Title

Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands

Authors

Irvin, Jeremy
Zhou, Sharon
McNicol, Gavin
et al.

Publication Date

2021-10-01

DOI

10.1016/j.agrformet.2021.108528

Title

Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands

Corresponding Author:

Gavin McNicol

Address: 845 West Taylor Street, SES 2440, MC 186, Chicago, Illinois 60607

Email: gmcnicol@uic.edu

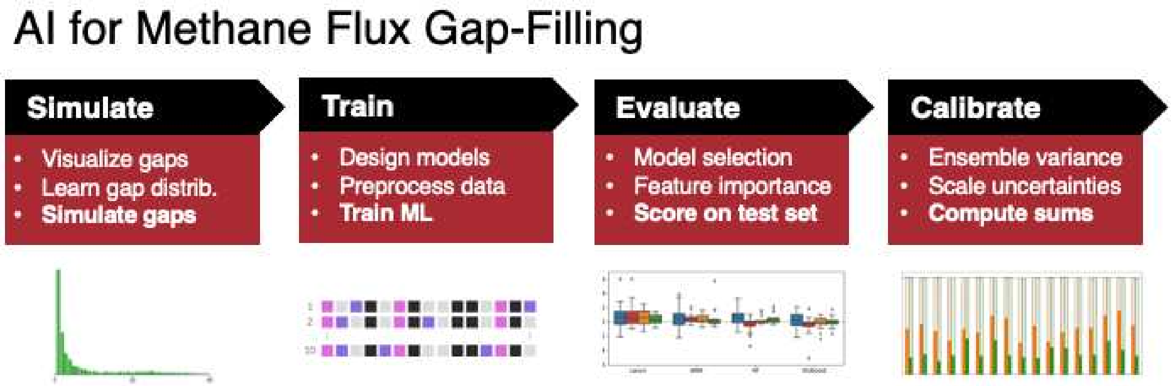
Keywords

Machine learning; timeseries; imputation; gap-filling; methane; flux; wetlands

Highlights

- We evaluate methane flux gap-filling methods across 17 boreal-to-tropical wetlands
- New methods for generating realistic artificial gaps and uncertainties are proposed
- Decision tree algorithms perform slightly better than neural networks on average
- Soil temperature and generic seasonality are the most important predictors
- Open-source code is released for gap-filling steps and uncertainty evaluation

Graphical Abstract



Abstract

Time series of wetland methane fluxes measured by eddy covariance require gap-filling to estimate daily, seasonal, and annual emissions. Gap-filling methane fluxes is challenging because of high variability and complex responses to multiple drivers. To date, there is no widely established gap-filling standard for wetland methane fluxes, with regards both to the best model algorithms and predictors. This study synthesizes results of different gap-filling methods systematically applied at 17 wetland sites spanning boreal to tropical regions and including all major wetland classes and two rice paddies. Procedures are proposed for: 1) creating realistic artificial gap scenarios, 2) training and evaluating gap-filling models without overstating performance, and 3) predicting half-hourly methane fluxes and annual emissions with realistic uncertainty estimates. Performance is compared between a conventional method (marginal distribution sampling) and four machine learning algorithms. The conventional method achieved similar median performance as the machine learning models but was worse than the best machine learning models and relatively insensitive to predictor choices. Of the machine learning models, decision tree algorithms performed the best in cross-validation experiments, even with a baseline predictor set, and artificial neural networks showed comparable performance when using all predictors. Soil temperature was frequently the most important predictor whilst water table depth was important at sites with substantial water table fluctuations, highlighting the value of data on wetland soil conditions. Raw gap-filling uncertainties from the machine learning models were underestimated and we propose a method to calibrate uncertainties to observations. The python code for model development, evaluation, and uncertainty estimation is publicly available. This study outlines a modular and robust machine learning workflow and makes recommendations for, and evaluates an improved baseline of, methane gap-filling models that can be implemented in multi-site syntheses or standardized products from regional and global flux networks (e.g., FLUXNET).

Main Text

1 Introduction

Globally, wetlands emit 102-200 teragrams (Tg) of the greenhouse gas methane (CH_4) to the atmosphere and the scarcity of wetland CH_4 flux data has hindered efforts to better constrain emission uncertainties (Saunois et al. 2020). Eddy covariance-based measurements of CH_4 fluxes have increased rapidly over the last two decades, leading to the release of the first global compilation of CH_4 flux data from 81 sites in 2020 (FLUXNET-CH4 community product Version 1.0; Knox et al. 2019; Delwiche et al. 2021). The growth in available CH_4 data can help improve bottom-up estimates of regional-to-global wetland CH_4 sources (Treat et al. 2018; Peltola et al. 2019; Rosentreter et al. 2021) but this requires data processing standards that ensure eddy covariance CH_4 flux data products are of the same quality and provenance as carbon dioxide (CO_2) and energy fluxes (e.g., FLUXNET2015; Pastorello et al. 2020). Gap-filling is a particularly important step during data processing as it impacts estimates of ecosystem carbon balance and net ecosystem radiative forcing at individual sites, due to the potency of CH_4 as a greenhouse gas (Neubauer and Megonigal 2015; Hemes et al. 2019; Günther et al. 2020), and can alter upscaled predictions in data driven CH_4 flux models (Turetsky et al. 2014; Treat et al. 2018; Peltola et al. 2019). Comprehensive evaluations of gap-filling methods for CH_4 fluxes across many wetland sites are still lacking and needed in order to advance existing methods (Nemitz et al. 2018; Mammarella et al. 2020).

Gaps of various lengths arise in time series of eddy covariance CH_4 fluxes because of system failure (including signal degradation due to sensor soiling), insufficient turbulent mixing, extreme weather conditions, irregular maintenance, and wind direction filtering, among other reasons. Technical challenges remain in precise and accurate measurement of eddy covariance CH_4 fluxes (Morin 2019; Knox et al. 2019) despite recent technological advances in spectra-based gas analyzers (Nemitz et al. 2018). After filtering, annual data coverage can be low for CH_4 (25-40%; Delwiche et al. 2021). Therefore gap-filling procedures are required to construct the continuous time series for quantifying continuous daily, seasonally, and annually integrated CH_4 emission estimates. Gap-filling techniques used to impute half-hourly eddy covariance fluxes at individual sites include look-up tables (Reichstein et al. 2005), machine learning and genetic algorithms (Ooba et al. 2006; Moffat et al. 2007; Kim et al. 2019), multiple imputation (Hui et al. 2004; Vitale et al. 2018), and process models (Oikawa et al. 2017). Any bias tied to a given method propagates to seasonal and annual CH_4 emissions and can therefore impact CH_4 emission estimates at regional to global scales (Falge et al. 2001; Moffat et al. 2007; Peltola et al. 2019; Vitale et al. 2019).

Marginal distribution sampling (MDS) (Reichstein et al. 2005; Moffat et al. 2007; Pastorello et al. 2020) and machine-learning (ML) have become the standard gap-filling methods for CO_2 fluxes in the eddy covariance community (Wutzler et al. 2018), while no similar standard has yet been established for CH_4 fluxes. MDS is a multi-step sampling scheme, akin to a complex decision tree, and uses look-up tables to identify similar predictor conditions within a given time window, which conservatively expands around the gap, only as is necessary. MDS is an efficient gap-filling method that supplements the look-up tables with diurnal cycle interpolation, allowing it to function when there are gaps in predictors. However, MDS performance can be limited by the number of permissible predictors and current predictor choices are optimized for CO_2, not CH_4 fluxes (Falge et al. 2001). Moreover, unlike CO_2 fluxes, CH_4 fluxes at many sites appear to lack a consistent diel cycle and display different diel patterns (Bansal et al. 2018). In contrast, ML is well suited to high-dimensional datasets and can capture nonlinear relationships between predictors and fluxes (Tramontana et al. 2016; Bodesheim et al. 2018) albeit they generally need more time to train and evaluate. A summary of some of the methodological considerations for MDS and four different ML algorithms considered in this study are shown in **Table 1**.

Table 1 An overview of marginal distribution sampling and potential machine learning algorithms for gap-filling of CH_4 flux in wetlands.

Method	**Marginal Distribution Sampling (MDS)**	**Lasso Regression (Lasso)**	**Artificial Neural Network (ANN)**	**Random Forest (RF)**	**XGBoost**
Justification	Simple alternative to ML	Interpretable baseline	Most common current method	Fast and promising for tabular data	Strong in other ML applications with tabular data
Class	Multi-step sampling scheme	Linear regression	Regression	Regression (Decision tree)	Regression (Decision tree)
Algorithm	Multi-step look-up table with backup of diurnal cycle interpolation	Least squares regression with regularization penalty on coefficients to "shrink"	Layers of nodes performing linear transformations with nonlinear transfer functions	Ensemble of decision trees learned independently on randomly bagged data	Similar to random forest but decision trees learn iteratively using gradient boosting

		unimportant coefficients to zero		subsets	
Pre-processing	Predictor choice (combinations of 3)	Imputation	Normalization & imputation	Imputation	None (Imputation optional)
Hyperparameter Tuning	None	Yes (minimal)	Yes	Yes	Yes (few)
Interpretability	Low	High (coefficients)	Low	High (importances)	High (importances)
Uncertainty	Variance of observations	Bootstrap ensembles	Bootstrap ensembles	Bootstrap ensembles	Bootstrap ensembles
References	(Falge et al. 2001; Reichstein et al. 2005)	(Tibshirani 1996)	(Rojas 2013)	(Breiman 2001)	(Chen and Guestrin 2016)

To date, artificial neural networks (ANN) have been found to be effective for gap-filling CH_4 fluxes across six high-latitude wetlands (Dengel et al. 2013). ANN have since been used across a variety of eddy covariance sites at natural, rewetted, and urban wetlands (Morin et al. 2014; Goodrich et al. 2015; Rey-Sanchez et al. 2018; Hemes et al. 2019; Li et al. 2020; Koebsch et al. 2020), tidal salt marshes (Vázquez-Lule and Vargas 2021), and rice paddies (Knox et al. 2016; Runkle et al. 2019), as well as in a FLUXNET-CH_4 synthesis and the FLUXNET-CH4 community product Version 1.0 (Knox et al. 2019; Delwiche et al. 2021). However, the ANN algorithms developed by Dengel et al. (2013) and Moffat et al. (2007) were only inter-compared in detail among six high-latitude sites and were only evaluated on single site-growing-seasons of data. More recently, random forests (RF) were found to match or outperform both MDS and ANN at five wetlands and rice paddies, with strengths in predicting interannual variability from a single multi-year model (Kim et al. 2019). Overall, although some important insights into CH_4 gap-filling strategies with ML have been made at individual, or small sets of sites, comprehensive experiments are still needed to identify the best approaches across the global distribution of wetlands.

In addition to algorithm choice, investigators need to consider the causes of spatial and temporal variability and the effects of biases between training and test data. The complexity of wetland CH_4 production, consumption, and transport processes can lead to high temporal and spatial variability in fluxes across flux tower footprints. Relationships between biophysical drivers and CH_4 flux can be nonlinear and obscured by lags and asynchronicity (Sturtevant et al.

2016). Additionally, the temporal signals in CH_4 flux time series are observed across a broad range of hourly, multi-day, and seasonal timescales (Knox et al. 2019; Knox et al. 2021), and can lack a clear diel cycle as observed for CO_2 (Moffat et al. 2007). Challenges also arise for standardization due to site uniqueness (Bridgham et al. 2013; Trifunovic et al. 2020). For example, Knox et al. (2019) showed that variation in water table depth, a well-established control on wetland CH_4 fluxes, only measurably affected CH_4 flux at sites where its range extended across the soil surface. Similarly, the spatial mosaic of inundation and vegetation varies both within and across wetland classes and affects wetland CH_4 flux via substrate supply and gas transport processes (Matthes et al. 2014; McNicol et al. 2017; Rey-Sanchez et al. 2018). This high spatial heterogeneity creates a wind direction (footprint) dependency rarely observed for CO_2 fluxes (Tuovinen et al. 2019). To be able to explain the complex dynamics of wetland CH_4 emissions, models need information on water table position, soil oxygen and moisture, and soil temperature (Bridgham et al. 2013). Other issues include biases in training observations introduced by low turbulence (friction velocity, USTAR) filters (Göckede et al. 2019) which might make gap-filling models more prone to errors during imputation of CH_4 flux from higher-to-lower turbulence conditions (Dengel et al. 2013), as is observed at some sites for daytime-to-nighttime imputation of CO_2 flux (Moffat et al. 2007). Conditions that lead to exceptional but short-lived fluxes (e.g., ebullition events) may also be less easy to capture in training and test data (Ueyama et al. 2020b; Taoka et al. 2020). In sum, the combination of high temporal variability of CH_4 flux within and across sites (Knox et al. 2019), high spatial variation of fluxes in some wetlands (Morin et al. 2017), and the sensitivity of fluxes to a suite of drivers at different timescales (Sturtevant et al. 2016), requires a thorough evaluation of CH_4 flux gap-filling models across a broad range of possible gap lengths.

This study provides a systematic evaluation of MDS and four ML algorithms for gap-filling CH_4 fluxes at 17 FLUXNET-CH4 sites. The 17 sites cover a wide range of wetland types, and climate and gap conditions (i.e., length and distribution). Collectively, these sites provide a large and fairly standard set of predictors, allowing for a robust across-site comparison of model performance and predictor importance. The overall ML workflow from artificial gap generation, to cross validation and testing, and to prediction uncertainty estimation, is robust and reproducible (Pastorello et al. 2020; Nemitz et al. 2018) and designed to be general and applicable to a wide range of gap-filling scenarios across terrestrial wetland ecosystems. The data and code are made public [https://github.com/stanfordmlgroup/methane-gapfill-ml].

2 Materials and Methods

2.1 Site Data

Seventeen managed agricultural (i.e., rice paddies) and natural wetlands were selected from Version 1 of the FLUXNET-CH4 database (Delwiche et al. 2021) for the comparison of gap-filling methods (**Table 2**). Selection criteria of the sites included: 1) at least one calendar year of measured fluxes; and 2) a complete set of measured physical and biological predictors, including soil temperature and water-table depth (**Table A.1)**. Although FLUXNET-CH4 contains other ecosystem types, including several upland cover types, lakes, and mangroves, these ecosystems were beyond the scope of the present study.

Table 2: Site information and data references for 17 wetland FLUXNET-CH4 sites. Sites are arranged in order of increasing mean of observed CH_4 flux (which is also sensitive to differences in temporal coverage between sites) and days refers to the number of days with some observed CH_4 fluxes. Data are the same as those published in the FLUXNET-CH4 community product Version 1.0 (https://fluxnet.org/data/fluxnet-ch4-community-product/) (Delwiche et al. 2021). Mean annual temperature and precipitation were extracted from respective WorldClim 2.0 gridded products at site locations (Fick and Hijmans 2017).

Site ID	Climate Zone	Mean Annual Temp. °C	Mean Annual Precip. mm	Mean FCH4, nmol m^{-2} s^{-1}	Days, n	Site DOI
US-Uaf	Boreal	-2.8	298	2.7	2922	(Iwata et al. 2020b)
US-Los	Temperate	4.1	833	18.4	1826	(Desai 2020)
SE-Deg	Boreal	1.7	620	31.7	1826	(Nilsson and Peichl 2020)
FI-Sii	Boreal	3.2	666	35.4	2191	(Vesala et al. 2020b)
US-Twt	Temperate	15.2	372	37.7	3016	(Knox et al. 2020)
FI-Si2	Boreal	3.2	664	46.1	1827	(Vesala et al. 2020a)

CA-SCB	Boreal	-2.8	414	46.3	1417	(Sonnentag and Helbig 2020)
NZ-Kop	Temperate	13.9	1343	47.0	1461	(Campbell and Goodrich 2020)
FI-Lom	Boreal	-0.4	484	49.7	1826	(Lohila et al. 2020)
JP-Mse	Temperate	14.1	1305	59.4	366	(Iwata 2020a)
JP-BBY	Temperate	6.7	1153	65.0	1461	(Ueyama et al. 2020a)
BR-Npw	Tropical	25.2	1318	69.7	1122	(Vourlitis et al. 2020)
US-Tw4	Temperate	15.4	370	97.5	2191	(Eichelmann et al. 2020)
US-WPT	Temperate	9.9	881	127.6	1096	(Chen and Chu 2020)
US-Myb	Temperate	15.4	346	142.8	3287	(Matthes et al. 2020)
US-Tw1	Temperate	15.4	371	166.7	2922	(Valach et al. 2020)
US-OWC	Temperate	9.9	898	627.3	669	(Bohrer et al. 2020)

The 17 sites span tropical to boreal climates and diverse and representative wetland types (**Figure 1**), including bogs (5), marshes (5), fens (4), a tropical swamp (1), and rice paddies (2). Altogether, 32.4 site-years of CH_4 flux data were used for gap-filling model development and validation, collected during 2010-2018. Data pre-processing steps prior to gap-filling were the same as described in (Delwiche et al. 2021). Each site was classified into a wetland class based on site investigator self-reporting.

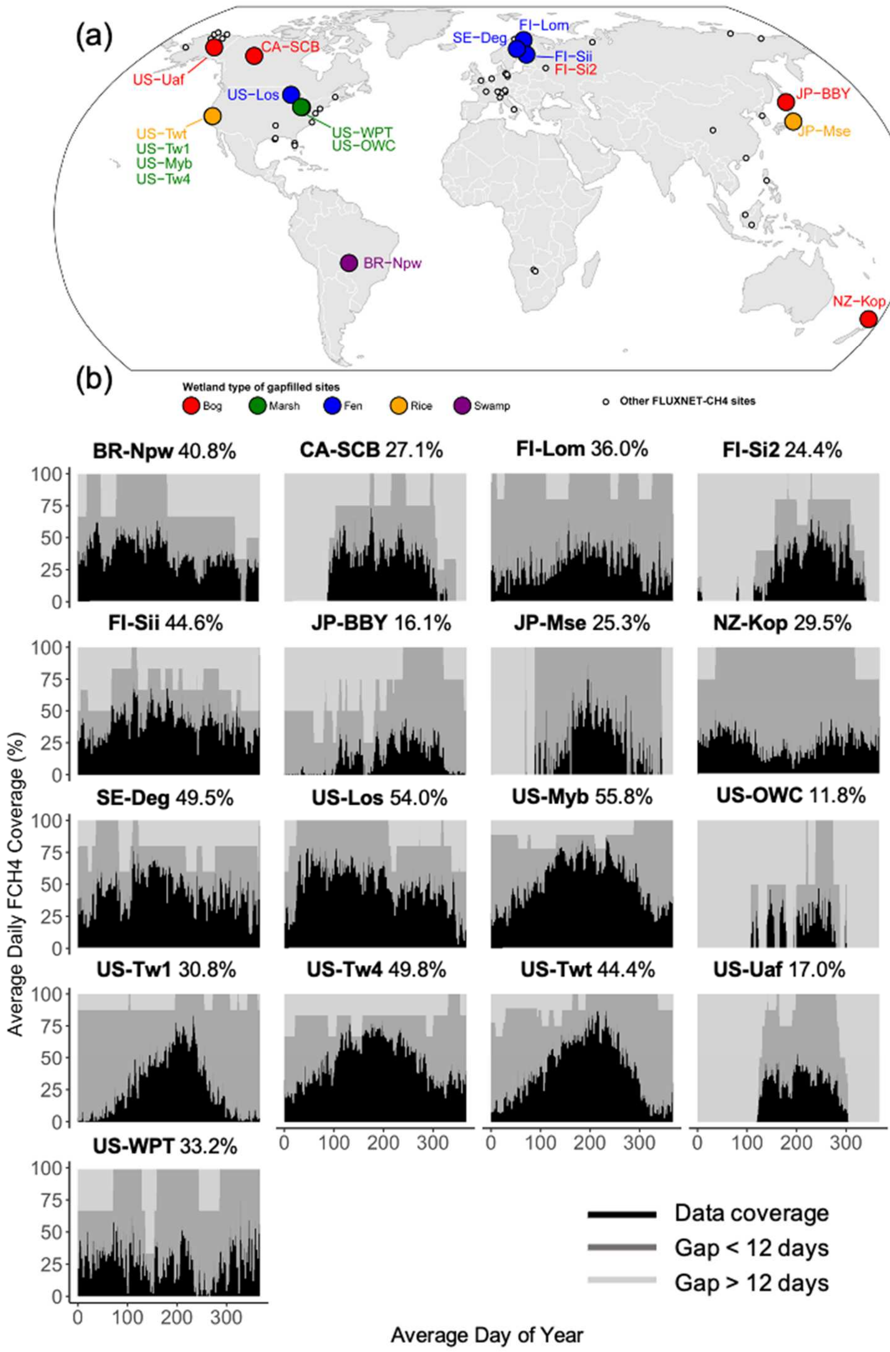


Figure 1: (a) Map of the 17 wetland sites used for the gap-filling experiment and (b) average daily data coverage (%) at each site. The average daily data coverage was computed at each site as the proportion of available to total (48) half-hourly flux periods per day, averaging across available years of data. In addition to spanning a wide geographic and climatic range, the temporal distribution of gaps and their lengths varied

greatly across sites providing a large range of conditions for model testing and evaluation.

2.2 Predictor Variables

For each site, four different combinations of input predictors were tested (**Table 3)**. The simple "temporal set" consisted of two variables that mimic a generic seasonal cycle (sine and cosine functions with yearly wavelengths and amplitude equal to 1) and decimal day of year (delta). The "meteorological set" included four variables (air temperature (TA), incoming shortwave radiation (SW_IN), wind speed (WS), and atmospheric pressure (PA)) measured at eddy covariance towers that were gap-filled using atmospheric reanalysis products (ERA-Interim reanalysis data; Vuichard and Papale 2015). The "baseline set" combined the temporal and meteorological sets, for a total of 7 predictors. These predictors were chosen as the baseline for comparison for their consistent availability as core eddy covariance measurements and were used to gap-fill the FLUXNET-CH4 Version 1.0 dataset (Knox et al. 2019; Delwiche et al. 2021).

Beyond the baseline predictors of Knox et al. (2019), the use of all predictors at each site was also tested, providing a large and comparable predictor set that always included soil temperature, and soil moisture, and/or water table position, among others (**Table 3**). Although availability of these additional predictors varied widely across other FLUXNET-CH4 sites, for these 17 sites, the additional predictors constituting the all-predictor set were highly consistent. Missing predictor data were mean-imputed and "imputed flag" predictors were created, which is standard in ML.

Table 3: Input predictor subsets with variables and their abbreviations used in the text and figures. Further details for predictors are provided in Table A.1.

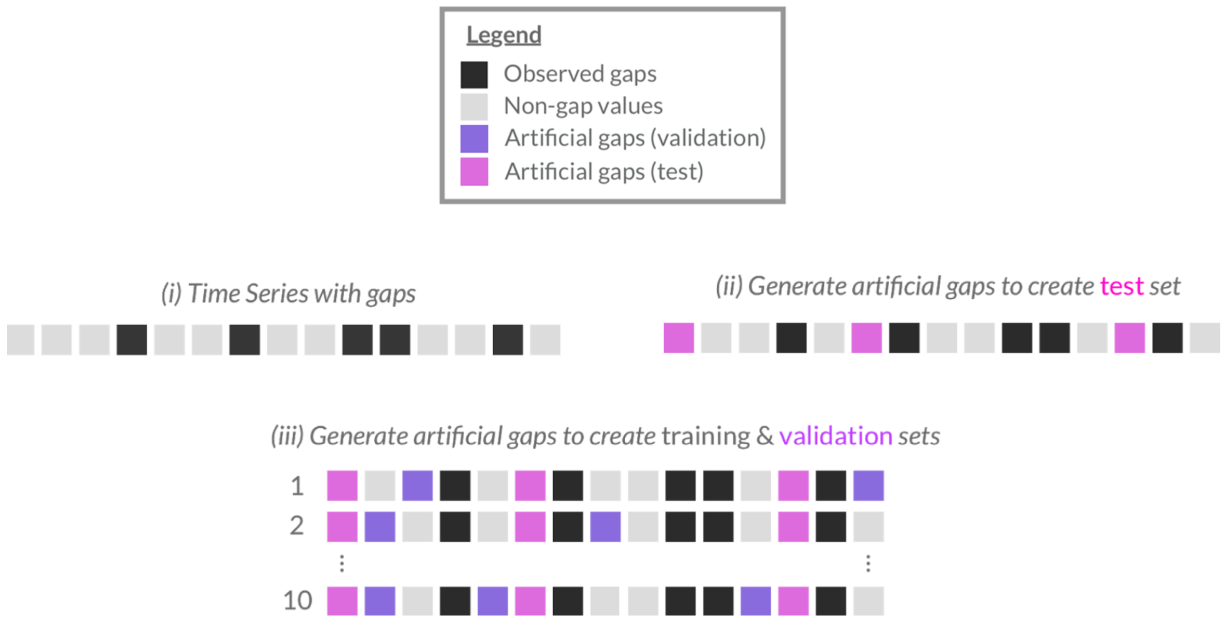
Predictor Subset	Predictor Variables
Temporal	Yearly sine Yearly cosine Delta (decimal day of year)
Meteorological	Air temperature (TA) Incoming shortwave radiation (SW_IN) Wind speed (WS) Atmospheric pressure (PA)

Baseline Applied in (Knox et al. 2019) and FLUXNET-CH4 Version 1.0 (Delwiche et al. 2021)	**Temporal + Meteorological**
All	**Baseline + all other available eddy covariance measurements, including:** ***Soil*** Soil temperature (TS) Water table depth (WTD) Soil water content (SWC) ***Carbon fluxes*** Net ecosystem exchange (NEE) Ecosystem respiration (RECO – day-and-night methods)* Gross primary productivity (GPP – day-and-night methods) ***Energy fluxes*** Latent heat (LE) Sensible heat (H) Soil heat (G) ***Additional meteorology*** Radiation fluxes (SW_OUT, LW_IN/OUT, NETRAD) Friction velocity (USTAR) Vapor pressure deficit (VPD) Precipitation (P) Relative humidity (RH) Snow depth (SD) Photosynthetic photon flux density (PPFD_IN/OUT) Wind direction (WD)

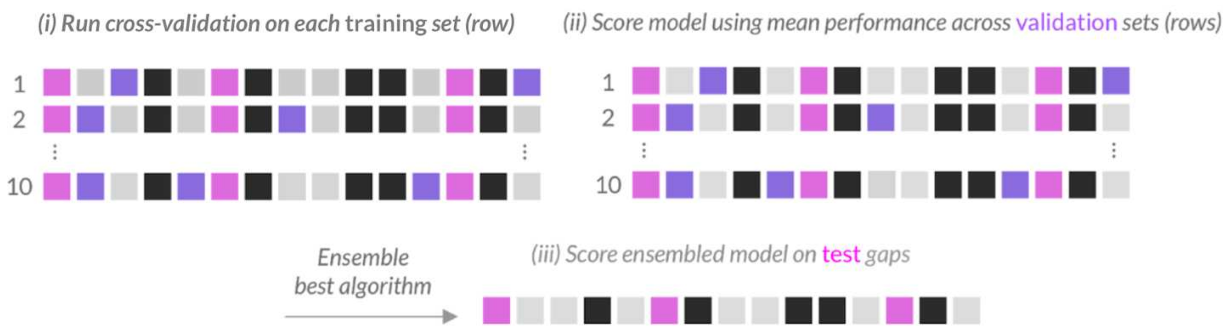
*Both conventional nighttime temperature extrapolation method (Reichstein et al. 2005) and more recent daytime method (Lasslop et al. 2010) variables were included.

2.3 Machine Learning Model Training Procedure

Four ML algorithms were trained with each of the four subsets of input predictors (**Table 3**), leading to a total of 16 algorithm-predictor combinations per site, which were evaluated using a nested cross validation procedure (**Figure 2**). In each algorithm-by-predictor set experiment, the following steps were repeated for each site. Firstly, artificial gaps were introduced which constituted a single, held-out test set. The test set was only used after model training and selection to evaluate the gap-filling performance of the selected models. Secondly, following Moffat et al. (2007), 10 additional pairs of training and validation sets of artificial gaps were created with several independent samples of artificial gaps to mitigate potential bias in model performance for any particular gap sequence. Thirdly, for each algorithm-by-predictor combination, a model was trained on each of the 10 training sets and the best ML hyperparameters were selected based on average model performance during 5-fold cross-validation. Cross-validation involved creating 5 random subsets (folds) of each training set, training the model multiple times with a broad hyperparameter grid search on 4 folds, and evaluating the models on one held-out fold. This hyperparameter search was repeated 5 times, changing the held-out fold each time. The best hyperparameters across all folds were then used to refit the model on the full training set, resulting in 10 trained models for each algorithm-by-predictor combination. Fourthly, each of the 10 models was evaluated using the corresponding validation set, and the mean and variance of model scores for the 10 validation sets were used to compare algorithm classes with different input predictor groups. Finally, the 10 models of the algorithm classes that scored highest on the validation sets were ensembled and the ensemble mean prediction was evaluated against the test set.



(a) Creating data (training, validation, and test) splits with artificial gaps



(b) Model development and validation procedure

Figure 2: Artificial gap generation and evaluation procedure. (a) Artificial gaps are introduced to create the test set, which is set aside, followed by several alternative validation sets. (b) One model is trained on each validation set, including a 5-fold cross validation step to tune hyperparameters. The validation set performance can be compared across the different algorithms. Then, for select algorithms (best on validation set), the 10-model ensemble is run on the test set to fill in gaps and mean predictions are used to obtain a final score while prediction variance is used to parameterized uncertainty distributions. With this procedure, no model tuning or predictor selection is performed on the test set.

2.4 Gap-filling Methods

Marginal Distribution Sampling and four ML algorithms were used for gap-filling, including lasso regression, artificial neural networks, random forests, and gradient boosted decision trees. Each ML algorithm was trained using the four different predictor subsets at each site. The “xgboost” package (Chen and Guestrin 2016) was used to implement the gradient boosted decision tree models and the “scikit-learn” package (Pedregosa et al. 2011) in python (Van Rossum and Drake 2009) was used to implement lasso regression, artificial neural networks, and random forests.

2.4.1 MDS

The Marginal Distribution Sampling method originally proposed by (Reichstein et al. 2005) is based on the construction of a look-up table around each single gap (half hour). The method considers three possible drivers, one identified as the main driver and the other two as additional drivers. For each driver, a threshold value is set to define the similarity conditions. For each gap, the missing value is replaced with the average of the measurements found in the time window around the gap with similar meteorological conditions (i.e., similar value of the drivers). The algorithm first tries to use all three drivers for a window which is kept as short as possible to avoid the confounding effects of other slow-changing drivers such as phenology. If no similar conditions are found, the window size is increased and only the main driver is considered, or alternatively, and as a last option, the mean diurnal cycle within adjacent days is used. More details on the overall strategy and compromise between having a larger window or only one driver included can be found in the appendix of (Reichstein et al. 2005). The original method, designed for CO_2 fluxes, uses SW_IN as the main driver, and TA, and VPD as additional drivers. In the current application of the method to wetland CH_4 fluxes, however, seven different driver combinations were tested as reported in **Table 4**.

Table 4: Driver combinations used for the MDS method. SW_IN = Incoming shortwave radiation (W m^{-2}), TA = air temperature (°C), PA = air pressure (hP), WTD = water table depth (m), WS = wind speed (m s^{-1}), RECO = ecosystem respiration (µmol CO_2 m^{-2} s^{-1}). The values in parenthesis are the thresholds used to define similar conditions (i.e., value ± threshold). In case of SW_IN, as in the original formulation of the method in (Reichstein et al. 2005), the thresholds are two (20, 50): similar conditions for a measured value V are considered in the range V ± 50 if V > 50, V ± 20 if V < 20 and V ± V for values of V between 20 and 50.

Combination	Main driver (threshold)	Secondary driver 1 (threshold)	Secondary driver 2 (threshold)
1	SW_IN (20, 50)	TA (2.5)	PA (0.2)
2	TA (2.5)	SW_IN (20, 50)	PA (0.2)
3	TA (2.5)	SW_IN (20, 50)	RECO (1)
4	TA (2.5)	SW_IN (20, 50)	WTD (0.02)
5	TA (2.5)	SW_IN (20, 50)	TS (1)
6	TA (2.5)	WS (1)	PA (0.2)
7	TA (2.5)	SW_IN (20, 50)	WS (1)

2.4.2 ML Algorithms

Serving as an interpretable and simple baseline model, penalized linear regression was tested for flux gap-filling, referred to here as Least Absolute Shrinkage and Selection Operator (Lasso; Tibshirani 1996). Lasso regression penalizes the sum of the absolute value of coefficients leading to a sparse selection of variables. The regularization coefficient (penalty) was selected during cross validation. Predictors were standardized after imputation by subtracting the mean and dividing by the standard deviation which is necessary for methods that are not scale-invariant such as Lasso which are sensitive to predictor data ranges.

Artificial neural networks (ANN, i.e., shallow multilayer perceptrons) were tested and have been used in previous works for CO_2 and CH_4 fluxes (Goodrich et al. 2015; Dengel et al. 2013; Knox et al. 2016; Hemes et al. 2019; Li et al. 2020). Neural networks consist of a few layers, with each layer containing different numbers of nodes that sequentially apply linear transformations with parameters that are learned during model training. These layers are separated by nonlinear activation functions that enable the neural network to model more complex functions. During training, the parameters of each layer's transformation were adjusted to minimize the squared loss between the predicted and observed flux values. Hyperparameters tuned during cross validation included the optimization method for adjusting parameters (LBFGS or Adam), learning rate (0.01, 0.001, 0.0001), the nonlinear activation function (hyperbolic tangent or rectified linear unit), the numbers of hidden layers (1 or 2; Knox et al. 2019), and the number of nodes per layer (5-30). Normalization was the same as Lasso.

Random forests have been commonly used to model tabular data and have recently emerged for gap-filling CH_4 fluxes (Kim et al. 2019). Random forests are an ensemble of decision trees which are each learned independently on bootstrapped data (Breiman 2001). The mean of the predictions across the ensemble of trees is taken as the final prediction. Hyperparameters tuned during cross validation included the number of trees (50-500), the maximum depth per tree (10-110, as well as no maximum depth), the number of predictors considered at each split (n or square-root of n), the minimum number of samples required to split a node (2, 5, or 10), the minimum number of samples required at each leaf node (1, 2, or 4), and whether to bootstrap the data when building trees. Normalization is not required for RF. Predictor importance was computed as reduction in Gini impurity (Breiman 2001).

Boosting enables decision trees to be grown iteratively based on the mistakes of prior trees (Freund and Schapire 1999). XGBoost was tested as a widely used and efficient gradient boosted decision tree framework that builds decision trees sequentially (Chen and Guestrin 2016) and has demonstrated success in a wide variety of ML applications. A squared loss was used as the objective function with the default learning rate of 0.1. The number of decision trees, the maximum depth per tree, and the minimum number of samples required to split a node used the same ranges as RF. Other hyperparameters tuned included the proportion of the training data to subsample prior to growing trees (0.75, 0.85, or 0.95), the minimum loss reduction required to split a leaf node (0, 0.2, or 0.4), and the fraction of predictors that were randomly selected for the construction of each tree (0.6, 0.7, 0.8, or 0.9). XGBoost handles predictor imputation during training using sparsity-aware split finding, which provides a default direction on each node in the decision tree and allows for skipping over missing values (Chen and Guestrin 2016). Normalization is not required for XGBoost.

2.5 Artificial Gap Generation

Different gap lengths occur naturally in the time series of eddy covariance flux measurements, for reasons that include instrument malfunction, power outages, seasonal changes (winter), and data QA/QC (Moffat et al. 2007). Introducing artificial gaps into the flux data, across this range of observed gap lengths is necessary to provide scorable validation and test cases. Previous studies have achieved this by evaluating models on different artificial gap-length scenarios. In each scenario, gaps of a limited range of lengths (e.g., 1-8 half-hours) are introduced and model performance is compared among the different gap-length scenarios (Moffat et al. 2007; Kim et al. 2019). This approach ensures gaps of all lengths are evaluated because it relies on sampling gaps randomly or uniformly within fixed gap length scenarios. However, the resulting gap distributions also become skewed when longer gaps form due to artificial gaps merging with observed gaps. This may incorrectly favor models that perform better on longer gaps which are less common in eddy covariance flux data.

To retain the observed gap length distribution, a new artificial gap generation procedure was developed. The new procedure takes into account the locations of the observed gaps when generating artificial gaps of varying lengths, such that the observed plus artificial gap length distribution resembles the observed distribution. Formally, the artificial gap generation procedure finds a distribution q of artificial gap lengths for each site such that the true empirical distribution p of gap lengths at that site is approximated by the union of q and p, which is denoted $r = q \cup p$. In order to obtain a distribution r which is close to p, a method is proposed for finding q. Intuitively, the histogram of q should look "compressed" compared to the histogram of p; that is, it places more weight on shorter gap lengths and has lighter tails: while shorter gap lengths will be sampled more from q, longer gaps will still form from the merging that occurs between newly sampled and observed existing gaps. A detailed description and parameterization of the artificial gap generation algorithm are provided in **Appendix B**.

The proposed method thus maintains a similar distribution of gap lengths to the observed distribution, aiming to strike a balance between having enough scorable (artificial) gaps for model training and ensuring the distribution of gaps input to the model is similar to that of the observed data. As this method does not use prescribed gap scenarios, it is important to inspect the resulting artificial gap distributions. For this study, site-specific gap sampling details and gap length distributions are provided in **Appendix C**.

2.6 Evaluation

For each site, MDS-and the ML algorithm-predictor combinations were compared by evaluating predictive performance on the 10 validation sets. The best two algorithms and their ensemble performance were then evaluated on the test set using both baseline and all predictors to: 1) measure absolute improvements over previously implemented standards (ANN plus baseline predictors; Knox et al. 2019); 2) understand how each algorithm benefited (if at all) from using all, rather than only baseline, predictors; and 3) measure the effect that the different algorithm predictions had on cumulative annual and growing season CH_4 emissions estimates for each site, and associated uncertainties.

2.6.1 Performance Measures

Model performance was measured using the coefficient of determination (R^2), mean absolute error normalized by the standard deviation of CH_4 flux (nMAE), mean bias (Bias), root mean squared error (RMSE), and standard deviation. R^2 was used to measure the ability of the gap-filling model to reproduce the time series pattern, after confirming that Pearson correlations were all positive (Taylor 1990). nMAE was used to measure the difference between predictions from observations regardless of the direction of the error; the normalization allows us to compare across sites despite large differences in flux variability. Finally, Bias was used to measure the average direction of error, which will have the largest consequence on site emission sums. The nonparametric basic bootstrap with 5,000 bootstrap replicates was used to compute variability around the performance metrics on the test set (Efron and Tibshirani 1994); and 95% confidence intervals for each measure were reported. Taylor diagrams were used to visually compare the performance of each of the models with different input predictors. Taylor diagrams provide a visually intuitive way of displaying the performance of each model in terms of three metrics: R^2, root mean squared error (RMSE), and standard deviation (Taylor, 2001). Finally, nMAE and Bias were used to assess the performance of the models across different gap lengths similar to Moffat et al. (2007), Nemitz et al. (2018), Kim et al. (2019), and Knox et al. (2019): very short gaps (1 half hour), short gaps (2-8 half hours), medium gaps (9-64 half hours, i.e., 1.5 days), long gaps (1.5-12 consecutive days), and extremely long gaps (> 12 consecutive days).

2.6.2 Statistical Analysis

Validation set performance was evaluated coarsely using differences in median model metrics and was only used to select models for the more detailed statistical comparison on the test set. Then, for each site, the test set performance of the best two algorithms was compared (RF, as the faster of the two decision tree algorithms, and ANN) with two predictor sets (baseline and all). The performance metrics showed significant non-normality across the 17 sites according to the Shapiro-Wilk test. As a result, the Friedman test followed by post hoc Nemenyi was used for evaluating pairwise comparisons. This pair of tests is the nonparametric equivalent of the one-way ANOVA with repeated measures (followed by Tukey's test) and is the standard procedure when the assumptions of ANOVA are not met (normality in this case; Derrac et al. 2011; Schuurmans 2006). Performance metric comparisons were implemented in R (R Core Team 2019) using the PMCMR package (Pohlert 2014).

To evaluate whether the gap-filling performance is related to the characteristics of CH_4 flux, Pearson correlation coefficient between the best model performance metrics (RF and all predictors) and the annual mean and variance of the fluxes were analyzed. Correlation analyses were performed in Python using the 'scipy' package (Virtanen et al. 2020).

2.6.3 Evaluating Systematic USTAR Bias

Filtering to remove eddy covariance CH_4 fluxes during low turbulence conditions (using friction velocity, USTAR, as a measure of turbulence) may introduce a systematic bias into ML training because the efficiency of CH_4 gas transport mechanisms such as plant mediated flow can increase with wind speed (Laanbroek 2010). To approximate an evaluation of biases introduced from low USTAR filtering, the amount of filtered data across each site was quantified (0-21%) and the same fraction of high USTAR conditions (top percentile) was removed from each paired training and validation set. The original and high-USTAR-filtered model performance was then evaluated on the scorable gaps created with the high USTAR filter. Although an imperfect analogue, this test therefore simulated model extrapolation to very low USTAR conditions by evaluating performance during extrapolations to high USTAR conditions.

2.7 Uncertainty Estimation

2.7.1 Uncertainty Evaluation

Machine learning model (gap-filling) uncertainty for each half-hour flux prediction was estimated using the variation of the model ensemble predictions. For each input, the mean and variance of the ensemble predictions were used to parameterize a double exponential distribution (a *probabilistic* prediction) (Hollinger and Richardson 2005). The confidence intervals of the specified confidence level are computed using this full distribution. Similar to Richardson and Hollinger (2007), Lasslop et al. (2008), Richardson et al. (2012), Menzer et al. (2013), Vitale et al. (2019), the model ensemble uncertainty was used to approximate random flux uncertainty. It is acknowledged, however, that because the contribution of missing values in input predictors is not taken into account, the derived uncertainties only approximate the total random uncertainties that can be better accounted for with alternative multiple imputation methods (Vitale et al. 2018). The described method focuses on providing a method to robustly evaluate gap-filling uncertainties in a manner suitable for ML ensemble workflows.

The consistency of the uncertainty estimates was evaluated using standard probabilistic forecasting evaluation measures, namely calibration and sharpness (Gneiting et al. 2007). Calibration captures the consistency between probabilistic forecasts and observations, and measures whether predicted distributions correctly capture confidence levels as validated against observed data. A well-calibrated model produces predictive distributions such that P% confidence interval (CI) contains the observations P% of the time. A model can be well calibrated only at specific percentiles (e.g., 95%) or across multiple percentiles. At a minimum, models should be well calibrated at the specific desired percentile before uncertainty estimates at that percentile can be reliably used. Once models are shown to be well calibrated, they can be compared using sharpness - a property that measures the concentration of the predictive distributions. The approach of maximizing sharpness subject to calibration is widely adopted in meteorology (Gneiting and Katzfuss 2014). Model improvement is captured by increasing sharpness, subject to calibration. For each site, performance was evaluated at the 95% CI. Calibration was measured by computing the proportion of the observed values within the 95% CIs and measured sharpness using the mean width of the 95% CIs across the test set. A normalized sharpness metric is reported by dividing by the standard deviation of flux to account for the differing flux variance at each site.

2.7.2 Uncertainty Interval Scaling

Models that produce predictive distributions, such as the ML ensemble in the present study, are not necessarily well calibrated by default. Several techniques have been proposed to calibrate models after they are trained (post-processing calibration), most often using Platt scaling (Platt 1999) and isotonic regression (Zadrozny and Elkan 2002). In this work, Platt scaling is adopted to calibrate the ensemble predictions. Platt scaling learns a scaling parameter that is used to scale the variance uniformly for every input. This parameter is learned by assuming a distribution (e.g., double exponential) and using maximum likelihood estimation to derive a value from observed data. A double exponential distribution was assumed and derived a closed-form expression for the scaling parameter (see **Appendix D** for derivation). Following this calibration procedure, the probabilistic predictions of different models were compared by measuring the sharpness of the calibrated distributions.

2.8 Annual and Growing Season Emissions

Annual CH_4 emissions were computed as the mean cumulative sum of the 10 gap-filled flux time series, predicted by each ML model ensemble. To account for the uncertainty calibration procedure, ensemble predictions were rescaled (spread out) around the mean in proportion to the Platt scaling value. Annual sums and uncertainties (uncalibrated and calibrated) were quantified from the mean and variance of the cumulative sums, respectively. As is standard for CO_2 gap-filling, site-years with a gap of 60 days or longer during the growing or shoulder seasons were excluded (Richardson and Hollinger 2007; Richardson et al. 2012), except for US-Uaf, which only had one site-year available, and for US-OWC, which had large shoulder or growing season gaps during both available years. Additional date thresholds were applied for the two rice paddies (US-Twt and JP-Mse) to only sum fluxes during the rice growing season based on rice management information (Knox et al. 2016; Miyata et al. 2000). All other gap-filled values for gap lengths < 60 days were included. Annual or growing season CH_4 emissions estimates were also computed for each of the seven MDS models (different predictor sets) as the cumulative sum of the gap-filled time series. Similar to ML, summed uncertainties were taken as the variance of the sums from the seven MDS models, however no calibration method was applied.

3 Results

3.1 Scorable Gap Conditions

In addition to their wide geographical distribution (**Figure 1a**), the 17 wetland sites also covered a wide range of biophysical conditions. Across all sites, water table depth (WTD) ranged from < -1 m to > 1 m relative to the soil surface, while gross primary production (GPP) ranged from zero in winter to > 40 µmol m^{-2} s^{-1} (**Figure 3a**). Unlike GPP, within site variation in WTD was small relative to across site variation, with the WTD range at some sites being either above (e.g., US-Myb) or below (e.g., US-Uaf) the soil surface. Rice paddies and one tropical swamp (i.e., JP-Mse, US-Tw1, US-Twt, and BR-Npw) showed larger fluctuations that crossed the soil surface (± 50 cm or more). In addition, soil temperature (TS) spanned from -10 °C to > 40 °C across sites, and CH_4 fluxes ranged across 5 orders of magnitude from < 0.01 to > 1,000 nmol m^{-2} s^{-1} (**Figure 3c**). Sites tended to overlap more in their range of TS and CH_4 flux (FCH4), but more distinctive in WTD and GPP. The biophysical conditions for scorable test conditions introduced as artificial gaps in the test set (**Figure 3b, d**) displayed a similar range, indicating that models were evaluated on the full range of observed data conditions.

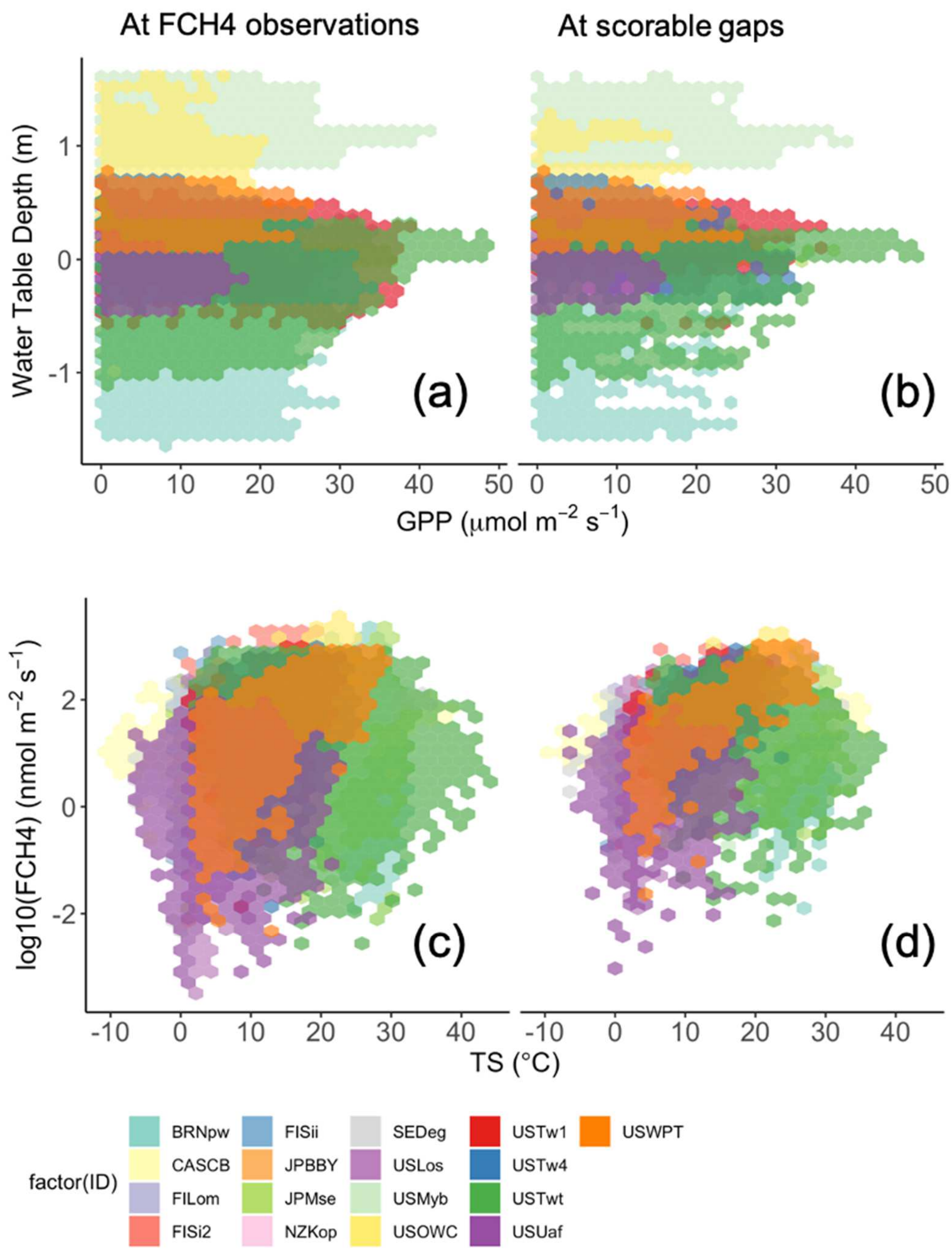


Figure 3: The coverage of training and test data for select predictor and CH_4 flux conditions. All observations (a, c), and scorable gaps (b, d) spanned a wide range of (a, b) water table depth and gross primary production (GPP), and (c, d) CH_4 flux (FCH4) and soil temperature (TS).

3.2 Performance Patterns on the Validation Set

Median MDS performance (R^2 = 0.65; nMAE = 0.35; Bias = -0.03 nmol m^{-2} s^{-1}) was better than median ML performance (R^2 = 0.56; nMAE = 0.39; Bias = 0.01 nmol m^{-2} s^{-1}). However, predictor subsets had little effect on MDS performance (**Figure 4a, c, e**). Only slight improvements were seen over baseline meteorological predictors (i.e., SW_IN, TA, and PA) when one of the CH_4-centric predictors (i.e., WTD, TS, RECO, or WS) was included. Overall, the best performing predictor combination for MDS was TA, PA, and WS (R^2 = 0.66; nMAE = 0.34; Bias = -0.07 nmol m^{-2} s^{-1}), which was used subsequently to compute annual and growing season sums and uncertainties.

There was a larger spread in performance across the ML (**Figure 4b, d, f**). Median performance increased from Lasso (R^2 = 0.37; nMAE = 0.51; Bias = 0.10 nmol m^{-2} s^{-1}), to ANN (R^2 = 0.58; nMAE = 0.39; Bias = 0.06 nmol m^{-2} s^{-1}), to XGBoost (R^2 = 0.65; nMAE = 0.35; Bias = -0.11 nmol m^{-2} s^{-1}) and RF (R^2 = 0.67; nMAE = 0.32; Bias = 0.01 nmol m^{-2} s^{-1}). Unlike MDS, ML performance was strongly dependent on the predictor set. Using all predictors was consistently the best choice across all sites and all classes of models, while using the meteorological subset alone performed the worst. Median model performance ranged from R^2 of 0.27, nMAE of 0.60, and mean Bias of 0.08 nmol m^{-2} s^{-1} for Lasso model class with the meteorological predictors only, to R^2 of 0.79, nMAE of 0.26, and Bias of 0.12 nmol m^{-2} s^{-1} for the RF model class with all predictors. Notably, decision tree models using the baseline predictor set (e.g., RF R^2 = 0.75; nMAE = 0.29; Bias = 0.02 nmol m^{-2} s^{-1}) still outperformed ANN using all predictors (R^2 = 0.70; nMAE = 0.31; Bias = 0.05 nmol m^{-2} s^{-1}). For both decision tree and ANN models, the temporal set was much more important for baseline performance than the meteorological set. As the temporal set can be created for any CH_4 gap-filling effort, the meteorological set is unlikely to be used alone in practice and is therefore only distinguished here to understand its relative contribution to the baseline set.

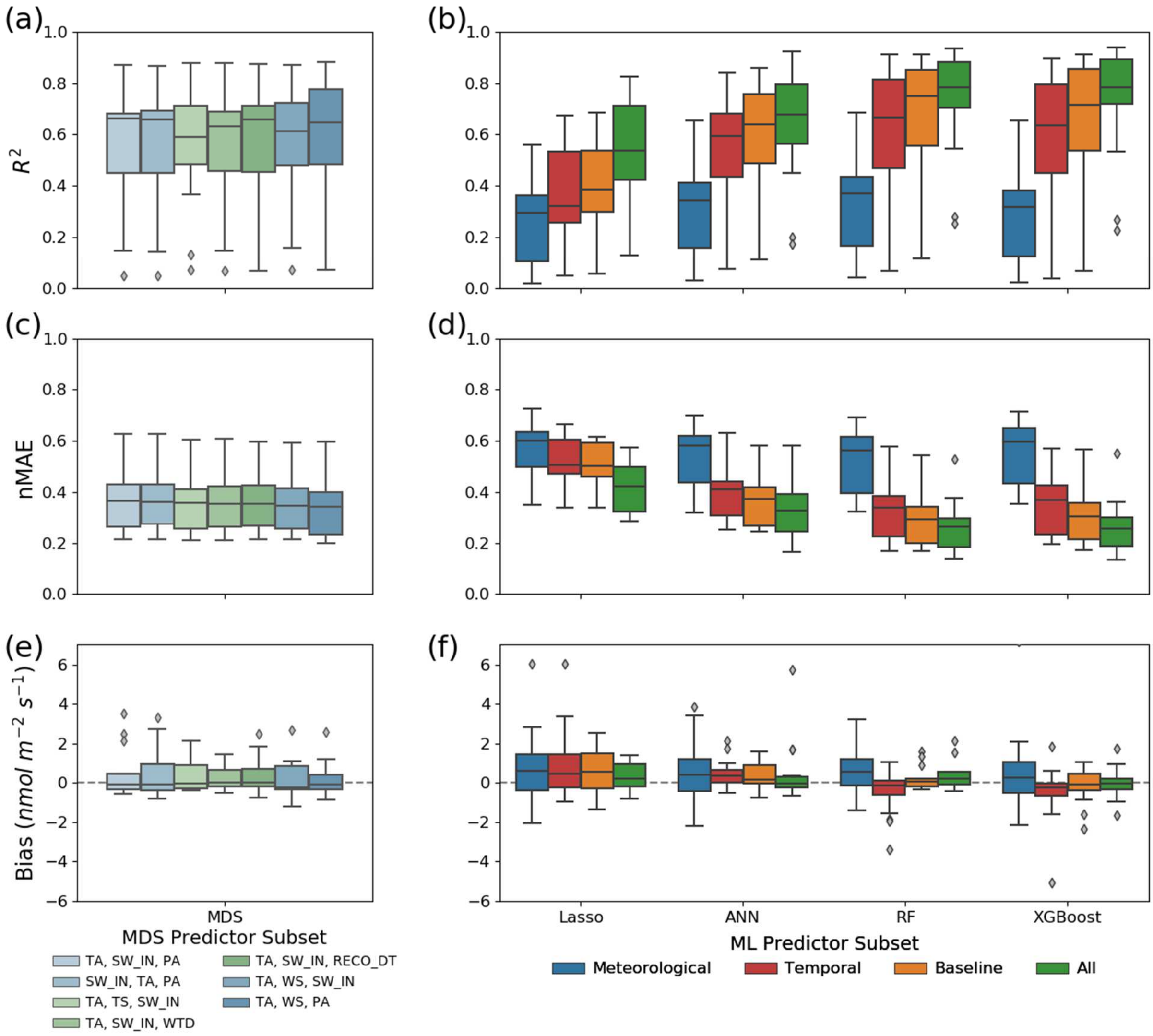


Figure 4: Boxplots illustrating 10 validation set performance metrics for each of the models (Lasso regression (Lasso), artificial neural networks (ANN), random forests (RF), and gradient boosted decision trees (XGBoost)) and predictor subsets across the 17 sites: (a, b) R^2, (c, d) normalized mean absolute error (nMAE), (e, f) bias, where the left column is Marginal Distribution Sampling and the right column is machine learning. Each colored box shows the quartiles of the performance metrics and the whiskers show the rest of the distribution, excluding points determined to be outliers that are presented individually.

3.3 Test Set Performance Patterns

The ANN and RF (as the faster of the two decision tree algorithms) achieved the best performance on the validation set and were then evaluated on the test set for each site. Test set performance patterns were similar to the validation set, confirming that the models were not over-fit. Median performance on the test set was better overall for RF (R^2 = 0.79; nMAE = 0.27; Bias = 0.24 nmol m^{-2} s^{-1}) than ANN (R^2 = 0.73; nMAE = 0.30; Bias = 0.18 nmol m^{-2} s^{-1}). Median nMAE and R^2 both improved when ANN used all rather than baseline predictors (p = 0.0007 and p = 0.0004, respectively). Similarly, median nMAE and R^2 both improved when RF used all rather than baseline predictors (p = 0.0031 and p = 0.0050, respectively). Test set evaluation also provided some evidence of RF outperforming ANN in general. Using all predictors, median nMAE for the RF was smaller than that of the ANN (p = 1.40e-8) although there was no significant difference between the median R^2 of RF and ANN (p = 0.191). Similarly, on baseline predictors, median nMAE for the RF was smaller than that of the ANN (p = 0.0078) but there was no significant difference between the median R^2 of RF and ANN (p = 0.056).

A large spread in performance was observed within most wetland classes, suggesting a high level of site uniqueness, rather than generalizability, within a particular wetland class (**Figure 5**). The large spread was especially apparent for bogs and fens, whereas marshes and the two rice paddies were clustered at intermediate to high performance. To better understand the patterns of performance within and among wetland classes, correlations were examined between best model performance metrics and the annual mean and variance of the fluxes. There was no significant relationship between model performance and the annual mean of site CH_4 fluxes, however, there was a clear negative relationship between performance and the coefficient of variation of CH_4 fluxes (p = 0.001; ρ = 0.72) and an even stronger negative correlation with the flux variance at short (hourly) timescales (p = 1.44e-6; ρ = 0.89) (**Figure E.1)**.

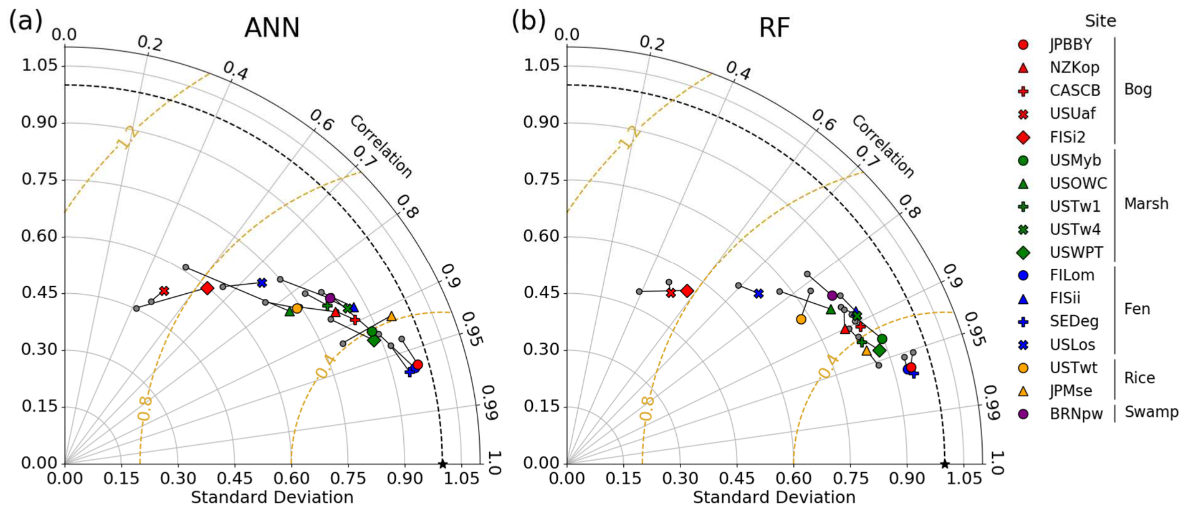


Figure 5: Taylor diagram visualizing artificial neural network (ANN) and random forest (RF) performance improvements on the test set between the baseline and all predictor sets for each of the 17 primary sites. The baseline set metrics for each algorithm are shown in small grey circle symbols and the all predictor set metrics are shown in larger color-filled symbols. Model improvements can be measured in the Taylor diagram in proportion to 2D shifts towards the black star at *(1, 0)*. Taylor diagrams display the ratio of the standard deviation of predictions to observations on the x and y axes, the correlation of predictions to the observed temporal pattern on the curved right axis, and the root mean square error of predictions on the diagram surface as concentric (orange) circles around the origin.

ANN performance showed larger improvements when all predictors were used rather than only baseline predictors (**Figure 6**) and RF performance showed small or negligible improvements. However, absolute RF performance was already relatively high using only the baseline predictors. Overall, the largest ANN and RF performance improvements were observed in marshes, with exceptionally large gains at one site (US-OWC). Several other bog, rice paddy and swamp sites achieved moderate improvements from the additional predictors (i.e., 0.1 to 0.2 increase in R^2), whereas only small improvements were observed at fens, with less than a 0.05 increase in R^2.

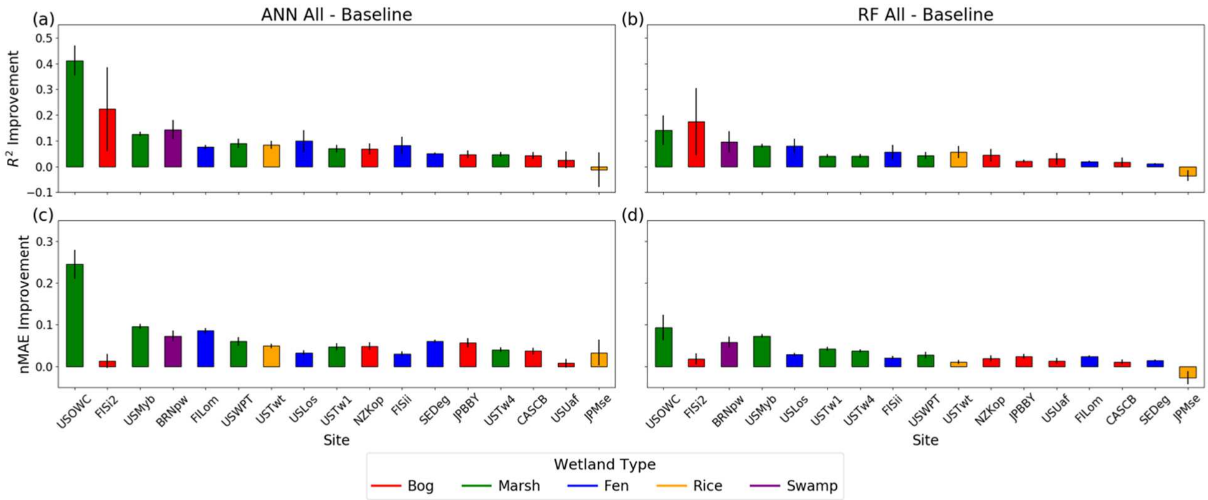


Figure 6: Improvements in test set performance metrics for the artificial neural network (ANN) and random forest (RF) algorithms between the baseline and all predictor sets on the 17 wetland sites. Vertical error bars show the 95% confidence interval around the improvement, computed using the nonparametric basic bootstrap with 5,000 replicates. Sites are plotted in order of the total of R^2 and nMAE improvement.

Across all very short (1 half-hour), short (2-8 half-hours), medium (9-64 half-hours), and long (65-576 half-hours) gap lengths, bias was low for both the ANN and RF models. Errors (nMAE) and biases were typically smaller for RF than ANN, and biases were generally larger at marshes and the swamp (**Figure 7**). For the longest gaps (577+ half-hours), RF and ANN performance was less consistent and the largest biases were introduced at marsh sites when using RF.

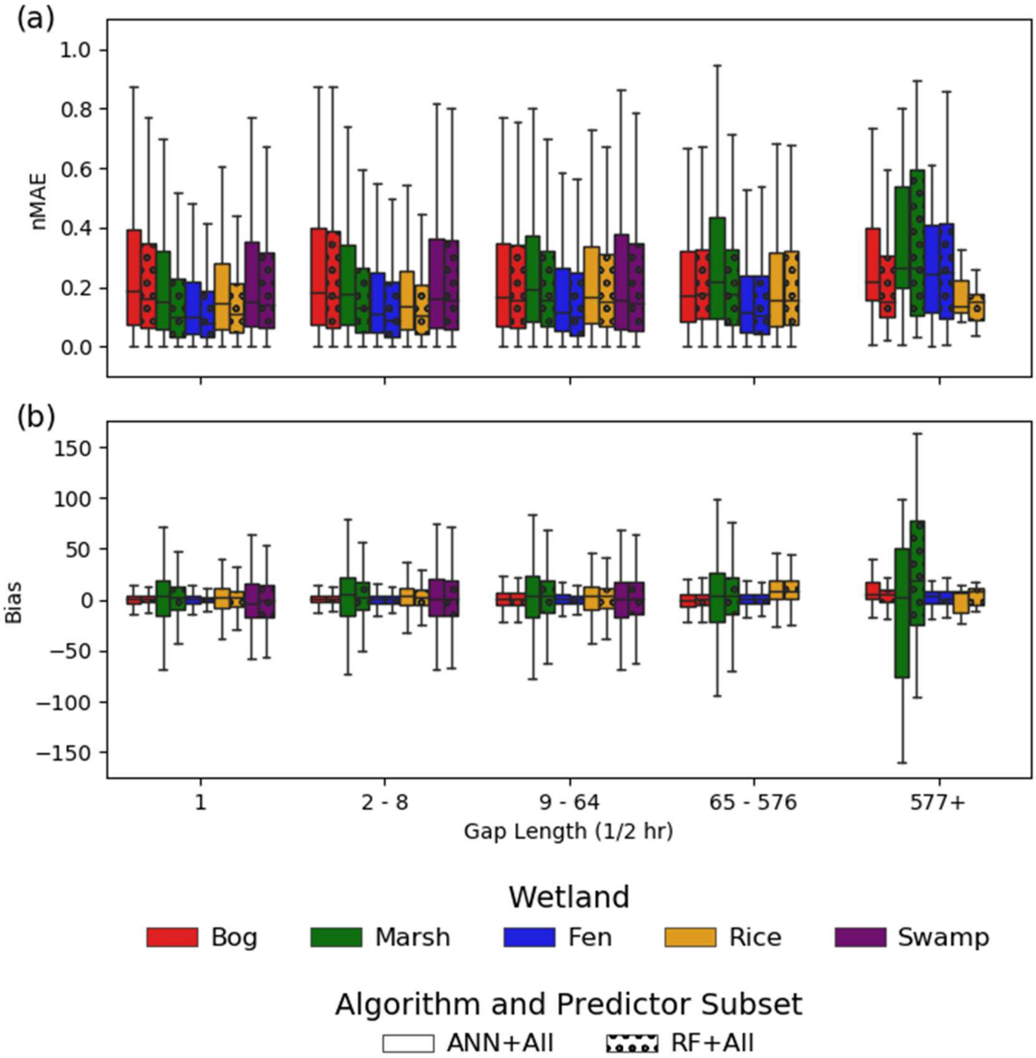


Figure 7: Performance of the two best algorithms (ANN+All and RF+All) on the test sets, broken down by gap length for the 17 primary sites. Swamp values on long gaps (> 65 half-hours) are not shown here as the R^2 is not well-defined on single samples. Gap length values indicate merged gap lengths after test gap generation.

Finally, an exploratory evaluation of errors that may be introduced due to USTAR filtering was conducted. The test set was used with the best model formulations (RF and all predictors). Model performance showed a slight reduction in performance when extrapolating to high USTAR conditions (**Table E.2),** suggesting that similar extrapolations to low USTAR conditions may introduce small but non-negligible errors. Average Bias across all 17 sites increased by 9%, average nMAE by 10%, and R^2 decreased by 8%.

3.4 Predictor Importance

Variable importance rankings are readily retrievable from RF models. The most important predictors of the RF model (in order) across all 17 sites were temporal, TS, radiation (aggregate of SW_IN, SW_OUT, LW_IN, LW_OUT, and NETRAD), and RECO (**Figure 8**), with TS being the single most important predictor for many sites. Air temperature (TA) and turbulence (WS and USTAR), GPP and NEE, and WTD were useful for some sites, but not universally. Wind direction (WD) was important at 2 sites (US-OWC and US-Myb). Generally, there were few strong patterns within bogs, fens and marshes (which were the only classes with at least 4 representative sites), suggesting that predictor groups are not necessarily tied to wetland classification, although TS was important at all of the bogs. Notably, the baseline set captured several of the key predictors and all of the important meteorological predictors, except wind direction. Of the two partitioning methods for RECO and GPP (nighttime and daytime), the nighttime method ranked higher at 15 and 13 (of 17 total) sites, respectively.

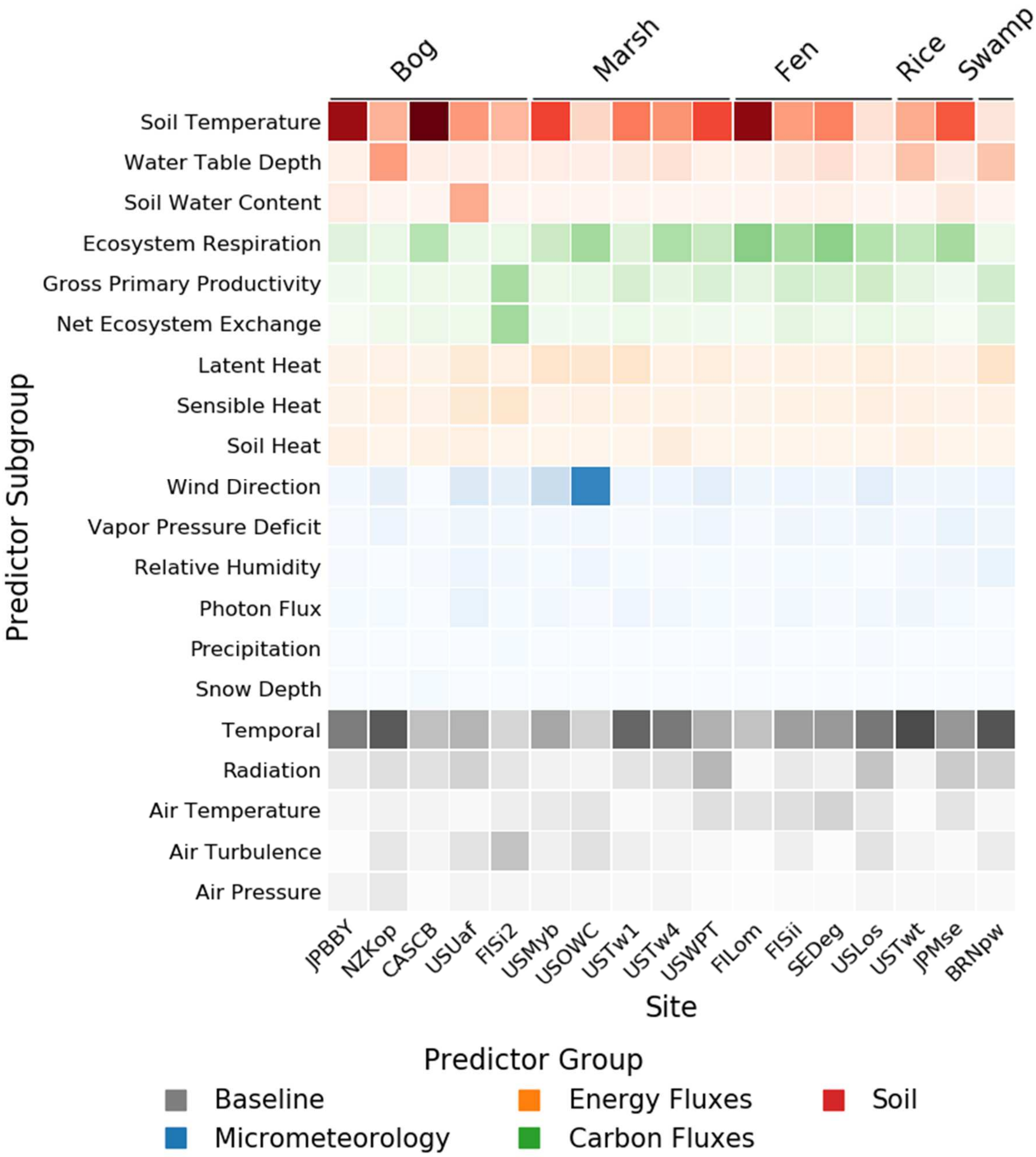


Figure 8: Predictor importance of the best model (RF+All) on each of the 17 primary sites. Darker color indicates higher importance assigned to that predictor for that site. The predictors within each group were arranged in descending order by the sum of the importance values across the sites. Note that all radiation predictors were grouped (e.g., incoming shortwave radiation (SW_IN), outgoing shortwave radiation

(SW_OUT), net radiation (NETRAD), etc.), as were air turbulence (friction velocity (USTAR) and wind speed (WS)). Similarly, predictors with alternative methods (e.g., daytime/nighttime partitioning) were grouped as were those with multiple depths of measurement (e.g. soil temperature (TS)). For full details please refer to **Table A.1**.

3.5 Uncertainty Estimation

The gap-filling prediction uncertainties for the two best ML algorithms (ANN and RF) were evaluated with respect to the concepts of calibration and sharpness. For ANN, the baseline predictor set model ensemble was evaluated because it most closely approximates a previously described method (Knox et al. 2019) which was used to gap-fill the FLUXNET-CH4 Version 1.0 community product (Delwiche et al. 2021). The prediction uncertainties of both the ANN and RF were not well-calibrated by default (**Figure 9**). In other words, without calibration by scaling, the 95% CI of the estimates for both models contained significantly less than 95% of the observed values (56.6% on average for ANN, 28.4% on average for RF), indicating that the models produced overly tight uncertainties across all sites. The ANN produced wider (less sharp) uncertainty estimates than the RF without calibration.

At all sites, both ANN and RF model prediction uncertainties were well-calibrated after performing the calibration step (**Figure 9**). In other words, the 95% CI of the estimates contained close to 95% of the observed values in the test set (95.6% on average for ANN, 95.2% on average for RF). Notably, once calibrated, the RF model made sharper predictions across all of the sites than the ANN model. The sites where predictions remained the widest (least sharp) after normalizing by the standard deviation of flux were US-Uaf, US-Twt, US-OWC, BR-Npw, and US-Los, which were the sites with the worst performance in terms of R^2 on the test set. These sites had one or more of a site-specific combination of low seasonality and/or extremely long gaps and/or highly variable fluxes. Similarly, the sites whose predictions were the sharpest corresponded to the sites with the best performance on the test set. Examples of pre- and post-calibration uncertainty ranges are shown in **Figure E.3**.

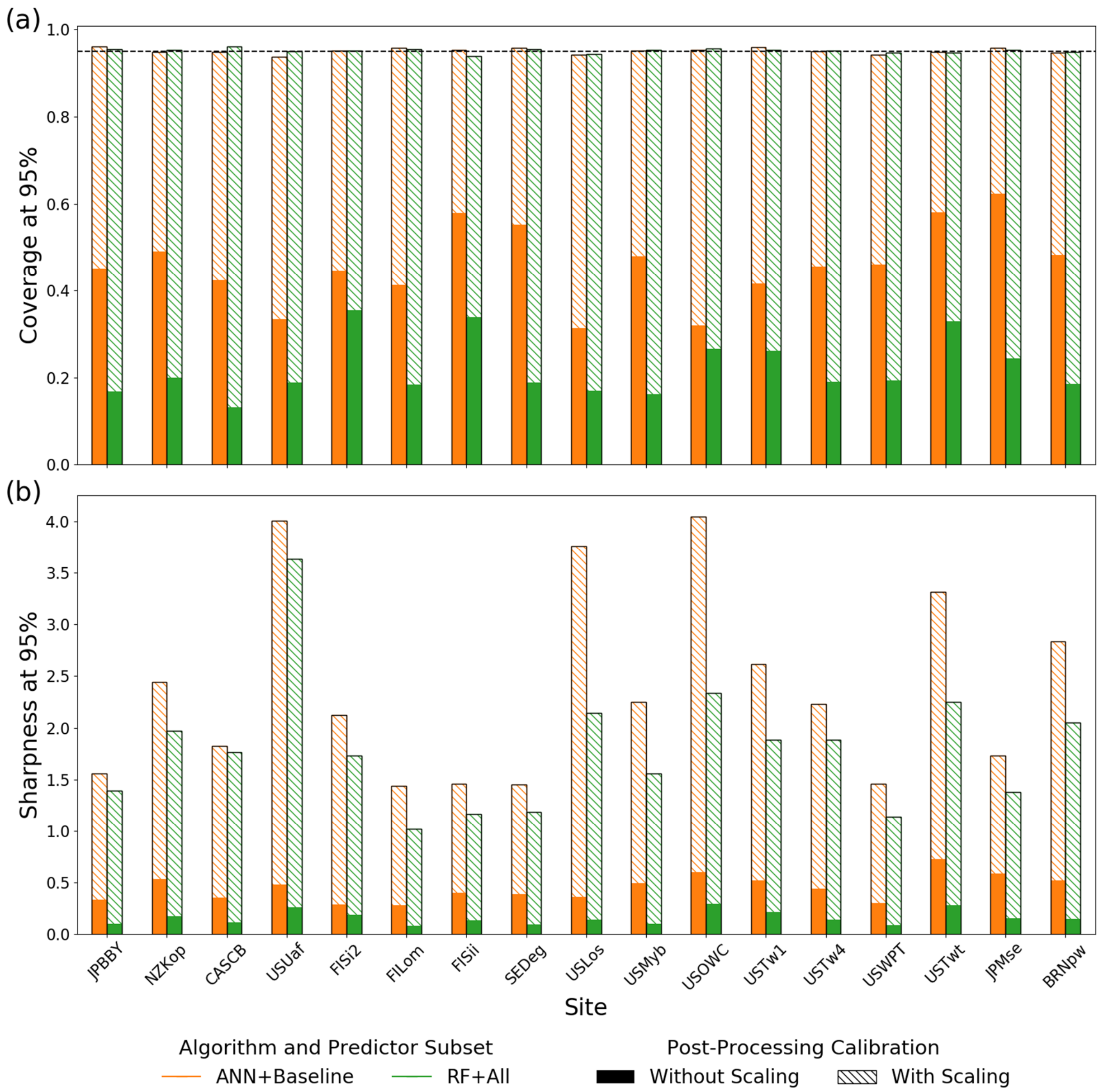


Figure 9. Per-site calibration and sharpness for the baseline model (ANN+Baseline) and best model (RF+All) before and after Platt scaling on the test set. The results without scaling (filled bar) represent the previous way of constructing uncertainty estimates, by training an ensemble of models and using the variation of the predictions without any adjustment, which leads to overly sharp confidence intervals measured by coverage. The results with scaling (hashed bar) incorporate a scaling factor which is learned from the data to adjust the ensemble uncertainty estimates and

yield calibrated uncertainties. Sharpness was measured as the mean width of the 95% uncertainty estimates on the test set normalized by the standard deviation of flux at the site.

3.6 Annual and Growing Season Emissions

A total of 30.4 site years were gap-filled with MDS with best (TA, WS, and PA) predictors, and the baseline ML (ANN plus baseline predictors) and best ML (RF plus all predictors) models and summed for annual or growing season CH_4 emissions. Note that reported uncertainties around summed emissions reflect only gap-filling uncertainties and exclude additional random uncertainties which, though tending to be small, can be considered separately (Knox et al. 2019) or in an integrated manner (Vitale et al. 2018).

Annual and growing season emissions did not differ significantly (measured by overlapping 95% CI) at any of the sites when comparing the two ML gap-filling methods (**Table 5**). Calibrated prediction uncertainties for ANN and RF resulted in less sharp, but more plausible, 95% CI around the annual sum. For all sites except US-OWC and BR-Npw, emissions from the best ML model (RF and All) fell within the unscaled 95% CI of the baseline model (ANN and Baseline; approximating Knox et al. 2019), supporting a generally high level of accuracy for the baseline method under the majority of site and gap conditions in this analysis. At the highly variable US-OWC marsh and BR-Npw swamp sites, the best model predictions fell outside the unscaled but within the scaled baseline CI, which underscores the implausible sharpness of unscaled ML ensemble predictions but does not support greater accuracy of RF than ANN. Uncertainties around MDS were much sharper (median 95% CI was ± 3% of annual emissions) than the scaled ML methods for ANN (± 38%) and RF (± 18%). The sharp uncertainties resulted in small but significant differences between annual and growing season sums from MDS and one ML model (e.g., JP-BBY, BR-Npw, US-Tw1) or both ML models (e.g., CA-SCB, US-Los).

Table 5: Mean annual and growing season emissions estimates for three methods (MDS, ANN, and RF) and their uncalibrated and calibrated uncertainties (95% CI) across the 17 sites. Calibration is only applicable to ML model ensemble methods and therefore cannot be reported for MDS.

Site	Annual or Growing Season Date Ranges	Mean Annual or Growing Season Methane Emissions ± Gap-

(class)	(Annual means only computed on years with good or comparable data coverage)	Filling Uncertainty (95% CI) (g CH_4-C m^{-2} y^{-1})		
		Best MDS, (TA, WS, PA) Unc. not scaled	**ANN+Baseline, (as in Knox et al. 2019)** Unc. not scaled Calibrated (lower)	**RF+All, Best model** Unc. not scaled Calibrated (lower)
JP-BBY (bog)	March 2016 - December 2017	17.84 ± 0.29	18.15 ± 0.86 18.22 ± 3.93	17.65 ± 0.13 17.65 ± 1.75
NZ-Kop (bog)	January 2012 - December 2014	17.57 ± 0.38	15.39 ± 1.78 17.97 ± 9.57	17.98 ± 0.28 17.98 ± 3.22
CA-SCB (bog)	April 2014 - November 2014 March 2016 - December 2016 March 2017 - November 2017	11.33 ± 0.24	11.21 ± 0.60 11.61 ± 2.82	11.60 ± 0.16 11.71 ± 2.05
US-Uaf (bog)	April 2011 - October 2011 May - October, 2012 - 2017 May 2018 - November 2018	0.57 ± 0.03	0.50 ± 0.09 0.57 ± 0.58	0.54 ± 0.03 0.56 ± 0.40
FI-Si2 (bog)	April - November, 2012 - 2013	11.36 ± 0.56	12.33 ± 1.23 12.60 ± 8.63	11.68 ± 0.91 11.81 ± 8.35
FI-Lom (fen)	January 2006 - December 2010	15.61 ± 0.16	15.75 ± 0.74 15.76 ± 3.83	15.63 ± 0.09 15.63 ± 1.12
FI-Sii (fen)	January 2013 - November 2014 March 2016 - December 2018	12.09 ± 0.36	12.47 ± 0.80 12.52 ± 2.9	12.07 ± 0.25 12.10 ± 2.12
SE-Deg (fen)	January 2014 - December 2016 January 2018 - December 2018	11.63 ± 0.14	11.44 ± 0.68 11.58 ± 2.18	11.30 ± 0.05 11.31 ± 0.60
US-Los (fen)	January 2014 - December 2018	6.56 ± 0.49	6.25 ± 1.29 7.79 ± 10.09	6.28 ± 0.20 6.63 ± 3.2
US-Myb (marsh)	January 2011 - December 2018	49.18 ± 0.79	47.97 ± 3.76 48.43 ± 16.71	49.14 ± 0.29 49.15 ± 4.44
US-OWC (marsh)	April 2016 - October 2016	116.85 ± 2.15	117.09 ± 7.56 120.07 ± 46.19	131.69 ± 7.97 132.44 ± 60.2
US-Tw1	January 2013 -	47.42 ± 2.09	44.81 ± 6.62	44.88 ± 0.87

(marsh)	December 2018		46.14 ± 32.2	44.89 ± 7.52
US-Tw4 (marsh)	January 2014 - December 2018	32.86 ± 0.70	32.32 ± 2.81 32.66 ± 13.87	32.63 ± 0.23 32.64 ± 3.05
US-WPT (marsh)	March 2011 - December 2013	50.45 ± 1.55	48.88 ± 3.02 49.21 ± 14.17	52.28 ± 0.66 52.27 ± 8.61
US-Twt (rice paddy)	April - October, 2010 - 2016	7.90 ± 0.44	8.06 ± 1.96 8.58 ± 8.41	8.44 ± 0.66 8.56 ± 5.07
JP-Mse (rice paddy)	May 2012 - September 2012	9.39 ± 0.44	8.88 ± 0.66 8.99 ± 1.75	9.51 ± 0.17 9.51 ± 1.57
BR-Npw (swamp)	January 2014 - December 2016	25.90 ± 1.61	19.22 ± 2.52 21.85 ± 14.23	24.73 ± 0.63 25.01 ± 8.01

4 Discussion

4.1 Methods & Algorithms

The gap-filling approach outlined in this study optimizes for the training and evaluation of ML gap-filling models. A new technique is proposed for generating artificial gap scenarios that resemble the true observed gap distributions. This is important to ensure that ML models are trained and scored on unbiased distributions of gap lengths. Using this artificial gap generation procedure, one can generate many site-specific scenarios and reliably evaluate models on their ability to fill data gaps. There are trade-offs between this approach and the introduction of uniform gap-length scenarios (e.g., (Moffat et al. 2007), which alternatively ensures a consistent number of scorable gaps (even extremely long gaps) at the expense of unbiased training conditions. However, the proposed method is recommended for ML-focused studies given that the gap-filling of extremely long gaps (e.g., multiple months) is much less reliable, regardless of the method used, and are best avoided entirely, if possible.

Decision tree-based models (RF and XGBoost) showed better performance than ANN and Lasso models across the majority of the 17 wetland and rice paddy sites. This is consistent with recent work on CH_4 gap-filling which demonstrated that a RF gap-filling model outperformed both ANN and support vector regression models across five wetland and rice paddy sites (Nemitz et al. 2018; Kim et al. 2019; Knox et al. 2019). RF models are also relatively easy to tune, fast to train even on large datasets, and require little preprocessing. Furthermore, decision-tree-based models are more interpretable (presently) than ANN (Russell and Norvig 1995), which enables analysis of important predictors. In comparison to ML approaches, MDS was tested as an easy and fast method that makes use of only three predictors. MDS scored highly on average although still much lower than the best ML models. Kim et al. (2019) also found that MDS more frequently introduced statistical bias in annual sums than ML models.

Although RF and ANN models are recommended ML methods, there is still room to improve their gap-filling performance, especially on long gaps. Recent deep neural network architectures have shown impressive results in modeling long sequences in natural language processing, particularly recurrent neural network variants (Lipton et al. 2015) and Transformers (Vaswani et al. 2017). These models have the potential to reproduce highly nonlinear variable interactions using large datasets including half-hourly time series flux data and may be able to capture lagged relationships between predictors and CH_4 flux without further manual revision. However, representing non-stationary conditions such as pulse events has proven to be challenging for ML approaches (Vargas et al. 2018). Future work could explore the effectiveness of deep neural network architectures for gap-filling CH_4. It is likely, however, that problems of non-stationarity during long gaps will apply for CH_4 as they do for CO_2 imputation (Richardson and Hollinger 2007) and are best handled during data collection.

4.2 Methane Predictors

The inclusion of soil temperature (TS) and ecosystem carbon flux predictors (NEE, RECO, and GPP) improved gap-filling performance over the baseline set (three temporal, plus TA, PA, SW_IN, and WS), in broad agreement with known controls by temperature (Yvon-Durocher et al. 2014) and substrate availability (Whiting and Chanton 1993; Matthes et al. 2014; McNicol et al. 2020; Laanbroek 2010). Soil temperature was the single most important additional predictor over the baseline set at most sites, followed by RECO. While TS was available at all sites in this study, it is not available across all FLUXNET sites. Although NEE and its component ecosystem carbon fluxes (GPP and RECO) are highly correlated, the consistent favoring of RECO suggests they are not perfectly interchangeable for gap-filling performance, and RECO and CH_4 flux are both largely the result of microbial metabolism, and are similarly affected by environmental drivers (Morin et al. 2014), However, partitioned fluxes (RECO and GPP) are overall less practical than measured NEE as predictors because they are typically partitioned from NEE as a function of TS, and thus its importance may largely reflect its correlation with TS

(Reichstein et al. 2005; Keenan et al. 2019) while RECO is limited in its ability to represent respiration fluxes across different ecosystems (Barba et al. 2018).

Water table depth, a proxy for the balance of anaerobic CH_4-producing and aerobic CH_4-consuming soil volumes (Bridgham et al. 2013), was an important predictor at rice and swamp sites that undergo larger changes in seasonal inundation (Dalmagro et al. 2018; Muramatsu et al. 2017), but not at other wetland types. Although WTD has been found to be important in bogs and fens (Moore et al. 2011; Goodrich et al. 2015; Koebsch et al. 2020), it was only an important gap-filling predictor at one of the five bogs in this study. This is consistent with prior work showing that WTD becomes important when its range is large and/or crosses above and below the soil surface (Knox et al. 2019; Alekseychik et al. 2021; Knox et al. 2021). Moreover, in some wetlands, WTD is only a coarse proxy for anaerobic volume activity due to the presence of anaerobic microsites in drained layers and anaerobic methane oxidation in saturated layers (Yang et al. 2017). Although WTD was available at all 17 sites, it is only currently reported for half of wetland sites in FLUXNET-CH4 (Knox et al. 2019). The moderate importance of WTD measurements as a predictor in many sites, and high importance in some, suggests it should be widely collected and reported to ensure optimal CH_4 gap-filling when using ML models. The predictor experiments also allowed us to investigate the usefulness of broad classes of predictors. As "fuzzy" temporal predictors (cosine year, sine year, and delta) (Moffat et al. 2007), can be computed, they are always recommended for gap-filling. It was also confirmed that the most useful meteorological predictors (TA, SW_IN, WS and PA) were already included in the baseline model of a recent synthesis (Knox et al. 2019).

The performance improvements using all predictors in this study suggests a moderate amount of predictor redundancy does not harm ML performance and predictor curation may be less important for ML than in other modeling approaches. Kim et al. (2019) similarly showed that ML models can benefit from a large predictor set that includes soil variables and that dimension-reduction via principal component analysis was not necessary to achieve good performance. However, site uniqueness may also necessitate the tailoring of models for optimal performance at individual sites, illustrated in this study by the ranges in 1) observed CH_4 fluxes, 2) model performance, and 3) predictor importance within bog, fen, and marsh classes. For instance, despite high spatial variability in CH_4 fluxes at some wetlands (Rey-Sanchez et al. 2018; Matthes et al. 2014), WD (which determines the flux footprint) was only an important predictor at one marsh site (US-OWC), which has very high spatial variation in flux between different cover types (Rey-Sanchez et al. 2018). The site-specificity of WD for heterogeneous sites was also reported in a recent study that used a ML approach to partition NEE (Tramontana et al. 2020). Entirely new predictors may also be necessary at some sites, such as salinity, which is likely an important predictor for gap-filling at estuaries or other coastal locations with a (tidal) salinity influence (Holmquist et al. 2018; Poffenbarger et al. 2011). Although not prioritized in the present study, a more parsimonious predictor set may be identified via a combination of site-specific and process knowledge, as well as automated feature selection methods (Kumar and Minz 2014). Curated predictor sets should, however, be reevaluated when gap-filling new data (e.g., site-years, or across multiple sites) as past models may be overfit with respect to new data conditions.

Future work could also explore the use of led or lagged predictors, which could be used to engineer predictors with greater coherence with CH_4 flux (Vitale et al. 2018). For example, recent syntheses have demonstrated that the timing and seasonality of CH_4 fluxes lags TS across several FLUXNET-CH4 sites (Delwiche et al. 2021), leading to an apparent hysteretic dependency (Chang et al. 2021), and therefore using lagged TS predictors may improve ML gap-filling performance. More sophisticated feature selection methods are possible, such as information theory, which can be used to first identify the predictor and timescale of the lag (or lead), and then curate a more parsimonious predictor set (e.g., Sturtevant et al. 2016; Knox et al. 2021). Overall, improvements in the measurement and coverage of key soil predictors, especially high-quality soil temperature and water table depth data, is recommended.

4.3 Integrated Emissions & Uncertainties

Computing annual or growing season CH_4 emissions requires gap-filling because filtering of EC data and other acquisition issues typically creates gaps of a wide variety of lengths, and especially an abundance of short gaps (**Table C2**). Gaps are not normally distributed in time and therefore FCH4 observations are likely to be biased, which will propagate to the time-integrated flux. However, the investigator must decide: 1) which gap-filled values are likely to be of sufficient accuracy to be retained, and 2) whether the retained gap-filled plus observed values are sufficient to integrate emissions over an annual, seasonal, or other timeframe. As a rough guide, filled values should be treated with greater scrutiny as they become longer and less frequent in the scorable dataset. The most abundant scorable gaps of length one half-hour to approximately 12 days can be filled confidently, given performance metric checks as described in this study. Investigators should, however, be aware that episodic fluxes, perhaps due to ebullition events, may not always be captured and instead may be filled with average fluxes for the most comparable conditions (e.g., FCH4 and MAE spikes in **Figure E.2**). Greater scrutiny of evaluation metrics is recommended for gaps longer than approximately 12 days, but less than multiple months, whereas, filled values in gaps of multiple months (> 60 days) should generally be excluded, as is done in CO_2 gap-filling (Wutzler et al. 2018). The exception may be very long (decadal) datasets where the monthly-scale gap occurs in a season with ample data from other sites-years and can be reasonably evaluated. After determining which filled values to retain, the coverage of filled plus observed fluxes should be considered with respect to the integration period. For rice paddies (e.g., US-Twt, JP-Mse), and sites with low winter season fluxes due to frozen soils (US-OWC or US-Uaf), it may be adequate and interesting to report a growing season flux as is done in this study and the FLUXNET-CH4 synthesis (Delwiche et al. 2021). Time-integrated uncertainties from ML gap-filling methods will also widen significantly as more gap-filling is required and should always be reported alongside long-term sums.

The improvement in performance gained by using ML over MDS, and all predictors over baseline predictors, did not have a significant effect on annual CH_4 emissions estimates at most sites. However, seemingly minor changes in CH_4 fluxes can have disproportionate impacts when calculating greenhouse gas emissions due to the high radiative forcing effects of CH_4 or when sparsely distributed sites are used in data-driven regional or global upscaling efforts (Tramontana et al. 2016; Roberts et al. 2017). Specifically, absolute differences in annual emissions among the gap-filling methods were larger at high-emitting sites which could lead to larger upscaling errors in high-emitting tropical regions that account for > 60% of global wetland sources (Wania et al. 2013; Bloom et al. 2017; Saunois et al. 2020). These results therefore highlight the need for robust methods for estimating and propagating uncertainty from flux gap-filling to upscaling.

Machine learning model-generated uncertainties around both half-hourly predictions and annual emissions have been underestimated. A scaling procedure (Platt scaling) which expands the uncertainty estimates can be used to produce well-calibrated predictions. Well-calibrated models can be compared using the sharpness of their predictions, where sharper predictions corresponded to better models. Using this method, sharper uncalibrated RF (compared to ANN) prediction uncertainties were retained post-calibration, indicating greater precision of predictions. However, the frequent overlap between uncalibrated and calibrated for both algorithms means a firm conclusion about algorithm differences in accuracy is not possible. It is also acknowledged that this uncertainty does not capture all sources of uncertainty that could arise from random measurement errors, unseen events, uncertainties in the predictors, or other systematic bias, among others. However, calibrating predictive ML models to avoid underestimating gap-filling uncertainties is strongly recommended.

Other calibration methods have the potential to achieve calibration while producing sharper predictions (Kuleshov et al. 2018). Furthermore, probabilistic models like Gaussian processes or multiple imputation methods may be able to produce well-calibrated models without the need for post-processing calibration procedures (Vitale et al. 2018; Camps-Valls et al. 2019). Recently, a method for producing uncertainty estimates from any gradient boosting model was introduced which may enable decision tree models to produce well-calibrated, probabilistic predictions without requiring a model ensemble or post-processing calibration (Duan et al. 2019). Finally, deep learning models can capture highly nonlinear relationships in large datasets and make probabilistic predictions which have the potential to outperform other gap-filling methods.

5 Conclusions

This study outlines a robust and reproducible ML workflow for CH_4 gap-filling models that can be applied at individual wetland sites or in multi-site syntheses. Specifically, the study advances CH_4 gap-filling in wetlands using ML by: 1) introducing a thorough gap-filling model development and validation procedure that reliably generates gaps and splits the data into training, validation, and test sets; 2) experimentally evaluating conventional MDS (with drivers adapted for wetland CH_4 fluxes) against combinations of ML algorithms and predictor sets; and 3) proposing a model calibration method to estimate, evaluate, and calibrate model uncertainties. This study also provides insights into methodological choices. Decision tree algorithms (RF and XGBoost) offer the best performance on average; using all predictors (or best set for MDS), median nMAE followed the order Lasso (0.42) > MDS (0.34) > ANN (0.31) > RF/XGBoost (0.26), and median R^2 followed the order Lasso (0.57) < MDS (0.66) < ANN (0.70) < RF/XGBoost (0.79). Overall, RF is recommended as it benefits from less pre-processing and faster run-time than XGBoost. ANN predictions had less bias when filling the longest gaps and performance improved when using all rather than baseline predictors, suggesting ANN may benefit from additional predictor curation and feature engineering. Using all available variables collected at eddy covariance towers as predictors is also fast, effective, and reasonable, given the large ratio of observations to predictors (favorable data dimensionality). Conventional MDS also proved to be a fast method that provides reasonable performance when CH_4 predictors (air temperature, air pressure, and wind speed) are selected, however, the lack of post-calibration results in uncertainties that are very sharp (unrealistic). ML prediction uncertainties, in contrast, can be calibrated to observations using Platt scaling. Finally, based on variable importance results, it is recommended that soil temperature and water table depth are measured at all wetland eddy covariance sites. The python code for developing gap-filling methods, comparing predictions, and calibrating uncertainties is available [https://github.com/stanfordmlgroup/methane-gapfill-ml]. For future evaluations at wetlands and other ecosystems, this code can provide a foundation for the development of standardized eddy covariance CH_4 processing by different teams and Regional Flux Networks which can also be tested on nitrous oxide fluxes as longer time series become available (Papale 2020).

Acknowledgments

This study was supported by the Gordon and Betty Moore Foundation through Grant GBMF5439 “Advancing Understanding of the Global Methane Cycle” to Stanford University supporting the Methane Budget activity for the Global Carbon Project (globalcarbonproject.org). BRKR was supported by NSF Award 1752083. OS was supported through the Canada Research Chairs and Natural Sciences and Engineering Research Council Discovery Grants programs. TS and CW were supported by the Helmholtz Association of German Research Centres (Grant No. VH-NG-821). GB was supported by a US Department of Energy (Grant DE-SC0021067) and NOAA Davidson Fellowship Award administered by ODNR OWC-NERR (Subaward N18B 315-11). Funding from the SNF projects DiRad and InnoFarm (146373 and 407340_172433), from the ETH Board and from ETH Zurich is greatly acknowledged. DP, CT and DV were supported by the Department of Excellence 2018 Program MIUR Project “Landscape 4.0 – food, wellbeing and environment" and the ICOS Ecosystem Thematic Centre. CT was supported by the E-SHAPE (GA820852) H2020 European project. DB, JV, DS, CS, SK, EE, KSH, KK, AV, CRS, RS (or sites US-MYB, US-TW1, US-TW3, US-TW4, US-TW5, US-TWT, US-SND, US-SNE) were supported by the California Department of Water Resources through a contract from the California Department of Fish and Wildlife and the United States Department of Agriculture (NIFA grant #2011-67003-30371). Funding for the AmeriFlux core sites was provided by the U.S. Department of Energy’s Office of Science (AmeriFlux contract #7079856). KK was supported by the Estonian Research Council grants No. PSG631 and

PRG352. KSH was supported by the California Sea Grant Delta Science Fellowship (programs R/SF-70, grant no. 2271). The contents of this material do not necessarily reflect the views and policies of the Delta Stewardship Council or California Sea Grant, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. MU was supported by the Arctic Challenge for Sustainability II (JPMXD1420318865) and the JSPS KAKENHI (20K21849). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. IM thanks H2020 RINGO project (Grant Agreement 730944), the Academy of Finland Flagship funding (grant no. 337549) and ICOS-Finland by University of Helsinki funding.

Author Contributions

AN, BP, RBJ, SHK, and LWM acquired funding for and conceived of the project. EFC, FL, GM, JI, SZ, VL, and ZO conceived of, and AA, ALM, CT, DP, DV, and IM contributed to, the design and execution of the machine learning analysis. ALM was consulted with on machine learning model and artificial gap evaluation. CT and DP contributed the marginal distribution sampling analysis. DV was consulted with on multi-imputation methods. EFC, FL, GM, JI, SZ VL, and ZO wrote the initial draft of the manuscript and ALM, ACRS, ACV, ADR, AK, AL, AM, AN, ARD, BM, BRKR, CH, CS, CT, DDB, DIC, DP, DV, DYFL, DZ, EE, EJW, ESH, GB, GJ, GLV, GXW, HC, HI, HJD, IM, JC, KBD, KSH, KVRS, LM, LWM, MA, MBN, MG, MHelbig, MHeimann, MP, MU, OS, PA, RBJ, RV, SB, SF, SHK, and TH contributed edits to subsequent drafts. ADR, ARD, CH, DDB, DIC, DYFL, GB, GLV, HI, HJD, IM, JC, KVRS, MA, MBN, MH, MU, OS, TF, TH, and TS were affiliated as principal investigators for the 17 core analysis sites. All other coauthors contributed data as principal investigators or were named as affiliated team members at other FLUXNET-CH4 sites.

References

Alekseychik, P., Korrensalo, A., Mammarella, I., Launiainen, S., Tuittila, E.-S., Korpela, I., Vesala, T., 2021. Carbon balance of a Finnish bog: temporal variability and limiting factors. https://doi.org/10.5194/bg-2020-488

Bansal, S., Tangen, B., Finocchiaro, R., 2018. Diurnal Patterns of Methane Flux from a Seasonal Wetland: Mechanisms and Methodology. Wetlands 38, 933–943. https://doi.org/10.1007/s13157-018-1042-5

Barba, J., Cueva, A., Bahn, M., Barron-Gafford, G.A., Bond-Lamberty, B., Hanson, P.J., Jaimes, A., Kulmala, L., Pumpanen, J., Scott, R.L., Wohlfahrt, G., Vargas, R., 2018. Comparing ecosystem and soil respiration: Review and key challenges of tower-based and soil measurements. Agric. For. Meteorol. 249, 434–443. https://doi.org/10.1016/j.agrformet.2017.10.028

Bloom, A.A., Bowman, K.W., Lee, M., Turner, A.J., Schroeder, R., Worden, J.R., Weidner, R.J., Mcdonald, K.C., Jacob, D.J., 2017. CMS: Global 0.5-deg Wetland Methane Emissions and Uncertainty (WetCHARTs v1. 0). https://doi.org/10.3334/ORNLDAAC/1502

Bodesheim, P., Jung, M., Gans, F., Mahecha, M.D., Reichstein, M., 2018. Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product. Earth Syst. Sci. Data 10, 1327–1365. https://doi.org/10.5194/essd-10-1327-2018

Bohrer, G., Kerns, J., Morin, T., Rey-Sanchez, A., Villa, J., Ju, Y., 2020. FLUXNET-CH4 US-OWC Old Woman Creek. https://doi.org/10.18140/FLX/1669690

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Bridgham, S.D., Cadillo-Quiroz, H., Keller, J.K., Zhuang, Q., 2013. Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. Glob. Chang. Biol. 19, 1325–1346. https://doi.org/10.1111/gcb.12131

Campbell, D., Goodrich, J., 2020. FLUXNET-CH4 NZ-Kop Kopuatai. https://doi.org/10.18140/FLX/1669652

Camps-Valls, G., Sejdinovic, D., Runge, J., Reichstein, M., 2019. A perspective on Gaussian processes for Earth observation. Natl Sci Rev 6, 616–618. https://doi.org/10.1093/nsr/nwz028

Chang, K.-Y., Riley, W.J., Knox, S.H., Jackson, R.B., McNicol, G., Poulter, B., Aurela, M., Baldocchi, D., Bansal, S., Bohrer, G., Campbell, D.I., Cescatti, A., Chu, H., Delwiche, K.B., Desai, A.R., Euskirchen, E., Friborg, T., Goeckede, M., Helbig, M., Hemes, K.S., Hirano, T., Iwata, H., Kang, M., Keenan, T., Krauss, K.W., Lohila, A., Mammarella, I., Mitra, B., Miyata, A., Nilsson, M.B., Noormets, A., Oechel, W.C., Papale, D., Peichl, M., Reba, M.L., Rinne, J., Runkle, B.R.K., Ryu, Y., Sachs, T., Schäfer, K.V.R., Schmid, H.P., Shurpali, N., Sonnentag, O., Tang, A.C.I., Torn, M.S., Trotta, C., Tuittila, E.-S., Ueyama, M., Vargas, R., Vesala, T., Windham-Myers, L., Zhang, Z., Zona, D., 2021. Substantial hysteresis in emergent temperature sensitivity of global wetland CH4 emissions. Nat. Commun. 12, 1–10. https://doi.org/10.1038/s41467-021-22452-1

Chen, J., Chu, H., 2020. FLUXNET-CH4 US-WPT Winous Point North Marsh. https://doi.org/10.18140/FLX/1669702

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG].

Dalmagro, H.J., Lathuillière, M.J., Hawthorne, I., Morais, D.D., Pinto, O.B., Jr, Couto, E.G., Johnson, M.S., 2018. Carbon biogeochemistry of a flooded Pantanal forest over three annual flood cycles. Biogeochemistry 139, 1–18. https://doi.org/10.1007/s10533-018-0450-1

Delwiche, K.B., Knox, S.H., Malhotra, A., Fluet-Chouinard, E., McNicol, G., Feron, S., Ouyang, Z., Papale, D., Trotta, C., Canfora, E., Cheah, Y.-W., Christianson, D., Alberto, M.C.R., Alekseychik, P., Aurela, M., Baldocchi, D., Bansal, S., Billesbach, D.P., Bohrer, G., Bracho, R., Buchmann, N., Campbell, D.I., Celis, G., Chen, J., Chen, W., Chu, H., Dalmagro, H.J., Dengel, S., Desai, A.R., Detto, M., Dolman, H., Eichelmann, E., Euskirchen, E., Famulari, D., Friborg, T., Fuchs, K., Goeckede, M., Gogo, S., Gondwe, M.J., Goodrich, J.P., Gottschalk, P., Graham, S.L., Heimann, M., Helbig, M., Helfter, C., Hemes, K.S., Hirano, T., Hollinger, D., Hörtnagl, L., Iwata, H., Jacotot, A., Jansen, J., Jurasinski, G., Kang, M., Kasak, K., King, J., Klatt, J., Koebsch, F., Krauss, K.W., Lai, D.Y.F., Mammarella, I., Manca, G., Marchesini, L.B., Matthes, J.H., Maximon, T., Merbold, L., Mitra, B., Morin, T.H., Nemitz, E., Nilsson, M.B., Niu, S., Oechel, W.C., Oikawa, P.Y., Ono, K., Peichl, M., Peltola, O., Reba, M.L., Richardson, A.D., Riley, W., Runkle, B.R.K., Ryu, Y., Sachs, T., Sakabe, A., Sanchez, C.R., Schuur, E.A., Schäfer, K.V.R., Sonnentag, O., Sparks, J.P., Stuart-Haëntjens, E., Sturtevant, C., Sullivan, R.C., Szutu, D.J., Thom, J.E., Torn, M.S., Tuittila, E.-S., Turner, J., Ueyama, M., Valach, A.C., Vargas, R., Varlagin, A., Vazquez-Lule, A., Verfaillie, J.G., Vesala, T., Vourlitis, G.L., Ward, E.J., Wille, C., Wohlfahrt, G., Wong, G.X., Zhang, Z., Zona, D., Windham-Myers, L., Poulter, B., Jackson, R.B., 2021. FLUXNET-CH4: A global, multi-ecosystem dataset and analysis of methane seasonality from freshwater wetlands. Earth Syst. Sci. Data. https://doi.org/10.5194/essd-2020-307

Dengel, S., Zona, D., Sachs, T., Aurela, M., Jammet, M., Parmentier, F.J.W., Oechel, W., Vesala, T., 2013. Testing the applicability of neural networks as a gap-filling method using CH_4 flux data from high latitude wetlands. Biogeosciences 10, 8185–8200. https://doi.org/10.5194/bg-10-8185-2013

Derrac, J., García, S., Molina, D., Herrera, F., 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evol. Comput. 1, 3–18. https://doi.org/10.1016/j.swevo.2011.02.002

Desai, A., 2020. FLUXNET-CH4 US-Los Lost Creek. https://doi.org/10.18140/FLX/1669682

Duan, T., Avati, A., Ding, D.Y., Basu, S., Ng, A.Y., Schuler, A., 2020. NGBoost: Natural Gradient Boosting for Probabilistic Prediction, in: International Conference on Machine Learning. PMLR, pp. 2690–2700.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press.

Eichelmann, E., Knox, S., Sanchez, C., Valach, A., Sturtevant, C., Szutu, D., Verfaillie, J., Baldocchi, D., 2020. FLUXNET-CH4 US-Tw4 Twitchell East End Wetland. https://doi.org/10.18140/FLX/1669698

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., Granier, A., Gross, P., Grünwald, T., Hollinger, D., Jensen, N.-O., Katul, G., Keronen, P., Kowalski, A., Lai, C.T., Law, B.E., Meyers, T., Moncrieff, J., Moors, E., Munger, J.W., Pilegaard, K., Rannik, Ü., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agric. For. Meteorol. 107, 43–69. https://doi.org/10.1016/s0168-1923(00)00225-2

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086

Freund, Y., Schapire, R., Abe, N., 1999. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14, 1612.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Series B Stat. Methodol. 69, 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annu. Rev. Stat. Appl. 1, 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Göckede, M., Kittler, F., Schaller, C., 2019. Quantifying the impact of emission outbursts and non-stationary flow on eddy-covariance CH_4 flux measurements using wavelet techniques. Biogeosciences 16, 3113–3131. https://doi.org/10.5194/bg-16-3113-2019

Goodrich, J.P., Campbell, D.I., Roulet, N.T., Clearwater, M.J., Schipper, L.A., 2015. Overriding control of methane flux temporal variability by water table dynamics in a Southern Hemisphere, raised bog: Methane fluxes from a S.H. bog. J. Geophys. Res. Biogeosci. 120, 819–831. https://doi.org/10.1002/2014jg002844

Günther, A., Barthelmes, A., Huth, V., Joosten, H., Jurasinski, G., Koebsch, F., Couwenberg, J., 2020. Prompt rewetting of drained peatlands reduces climate warming despite methane emissions. Nat. Commun. 11, 1644. https://doi.org/10.1038/s41467-020-15499-z

Hatala, J.A., Detto, M., Baldocchi, D.D., 2012. Gross ecosystem photosynthesis causes a diurnal pattern in methane emission from rice. Geophys. Res. Lett. 39, L06409. https://doi.org/10.1029/2012gl051303

Hemes, K.S., Chamberlain, S.D., Eichelmann, E., Anthony, T., Valach, A., Kasak, K., Szutu, D., Verfaillie, J., Silver, W.L., Baldocchi, D.D., 2019. Assessing the carbon and climate benefit of restoring degraded agricultural peat soils to managed wetlands. Agric. For. Meteorol. 268, 202–214. https://doi.org/10.1016/j.agrformet.2019.01.017

Hollinger, D.Y., Richardson, A.D., 2005. Uncertainty in eddy covariance measurements and its application to physiological models. Tree Physiol. 25, 873–885. https://doi.org/10.1093/treephys/25.7.873

Holmquist, J.R., Windham-Myers, L., Bernal, B., Byrd, K.B., Crooks, S., Gonneea, M.E., Herold, N., Knox, S.H., Kroeger, K.D., McCombs, J., Megonigal, J.P., Lu, M., Morris, J.T., Sutton-Grier, A.E., Troxler, T.G., Weller, D.E., 2018. Uncertainty in United States coastal wetland greenhouse gas inventorying. Environ. Res. Lett. 13, 115005. https://doi.org/10.1088/1748-9326/aae157

Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y., 2004. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. Agric. For. Meteorol. 121, 93–111. https://doi.org/10.1016/s0168-1923(03)00158-8

Iwata, H., 2020a. FLUXNET-CH4 JP-Mse Mase rice paddy field. https://doi.org/10.18140/FLX/1669647

Iwata, H., Ueyama, M., Harazono, Y., 2020b. FLUXNET-CH4 US-Uaf University of Alaska, Fairbanks. https://doi.org/10.18140/FLX/1669701

Keenan, T.F., Migliavacca, M., Papale, D., Baldocchi, D., Reichstein, M., Torn, M., Wutzler, T., 2019. Widespread inhibition of daytime ecosystem respiration. Nat Ecol Evol 3, 407–415. https://doi.org/10.1038/s41559-019-0809-2

Kim, Y., Johnson, M.S., Knox, S.H., Black, T.A., Dalmagro, H.J., Kang, M., Kim, J., Baldocchi, D., 2020. Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. Glob. Chang. Biol. 26, 1499–1518. https://doi.org/10.1111/gcb.14845

Knox, S., Matthes, J., Verfaillie, J., Baldocchi, D., 2020. FLUXNET-CH4 US-Twt Twitchell Island. https://doi.org/10.18140/FLX/1669700

Knox, S.H., Bansal, S., McNicol, G., Schafer, K., Sturtevant, C., Ueyama, M., Valach, A.C., Baldocchi, D., Delwiche, K., Desai, A.R., Euskirchen, E., Liu, J., Lohila, A., Malhotra, A., Melling, L., Riley, W., Runkle, B.R.K., Turner, J., Vargas, R., Zhu, Q., Alto, T., Fluet-Chouinard, E., Goeckede, M., Melton, J.R., Sonnentag, O., Vesala, T., Ward, E., Zhang, Z., Feron, S., Ouyang, Z., Alekseychik, P., Aurela, M., Bohrer, G., Campbell, D.I., Chen, J., Chu, H., Dalmagro, H.J., Goodrich, J.P., Gottschalk, P., Hirano, T., Iwata, H., Jurasinski, G., Kang, M., Koebsch, F., Mammarella, I., Nilsson, M.B., Ono, K., Peichl, M., Peltola, O., Ryu, Y., Sachs, T., Sakabe, A., Sparks, J., Tuittila, E.-S., Vourlitis, G.L., Wong, G.X., Windham-Myers, L., Poulter, B., Jackson, R.B., 2021. Identifying dominant environmental predictors of freshwater wetland methane fluxes across diurnal to seasonal time scales. Glob. Chang. Biol. https://doi.org/10.1111/gcb.15661

Knox, S.H., Jackson, R.B., Poulter, B., McNicol, G., Fluet-Chouinard, E., Zhang, Z., Hugelius, G., Bousquet, P., Canadell, J.G., Saunois, M., Papale, D., Chu, H., Keenan, T.F., Baldocchi, D., Torn, M.S., Mammarella, I., Trotta, C., Aurela, M., Bohrer, G., Campbell, D.I., Cescatti, A., Chamberlain, S., Chen, J., Chen, W., Dengel, S., Desai, A.R., Euskirchen, E., Friborg, T., Gasbarra, D., Goded, I., Goeckede, M., Heimann, M., Helbig, M., Hirano, T., Hollinger, D.Y., Iwata, H., Kang, M., Klatt, J., Krauss, K.W., Kutzbach, L., Lohila, A., Mitra, B., Morin, T.H., Nilsson, M.B., Niu, S., Noormets, A., Oechel, W.C., Peichl, M., Peltola, O., Reba, M.L., Richardson, A.D., Runkle, B.R.K., Ryu, Y., Sachs, T., Schäfer, K.V.R., Schmid, H.P., Shurpali, N., Sonnentag, O., Tang, A.C.I., Ueyama, M., Vargas, R., Vesala, T., Ward, E.J., Windham-Myers, L., Wohlfahrt, G., Zona, D., 2019. FLUXNET-CH4 synthesis activity: Objectives, observations, and future directions. Bull. Am. Meteorol. Soc. 100, 2607–2632. https://doi.org/10.1175/bams-d-18-0268.1

Knox, S.H., Matthes, J.H., Sturtevant, C., Oikawa, P.Y., Verfaillie, J., Baldocchi, D., 2016. Biophysical controls on interannual variability in ecosystem-scale CO2 and CH4 exchange in a California rice paddy. J. Geophys. Res. Biogeosci. 121, 978–1001. https://doi.org/10.1002/2015jg003247

Koebsch, F., Gottschalk, P., Beyer, F., Wille, C., Jurasinski, G., Sachs, T., 2020. The impact of occasional drought periods on vegetation spread and greenhouse gas exchange in rewetted fens. Philos. Trans. R. Soc. Lond. B Biol. Sci. 375, 20190685. https://doi.org/10.1098/rstb.2019.0685

Kuleshov, V., Fenner, N., Ermon, S., 2018. Accurate Uncertainties for Deep Learning Using Calibrated Regression. arXiv [cs.LG].

Kumar, V., Minz, S., 2014. Feature Selection: A literature review. Smart Computing Review 4, 211–229. https://doi.org/10.6029/smartcr.2014.03.007

Laanbroek, H.J., 2010. Methane emission from natural wetlands: interplay between emergent macrophytes and soil microbial processes. A mini-review. Ann. Bot. 105, 141–153. https://doi.org/10.1093/aob/mcp201

Lasslop, G., Reichstein, M., Kattge, J., Papale, D., 2008. Influences of observation errors in eddy flux data on inverse model parameter estimation. Biogeosciences 5, 1311–1324. https://doi.org/10.5194/bg-5-1311-2008

Lasslop, G., Reichstein, M., Papale, D., Richardson, A.D., Arneth, A., Barr, A., Stoy, P., Wohlfahrt, G., 2010. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. Glob. Chang. Biol. 16, 187–208. https://doi.org/10.1111/j.1365-2486.2009.02041.x

Li, X., Wahlroos, O., Haapanala, S., Pumpanen, J., Vasander, H., Ojala, A., Vesala, T., Mammarella, I., 2020. Carbon dioxide and methane fluxes from different surface types in a created urban wetland. Biogeosciences 17, 3409–3425. https://doi.org/10.5194/bg-17-3409-2020

Lipton, Z.C., Berkowitz, J., Elkan, C., 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. arXiv [cs.LG].

Lohila, A., Aurela, M., Tuovinen, J.-P., Laurila, T., Hatakka, J., Rainne, J., Mäkelä, T., 2020. FLUXNET-CH4 FI-Lom Lompolojankka. https://doi.org/10.18140/FLX/1669638

Mammarella, I., Aslan, T., Burba, G., Cowan, N., Helfter, C., Herbst, M., Hörtnagl, L., Ibrom, A., Lucas-Moffat, A.M., Nicolini, G., Papale, D., Peltola, O., Rannik, Ü., Vitale, D., Yeung, K., Nemitz, E., 2020. Protocol for non-CO2 eddy covariance measurements, QA/QC, data processing and gap-filling. Readiness of ICOS for Necessities of integrated Global Observations (RINGO).

Matthes, J., Sturtevant, C., Oikawa, P., Chamberlain, S., Szutu, D., Ortiz, A., Verfaillie, J., Baldocchi, D., 2020. FLUXNET-CH4 US-Myb Mayberry Wetland. https://doi.org/10.18140/FLX/1669685

Matthes, J.H., Sturtevant, C., Verfaillie, J., Knox, S., Baldocchi, D., 2014. Parsing the variability in CH4 flux at a spatially heterogeneous wetland: Integrating multiple eddy covariance towers with high-resolution flux footprint analysis. J. Geophys. Res. Biogeosci. 119, 1322–1339. https://doi.org/10.1002/2014jg002642

McNicol, G., Knox, S.H., Guilderson, T.P., Baldocchi, D.D., Silver, W.L., 2020. Where old meets new: An ecosystem study of methanogenesis in a reflooded agricultural peatland. Glob. Chang. Biol. 26, 772–785. https://doi.org/10.1111/gcb.14916

McNicol, G., Sturtevant, C.S., Knox, S.H., Dronova, I., Baldocchi, D.D., Silver, W.L., 2017. Effects of seasonality, transport pathway, and spatial structure on greenhouse gas fluxes in a restored wetland. Glob. Chang. Biol. 23, 2768–2782. https://doi.org/10.1111/gcb.13580

Menzer, O., Moffat, A.M., Meiring, W., Lasslop, G., Schukat-Talamazzini, E.G., Reichstein, M., 2013. Random errors in carbon and water vapor fluxes assessed with Gaussian Processes. Agric. For. Meteorol. 178-179, 161–172. https://doi.org/10.1016/j.agrformet.2013.04.024

Miyata, A., Leuning, R., Denmead, O.T., Kim, J., Harazono, Y., 2000. Carbon dioxide and methane fluxes from an intermittently flooded paddy field. Agric. For. Meteorol. 102, 287–303. https://doi.org/10.1016/S0168-1923(00)00092-7

Moffat, A.M., Papale, D., Reichstein, M., Hollinger, D.Y., Richardson, A.D., Barr, A.G., Beckstein, C., Braswell, B.H., Churkina, G., Desai, A.R., Falge, E., Gove, J.H., Heimann, M., Hui, D., Jarvis, A.J., Kattge, J., Noormets, A., Stauch, V.J., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric. For. Meteorol. 147, 209–232. https://doi.org/10.1016/j.agrformet.2007.08.011

Moore, T.R., De Young, A., Bubier, J.L., Humphreys, E.R., Lafleur, P.M., Roulet, N.T., 2011. A multi-year record of methane flux at the Mer bleue bog, southern Canada. Ecosystems 14, 646–657. https://doi.org/10.1007/s10021-011-9435-9

Morin, T.H., Bohrer, G., Frasson, R.P. d. M., Naor-Azreli, L., Mesi, S., Stefanik, K.C., Schäfer, K.V.R., 2014. Environmental drivers of methane fluxes from an urban temperate wetland park. J. Geophys. Res. Biogeosci. 119, 2188–2208. https://doi.org/10.1002/2014jg002750

Morin, T.H., Bohrer, G., Stefanik, K.C., Rey-Sanchez, A.C., Matheny, A.M., Mitsch, W.J., 2017. Combining eddy-covariance and chamber measurements to determine the methane budget from a small, heterogeneous urban floodplain wetland park. Agric. For. Meteorol. 237-238, 160–170. https://doi.org/10.1016/j.agrformet.2017.01.022

Muramatsu, K., Ono, K., Soyama, N., Thanyapraneedkul, J., Miyata, A., Mano, M., 2017. Determination of rice paddy parameters in the global gross primary production capacity estimation algorithm using 6 years of JP-MSE flux observation data. Journal of Agricultural Meteorology 73, 119–132. https://doi.org/10.2480/agrmet.D-16-00017

Nemitz, E., Mammarella, I., Ibrom, A., Aurela, M., Burba, G.G., Dengel, S., Gielen, B., Grelle, A., Heinesch, B., Herbst, M., Hörtnagl, L., Klemedtsson, L., Lindroth, A., Lohila, A., McDermitt, D.K., Meier, P., Merbold, L., Nelson, D., Nicolini, G., Nilsson, M.B., Peltola, O., Rinne, J., Zahniser, M., 2018. Standardisation of eddy-covariance flux measurements of methane and nitrous oxide. Int. Agrophys 32, 517–549. https://doi.org/10.1515/intag-2017-0042

Neubauer, S.C., Megonigal, J.P., 2015. Moving beyond global warming potentials to quantify the climatic role of ecosystems. Ecosystems 18, 1000–1013. https://doi.org/10.1007/s10021-015-9879-4

Nilsson, M., Peichl, M., 2020. FLUXNET-CH4 SE-Deg Degero. https://doi.org/10.18140/FLX/1669659

Oikawa, P.Y., Sturtevant, C., Knox, S.H., Verfaillie, J., Huang, Y.W., Baldocchi, D.D., 2017. Revisiting the partitioning of net ecosystem exchange of CO2 into photosynthesis and respiration with simultaneous flux measurements of 13CO2 and CO2, soil respiration and a biophysical model, CANVEG. Agric. For. Meteorol. 234-235, 149–163. https://doi.org/10.1016/j.agrformet.2016.12.016

Ooba, M., Hirano, T., Mogami, J.-I., Hirata, R., Fujinuma, Y., 2006. Comparisons of gap-filling methods for carbon flux dataset: A combination of a genetic algorithm and an artificial neural network. Ecol. Modell. 198, 473–486. https://doi.org/10.1016/j.ecolmodel.2006.06.006

Papale, D., 2020. Ideas and perspectives: enhancing the impact of the FLUXNET network of eddy covariance sites. Biogeosciences 17, 5587–5598. https://doi.org/10.5194/bg-17-5587-2020

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B., Ammann, C., Arain, M.A., Ardö, J., Arkebauer, T., Arndt, S.K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L.B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T.A., Blanken, P.D., Bohrer, G., Boike, J., Bolstad, P.V., Bonal, D., Bonnefond, J.-M., Bowling, D.R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S.P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T.R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B.D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P.S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K.J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R.C., Delpierre, N., Desai, A.R., Di Bella, C.M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H.A.M., Eugster, W., Ewenz, C.M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A.H., Gough, C.M., Goulden, M.L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B.U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W.L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A.J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H.A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J.H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A.M.S., Merbold, L., Meyer, W., Meyers, T., Miller, S.D., Minerbi, S., Moderow, U., Monson, R.K., Montagnani, L., Moore, C.E., Moors, E., Moreaux, V., Moureaux, C., Munger, J.W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J.E., Ourcival, J.-M., Papuga, S.A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R.P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S.M., Rambal, S., Rannik, Ü., Raz-Yaseef, N., Reed, D., de Dios, V.R., Restrepo-Coupe, N., Reverter, B.R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S.R., Sánchez-Cañete, E.P., Sanchez-Mejia, Z.M., Schmid, H.P., Schmidt, M., Schneider, K., Schrader, F., Schroder, I., Scott, R.L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R.M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tiedemann, F., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J.P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., Papale, D., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci Data 7, 225. https://doi.org/10.1038/s41597-020-0534-3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Others, 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peltola, O., Vesala, T., Gao, Y., Räty, O., Alekseychik, P., Aurela, M., Chojnicki, B., Desai, A.R., Dolman, A.J., Euskirchen, E.S., Friborg, T., Göckede, M., Helbig, M., Humphreys, E., Jackson, R.B., Jocher, G., Joos, F., Klatt, J., Knox, S.H., Kowalska, N., Kutzbach, L., Lienert, S., Lohila, A., Mammarella, I., Nadeau, D.F., Nilsson, M.B., Oechel, W.C., Peichl, M., Pypker, T., Quinton, W., Rinne, J., Sachs, T., Samson, M., Schmid, H.P., Sonnentag, O., Wille, C., Zona, D., Aalto, T., 2019. Monthly gridded data product of northern wetland methane emissions based on upscaling eddy covariance observations. Earth Syst. Sci. Data 11, 1263–1289. https://doi.org/10.5194/essd-11-1263-2019

Platt, J.C., 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, in: Advances in Large Margin Classifiers.

Poffenbarger, H.J., Needelman, B.A., Megonigal, J.P., 2011. Salinity Influence on Methane Emissions from Tidal Marshes. Wetlands 31, 831–842. https://doi.org/10.1007/s13157-011-0197-0

Pohlert, T., 2014. The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR).

R Core Team, 2021. R: A Language and Environment for Statistical Computing.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grunwald, T., Havrankova, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., Valentini, R., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Glob. Chang. Biol. 11, 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x

Rey-Sanchez, A.C., Morin, T.H., Stefanik, K.C., Wrighton, K., Bohrer, G., 2018. Determining total emissions and environmental drivers of methane flux in a Lake Erie estuarine marsh. Ecol. Eng. 114, 7–15. https://doi.org/10.1016/j.ecoleng.2017.06.042

Richardson, A.D., Aubinet, M., Barr, A.G., Hollinger, D.Y., Ibrom, A., Lasslop, G., Reichstein, M., 2012. Uncertainty Quantification, in: Aubinet, M., Vesala, T., Papale, D. (Eds.), Eddy Covariance: A Practical Guide to Measurement and Data Analysis. Springer Netherlands, Dordrecht, pp. 173–209.

Richardson, A.D., Hollinger, D.Y., 2007. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO2 flux record. Agric. For. Meteorol. 147, 199–208. https://doi.org/10.1016/j.agrformet.2007.06.004

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881

Rojas, R., 2013. Neural Networks: A Systematic Introduction. Springer Science & Business Media.

Rosentreter, J.A., Borges, A.V., Deemer, B.R., Holgerson, M.A., Liu, S., Song, C., Melack, J., Raymond, P.A., Duarte, C.M., Allen, G.H., Olefeldt, D., Poulter, B., Battin, T.I., Eyre, B.D., 2021. Half of global methane emissions come from highly variable aquatic ecosystem sources. Nat. Geosci. 14, 225–230. https://doi.org/10.1038/s41561-021-00715-2

Runkle, B.R.K., Suvočarev, K., Reba, M.L., Reavis, C.W., Smith, S.F., Chiu, Y.-L., Fong, B., 2019. Methane Emission Reductions from the Alternate Wetting and Drying of Rice Fields Detected Using the Eddy Covariance Method. Environ. Sci. Technol. 53, 671–681. https://doi.org/10.1021/acs.est.8b05535

Russell, S.J., Norvig, P., 1995. Artificial Intelligence: A Modern Approach. Prentice Hall.

Saunois, M., Stavert, A.R., Poulter, B., Bousquet, P., Canadell, J.G., Jackson, R.B., Raymond, P.A., Dlugokencky, E.J., Houweling, S., Patra, P.K., Ciais, P., Arora, V.K., Bastviken, D., Bergamaschi, P., Blake, D.R., Brailsford, G., Bruhwiler, L., Carlson, K.M., Carrol, M., Castaldi, S., Chandra, N., Crevoisier, C., Crill, P.M., Covey, K., Curry, C.L., Etiope, G., Frankenberg, C., Gedney, N., Hegglin, M.I., Höglund-Isaksson, L., Hugelius, G., Ishizawa, M., Ito, A., Janssens-Maenhout, G., Jensen, K.M., Joos, F., Kleinen, T., Krummel, P.B., Langenfelds, R.L., Laruelle, G.G., Liu, L., Machida, T., Maksyutov, S., McDonald, K.C., McNorton, J., Miller, P.A., Melton, J.R., Morino, I., Müller, J., Murguia-Flores, F., Naik, V., Niwa, Y., Noce, S., O'Doherty, S., Parker, R.J., Peng, C., Peng, S., Peters, G.P., Prigent, C., Prinn, R., Ramonet, M., Regnier, P., Riley, W.J., Rosentreter, J.A., Segers, A., Simpson, I.J., Shi, H., Smith, S.J., Steele, L.P., Thornton, B.F., Tian, H., Tohjima, Y., Tubiello, F.N., Tsuruta, A., Viovy, N., Voulgarakis, A., Weber, T.S., van Weele, M., van der Werf, G.R., Weiss, R.F., Worthy, D., Wunch, D., Yin, Y., Yoshida, Y., Zhang, W., Zhang, Z., Zhao, Y., Zheng, B., Zhu, Q., Zhu, Q., Zhuang, Q., 2020. The global methane budget 2000–

2017. Earth Syst. Sci. Data 12, 1561–1623. https://doi.org/10.5194/essd-12-1561-2020

Schuurmans, E.D., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7, 1–30.

Sonnentag, O., Helbig, M., 2020. FLUXNET-CH4 CA-SCB Scotty Creek Bog. https://doi.org/10.18140/FLX/1669613

Sturtevant, C., Ruddell, B.L., Knox, S.H., Verfaillie, J., Matthes, J.H., Oikawa, P.Y., Baldocchi, D., 2016. Identifying scale-emergent, nonlinear, asynchronous processes of wetland methane exchange. J. Geophys. Res. Biogeosci. 121, 188–204. https://doi.org/10.1002/2015jg003054

Taoka, T., Iwata, H., Hirata, R., Takahashi, Y., Miyabara, Y., Itoh, M., 2020. Environmental controls of diffusive and ebullitive methane emissions at a subdaily time scale in the littoral zone of a midlatitude shallow lake. J. Geophys. Res. Biogeosci. 125. https://doi.org/10.1029/2020jg005753

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res., WMO TD-732 106, 7183–7192. https://doi.org/10.1029/2000jd900719

Taylor, R., 1990. Interpretation of the Correlation Coefficient: A Basic Review. J. Diagn. Med. Sonogr. 6, 35–39. https://doi.org/10.1177/875647939000600106

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Series B Stat. Methodol. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tramontana, G., Jung, M., Schwalm, C.R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M.A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., Papale, D., 2016. Predicting carbon dioxide and energy fluxes across global

FLUXNET sites with regression algorithms. Biogeosciences 13, 4291–4313. https://doi.org/10.5194/bg-13-4291-2016

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T.F., Camps-Valls, G., Ogee, J., Verrelst, J., Papale, D., 2020. Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. Glob. Chang. Biol. 26, 5235–5253. https://doi.org/10.1111/gcb.15203

Treat, C.C., Bloom, A.A., Marushchak, M.E., 2018. Nongrowing season methane emissions-a significant component of annual emissions across northern ecosystems. Glob. Chang. Biol. 24, 3331–3343. https://doi.org/10.1111/gcb.14137

Trifunovic, B., Vázquez-Lule, A., Capooci, M., Seyfferth, A.L., Moffat, C., Vargas, R., 2020. Carbon dioxide and methane emissions from temperate salt marsh tidal creek. J. Geophys. Res. Biogeosci., NOAA National Estuarine Research Reserve, Central Data Management Office, Baruch Marine Laboratory, University of South Carolina 125, 84. https://doi.org/10.1029/2019jg005558

Tuovinen, J.-P., Aurela, M., Hatakka, J., Räsänen, A., Virtanen, T., Mikola, J., Ivakhov, V., Kondratyev, V., Laurila, T., 2019. Interpreting eddy covariance data from heterogeneous Siberian tundra: land-cover-specific methane fluxes and spatial representativeness. Biogeosciences 16, 255–274. https://doi.org/10.5194/bg-16-255-2019

Turetsky, M.R., Kotowska, A., Bubier, J., Dise, N.B., Crill, P., Hornibrook, E.R.C., Minkkinen, K., Moore, T.R., Myers-Smith, I.H., Nykänen, H., Olefeldt, D., Rinne, J., Saarnio, S., Shurpali, N., Tuittila, E.-S., Waddington, J.M., White, J.R., Wickland, K.P., Wilmking, M., 2014. A synthesis of methane emissions from 71 northern, temperate, and subtropical wetlands. Glob. Chang. Biol. 20, 2183–2197. https://doi.org/10.1111/gcb.12580

Ueyama, M., Hirano, T., Kominami, Y., 2020a. FLUXNET-CH4 JP-BBY Bibai bog. https://doi.org/10.18140/FLX/1669646

Ueyama, M., Yazaki, T., Hirano, T., Futakuchi, Y., Okamura, M., 2020b. Environmental controls on methane fluxes in a cool temperate bog. Agric. For. Meteorol. 281, 107852. https://doi.org/10.1016/j.agrformet.2019.107852

Valach, A., Szutu, D., Eichelmann, E., Knox, S., Verfaillie, J., Baldocchi, D., 2020. FLUXNET-CH4 US-Tw1 Twitchell Wetland West Pond. https://doi.org/10.18140/FLX/1669696

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform.

Vargas, R., Sánchez-Cañete P., E., Serrano-Ortiz, P., Curiel Yuste, J., Domingo, F., López-Ballesteros, A., Oyonarte, C., 2018. Hot-Moments of Soil CO2 Efflux in a Water-Limited Grassland. Soil Systems 2, 47. https://doi.org/10.3390/soilsystems2030047

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł.U., Polosukhin, I., 2017. Attention is All you Need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5998–6008.

Vázquez-Lule, A., Vargas, R., 2021. Biophysical drivers of net ecosystem and methane exchange across phenological phases in a tidal salt marsh. Agric. For. Meteorol. 300, 108309. https://doi.org/10.1016/j.agrformet.2020.108309

Vesala, T., Tuittila, E.-S., Mammarella, I., Alekseychik, P., 2020a. FLUXNET-CH4 FI-Si2 Siikaneva-2 Bog. https://doi.org/10.18140/FLX/1669639

Vesala, T., Tuittila, E.-S., Mammarella, I., Rinne, J., 2020b. FLUXNET-CH4 FI-Sii Siikaneva. https://doi.org/10.18140/FLX/1669640

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vitale, D., Bilancia, M., Papale, D., 2019. Modelling random uncertainty of eddy covariance flux measurements. Stoch. Environ. Res. Risk Assess. 33, 725–746. https://doi.org/10.1007/s00477-019-01664-4

Vitale, D., Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF), University of Tuscia, via San Camillo de Lellis, 01100 Viterbo, Italy, Bilancia, M., Papale, D., Ionian Department of Law, Economics and Environment, University of Bari Aldo Moro, Via Lago Maggiore angolo Via Ancona, 74121 Taranto, Italy, Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF), University of Tuscia, via San Camillo de Lellis, 01100 Viterbo, Italy, 2018. A Multiple Imputation Strategy for Eddy Covariance Data. J. Environ. Inf. 1–20. https://doi.org/10.3808/jei.201800391

Vourlitis, G., Dalmagro, H., de S. Nogueira, J., Johnson, M., Arruda, P., 2020. FLUXNET-CH4 BR-Npw Northern Pantanal Wetland. https://doi.org/10.18140/FLX/1669368

Vuichard, N., Papale, D., 2015. Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis. Earth Syst. Sci. Data 7, 157–171. https://doi.org/10.5194/essd-7-157-2015

Wania, R., Melton, J.R., Hodson, E.L., Poulter, B., Ringeval, B., Spahni, R., Bohn, T., Avis, C.A., Chen, G., Eliseev, A.V., Hopcroft, P.O., Riley, W.J., Subin, Z.M., Tian, H., van Bodegom, P.M., Kleinen, T., Yu, Z.C., Singarayer, J.S., Zürcher, S., Lettenmaier, D.P., Beerling, D.J., Denisov, S.N., Prigent, C., Papa, F., Kaplan, J.O., 2013. Present state of global wetland extent and wetland methane modelling: methodology of a model inter-comparison project (WETCHIMP). Geosci. Model Dev. 6, 617–641. https://doi.org/10.5194/gmd-6-617-2013

Whiting, G.J., Chanton, J.P., 1993. Primary production control of methane emission from wetlands. Nature 364, 794–795. https://doi.org/10.1038/364794a0

Wutzler, T., Lucas-Moffat, A., Migliavacca, M., Knauer, J., Sickel, K., Šigut, L., Menzer, O., Reichstein, M., 2018. Basic and extensible post-processing of eddy covariance flux data with REddyProc. Biogeosciences 15, 5015–5030. https://doi.org/10.5194/bg-15-5015-2018

Yang, W.H., McNicol, G., Teh, Y.A., Estera-Molina, K., Wood, T.E., Silver, W.L., 2017. Evaluating the classical versus an emerging conceptual model of peatland methane dynamics: Peatland methane dynamics. Global Biogeochem. Cycles 31, 1435–1453. https://doi.org/10.1002/2017gb005622

Yvon-Durocher, G., Allen, A.P., Bastviken, D., Conrad, R., Gudasz, C., St-Pierre, A., Thanh-Duc, N., del Giorgio, P.A., 2014. Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. Nature 507, 488–491. https://doi.org/10.1038/nature13164

Zadrozny, B., Elkan, C., 2002. Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02. Association for Computing Machinery, New York, NY, USA, pp. 694–699.