**Machine Learning for Carbon Fluxes Prediction**

The sharp increase in greenhouse gas (GHG) emissions has been a major contributor to the global rise in temperatures. In response, the Paris Agreement, established in 2015, set the ambitious goal of achieving carbon neutrality by 2050. Consequently, accurate forecasting of carbon fluxes—the exchange of carbon between terrestrial ecosystems and the atmosphere—has become a crucial area of research. This has important implications for sustainable agriculture, forest management, and climate change mitigation efforts. Although physical methods such as the eddy covariance technique provide highly accurate carbon flux measurements [1], they are costly and spatially limited. For example, a single eddy covariance tower costs over £200,000, stands more than 50 metres tall, and can only monitor an area of around 10 km². This makes it difficult to measure carbon fluxes over larger spatial and temporal scales. In contrast, recent advancements in machine learning (ML) offer a scalable alternative by combining various datasets with carbon dynamics to predict carbon fluxes. ML models utilise affordable and widely available data sources such as climate records and remote sensing imagery, making it a promising approach for broader carbon flux estimation.

**Research Aim**

The aim of this research is to develop an advanced machine learning (ML) framework for accurate prediction of carbon fluxes. We collected carbon fluxes data from six eddy covariance towers across Europe, alongside corresponding climate and remote sensing data for each station. Climate and remote sensing data will serve as input features, while Net Ecosystem Exchange (NEE) will be the prediction target. The model can be built using various machine learning algorithms such as Random Forest (RF) [2] and XGBoost [7], or advanced deep learning approaches like Long Short-Term Memory (LSTM) networks [3][4], Transformers [5][7], Knowledge-guided Neural Networks (KGML) [6] to capture both temporal and spatial dynamics in carbon flux patterns.

**Dataset Description**

Notice: the IGBP is the types of terrestrial ecosystem, include permanent wetlands (WET) and evergreen needleleaf forest (ENF).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SITE\_ID | Country/Region | LATITUDE | LONGITUDE | ELEVATION | IGBP | START Time | END TIME |
| FI-Lom | Finland | 67.9972 | 24.2092 | 274 | WET | 21/02/2009 00:00:00 | 20/10/2009 23:00:00 |
| GL-ZaF | Denmark (Greenland) | 74.4814 | -20.5545 | 38 | WET | 01/05/2011 18:00:00 | 10/10/2011 23:00:00 |
| IE-Cra | Ireland | 53.32309 | -7.641774 | null | WET | 01/01/2020 01:00:00 | 31/12/2020 23:00:00 |
| DE-Akm | Germany | 53.8662 | 13.6834 | -1 | WET | 25/03/2020 08:00:00 | 07/10/2020 12:00:00 |
| FR-LGt | France | 47.3229 | 2.2841 | null | WET | 01/01/2022 00:00:00 | 02/11/2022 23:00:00 |
| UK-AMo | United Kingdom | 55.792546 | -3.2436918 | 270 | WET | 16/11/2021 17:00:00 | 31/12/2022 23:00:00 |
| SE-Htm | Sweden | 56.09763 | 13.41897 | 115 | ENF | 01/01/2016 01:00:00 | 14/08/2016 14:00:00 |

Input data：

|  |  |  |
| --- | --- | --- |
| Category | Variables | Description |
| Temporal information | timestamp | Pandas datetime object used for indexing |
| DOY | Day of year |
| TOD | Time of day |
| Climate information | CO2\_F\_MDS | CO2 concentration |
| G\_F\_MDS | Soil heat flux |
| H\_F\_MDS | Sensible heat flux |
| LE\_F\_MDS | Latent heat flux |
| LW\_IN\_F | Incoming longwave radiation |
| LW\_OUT | Outgoing longwave radiation |
| NETRAD | Net radiation |
| PA\_F | Air pressure |
| PPFD\_IN | Incoming photosynthetic photon flux density |
| PPFD\_OUT | Outgoing photosynthetic photon flux density |
| P\_F | Precipitation |
| RH | Relative humidity |
| SW\_IN\_F | Incoming shortwave radiation |
| SW\_OUT | Outgoing shortwave radiation |
| TA\_F | Air temperature |
| USTAR | Wind friction velocity |
| VPD\_F | Vapor pressure deficit |
| WD | Wind direction |
| WS\_F | Wind speed |
| Remote sensing data | MODIS\_band\_1 | MCD43A4 Band 1 |
| MODIS\_band\_2 | MCD43A4 Band 2 |
| MODIS\_band\_3 | MCD43A4 Band 3 |
| MODIS\_band\_4 | MCD43A4 Band 4 |
| MODIS\_band\_5 | MCD43A4 Band 5 |
| MODIS\_band\_6 | MCD43A4 Band 6 |
| MODIS\_band\_7 | MCD43A4 Band 7 |
| MODIS\_snow | MCD43A2 Band 1 |
| MODIS\_water | MCD43A2 Band 2 |

Target:

|  |  |
| --- | --- |
| NEE\_VUT\_REF | Net ecosystem exchange |

**Experimental settings:**

The dataset for this study includes data from seven eddy covariance flux towers. Specifically, data from the FI-Lom, GL-ZaF, IE-Cra, and DE-Akm sites will be used for training purposes, the FR-LGt site will serve as validation data, and the UK-Amo and SE-Htm sites will be reserved for testing. To evaluate the performance of different prediction algorithms, we will use root mean square error (rMSE) as the primary metric for assessing accuracy. Additionally, the R-squared (R²) statistic will be used to evaluate the overall regression performance, providing a measure of how well the predicted values align with the actual observations. The definitions of the rMSE and R2 metrics are as follows:

where for the rMSE, y represents the actual target value, and denotes the corresponding model prediction. In the case of , is the true target value for instance i, is the predicted value, and is the mean of the actual target values.

**Reference:**

[1] Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M. and Isaac, P., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific data, 7(1), p.225.

[2] Zeng, J., Matsunaga, T., Tan, Z.H., Saigusa, N., Shirai, T., Tang, Y., Peng, S. and Fukuda, Y., 2020. Global terrestrial carbon fluxes of 1999–2019 estimated by upscaling eddy covariance data with a random forest. Scientific data, 7(1), p.313.

[3] Shangguan, W., Xiong, Z., Nourani, V., Li, Q., Lu, X., Li, L., Huang, F., Zhang, Y., Sun, W. and Dai, Y., 2023. A 1 km global carbon flux dataset using in situ measurements and deep learning. Forests, 14(5), p.913.

[4] Nathaniel, J., Liu, J. and Gentine, P., 2023. MetaFlux: Meta-learning global carbon fluxes from sparse spatiotemporal observations. Scientific Data, 10(1), p.440.

[5] Phan, A. and Fukui, H., 2023. FluxFormer: Upscaled global carbon fluxes from eddy covariance data with multivariate timeseries Transformer.

[6] Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., Zhu, Q., Till, J., Jia, X., Jiang, C. and Wang, S., 2024. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. Nature communications, 15(1), p.357.

[7] Fortier, M., Richter, M.L., Sonnentag, O. and Pal, C., 2024. CarbonSense: A Multimodal Dataset and Baseline for Carbon Flux Modelling. arXiv preprint arXiv:2406.04940.