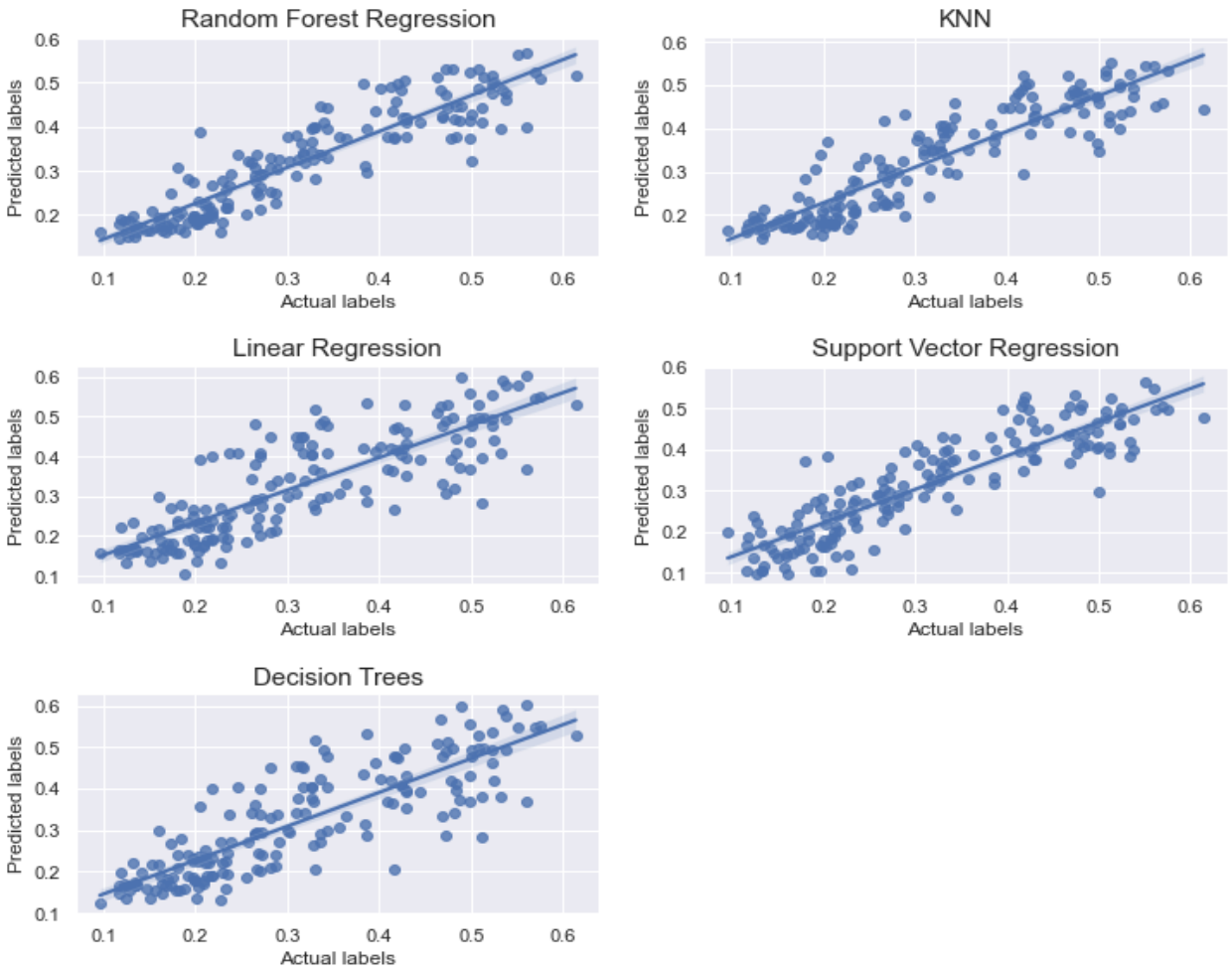


Fig 5.3.3b plots representing actual vs predicted labels.

The resulting plot Fig 5.3.3b demonstrates how well the model is predicting the output values. If the points on the plots are scattered around a diagonal line, it indicates that the model is performing good at predicting the output values. On the other hand, if the points are scattered randomly, it indicates that the model is not performing well. Therefore, for Random Forest regressor from Fig 5.3.3b the points are scattered around the diagonal line which means model is making accurate predictions compared to all the other models. The diagonal line (line of perfect fit) represents perfect predictions, where the predicted values are exactly equal to actual labels.

7. DELIVERABLES

We successfully predicted the ecosystem respiration and MODIS EVI from the available environmental variables and weather data. Our analysis reveals some interesting insights into the impact of weather data and environmental data on MODIS EVI and ER data. After getting access to the USDA data, data cleaning and pre-processing is performed to prepare the data for both datasets. Once the data is prepared, exploratory data analysis is performed on the processed data to observe and capture the patterns. Followed by machine learning models are applied on training data and evaluated it on unseen test data. Based on the results obtained, the model is performing well with decent accuracy on unseen data which can be a good resolution to the problem. The results of our analysis also suggest that there is a significant positive correlation between temperature and MODIS EVI in grasslands. Similarly for the ER dataset the soil temperature is much correlated with ER data.

Moreover, for methane dataset we applied data normalization, data standardization and regression techniques apart from time-series algorithms. But we could not find decent results from our analysis. However, we did gain valuable experience working with complex and diverse datasets, as well as developing and testing models to predict ecological phenomena.

We believe that our findings in this project could have important implications for the agricultural and ecological research community, and we plan to share our insights and methodology with other researchers in the field. To take this research further, we can explore additional factors that may influence methane flux, ecosystem respiration and MODIS EVI such as soil quality, crop type, precipitation, water quality.

8. REFERENCES

1. M. A. K. Khalil, R. A. Rasmussen, M. J. Shearer, R. W. Dalluge, Lixin Ren, Chang-Lin Duan (1998). Factors affecting methane emissions from rice fields. https://doi.org/10.1029/98JD01115
2. Elaine Wheaton and Suren Kulshreshtha (2017). Environmental Sustainability of Agriculture Stressed by Changing Extremes of Drought and Excess Moisture: A Conceptual Review. https://doi.org/10.3390/su9060970.
3. Munkhnasan Lamchin (2018). Long-term trend and correlation between vegetation greenness and climate variables in Asia based on satellite data. https://doi.org/10.1016/j.scitotenv.2017.09.145
4. Cai Y, Zheng W, Zhang X, Zhangzhong L, Xue X (2019) Research on soil moisture prediction model based on deep learning. PLoS ONE 14(4): e0214508. https://doi.org/10.1371/journal.pone.0214508.
5. Time Series Analysis in Python - A Comprehensive Guide with Examples (2019). URL: https://www.machinelearningplus.com/time-series/time-series-analysis-python/.
6. ARIMA vs Prophet vs LSTM for Time Series Prediction (2023).

URL: https://neptune.ai/blog/arima-vs-prophet-vs-lstm#:~:text=We%20see%20that%20ARIMA%20yields,seen%20in%20the%20following%20images.

9. SELF-ASSESSMENT

In this project, I became familiar with working on a real-world problem in sustainable agriculture. I also feel that I have honed my skills in data exploration, data cleaning, pre-processing, data modelling and as well as in visualizing and interpreting datasets using various techniques such as box plots and correlation matrices. However, one area where I could have improved is Time series techniques. Though, I was able to successfully implement regression models to predict methane emissions and ecosystem respiration, I struggled with applying time series techniques on methane data. Moving forward, I plan to dedicate more time to learning and practicing these techniques.