Kevin Matthew Duyst

ENV-745

Final Project

20 December 2022

Forecasting Methane Flux (FCH_4) Measurements: OLS Regression vs Machine Learning

Project Summary:

I aim to evaluate and process eddy covariance and flux variance data (specifically atmospheric CH_4 and CO_2) to better understand the ecosystem processes of a peatland site in Marcell Experimental Forest, Bog Lake Fen. Peatlands serve an integral role in the global biogeochemical cycles – they're natural sources of two leading greenhouse gases, CO_2 and CH_4. These sites are susceptible to warming temperatures: peat in this northern region thaws and causes natural vegetation to drown and lakes to brown with solid organic matter that dissolve into the water. This results in vast quantities of CO_2 and CH_4 emissions being released into aquatic bodies and the atmosphere. The overall objective of this study is to analyze sources and responses within aquatic ecosystems and feedbacks with the atmosphere through CO_2 and CH_4 flux measurements.

We created various forecasting models of CH_4 flux measurements with simultaneous observations of CO_2 flux values, precipitation values, and temperature values (atmospheric and 10CM soil readings). Multiple time series of our data were produced to showcase trends over the years (2009 – 2021) and the influence of seasonality for increasing emissions. Ultimately, we were testing and comparing various models (traditional statistical methods like multiple linear regression versus machine learning methods like Microsoft Light Gradient Boosting Machine, SARIMAX, Facebook Prophet, and Xgboost) and their accuracy assessments for future flux measurements. Univariate models (just looking at FCH_4 or just looking at FCO_2) and multivariate models (the incorporation of correlated predictor variables) were compared. We were curious what variables garner the largest influence over CH_4 emissions, and if those variables could be integrated into a model that accurately forecasts atmospheric CH_4 emissions up to 95% confidence.

CH_4 flux values were assigned the role of our indicator variable. Our predictor variables for the model (in order of influence) were as follows: Soil Temperature, ATM Temperature, CO_2 flux values, and Precipitation. We used data captured over the years of 2009 – 2019 as testing values for our model and created different periods for forecasting (one-year into the future, two-year predictions, and so forth). We then compared the accuracy of these forecasts with the actual values of flux measurements.

Research Design:

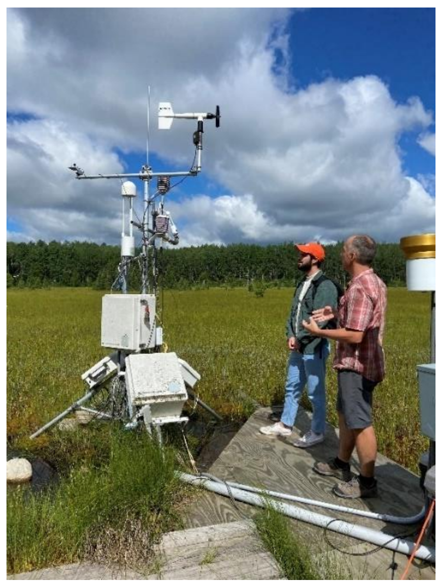
Three micrometeorological-towers have been deployed above Bog Lake Fen during the ice-free period, along with several other *in situ* sensors at the lake's surface (1-2 meters in depth). Thermistor chains extending the surface to the lake's bottom are regularly deployed throughout the year. If conditions are ideal, a water quality sonde is deployed. Variance fluxes will be estimated across the lake-atmosphere interface using an aerodynamic flux-gradient approach.

A 3-dimensionsal sonic anemometer-thermometer will be used to obtain the eddy diffusivity. The CH_4 vertical concentration gradient above the lake will be measured at intervals of 1.5 meters and 3.5 meters. Atmospheric emissions are recorded using an Ultraportable GHG Analyzer, which quantifies CH_4, CO_2, and water vapor concentrations near-simultaneously. Air sampling is performed using a custom designed manifold; air will be pulled through this manifold at 406 SLPM using a KNF diaphragm pump. This sampling sequence has two cycles. The first cycle calibration will be set every 6 hours using a zero-air standard measured for 30 seconds and a standard calibration gas measured for 30 seconds. The second cycle will conduct atmospheric sampling through alternating sampled inlets at 15 second intervals.

The raw mixing ratios will be calibrated using the interpolated calibration data. The final mixing ratios consider flux calculations and will be block-averaged at 30-minute intervals. All sonic anemometer data will be recorded at 10 Hz. AmeriFlux protocols will be followed for routine calculations, corrections, and data quality estimations.

A buoy system has been implemented to automate in-lake measurements and quantify oscillations in biogeochemical pools, such as carbon and oxygen. Carbon fluxes will be measured through three factors: pH, alkalinity, and dissolved inorganic carbon (DIC). Dissolved oxygen (DO) is a product of photosynthesis and can be used to assess carbon pools. A chlorophyll sensor will be used to analyze changes in phytoplankton biomass – and a fluoroprobe will be used to further differentiate algae into varying classes based on pigment signatures. These algal sensors allow for an analysis of phytoplankton's role on mediating greenhouse gas fluxes. High frequency CO_2 sensors will also be used to assess short-term oscillations in the CO_2 pool of Bog Lake Fen.

Study Area: Bog Lake Fen, MN



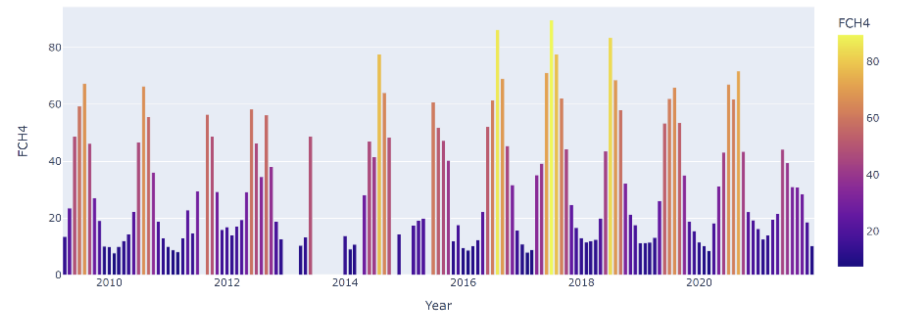
Eddy Covariance Tower (Research Site)

Project Objective(s):

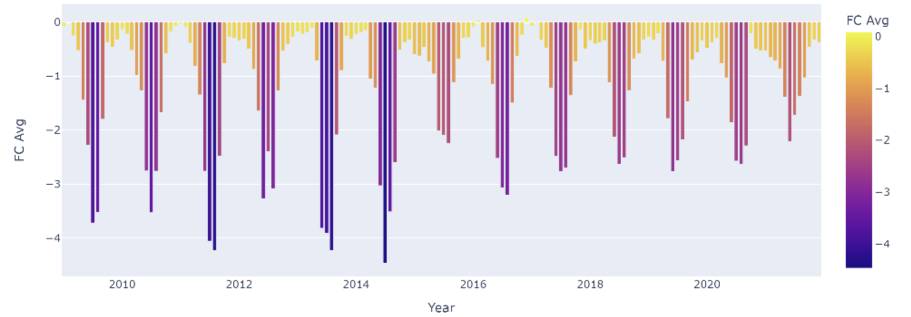
- Plot Monthly (2009 – 2021) Variable Averages (Emission Trends & Seasonality)
- Multiple Linear Regression & Correlation Matrices
- Forecast Modeling Methane Flux values with Seasonal Variability
- Time Series Analysis (Univariate & Multivariate)

Results:

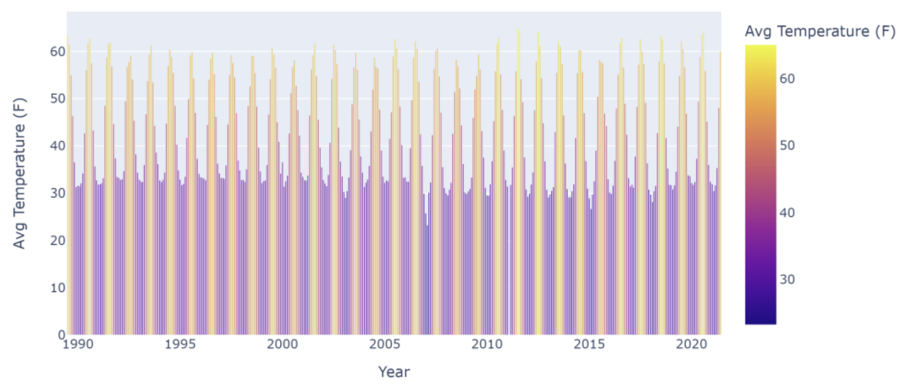
1. *Monthly Variation Plots (Seasonality Incorporations of Averaged Monthly Emissions)*



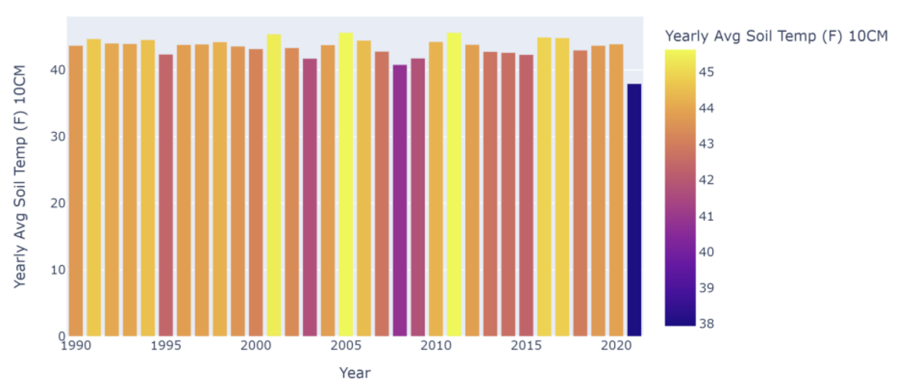
FCH_4



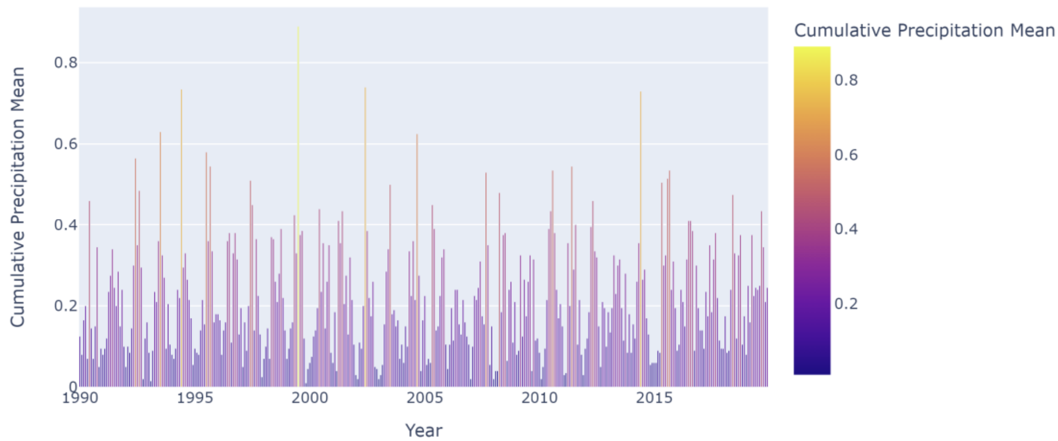
FCO_2



Atmospheric Temperature (°F)



Soil Temperature (°F)

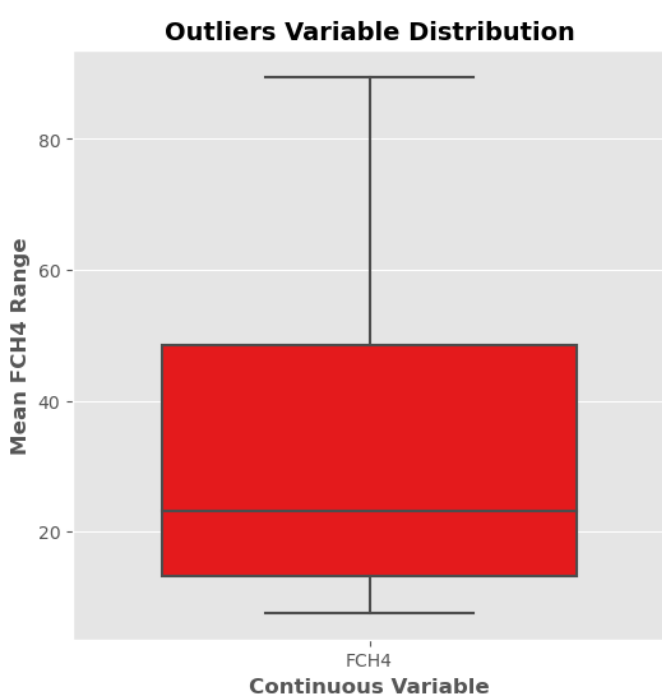


Cumulative Precipitation Values (mm)

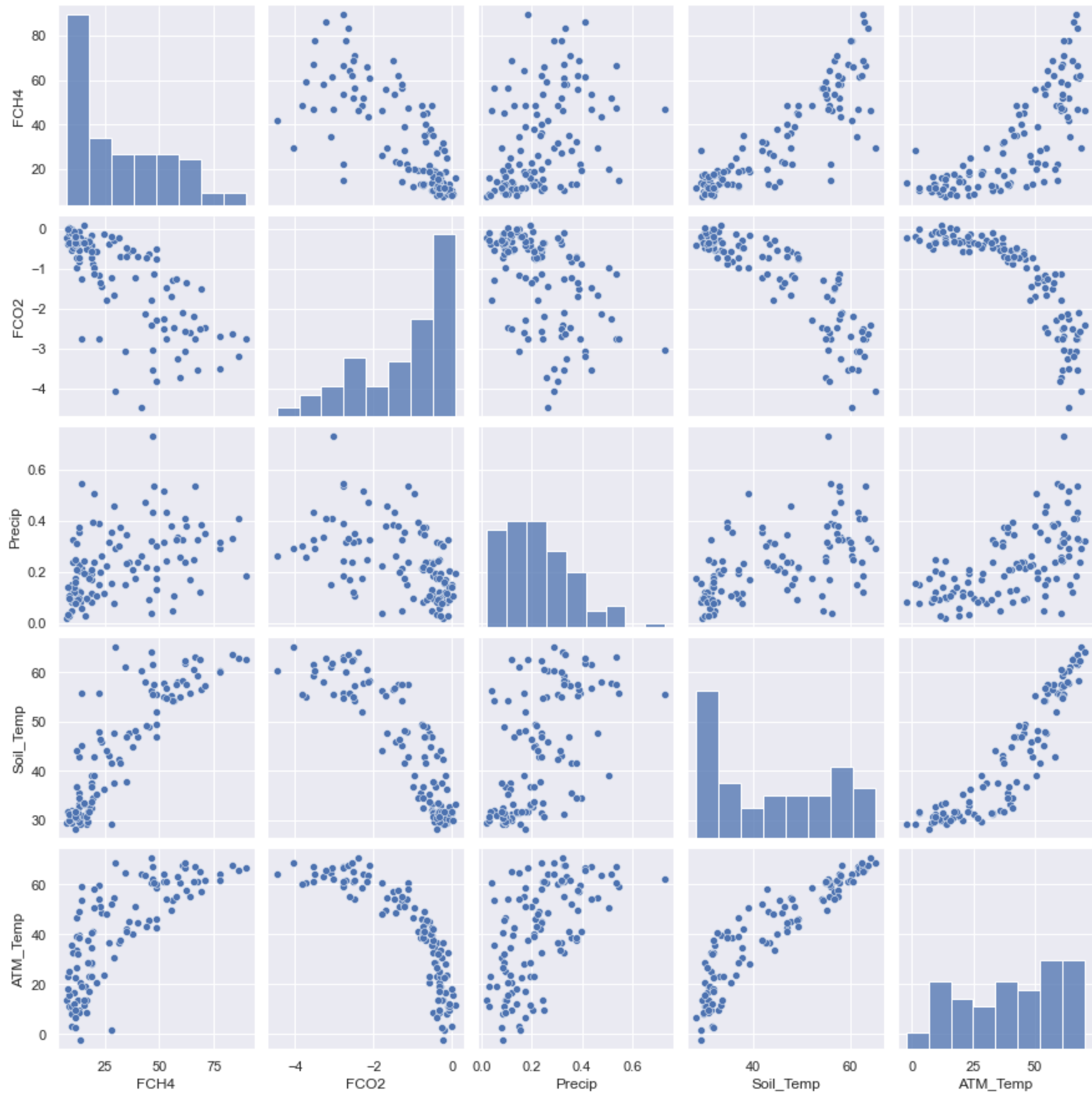
2. *Linear Regression Process*

	FCH4	FCO2	Precip	Soil_Temp	ATM_Temp
FCH4	1.000000	-0.735146	0.433535	0.873856	0.762159
FCO2	-0.735146	1.000000	-0.511046	-0.869557	-0.830535
Precip	0.433535	-0.511046	1.000000	0.574996	0.591479
Soil_Temp	0.873856	-0.869557	0.574996	1.000000	0.915810
ATM_Temp	0.762159	-0.830535	0.591479	0.915810	1.000000

Correlation Matrix



Outlier Assessment

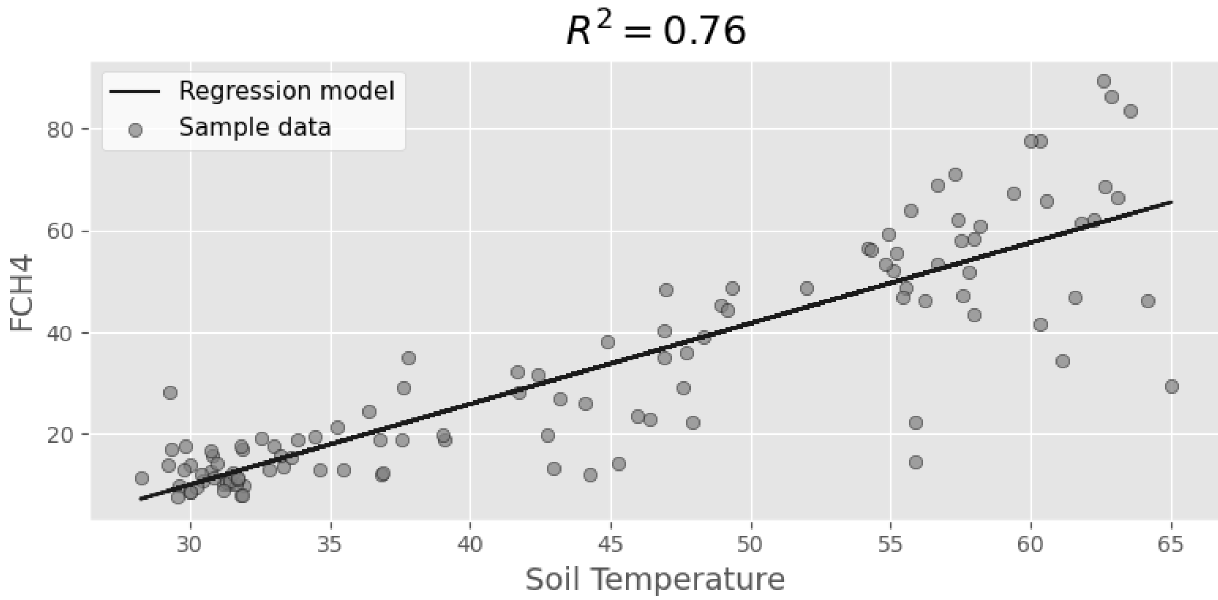


Scatterplot of all Variables

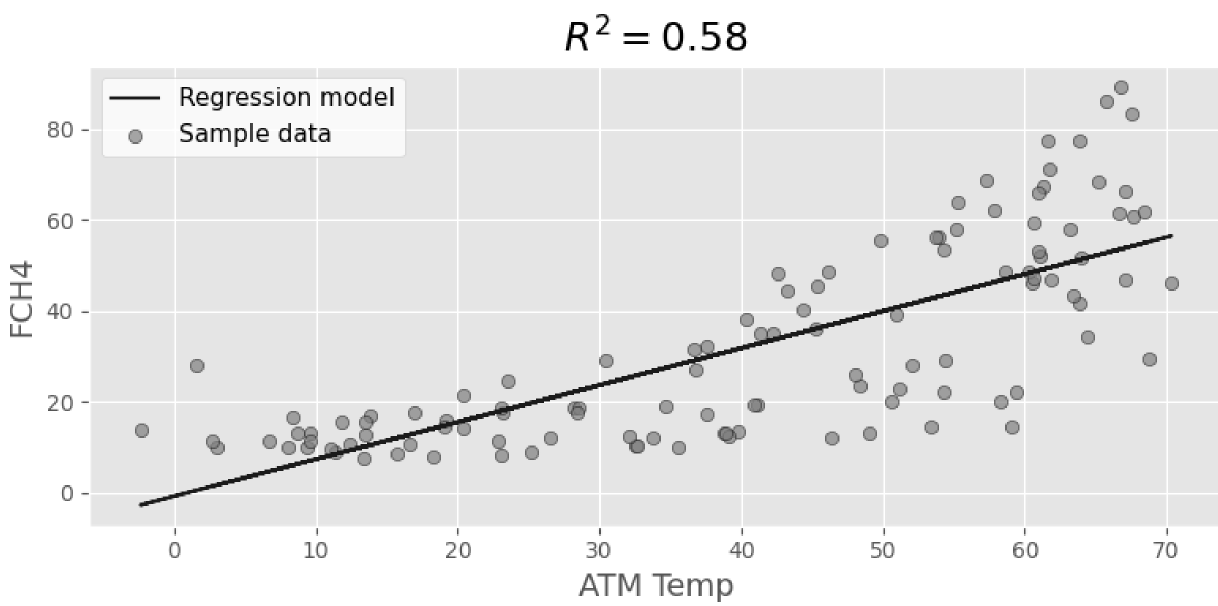
The Correlation Matrix indicates all predictor variables are strongly correlated with our indicator variable (FCH_4). Soil Temperature is the highest correlated variable (0.87), followed by Atmospheric Temperature (0.76), FCO_2 (-0.74), and Precipitation (0.43). Soil Temperature and Atmospheric Temperature are highly dependent on each other – the Soil Temperature measurements at the site were captured at 10CM. FCO_2 was the only negative association variable in the model, but its influence is significant (Carbon associated with both CH_4 & CO_2).

An Outlier Assessment underscores the influence of seasonality with FCH_4 measurements; during the summer months, with augmented temperatures, increased measurements of FCH_4 were reported. This simple assessment suggests the importance of considering seasonality and capturing trends in environmental data. This is integral for time-series analyses and predictive models.

The mass scatterplot visualizes the relationships between our variables – another way of expressing the results of our correlation matrix (a visualization to assess linearity, relationships, and the spread of the data).



Simple Linear Regression: FCH_4 vs Soil Temperature



Simple Linear Regression: FCH_4 vs Atmospheric Temperature

Building a simple linear regression model is a first-step in assessing the fit of our ML Regression model – I chose to look at Soil Temperature and Atmospheric Temperature, first, because they were the most correlated variables with FCH_4. Simply using Soil Temperature returned an R^2 value of 0.76 and BIC of 608.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   FCH4   R-squared:                       0.750
Model:                            OLS   Adj. R-squared:                  0.747
Method:                 Least Squares   F-statistic:                     230.9
Date:                Wed, 07 Dec 2022   Prob (F-statistic):           6.99e-25
Time:                        09:54:41   Log-Likelihood:                -299.63
No. Observations:                  79   AIC:                             603.3
Df Residuals:                      77   BIC:                             608.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -37.1362      4.735     -7.842      0.000     -46.566     -27.707
Soil_Temp      1.5832      0.104     15.195      0.000       1.376       1.791
```

Simple Linear Regression: OLS Results (FCH_4 vs Soil Temperature)

```
# Build a ML Model

# Set independent and dependent variables
X = df[['Soil_Temp', 'ATM_Temp', 'FCO2', 'Precip']]
y = df['FCH4']

# Initialize model from sklearn and fit it into our data
regr = linear_model.LinearRegression()
model = regr.fit(X, y)

print('Intercept:', model.intercept_)
print('Coefficients:', model.coef_)
```

✓ 0.1s

```
Intercept: -47.08696940701227
Coefficients: [  2.09416426  -0.20092428   1.35468969 -13.26701851]
```

Intercept = estimated average value of FCH4 Measurement when all independent variables are set to 0.

Coefficients = relationship of independent variables to depedent variable (FCH4)

Result of Fit of ML Regression Model:

Avg FCH4 Measurement = (2.1 * Soil Temp) - (0.2 * ATM Temp) + (1.4 * FCO2) - (13.3 * Precip)

Fit of ML Regression Model

$$AVG\ FCH_4 = (2.1 * Soil\ Temp) - (0.2 * ATM\ Temp) + (1.4 * FCO_2) - (13.3 * Precip)$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   FCH4   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.770
Method:                 Least Squares   F-statistic:                     95.67
Date:                Wed, 07 Dec 2022   Prob (F-statistic):           9.58e-35
Time:                        10:04:54   Log-Likelihood:                -426.35
No. Observations:                 114   AIC:                             862.7
Df Residuals:                     109   BIC:                             876.4
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -47.0870      6.202     -7.592      0.000     -59.380     -34.794
Soil_Temp      2.0942      0.234      8.950      0.000       1.630       2.558
ATM_Temp      -0.2009      0.124     -1.614      0.109      -0.448       0.046
FCO2           1.3547      1.762      0.769      0.444      -2.138       4.847
Precip       -13.2670      8.923     -1.487      0.140     -30.952       4.418
==============================================================================
Omnibus:                       18.663   Durbin-Watson:                   1.247
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               32.222
Skew:                          -0.716   Prob(JB):                     1.01e-07
Kurtosis:                       5.176   Cond. No.                         582.
```

ML Regression Model (OLS Results)

Our Multiple Linear Regression Model (built using all predictor variables – Soil Temperature, Atmospheric Temperature, FCO_2 and Precipitation) returned an R^2 of 0.78 and BIC of 876.4. The predictive power of this model is marginally higher than our simple linear regression model (just using Soil Temperature as a predictor variable for FCH_4), with an increased BIC of 268.4 (our simple linear regression model returned a BIC value of 608, compared to this model's BIC of 876.4).

```python
# F-Test (ANOVA): Analysis of Variance

print('F-statistic:', olsmod.fvalue)
print('Probability of observing value at least as high as F-statistic:', olsmod.f_pvalue)
```

✓ 0.2s

```
F-statistic: 95.66583514754795
Probability of observing value at least as high as F-statistic: 9.583249586791431e-35
```

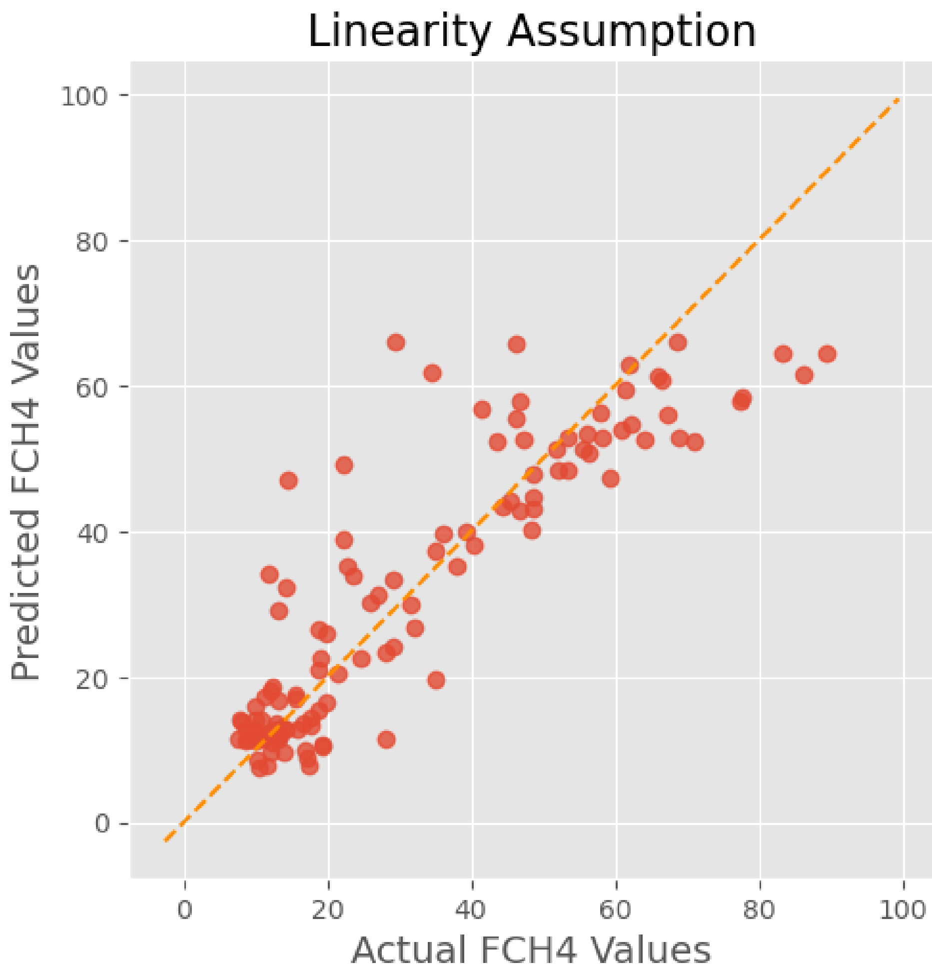
F-values represent the ratio between group-variation and within-group variation.

Large f-values indicate between-group variation to be larger than within-group variation. This implies there is a statistically significant difference in our group means.

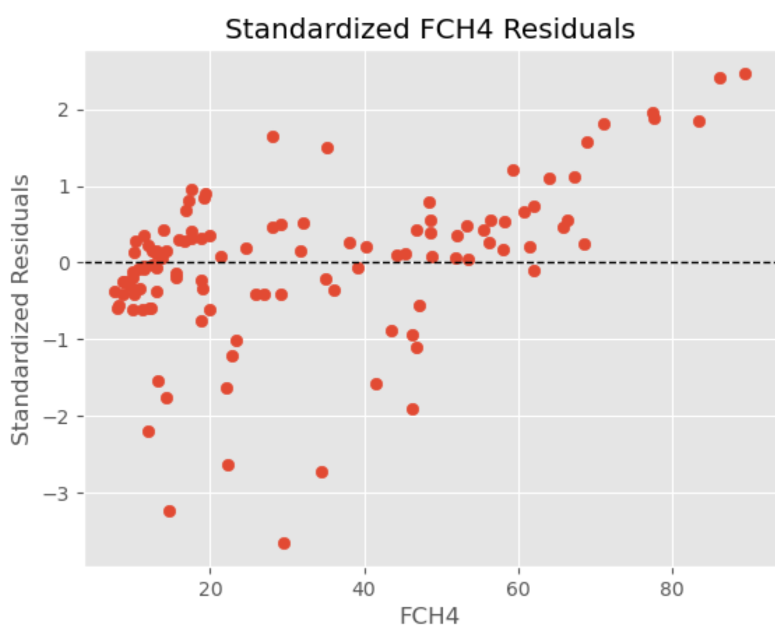
Our model returned an F-statistic of 95.7. This implies we have statistically signficant predictor variables in the model.

F-Test (ANOVA)

An F-Test (ANOVA) was performed to analyze the variance of our ML Regression Model. Our model returned an F-statistics of 95.7, implying statistical significance of our predictor variables.



Linearity Assumption: Actual FCH_4 Values vs Predicted FCH_4 Values



Standardized Residuals

Standardized variables (either the predicted values or the residuals) have a mean of zero and standard deviation of one. If the residuals are normally distributed, 95% of them will fall between -2 and 2. If they fall above or below 2 (in this case, four observations) they are considered unusual.

Almost all observed residuals of FCH_4 measurements fall between -2 and 2 and can therefore be considered normally distributed.

```
# Assess the autocorrelation of our model:

# Autocorrelation - correlation of errors (residuals) over time

from statsmodels.stats.stattools import durbin_watson

durbinWatson = durbin_watson(df['residual'])

print('Durbin-Watson:', durbinWatson)
if durbinWatson < 1.5:
    print('Signs of positive autocorrelation', '\n')
    print('Assumption not satisfied')
elif durbinWatson > 2.5:
    print('Signs of negative autocorrelation', '\n')
    print('Assumption not satisfied')
else:
    print('Little to no autocorrelation', '\n')
    print('Assumption satisfied')
```

✓ 0.3s

```
Durbin-Watson: 1.2466964078481961
Signs of positive autocorrelation
```

Autocorrelation Assessment: Durbin-Watson Analysis

Our Durbin-Watson analysis showed signs of positive autocorrelation - this means that the increased observed in a time interval will lead to a proportionate increase in the lagged time interval.

The outcome of a Durbin-Watson test ranges from 0 to 4. An outcome closely around 2 means a very low level of autocorrelation. An outcome closer to 0 suggests a stronger positive autocorrelation, and an outcome closer to 4 suggests a stronger negative autocorrelation.

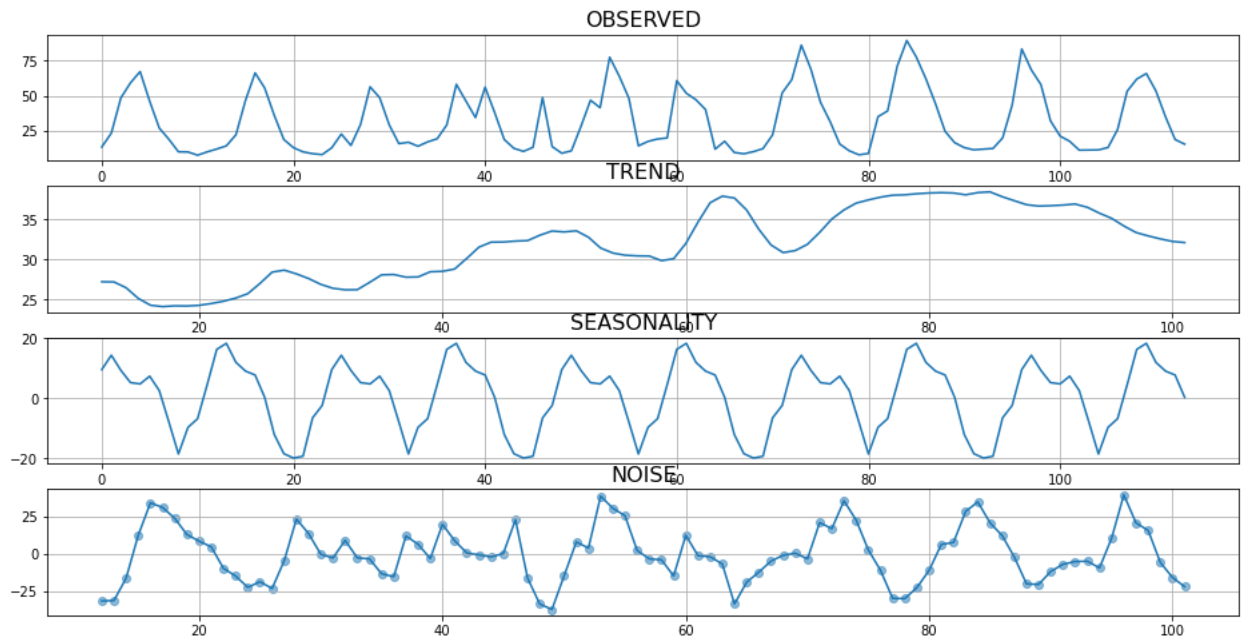
Our Durbin-Watson result returned a value of 1.2.

3. *Forecast Modeling – Predictive Estimates of FCH_4 Measurements*

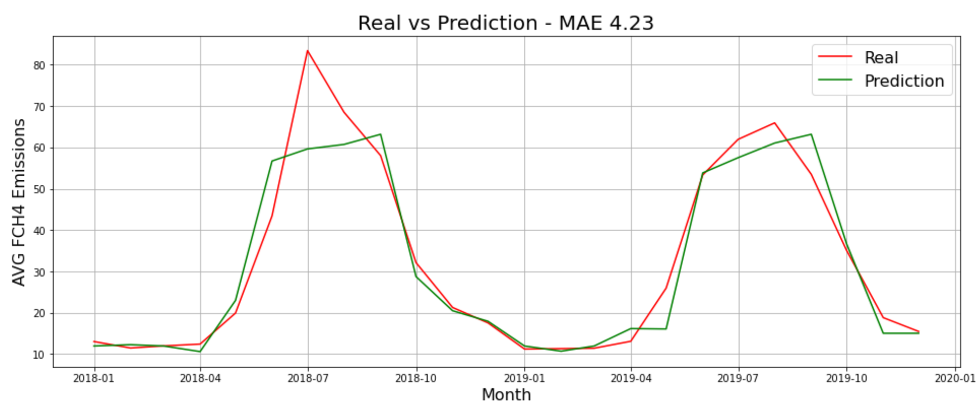
Supervised Machine Learning (Microsoft Light Gradient Boosting Machine Model)

- Variables incorporated were only those indicated to have correlation with FCH4 (ATM temperature, Soil temperature, FCO2, and Precipitation)
- We used data values from the years 2009 – 2019 to predict 2020 values.
- We looked for seasonalities and trends by decomposing the time series into 4 states: observed, trend, seasonality, and noise (residuals).

- We assessed the influence of incorporating lags into our predictive model and created a dataframe with variables ranked in importance on our model. We trained a model without lags, and we trained a model with lags.
- Our first attempt at forecasting FCH4 emissions was set to one week, meaning we plotted the real values versus predicted values for the last week of the dataset. This was a test to see if the SML model worked.
- We then created a 1-year lag variable by shifting the target value (FCH4) back a year



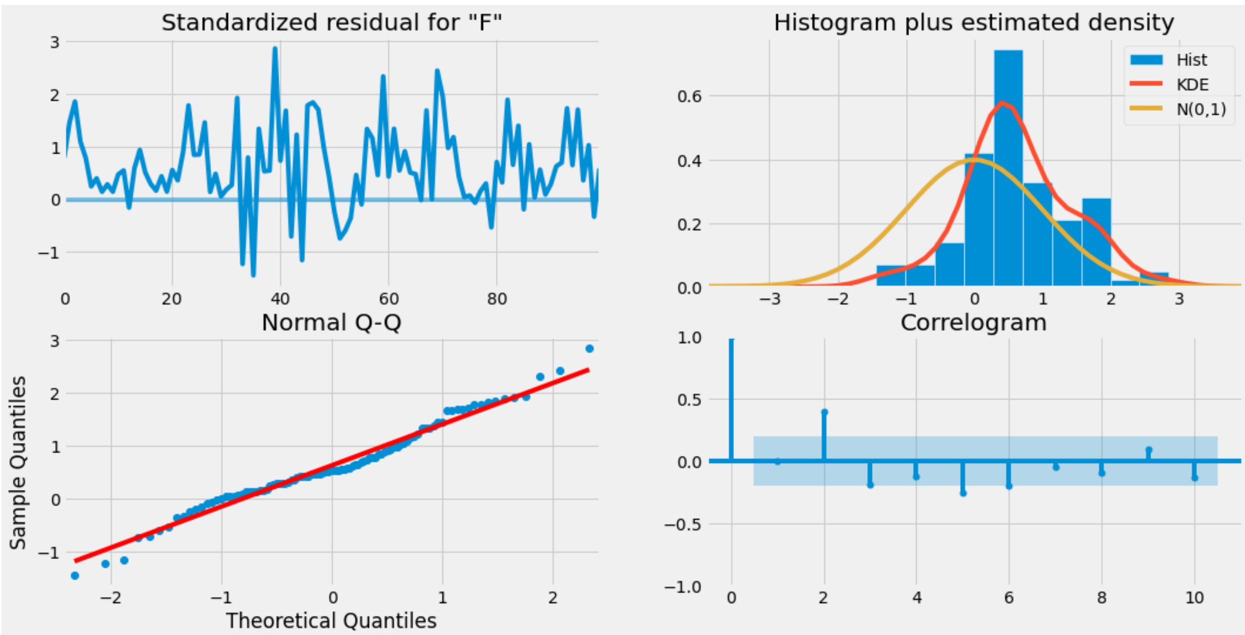
Decomposed Time-Series: Observed, Trend, Seasonality, and Noise (Residuals)



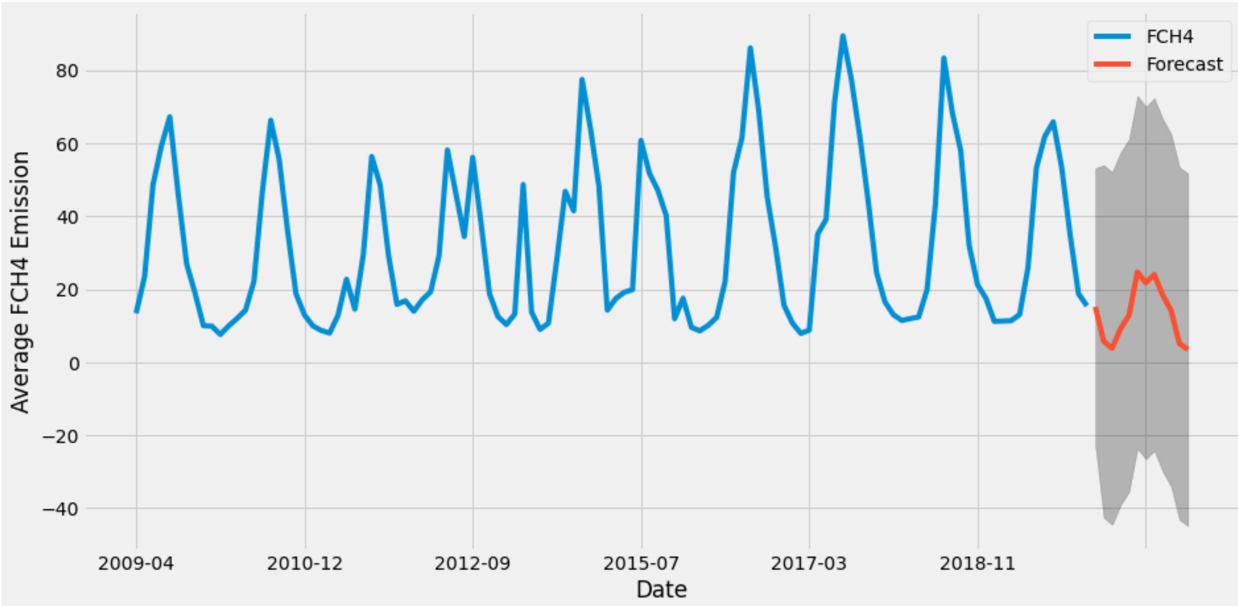
Real vs Predicted (Mean Average Error of 4.23)

SARIMAX: Univariate (FCH_4) Predictive Modeling

- Seasonal Auto-Regressive Integrated Moving Average w/ eXogenous factors (SARIMAX)
- Includes seasonal effects & eXogenous factors w/ the autoregressive and moving average computed in the model
- Only looking at FCH4 data for the years 2009 – 2019 to predict 2020 emissions.
- Decompose dataframe by an 'additive' model to isolate 4 factors: observed, trend, seasonality, and residuals.
- Compute parameter combinations for Seasonal Arima through iterative process. Use these parameters to incorporate seasonality.
- Plot the predictive model for FCH4 – observed vs predicted – based on a 1-year forecast.



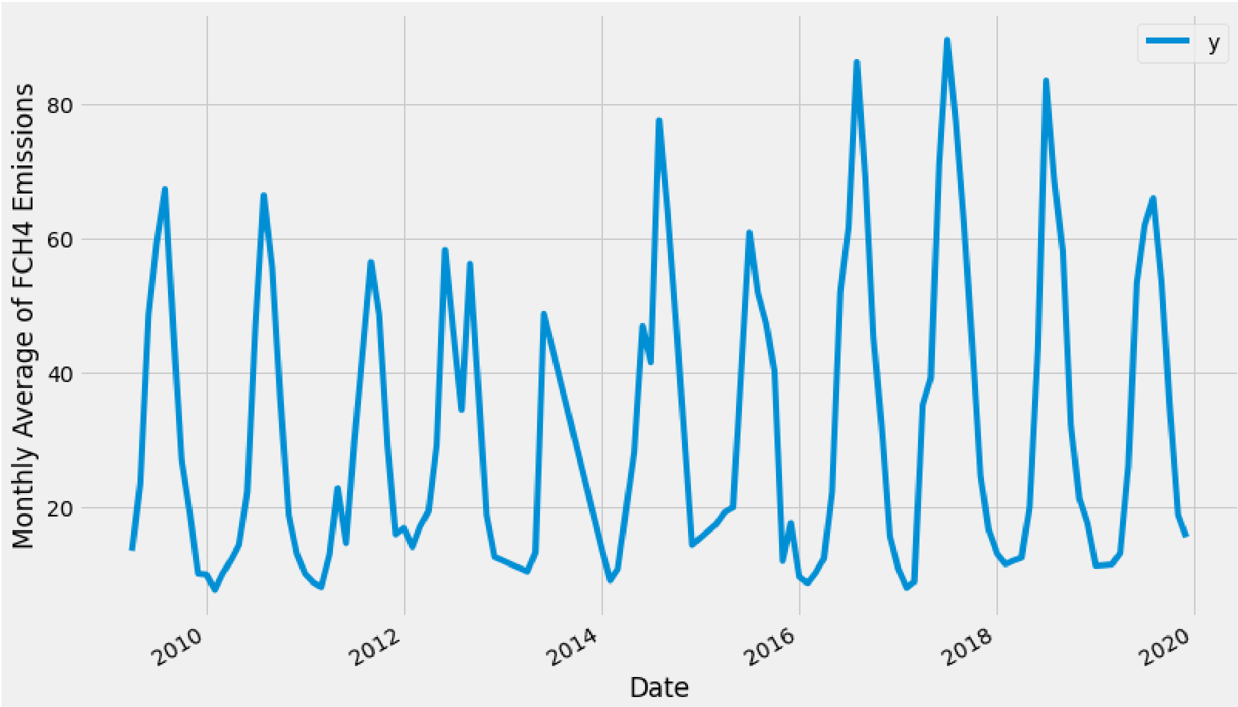
Four-Series Assessment: Residuals, Q-Q Plotting, Histogram, Correlogram



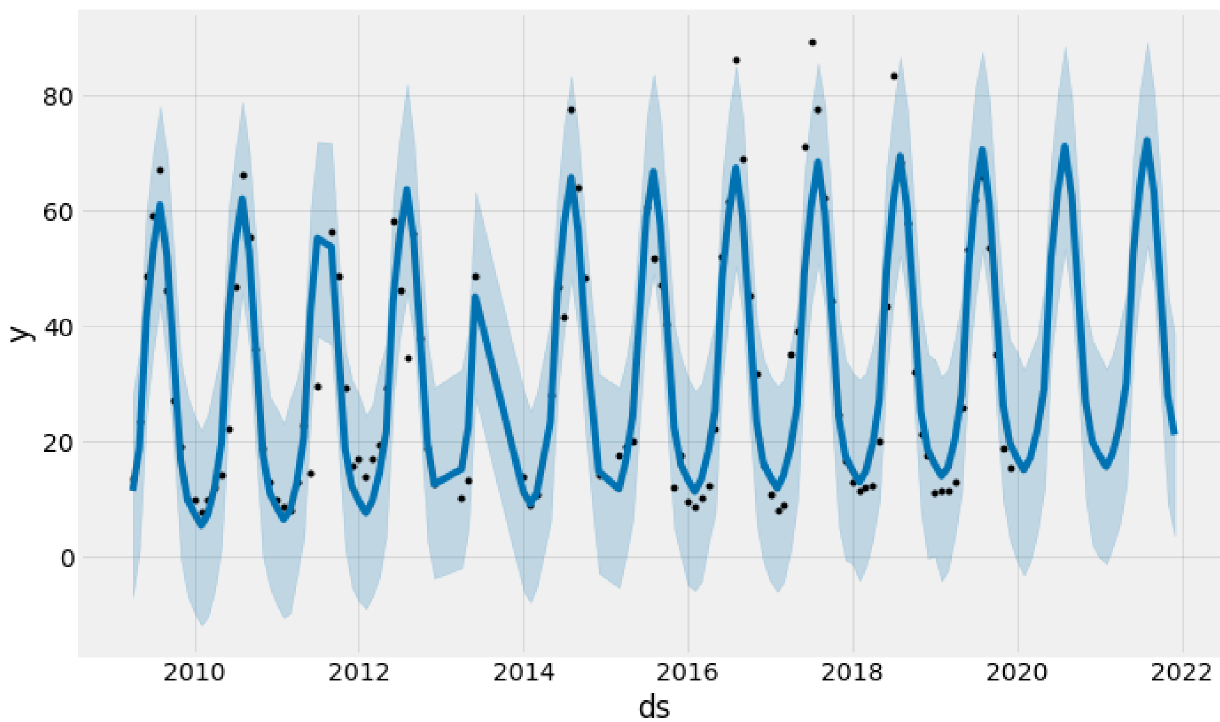
SARIMAX Predictive Modeling for 2020 FCH_4 Emissions

Facebook Prophet: Univariate (FCH_4) Two-Year Forecasting (2022 Emissions)

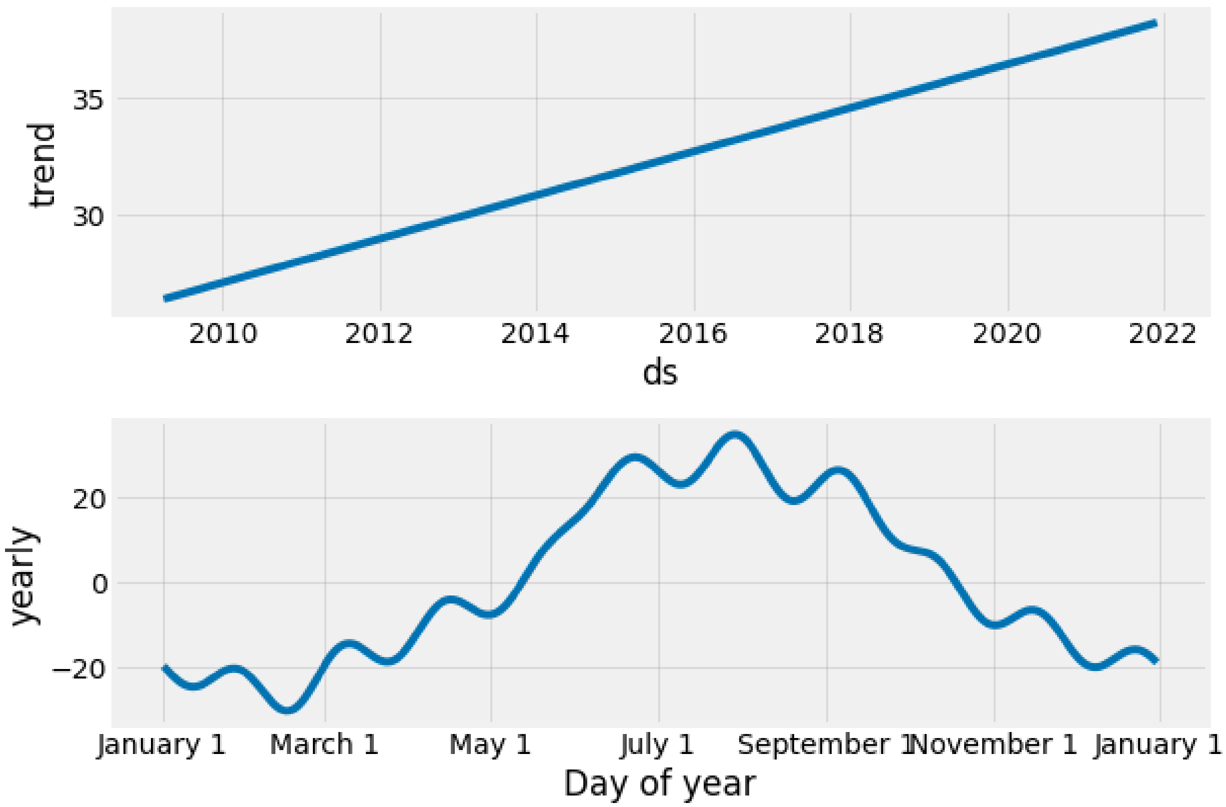
- Forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.
- Univariate analysis (FCH4) from 2009 – 2019 emissions to predict 2022 emissions.
- Visualize monthly average of FCH4 emissions (y)
- Set the uncertainty interval to 95%
- Assess overall trend in FCH4 emissions (completely linear) and monthly level. Yearly seasonality shows FCH4 is highest in Summer (July) and tapers off in January.
- Automatically isolate largest quantity of change points where the rate has highest amount of change.



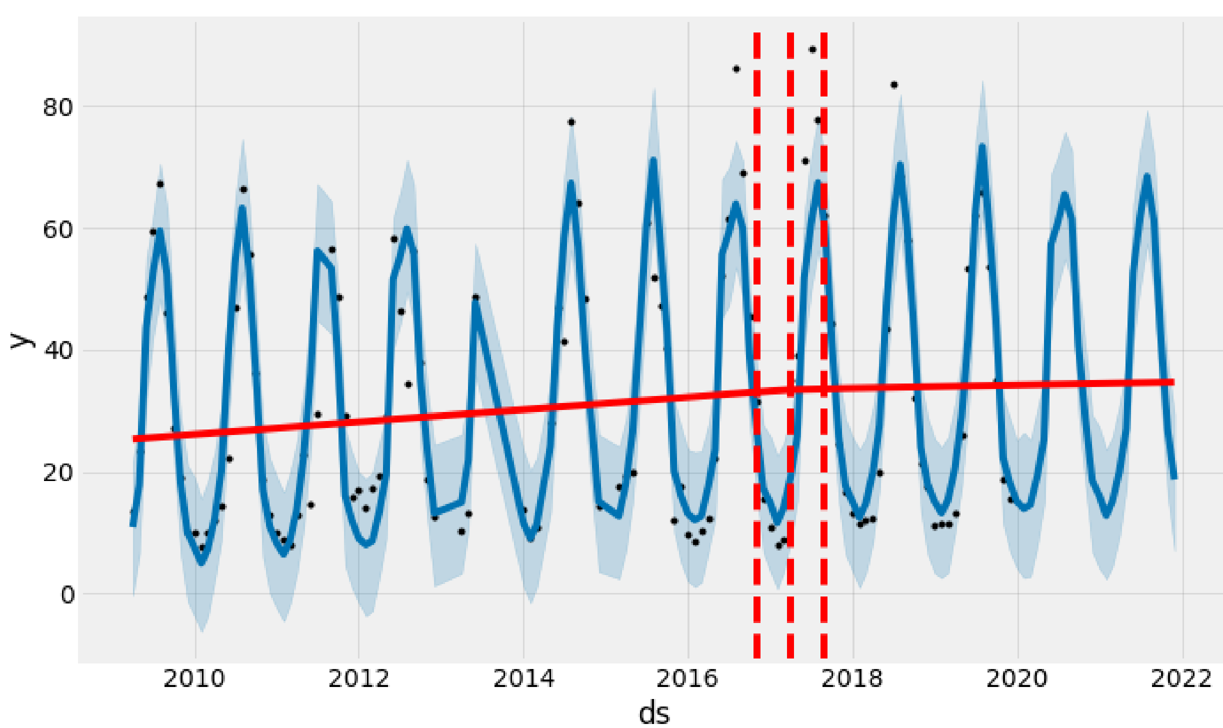
Visualizing Monthly Averages of FCH_4 (y)



Forecasting Model (2022 Predictions Based on 95% C.I. and Outlier Detection)



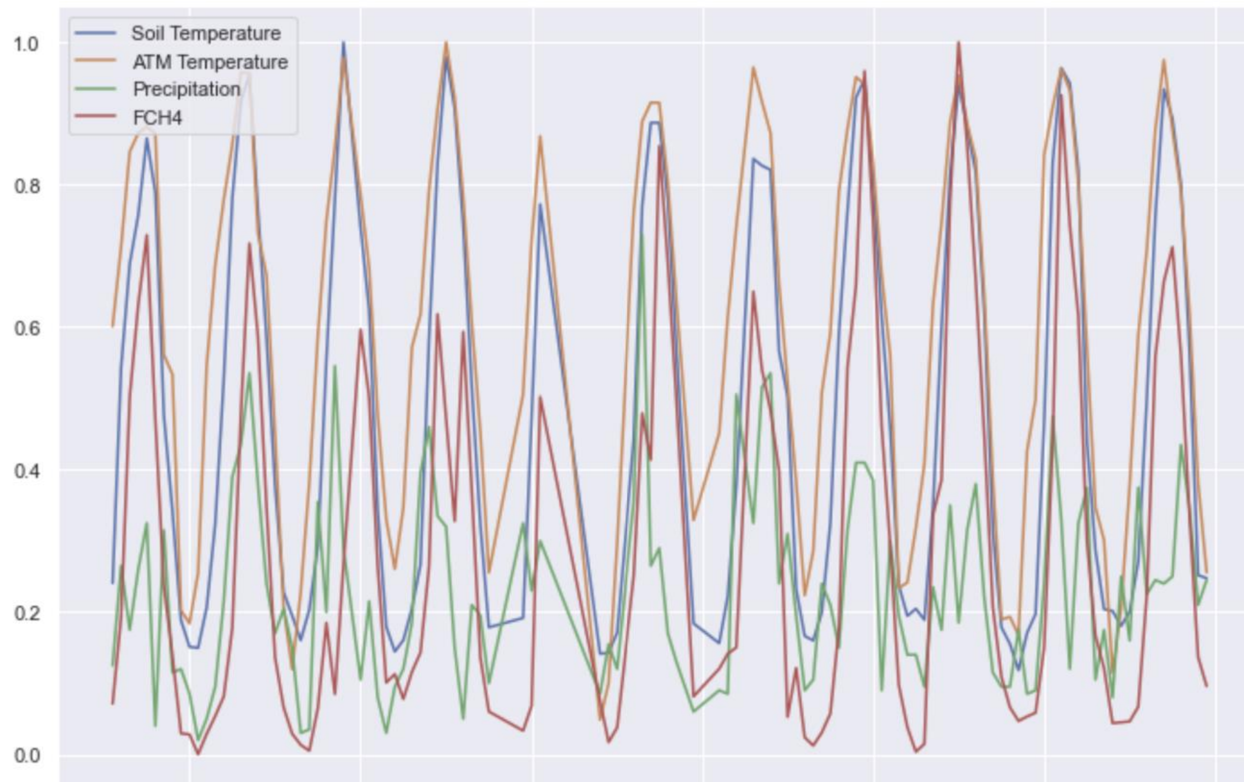
Trend: Linear Increase of FCH_4 Estimates (with Monthly Peaks and Tapers)



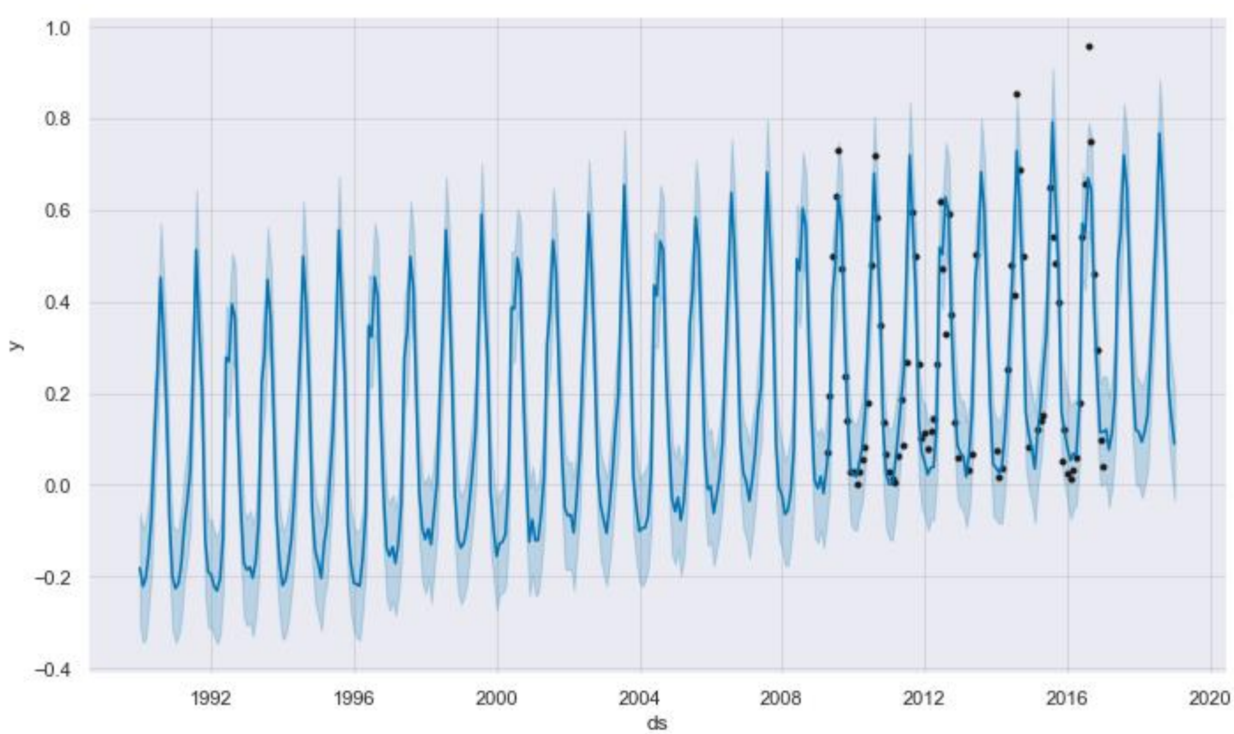
Facebook Prophet Predictive Model: Trends and Year of Highest Variability (2017)

Facebook Prophet: Multivariate Time-Series Analysis

- Set start date for model (1990-01) and end date for model (2022-01). Choose a date for splitting, training, and testing dataset (2017-01).
- Variables incorporated into model: Soil Temperature, ATM Temperature, Precipitation, and FCH4.
- Trained the model with all values <= the training end date (2017-01). Tested the model with all values > the training end date (2017-01). Basically, our training dataset captured all values from 1990-01-01 through 2017-01-01. Our testing dataset captured all values from 2019-02-01 through 2021-12-01. Based on the threshold data set, there are 325 data points in the training dataset and 59 data points in the testing dataset.
- We then created a univariate baseline model using the default Prophet hyperparameters – and then fit the model using the training dataset. Prophet automatically fits daily, weekly, and yearly seasonalities.
- To make a forecast, we need to first create a time range for future predictions (periods). Black dots are the actual values. Blue line is the prediction. Blue shades are the uncertainty interval (80% is FB Prophet default). The uncertainty interval is calculated based on the assumption that the average frequency and magnitude of trend changes in the future will be the same as the historical data. The historical data trend changes are projected forward to get the uncertainty intervals.
- Next step is to assess the model's performance. The forecast dataframe doesn't include the actual values, so we need to merge the forecast dataframe with the test dataframe to compare the actual values with the predicted values.
- The mean absolute percent error (MAPE) for the baseline model was 5.5%, meaning that on average the forecast is off by 5.5 of the FCH4 emission value.
- We then tuned the model to make better estimations by forcing the model to consider yearly seasonality.
- We created a multivariate model that considered seasonality and added our predictor variables as regressors.



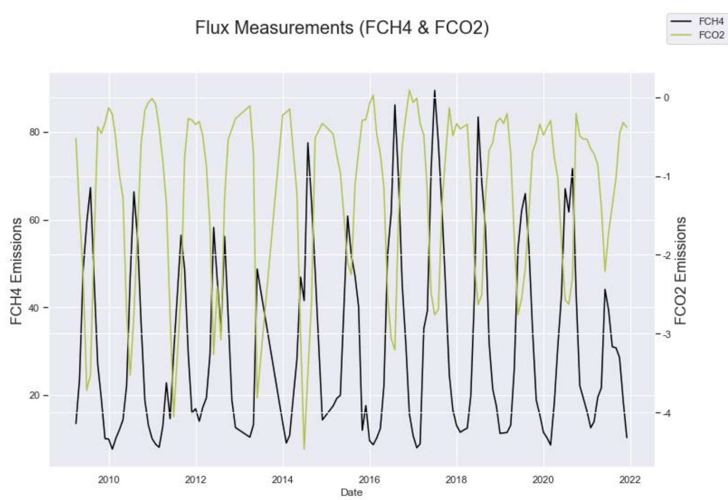
Standardized Averages of All Variables (For a 'Same-Scale' Analysis)



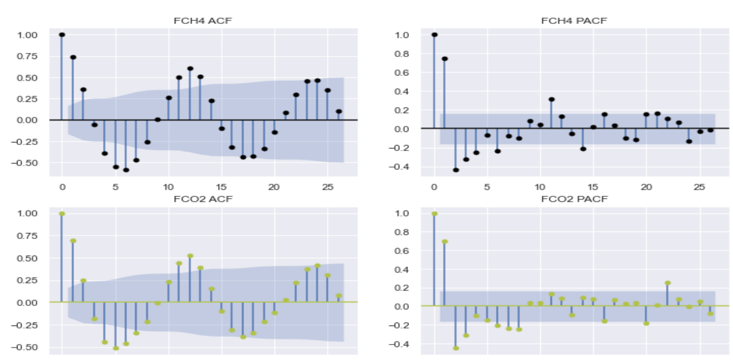
Facebook Prophet Multivariate Predictive Model (2020 Emission Estimates)

The Final Dive: Comparing Machine Learning Models for FCH_4 & FCO_2 Estimates

- Forecast 52 weeks into future (2022).
- Plot the monthly flux measurements.
- Examine the correlation between the series (FCH4 and FCO2). Correlation = -74.90%.
- Assess autocorrelation for FCH4 and FCO2.
- Check stationarity for FCH4 and FCO2 using Augmented Dicky-Fuller method
- Tune forecast for 52-week prediction and create baseline model that doesn't account for seasonality (no lags).
- Optimize model to incorporate seasonality (lags) and use a weighted average to determine best model by isolating lowest test level MAPE.
- Plot updated forecast for FCH4 and FCO2 emissions and compare the MAPE results for each machine learning method.
- Create a Layer Stacking for the 3 best methods (the top 3 with lowest MAPE results) for FCH4 and FCO2.
- Update forecast by comparing MAPE results for layer stacked results.
- Create a single forecast using the best machine learning method for FCH4 and FCO2 (XGBOOST)



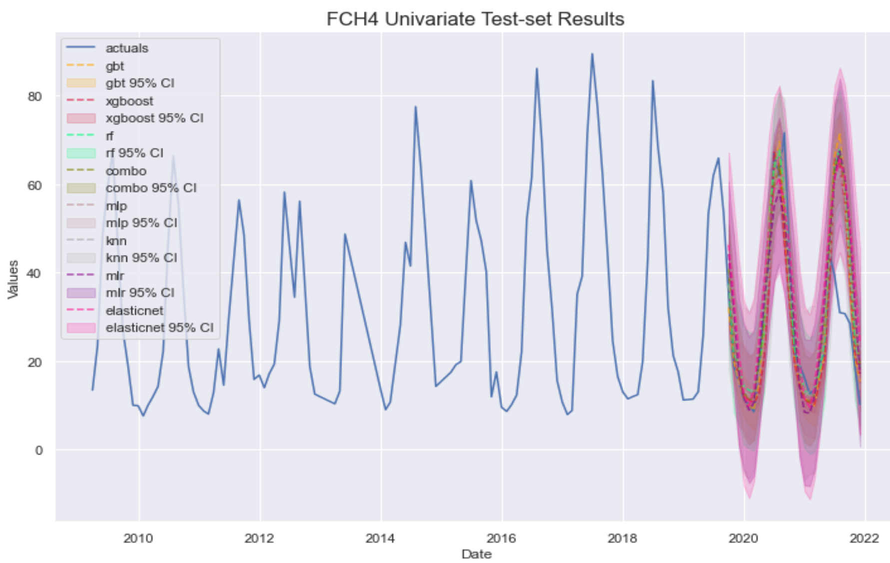
FCH_4 & FCO_2 Measurements (Time-Series) into 2022



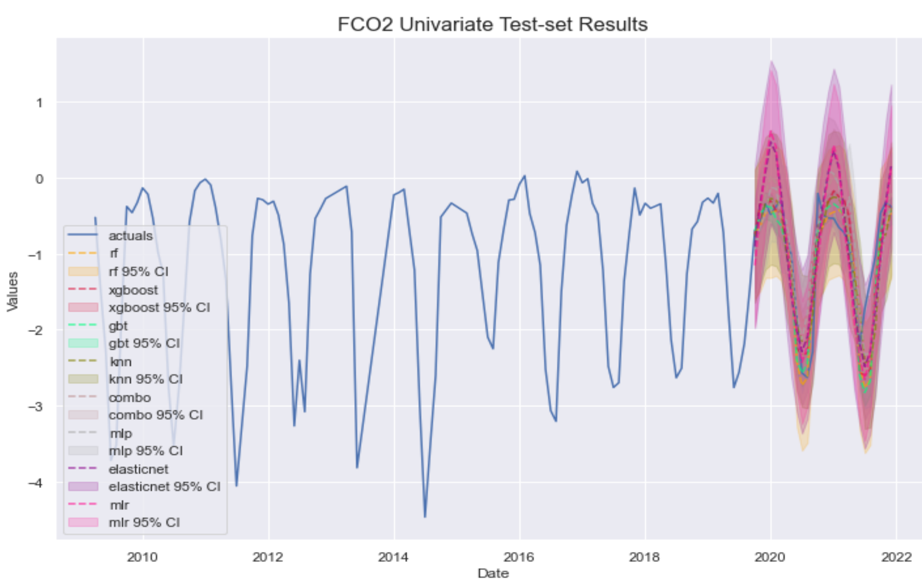
Autocorrelation Assessment

```
FCH4 Augmented Dickey-Fuller results:
the test-stat value is: -2.79
the p-value is 0.0603
the series is not stationary
-------------------------------------
-------------------------------------
FCO2 Augmented Dickey-Fuller results:
the test-stat value is: -8.26
the p-value is 0.0000
the series is stationary
```

Dickey-Fuller Analysis (Stationarity in FCO_2… but not FCH_4)



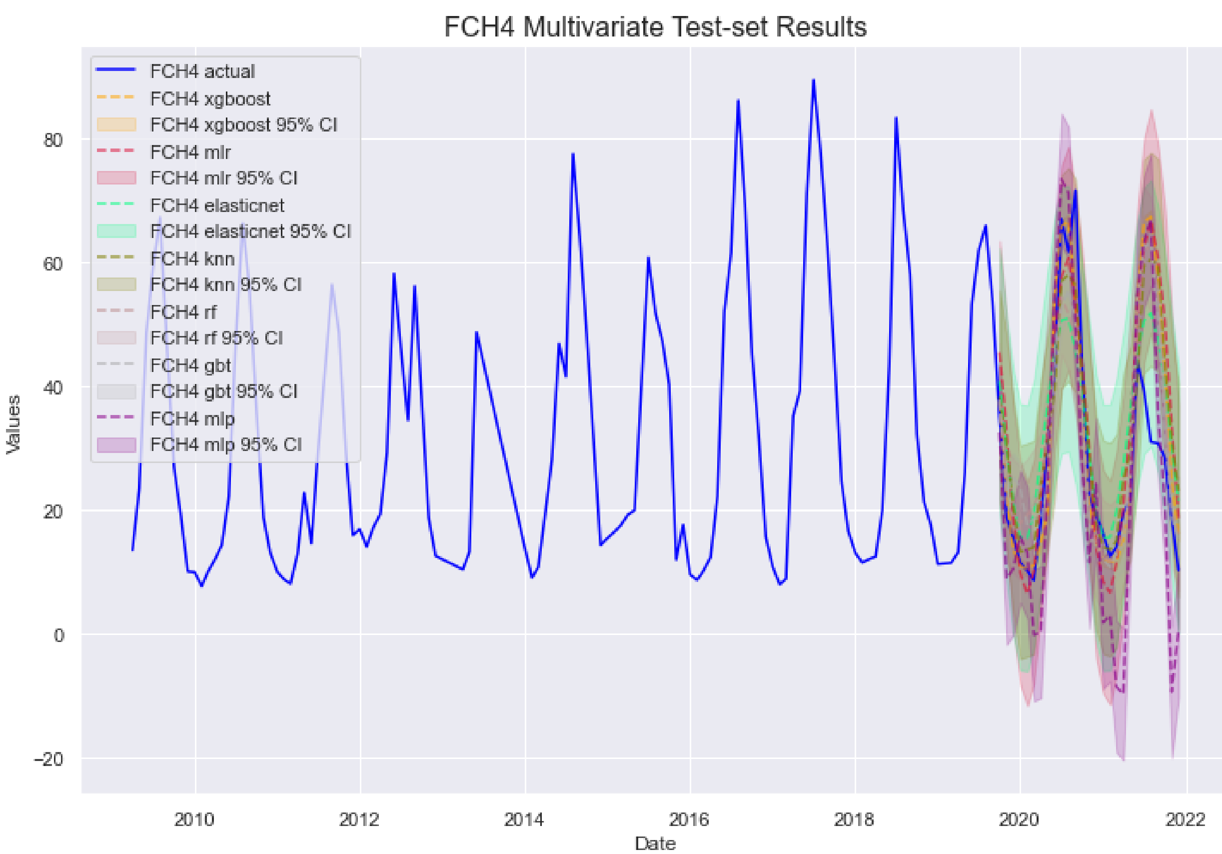
FCH_4 Forecasting (Comparing Deep Learning Models)



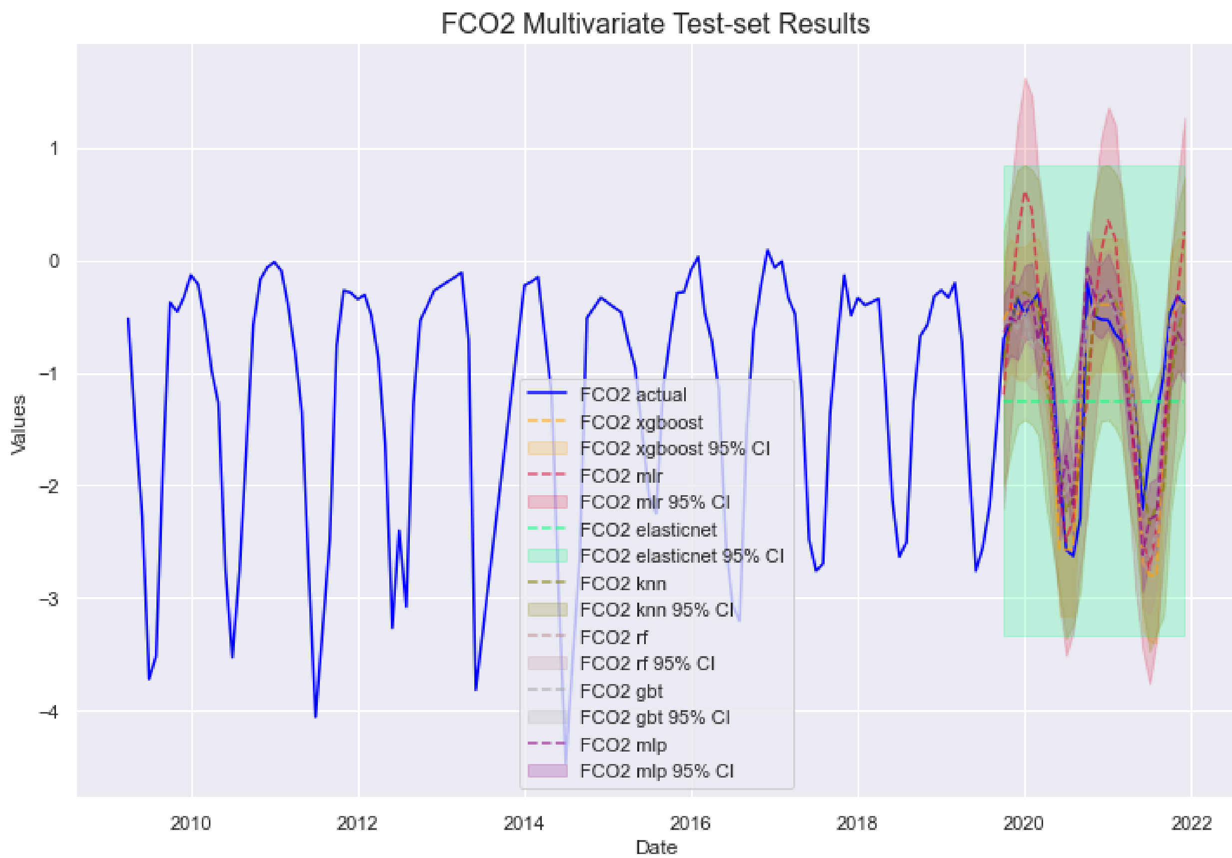
FCO_2 Forecasting (Comparing Deep Learning Models)

	ModelNickname	Series	Integration	LevelTestSetMAPE
0	gbt	FCH4	0	0.2401
1	xgboost	FCH4	0	0.2697
2	rf	FCH4	0	0.2769
3	combo	FCH4	0	0.3032
4	mlp	FCH4	0	0.3211
5	knn	FCH4	0	0.3355
6	mlr	FCH4	0	0.3487
7	elasticnet	FCH4	0	0.4254
8	rf	FCO2	0	0.3766
9	xgboost	FCO2	0	0.3816
10	gbt	FCO2	0	0.3851
11	knn	FCO2	0	0.4634
12	combo	FCO2	0	0.5229
13	mlp	FCO2	0	0.5870
14	elasticnet	FCO2	0	0.7649
15	mlr	FCO2	0	0.8204

Comparison of Deep Learning Models (Choose Models with Lowest MAPE)



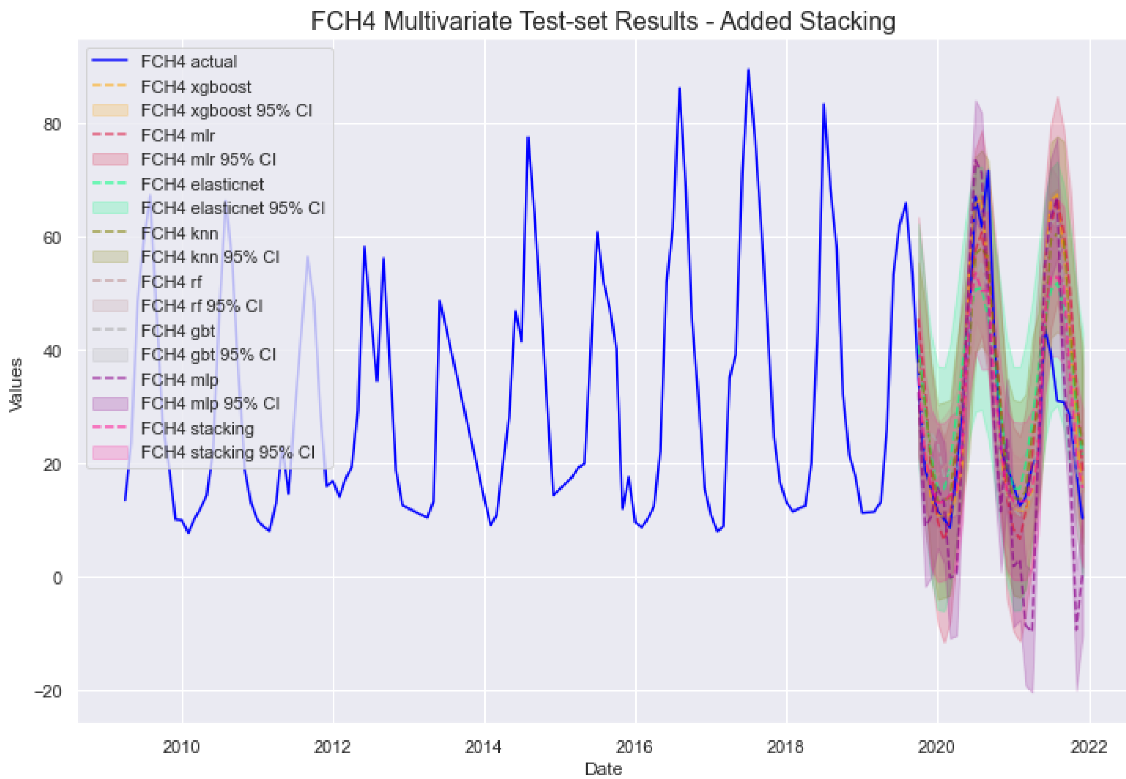
FCH_4 Multi-Variate Forecasting (Comparing Deep Learning Models)



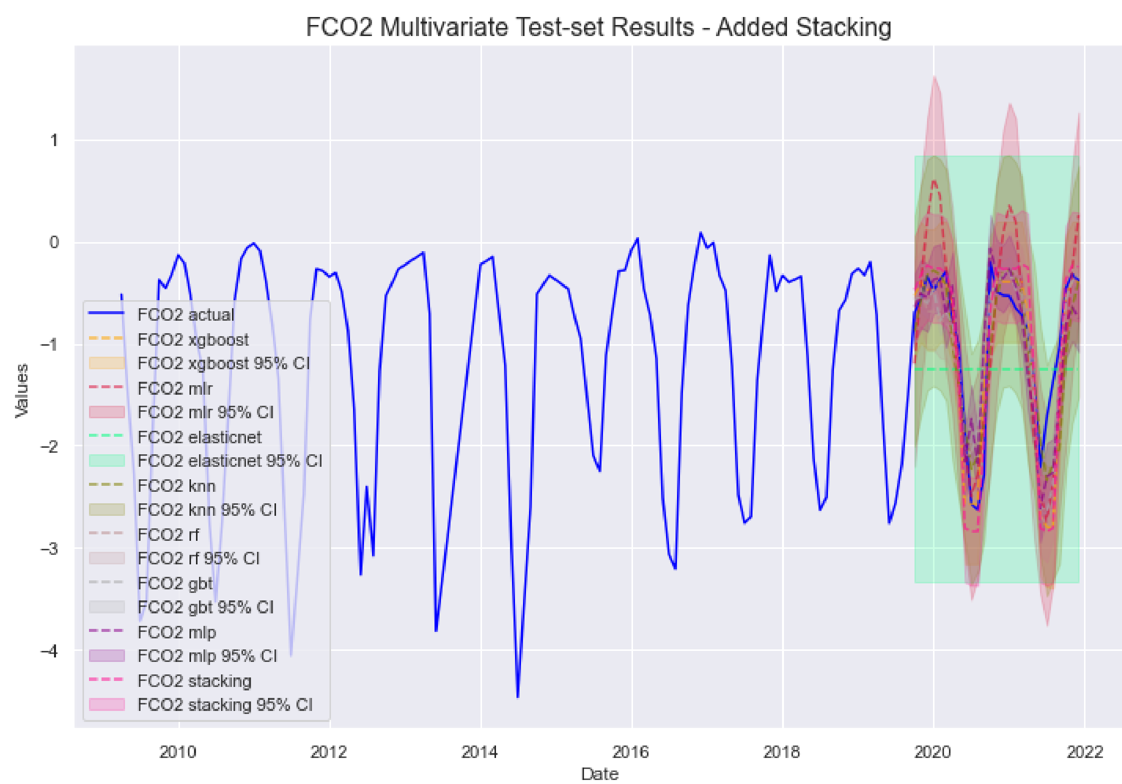
FCO_2 Multi-Variate Forecasting (Comparing Deep Learning Models)

	ModelNickname	Series	HyperParams	LevelTestSetMAPE
0	xgboost	FCH4	{'n_estimators': 250, 'scale_pos_weight': 5, 'learning_rate': 0.2, 'gamma': 3, 'subsample': 0.8}	0.2764
1	mlr	FCH4	{}	0.3859
2	elasticnet	FCH4	{'alpha': 2, 'l1_ratio': 1}	0.3991
3	knn	FCH4	{'n_neighbors': 24}	0.3445
4	rf	FCH4	{'max_depth': 5, 'n_estimators': 500, 'max_features': 'sqrt', 'max_samples': 0.9}	0.3352
5	gbt	FCH4	{'max_depth': 3, 'max_features': 'sqrt'}	0.3558
6	mlp	FCH4	{'activation': 'relu', 'hidden_layer_sizes': (25,), 'solver': 'lbfgs'}	0.5709
7	xgboost	FCO2	{'n_estimators': 250, 'scale_pos_weight': 5, 'learning_rate': 0.2, 'gamma': 3, 'subsample': 0.8}	0.3249
8	mlr	FCO2	{}	0.8575
9	elasticnet	FCO2	{'alpha': 2, 'l1_ratio': 1}	1.2545
10	knn	FCO2	{'n_neighbors': 24}	0.4767
11	rf	FCO2	{'max_depth': 5, 'n_estimators': 500, 'max_features': 'sqrt', 'max_samples': 0.9}	0.4603
12	gbt	FCO2	{'max_depth': 3, 'max_features': 'sqrt'}	0.3785
13	mlp	FCO2	{'activation': 'relu', 'hidden_layer_sizes': (25,), 'solver': 'lbfgs'}	0.4291

Model Results: Improvements & MAPE Scores



FCH_4 Multi-Variate (Layer Stacking)

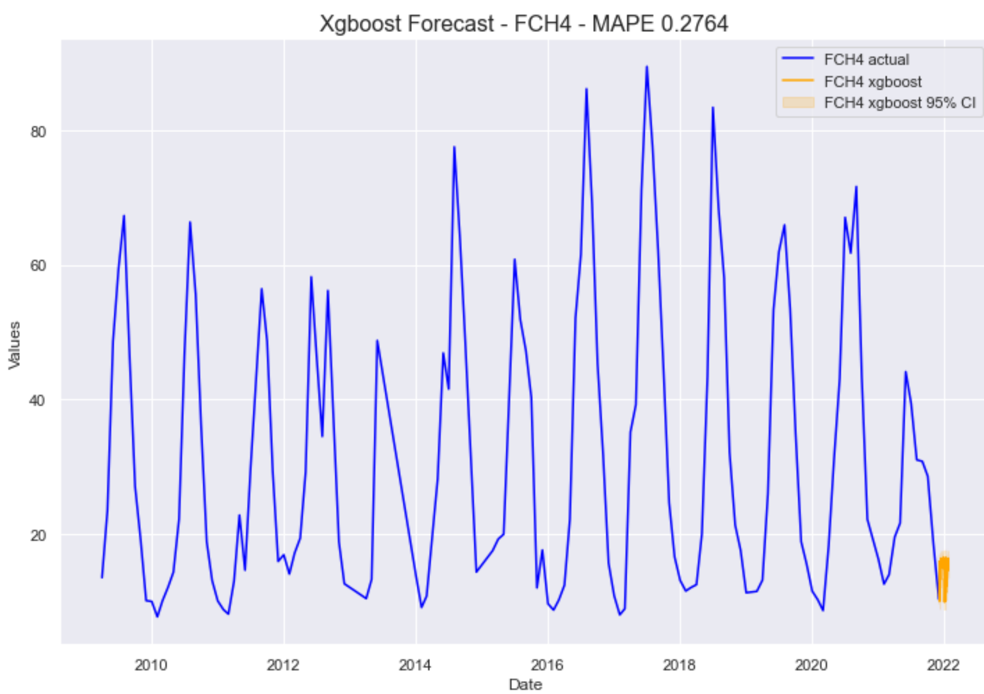


FCO_2 Multi-Variate (Layer Stacking)

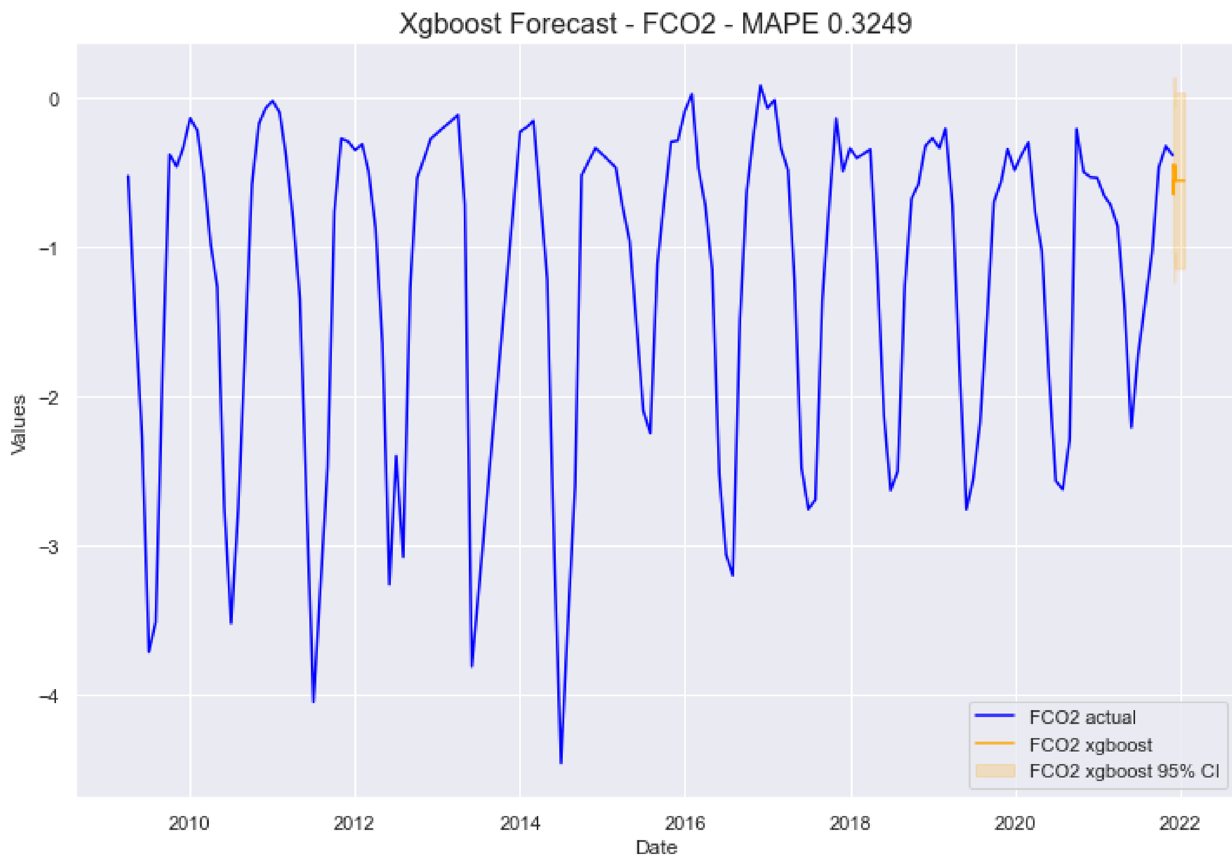
	ModelNickname	Series	LevelTestSetMAPE
0	xgboost	FCH4	0.2764
1	mlr	FCH4	0.3859
2	elasticnet	FCH4	0.3991
3	knn	FCH4	0.3445
4	rf	FCH4	0.3352
5	gbt	FCH4	0.3558
6	mlp	FCH4	0.5709
7	stacking	FCH4	0.2229
8	xgboost	FCO2	0.3249
9	mlr	FCO2	0.8575
10	elasticnet	FCO2	1.2545
11	knn	FCO2	0.4767
12	rf	FCO2	0.4603
13	gbt	FCO2	0.3785
14	mlp	FCO2	0.4291
15	stacking	FCO2	0.4488

```
multivariate average test MAPE for FCH4: 0.3613
multivariate average test R2 for FCH4: 0.52
-----------------------------------------------
multivariate average test MAPE for FCO2: 0.5788
multivariate average test R2 for FCO2: 0.49
```

Model Cross-Comparison (MAPE Scores vs R2 Results) for Multi-Variate Models



XGBOOST: Best Deep Learning Model for Forecasting FCH_4 (Based on ↓ MAPE)



XGBOOST: Best Deep Learning Model for Forecasting FCO_2 (Based on ↓ MAPE)