CarbonSense: A Multimodal Dataset and Baseline for Carbon Flux Modelling

Matthew Fortier*
Mila Quebec AI Institute
& Polytechnique Montréal

Mats L. Richter
ServiceNow

Oliver Sonnentag
Université de Montréal

Chris Pal
Mila Quebec AI Institute
& Polytechnique Montréal

Abstract

Terrestrial carbon fluxes provide vital information about our biosphere's health and its capacity to absorb anthropogenic CO_2 emissions. The importance of predicting carbon fluxes has led to the emerging field of data-driven carbon flux modelling (DDCFM), which uses statistical techniques to predict carbon fluxes from meteorological and geospatial data. However, the field lacks a standardized dataset to promote comparisons between models. To address this gap, we present CarbonSense, the first machine learning-ready dataset for DDCFM. CarbonSense integrates carbon fluxes, meteorological predictors, and satellite imagery from 385 locations across the globe, offering comprehensive coverage and facilitating robust model training. Additionally, we provide a baseline model using a current state of the the art DDCFM approach and a novel transformer based model. Our experiments illustrate the potential gains that multimodal deep learning techniques can bring to this domain. By providing these resources, we aim to lower the barrier to entry for other deep learning researchers to develop new models and drive new advances in carbon flux modelling.

1 Introduction

The biosphere plays a critical role in regulating Earth's climate. Since the mid-20th century, terrestrial ecosystems have absorbed up to a third of anthropogenic carbon emissions [1]. However, climate change introduces uncertainty about the future resilience and capacity of these ecosystems. Understanding the carbon dynamics of our biosphere and how those dynamics are changing in response to climate change, will give crucial insight into the health of our ecosystems and their ability to sequester carbon in the future.

Carbon fluxes describe the movement of carbon into and out of these ecosystems resulting from processes like photosynthesis and cellular respiration. They are often analogised as the "breathing of the biosphere" [2]. However, the large number of dependent processes makes these fluxes challenging to simulate with process-based models - traditional climate models which simulate biological, chemical, and physical processes with parameterized equations. Normally, large scale

*Address correspondence to matthew.fortier@mila.quebec

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

observational data helps climate scientists to parameterize their models, but carbon fluxes can only be measured in small areas with long-term sensor deployment, creating a data bottleneck.

In recent years however, machine learning techniques have been used to address this bottleneck. Researchers now use meteorological and geospatial predictors to upscale carbon flux data. This is known as data-driven carbon flux modelling (DDCFM), and it can be used to produce global carbon flux maps that are useful both as an end product for studying ecosystem health and as a benchmark for improving process-based models [3]. Most studies perform DDCFM with random forests [4]–[6], XGBoost [7], or ensembles of similar methods [8], [9]. While an XGBoost approach presently yields state of the art results for DDCFM, multimodal models are an extremely active area of research in the deep learning community, and further advances have the potential to continue to drive the evolution of these techniques.

Deep learning-based multimodal models have been successfully applied to many domains [10], including clinical diagnostics [11], land use cover classification [12], and wildfire surface fuel estimates [13]. The development of such techniques for improved carbon flux modelling could have important implications on the quality of information provided to decision makers, and significantly impact our ability to deal with climate change.

Our work seeks to bridge the gap between Deep Learning researchers and the DDCFM community by making the following contributions:

- We provide an overview of DDCFM for deep learning researchers (Section 2)
- We publish a multimodal ML-ready dataset for DDCFM (Section 3)
- Following best practices and prior work from the DDCFM community we provide a SOTA XGBoost based baseline and compare it with a novel multimodal deep learning model which achieves improved performance for DDCFM (Section 4)

We will discuss our experiments in section 5 and provide guidelines for reporting results in this domain.

2 Data-Driven Carbon Flux Modelling

At its core, DDCFM is a regression problem. The target (carbon flux) depends on many factors including ecosystem makeup, meteorological conditions, local topography and geology, and disturbances (fires, animal activity, etc). Meteorological data is easy to obtain, but the other predictors are challenging to measure and represent, especially at a global scale. Geospatial and semantic data are commonly employed as a proxy for the other predictors.

2.1 Measuring Fluxes

The most common technique for measuring fluxes is eddy covariance (EC) [14]. This is a micrometeorological technique where researchers erect a tower (typically above canopy height) and mount sensors that measure atmospheric gas fluxes across small turbulent vortices ("eddies"). A simplified EC station is depicted in Figure 1. CO2 and water vapour are the most widely measured, but some towers also measure methane (CH4) [4], [5] or nitrous oxide (N2O) [15]. Our work focuses on CO2 due to the prevalance of standardized data collections.

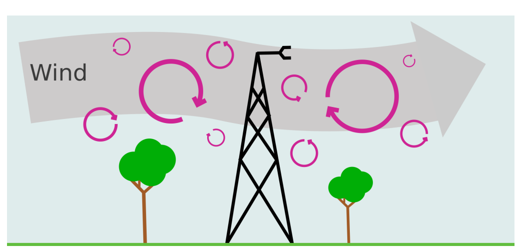

Figure 1: Simplified EC station. Sensors measure atmospheric CO2 gradients across small atmospheric vortices (eddies).

Carbon fluxes can be recorded and expressed in many ways, but most represent the movement of carbon in terms of mass / area / time (ex $g \cdot C \cdot m^{-2} \cdot hr$). Gross primary productivity (GPP) refers to the total carbon uptake by plants for photosynthesis. Ecosystem respiration (RECO) is the total carbon returned to the atmosphere through both plant and microbial respiration. Net ecosystem exchange (NEE) is the net carbon flux from GPP and RECO; a carbon sink will have a negative NEE as more carbon is being consumed through GPP than released through RECO. NEE is the flux that is directly measured by EC stations and is the main focus of our experiments, but GPP and RECO data are also provided in our dataset.

2.2 Flux Predictors

Meteorological Data DDCFM meteorological data comes from EC stations. In addition to carbon fluxes, EC stations measure local atmospheric conditions such as wind velocity and direction, precipitation, surface temperature, soil moisture, and others. The exact number and type of variables depends on the site, but regional networks maintain a minimum mandatory set for researchers wishing to submit their data [16]. For trained models looking to predict fluxes at the global scale, meteorological data can be obtained from publicly available reanalysis products such as ERA5 [17] which provides the variables on a 0.05 degree grid.

Geospatial Data Satellite imagery of the area surrounding an EC station can give useful information about the land cover and ecosystem makeup. The most common products for DDCFM are based on Moderate Resolution Imaging Spectroradiometer (MODIS) data [18]. This satellite constellation produces new imagery for Earth's surface every 1-2 days and has 36 spectral bands with resolutions varying between 250m and 1km. The MCD43A4 derived product is particularly common - it fuses MODIS data in a 16-day sliding window to produce a single image each day. This not only helps to address cloud coverage, but produces nadir BRDF-adjusted reflectance (NBAR) images which remove angle effects from directional reflectance [19]. Each image therefore appears as it would from directly overhead at solar noon. MCD43A2 is also widely used, which contains categorical values for each pixel indicating snow and water cover [20]. The terms "geospatial data", "satellite data" and "remote sensing data" are often used interchangeably in this domain.

Semantic Data Some models ingest semantic data such as ecosystem classification ("Croplands", "Evergreen needleleaf forest", "Snow and ice", etc). These classifications follow standardized schemes such as the International Geosphere-Biosphere Programme (IGBP). Ecosystem classification is performed by domain experts, but some MODIS products coarsely approximate this information on a global grid [21], allowing this data to also be used for global inference.

3 The CarbonSense Dataset

We present the first ML-ready dataset for DDCFM, CarbonSense. CarbonSense consists of EC station data and corresponding MODIS geospatial data for 385 sites across the globe, totalling over 27 million hourly observations. This section provides a brief overview of the dataset structure, processing pipeline, and usage guidelines. A more comprehensive guide is given in the supplementary material. For a detailed list of the 385 locations and their respective ecosystem types, see Appendix A.

3.1 Data Collection

All meteorological data was aggregated from major EC data networks, including FLUXNET 2015 [16], the Integrated Carbon Observation System (ICOS) 2023 release [22], ICOS Warm Winter release [23], and Ameriflux 2023 release [24]. These source datasets were chosen due to their use of the ONEFlux processing pipeline [16], ensuring standardized coding and units. A map of EC sites and their source networks is shown in Figure 2. North America and Europe are over-represented in this site list due greater data accessibility, and we discuss the implications of this in Section 3.3.

Geospatial data in CarbonSense are sourced from MODIS products. Specifically, we utilize the seven spectral bands from the MCD43A4 product [19], as well as the water and snow cover bands from MCD43A2 [20]. Following the guidelines from [18], we extract images in a 4km by 4km square centered on each EC station. Given a spatial resolution of 500m per pixel, this yields an 8x8 pixel image with 9 channels for every site-day.

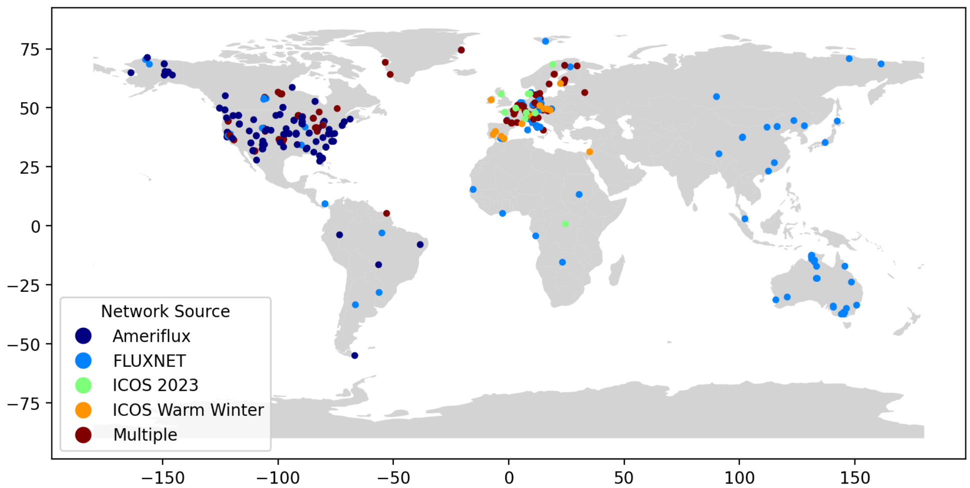

Figure 2: Global map of eddy covariance sites used in CarbonSense, with corresponding source networks. Some sites were present in multiple networks.

3.2 Data Pipeline

The first stage in the pipeline is EC data fusion. Many sites had overlapping data from different source networks. For example, the site Degero in Sweden (SE-Deg) had data from 2001-2020 in the ICOS Warm Winter release, and data from 2019-2022 in the ICOS 2023 release. Data were fused with overlapping values taken from the more recent release as in previous DDCFM work [7]. Any sites which report half-hourly data were downsampled to hourly at this stage, and daily and monthly recordings were discarded.

Once fused, we extracted the relevant time blocks for each EC station along with its geographic location. This metadata was used to obtain the appropriate MODIS data for each site. Data was pulled procedurally from Google Earth Engine [25].

Meteorological data was pruned to remove unwanted variables. Some, like soil moisture and temperature, were either unavailable for most sites or were heavily gap-filled. We removed these variables to reduce the risk of compounding errors on the underlying pipeline gapfilling techniques. A full list of variables at this stage is given in Table 6.

As a final stage in the pipeline, we apply a min-max normalization on predictor variables. We map cyclic variables (those with a cyclic range such as wind direction) to the range $[-1, 1)$ and acyclic variables to the range $[-0.5, 0.5)$. This normalization procedure is conducive to our Fourier encoding method discussed in Section 4.1.

We offer CarbonSense both before ("raw") and after ("normalized") our normalization step. Our pipeline can be configured to have variable "leniency" for gap-filled values. Those who wish to use CarbonSense with strictly observed values may do so at the cost of a smaller number of samples. The full pipeline code is also available so that researchers can add additional source networks with minimal modifications. A diagram of the entire pipeline is shown in Figure 3.

3.3 Using the Dataset

Site Sampling The geographic and ecological distribution of sites remains a challenge in DDCFM, and CarbonSense is no different. Given the significant overrepresentation of certain regions (North America, Europe) and ecosystems (evergreen needleleaf forests, grasslands), we maintain a partitioned

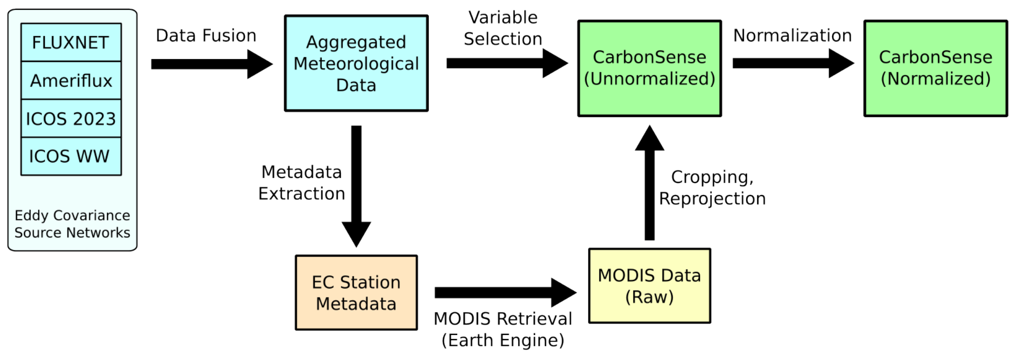

Figure 3: Data pipeline used to create CarbonSense from EC and MODIS data.

structure where each site has its own directory containing EC data, geospatial data, and metadata. Researchers are encouraged to select sites for training and testing based on their experiment objectives such as high performance on particular ecosystems, or out-of-distribution generalization. Our experiments in section 5 are an example of the latter.

Dataloader We supply an example PyTorch dataloader for CarbonSense specifically tailored to our baseline model. Using the dataloader requires specifying which carbon flux to use as the target, which sites to include in each dataloader instance, and the context window length for multi-timestep training.

Licensing CarbonSense is available under the CC-BY-4.0 license, meaning it can be shared, transformed, and used for any purpose given proper attribution. This is an extension of the same license for all three source networks, and MODIS data is provided under public domain. We feel that permissive licensing is essential in order to foster greater scientific interest in DDCFM.

4 The EcoPerceiver Architecture

In this section we present EcoPerceiver, a multimodal architecture for DDCFM. The SOTA for DDCFM are tabular methods, and we felt it would be appropriate to include a baseline model which demonstrates how deep learning concepts can be leveraged for this unique problem domain.

EcoPerceiver is based on the Perceiver architecture [26], which cross attends a variable number of inputs onto a compact latent space, allowing for extreme input flexibility. Missing inputs are common in DDCFM due to coverage gaps, outlier values, or failing sensors. Rather than rely on gapfilling techniques, we chose this architecture for its robustness to missing inputs.

We also wanted a model which could ingest data from a varying time window. To our knowledge, this is the first DDCFM model to treat carbon dynamics as non-Markovian with respect to predictors. We feel this more accurately reflects biological processes, since a plant's rate of photosynthesis may depend on conditions hours or days into the past. Our ablation experiments in Appendix B explore this idea further.

4.1 Data Ingestion

Small fluctuations in meteorological variables have the potential to influence ecological processes. For this reason, it is important that the model is sensitive to small changes in input values. We take inspiration from NeRF's Fourier encoding [27] which maps continuous values to higher dimensional space with high frequency sinusoids. As discussed previously, cyclic variables in CarbonSense are mapped to $[-0.5, 0.5)$ and acyclic variables are mapped to $[-1.0, 1.0)$. We start by taking each

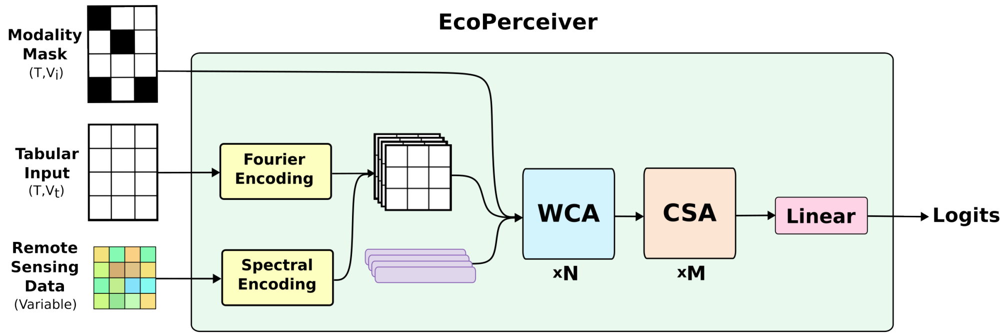

Figure 4: Overview of EcoPerceiver architecture.

variable x, and applying it to a series of sinusoids to produce an encoded vector with

$$f(x;K) = \left[\ldots, \sin(2^k \pi x), \cos(2^k \pi x), \ldots \,\middle|\, k \in [0, K)\right], \tag{1}$$

where K is a hyperparameter indicating the maximum sampling frequency. Higher values of K allow the model to better discern between small differences in input. With our normalization scheme, cyclic variables at values of -1 and 1 will produce identical vectors under this transform as intended. Each input is given a learned embedding specific to the underlying variable. This is then concatenated with the Fourier encoding to produce a final input vector of length $H_i = 2K + l_{emb}$ for each input.

Geospatial data is similarly processed, except that each spectral band is flattened into a vector of length $2K$ via linear transformation instead of Fourier encoding. Each band is then given an embedding to produce a vector of length H_i. We then stack the encoded data to create a matrix of shape (T, V_t, H_i) where V_t is the sum of the total number of variables and T is the context window length. Figure 5 depicts the encoding procedure.

To account for missing values and timesteps without geospatial data, EcoPerceiver takes a modality mask indicating which values to ignore in the cross attentive layers. This modality mask doubles as a dropout mechanism which reduces over-reliance on a small subset of variables (observational dropout).

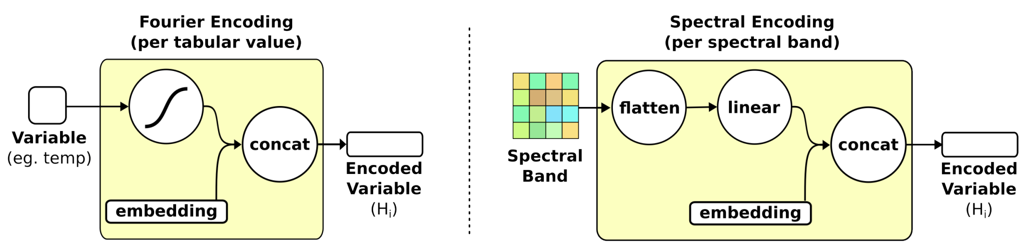

Figure 5: Input encoding for EcoPerceiver. Left: Tabular values are fed into the Fourier encoding function (1) and concatenated with an embedding. Right: Spectral bands are similarly processed except with linear projection instead of Fourier encoding.

4.2 Windowed Cross Attention

We build on Perceiver's core concept of cross-attending data onto a compact latent space for processing. EcoPerceiver uses a latent space of size (T, H_l) where H_l is the latent hidden dimension. Each token extracts input data via cross-attention from its respective timestep's observations. Intuitively, each token may representing the ecosystem's "state" at a particular time, and ingests observations from that timestep.

This operation would be inefficient with vanilla cross attention, as each token would use at most $\frac{1}{T}$ observations with an attention mask removing the rest. We take inspiration from SWin Transformer [28] and instead push the context window dimension (T) into the batch dimension for both input and latent space. The resulting Windowed Cross Attention (WCA) has a runtime of $O(T \cdot V_t \cdot H_a)$ where H_a is the projection dimension.

In keeping with Perceiver, each WCA operation is followed by a self-attention operation in the latent space. We pass a causal mask to the self attention so each timestep is conditioned only on past and present observations. We refer to this as Causal Self Attention (CSA). This constitutes a full WCA block as shown in Figure 6.

WCA blocks are repeated N times, repeatedly cross attending inputs onto the latent space with self attention in between. We then apply a series of M CSA operations and use the final timestep's token as input to a linear layer. The output of this is the estimate of the desired carbon flux.

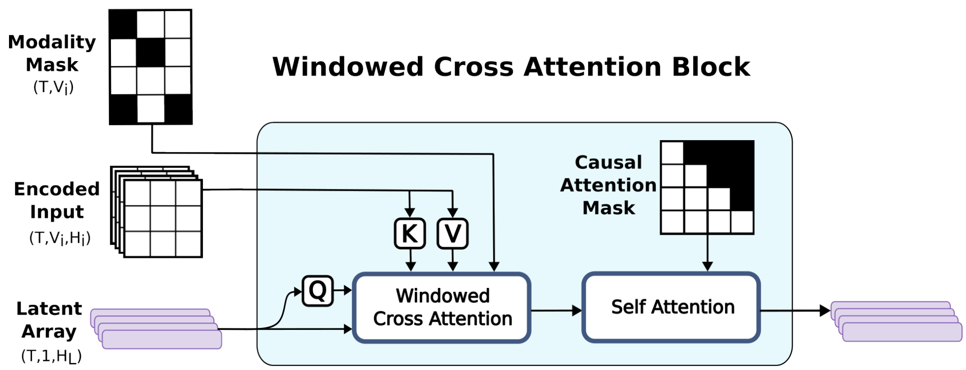

Figure 6: Windowed Cross Attention (WCA) block. Encoded inputs are cross-attended onto the latent space with a modality mask to indicate missing values. The time dimension is pushed into the batch dimension, so this operation is performed $B \cdot T$ times per batch. Causal self attention proceeds as normal.

5 Experiments

In this section, we present a series of experiments using CarbonSense. Our analysis includes two models: an EcoPerceiver model as introduced in 4, and an XGBoost model implemented to mimic current SOTA approaches in DDCFM. We demonstrate the power of tailored deep architectures for DDCFM and establish a robust baseline that will support and inspire future research efforts. We also present guidelines for running similar experiments and presenting results.

5.1 Data Splitting

EC stations were randomly divided into train and test sets based on their IGBP ecosystem classification (IGBP type). We mostly refer to IGBP types by their acronyms for brevity, but a list of IGBP types with expanded names is found in Appendix A.

Despite the imbalance of IGBP types in CarbonSense, we wanted the test set to be as balanced as possible. The number of sites in the test set were determined with $\min(5, \lceil 0.2 * \text{num\_sites} \rceil)$. This provided between 1 and 5 sites per IGBP type as shown in Table 1. As a consequence, SNO and DNF provide information about zero-shot generalization, and CVM and WAT about one-shot generalization.

The main focus of this research is on models trained *across* different ecosystem types, as opposed to other research studying DDCFM *within a single* type (ex: [4] [8] [9]). However, the partitioned nature

Table 1: Train / test split distribution by IGBP type

IGBP	CRO	CVM	CSH	DBF	DNF	EBF	ENF	GRA	MF	OSH
Train	44	1	5	42	0	10	80	59	10	25
Test	5	1	2	5	1	3	5	5	3	5
IGBP	SAV	SNO	WAT	WET	WSA					
Train	11	0	1	42	8					
Test	3	1	1	5	2					

of CarbonSense makes it flexible for different modelling objectives, such as individual ecosystems. We give an example of this in Appendix B.

5.2 Model Configurations

EcoPerceiver experiments were each run on 4 A100 GPUs using dataset parallelization. The train sites were further divided into train and validation splits at a 0.8 / 0.2 ratio respectively. We used the AdamW optimizer [29] with a learning rate of 4e-5 and a batch size of 4096. A single warm-up epoch was performed followed by a cosine annealing learning rate schedule over 20 epochs. All experiments converged between 6 and 13 epochs.

XGBoost experiments were run on CPU nodes. We designed our XGBoost experiments to resemble [7] as closely as possible. This allows us to compare EcoPerceiver's relative performance against a stand-in for the SOTA. Appendix B provides a detailed description of XGBoost data preprocessing, hyperparameters for all models, ablation studies, and more.

5.3 Metrics

The most commonly used performance metric in DDCFM (and any form of hydrologic modelling) is the Nash-Sutcliffe Modelling Efficiency (NSE) [30], described with the following equation:

$$\text{NSE}(x) = 1 - \frac{\sum_i (y_i - x_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{2}$$

where a value of 1 represents perfect correlation between x and y. A value of 0 represents the same performance as guessing the mean of y, and negative values indicate that the mean of y is a better predictor than x. NSE is more challenging to use directly as a loss function since it would require the dataloader to also provide the mean of the data for a given site or ecosystem type. We therefore use mean squared error (MSE) as a loss function and report its root (RMSE) as well as NSE in our results.

Data balance in results reporting is also a concern. At first glance, the data appears very imbalanced with respect to ecosystem prevalence. CarbonSense contains 64 grasslands (GRA) sites, but only 1 deciduous needleleaf forest (DNF). While this is an extreme gap, ecosystems are more diverse than IGBP types can capture; grasslands in central North America will differ significantly from those in Europe or Asia. Still, it is prudent to separate results by ecosystem type to give a better picture of model performance.

5.4 Results

We ran 10 experiments with each model using different seeds to get an accurate picture of performance. Table 2 shows the mean performance on the test set for each model over every IGBP type. We provide a breakdown with boxplots in Appendix B.

EcoPerceiver outperformed the XGBoost baseline across most IGBP types. XGBoost performed better in permanent wetlands (WET), water bodies (WAT), and evergreen needleleaf forests (ENF) by a slim margin. WAT is particularly far out of distribution (EC stations are mounted above lakes) and both models did worse than predicting the mean, indicating this could be an issue with data quantity and quality.

Besides WAT, EcoPerceiver did substantially better on zero- and one-shot tests. The NSE differential between the two models was +0.0486 on CVM, +0.1533 on DNF, and +0.3268 on SNO. High predictive power on out-of-distribution sites like this is especially important for modellers wishing to run inference on global data, where each grid cell is likely to be quite different from any of the training sites.

Table 2: NSE and RMSE by model and IGBP type

	XGBoost		EcoPerceiver	
IGBP	NSE	RMSE	NSE	RMSE
CRO	0.8066	3.2381	**0.8482**	**2.8677**
CSH	0.7510	1.5224	**0.7670**	**1.4709**
CVM	0.5277	5.5157	**0.5763**	**5.2236**
DBF	0.7250	4.0959	**0.7547**	**3.8678**
DNF	0.2803	4.0974	**0.4336**	**3.6322**
EBF	0.7966	4.6050	**0.8220**	**4.3070**
ENF	**0.7765**	**2.8141**	0.7694	2.8579
GRA	0.7461	3.2487	**0.7967**	**2.9059**
MF	0.7559	3.8633	**0.7717**	**3.7361**
OSH	0.5451	1.8796	**0.6060**	**1.7475**
SAV	0.5802	1.6514	**0.7368**	**1.3070**
SNO	-0.0370	1.4291	**0.2898**	**1.1816**
WAT	**-11.0524**	**3.1838**	-14.4010	3.5802
WET	**0.4530**	**2.2073**	0.4137	2.2830
WSA	0.6132	2.5153	**0.6267**	**2.4706**

Our results also underline the importance of using NSE as the main metric for evaluation. Consider the models' performance on open savannas (SAV). XGBoost had an RMSE of 1.6514 versus EcoPerceiver's 1.3070. The magnitude of difference is small, and both values are significantly lower than the RMSE of many other IGBP types. But XGBoost had an NSE of 0.5802 while EcoPerceiver achieved 0.7368 which is a significant improvement. Different ecosystems may have wildly different variances in their carbon fluxes, and NSE accounts for this by dividing the performance by the variance of the target.

6 Conclusion

Our work establishes a foothold for deep learning in the field of DDCFM. We provide an open source ML-ready dataset, CarbonSense, using EC station data and geospatial data from a variety of ecosystems. DDCFM is an inherently a time-dependent and multimodal task, and our baseline model EcoPerceiver demonstrates that recent advances in deep learning can unlock substantial performance gains in this domain. We implore more deep learning researchers to help develop this field further, because the potential of artificial intelligence to improve our world can only be realized if we actively apply it to solve pressing social and environmental issues.

Limitations Data diversity remains the biggest challenge in this domain. CarbonSense has a data imbalance in not only ecosystem types, but geographic location. Africa, Central Asia, and South America are all underrepresented. While these areas contain many EC stations, most do not have readily available data in ONEFlux format, presenting a barrier to their inclusion. Researchers should be aware of the consequences of developing models with imbalanced data, including poor performance in underrepresented areas.

References

[1] C. Seiler, J. Melton, V. Arora, *et al.*, "Are terrestrial biosphere models fit for simulating the global land carbon sink?" *Journal of Advances in Modeling Earth Systems*, vol. 14, 2021. DOI: 10.1029/2021MS002946.

[2] D. Baldocchi, K. Novick, T. Keenan, and M. Torn, "Ameriflux: Its impact on our understanding of the 'breathing of the biosphere', after 25 years," *Agricultural and Forest Meteorology*, vol. 348, p. 109 929, 2024, ISSN: 0168-1923. DOI: https://doi.org/10.1016/j.agrformet.2024.109929. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192324000443.

[3] A. Anav, P. Friedlingstein, M. Kidston, *et al.*, "Evaluating the land and ocean components of the global carbon cycle in the cmip5 earth system models," *Journal of Climate*, vol. 26, no. 18, pp. 6801–6843, 2013. DOI: 10.1175/JCLI-D-12-00417.1. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/26/18/jcli-d-12-00417.1.xml.

[4] O. Peltola, T. Vesala, Y. Gao, *et al.*, "Monthly gridded data product of northern wetland methane emissions based on upscaling eddy covariance observations," *Earth System Science Data*, vol. 11, no. 3, pp. 1263–1289, 2019. DOI: 10.5194/essd-11-1263-2019. [Online]. Available: https://essd.copernicus.org/articles/11/1263/2019/.

[5] G. McNicol, E. Fluet-Chouinard, Z. Ouyang, *et al.*, "Upscaling wetland methane emissions from the fluxnet-ch4 eddy covariance network (upch4 v1.0): Model development, network assessment, and budget comparison," *AGU Advances*, vol. 4, no. 5, e2023AV000956, 2023, e2023AV000956 2023AV000956. DOI: https://doi.org/10.1029/2023AV000956. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023AV000956. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023AV000956.

[6] A.-M. Virkkala, B. M. Rogers, J. D. Watts, *et al.*, "An increasing arctic-boreal co2 sink despite strong regional sources," *bioRxiv*, 2024. DOI: 10.1101/2024.02.09.579581. [Online]. Available: https://www.biorxiv.org/content/early/2024/02/12/2024.02.09.579581.

[7] J. A. Nelson, S. Walther, F. Gans, *et al.*, "X-base: The first terrestrial carbon and water flux products from an extended data-driven scaling framework, fluxcom-x," *EGUsphere*, vol. 2024, pp. 1–51, 2024. DOI: 10.5194/egusphere-2024-165. [Online]. Available: https://egusphere.copernicus.org/preprints/2024/egusphere-2024-165/.

[8] A.-M. Virkkala, J. Aalto, B. M. Rogers, *et al.*, "Statistical upscaling of ecosystem co2 fluxes across the terrestrial tundra and boreal domain: Regional patterns and uncertainties," *Global Change Biology*, vol. 27, no. 17, pp. 4040–4059, 2021. DOI: https://doi.org/10.1111/gcb.15659. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15659.

[9] C. Zhang, D. Brodylo, M. Rahman, M. A. Rahman, T. A. Douglas, and X. Comas, "Using an object-based machine learning ensemble approach to upscale evapotranspiration measured from eddy covariance towers in a subtropical wetland," *Science of The Total Environment*, vol. 831, p. 154 969, 2022, ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2022.154969. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969722020629.

[10] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023. DOI: 10.1109/TPAMI.2023.3275156.

[11] H.-Y. Zhou, Y. Yu, C. Wang, *et al.*, "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 743–755, Jun. 2023, ISSN: 2157-846X. DOI: 10.1038/s41551-023-01045-x. [Online]. Available: https://doi.org/10.1038/s41551-023-01045-x.

[12] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023. DOI: 10.1109/TGRS.2023.3284671.

[13] M. Alipour, I. La Puma, J. Picotte, *et al.*, "A multimodal data fusion and deep learning framework for large-scale wildfire surface fuel mapping," *Fire*, vol. 6, no. 2, 2023, ISSN: 2571-6255. DOI: 10.3390/fire6020036. [Online]. Available: https://www.mdpi.com/2571-6255/6/2/36.

[14] D. D. Baldocchi, "How eddy covariance flux measurements have contributed to our understanding of global change biology," *Global Change Biology*, vol. 26, no. 1, pp. 242–260, 2020. DOI: https://doi.org/10.1111/gcb.14807. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14807.

[15] M. Khalil, A. AlSayed, Y. Liu, and P. A. Vanrolleghem, "Machine learning for modeling n2o emissions from wastewater treatment plants: Aligning model performance, complexity, and interpretability," *Water Research*, vol. 245, p. 120 667, 2023, ISSN: 0043-1354. DOI: https://doi.org/10.1016/j.watres.2023.120667. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0043135423011077.

[16] G. Pastorello, C. Trotta, E. Canfora, *et al.*, "The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data," *Scientific Data*, vol. 7, no. 1, p. 225, Jul. 2020, ISSN: 2052-4463. DOI: 10.1038/s41597-020-0534-3. [Online]. Available: https://doi.org/10.1038/s41597-020-0534-3.

[17] H. Hersbach, B. Bell, P. Berrisford, *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. DOI: https://doi.org/10.1002/qj.3803. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

[18] S. Walther, S. Besnard, J. A. Nelson, *et al.*, "Technical note: A view from space on global flux towers by modis and landsat: The fluxneteo data set," *Biogeosciences*, vol. 19, no. 11, pp. 2805–2840, 2022. DOI: 10.5194/bg-19-2805-2022. [Online]. Available: https://bg.copernicus.org/articles/19/2805/2022/.

[19] C. Schaaf and Z. Wang, "MCD43A4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global - 500m V006," 2015b. [Online]. Available: https://www.umb.edu/spectralmass/v006/mcd43a4-nbar-product/.

[20] C. Schaaf and Z. Wang, "MCD43A2 MODIS/Terra+Aqua BRDF/Albedo Quality Daily L3 Global - 500m V006," 2015a. [Online]. Available: https://www.umb.edu/spectralmass/v006/mcd43a2-albedo-product/.

[21] D. Sulla-Menashe and M. A. Friedl, "User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product," May 2018. [Online]. Available: https://www.umb.edu/spectralmass/v006/mcd43a4-nbar-product/.

[22] ICOS RI, F. Apadula, S. Arnold, *et al.*, "ICOS Atmosphere Release 2023-1 of Level 2 Greenhouse Gas Mole Fractions of CO2, CH4, N2O, CO, meteorology and 14CO2, and flask samples analysed for CO2, CH4, N2O, CO, H2 and SF6.," 2023. DOI: https://doi.org/10.18160/VXCS-95EV. [Online]. Available: https://www.icos-cp.eu/data-products/atmosphere-release.

[23] Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre, "Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format—release 2022-1 (Version 1.0).," *ICOS Carbon Portal*, 2022. DOI: 10.18160/VXCS-95EV. [Online]. Available: https://www.icos-cp.eu/data-products/2G60-ZHAK.

[24] H. Chu, D. S. Christianson, Y.-W. Cheah, *et al.*, "Ameriflux base data pipeline to support network growth and data sharing," *Scientific Data*, vol. 10, no. 1, p. 614, Sep. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-02531-2. [Online]. Available: https://doi.org/10.1038/s41597-023-02531-2.

[25] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017. DOI: `10.1016/j.rse.2017.06.031`. [Online]. Available: `https://doi.org/10.1016/j.rse.2017.06.031`.

[26] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 18–24 Jul 2021, pp. 4651–4664.

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[28] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, (ICLR) 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. [Online]. Available: `https://openreview.net/forum?id=Bkg6RiCqY7`.

[30] R. McCuen, Z. Knight, and A. Cutter, "Evaluation of the nash–sutcliffe efficiency index," *Journal of Hydrologic Engineering - J HYDROL ENG*, vol. 11, Nov. 2006. DOI: `10.1061/(ASCE)1084-0699(2006)11:6(597)`.

[31] C. Wagner-Riddle, "AmeriFlux FLUXNET-1F CA-ER1 Elora Research Station, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832154`.

[32] B. Amiro, "AmeriFlux FLUXNET-1F CA-MA1 Manitoba Agricultural Site 1, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007165`.

[33] B. Amiro, "AmeriFlux FLUXNET-1F CA-MA2 Manitoba Agricultural Site 2, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007166`.

[34] D. Billesbach, L. Kueppers, M. Torn, and S. Biraud, "AmeriFlux FLUXNET-1F US-A74 ARM SGP milo field, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006963`.

[35] S. Biraud, M. Fischer, S. Chan, and M. Torn, "AmeriFlux FLUXNET-1F US-ARM ARM Southern Great Plains site- Lamont, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1854366`.

[36] C. Rey-Sanchez, C. T. Wang, D. Szutu, *et al.*, "AmeriFlux FLUXNET-1F US-Bi1 Bouldin Island Alfalfa, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871134`.

[37] C. Rey-Sanchez, C. T. Wang, D. Szutu, K. Hemes, J. Verfaillie, and D. Baldocchi, "AmeriFlux FLUXNET-1F US-Bi2 Bouldin Island corn, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871135`.

[38] C. L. Phillips and D. Huggins, "AmeriFlux FLUXNET-1F US-CF1 CAF-LTAR Cook East, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832158`.

[39] D. Huggins, "AmeriFlux FLUXNET-1F US-CF2 CAF-LTAR Cook West, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881573`.

[40] D. Huggins, "AmeriFlux FLUXNET-1F US-CF3 CAF-LTAR Boyd North, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881574`.

[41] D. Huggins, "AmeriFlux FLUXNET-1F US-CF4 CAF-LTAR Boyd South, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881575`.

[42] J. Chen and H. Chu, "AmeriFlux FLUXNET-1F US-CRT Curtice Walter-Berger cropland, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006974`.

[43] A. Desai, "AmeriFlux FLUXNET-1F US-CS1 Central Sands Irrigated Agricultural Field, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881576`.

[44] A. Desai, "AmeriFlux FLUXNET-1F US-CS3 Central Sands Irrigated Agricultural Field, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881578`.

[45] A. Desai, “AmeriFlux FLUXNET-1F US-CS4 Central Sands Irrigated Agricultural Field, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881579.

[46] A. Duff and A. Desai, “AmeriFlux FLUXNET-1F US-DFC US Dairy Forage Research Center, Prairie du Sac, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2006975.

[47] M. R. Schuppenhauer, S. C. Biraud, S. Deverel, and S. Chan, “AmeriFlux FLUXNET-1F US-DS3 Staten Rice 1, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229376.

[48] S. Fares, “AmeriFlux FLUXNET-1F US-Lin Lindcove Orange Orchard, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229381.

[49] A. Schreiner-McGraw, “AmeriFlux FLUXNET-1F US-Mo1 LTAR CMRB Field 1 (CMRB ASP), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229382.

[50] A. Schreiner-McGraw, “AmeriFlux FLUXNET-1F US-Mo3 LTAR CMRB Field 3 (CMRB BAU), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229384.

[51] A. Suyker, “AmeriFlux FLUXNET-1F US-Ne1 Mead - irrigated continuous maize site, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1871140.

[52] M. R. Schuppenhauer, S. C. Biraud, and S. Chan, “AmeriFlux FLUXNET-1F US-RGA Arkansas Corn Farm, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2204873.

[53] M. Schuppenhauer, S. C. Biraud, and S. Chan, “AmeriFlux FLUXNET-1F US-RGB Butte County Rice Farm, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2204874.

[54] M. R. Schuppenhauer, S. C. Biraud, and S. Chan, “AmeriFlux FLUXNET-1F US-RGo Glenn County Organic Rice Farm, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2204875.

[55] J. Baker, T. Griffis, and T. Griffis, “AmeriFlux FLUXNET-1F US-Ro1 Rosemount- G21, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881588.

[56] J. Baker and T. Griffis, “AmeriFlux FLUXNET-1F US-Ro2 Rosemount- C7, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2204876.

[57] J. Baker and T. Griffis, “AmeriFlux FLUXNET-1F US-Ro5 Rosemount I18$_{S}outh, Ver.3-5$,” *AmeriFlux AMP*, 2021. DOI: http://doi.org/10.17190/AMF/1818371.

[58] J. Baker and T. Griffis, “AmeriFlux FLUXNET-1F US-Ro6 Rosemount I18$_{N}orth, Ver.3-5$,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881590.

[59] C. Sturtevant, J. Verfaillie, and D. Baldocchi, “AmeriFlux FLUXNET-1F US-Tw2 Twitchell Corn, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881593.

[60] S. D. Chamberlain, P. Oikawa, C. Sturtevant, D. Szutu, J. Verfaillie, and D. Baldocchi, “AmeriFlux FLUXNET-1F US-Tw3 Twitchell Alfalfa, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881594.

[61] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xSL NEON North Sterling, CO (STER), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229411.

[62] B. Drake, R. Hinkle, R. Bracho, S. Dore, and T. Powell, “AmeriFlux FLUXNET-1F US-KS2 Kennedy Space Center (scrub oak), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229380.

[63] G. Flerchinger, “AmeriFlux FLUXNET-1F US-Rls RCEW Low Sagebrush, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229387.

[64] G. Flerchinger, “AmeriFlux FLUXNET-1F US-Rms RCEW Mountain Big Sagebrush, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881587.

[65] G. Flerchinger and M. L. Reba, “AmeriFlux FLUXNET-1F US-Rwe RCEW Reynolds Mountain East, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1871143.

[66] G. Flerchinger, “AmeriFlux FLUXNET-1F US-Rwf RCEW Upper Sheep Prescibed Fire, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881591.

[67] S. Goslee, “AmeriFlux FLUXNET-1F US-HWB USDA ARS Pasture Sytems and Watershed Management Research Unit- Hawbecker Site, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881582.

[68] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xDS NEON Disney Wilderness Preserve (DSNY), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/1985439.

[69] R. Staebler, “AmeriFlux FLUXNET-1F CA-Cbo Ontario - Mixed Deciduous, Borden Forest Site, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1854365.

[70] M. A. Arain, “AmeriFlux FLUXNET-1F CA-TPD Ontario - Turkey Point Mature Deciduous, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881567.

[71] E. A. Yepez and J. Garatuza, “AmeriFlux FLUXNET-1F MX-Tes Tesopaco, secondary tropical dry forest, Ver. 3-5,” *AmeriFlux AMP*, 2021. DOI: http://doi.org/10.17190/AMF/1832156.

[72] A. Richardson and D. Hollinger, “AmeriFlux FLUXNET-1F US-Bar Bartlett Experimental Forest, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2006969.

[73] J. W. Munger, “AmeriFlux FLUXNET-1F US-Ha1 Harvard Forest EMS Tower (HFR1), Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1871137.

[74] K. Novick and R. Phillips, “AmeriFlux FLUXNET-1F US-MMS Morgan Monroe State Forest, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1854369.

[75] J. Wood and L. Gu, “AmeriFlux FLUXNET-1F US-MOz Missouri Ozark Site, Ver. 3-5,” *AmeriFlux AMP*, 2021. DOI: http://doi.org/10.17190/AMF/1854370.

[76] J. Chen, H. Chu, and A. Noormets, “AmeriFlux FLUXNET-1F US-Oho Oak Openings, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229385.

[77] M. Ueyama, H. Iwata, and Y. Harazono, “AmeriFlux FLUXNET-1F US-Rpf Poker Flat Research Range: Succession from fire scar to deciduous forest, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229388.

[78] C. Gough, G. Bohrer, and P. Curtis, “AmeriFlux FLUXNET-1F US-UMB Univ. of Mich. Biological Station, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2204882.

[79] C. Gough, G. Bohrer, and P. Curtis, “AmeriFlux FLUXNET-1F US-UMd UMBS Disturbance, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881597.

[80] J. Chen, “AmeriFlux FLUXNET-1F US-Wi1 Intermediate hardwood (IHW), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229394.

[81] J. Chen, “AmeriFlux FLUXNET-1F US-Wi3 Mature hardwood (MHW), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229395.

[82] J. Chen, “AmeriFlux FLUXNET-1F US-Wi8 Young hardwood clearcut (YHW), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229400.

[83] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xBL NEON Blandy Experimental Farm (BLAN), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229405.

[84] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xBR NEON Bartlett Experimental Forest (BART), Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881598.

[85] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xGR NEON Great Smoky Mountains National Park, Twin Creeks (GRSM), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/1985440.

[86] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xHA NEON Harvard Forest (HARV), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/1985441.

[87] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xML NEON Mountain Lake Biological Station (MLBS), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985447`.

[88] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xSC NEON Smithsonian Conservation Biology Institute (SCBI), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229409`.

[89] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xSE NEON Smithsonian Environmental Research Center (SERC), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985452`.

[90] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xST NEON Steigerwaldt Land Services (STEI), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985454`.

[91] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xTR NEON Treehaven (TREE), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985456`.

[92] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xUK NEON The University of Kansas Field Station (UKFS), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985457`.

[93] A. Antonino, "AmeriFlux FLUXNET-1F BR-CST Caatinga Serra Talhada, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902820`.

[94] T. A. Black, "AmeriFlux FLUXNET-1F CA-Ca1 British Columbia - 1949 Douglas-fir stand, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007163`.

[95] T. A. Black, "AmeriFlux FLUXNET-1F CA-Ca2 British Columbia - Clearcut Douglas-fir stand (harvested winter 1999/2000), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007164`.

[96] T. A. Black, "AmeriFlux FLUXNET-1F CA-LP1 British Columbia - Mountain pine beetle-attacked lodgepole pine stand , Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832155`.

[97] M. Goulden, "AmeriFlux FLUXNET-1F CA-NS1 UCI-1850 burn site, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902824`.

[98] M. Goulden, "AmeriFlux FLUXNET-1F CA-NS2 UCI-1930 burn site, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902825`.

[99] M. Goulden, "AmeriFlux FLUXNET-1F CA-NS3 UCI-1964 burn site, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902826`.

[100] M. Goulden, "AmeriFlux FLUXNET-1F CA-NS4 UCI-1964 burn site wet, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902827`.

[101] M. Goulden, "AmeriFlux FLUXNET-1F CA-NS5 UCI-1981 burn site, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902828`.

[102] H. A. Margolis, "AmeriFlux FLUXNET-1F CA-Qfo Quebec - Eastern Boreal, Mature Black Spruce, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006960`.

[103] B. Amiro, "AmeriFlux FLUXNET-1F CA-SF1 Saskatchewan - Western Boreal, forest burned in 1977, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902831`.

[104] B. Amiro, "AmeriFlux FLUXNET-1F CA-SF2 Saskatchewan - Western Boreal, forest burned in 1989, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006961`.

[105] M. A. Arain, "AmeriFlux FLUXNET-1F CA-TP1 Ontario - Turkey Point 2002 Plantation White Pine, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006962`.

[106] M. A. Arain, “AmeriFlux FLUXNET-1F CA-TP3 Ontario - Turkey Point 1974 Plantation White Pine, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881566`.

[107] E. Euskirchen, “AmeriFlux FLUXNET-1F US-BZS Bonanza Creek Black Spruce, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881572`.

[108] A. Desai, “AmeriFlux FLUXNET-1F US-CS2 Tri county school Pine Forest, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881577`.

[109] S. Dore and T. Kolb, “AmeriFlux FLUXNET-1F US-Fmf Flagstaff - Managed Forest, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007173`.

[110] S. Dore and T. Kolb, “AmeriFlux FLUXNET-1F US-Fuf Flagstaff - Unmanaged Forest, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007174`.

[111] J. Frank and B. Massman, “AmeriFlux FLUXNET-1F US-GLE GLEES, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871136`.

[112] J. D. Forsythe, M. A. Kline, and T. L. O’Halloran, “AmeriFlux FLUXNET-1F US-HB2 Hobcaw Barony Mature Longleaf Pine, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229377`.

[113] J. D. Forsythe, M. A. Kline, and T. L. O’Halloran, “AmeriFlux FLUXNET-1F US-HB3 Hobcaw Barony Longleaf Pine Restoration, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229378`.

[114] D. Hollinger, “AmeriFlux FLUXNET-1F US-Ho2 Howland Forest (west tower), Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881581`.

[115] B. Drake, R. Hinkle, R. Bracho, T. Powell, and S. Dore, “AmeriFlux FLUXNET-1F US-KS1 Kennedy Space Center (slash pine), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229379`.

[116] B. Law, “AmeriFlux FLUXNET-1F US-Me1 Metolius - Eyerly burn, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902834`.

[117] B. Law, “AmeriFlux FLUXNET-1F US-Me2 Metolius mature ponderosa pine, Ver. 3-5,” *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1854368`.

[118] B. Law, “AmeriFlux FLUXNET-1F US-Me3 Metolius-second young aged pine, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902835`.

[119] B. Law, “AmeriFlux FLUXNET-1F US-Me6 Metolius Young Pine Burn, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2204871`.

[120] A. Noormets, G. Sun, M. Gavazzi, S. McNulty, J.-C. Domec, and J. King, “AmeriFlux FLUXNET-1F US-NC1 NC$_Clearcut, Ver.3 - 5$,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902836`.

[121] A. Noormets, M. Gavazzi, M. Aguilos, J. King, B. Mitra, and J.-C. Domec, “AmeriFlux FLUXNET-1F US-NC3 NC$_Clearcut3, Ver.3 - 5$,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2204872`.

[122] P. D. Blanken, R. K. Monson, S. P. Burns, D. R. Bowling, and A. A. Turnipseed, “AmeriFlux FLUXNET-1F US-NR1 Niwot Ridge Forest (LTER NWT1), Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871141`.

[123] M. Litvak, “AmeriFlux FLUXNET-1F US-Vcm Valles Caldera Mixed Conifer, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229391`.

[124] M. Litvak, “AmeriFlux FLUXNET-1F US-Vcp Valles Caldera Ponderosa Pine, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229392`.

[125] J. Chen, “AmeriFlux FLUXNET-1F US-Wi0 Young red pine (YRP), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229393`.

[126] J. Chen, “AmeriFlux FLUXNET-1F US-Wi4 Mature red pine (MRP), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229396`.

[127] J. Chen, “AmeriFlux FLUXNET-1F US-Wi5 Mixed young jack pine (MYJP), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229397`.

[128] J. Chen, "AmeriFlux FLUXNET-1F US-Wi9 Young Jack pine (YJP), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229401`.

[129] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xAB NEON Abby Road (ABBY), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229403`.

[130] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xBN NEON Caribou Creek - Poker Flats Watershed (BONA), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229406`.

[131] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xDJ NEON Delta Junction (DEJU), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229407`.

[132] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xJE NEON Jones Ecological Research Center (JERC), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985443`.

[133] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xRM NEON Rocky Mountain National Park, CASTNET (RMNP), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985450`.

[134] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xSB NEON Ordway-Swisher Biological Station (OSBS), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985451`.

[135] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xTA NEON Talladega National Forest (TALL), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985455`.

[136] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xYE NEON Yellowstone Northern Range (Frog Rock) (YELL), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985459`.

[137] B. Amiro, "AmeriFlux FLUXNET-1F CA-MA3 Manitoba Agricultural Site 3, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007167`.

[138] D. Billesbach, L. Kueppers, M. Torn, and S. Biraud, "AmeriFlux FLUXNET-1F US-A32 ARM-SGP Medford hay pasture, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881568`.

[139] D. Billesbach, J. Bradford, and M. Torn, "AmeriFlux FLUXNET-1F US-AR1 ARM USDA UNL OSU Woodward Switchgrass 1, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006965`.

[140] D. Billesbach, J. Bradford, and M. Torn, "AmeriFlux FLUXNET-1F US-AR2 ARM USDA UNL OSU Woodward Switchgrass 2, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006966`.

[141] M. Torn, "AmeriFlux FLUXNET-1F US-ARb ARM Southern Great Plains burn site- Lamont, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006967`.

[142] M. Torn, "AmeriFlux FLUXNET-1F US-ARc ARM Southern Great Plains control site-Lamont, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006968`.

[143] K. Novick, "AmeriFlux FLUXNET-1F US-BRG Bayles Road Grassland Tower, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006970`.

[144] D. Bowling, "AmeriFlux FLUXNET-1F US-Cop Corral Pocket, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006972`.

[145] H. Liu, M. Huang, and X. Chen, "AmeriFlux FLUXNET-1F US-Hn2 Hanford 100H grassland, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902832`.

[146] N. Brunsell, "AmeriFlux FLUXNET-1F US-KFS Kansas Field Station, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881585`.

[147] N. Brunsell, "AmeriFlux FLUXNET-1F US-KLS Kansas Land Institute, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1854367`.

[148] N. Brunsell, "AmeriFlux FLUXNET-1F US-Kon Konza Prairie LTER (KNZ), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2316062`.

[149] A. Schreiner-McGraw, "AmeriFlux FLUXNET-1F US-Mo2 LTAR CMRB Tucker Prairie (CMRB TP), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229383`.

[150] M. Torn and S. Dengel, "AmeriFlux FLUXNET-1F US-NGC NGEE Arctic Council, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902838`.

[151] M. L. Silveira and R. Bracho, "AmeriFlux FLUXNET-1F US-ONA Florida pine flatwoods, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832163`.

[152] J. Baker and T. Griffis, "AmeriFlux FLUXNET-1F US-Ro4 Rosemount Prairie, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881589`.

[153] R. Scott, "AmeriFlux FLUXNET-1F US-SRG Santa Rita Grassland, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2204877`.

[154] M. Litvak, "AmeriFlux FLUXNET-1F US-Seg Sevilleta grassland, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1984572`.

[155] R. Shortt, K. Hemes, D. Szutu, J. Verfaillie, and D. Baldocchi, "AmeriFlux FLUXNET-1F US-Sne Sherman Island Restored Wetland, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871144`.

[156] K. Kusak, C. R. Sanchez, D. Szutu, and D. Baldocchi, "AmeriFlux FLUXNET-1F US-Snf Sherman Barn, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1854371`.

[157] S. Ma, L. Xu, J. Verfaillie, and D. Baldocchi, "AmeriFlux FLUXNET-1F US-Var Vaira Ranch-Ione, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1993904`.

[158] R. Scott, "AmeriFlux FLUXNET-1F US-Wkg Walnut Gulch Kendall Grasslands, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1984575`.

[159] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xAE NEON Klemme Range Research Station (OAES), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985434`.

[160] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xCL NEON LBJ National Grassland (CLBJ), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985435`.

[161] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xCP NEON Central Plains Experimental Range (CPER), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985436`.

[162] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xDC NEON Dakota Coteau Field School (DCFS), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985437`.

[163] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xKA NEON Konza Prairie Biological Station - Relocatable (KONA), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985444`.

[164] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xKZ NEON Konza Prairie Biological Station (KONZ), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985445`.

[165] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xNG NEON Northern Great Plains Research Laboratory (NOGP), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985448`.

[166] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xWD NEON Woodworth (WOOD), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229412`.

[167] H. McCaughey, "AmeriFlux FLUXNET-1F CA-Gro Ontario - Groundhog River, Boreal Mixedwood Forest, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902823`.

[168] A. Desai, “AmeriFlux FLUXNET-1F US-Syv Sylvania Wilderness Area, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2204879`.

[169] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xDL NEON Dead Lake (DELA), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985438`.

[170] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xUN NEON University of Notre Dame Environmental Research Center (UNDE), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985458`.

[171] M. Goulden, “AmeriFlux FLUXNET-1F CA-NS6 UCI-1989 burn site, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902829`.

[172] no$_{i}nfo_{a}vailable$, “no$_{i}nfo_{a}vailable$,” *no*$_{i}nfo_{a}vailable$, no$_{i}nfo_{a}vailable$. DOI: `no_info_available`.

[173] R. Bracho, G. Celis, H. Rodenhizer, C. See, and E. A. Schuur, “AmeriFlux FLUXNET-1F US-EML Eight Mile Lake Permafrost thaw gradient, Healy Alaska., Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007170`.

[174] M. Ueyama, H. Iwata, and Y. Harazono, “AmeriFlux FLUXNET-1F US-Fcr Cascaden Ridge Fire Scar, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007172`.

[175] H. Liu, M. Huang, and X. Chen, “AmeriFlux FLUXNET-1F US-Hn3 Hanford 100H sagebrush, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881580`.

[176] E. Euskirchen, G. Shaver, and S. Bret-Harte, “AmeriFlux FLUXNET-1F US-ICh Imnavait Creek Watershed Heath Tundra, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2007175`.

[177] E. Euskirchen, G. Shaver, and S. Bret-Harte, “AmeriFlux FLUXNET-1F US-ICt Imnavait Creek Watershed Tussock Tundra, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881583`.

[178] C. Tweedie, “AmeriFlux FLUXNET-1F US-Jo1 Jornada Experimental Range Bajada Site, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902833`.

[179] E. R. Vivoni and E. R. Perez-Ruiz, “AmeriFlux FLUXNET-1F US-Jo2 Jornada Experimental Range Mixed Shrubland, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881584`.

[180] G. Flerchinger, “AmeriFlux FLUXNET-1F US-Rws Reynolds Creek Wyoming big sagebrush, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881592`.

[181] S. Kurc, “AmeriFlux FLUXNET-1F US-SRC Santa Rita Creosote, Ver. 3-5,” *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871145`.

[182] M. Litvak, “AmeriFlux FLUXNET-1F US-Ses Sevilleta shrubland, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1984573`.

[183] R. Scott, “AmeriFlux FLUXNET-1F US-Whs Walnut Gulch Lucky Hills Shrub, Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1984574`.

[184] J. Chen, “AmeriFlux FLUXNET-1F US-Wi6 Pine barrens 1 (PB1), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229398`.

[185] J. Chen, “AmeriFlux FLUXNET-1F US-Wi7 Red pine clearcut (RPCC), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229399`.

[186] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xHE NEON Healy (HEAL), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985442`.

[187] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xJR NEON Jornada LTER (JORN), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229408`.

[188] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xMB NEON Moab (MOAB), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985446`.

[189] N. (E. O. Network), “AmeriFlux FLUXNET-1F US-xNQ NEON Onaqui-Ault (ONAQ), Ver. 3-5,” *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985449`.

[190] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xSR NEON Santa Rita Experimental Range (SRER), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/1985453`.

[191] R. Scott, "AmeriFlux FLUXNET-1F US-LS2 San Pedro River Lewis Springs Savanna, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2204870`.

[192] M. Litvak, "AmeriFlux FLUXNET-1F US-Wjs Willard Juniper Savannah, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871146`.

[193] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xSJ NEON San Joaquin Experimental Range (SJER), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229410`.

[194] M. Torn and S. Dengel, "AmeriFlux FLUXNET-1F US-NGB NGEE Arctic Barrow, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832162`.

[195] A. Desai, "AmeriFlux FLUXNET-1F US-Pnp Lake Mendota, Picnic Point Site, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229386`.

[196] G. Bohrer, "AmeriFlux FLUXNET-1F US-UM3 Douglas Lake, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881596`.

[197] L. Kutzbach, "AmeriFlux FLUXNET-1F AR-TF1 Rio Moat bog, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1818370`.

[198] A. Todd and E. Humphreys, "AmeriFlux FLUXNET-1F CA-ARB Attawapiskat River Bog, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902821`.

[199] A. Todd and E. Humphreys, "AmeriFlux FLUXNET-1F CA-ARF Attawapiskat River Fen, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1902822`.

[200] T. Papakyriakou, "AmeriFlux FLUXNET-1F CA-CF1 Churchill Fen Site 1, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2229375`.

[201] S. Knox, "AmeriFlux FLUXNET-1F CA-DB2 Delta Burns Bog 2, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881564`.

[202] A. Christen and S. Knox, "AmeriFlux FLUXNET-1F CA-DBB Delta Burns Bog, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881565`.

[203] T. Roman, T. Griffis, R. Kolka, *et al.*, "AmeriFlux FLUXNET-1F PE-QFR Quistococha Forest Reserve, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832157`.

[204] B. Olson, "AmeriFlux FLUXNET-1F US-ALQ Allequash Creek Site, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: `http://doi.org/10.17190/AMF/2006964`.

[205] E. Euskirchen, "AmeriFlux FLUXNET-1F US-BZB Bonanza Creek Thermokarst Bog, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881569`.

[206] E. Euskirchen, "AmeriFlux FLUXNET-1F US-BZF Bonanza Creek Rich Fen, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881570`.

[207] E. Euskirchen, "AmeriFlux FLUXNET-1F US-BZo Bonanza Creek Old Thermokarst Bog, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881571`.

[208] P. Oikawa, "AmeriFlux FLUXNET-1F US-EDN Eden Landing Ecological Reserve, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832159`.

[209] J. D. Forsythe, M. A. Kline, and T. L. O'Halloran, "AmeriFlux FLUXNET-1F US-HB1 North Inlet Crab Haul Creek, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: `http://doi.org/10.17190/AMF/1832160`.

[210] E. Euskirchen, G. Shaver, and S. Bret-Harte, "AmeriFlux FLUXNET-1F US-ICs Imnavait Creek Watershed Wet Sedge Tundra, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1871138`.

[211] R. Bracho and C. R. Hinkle, "AmeriFlux FLUXNET-1F US-KS3 Kennedy Space Center (salt marsh), Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: `http://doi.org/10.17190/AMF/1881586`.

[212] J. H. Matthes, C. Sturtevant, P. Oikawa, *et al.*, "AmeriFlux FLUXNET-1F US-Myb Mayberry Wetland, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1871139.

[213] A. Noormets, J. King, B. Mitra, *et al.*, "AmeriFlux FLUXNET-1F US-NC4 NC$_{A}lligator River, Ver. 3-5$," *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1902837.

[214] G. Bohrer, "AmeriFlux FLUXNET-1F US-ORv Olentangy River Wetland Research Park, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: http://doi.org/10.17190/AMF/1832164.

[215] G. Bohrer and J. Kerns, "AmeriFlux FLUXNET-1F US-OWC Old Woman Creek, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1871142.

[216] B. Bergamaschi and L. Windham-Myers, "AmeriFlux FLUXNET-1F US-Srr Suisun marsh - Rush Ranch, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229389.

[217] R. Vargas, "AmeriFlux FLUXNET-1F US-StJ St Jones Reserve, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229390.

[218] A. Valach, R. Shortt, D. Szutu, *et al.*, "AmeriFlux FLUXNET-1F US-Tw1 Twitchell Wetland West Pond, Ver. 3-5," *AmeriFlux AMP*, 2021. DOI: http://doi.org/10.17190/AMF/1832165.

[219] E. Eichelmann, R. Shortt, S. Knox, *et al.*, "AmeriFlux FLUXNET-1F US-Tw4 Twitchell East End Wetland, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2204881.

[220] A. Valach, K. Kasak, D. Szutu, J. Verfaillie, and D. Baldocchi, "AmeriFlux FLUXNET-1F US-Tw5 East Pond Wetland, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881595.

[221] J. Chen and H. Chu, "AmeriFlux FLUXNET-1F US-WPT Winous Point North Marsh, Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229402.

[222] N. (E. O. Network), "AmeriFlux FLUXNET-1F US-xBA NEON Barrow Environmental Observatory (BARR), Ver. 3-5," *AmeriFlux AMP*, 2023. DOI: http://doi.org/10.17190/AMF/2229404.

[223] G. Vourlitis, H. Dalmagro, J. de S. Nogueira, M. Johnson, and P. Arruda, "AmeriFlux FLUXNET-1F BR-Npw Northern Pantanal Wetland, Ver. 3-5," *AmeriFlux AMP*, 2022. DOI: http://doi.org/10.17190/AMF/1881563.

[224] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[225] D. PAPALE and R. VALENTINI, "A new assessment of european forests carbon exchanges by eddy fluxes and artificial neural network spatialization," *Global Change Biology*, vol. 9, no. 4, pp. 525–535, 2003. DOI: https://doi.org/10.1046/j.1365-2486.2003.00609.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2486.2003.00609.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2486.2003.00609.x.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Our limitations cover this
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] There is no PII in our dataset, and the applications are largely non-commercial.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Briefly discussed in section 5, expanded upon in section B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix B
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section 3.1 for data attributions. Appendix contains full EC site list.
   (b) Did you mention the license of the assets? [Yes] Section 3.3 has info on licensing of EC data and MODIS data
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Covered under licensing
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our dataset contains no PII.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices

A Eddy Covariance Site Details

Here we provide an exhaustive list of EC sites used in CarbonSense along with their most recent publication. As per Ameriflux's data policy, each site has an individual citation with DOI; other networks simply required citation of the unified release. It would be impractical to have each site's full description in these tables, but the first two letters of each code represent the country where the site is located (ex. "DE" for Germany).

We also enumerate all meteorological predictors and targets in table 6.

Table 3: EC Sites

Croplands (CRO)

BE-Lon [22]	CA-ER1 [31]	CA-MA1 [32]	CA-MA2 [33]	CH-Oe2 [23]	CZ-KrP [23]
DE-Geb [22]	DE-Kli [22]	DE-RuS [22]	DE-Seh [16]	DK-Fou [16]	DK-Vng [22]
FI-Jok [16]	FI-Qvd [23]	FR-Aur [22]	FR-EM2 [22]	FR-Gri [22]	FR-Lam [22]
IT-BCi [23]	IT-CA2 [16]	US-A74 [34]	US-ARM [35]	US-Bi1 [36]	US-Bi2 [37]
US-CF1 [38]	US-CF2 [39]	US-CF3 [40]	US-CF4 [41]	US-CRT [42]	US-CS1 [43]
US-CS3 [44]	US-CS4 [45]	US-DFC [46]	US-DS3 [47]	US-Lin [48]	US-Mo1 [49]
US-Mo3 [50]	US-Ne1 [51]	US-RGA [52]	US-RGB [53]	US-RGo [54]	US-Ro1 [55]
US-Ro2 [56]	US-Ro5 [57]	US-Ro6 [58]	US-Tw2 [59]	US-Tw3 [60]	US-Twt [16]
US-xSL [61]					

Closed Shrublands (CSH)

BE-Maa [22]	IT-Noe [16]	US-KS2 [62]	US-Rls [63]	US-Rms [64]	US-Rwe [65]
US-Rwf [66]					

Cropland/Natural Vegetation Mosaics (CVM)

US-HWB [67]	US-xDS [68]				

Deciduous Broadleaf Forests (DBF)

AU-Lox [16]	BE-Lcr [22]	CA-Cbo [69]	CA-Oas [16]	CA-TPD [70]	CZ-Lnz [22]
CZ-Stn [23]	DE-Hai [22]	DE-Hzd [23]	DE-Lnf [16]	DK-Sor [22]	FR-Fon [22]
FR-Hes [22]	IT-BFt [22]	IT-CA1 [16]	IT-CA3 [16]	IT-Col [16]	IT-Isp [16]
IT-PT1 [16]	IT-Ro1 [16]	IT-Ro2 [16]	JP-MBF [16]	MX-Tes [71]	PA-SPn [16]
US-Bar [72]	US-Ha1 [73]	US-MMS [74]	US-MOz [75]	US-Oho [76]	US-Rpf [77]
US-UMB [78]	US-UMd [79]	US-WCr [16]	US-Wi1 [80]	US-Wi3 [81]	US-Wi8 [82]
US-xBL [83]	US-xBR [84]	US-xGR [85]	US-xHA [86]	US-xML [87]	US-xSC [88]
US-xSE [89]	US-xST [90]	US-xTR [91]	US-xUK [92]	ZM-Mon [16]	

Deciduous Needleleaf Forests (DNF)

BR-CST [93]					

Evergreen Broadleaf Forests (EBF)

AU-Cum [16]	AU-Rob [16]	AU-Wac [16]	AU-Whr [16]	AU-Wom [16]	BR-Sa3 [16]
CN-Din [16]	FR-Pue [22]	GF-Guy [22]	GH-Ank [16]	IT-Cp2 [22]	IT-Cpz [16]
MY-PSO [16]					

Table 4: EC Sites (cont'd)

Evergreen Needleleaf Forests (ENF)

AR-Vir [16]	CA-Ca1 [94]	CA-Ca2 [95]	CA-LP1 [96]	CA-Man [16]	CA-NS1 [97]
CA-NS2 [98]	CA-NS3 [99]	CA-NS4 [100]	CA-NS5 [101]	CA-Obs [16]	CA-Qfo [102]
CA-SF1 [103]	CA-SF2 [104]	CA-TP1 [105]	CA-TP2 [16]	CA-TP3 [106]	CA-TP4 [16]
CH-Dav [22]	CN-Qia [16]	CZ-BK1 [22]	CZ-RAJ [23]	DE-Lkb [16]	DE-Msr [22]
DE-Obe [23]	DE-RuW [22]	DE-Tha [22]	DK-Gds [22]	FI-Hyy [22]	FI-Ken [22]
FI-Let [22]	FI-Sod [16]	FI-Var [22]	FR-Bil [22]	FR-FBn [23]	FR-LBr [16]
IL-Yat [23]	IT-La2 [16]	IT-Lav [23]	IT-Ren [22]	IT-SR2 [22]	IT-SRo [16]
NL-Loo [16]	RU-Fy2 [23]	RU-Fyo [23]	SE-Htm [22]	SE-Nor [22]	SE-Ros [23]
SE-Svb [22]	US-BZS [107]	US-Blo [16]	US-CS2 [108]	US-Fmf [109]	US-Fuf [110]
US-GBT [16]	US-GLE [111]	US-HB2 [112]	US-HB3 [113]	US-Ho2 [114]	US-KS1 [115]
US-Me1 [116]	US-Me2 [117]	US-Me3 [118]	US-Me4 [16]	US-Me5 [16]	US-Me6 [119]
US-NC1 [120]	US-NC3 [121]	US-NR1 [122]	US-Prr [16]	US-Vcm [123]	US-Vcp [124]
US-Wi0 [125]	US-Wi2 [16]	US-Wi4 [126]	US-Wi5 [127]	US-Wi9 [128]	US-xAB [129]
US-xBN [130]	US-xDJ [131]	US-xJE [132]	US-xRM [133]	US-xSB [134]	US-xTA [135]
US-xYE [136]					

Grasslands (GRA)

AT-Neu [16]	AU-DaP [16]	AU-Emr [16]	AU-Rig [16]	AU-Stp [16]	AU-TTE [16]
AU-Ync [16]	BE-Dor [23]	CA-MA3 [137]	CH-Aws [23]	CH-Cha [23]	CH-Fru [23]
CH-Oe1 [16]	CN-Cng [16]	CN-Dan [16]	CN-Du2 [16]	CN-Du3 [16]	CN-HaM [16]
CN-Sw2 [16]	CZ-BK2 [16]	DE-Gri [22]	DE-RuR [22]	DK-Eng [16]	FR-Mej [22]
FR-Tou [22]	GL-ZaH [22]	IT-MBo [23]	IT-Niv [22]	IT-Tor [22]	NL-Hor [16]
PA-SPs [16]	RU-Ha1 [16]	SE-Deg [22]	US-A32 [138]	US-AR1 [139]	US-AR2 [140]
US-ARb [141]	US-ARc [142]	US-BRG [143]	US-Cop [144]	US-Goo [16]	US-Hn2 [145]
US-IB2 [16]	US-KFS [146]	US-KLS [147]	US-Kon [148]	US-Mo2 [149]	US-NGC [150]
US-ONA [151]	US-Ro4 [152]	US-SRG [153]	US-Seg [154]	US-Sne [155]	US-Snf [156]
US-Var [157]	US-Wkg [158]	US-xAE [159]	US-xCL [160]	US-xCP [161]	US-xDC [162]
US-xKA [163]	US-xKZ [164]	US-xNG [165]	US-xWD [166]		

Mixed Forests (MF)

AR-SLu [16]	BE-Bra [22]	BE-Vie [22]	CA-Gro [167]	CD-Ygb [22]	CH-Lae [23]
CN-Cha [16]	DE-Har [22]	DE-HoH [22]	JP-SMF [16]	US-Syv [168]	US-xDL [169]
US-xUN [170]					

Open Shrublands (OSH)

CA-NS6 [171]	CA-NS7 [172]	CA-SF3 [16]	ES-Agu [23]	ES-Amo [16]	ES-LJu [23]
ES-LgS [16]	ES-Ln2 [16]	GL-Dsk [22]	IT-Lsn [22]	RU-Cok [16]	US-EML [173]
US-Fcr [174]	US-Hn3 [175]	US-ICh [176]	US-ICt [177]	US-Jo1 [178]	US-Jo2 [179]
US-Rws [180]	US-SRC [181]	US-Ses [182]	US-Sta [16]	US-Whs [183]	US-Wi6 [184]
US-Wi7 [185]	US-xHE [186]	US-xJR [187]	US-xMB [188]	US-xNQ [189]	US-xSR [190]

Savannas (SAV)

AU-ASM [16]	AU-Cpr [16]	AU-DaS [16]	AU-Dry [16]	AU-GWW [16]	CG-Tch [16]
ES-Abr [23]	ES-LM1 [23]	ES-LM2 [23]	SD-Dem [16]	SN-Dhr [16]	US-LS2 [191]
US-Wjs [192]	US-xSJ [193]				

Snow and Ice (SNO)

US-NGB [194]					

Table 5: EC Sites (cont'd)

Water Bodies (WAT)

US-Pnp [195]	US-UM3 [196]				

Permanent Wetlands (WET)

AR-TF1 [197]	AU-Fog [16]	CA-ARB [198]	CA-ARF [199]	CA-CF1 [200]	CA-DB2 [201]
CA-DBB [202]	CN-Ha2 [16]	CZ-wet [22]	DE-Akm [23]	DE-SfN [16]	DE-Spw [16]
DE-Zrk [16]	DK-Skj [22]	FI-Lom [16]	FI-Sii [22]	FR-LGt [22]	GL-NuF [22]
GL-ZaF [16]	IE-Cra [23]	PE-QFR [203]	RU-Che [16]	SE-Sto [22]	SJ-Adv [16]
UK-AMo [22]	US-ALQ [204]	US-Atq [16]	US-BZB [205]	US-BZF [206]	US-BZo [207]
US-EDN [208]	US-HB1 [209]	US-ICs [210]	US-Ivo [16]	US-KS3 [211]	US-Los [16]
US-Myb [212]	US-NC4 [213]	US-ORv [214]	US-OWC [215]	US-Srr [216]	US-StJ [217]
US-Tw1 [218]	US-Tw4 [219]	US-Tw5 [220]	US-WPT [221]	US-xBA [222]	

Woody Savannas (WSA)

AU-Ade [16]	AU-Gin [16]	AU-How [16]	AU-RDF [16]	BR-Npw [223]	ES-Cnd [23]

Table 6: Meteorological Variables in CarbonSense

Code	Description	Units
Predictors		
TA_F	Air temperature	deg C
PA_F	Atmospheric pressure	kPa
P_F	Precipitation	mm
RH	Relative humidity	%
VPD_F	Vapor pressure deficit	hPa
WS_F	Wind speed	m s-1
WD	Wind direction	decimal degrees
USTAR	Frictional wind velocity	m s^{-1}
NETRAD	Net radiation	W m^{-2}
SW_IN_F	Incoming shortwave radiation	W m^{-2}
SW_OUT	Outgoing shortwave radiation	W m^{-2}
SW_DIF	Incoming diffuse shortwave radiation	W m^{-2}
LW_IN_F	Incoming longwave radiation	W m^{-2}
LW_OUT	Outgoing longwave radiation	W m^{-2}
PPFD_IN	Incoming photosynthetic photon flux density	µmol Photon m^{-2} s^{-1}
PPFD_OUT	Outgoing photosynthetic photon flux density	µmol Photon m^{-2} s^{-1}
PPFD_DIF	Incoming diffuse photosynthetic photon flux density	µmol Photon m^{-2} s^{-1}
CO2_F_MDS	CO2 atmospheric concentration	µmol CO_2 mol^{-1}
G_F_MDS	Soil heat flux	W m^{-2}
LE_F_MDS	Latent heat flux	W m^{-2}
H_F_MDS	Sensible heat flux	W m^{-2}
Targets		
NEE_VUT_REF	Net Ecosystem Exchange (variable USTAR)	µmol CO_2 m^{-2} s^{-1}
GPP_DT_VUT_REF	Gross Primary Production (daytime partitioning)	µmol CO_2 m^{-2} s^{-1}
GPP_NT_VUT_REF	Gross Primary Production (nighttime partitioning)	µmol CO_2 m^{-2} s^{-1}
RECO_DT_VUT_REF	Ecosystem Respiration (daytime partitioning)	µmol CO_2 m^{-2} s^{-1}
RECO_NT_VUT_REF	Ecosystem Respiration (nighttime partitioning)	µmol CO_2 m^{-2} s^{-1}

B Experiment Details

B.1 Dataset Configuration

As part of the CarbonSense pipeline, we filter out poorly gapfilled values. Each variable in the EC data has a corresponding "quality check" (QC) flag indicating if it was directly measured, gap filled during the ONEFlux pipeline (with varying gap fill quality levels), or simply taken from ERA5 reanalysis products. The tolerance level can be configured during the CarbonSense normalization process, and we discuss this further in the supplementary material.

We chose to use a maximum QC flag of 1, indicating all values in the dataset were either directly measured, or gap filled with high confidence. We found this had the best trade-off of data quality and quantity, as setting the maximum QC flag to 0 (only directly measured values) reduced the dataset size by 55%, while including medium-confidence values only increased it by 9%.

The data split was randomized within each ecosystem type. We held out 20% or 5 sites for each type, whichever is lower. The remaining sites were divided 80/20 between training and validation sets for EcoPerceiver, while our XGBoost model used all the training data for a cross-validation procedure.

B.2 EcoPerceiver Configuration

Hyperparameter tuning for EcoPerceiver was performed with our train and validation splits, and comprised the bulk of the experiment efforts. Where possible, we started with our best guesses and ran a pseudo-random search based on intuition. A true random search of the parameter space would have been extremely sparse given the available compute resources.

We set our latent hidden size to 128, our input embedding size to 16, and the number of Fourier encoding frequencies to 12. This gave a total input hidden size of 40. Our context window is 32, meaning our model sees the previous 32 hours of observations. We use 8 WCA blocks followed by 4 CSA blocks. We set our observational dropout at 0.3 and use causal masking in all self-attention blocks. In keeping with [26], we employ weight sharing between all WCA blocks. With these hyperparameters our model weighs in at a very reasonable 988,633 parameters.

Heavier configurations were considered, but performance gains were minimal (see ablation studies) and the compute tradeoff made it impractical for anyone without multi-GPU cluster access to use the model. This is especially true for increasing the context window or latent hidden dimension.

B.3 XGBoost Configuration

Hyperparameters were found by random search. We used the same train/test split as the EcoPerceiver experiment; the train set was used in a 5-fold cross validation framework with 50 iterations. Once hyperparameters were found, we retrained XGBoost on all training data before running inference on the test set. Table 7 details the parameterization of our final model.

Since XGBoost is a tabular algorithm, we prepared geospatial data in a similar fashion to XBASE [7]; each spectral band represents a single input value to the model. The value is obtained by taking a weighted average of pixels based on Euclidean distance from the center of the image. Missing pixels were removed from this process, and the weights of the remaining pixels were increased to accommodate for this. The code for this procedure is provided with CarbonSense.

Table 7: XGBoost Hyperparameters

Parameter	Value
learning_rate	0.1
alpha	0.1
gamma	0.4
lambda	0.0
max_depth	9
min_child_weight	9
n_estimators	150
subsample	0.7
scale_pos_weight	0.5
colsample_bytree	0.7
colsample_bylevel	0.8

B.4 Reproducibility and Reliability

Both EcoPerceiver and XGBoost were trained with reproducibility in mind. Once optimal hyperparameters were found, we performed 10 experiments with each model in order to obtain a reliable measure of performance (inspired by [224]). Set seeds were provided to all frameworks utilizing RNG, and distributed dataloader workers were also seed-controlled to ensure full reproducibility of our results.

The seeds for our experiments were simple integer values (0, 10, 20, ... , 90) and were provided for the final training runs after hyperparameters had already been chosen.

B.5 Detailed Results

Here we take a closer look at the performance of our models across different IGBP types. Figure 7 and Figure 8 show box plots of our models' test set performance using NSE and RMSE respectively. Note that the y-axis changes between each plot. We do this because the variance in model performance across different seeds was generally small, and charting all box plots on the same axis makes them challenging to interpret.

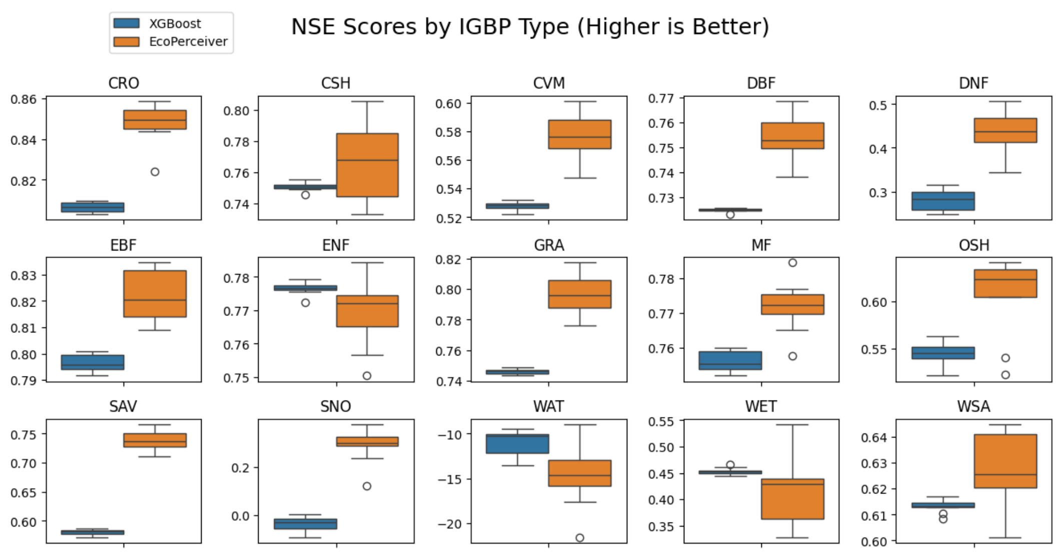

Figure 7: NSE scores of EcoPerceiver and XGBoost across different IGBP types. Each chart represents 10 experiments with different seeds.

As discussed in 5, EcoPerceiver performs better than the XGBoost model in 12 out of 15 IGBP types. In 1 of the 3 that XGBoost wins, both models do substantially worse than simply guessing the mean, which makes the results for WAT challenging to interpret. The other 2, ENF and WET, are not significantly better than EcoPerceiver's performance and have mean NSE advantages of $+0071$ and $+0.0393$ respectively. This could be explained by the nature of data splitting; once hyperparameters were obtained for XGBoost, it was able to train on the entirety of the train split, while EcoPerceiver still had to reserve 20% of the split for validation testing to measure convergence. Both ENF and WET had significant train set prevalence, so these represent IGBP types where XGBoost was most able to take advantage of additional data.

Imperfect hyperparameter selection could also account for the lack of consistent performance. While XGBoost is lightweight enough for a virtually exhaustive parameter search with cross validation, deep models have significantly higher experiment overhead. Due to compute limitations, we were limited in how exhaustively we could explore model configurations for EcoPerceiver.

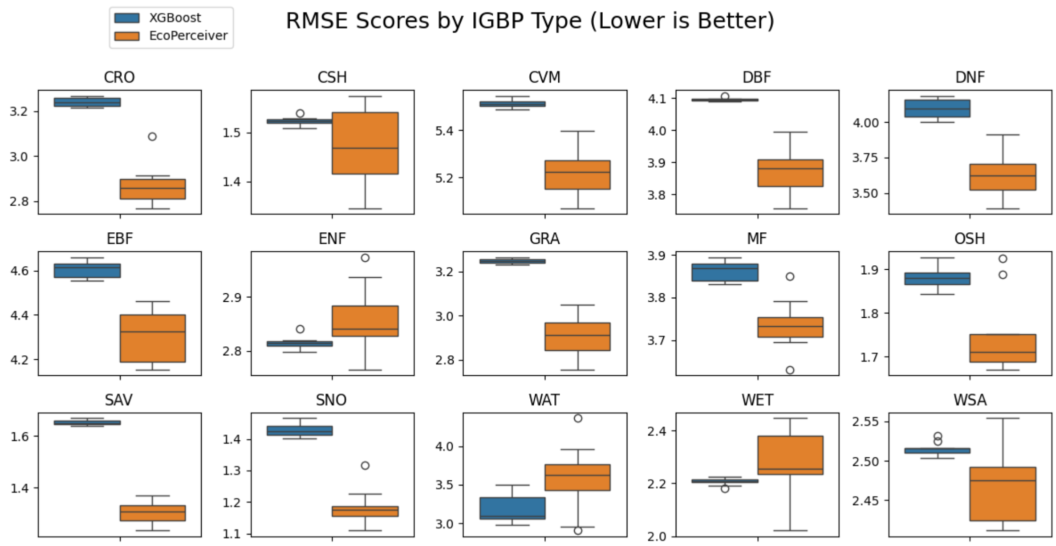

Figure 8: RMSE scores of EcoPerceiver and XGBoost across different IGBP types. Each chart represents 10 experiments with different seeds.

B.6 Ablation Studies

Here we present ablation studies on EcoPerceiver. While we seek to provide a broad view of the impacts of our architectural decisions, note that running all these tests with a large number of seeds was computationally infeasible for us. The boxplots in this section instead represent scores across different IGBP types. We truncated the y axes to positive values due to the outlier IGBP (SNO) making the charts impossible to read.

Running only one seed means the uncertainty of these tests are high, but we felt it would still be beneficial to get an idea of the model's ablated performance.

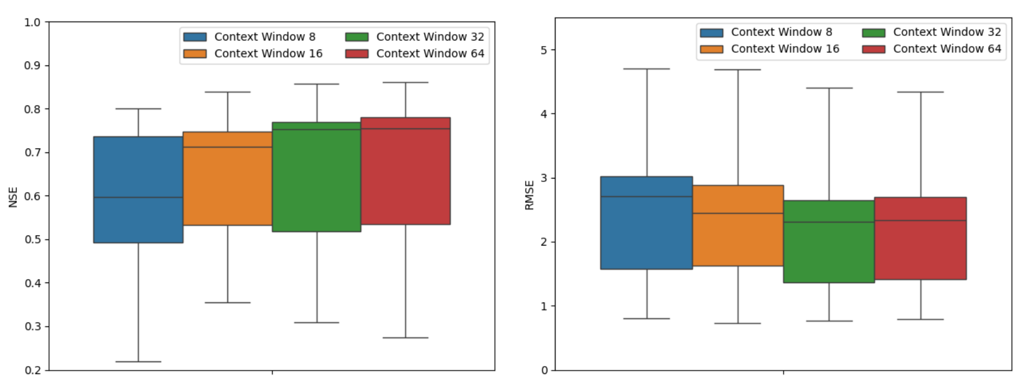

Figure 9: Effects of context window length on NSE (left) and RMSE (right) of the test site for EcoPerceiver.

We first conduct an experiment to test the assumption that DDCFM benefits from ingesting data in a temporal context window. The results are shown in Figure 9. The box plots here represent not different seeds, but the NSE / RMSE values across different IGBP types (it is still important to separate results in this way). There is a clear performance advantage going from a context window of 8 hours to 16, and a small advantage going from 16 to 32. Anecdotally, this reflects our findings from early hyperparameter tuning experiments. Going from 32 to 64 did not meaningfully improve performance, but it did significantly increase our wall time. We therefore used a context window of 32 for our main experiments.

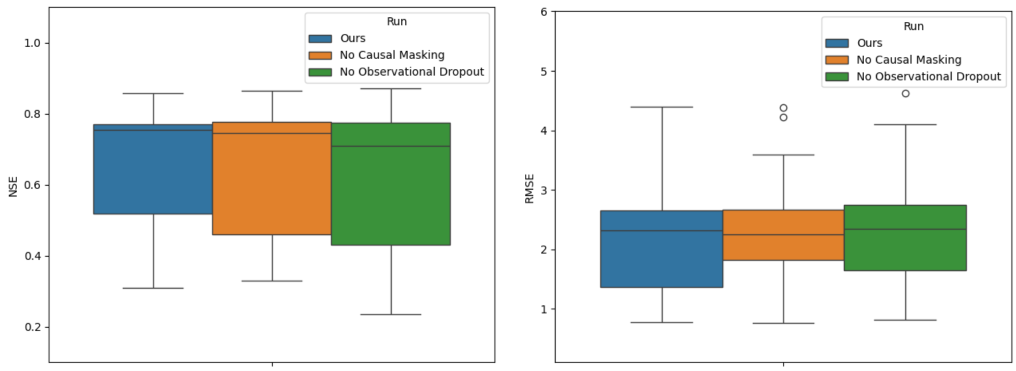

Figure 10: Effects of causal masking and observational dropout on NSE (left) and RMSE (right) of the test site for EcoPerceiver.

Figure 10 shows the effects of both causal masking and observational dropout in EcoPerceiver. We found observational dropout to have a small but notable improvement on test performance. Causal masking very slightly improved NSE but not RMSE, and we feel the uncertainty here makes it impossible to make any statements about its value. Nonetheless, it was kept in the final model since it does not impact wall time, does not appear to make results worse, but does align the model more closely with real-world ecological processes. By ensuring that carbon fluxes depend only on past meteorological conditions and not future ones, causal masking enhances the ecological validity of our model, even if its impact on performance metrics is minimal.

B.7 Ecosystem-Specific DDCFM

DDCFM is not always performed at the global scale; many research teams have studied it in the context of specific regions and ecosystem types [4], [5], [9], [225]. This use case is one of the reasons for CarbonSense's partitioned structure. As a proof of concept, we ran a single experiment with EcoPerceiver where we use the same parameters and train/test split as our main experiment, but only include the DBF sites. We then compared the test set performance against our main model. Table 8 shows the results - our model trained on multiple ecosystem types had notably better performance despite a similar convergence time.

Table 8: EcoPerceiver performance on DBF sites when trained on only DBF data vs trained on all sites

Only DBF		All Sites	
NSE	RMSE	NSE	RMSE
0.7405	3.9782	**0.7532**	**3.8806**

As with the ablation studies, we did not have the compute resources to do 10 seeds for every experiment variation - but it shows the flexibility of CarbonSensefor different research scenarios. It also provides preliminary evidence that DDCFM with multimodal models may benefit from adding more training data even if it is relatively out of distribution.

B.8 Qualitative Analysis

While error metrics are useful for assessing the aggregate performance of the model, we encourage researchers to inspect the model outputs in comparison to the observed data. As an example, consider Figures 11 and 12 below. Both of these were randomly selected 4-day stretches of data from their respective sites. Both models appear able to model GF-Guy very well, but not CA-LP1, and this may be counterintuitive at first glance. But there's quite a bit going on here.

GF-Guy is an evergreen broadleaf forest in the tropics (which is not highly prevalent in CarbonSense). Even so, its carbon fluxes appear quite stable from day to day. We note that ecosystems with this interseasonal stability tend to be more easily modelled in our experiments, though this is not an easy metric to quantify. It should be noted that the y-axis has a much larger scale, so while the models appear close to the ground truth, they often have an error in excess of 5 μmol CO_2 m^{-2}. This again highlights the importance of using NSE as an error metric - RMSE will unfairly punish highly active ecosystems like this due to higher natural variance in carbon fluxes.

CA-LP1 is an evergreen needleleaf forest in the temperate region, which is one of the best represented ecosystems in CarbonSense, yet both models struggle with it (despite low *absolute* error, especially in the winter. Reading into this site reveals it is a pine beetle-attacked forest [96]; disturbances like these can be challenging to model as we discussed in 2. This gives future work in DDCFM a vector for potential improvement.

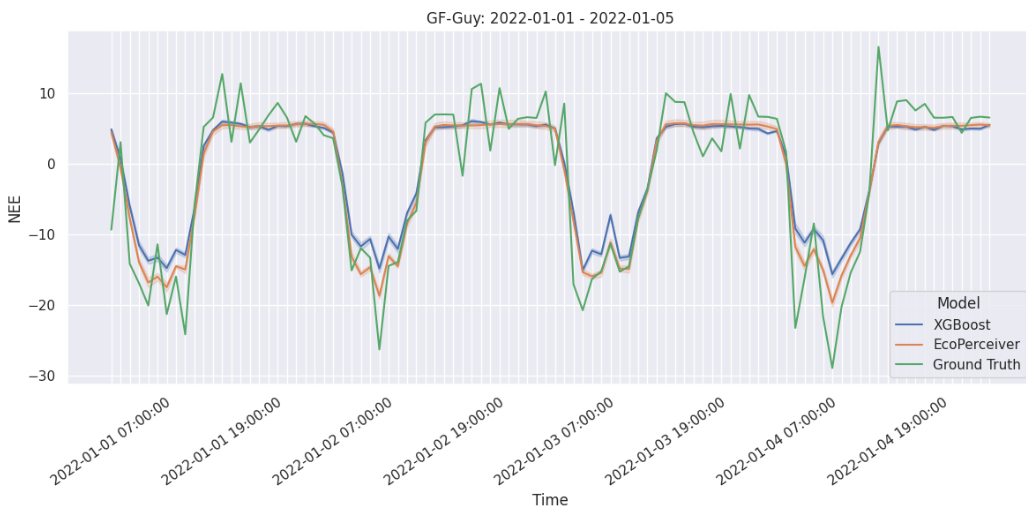

Figure 11: Hourly data and model results for Guyaflux (GF-Guy), an evergreen broadleaf forest station in French Guiana.

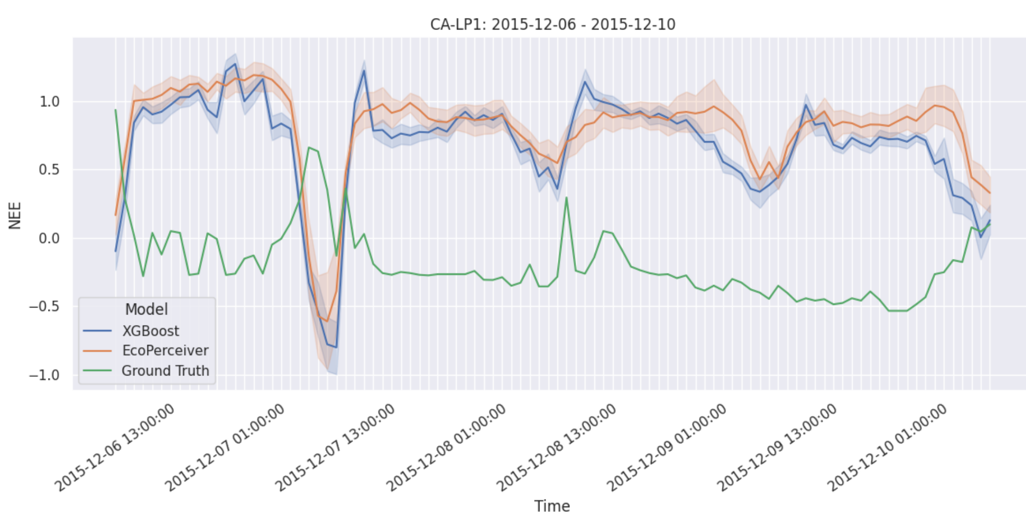

Figure 12: Hourly data and model results for CA-LP1, a pine beetle-attacked evergreen needleleaf forest in northern British Columbia.