STAT 430 Project 2

Predicting Methane Flux Emissions using Deep Learning

Kashyap Ava
vava2

May 07 2025

Abstract

Methane, a potent greenhouse gas, is critical to global warming. While its emissions are often analyzed in pasture lands due to cattle presence, croplands may also act as sources or sinks of methane. This study investigates methane flux emissions in a maize field to develop predictive deep learning models for accurately estimating the annual methane flux. Using time-series data, deep learning techniques, including Feed Forward Neural Network (FF), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM), were applied to assess methane flux variations. LSTM demonstrated the highest predictive reliability among these models with the lowest RMSE of 0.0111 and the highest R-squared of 0.84, effectively capturing seasonal dependencies in methane emissions. The results highlight the significance of incorporating temporal patterns into methane flux modeling and suggest further refinements to improve predictive accuracy.

Introduction

Motivation

Global warming is primarily driven by greenhouse gases such as carbon dioxide (CO2), carbon monoxide (CO), and methane (CH4). While CO2 and CO have been extensively studied, methane emissions, mainly from agricultural croplands, require further investigation due to their potential impact on climate change. Though the methane emissions were primarily analyzed in the context of pasture lands (in the presence of cattle), they can also be a significant factor in non-pasture croplands. In the long term, the study aims to model the temporal dependencies of the methane flux emissions based on a subset of the data to effectively gap-fill the missing data throughout the year, as depicted in Figure 1, to get a reliable annual methane flux estimate. This project employs deep learning models to predict methane flux emissions in a maize field from May to July, enabling effective gap-filling of missing data. Gap-filling is crucial for determining whether the cropland acts as a methane sink or source (based on the annual methane flux

estimate), which can inform strategies to either enhance maize production or mitigate emissions by targeting significant environmental predictors.

To summarize, the long-term objective of this study is to estimate annual methane emissions in the maize field accurately. At the same time, the short-term goal focuses on modeling the temporal dependencies of methane flux. The significance of both goals is reflected in the selection of environmental predictors, which play a key role in understanding methane flux variations.

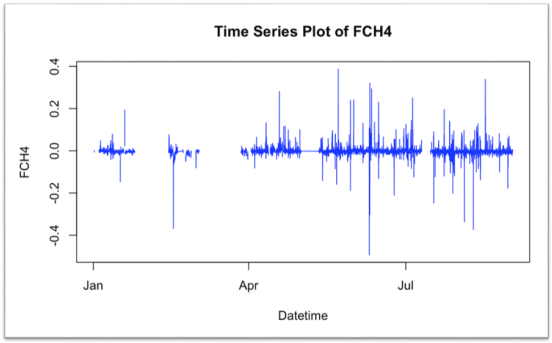


Figure 1: Time Series Plot of FCH4 in 2023

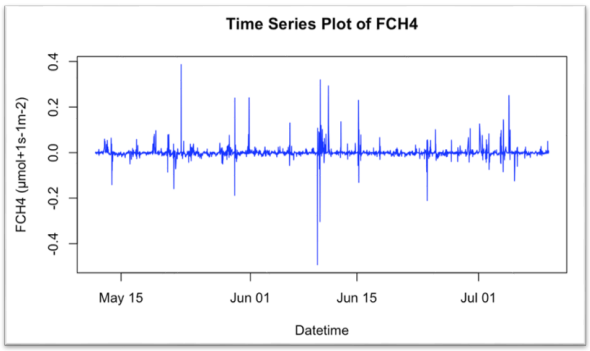


Figure 2: Time Series plot of FCH4 for the dataset

Data Description

The primary predictors selected from the literature review of gap-filling CO2 in Vekuri et al. (2023), which were also used previously in Project 1, are listed in Table 1. These predictors represent key environmental and meteorological variables influencing methane flux emissions. Radiation-related variables (SWin, SWout, LWin, LWout, Rn) measure the balance of incoming and outgoing shortwave and longwave radiation, affecting surface energy dynamics. Temperature variables (Ta, Tc, Ts) indicate air, canopy, and soil temperatures, influencing methane production and oxidation rates. Humidity and moisture-related factors (RH, SWC, P_rain) impact soil conditions and microbial activity and are critical for methane flux. Photosynthetic Photon Flux Density (PPFD, PPFDr) represents light availability for plant processes, which may indirectly affect methane emissions. Soil Heat Flux (SHF) captures heat transfer in the soil, influencing biological activity and gas exchange. Understanding these predictors helps in modeling the temporal dependencies of methane emissions and identifying the primary environmental drivers.

Abbreviation	Full Name	Unit
SWin	Shortwave Incoming Radiation	W/m^2

SWout	Shortwave Outgoing Radiation	W/m^2
LWin	Longwave Incoming Radiation	W/m^2
LWout	Longwave Outgoing Radiation	W/m^2
Rn	Net Radiation	W/m^2
Ta	Air Temperature	K
RH	Relative Humidity	%
Tc	Canopy Temperature	K
PPFD	Photosynthetic Photon Flux Density	$\mu mol/m^2/s$
PPFDr	Reflected Photosynthetic Photon Flux Density	$\mu mol/m^2/s$
Ts	Soil Temperature	K
SHF	Soil Heat Flux	W/m^2
SWC	Soil Water Content	m^3/m^3
P_rain	Rain Precipitation	mm

Table 1: Primary environmental predictors chosen from the literature review

Let's look at the Spearman correlation of the predictors with the methane flux emission (target). Spearman correlation was chosen for this analysis because it is a non-parametric rank-based measure that captures monotonic relationships between variables, making it robust to outliers and non-linear dependencies. Unlike Pearson correlation, which assumes a linear relationship and is sensitive to extreme values, Spearman correlation can identify strong associations even when the relationship is not strictly linear. Additionally, unlike the normality assumption in the Pearson correlation, there are no distributional assumptions for the Spearman correlation. Given that methane flux data and environmental variables may have complex, non-linear interactions, Spearman correlation provides a more reliable measure of association without imposing strict distributional assumptions. The Spearman correlation makes it a better

choice for feature selection and understanding predictor relevance in this study. The result of the Spearman correlation of the primary set of predictors (listed in Table 1) with the target variable is shown in Table 2.

	FCH4
FCH4	1.00000000
SWin	-0.27727763
SWout	-0.26439627
LWin	-0.06107978
LWout	-0.24063136
Rn	-0.26952222
Ta	-0.16952207
RH	0.17767913
Tc	-0.24060614
Ts	-0.09664943
P_rain	0.04787560
PPFD	-0.26480603
PPFDr	-0.26195552
SHF	-0.25116218
SWC	-0.02732446

Variable	Value
ch4_flux	1.0000000
un_ch4_flux	0.6289260
w/ch4_cov	0.6289260
T*	0.3280700
LE	0.3272453
un_LE	0.3270026
h2o_flux	0.3262913
ET	0.3262911
un_h2o_flux	0.3259612
w/h2o_cov	0.3259612
un_co2_flux	0.3156560
w/co2_cov	0.3156560
co2_flux	0.3054115
(z-d)/L	0.3033668
v_var	0.3012782

Table 2: Spearman Correlation between the primary set of predictors (left) and the added set of predictors (right) with the target (methane flux emission)

The primary set of predictors does not show a strong relationship with the methane flux emitted (target), as the maximum absolute value of the correlation observed was 0.277, which is low. It led me to consider adding more features, and I identified the predictors depicted in Table 2 based on a threshold of 0.3 for the Spearman correlation with the target. Out of the 15 newly identified predictors, 10 were chosen to be part of the predictors, as some of them were redundant in the sense that, after the initial preprocessing from the EddyPro for the crop data, both the uncorrected and the corrected fluxes were present and hence only one of the two was chosen. The details of the newly added predictors are given below in Table 3.

Abbreviation	**Full Name**	**Unit**
w/ch4_cov	Covariance between vertical wind speed and CH_4	$\mu mol/m^2/s$
T*	Scaling temperature related to turbulence	K
LE	Corrected latent heat flux	W/m^2
h2o_flux	Water vapor flux	$\mu mol/m^2/s$

ET	Evapotranspiration rate	mm/hour
w/h2o_cov	Covariance between vertical wind speed and H_2O	µmol/m²/s
w/co2_cov	Covariance between vertical wind speed and CO_2	µmol/m²/s
co2_flux	Net carbon dioxide flux	µmol/m²/s
(z-d)/L	Monin-Obukhov stability parameter	
v_var	Variance of the lateral (v) wind component	m^2/s^{-2}

Table 3: Newly identified and additional environmental predictors chosen for the analysis

The above newly added predictors are critical atmospheric and biophysical indicators used to understand trace gas and energy exchange in ecosystems. Covariance terms like w/ch4_cov, w/h2o_cov, and w/co2_cov quantify the exchange of gases (CH_4, H_2O, CO_2) between the surface and the atmosphere, based on their interaction with vertical wind speed. Flux variables (h2o_flux, co2_flux, LE) capture the net movement of water vapor, carbon dioxide, and latent heat, essential for quantifying evapotranspiration and photosynthesis. T* represents turbulent temperature scaling, and (z-d)/L measures atmospheric stability, helping classify the boundary layer conditions. Lastly, v_var offers insight into wind turbulence, influencing flux transport efficiency. Given that the appropriate features are chosen, after data cleaning, the dataset contains 2,726 rows of half-hourly readings of the target and the 26 predictors collected from a maize field between 15:00 on May 11, 2023, and 01:30 on July 10, 2023. The summary of the target variable (methane flux emitted), denoted as “FCH4”, is depicted in Appendix A.1. The time series data is plotted and shown in Figure 2.

In the next section, we look into some related inferences from Project 1 that would motivate us to choose the appropriate deep learning models.

Statistical Methods

Inferences from Project 1

Firstly, let's look at whether the data is a stationary time series, as it is a common assumption in time series modeling. I used the augmented Dickey-Fuller test to test for stationarity, and the test summary is depicted in Figure 3. I also checked if the data had any trends using the Mann-Kendall trend test, with the test summary in Figure 4.

```
        Augmented Dickey-Fuller Test

data:  backup_clean$FCH4
Dickey-Fuller = -11.532, Lag order = 13, p-value = 0.01
alternative hypothesis: stationary
```

Figure 3: Augmented Dickey-Fuller Test for Stationarity

```
        Mann-Kendall trend test

data:  backup_clean$FCH4
z = 1.3489, n = 2741, p-value = 0.1774
alternative hypothesis: true S is not equal to 0
sample estimates:
           S         varS          tau
6.454300e+04 2.289402e+09 1.718777e-02
```

Figure 4: Mann-Kendall Trend test

Since the p-value of the augmented Dickey-Fuller test is less than 0.05, we conclude that the data is stationary. From the trend test, we conclude that the data has no trends as the p-value is more significant than 0.05. Now, let's look at the plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to further analyze the plausible seasonality in the target variable (methane flux) in Figures 5 and 6.

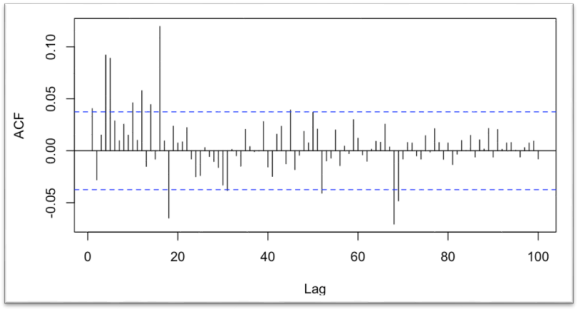


Figure 5: ACF plot of FCH4 plotted till lag 100

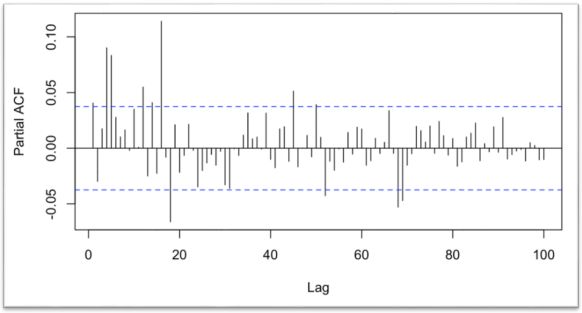


Figure 6: PACF plot of FCH4 plotted till lag 100

The ACF and PACF plots do not cut off at a certain lag, motivating us to look at the potential periods and seasonality. So, I used spectral analysis, preferably a periodogram, as depicted in Figure 7, to identify such periods. The periodogram was smoothed with a span parameter of 12 in the spectrum function in R.

From the periodogram, I have considered the highest peak frequency to be the dominant period in my analysis. This period was identified as 2.732 half-hours (since I am working with the data taken half hourly). This periodic nature has motivated me to use the SARIMA model in Project 1. While SARIMA demonstrated strong potential, alternative deep learning approaches could be explored to enhance predictive power and address some of the limitations observed. So, reliable predictive deep learning models were developed to improve the proportion of variance explained (approximately 0 for SARIMA).

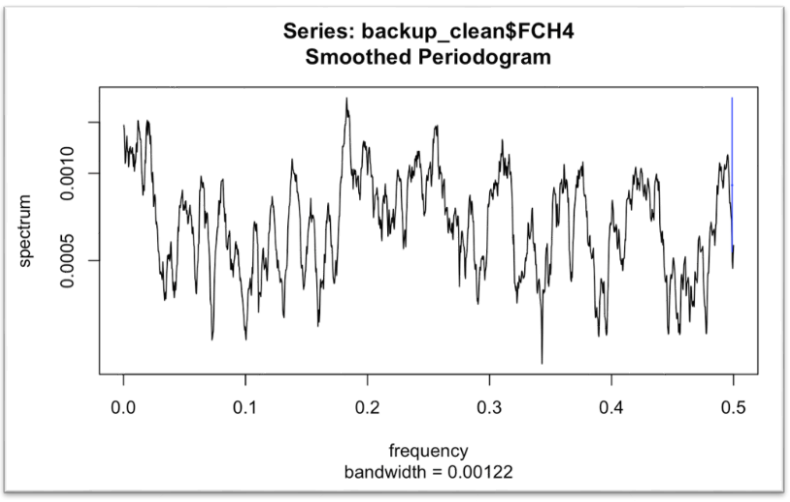


Figure 7: Smoothed Periodogram of FCH4 (methane flux emission) with a span of 12

The fully connected feed-forward neural network would be a great baseline deep learning model as it can capture non-linear relationships between environmental predictors and methane flux, which traditional statistical models often struggle to represent. Building on this, the Recurrent Neural Network (RNN) was introduced with feedback loops in a neural network, enabling the model to retain short-term memory and learn temporal dependencies across sequential data. The motivating factor was based on the idea that past precipitation can influence the methane flux spike in the present. Finally, the Long Short-Term Memory (LSTM) network was implemented to overcome the limitations of RNNs, using gated mechanisms to selectively retain relevant long-term patterns and forget less useful information, thereby significantly enhancing the model's ability to track temporal variability in methane emissions. The logic is supported by the possible idea that the time frame of the past precipitation event might be extended.

The following subsection will examine the modeling approaches based on the predictors' specific motivations and the data.

Modeling Approaches

Feed Forward Neural Network (FF)

A fully connected feed-forward neural network processes input data in a single direction, from the input layer, through one or more hidden layers, to the output layer. Each neuron in a layer is connected to every neuron in the subsequent layer, allowing the network to model complex, non-linear relationships between the environmental predictors and the methane flux. However, FFs do not account for temporal ordering in data, meaning they treat each time point as an independent observation. The architecture of a fully connected feed-forward neural network with two hidden layers is shown below in Figure 8.

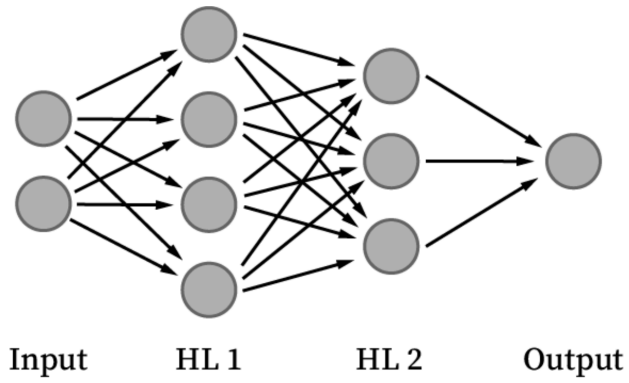


Figure 8: A fully connected feed-forward neural network with two hidden layers

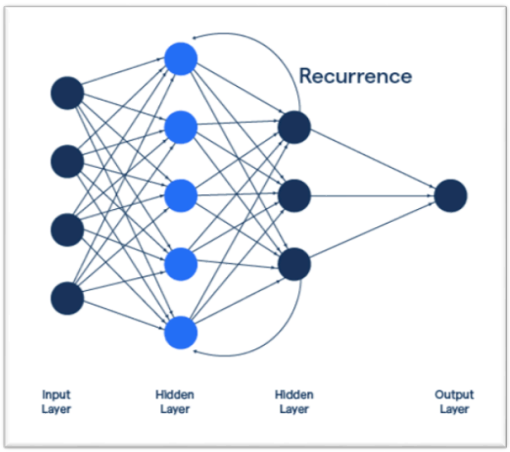


Figure 9: A recurrent neural network with two hidden layers

Recurrent Neural Network (RNN)

The RNN is an extension of the FF with recurrent connections between the units in different layers. The RNN architecture introduces recurrent connections, where outputs from the hidden layer at one time step are fed back into the network as inputs for the next step. This feedback loop, shown in Figure 9, allows the RNN to maintain a memory of previous time steps, making it suitable for modeling short-term temporal dependencies. However, RNNs often struggle with learning long-range dependencies due to issues like vanishing/ exploding gradients. This issue is primarily because the RNN continues to hold onto the coefficients while going deeper into the network, shrinking or expanding them repeatedly.

Long Short-Term Memory (LSTM)

The LSTM, illustrated in Figure 10, is a specialized type of RNN designed to retain long-term dependencies using a gated cell structure. Each LSTM unit contains gates that control how information is added, retained, or discarded over time. This structure helps the model selectively remember important patterns and forget irrelevant data, enabling it to effectively learn complex temporal relationships in methane flux time series data. LSTM's robust memory management makes it especially powerful for longer sequences. So, by overcoming the issues of the RNN, LSTM tends to capture the temporal dependencies of methane flux more accurately.

Model Evaluation Metrics

The deep learning models were trained using a fixed window approach, where the first 70% of the data was used for training, 20% for validation, and the final 10% reserved for testing. This chronological split respects the temporal structure of the data and ensures that future points are never used to predict the past,

avoiding data leakage. Unlike a rolling window, which updates the model incrementally and can accumulate errors over time, a fixed window allows the model to learn from a stable historical context and make robust predictions on unseen future data. This approach aligns with the project's long-term goal of gap-filling methane flux measurements, where the objective is to generate reliable estimates across extended periods, including for missing or unmeasured time intervals.

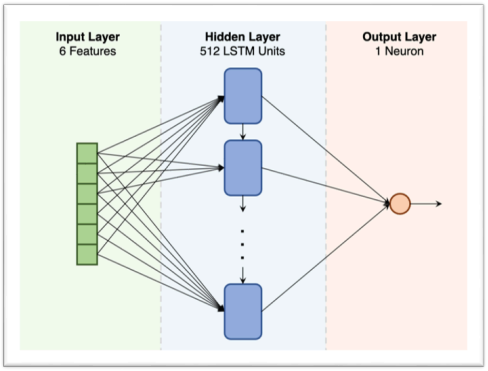


Figure 10: An LSTM with one hidden layer

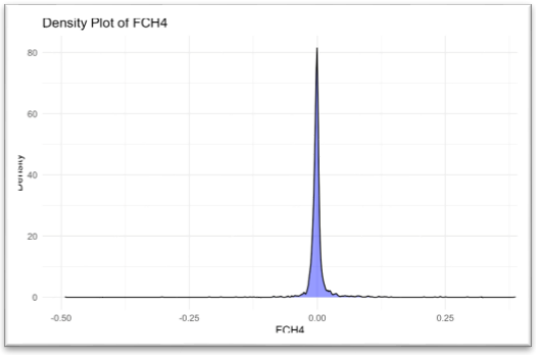


Figure 11: Density plot of FCH4

Since this is a regression problem, several performance metrics were employed to assess model accuracy and effectiveness. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) measure the absolute differences between predicted and actual values, providing insight into overall error magnitudes. Though RMSE and MAE work well in regression problems, the distribution of the target (methane flux emission) is shown above in Figure 11. So even if there is an error in prediction, both RMSE and MAE won't penalize them heavily as they are small in magnitude. So, Mean Absolute Percentage Error (MAPE) was particularly emphasized due to the small magnitude of methane flux values, offering a relative error measure to assess prediction accuracy more meaningfully. R-squared (R^2) was used to quantify the proportion of variance explained by the models. Together, these metrics provided a comprehensive model performance evaluation, helping identify the most effective approach for methane flux prediction.

Results

Model Performance

To ensure a fair comparison across model types, a consistent architecture was maintained with two hidden layers, comprising 26 units in the first layer and 13 in the second. All models used the ReLU activation function, the Adam optimizer, and were trained for 100 epochs with mean squared error (MSE) as the loss function. For the fully connected feed-forward (FF) model, a batch size of 13 was used to update weights after every mini-batch. In contrast, for RNN and LSTM models, training was conducted using the full sequence input, where the batch size is internally managed by the sequential data structure and not explicitly defined in the model.

Feed Forward Neural Network (FF)

The FF achieved an RMSE of 0.0138, MAE of 0.0085, and MAPE of 14.22%, with an R-squared value of 0.75. These results indicate that the FF moderately successfully captured the non-linear relationships between the predictors and methane flux. However, its inability to account for temporal dependencies likely limited its predictive performance, especially compared to sequential models like RNN and LSTM.

Recurrent Neural Network (RNN)

The RNN model improved performance over FF, with an RMSE of 0.0129, MAE of 0.0077, and a higher R-squared of 0.79. This improvement suggests that incorporating short-term memory through recurrent connections helped the model better track the temporal patterns in the methane flux data. However, the MAPE increased to 17.63%, indicating slightly higher relative errors, particularly on smaller flux values.

Long Short-Term Memory (LSTM)

LSTM outperformed FF and RNN in almost all metrics, with the lowest RMSE of 0.0111, MAE of 0.0067, and the highest R-squared value of 0.84. While its MAPE (17.60%) was slightly higher than FF's, this can be attributed to marginally larger errors on smaller values. LSTM's ability to retain long-term dependencies enhanced its predictive accuracy, making it the most robust model among the three. The model evaluation results are given in Table 4.

Model Comparison

Model	RMSE	MAE	MAPE	R-Squared
FF	0.0138	0.0085	**14.22%**	0.75
RNN	0.0129	0.0077	17.63%	0.79
LSTM	**0.0111**	**0.0067**	17.60%	**0.84**

Table 4: Model Evaluation Results

Overall, the LSTM model outperforms both FF and RNN, achieving the lowest RMSE (0.0111) and MAE (0.0067), and the highest R^2 score (0.84), demonstrating its superior ability to capture both the magnitude and variance of methane flux. This performance stems from LSTM's gated architecture, which enables it to retain long-term dependencies while filtering out irrelevant temporal noise. In contrast, the RNN shows a marked improvement over the FF, with better RMSE and R^2, reinforcing the importance of modeling temporal dependencies in time series data. While the LSTM's MAPE (17.60%) is slightly higher than FF's (14.22%), this primarily reflects minor relative errors on smaller flux values. Crucially, LSTM excels at predicting larger and more variable flux events, which are more consequential in environmental modeling. As such, despite the slight MAPE trade-off, LSTM offers a more robust and reliable model overall, balancing both temporal context and predictive precision better than its counterparts.

Discussion

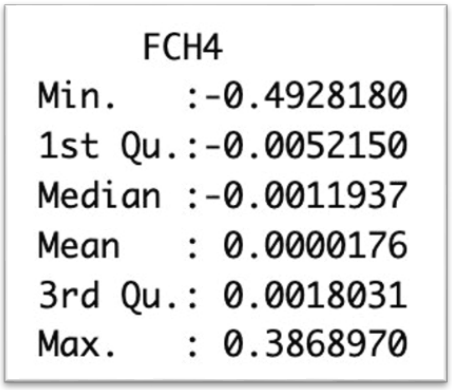
The results of this study underscore the effectiveness of deep learning architectures in modeling methane flux. Among the three models evaluated, the Long Short-Term Memory (LSTM) network emerged as the most accurate, achieving the lowest RMSE and MAE, and the highest R-squared value (0.84). Its gated architecture allowed it to retain long-term dependencies while selectively forgetting less relevant information, making it well-suited for sequential data with fluctuating trends. The Recurrent Neural Network (RNN) demonstrated improved performance over the Fully Connected Feed-Forward Network (FF) by incorporating short-term memory, highlighting the value of temporal modeling. However, both models fell short of LSTM's generalization ability across variable flux conditions. While LSTM exhibited a slightly higher MAPE, this was primarily attributed to relative errors on small flux values. It did not detract from its robustness, especially in capturing significant emission events. These findings suggest that deep learning models, particularly LSTM, can significantly improve over traditional models in forecasting methane flux, even in noisy and irregular patterns.

Despite their promising performance, deep learning models often act as "black boxes", offering limited insight into the influence of individual predictors. A key direction for future work involves developing interpretability tools that can help quantify the impact of environmental drivers on methane emissions. Exploring deeper and more advanced architectures, such as multi-layer LSTM or Gated Recurrent Units (GRU), could enhance model performance, particularly for multi-step forecasting. Given that the ultimate goal of this research is to generate a reliable, year-round estimate of methane flux, these models can also be deployed for gap-filling missing measurements in field observations, thereby supporting long-term flux integration and carbon budgeting. Finally, combining the strengths of deep learning with interpretable statistical techniques may lead to hybrid models that balance predictive power with scientific insight, contributing to both environmental modeling and decision-making in agricultural ecosystem management.

References

1. Vekuri H, Tuovinen JP, Kulmala L, Papale D, Kolari P, Aurela M, Laurila T, Liski J, Lohila A. A widely used eddy covariance gap-filling method creates a systematic bias in carbon balance estimates. Sci Rep. 2023 Jan 31;13(1):1720. doi: 10.1038/s41598-023-28827-2. PMID: 36720968; PMCID: PMC9889393.
2. R code for Data Cleaning: https://drive.google.com/file/d/1U3QisrHDSDUxWHoLhvZRPGr41iIM0B_6/view?usp=sharing
3. Python code for Modeling and Analysis: https://drive.google.com/file/d/1OPQcSD1_gM0KwoS-_Zw5Trj5kEJ7PRQY/view?usp=drive_link

Appendix



A.1: Summary of the target variable (methane flux emission)