



HAL
open science

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

Jeremy Irvin, Sharon Zhou, Gavin McNicol, Fred Lu, Vincent Liu, Etienne Fluet-Chouinard, Zutao Ouyang, Sara Helen Knox, Antje Lucas-Moffat, Carlo Trotta, et al.

► To cite this version:

Jeremy Irvin, Sharon Zhou, Gavin McNicol, Fred Lu, Vincent Liu, et al.. Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH₄ wetlands. *Agricultural and Forest Meteorology*, 2021, 308-309, pp.108528. <10.1016/j.agrformet.2021.108528>. <insu-03286676>

HAL Id: insu-03286676

<https://insu.hal.science/insu-03286676v1>

Submitted on 27 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

1 **Title**

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

4

5

6 **Corresponding Author:**

7 Gavin McNicol

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

9 Email: gmcnicol@uic.edu

10

11 **Keywords**

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

13

14 **Highlights**

15 • We evaluate methane flux gap-filling methods across 17 boreal-to-tropical wetlands

16 • New methods for generating realistic artificial gaps and uncertainties are proposed

17 • Decision tree algorithms perform slightly better than neural networks on average

18 • Soil temperature and generic seasonality are the most important predictors

19 • Open-source code is released for gap-filling steps and uncertainty evaluation

20

21 Graphical Abstract

AI for Methane Flux Gap-Filling



22

23

24

25

26

27

28

29

30

31 Abstract

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

56

57

58 Main Text

59 1 Introduction

60 Globally, wetlands emit 102-200 teragrams (Tg) of the greenhouse gas methane (CH₄) to the
61 atmosphere and the scarcity of wetland CH₄ flux data has hindered efforts to better constrain
62 emission uncertainties ([Saunois et al. 2020](#)). Eddy covariance-based measurements of CH₄
63 fluxes have increased rapidly over the last two decades, leading to the release of the first global
64 compilation of CH₄ flux data from 81 sites in 2020 (FLUXNET-CH4 community product Version
65 1.0; [Knox et al. 2019](#); [Delwiche et al. 2021](#)). The growth in available CH₄ data can help improve
66 bottom-up estimates of regional-to-global wetland CH₄ sources ([Treat et al. 2018](#); [Peltola et al.](#)
67 [2019](#); [Rosentreter et al. 2021](#)) but this requires data processing standards that ensure eddy
68 covariance CH₄ flux data products are of the same quality and provenance as carbon dioxide
69 (CO₂) and energy fluxes (e.g., FLUXNET2015; [Pastorello et al. 2020](#)). Gap-filling is a
70 particularly important step during data processing as it impacts estimates of ecosystem carbon
71 balance and net ecosystem radiative forcing at individual sites, due to the potency of CH₄ as a
72 greenhouse gas ([Neubauer and Megonigal 2015](#); [Hemes et al. 2019](#); [Günther et al. 2020](#)), and
73 can alter upscaled predictions in data driven CH₄ flux models ([Turetsky et al. 2014](#); [Treat et al.](#)
74 [2018](#); [Peltola et al. 2019](#)). Comprehensive evaluations of gap-filling methods for CH₄ fluxes
75 across many wetland sites are still lacking and needed in order to advance existing methods
76 ([Nemitz et al. 2018](#); [Mammarella et al. 2020](#)).

77
78 Gaps of various lengths arise in time series of eddy covariance CH₄ fluxes because of system
79 failure (including signal degradation due to sensor soiling), insufficient turbulent mixing, extreme
80 weather conditions, irregular maintenance, and wind direction filtering, among other reasons.
81 Technical challenges remain in precise and accurate measurement of eddy covariance CH₄
82 fluxes ([Morin 2019](#); [Knox et al. 2019](#)) despite recent technological advances in spectra-based
83 gas analyzers ([Nemitz et al. 2018](#)). After filtering, annual data coverage can be low for CH₄ (25-
84 40%; [Delwiche et al. 2021](#)). Therefore gap-filling procedures are required to construct the
85 continuous time series for quantifying continuous daily, seasonally, and annually integrated CH₄
86 emission estimates. Gap-filling techniques used to impute half-hourly eddy covariance fluxes at
87 individual sites include look-up tables ([Reichstein et al. 2005](#)), machine learning and genetic
88 algorithms ([Ooba et al. 2006](#); [Moffat et al. 2007](#); [Kim et al. 2019](#)), multiple imputation ([Hui et al.](#)
89 [2004](#); [Vitale et al. 2018](#)), and process models ([Oikawa et al. 2017](#)). Any bias tied to a given

90 method propagates to seasonal and annual CH₄ emissions and can therefore impact CH₄
 91 emission estimates at regional to global scales ([Falge et al. 2001](#); [Moffat et al. 2007](#); [Peltola et](#)
 92 [al. 2019](#); [Vitale et al. 2019](#)).

93
 94 Marginal distribution sampling (MDS) ([Reichstein et al. 2005](#); [Moffat et al. 2007](#); [Pastorello et al.](#)
 95 [2020](#)) and machine-learning (ML) have become the standard gap-filling methods for CO₂ fluxes
 96 in the eddy covariance community ([Wutzler et al. 2018](#)), while no similar standard has yet been
 97 established for CH₄ fluxes. MDS is a multi-step sampling scheme, akin to a complex decision
 98 tree, and uses look-up tables to identify similar predictor conditions within a given time window,
 99 which conservatively expands around the gap, only as is necessary. MDS is an efficient gap-
 100 filling method that supplements the look-up tables with diurnal cycle interpolation, allowing it to
 101 function when there are gaps in predictors. However, MDS performance can be limited by the
 102 number of permissible predictors and current predictor choices are optimized for CO₂, not CH₄
 103 fluxes ([Falge et al. 2001](#)). Moreover, unlike CO₂ fluxes, CH₄ fluxes at many sites appear to lack
 104 a consistent diel cycle and display different diel patterns ([Bansal et al. 2018](#)). In contrast, ML is
 105 well suited to high-dimensional datasets and can capture nonlinear relationships between
 106 predictors and fluxes ([Tramontana et al. 2016](#); [Bodesheim et al. 2018](#)) albeit they generally
 107 need more time to train and evaluate. A summary of some of the methodological considerations
 108 for MDS and four different ML algorithms considered in this study are shown in **Table 1**.

109
 110

111 **Table 1** An overview of marginal distribution sampling and potential machine learning
 112 algorithms for gap-filling of CH₄ flux in wetlands.
 113

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)

114

115 To date, artificial neural networks (ANN) have been found to be effective for gap-filling CH₄
116 fluxes across six high-latitude wetlands [\(Dengel et al. 2013\)](#). ANN have since been used across
117 a variety of eddy covariance sites at natural, rewetted, and urban wetlands [\(Morin et al. 2014;](#)
118 [Goodrich et al. 2015; Rey-Sanchez et al. 2018; Hemes et al. 2019; Li et al. 2020; Koebisch et al.](#)
119 [2020\)](#), tidal salt marshes [\(Vázquez-Lule and Vargas 2021\)](#), and rice paddies [\(Knox et al. 2016;](#)
120 [Runkle et al. 2019\)](#), as well as in a FLUXNET-CH₄ synthesis and the FLUXNET-CH₄ community
121 product Version 1.0 [\(Knox et al. 2019; Delwiche et al. 2021\)](#). However, the ANN algorithms
122 developed by Dengel et al. (2013) and Moffat et al. (2007) were only inter-compared in detail
123 among six high-latitude sites and were only evaluated on single site-growing-seasons of data.
124 More recently, random forests (RF) were found to match or outperform both MDS and ANN at
125 five wetlands and rice paddies, with strengths in predicting interannual variability from a single
126 multi-year model [\(Kim et al. 2019\)](#). Overall, although some important insights into CH₄ gap-filling
127 strategies with ML have been made at individual, or small sets of sites, comprehensive
128 experiments are still needed to identify the best approaches across the global distribution of
129 wetlands.

130

131 In addition to algorithm choice, investigators need to consider the causes of spatial and
132 temporal variability and the effects of biases between training and test data. The complexity of
133 wetland CH₄ production, consumption, and transport processes can lead to high temporal and
134 spatial variability in fluxes across flux tower footprints. Relationships between biophysical
135 drivers and CH₄ flux can be nonlinear and obscured by lags and asynchronicity [\(Sturtevant et al.](#)

136 [2016](#)). Additionally, the temporal signals in CH₄ flux time series are observed across a broad
137 range of hourly, multi-day, and seasonal timescales ([Knox et al. 2019](#); [Knox et al. 2021](#)), and
138 can lack a clear diel cycle as observed for CO₂ ([Moffat et al. 2007](#)). Challenges also arise for
139 standardization due to site uniqueness ([Bridgman et al. 2013](#); [Trifunovic et al. 2020](#)). For
140 example, [Knox et al. \(2019\)](#) showed that variation in water table depth, a well-established
141 control on wetland CH₄ fluxes, only measurably affected CH₄ flux at sites where its range
142 extended across the soil surface. Similarly, the spatial mosaic of inundation and vegetation
143 varies both within and across wetland classes and affects wetland CH₄ flux via substrate supply
144 and gas transport processes ([Matthes et al. 2014](#); [McNicol et al. 2017](#); [Rey-Sanchez et al.
145 2018](#)). This high spatial heterogeneity creates a wind direction (footprint) dependency rarely
146 observed for CO₂ fluxes ([Tuovinen et al. 2019](#)). To be able to explain the complex dynamics of
147 wetland CH₄ emissions, models need information on water table position, soil oxygen and
148 moisture, and soil temperature ([Bridgman et al. 2013](#)). Other issues include biases in training
149 observations introduced by low turbulence (friction velocity, USTAR) filters ([Göckede et al.
150 2019](#)) which might make gap-filling models more prone to errors during imputation of CH₄ flux
151 from higher-to-lower turbulence conditions ([Dengel et al. 2013](#)), as is observed at some sites for
152 daytime-to-nighttime imputation of CO₂ flux ([Moffat et al. 2007](#)). Conditions that lead to
153 exceptional but short-lived fluxes (e.g., ebullition events) may also be less easy to capture in
154 training and test data ([Ueyama et al. 2020b](#); [Taoka et al. 2020](#)). In sum, the combination of high
155 temporal variability of CH₄ flux within and across sites ([Knox et al. 2019](#)), high spatial variation
156 of fluxes in some wetlands ([Morin et al. 2017](#)), and the sensitivity of fluxes to a suite of drivers at
157 different timescales ([Sturtevant et al. 2016](#)), requires a thorough evaluation of CH₄ flux gap-
158 filling models across a broad range of possible gap lengths.

159
160 This study provides a systematic evaluation of MDS and four ML algorithms for gap-filling CH₄
161 fluxes at 17 FLUXNET-CH₄ sites. The 17 sites cover a wide range of wetland types, and climate
162 and gap conditions (i.e., length and distribution). Collectively, these sites provide a large and
163 fairly standard set of predictors, allowing for a robust across-site comparison of model
164 performance and predictor importance. The overall ML workflow from artificial gap generation,
165 to cross validation and testing, and to prediction uncertainty estimation, is robust and
166 reproducible ([Pastorello et al. 2020](#); [Nemitz et al. 2018](#)) and designed to be general and
167 applicable to a wide range of gap-filling scenarios across terrestrial wetland ecosystems. The
168 data and code are made public [<https://github.com/stanfordmlgroup/methane-gapfill-ml>].

169 2 Materials and Methods

170 2.1 Site Data

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

178
179 **Table 2: Site information and data references for 17 wetland FLUXNET-CH₄ sites.** Sites
180 are arranged in order of increasing mean of observed CH₄ flux (which is also sensitive to
181 differences in temporal coverage between sites) and days refers to the number of days with
182 some observed CH₄ fluxes. Data are the same as those published in the FLUXNET-CH₄
183 community product Version 1.0 (<https://fluxnet.org/data/fluxnet-ch4-community-product/>)
184 ([Delwiche et al. 2021](#)). Mean annual temperature and precipitation were extracted from
185 respective WorldClim 2.0 gridded products at site locations ([Fick and Hijmans 2017](#)).

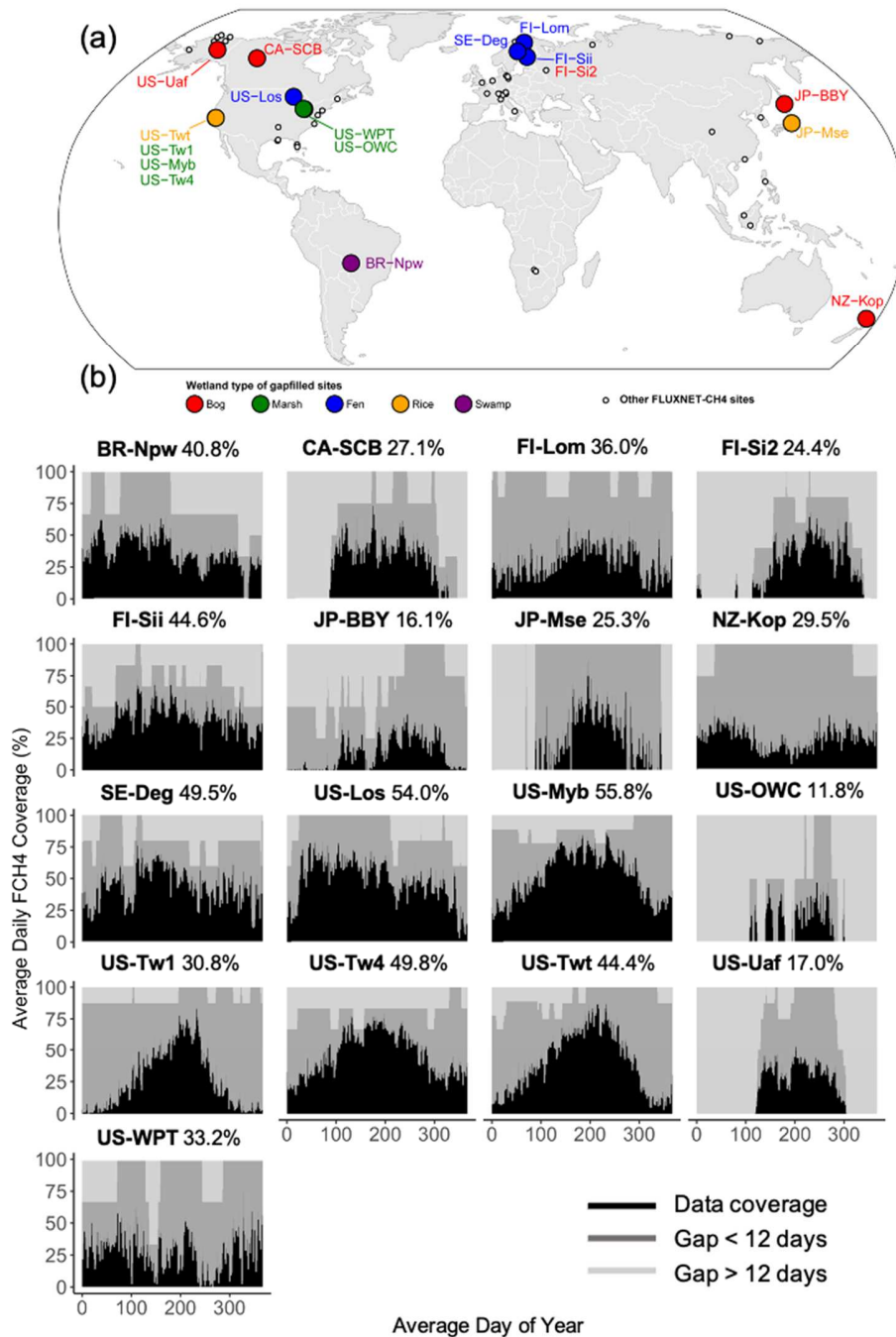
186

Site ID	Climate Zone	Mean Annual Temp. °C	Mean Annual Precip. mm	Mean FCH ₄ , 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	(Sonntag 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)

187

188 The 17 sites span tropical to boreal climates and diverse and representative wetland types
189 (**Figure 1**), including bogs (5), marshes (5), fens (4), a tropical swamp (1), and rice paddies (2).
190 Altogether, 32.4 site-years of CH₄ flux data were used for gap-filling model development and
191 validation, collected during 2010-2018. Data pre-processing steps prior to gap-filling were the
192 same as described in [\(Delwiche et al. 2021\)](#). Each site was classified into a wetland class based
193 on site investigator self-reporting.



194
 195
 196
 197
 198
 199

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

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

202 2.2 Predictor Variables

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

213

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

221

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

224

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)

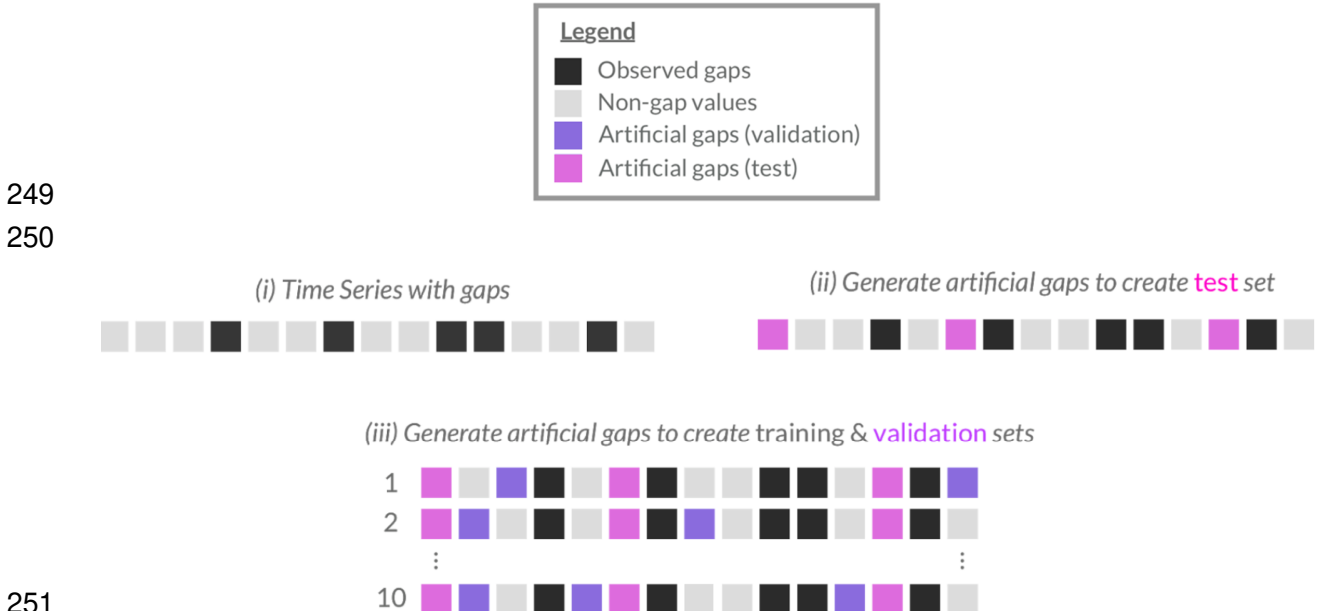
*Both conventional nighttime temperature extrapolation method [\(Reichstein et al. 2005\)](#) and more recent daytime method [\(Lasslop et al. 2010\)](#) variables were included.

225

226 2.3 Machine Learning Model Training Procedure

227 Four ML algorithms were trained with each of the four subsets of input predictors (**Table 3**),
 228 leading to a total of 16 algorithm-predictor combinations per site, which were evaluated using a
 229 nested cross validation procedure (**Figure 2**). In each algorithm-by-predictor set experiment, the
 230 following steps were repeated for each site. Firstly, artificial gaps were introduced which
 231 constituted a single, held-out test set. The test set was only used after model training and
 232 selection to evaluate the gap-filling performance of the selected models. Secondly, following
 233 [Moffat et al. \(2007\)](#), 10 additional pairs of training and validation sets of artificial gaps were
 234 created with several independent samples of artificial gaps to mitigate potential bias in model
 235 performance for any particular gap sequence. Thirdly, for each algorithm-by-predictor
 236 combination, a model was trained on each of the 10 training sets and the best ML

237 hyperparameters were selected based on average model performance during 5-fold cross-
 238 validation. Cross-validation involved creating 5 random subsets (folds) of each training set,
 239 training the model multiple times with a broad hyperparameter grid search on 4 folds, and
 240 evaluating the models on one held-out fold. This hyperparameter search was repeated 5 times,
 241 changing the held-out fold each time. The best hyperparameters across all folds were then used
 242 to refit the model on the full training set, resulting in 10 trained models for each algorithm-by-
 243 predictor combination. Fourthly, each of the 10 models was evaluated using the corresponding
 244 validation set, and the mean and variance of model scores for the 10 validation sets were used
 245 to compare algorithm classes with different input predictor groups. Finally, the 10 models of the
 246 algorithm classes that scored highest on the validation sets were ensembled and the ensemble
 247 mean prediction was evaluated against the test set.
 248



251

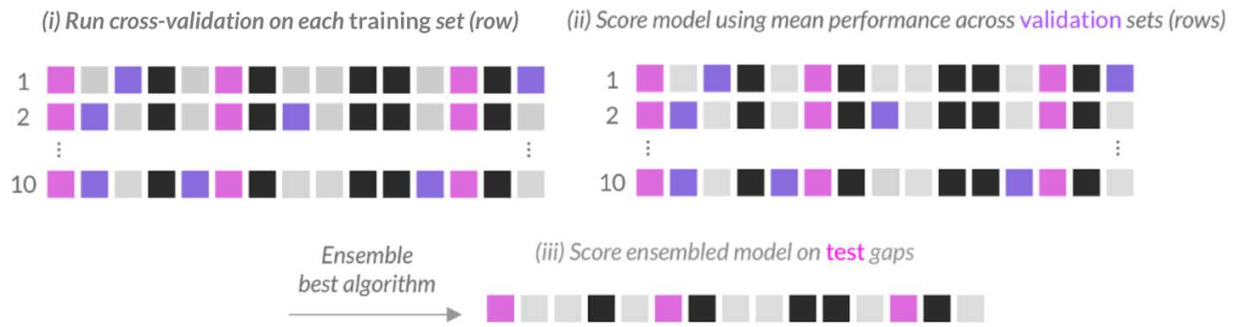
252 **(a) Creating data (training, validation, and test) splits with artificial gaps**

253

254

255

256



257

258

259

(b) Model development and validation procedure

260

261

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.

269

2.4 Gap-filling Methods

270

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.

277

2.4.1 MDS

278

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

281

282 each gap, the missing value is replaced with the average of the measurements found in the time
 283 window around the gap with similar meteorological conditions (i.e., similar value of the drivers).
 284 The algorithm first tries to use all three drivers for a window which is kept as short as possible to
 285 avoid the confounding effects of other slow-changing drivers such as phenology. If no similar
 286 conditions are found, the window size is increased and only the main driver is considered, or
 287 alternatively, and as a last option, the mean diurnal cycle within adjacent days is used. More
 288 details on the overall strategy and compromise between having a larger window or only one
 289 driver included can be found in the appendix of [\(Reichstein et al. 2005\)](#). The original method,
 290 designed for CO₂ fluxes, uses SW_IN as the main driver, and TA, and VPD as additional
 291 drivers. In the current application of the method to wetland CH₄ fluxes, however, seven different
 292 driver combinations were tested as reported in **Table 4**.

293

294 **Table 4: Driver combinations used for the MDS method.** SW_IN = Incoming
 295 shortwave radiation (W m⁻²), TA = air temperature (°C), PA = air pressure (hP), WTD =
 296 water table depth (m), WS = wind speed (m s⁻¹), RECO = ecosystem respiration (μmol
 297 CO₂ m⁻² s⁻¹). The values in parenthesis are the thresholds used to define similar
 298 conditions (i.e., value ± threshold). In case of SW_IN, as in the original formulation of the
 299 method in [\(Reichstein et al. 2005\)](#), the thresholds are two (20, 50): similar conditions for
 300 a measured value V are considered in the range V ± 50 if V > 50, V ± 20 if V < 20 and V
 301 ± V for values of V between 20 and 50.

302

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)

303

304

305 2.4.2 ML Algorithms

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

313
314 Artificial neural networks (ANN, i.e., shallow multilayer perceptrons) were tested and have been
315 used in previous works for CO₂ and CH₄ fluxes ([Goodrich et al. 2015](#); [Dengel et al. 2013](#); [Knox
316 et al. 2016](#); [Hemes et al. 2019](#); [Li et al. 2020](#)). Neural networks consist of a few layers, with
317 each layer containing different numbers of nodes that sequentially apply linear transformations
318 with parameters that are learned during model training. These layers are separated by nonlinear
319 activation functions that enable the neural network to model more complex functions. During
320 training, the parameters of each layer's transformation were adjusted to minimize the squared
321 loss between the predicted and observed flux values. Hyperparameters tuned during cross
322 validation included the optimization method for adjusting parameters (LBFGS or Adam),
323 learning rate (0.01, 0.001, 0.0001), the nonlinear activation function (hyperbolic tangent or
324 rectified linear unit), the numbers of hidden layers (1 or 2; [Knox et al. 2019](#)), and the number of
325 nodes per layer (5-30). Normalization was the same as Lasso.

326
327 Random forests have been commonly used to model tabular data and have recently emerged
328 for gap-filling CH₄ fluxes ([Kim et al. 2019](#)). Random forests are an ensemble of decision trees
329 which are each learned independently on bootstrapped data ([Breiman 2001](#)). The mean of the
330 predictions across the ensemble of trees is taken as the final prediction. Hyperparameters tuned
331 during cross validation included the number of trees (50-500), the maximum depth per tree (10-
332 110, as well as no maximum depth), the number of predictors considered at each split (n or
333 square-root of n), the minimum number of samples required to split a node (2, 5, or 10), the
334 minimum number of samples required at each leaf node (1, 2, or 4), and whether to bootstrap
335 the data when building trees. Normalization is not required for RF. Predictor importance was
336 computed as reduction in Gini impurity ([Breiman 2001](#)).

337

338 Boosting enables decision trees to be grown iteratively based on the mistakes of prior trees
339 ([Freund and Schapire 1999](#)). XGBoost was tested as a widely used and efficient gradient
340 boosted decision tree framework that builds decision trees sequentially ([Chen and Guestrin
341 2016](#)) and has demonstrated success in a wide variety of ML applications. A squared loss was
342 used as the objective function with the default learning rate of 0.1. The number of decision
343 trees, the maximum depth per tree, and the minimum number of samples required to split a
344 node used the same ranges as RF. Other hyperparameters tuned included the proportion of the
345 training data to subsample prior to growing trees (0.75, 0.85, or 0.95), the minimum loss
346 reduction required to split a leaf node (0, 0.2, or 0.4), and the fraction of predictors that were
347 randomly selected for the construction of each tree (0.6, 0.7, 0.8, or 0.9). XGBoost handles
348 predictor imputation during training using sparsity-aware split finding, which provides a default
349 direction on each node in the decision tree and allows for skipping over missing values ([Chen
350 and Guestrin 2016](#)). Normalization is not required for XGBoost.

351 2.5 Artificial Gap Generation

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

364

365 To retain the observed gap length distribution, a new artificial gap generation procedure was
366 developed. The new procedure takes into account the locations of the observed gaps when
367 generating artificial gaps of varying lengths, such that the observed plus artificial gap length
368 distribution resembles the observed distribution. Formally, the artificial gap generation
369 procedure finds a distribution q of artificial gap lengths for each site such that the true empirical
370 distribution p of gap lengths at that site is approximated by the union of q and p , which is

371 denoted $r = q \cup p$. In order to obtain a distribution r which is close to p , a method is proposed
372 for finding q . Intuitively, the histogram of q should look “compressed” compared to the histogram
373 of p ; that is, it places more weight on shorter gap lengths and has lighter tails: while shorter gap
374 lengths will be sampled more from q , longer gaps will still form from the merging that occurs
375 between newly sampled and observed existing gaps. A detailed description and
376 parameterization of the artificial gap generation algorithm are provided in **Appendix B**.

377

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

384 2.6 Evaluation

385 For each site, MDS-and the ML algorithm-predictor combinations were compared by evaluating
386 predictive performance on the 10 validation sets. The best two algorithms and their ensemble
387 performance were then evaluated on the test set using both baseline and all predictors to: 1)
388 measure absolute improvements over previously implemented standards (ANN plus baseline
389 predictors; [Knox et al. 2019](#)); 2) understand how each algorithm benefited (if at all) from using
390 all, rather than only baseline, predictors; and 3) measure the effect that the different algorithm
391 predictions had on cumulative annual and growing season CH₄ emissions estimates for each
392 site, and associated uncertainties.

393 2.6.1 Performance Measures

394 Model performance was measured using the coefficient of determination (R^2), mean absolute
395 error normalized by the standard deviation of CH₄ flux (nMAE), mean bias (Bias), root mean
396 squared error (RMSE), and standard deviation. R^2 was used to measure the ability of the gap-
397 filling model to reproduce the time series pattern, after confirming that Pearson correlations
398 were all positive ([Taylor 1990](#)). nMAE was used to measure the difference between predictions
399 from observations regardless of the direction of the error; the normalization allows us to
400 compare across sites despite large differences in flux variability. Finally, Bias was used to
401 measure the average direction of error, which will have the largest consequence on site

402 emission sums. The nonparametric basic bootstrap with 5,000 bootstrap replicates was used to
403 compute variability around the performance metrics on the test set ([Efron and Tibshirani 1994](#));
404 and 95% confidence intervals for each measure were reported. Taylor diagrams were used to
405 visually compare the performance of each of the models with different input predictors. Taylor
406 diagrams provide a visually intuitive way of displaying the performance of each model in terms
407 of three metrics: R^2 , root mean squared error (RMSE), and standard deviation (Taylor, 2001).
408 Finally, nMAE and Bias were used to assess the performance of the models across different
409 gap lengths similar to [Moffat et al. \(2007\)](#), [Nemitz et al. \(2018\)](#), [Kim et al. \(2019\)](#), and [Knox et al.
410 \(2019\)](#): very short gaps (1 half hour), short gaps (2-8 half hours), medium gaps (9-64 half hours,
411 i.e., 1.5 days), long gaps (1.5-12 consecutive days), and extremely long gaps (> 12 consecutive
412 days).

413 2.6.2 Statistical Analysis

414 Validation set performance was evaluated coarsely using differences in median model metrics
415 and was only used to select models for the more detailed statistical comparison on the test set.
416 Then, for each site, the test set performance of the best two algorithms was compared (RF, as
417 the faster of the two decision tree algorithms, and ANN) with two predictor sets (baseline and
418 all). The performance metrics showed significant non-normality across the 17 sites according to
419 the Shapiro-Wilk test. As a result, the Friedman test followed by post hoc Nemenyi was used for
420 evaluating pairwise comparisons. This pair of tests is the nonparametric equivalent of the one-
421 way ANOVA with repeated measures (followed by Tukey's test) and is the standard procedure
422 when the assumptions of ANOVA are not met (normality in this case; [Derrac et al. 2011](#);
423 [Schuurmans 2006](#)). Performance metric comparisons were implemented in R ([R Core Team
424 2019](#)) using the PMCMR package ([Pohlert 2014](#)).

425
426 To evaluate whether the gap-filling performance is related to the characteristics of CH_4 flux,
427 Pearson correlation coefficient between the best model performance metrics (RF and all
428 predictors) and the annual mean and variance of the fluxes were analyzed. Correlation analyses
429 were performed in Python using the 'scipy' package ([Virtanen et al. 2020](#)).

430 2.6.3 Evaluating Systematic USTAR Bias

431 Filtering to remove eddy covariance CH_4 fluxes during low turbulence conditions (using friction
432 velocity, USTAR, as a measure of turbulence) may introduce a systematic bias into ML training
433 because the efficiency of CH_4 gas transport mechanisms such as plant mediated flow can

434 increase with wind speed ([Laanbroek 2010](#)). To approximate an evaluation of biases introduced
435 from low USTAR filtering, the amount of filtered data across each site was quantified (0-21%)
436 and the same fraction of high USTAR conditions (top percentile) was removed from each paired
437 training and validation set. The original and high-USTAR-filtered model performance was then
438 evaluated on the scorable gaps created with the high USTAR filter. Although an imperfect
439 analogue, this test therefore simulated model extrapolation to very low USTAR conditions by
440 evaluating performance during extrapolations to high USTAR conditions.

441 2.7 Uncertainty Estimation

442 2.7.1 Uncertainty Evaluation

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

455
456 The consistency of the uncertainty estimates was evaluated using standard probabilistic
457 forecasting evaluation measures, namely calibration and sharpness ([Gneiting et al. 2007](#)).
458 Calibration captures the consistency between probabilistic forecasts and observations, and
459 measures whether predicted distributions correctly capture confidence levels as validated
460 against observed data. A well-calibrated model produces predictive distributions such that $P\%$
461 confidence interval (CI) contains the observations $P\%$ of the time. A model can be well
462 calibrated only at specific percentiles (e.g., 95%) or across multiple percentiles. At a minimum,
463 models should be well calibrated at the specific desired percentile before uncertainty estimates
464 at that percentile can be reliably used. Once models are shown to be well calibrated, they can
465 be compared using sharpness - a property that measures the concentration of the predictive

466 distributions. The approach of maximizing sharpness subject to calibration is widely adopted in
467 meteorology ([Gneiting and Katzfuss 2014](#)). Model improvement is captured by increasing
468 sharpness, subject to calibration. For each site, performance was evaluated at the 95% CI.
469 Calibration was measured by computing the proportion of the observed values within the 95%
470 CIs and measured sharpness using the mean width of the 95% CIs across the test set. A
471 normalized sharpness metric is reported by dividing by the standard deviation of flux to account
472 for the differing flux variance at each site.

473 2.7.2 Uncertainty Interval Scaling

474 Models that produce predictive distributions, such as the ML ensemble in the present study, are
475 not necessarily well calibrated by default. Several techniques have been proposed to calibrate
476 models after they are trained (post-processing calibration), most often using Platt scaling ([Platt
477 1999](#)) and isotonic regression ([Zadrozny and Elkan 2002](#)). In this work, Platt scaling is adopted
478 to calibrate the ensemble predictions. Platt scaling learns a scaling parameter that is used to
479 scale the variance uniformly for every input. This parameter is learned by assuming a
480 distribution (e.g., double exponential) and using maximum likelihood estimation to derive a value
481 from observed data. A double exponential distribution was assumed and derived a closed-form
482 expression for the scaling parameter (see **Appendix D** for derivation). Following this calibration
483 procedure, the probabilistic predictions of different models were compared by measuring the
484 sharpness of the calibrated distributions.

485 2.8 Annual and Growing Season Emissions

486 Annual CH₄ emissions were computed as the mean cumulative sum of the 10 gap-filled flux time
487 series, predicted by each ML model ensemble. To account for the uncertainty calibration
488 procedure, ensemble predictions were rescaled (spread out) around the mean in proportion to
489 the Platt scaling value. Annual sums and uncertainties (uncalibrated and calibrated) were
490 quantified from the mean and variance of the cumulative sums, respectively. As is standard for
491 CO₂ gap-filling, site-years with a gap of 60 days or longer during the growing or shoulder
492 seasons were excluded ([Richardson and Hollinger 2007](#); [Richardson et al. 2012](#)), except for
493 US-Uaf, which only had one site-year available, and for US-OWC, which had large shoulder or
494 growing season gaps during both available years. Additional date thresholds were applied for
495 the two rice paddies (US-Twt and JP-Mse) to only sum fluxes during the rice growing season
496 based on rice management information ([Knox et al. 2016](#); [Miyata et al. 2000](#)). All other gap-filled
497 values for gap lengths < 60 days were included. Annual or growing season CH₄ emissions

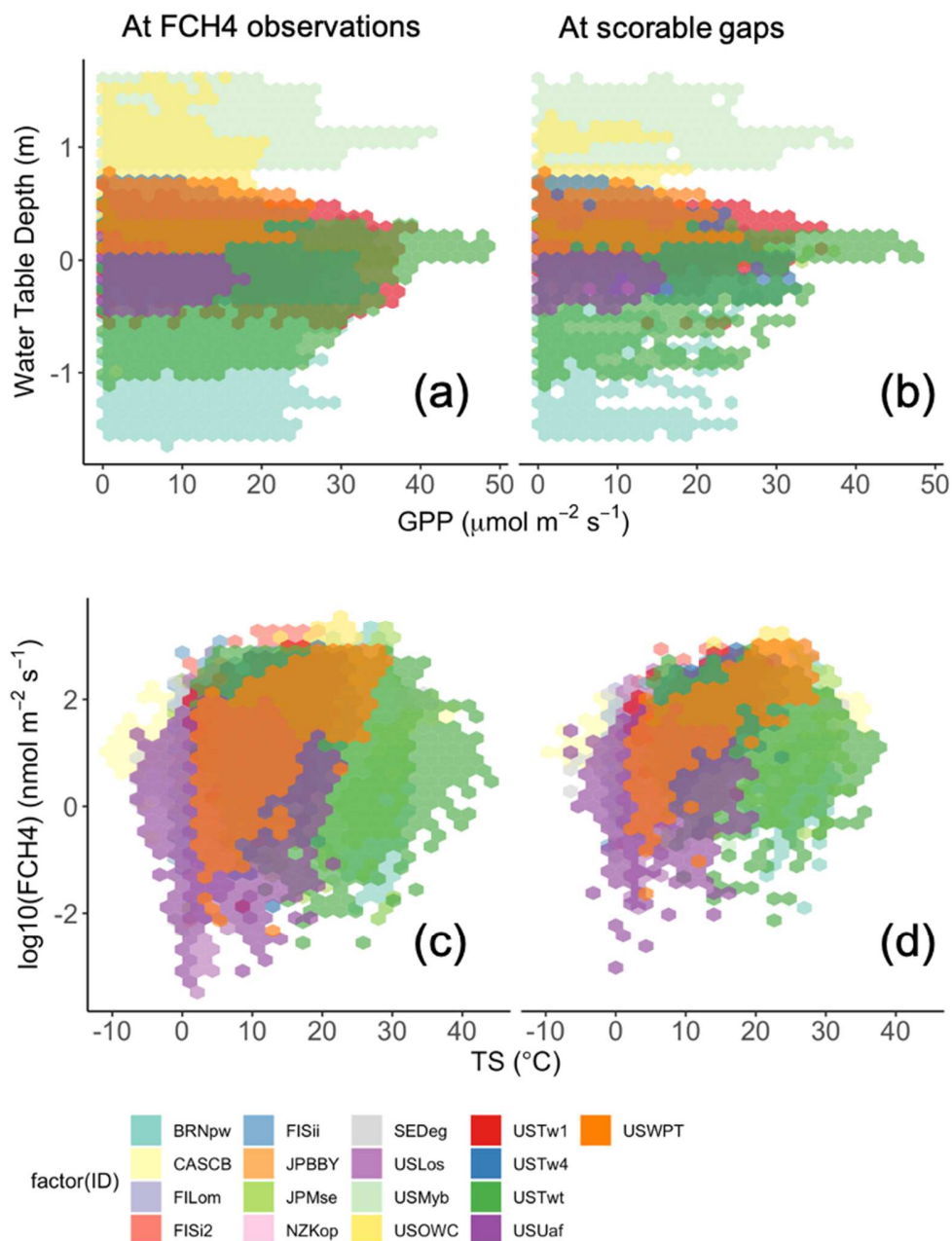
498 estimates were also computed for each of the seven MDS models (different predictor sets) as
499 the cumulative sum of the gap-filled time series. Similar to ML, summed uncertainties were
500 taken as the variance of the sums from the seven MDS models, however no calibration method
501 was applied.
502
503

504 **3 Results**

505 **3.1 Scorable Gap Conditions**

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

520



521
522
523
524
525

Figure 3: The coverage of training and test data for select predictor and CH₄ 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) CH₄ flux (FCH₄) and soil temperature (TS).

526 3.2 Performance Patterns on the Validation Set

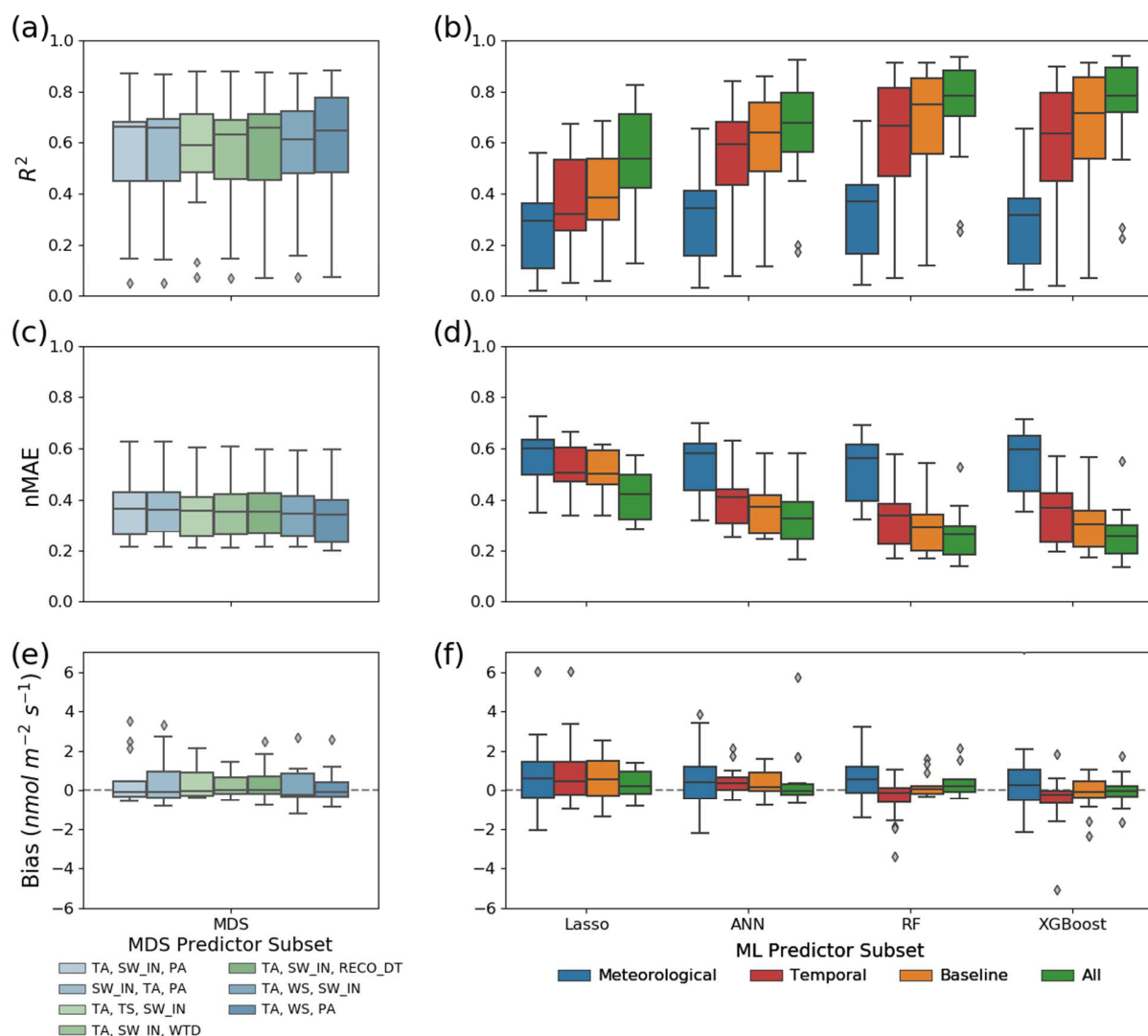
527 Median MDS performance ($R^2 = 0.65$; $nMAE = 0.35$; $Bias = -0.03 \text{ nmol m}^{-2} \text{ s}^{-1}$) was better than
528 median ML performance ($R^2 = 0.56$; $nMAE = 0.39$; $Bias = 0.01 \text{ nmol m}^{-2} \text{ s}^{-1}$). However, predictor
529 subsets had little effect on MDS performance (**Figure 4a, c, e**). Only slight improvements were
530 seen over baseline meteorological predictors (i.e., SW_IN, TA, and PA) when one of the CH₄-
531 centric predictors (i.e., WTD, TS, RECO, or WS) was included. Overall, the best performing
532 predictor combination for MDS was TA, PA, and WS ($R^2 = 0.66$; $nMAE = 0.34$; $Bias = -0.07$
533 $\text{nmol m}^{-2} \text{ s}^{-1}$), which was used subsequently to compute annual and growing season sums and
534 uncertainties.

535
536 There was a larger spread in performance across the ML (**Figure 4b, d, f**). Median performance
537 increased from Lasso ($R^2 = 0.37$; $nMAE = 0.51$; $Bias = 0.10 \text{ nmol m}^{-2} \text{ s}^{-1}$), to ANN ($R^2 = 0.58$;
538 $nMAE = 0.39$; $Bias = 0.06 \text{ nmol m}^{-2} \text{ s}^{-1}$), to XGBoost ($R^2 = 0.65$; $nMAE = 0.35$; $Bias = -0.11 \text{ nmol}$
539 $\text{m}^{-2} \text{ s}^{-1}$) and RF ($R^2 = 0.67$; $nMAE = 0.32$; $Bias = 0.01 \text{ nmol m}^{-2} \text{ s}^{-1}$). Unlike MDS, ML
540 performance was strongly dependent on the predictor set. Using all predictors was consistently
541 the best choice across all sites and all classes of models, while using the meteorological subset
542 alone performed the worst. Median model performance ranged from R^2 of 0.27, $nMAE$ of 0.60,
543 and mean Bias of $0.08 \text{ nmol m}^{-2} \text{ s}^{-1}$ for Lasso model class with the meteorological predictors
544 only, to R^2 of 0.79, $nMAE$ of 0.26, and Bias of $0.12 \text{ nmol m}^{-2} \text{ s}^{-1}$ for the RF model class with all
545 predictors. Notably, decision tree models using the baseline predictor set (e.g., RF $R^2 = 0.75$;
546 $nMAE = 0.29$; $Bias = 0.02 \text{ nmol m}^{-2} \text{ s}^{-1}$) still outperformed ANN using all predictors ($R^2 = 0.70$;
547 $nMAE = 0.31$; $Bias = 0.05 \text{ nmol m}^{-2} \text{ s}^{-1}$). For both decision tree and ANN models, the temporal
548 set was much more important for baseline performance than the meteorological set. As the
549 temporal set can be created for any CH₄ gap-filling effort, the meteorological set is unlikely to be
550 used alone in practice and is therefore only distinguished here to understand its relative
551 contribution to the baseline set.

552

553

554



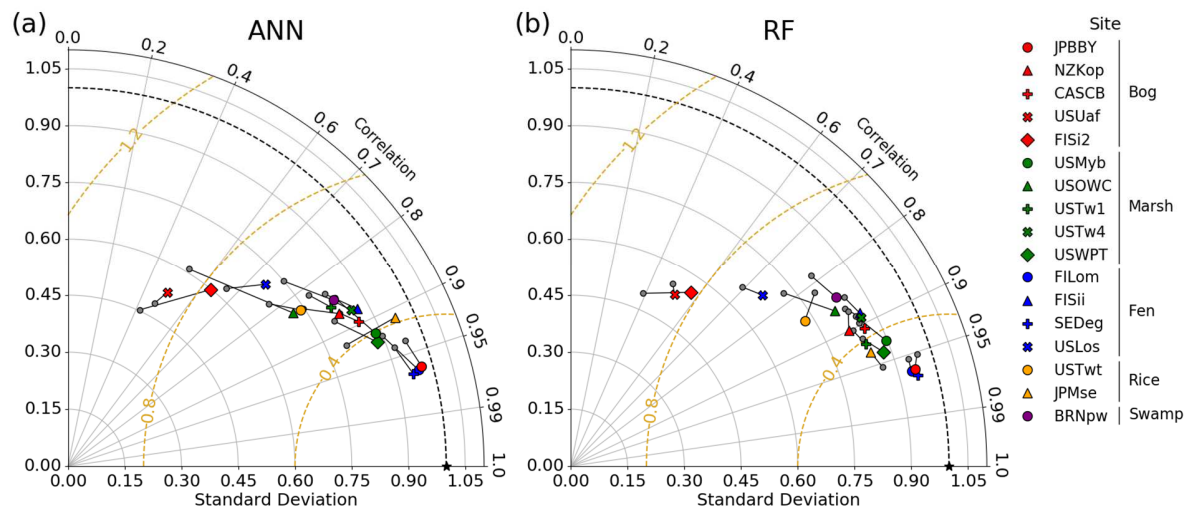
555
 556 **Figure 4: Boxplots illustrating 10 validation set performance metrics for each of**
 557 **the models (Lasso regression (Lasso), artificial neural networks (ANN), random**
 558 **forests (RF), and gradient boosted decision trees (XGBoost)) and predictor**
 559 **subsets across the 17 sites: (a, b) R^2 , (c, d) normalized mean absolute error**
 560 **(nMAE), (e, f) bias, where the left column is Marginal Distribution Sampling and**
 561 **the right column is machine learning. Each colored box shows the quartiles of the**
 562 **performance metrics and the whiskers show the rest of the distribution, excluding points**
 563 **determined to be outliers that are presented individually.**
 564

565 3.3 Test Set Performance Patterns

566 The ANN and RF (as the faster of the two decision tree algorithms) achieved the best
567 performance on the validation set and were then evaluated on the test set for each site. Test set
568 performance patterns were similar to the validation set, confirming that the models were not
569 over-fit. Median performance on the test set was better overall for RF ($R^2 = 0.79$; $nMAE = 0.27$;
570 $Bias = 0.24 \text{ nmol m}^{-2} \text{ s}^{-1}$) than ANN ($R^2 = 0.73$; $nMAE = 0.30$; $Bias = 0.18 \text{ nmol m}^{-2} \text{ s}^{-1}$). Median
571 $nMAE$ and R^2 both improved when ANN used all rather than baseline predictors ($p = 0.0007$ and
572 $p = 0.0004$, respectively). Similarly, median $nMAE$ and R^2 both improved when RF used all
573 rather than baseline predictors ($p = 0.0031$ and $p = 0.0050$, respectively). Test set evaluation
574 also provided some evidence of RF outperforming ANN in general. Using all predictors, median
575 $nMAE$ for the RF was smaller than that of the ANN ($p = 1.40e-8$) although there was no
576 significant difference between the median R^2 of RF and ANN ($p = 0.191$). Similarly, on baseline
577 predictors, median $nMAE$ for the RF was smaller than that of the ANN ($p = 0.0078$) but there
578 was no significant difference between the median R^2 of RF and ANN ($p = 0.056$).

579
580 A large spread in performance was observed within most wetland classes, suggesting a high
581 level of site uniqueness, rather than generalizability, within a particular wetland class (**Figure 5**).
582 The large spread was especially apparent for bogs and fens, whereas marshes and the two rice
583 paddies were clustered at intermediate to high performance. To better understand the patterns
584 of performance within and among wetland classes, correlations were examined between best
585 model performance metrics and the annual mean and variance of the fluxes. There was no
586 significant relationship between model performance and the annual mean of site CH_4 fluxes,
587 however, there was a clear negative relationship between performance and the coefficient of
588 variation of CH_4 fluxes ($p = 0.001$; $\rho = 0.72$) and an even stronger negative correlation with the
589 flux variance at short (hourly) timescales ($p = 1.44e-6$; $\rho = 0.89$) (**Figure E.1**).

590
591
592
593
594



595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

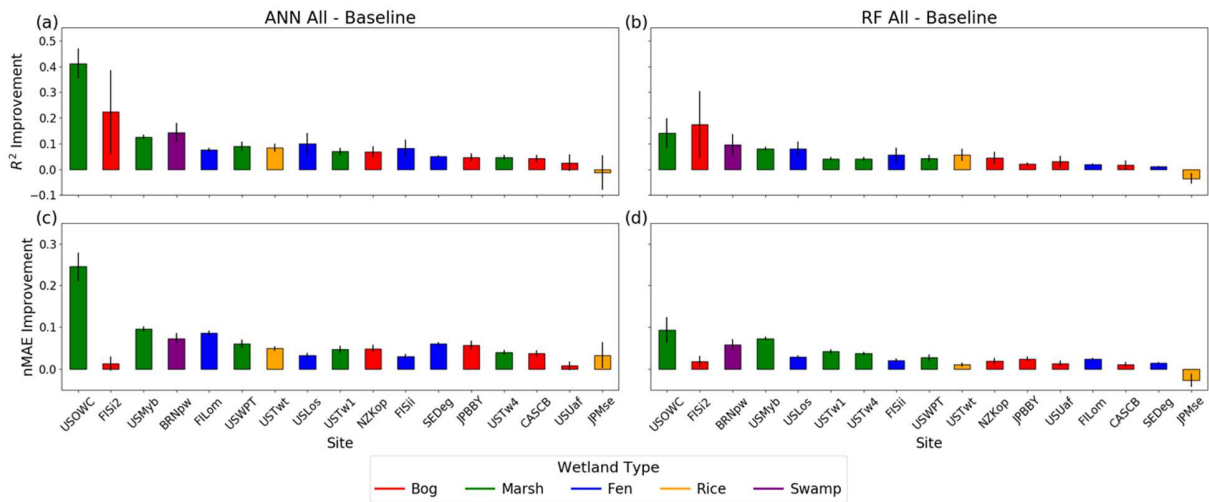
613

614

615

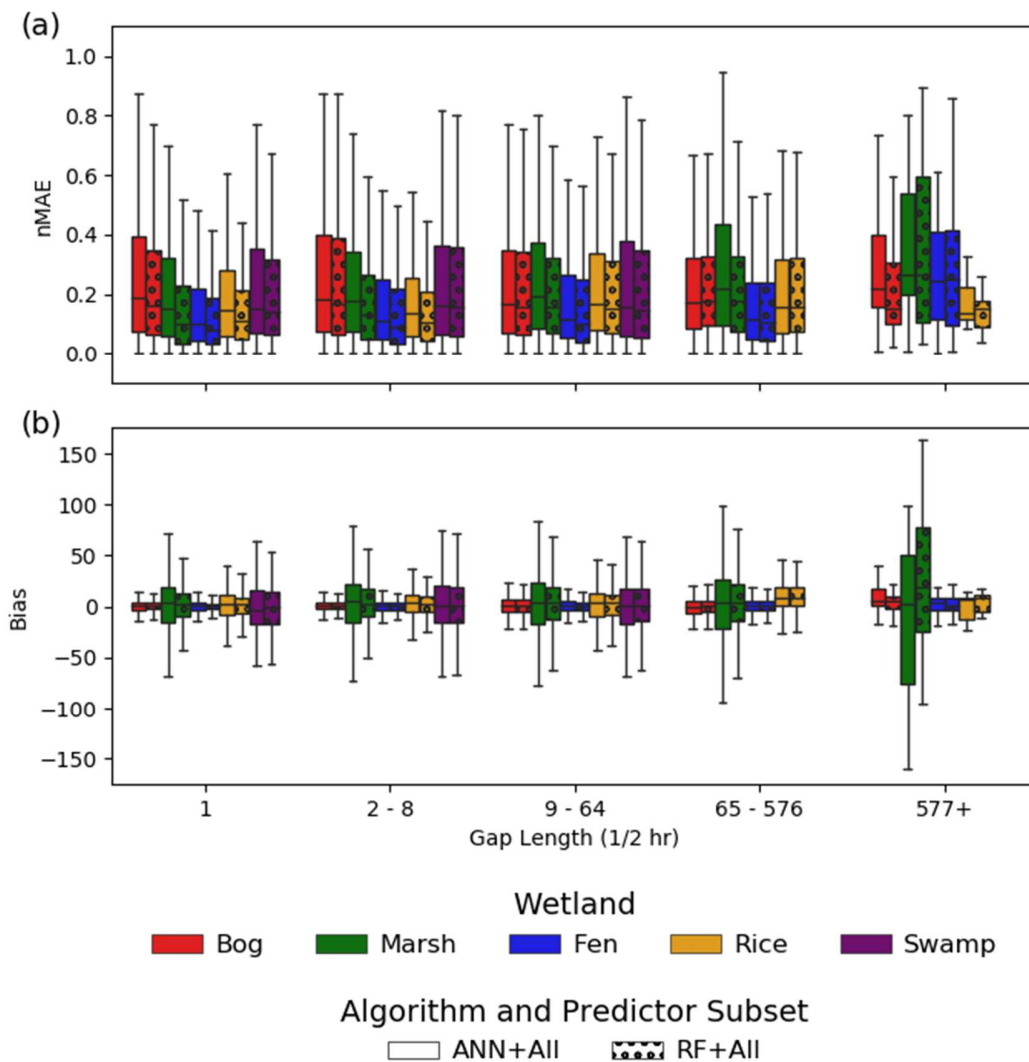
Figure 5: Taylor diagram visualizing artificial neural network (ANN) and random forest (RF) performance improvements on the test set between the baseline and all predictor sets for each of the 17 primary sites. The baseline set metrics for each algorithm are shown in small grey circle symbols and the 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 .



616
 617 **Figure 6: Improvements in test set performance metrics for the artificial neural**
 618 **network (ANN) and random forest (RF) algorithms between the baseline and all**
 619 **predictor sets on the 17 wetland sites.** Vertical error bars show the 95% confidence
 620 interval around the improvement, computed using the nonparametric basic bootstrap
 621 with 5,000 replicates. Sites are plotted in order of the total of R² and nMAE improvement.
 622

623 Across all very short (1 half-hour), short (2-8 half-hours), medium (9-64 half-hours), and long
 624 (65-576 half-hours) gap lengths, bias was low for both the ANN and RF models. Errors (nMAE)
 625 and biases were typically smaller for RF than ANN, and biases were generally larger at marshes
 626 and the swamp (**Figure 7**). For the longest gaps (577+ half-hours), RF and ANN performance
 627 was less consistent and the largest biases were introduced at marsh sites when using RF.
 628
 629

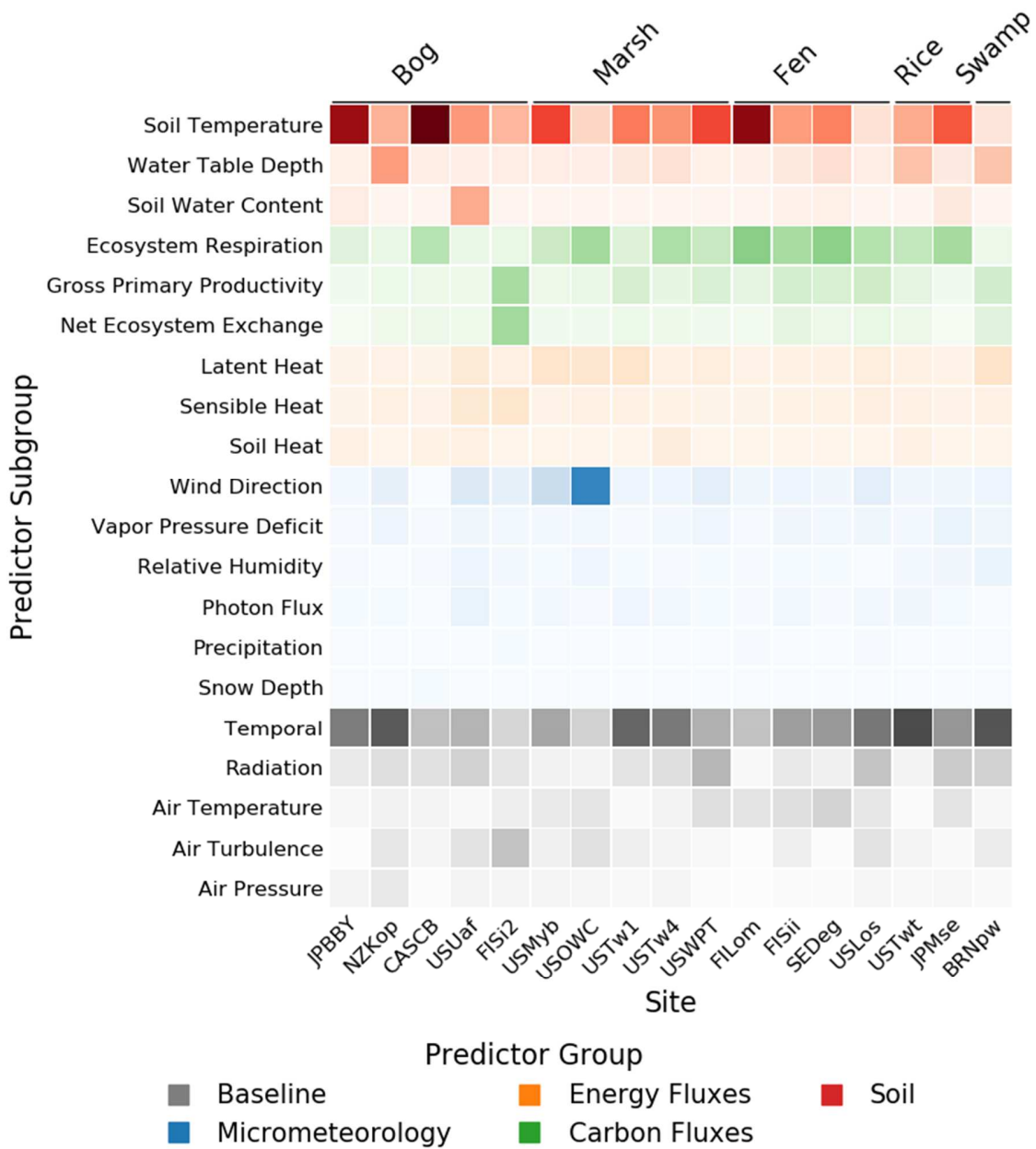


630
 631 **Figure 7: Performance of the two best algorithms (ANN+All and RF+All) on the test**
 632 **sets, broken down by gap length for the 17 primary sites.** Swamp values on long
 633 gaps (> 65 half-hours) are not shown here as the R^2 is not well-defined on single
 634 samples. Gap length values indicate merged gap lengths after test gap generation.

635
 636 Finally, an exploratory evaluation of errors that may be introduced due to USTAR filtering was
 637 conducted. The test set was used with the best model formulations (RF and all predictors).
 638 Model performance showed a slight reduction in performance when extrapolating to high
 639 USTAR conditions (**Table E.2**), suggesting that similar extrapolations to low USTAR conditions
 640 may introduce small but non-negligible errors. Average Bias across all 17 sites increased by
 641 9%, average nMAE by 10%, and R^2 decreased by 8%.

642 3.4 Predictor Importance

643 Variable importance rankings are readily retrievable from RF models. The most important
644 predictors of the RF model (in order) across all 17 sites were temporal, TS, radiation (aggregate
645 of SW_IN, SW_OUT, LW_IN, LW_OUT, and NETRAD), and RECO (**Figure 8**), with TS being
646 the single most important predictor for many sites. Air temperature (TA) and turbulence (WS
647 and USTAR), GPP and NEE, and WTD were useful for some sites, but not universally. Wind
648 direction (WD) was important at 2 sites (US-OWC and US-Myb). Generally, there were few
649 strong patterns within bogs, fens and marshes (which were the only classes with at least 4
650 representative sites), suggesting that predictor groups are not necessarily tied to wetland
651 classification, although TS was important at all of the bogs. Notably, the baseline set captured
652 several of the key predictors and all of the important meteorological predictors, except wind
653 direction. Of the two partitioning methods for RECO and GPP (nighttime and daytime), the
654 nighttime method ranked higher at 15 and 13 (of 17 total) sites, respectively.
655



657

658

659

660

661

662

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

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

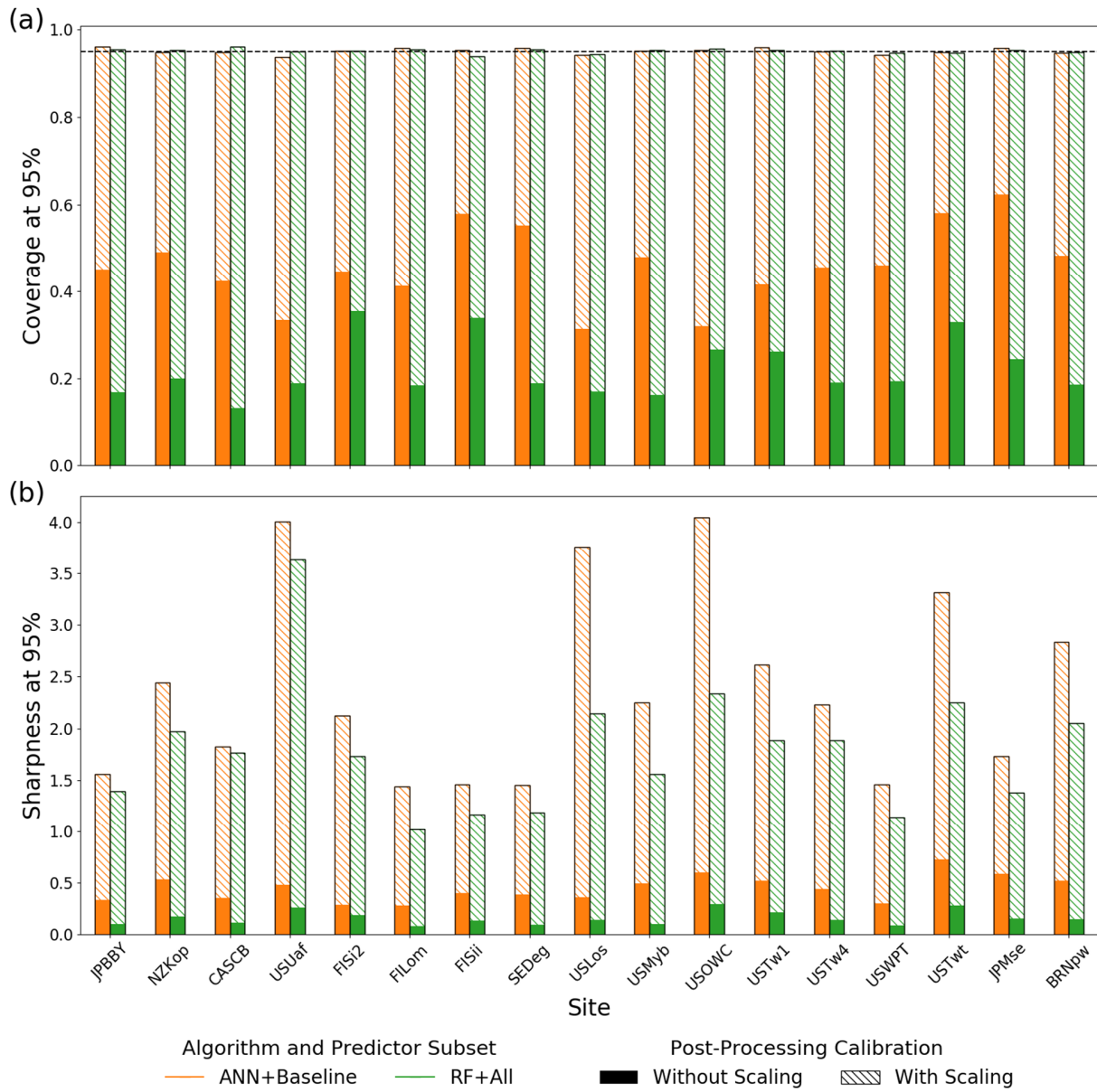
667 3.5 Uncertainty Estimation

668 The gap-filling prediction uncertainties for the two best ML algorithms (ANN and RF) were
669 evaluated with respect to the concepts of calibration and sharpness. For ANN, the baseline
670 predictor set model ensemble was evaluated because it most closely approximates a previously
671 described method ([Knox et al. 2019](#)) which was used to gap-fill the FLUXNET-CH4 Version 1.0
672 community product ([Delwiche et al. 2021](#)). The prediction uncertainties of both the ANN and RF
673 were not well-calibrated by default (**Figure 9**). In other words, without calibration by scaling, the
674 95% CI of the estimates for both models contained significantly less than 95% of the observed
675 values (56.6% on average for ANN, 28.4% on average for RF), indicating that the models
676 produced overly tight uncertainties across all sites. The ANN produced wider (less sharp)
677 uncertainty estimates than the RF without calibration.

678
679 At all sites, both ANN and RF model prediction uncertainties were well-calibrated after
680 performing the calibration step (**Figure 9**). In other words, the 95% CI of the estimates
681 contained close to 95% of the observed values in the test set (95.6% on average for ANN,
682 95.2% on average for RF). Notably, once calibrated, the RF model made sharper predictions
683 across all of the sites than the ANN model. The sites where predictions remained the widest
684 (least sharp) after normalizing by the standard deviation of flux were US-Uaf, US-Twt, US-OWC,
685 BR-Npw, and US-Los, which were the sites with the worst performance in terms of R^2 on the
686 test set. These sites had one or more of a site-specific combination of low seasonality and/or
687 extremely long gaps and/or highly variable fluxes. Similarly, the sites whose predictions were
688 the sharpest corresponded to the sites with the best performance on the test set. Examples of
689 pre- and post-calibration uncertainty ranges are shown in **Figure E.3**.

690

691



692

693

694

695

696

697

698

699

700

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

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

704 3.6 Annual and Growing Season Emissions

705 A total of 30.4 site years were gap-filled with MDS with best (TA, WS, and PA) predictors, and
 706 the baseline ML (ANN plus baseline predictors) and best ML (RF plus all predictors) models and
 707 summed for annual or growing season CH₄ emissions. Note that reported uncertainties around
 708 summed emissions reflect only gap-filling uncertainties and exclude additional random
 709 uncertainties which, though tending to be small, can be considered separately ([Knox et al.](#)
 710 [2019](#)) or in an integrated manner ([Vitale et al. 2018](#)).

711
 712 Annual and growing season emissions did not differ significantly (measured by overlapping 95%
 713 CI) at any of the sites when comparing the two ML gap-filling methods (**Table 5**). Calibrated
 714 prediction uncertainties for ANN and RF resulted in less sharp, but more plausible, 95% CI
 715 around the annual sum. For all sites except US-OWC and BR-Npw, emissions from the best ML
 716 model (RF and All) fell within the unscaled 95% CI of the baseline model (ANN and Baseline;
 717 approximating [Knox et al. 2019](#)), supporting a generally high level of accuracy for the baseline
 718 method under the majority of site and gap conditions in this analysis. At the highly variable US-
 719 OWC marsh and BR-Npw swamp sites, the best model predictions fell outside the unscaled but
 720 within the scaled baseline CI, which underscores the implausible sharpness of unscaled ML
 721 ensemble predictions but does not support greater accuracy of RF than ANN. Uncertainties
 722 around MDS were much sharper (median 95% CI was $\pm 3\%$ of annual emissions) than the
 723 scaled ML methods for ANN ($\pm 38\%$) and RF ($\pm 18\%$). The sharp uncertainties resulted in small
 724 but significant differences between annual and growing season sums from MDS and one ML
 725 model (e.g., JP-BBY, BR-Npw, US-Tw1) or both ML models (e.g., CA-SCB, US-Los).

726

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

731

Site	Annual or Growing Season Date Ranges	Mean Annual or Growing Season Methane Emissions \pm 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

732 4 Discussion

733 4.1 Methods & Algorithms

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

745
746 Decision tree-based models (RF and XGBoost) showed better performance than ANN and
747 Lasso models across the majority of the 17 wetland and rice paddy sites. This is consistent with
748 recent work on CH₄ gap-filling which demonstrated that a RF gap-filling model outperformed
749 both ANN and support vector regression models across five wetland and rice paddy sites
750 [\(Nemitz et al. 2018; Kim et al. 2019; Knox et al. 2019\)](#). RF models are also relatively easy to
751 tune, fast to train even on large datasets, and require little preprocessing. Furthermore,

752 decision-tree-based models are more interpretable (presently) than ANN ([Russell and Norvig](#)
753 [1995](#)), which enables analysis of important predictors. In comparison to ML approaches, MDS
754 was tested as an easy and fast method that makes use of only three predictors. MDS scored
755 highly on average although still much lower than the best ML models. [Kim et al. \(2019\)](#) also
756 found that MDS more frequently introduced statistical bias in annual sums than ML models.

757

758 Although RF and ANN models are recommended ML methods, there is still room to improve
759 their gap-filling performance, especially on long gaps. Recent deep neural network architectures
760 have shown impressive results in modeling long sequences in natural language processing,
761 particularly recurrent neural network variants ([Lipton et al. 2015](#)) and Transformers ([Vaswani et](#)
762 [al. 2017](#)). These models have the potential to reproduce highly nonlinear variable interactions
763 using large datasets including half-hourly time series flux data and may be able to capture
764 lagged relationships between predictors and CH₄ flux without further manual revision. However,
765 representing non-stationary conditions such as pulse events has proven to be challenging for
766 ML approaches ([Vargas et al. 2018](#)). Future work could explore the effectiveness of deep neural
767 network architectures for gap-filling CH₄. It is likely, however, that problems of non-stationarity
768 during long gaps will apply for CH₄ as they do for CO₂ imputation ([Richardson and Hollinger](#)
769 [2007](#)) and are best handled during data collection.

770 4.2 Methane Predictors

771 The inclusion of soil temperature (TS) and ecosystem carbon flux predictors (NEE, RECO, and
772 GPP) improved gap-filling performance over the baseline set (three temporal, plus TA, PA,
773 SW_IN, and WS), in broad agreement with known controls by temperature ([Yvon-Durocher et](#)
774 [al. 2014](#)) and substrate availability ([Whiting and Chanton 1993](#); [Matthes et al. 2014](#); [McNicol et](#)
775 [al. 2020](#); [Laanbroek 2010](#)). Soil temperature was the single most important additional predictor
776 over the baseline set at most sites, followed by RECO. While TS was available at all sites in this
777 study, it is not available across all FLUXNET sites. Although NEE and its component ecosystem
778 carbon fluxes (GPP and RECO) are highly correlated, the consistent favoring of RECO
779 suggests they are not perfectly interchangeable for gap-filling performance, and RECO and CH₄
780 flux are both largely the result of microbial metabolism, and are similarly affected by
781 environmental drivers ([Morin et al. 2014](#)). However, partitioned fluxes (RECO and GPP) are
782 overall less practical than measured NEE as predictors because they are typically partitioned
783 from NEE as a function of TS, and thus its importance may largely reflect its correlation with TS

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

786

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

806

807 The performance improvements using all predictors in this study suggests a moderate amount
808 of predictor redundancy does not harm ML performance and predictor curation may be less
809 important for ML than in other modeling approaches. [Kim et al. \(2019\)](#) similarly showed that ML
810 models can benefit from a large predictor set that includes soil variables and that dimension-
811 reduction via principal component analysis was not necessary to achieve good performance.
812 However, site uniqueness may also necessitate the tailoring of models for optimal performance
813 at individual sites, illustrated in this study by the ranges in 1) observed CH₄ fluxes, 2) model
814 performance, and 3) predictor importance within bog, fen, and marsh classes. For instance,
815 despite high spatial variability in CH₄ fluxes at some wetlands [\(Rey-Sanchez et al. 2018;](#)
816 [Matthes et al. 2014\)](#), WD (which determines the flux footprint) was only an important predictor at
817 one marsh site (US-OWC), which has very high spatial variation in flux between different cover

818 types ([Rey-Sanchez et al. 2018](#)). The site-specificity of WD for heterogeneous sites was also
819 reported in a recent study that used a ML approach to partition NEE ([Tramontana et al. 2020](#)).
820 Entirely new predictors may also be necessary at some sites, such as salinity, which is likely an
821 important predictor for gap-filling at estuaries or other coastal locations with a (tidal) salinity
822 influence ([Holmquist et al. 2018](#); [Poffenbarger et al. 2011](#)). Although not prioritized in the
823 present study, a more parsimonious predictor set may be identified via a combination of site-
824 specific and process knowledge, as well as automated feature selection methods ([Kumar and](#)
825 [Minz 2014](#)). Curated predictor sets should, however, be reevaluated when gap-filling new data
826 (e.g., site-years, or across multiple sites) as past models may be overfit with respect to new
827 data conditions.

828

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

839 4.3 Integrated Emissions & Uncertainties

840 Computing annual or growing season CH₄ emissions requires gap-filling because filtering of EC
841 data and other acquisition issues typically creates gaps of a wide variety of lengths, and
842 especially an abundance of short gaps (**Table C2**). Gaps are not normally distributed in time
843 and therefore FCH₄ observations are likely to be biased, which will propagate to the time-
844 integrated flux. However, the investigator must decide: 1) which gap-filled values are likely to be
845 of sufficient accuracy to be retained, and 2) whether the retained gap-filled plus observed values
846 are sufficient to integrate emissions over an annual, seasonal, or other timeframe. As a rough
847 guide, filled values should be treated with greater scrutiny as they become longer and less
848 frequent in the scorable dataset. The most abundant scorable gaps of length one half-hour to
849 approximately 12 days can be filled confidently, given performance metric checks as described
850 in this study. Investigators should, however, be aware that episodic fluxes, perhaps due to

851 ebullition events, may not always be captured and instead may be filled with average fluxes for
852 the most comparable conditions (e.g., FCH₄ and MAE spikes in **Figure E.2**). Greater scrutiny of
853 evaluation metrics is recommended for gaps longer than approximately 12 days, but less than
854 multiple months, whereas, filled values in gaps of multiple months (> 60 days) should generally
855 be excluded, as is done in CO₂ gap-filling ([Wutzler et al. 2018](#)). The exception may be very long
856 (decadal) datasets where the monthly-scale gap occurs in a season with ample data from other
857 sites-years and can be reasonably evaluated. After determining which filled values to retain, the
858 coverage of filled plus observed fluxes should be considered with respect to the integration
859 period. For rice paddies (e.g., US-Twt, JP-Mse), and sites with low winter season fluxes due to
860 frozen soils (US-OWC or US-Uaf), it may be adequate and interesting to report a growing
861 season flux as is done in this study and the FLUXNET-CH₄ synthesis ([Delwiche et al. 2021](#)).
862 Time-integrated uncertainties from ML gap-filling methods will also widen significantly as more
863 gap-filling is required and should always be reported alongside long-term sums.

864

865 The improvement in performance gained by using ML over MDS, and all predictors over
866 baseline predictors, did not have a significant effect on annual CH₄ emissions estimates at most
867 sites. However, seemingly minor changes in CH₄ fluxes can have disproportionate impacts
868 when calculating greenhouse gas emissions due to the high radiative forcing effects of CH₄ or
869 when sparsely distributed sites are used in data-driven regional or global upscaling efforts
870 ([Tramontana et al. 2016](#); [Roberts et al. 2017](#)). Specifically, absolute differences in annual
871 emissions among the gap-filling methods were larger at high-emitting sites which could lead to
872 larger upscaling errors in high-emitting tropical regions that account for > 60% of global wetland
873 sources ([Wania et al. 2013](#); [Bloom et al. 2017](#); [Saunois et al. 2020](#)). These results therefore
874 highlight the need for robust methods for estimating and propagating uncertainty from flux gap-
875 filling to upscaling.

876

877 Machine learning model-generated uncertainties around both half-hourly predictions and annual
878 emissions have been underestimated. A scaling procedure (Platt scaling) which expands the
879 uncertainty estimates can be used to produce well-calibrated predictions. Well-calibrated
880 models can be compared using the sharpness of their predictions, where sharper predictions
881 corresponded to better models. Using this method, sharper uncalibrated RF (compared to ANN)
882 prediction uncertainties were retained post-calibration, indicating greater precision of
883 predictions. However, the frequent overlap between uncalibrated and calibrated for both
884 algorithms means a firm conclusion about algorithm differences in accuracy is not possible. It is

885 also acknowledged that this uncertainty does not capture all sources of uncertainty that could
886 arise from random measurement errors, unseen events, uncertainties in the predictors, or other
887 systematic bias, among others. However, calibrating predictive ML models to avoid
888 underestimating gap-filling uncertainties is strongly recommended.

889
890 Other calibration methods have the potential to achieve calibration while producing sharper
891 predictions ([Kuleshov et al. 2018](#)). Furthermore, probabilistic models like Gaussian processes or
892 multiple imputation methods may be able to produce well-calibrated models without the need for
893 post-processing calibration procedures ([Vitale et al. 2018](#); [Camps-Valls et al. 2019](#)). Recently, a
894 method for producing uncertainty estimates from any gradient boosting model was introduced
895 which may enable decision tree models to produce well-calibrated, probabilistic predictions
896 without requiring a model ensemble or post-processing calibration ([Duan et al. 2019](#)). Finally,
897 deep learning models can capture highly nonlinear relationships in large datasets and make
898 probabilistic predictions which have the potential to outperform other gap-filling methods.

899 5 Conclusions

900 This study outlines a robust and reproducible ML workflow for CH₄ gap-filling models that can be
901 applied at individual wetland sites or in multi-site syntheses. Specifically, the study advances
902 CH₄ gap-filling in wetlands using ML by: 1) introducing a thorough gap-filling model development
903 and validation procedure that reliably generates gaps and splits the data into training, validation,
904 and test sets; 2) experimentally evaluating conventional MDS (with drivers adapted for wetland
905 CH₄ fluxes) against combinations of ML algorithms and predictor sets; and 3) proposing a model
906 calibration method to estimate, evaluate, and calibrate model uncertainties. This study also
907 provides insights into methodological choices. Decision tree algorithms (RF and XGBoost) offer
908 the best performance on average; using all predictors (or best set for MDS), median nMAE
909 followed the order Lasso (0.42) > MDS (0.34) > ANN (0.31) > RF/XGBoost (0.26), and median
910 R² followed the order Lasso (0.57) < MDS (0.66) < ANN (0.70) < RF/XGBoost (0.79). Overall,
911 RF is recommended as it benefits from less pre-processing and faster run-time than XGBoost.
912 ANN predictions had less bias when filling the longest gaps and performance improved when
913 using all rather than baseline predictors, suggesting ANN may benefit from additional predictor
914 curation and feature engineering. Using all available variables collected at eddy covariance
915 towers as predictors is also fast, effective, and reasonable, given the large ratio of observations
916 to predictors (favorable data dimensionality). Conventional MDS also proved to be a fast

917 method that provides reasonable performance when CH₄ predictors (air temperature, air
918 pressure, and wind speed) are selected, however, the lack of post-calibration results in
919 uncertainties that are very sharp (unrealistic). ML prediction uncertainties, in contrast, can be
920 calibrated to observations using Platt scaling. Finally, based on variable importance results, it is
921 recommended that soil temperature and water table depth are measured at all wetland eddy
922 covariance sites. The python code for developing gap-filling methods, comparing predictions,
923 and calibrating uncertainties is available [[https://github.com/stanfordmlgroup/methane-gapfill-
924 ml](https://github.com/stanfordmlgroup/methane-gapfill-ml)]. For future evaluations at wetlands and other ecosystems, this code can provide a
925 foundation for the development of standardized eddy covariance CH₄ processing by different
926 teams and Regional Flux Networks which can also be tested on nitrous oxide fluxes as longer
927 time series become available ([Papale 2020](#)).

928

929

930

931 **Acknowledgments**

932 This study was supported by the Gordon and Betty Moore Foundation through Grant
933 GBMF5439 “Advancing Understanding of the Global Methane Cycle” to Stanford University
934 supporting the Methane Budget activity for the Global Carbon Project (globalcarbonproject.org).
935 BRKR was supported by NSF Award 1752083. OS was supported through the Canada
936 Research Chairs and Natural Sciences and Engineering Research Council Discovery Grants
937 programs. TS and CW were supported by the Helmholtz Association of German Research
938 Centres (Grant No. VH-NG-821). GB was supported by a US Department of Energy (Grant DE-
939 SC0021067) and NOAA Davidson Fellowship Award administered by ODNR OWC-NERR
940 (Subaward N18B 315-11). Funding from the SNF projects DiRad and InnoFarm (146373 and
941 407340_172433), from the ETH Board and from ETH Zurich is greatly acknowledged. DP, CT
942 and DV were supported by the Department of Excellence 2018 Program MIUR Project
943 “Landscape 4.0 – food, wellbeing and environment” and the ICOS Ecosystem Thematic Centre.
944 CT was supported by the E-SHAPE (GA820852) H2020 European project. DB, JV, DS, CS, SK,
945 EE, KSH, KK, AV, CRS, RS (or sites US-MYB, US-TW1, US-TW3, US-TW4, US-TW5, US-
946 TWT, US-SND, US-SNE) were supported by the California Department of Water Resources
947 through a contract from the California Department of Fish and Wildlife and the United States
948 Department of Agriculture (NIFA grant #2011-67003-30371). Funding for the AmeriFlux core
949 sites was provided by the U.S. Department of Energy’s Office of Science (AmeriFlux contract
950 #7079856). KK was supported by the Estonian Research Council grants No. PSG631 and

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

960

961

962 **Author Contributions**

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

976

977 **References**

978

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

982

983 Bansal, S., Tangen, B., Finocchiaro, R., 2018. Diurnal Patterns of Methane Flux from a
984 Seasonal Wetland: Mechanisms and Methodology. *Wetlands* 38, 933–943.
985 <https://doi.org/10.1007/s13157-018-1042-5>
986

987 Barba, J., Cueva, A., Bahn, M., Barron-Gafford, G.A., Bond-Lamberty, B., Hanson, P.J.,
988 Jaimes, A., Kulmala, L., Pumpanen, J., Scott, R.L., Wohlfahrt, G., Vargas, R., 2018.
989 Comparing ecosystem and soil respiration: Review and key challenges of tower-
990 based and soil measurements. *Agric. For. Meteorol.* 249, 434–443.
991 <https://doi.org/10.1016/j.agrformet.2017.10.028>
992

993 Bloom, A.A., Bowman, K.W., Lee, M., Turner, A.J., Schroeder, R., Worden, J.R., Weidner,
994 R.J., McDonald, K.C., Jacob, D.J., 2017. CMS: Global 0.5-deg Wetland Methane
995 Emissions and Uncertainty (WetCHARTs v1. 0).
996 <https://doi.org/10.3334/ORNLDAAC/1502>
997

998 Bodesheim, P., Jung, M., Gans, F., Mahecha, M.D., Reichstein, M., 2018. Upscaled diurnal
999 cycles of land–atmosphere fluxes: a new global half-hourly data product. *Earth Syst.*
1000 *Sci. Data* 10, 1327–1365. <https://doi.org/10.5194/essd-10-1327-2018>
1001

1002 Bohrer, G., Kerns, J., Morin, T., Rey-Sanchez, A., Villa, J., Ju, Y., 2020. FLUXNET-CH4 US-
1003 OWC Old Woman Creek. <https://doi.org/10.18140/FLX/1669690>
1004

1005 Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32.
1006 <https://doi.org/10.1023/A:1010933404324>
1007

1008 Bridgman, S.D., Cadillo-Quiroz, H., Keller, J.K., Zhuang, Q., 2013. Methane emissions from
1009 wetlands: biogeochemical, microbial, and modeling perspectives from local to global
1010 scales. *Glob. Chang. Biol.* 19, 1325–1346. <https://doi.org/10.1111/gcb.12131>
1011

1012 Campbell, D., Goodrich, J., 2020. FLUXNET-CH4 NZ-Kop Kopuatai.
1013 <https://doi.org/10.18140/FLX/1669652>
1014

1015 Camps-Valls, G., Sejdinovic, D., Runge, J., Reichstein, M., 2019. A perspective on Gaussian
1016 processes for Earth observation. *Natl Sci Rev* 6, 616–618.
1017 <https://doi.org/10.1093/nsr/nwz028>
1018

1019 Chang, K.-Y., Riley, W.J., Knox, S.H., Jackson, R.B., McNicol, G., Poulter, B., Aurela, M.,
1020 Baldocchi, D., Bansal, S., Bohrer, G., Campbell, D.I., Cescatti, A., Chu, H., Delwiche,
1021 K.B., Desai, A.R., Euskirchen, E., Friborg, T., Goeckede, M., Helbig, M., Hemes, K.S.,
1022 Hirano, T., Iwata, H., Kang, M., Keenan, T., Krauss, K.W., Lohila, A., Mammarella, I.,
1023 Mitra, B., Miyata, A., Nilsson, M.B., Noormets, A., Oechel, W.C., Papale, D., Peichl,
1024 M., Reba, M.L., Rinne, J., Runkle, B.R.K., Ryu, Y., Sachs, T., Schäfer, K.V.R.,
1025 Schmid, H.P., Shurpali, N., Sonnentag, O., Tang, A.C.I., Torn, M.S., Trotta, C.,
1026 Tuittila, E.-S., Ueyama, M., Vargas, R., Vesala, T., Windham-Myers, L., Zhang, Z.,
1027 Zona, D., 2021. Substantial hysteresis in emergent temperature sensitivity of global
1028 wetland CH₄ emissions. *Nat. Commun.* 12, 1–10. [https://doi.org/10.1038/s41467-](https://doi.org/10.1038/s41467-021-22452-1)
1029 [021-22452-1](https://doi.org/10.1038/s41467-021-22452-1)
1030

1031 Chen, J., Chu, H., 2020. FLUXNET-CH₄ US-WPT Winous Point North Marsh.
1032 <https://doi.org/10.18140/FLX/1669702>
1033

1034 Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]*.
1035

1036 Dalmagro, H.J., Lathuilière, M.J., Hawthorne, I., Morais, D.D., Pinto, O.B., Jr, Couto, E.G.,
1037 Johnson, M.S., 2018. Carbon biogeochemistry of a flooded Pantanal forest over three
1038 annual flood cycles. *Biogeochemistry* 139, 1–18. [https://doi.org/10.1007/s10533-018-](https://doi.org/10.1007/s10533-018-0450-1)
1039 [0450-1](https://doi.org/10.1007/s10533-018-0450-1)
1040

1041 Delwiche, K.B., Knox, S.H., Malhotra, A., Fluet-Chouinard, E., McNicol, G., Feron, S.,
1042 Ouyang, Z., Papale, D., Trotta, C., Canfora, E., Cheah, Y.-W., Christianson, D.,
1043 Alberto, M.C.R., Alekseychik, P., Aurela, M., Baldocchi, D., Bansal, S., Billesbach,
1044 D.P., Bohrer, G., Bracho, R., Buchmann, N., Campbell, D.I., Celis, G., Chen, J.,
1045 Chen, W., Chu, H., Dalmagro, H.J., Dengel, S., Desai, A.R., Detto, M., Dolman, H.,
1046 Eichelmann, E., Euskirchen, E., Famulari, D., Friborg, T., Fuchs, K., Goeckede, M.,
1047 Gogo, S., Gondwe, M.J., Goodrich, J.P., Gottschalk, P., Graham, S.L., Heimann, M.,
1048 Helbig, M., Helfter, C., Hemes, K.S., Hirano, T., Hollinger, D., Hörtnagl, L., Iwata, H.,

1049 Jacotot, A., Jansen, J., Jurasinski, G., Kang, M., Kasak, K., King, J., Klatt, J.,
1050 Koebisch, F., Krauss, K.W., Lai, D.Y.F., Mammarella, I., Manca, G., Marchesini, L.B.,
1051 Matthes, J.H., Maximon, T., Merbold, L., Mitra, B., Morin, T.H., Nemitz, E., Nilsson,
1052 M.B., Niu, S., Oechel, W.C., Oikawa, P.Y., Ono, K., Peichl, M., Peltola, O., Reba,
1053 M.L., Richardson, A.D., Riley, W., Runkle, B.R.K., Ryu, Y., Sachs, T., Sakabe, A.,
1054 Sanchez, C.R., Schuur, E.A., Schäfer, K.V.R., Sonnentag, O., Sparks, J.P., Stuart-
1055 Haëntjens, E., Sturtevant, C., Sullivan, R.C., Szutu, D.J., Thom, J.E., Torn, M.S.,
1056 Tuittila, E.-S., Turner, J., Ueyama, M., Valach, A.C., Vargas, R., Varlagin, A.,
1057 Vazquez-Lule, A., Verfaillie, J.G., Vesala, T., Vourlitis, G.L., Ward, E.J., Wille, C.,
1058 Wohlfahrt, G., Wong, G.X., Zhang, Z., Zona, D., Windham-Myers, L., Poulter, B.,
1059 Jackson, R.B., 2021. FLUXNET-CH4: A global, multi-ecosystem dataset and analysis
1060 of methane seasonality from freshwater wetlands. *Earth Syst. Sci. Data*.
1061 <https://doi.org/10.5194/essd-2020-307>
1062
1063 Dengel, S., Zona, D., Sachs, T., Aurela, M., Jammot, M., Parmentier, F.J.W., Oechel, W.,
1064 Vesala, T., 2013. Testing the applicability of neural networks as a gap-filling method
1065 using CH₄ flux data from high latitude wetlands. *Biogeosciences* 10, 8185–8200.
1066 <https://doi.org/10.5194/bg-10-8185-2013>
1067
1068 Derrac, J., García, S., Molina, D., Herrera, F., 2011. A practical tutorial on the use of
1069 nonparametric statistical tests as a methodology for comparing evolutionary and
1070 swarm intelligence algorithms. *Swarm Evol. Comput.* 1, 3–18.
1071 <https://doi.org/10.1016/j.swevo.2011.02.002>
1072
1073 Desai, A., 2020. FLUXNET-CH4 US-Los Lost Creek. <https://doi.org/10.18140/FLX/1669682>
1074
1075 Duan, T., Avati, A., Ding, D.Y., Basu, S., Ng, A.Y., Schuler, A., 2020. NGBoost: Natural
1076 Gradient Boosting for Probabilistic Prediction, in: *International Conference on*
1077 *Machine Learning*. PMLR, pp. 2690–2700.
1078
1079 Efron, B., Tibshirani, R.J., 1994. *An Introduction to the Bootstrap*. CRC Press.
1080

1081 Eichelmann, E., Knox, S., Sanchez, C., Valach, A., Sturtevant, C., Szutu, D., Verfaillie, J.,
1082 Baldocchi, D., 2020. FLUXNET-CH4 US-Tw4 Twitchell East End Wetland.
1083 <https://doi.org/10.18140/FLX/1669698>
1084

1085 Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G.,
1086 Ceulemans, R., Clement, R., Dolman, H., Granier, A., Gross, P., Grünwald, T.,
1087 Hollinger, D., Jensen, N.-O., Katul, G., Keronen, P., Kowalski, A., Lai, C.T., Law, B.E.,
1088 Meyers, T., Moncrieff, J., Moors, E., Munger, J.W., Pilegaard, K., Rannik, Ü.,
1089 Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K.,
1090 Wofsy, S., 2001. Gap filling strategies for defensible annual sums of net ecosystem
1091 exchange. *Agric. For. Meteorol.* 107, 43–69. [https://doi.org/10.1016/s0168-](https://doi.org/10.1016/s0168-1923(00)00225-2)
1092 [1923\(00\)00225-2](https://doi.org/10.1016/s0168-1923(00)00225-2)
1093

1094 Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for
1095 global land areas. *Int. J. Climatol.* 37, 4302–4315. <https://doi.org/10.1002/joc.5086>
1096 Freund, Y., Schapire, R., Abe, N., 1999. A short introduction to boosting. *Journal-Japanese*
1097 *Society For Artificial Intelligence* 14, 1612.
1098

1099 Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and
1100 sharpness. *J. R. Stat. Soc. Series B Stat. Methodol.* 69, 243–268.
1101 <https://doi.org/10.1111/j.1467-9868.2007.00587.x>
1102

1103 Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. *Annu. Rev. Stat. Appl.* 1, 125–151.
1104 <https://doi.org/10.1146/annurev-statistics-062713-085831>
1105

1106 Göckede, M., Kittler, F., Schaller, C., 2019. Quantifying the impact of emission outbursts and
1107 non-stationary flow on eddy-covariance CH₄ flux measurements using wavelet
1108 techniques. *Biogeosciences* 16, 3113–3131. <https://doi.org/10.5194/bg-16-3113-2019>
1109

1110 Goodrich, J.P., Campbell, D.I., Roulet, N.T., Clearwater, M.J., Schipper, L.A., 2015.
1111 Overriding control of methane flux temporal variability by water table dynamics in a
1112 Southern Hemisphere, raised bog: Methane fluxes from a S.H. bog. *J. Geophys. Res.*
1113 *Biogeosci.* 120, 819–831. <https://doi.org/10.1002/2014jg002844>
1114

1115 Günther, A., Barthelmes, A., Huth, V., Joosten, H., Jurasinski, G., Koebisch, F., Couwenberg,
1116 J., 2020. Prompt rewetting of drained peatlands reduces climate warming despite
1117 methane emissions. *Nat. Commun.* 11, 1644. [https://doi.org/10.1038/s41467-020-](https://doi.org/10.1038/s41467-020-15499-z)
1118 [15499-z](https://doi.org/10.1038/s41467-020-15499-z)
1119

1120 Hatala, J.A., Detto, M., Baldocchi, D.D., 2012. Gross ecosystem photosynthesis causes a
1121 diurnal pattern in methane emission from rice. *Geophys. Res. Lett.* 39, L06409.
1122 <https://doi.org/10.1029/2012gl051303>
1123

1124 Hemes, K.S., Chamberlain, S.D., Eichelmann, E., Anthony, T., Valach, A., Kasak, K., Szutu,
1125 D., Verfaillie, J., Silver, W.L., Baldocchi, D.D., 2019. Assessing the carbon and
1126 climate benefit of restoring degraded agricultural peat soils to managed wetlands.
1127 *Agric. For. Meteorol.* 268, 202–214. <https://doi.org/10.1016/j.agrformet.2019.01.017>
1128

1129 Hollinger, D.Y., Richardson, A.D., 2005. Uncertainty in eddy covariance measurements and
1130 its application to physiological models. *Tree Physiol.* 25, 873–885.
1131 <https://doi.org/10.1093/treephys/25.7.873>
1132

1133 Holmquist, J.R., Windham-Myers, L., Bernal, B., Byrd, K.B., Crooks, S., Gonneea, M.E.,
1134 Herold, N., Knox, S.H., Kroeger, K.D., McCombs, J., Megonigal, J.P., Lu, M., Morris,
1135 J.T., Sutton-Grier, A.E., Troxler, T.G., Weller, D.E., 2018. Uncertainty in United States
1136 coastal wetland greenhouse gas inventorying. *Environ. Res. Lett.* 13, 115005.
1137 <https://doi.org/10.1088/1748-9326/aae157>
1138

1139 Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y., 2004. Gap-filling missing data in
1140 eddy covariance measurements using multiple imputation (MI) for annual estimations.
1141 *Agric. For. Meteorol.* 121, 93–111. [https://doi.org/10.1016/s0168-1923\(03\)00158-8](https://doi.org/10.1016/s0168-1923(03)00158-8)
1142

1143 Iwata, H., 2020a. FLUXNET-CH4 JP-Mse Mase rice paddy field.
1144 <https://doi.org/10.18140/FLX/1669647>
1145

1146 Iwata, H., Ueyama, M., Harazono, Y., 2020b. FLUXNET-CH4 US-Uaf University of Alaska,
1147 Fairbanks. <https://doi.org/10.18140/FLX/1669701>
1148

1149 Keenan, T.F., Migliavacca, M., Papale, D., Baldocchi, D., Reichstein, M., Torn, M., Wutzler,
1150 T., 2019. Widespread inhibition of daytime ecosystem respiration. *Nat Ecol Evol* 3,
1151 407–415. <https://doi.org/10.1038/s41559-019-0809-2>
1152

1153 Kim, Y., Johnson, M.S., Knox, S.H., Black, T.A., Dalmagro, H.J., Kang, M., Kim, J.,
1154 Baldocchi, D., 2020. Gap-filling approaches for eddy covariance methane fluxes: A
1155 comparison of three machine learning algorithms and a traditional method with
1156 principal component analysis. *Glob. Chang. Biol.* 26, 1499–1518.
1157 <https://doi.org/10.1111/gcb.14845>
1158

1159 Knox, S., Matthes, J., Verfaillie, J., Baldocchi, D., 2020. FLUXNET-CH4 US-Twt Twitchell
1160 Island. <https://doi.org/10.18140/FLX/1669700>
1161

1162 Knox, S.H., Bansal, S., McNicol, G., Schafer, K., Sturtevant, C., Ueyama, M., Valach, A.C.,
1163 Baldocchi, D., Delwiche, K., Desai, A.R., Euskirchen, E., Liu, J., Lohila, A., Malhotra,
1164 A., Melling, L., Riley, W., Runkle, B.R.K., Turner, J., Vargas, R., Zhu, Q., Alto, T.,
1165 Fluet-Chouinard, E., Goeckede, M., Melton, J.R., Sonnentag, O., Vesala, T., Ward,
1166 E., Zhang, Z., Feron, S., Ouyang, Z., Alekseychik, P., Aurela, M., Bohrer, G.,
1167 Campbell, D.I., Chen, J., Chu, H., Dalmagro, H.J., Goodrich, J.P., Gottschalk, P.,
1168 Hirano, T., Iwata, H., Jurasinski, G., Kang, M., Koepsch, F., Mammarella, I., Nilsson,
1169 M.B., Ono, K., Peichl, M., Peltola, O., Ryu, Y., Sachs, T., Sakabe, A., Sparks, J.,
1170 Tuittila, E.-S., Vourlitis, G.L., Wong, G.X., Windham-Myers, L., Poulter, B., Jackson,
1171 R.B., 2021. Identifying dominant environmental predictors of freshwater wetland
1172 methane fluxes across diurnal to seasonal time scales. *Glob. Chang. Biol.*
1173 <https://doi.org/10.1111/gcb.15661>
1174

1175 Knox, S.H., Jackson, R.B., Poulter, B., McNicol, G., Fluet-Chouinard, E., Zhang, Z.,
1176 Hugelius, G., Bousquet, P., Canadell, J.G., Saunio, M., Papale, D., Chu, H.,
1177 Keenan, T.F., Baldocchi, D., Torn, M.S., Mammarella, I., Trotta, C., Aurela, M.,
1178 Bohrer, G., Campbell, D.I., Cescatti, A., Chamberlain, S., Chen, J., Chen, W., Dengel,
1179 S., Desai, A.R., Euskirchen, E., Friborg, T., Gasbarra, D., Goded, I., Goeckede, M.,
1180 Heimann, M., Helbig, M., Hirano, T., Hollinger, D.Y., Iwata, H., Kang, M., Klatt, J.,
1181 Krauss, K.W., Kutzbach, L., Lohila, A., Mitra, B., Morin, T.H., Nilsson, M.B., Niu, S.,
1182 Noormets, A., Oechel, W.C., Peichl, M., Peltola, O., Reba, M.L., Richardson, A.D.,

1183 Runkle, B.R.K., Ryu, Y., Sachs, T., Schäfer, K.V.R., Schmid, H.P., Shurpali, N.,
1184 Sonntag, O., Tang, A.C.I., Ueyama, M., Vargas, R., Vesala, T., Ward, E.J.,
1185 Windham-Myers, L., Wohlfahrt, G., Zona, D., 2019. FLUXNET-CH₄ synthesis activity:
1186 Objectives, observations, and future directions. *Bull. Am. Meteorol. Soc.* 100, 2607–
1187 2632. <https://doi.org/10.1175/bams-d-18-0268.1>
1188
1189 Knox, S.H., Matthes, J.H., Sturtevant, C., Oikawa, P.Y., Verfaillie, J., Baldocchi, D., 2016.
1190 Biophysical controls on interannual variability in ecosystem-scale CO₂ and CH₄
1191 exchange in a California rice paddy. *J. Geophys. Res. Biogeosci.* 121, 978–1001.
1192 <https://doi.org/10.1002/2015jg003247>
1193
1194 Koebisch, F., Gottschalk, P., Beyer, F., Wille, C., Jurasinski, G., Sachs, T., 2020. The impact
1195 of occasional drought periods on vegetation spread and greenhouse gas exchange in
1196 rewetted fens. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375, 20190685.
1197 <https://doi.org/10.1098/rstb.2019.0685>
1198
1199 Kuleshov, V., Fenner, N., Ermon, S., 2018. Accurate Uncertainties for Deep Learning Using
1200 Calibrated Regression. *arXiv [cs.LG]*.
1201
1202 Kumar, V., Minz, S., 2014. Feature Selection: A literature review. *Smart Computing Review*
1203 4, 211–229. <https://doi.org/10.6029/smartcr.2014.03.007>
1204
1205 Laanbroek, H.J., 2010. Methane emission from natural wetlands: interplay between
1206 emergent macrophytes and soil microbial processes. A mini-review. *Ann. Bot.* 105,
1207 141–153. <https://doi.org/10.1093/aob/mcp201>
1208
1209 Lasslop, G., Reichstein, M., Kattge, J., Papale, D., 2008. Influences of observation errors in
1210 eddy flux data on inverse model parameter estimation. *Biogeosciences* 5, 1311–
1211 1324. <https://doi.org/10.5194/bg-5-1311-2008>
1212
1213 Lasslop, G., Reichstein, M., Papale, D., Richardson, A.D., Arneeth, A., Barr, A., Stoy, P.,
1214 Wohlfahrt, G., 2010. Separation of net ecosystem exchange into assimilation and
1215 respiration using a light response curve approach: critical issues and global

1216 evaluation. *Glob. Chang. Biol.* 16, 187–208. <https://doi.org/10.1111/j.1365->
1217 [2486.2009.02041.x](https://doi.org/10.1111/j.1365-2486.2009.02041.x)
1218
1219 Li, X., Wahlroos, O., Haapanala, S., Pumpanen, J., Vasander, H., Ojala, A., Vesala, T.,
1220 Mammarella, I., 2020. Carbon dioxide and methane fluxes from different surface
1221 types in a created urban wetland. *Biogeosciences* 17, 3409–3425.
1222 <https://doi.org/10.5194/bg-17-3409-2020>
1223
1224 Lipton, Z.C., Berkowitz, J., Elkan, C., 2015. A Critical Review of Recurrent Neural Networks
1225 for Sequence Learning. *arXiv [cs.LG]*.
1226
1227 Lohila, A., Aurela, M., Tuovinen, J.-P., Laurila, T., Hatakka, J., Rainne, J., Mäkelä, T., 2020.
1228 FLUXNET-CH4 FI-Lom Lompolojankka. <https://doi.org/10.18140/FLX/1669638>
1229
1230 Mammarella, I., Aslan, T., Burba, G., Cowan, N., Helfter, C., Herbst, M., Hörtnagl, L., Ibrom,
1231 A., Lucas-Moffat, A.M., Nicolini, G., Papale, D., Peltola, O., Rannik, Ü., Vitale, D.,
1232 Yeung, K., Nemitz, E., 2020. Protocol for non-CO₂ eddy covariance measurements,
1233 QA/QC, data processing and gap-filling. Readiness of ICOS for Necessities of
1234 integrated Global Observations (RINGO).
1235
1236 Matthes, J., Sturtevant, C., Oikawa, P., Chamberlain, S., Szutu, D., Ortiz, A., Verfaillie, J.,
1237 Baldocchi, D., 2020. FLUXNET-CH4 US-Myb Mayberry Wetland.
1238 <https://doi.org/10.18140/FLX/1669685>
1239
1240 Matthes, J.H., Sturtevant, C., Verfaillie, J., Knox, S., Baldocchi, D., 2014. Parsing the
1241 variability in CH₄ flux at a spatially heterogeneous wetland: Integrating multiple eddy
1242 covariance towers with high-resolution flux footprint analysis. *J. Geophys. Res.*
1243 *Biogeosci.* 119, 1322–1339. <https://doi.org/10.1002/2014jg002642>
1244
1245 McNicol, G., Knox, S.H., Guilderson, T.P., Baldocchi, D.D., Silver, W.L., 2020. Where old
1246 meets new: An ecosystem study of methanogenesis in a reflooded agricultural
1247 peatland. *Glob. Chang. Biol.* 26, 772–785. <https://doi.org/10.1111/gcb.14916>
1248

- 1249 McNicol, G., Sturtevant, C.S., Knox, S.H., Dronova, I., Baldocchi, D.D., Silver, W.L., 2017.
1250 Effects of seasonality, transport pathway, and spatial structure on greenhouse gas
1251 fluxes in a restored wetland. *Glob. Chang. Biol.* 23, 2768–2782.
1252 <https://doi.org/10.1111/gcb.13580>
1253
- 1254 Menzer, O., Moffat, A.M., Meiring, W., Lasslop, G., Schukat-Talamazzini, E.G., Reichstein,
1255 M., 2013. Random errors in carbon and water vapor fluxes assessed with Gaussian
1256 Processes. *Agric. For. Meteorol.* 178-179, 161–172.
1257 <https://doi.org/10.1016/j.agrformet.2013.04.024>
1258
- 1259 Miyata, A., Leuning, R., Denmead, O.T., Kim, J., Harazono, Y., 2000. Carbon dioxide and
1260 methane fluxes from an intermittently flooded paddy field. *Agric. For. Meteorol.* 102,
1261 287–303. [https://doi.org/10.1016/S0168-1923\(00\)00092-7](https://doi.org/10.1016/S0168-1923(00)00092-7)
1262
- 1263 Moffat, A.M., Papale, D., Reichstein, M., Hollinger, D.Y., Richardson, A.D., Barr, A.G.,
1264 Beckstein, C., Braswell, B.H., Churkina, G., Desai, A.R., Falge, E., Gove, J.H.,
1265 Heimann, M., Hui, D., Jarvis, A.J., Kattge, J., Noormets, A., Stauch, V.J., 2007.
1266 Comprehensive comparison of gap-filling techniques for eddy covariance net carbon
1267 fluxes. *Agric. For. Meteorol.* 147, 209–232.
1268 <https://doi.org/10.1016/j.agrformet.2007.08.011>
1269
- 1270 Moore, T.R., De Young, A., Bubier, J.L., Humphreys, E.R., Lafleur, P.M., Roulet, N.T., 2011.
1271 A multi-year record of methane flux at the Mer bleue bog, southern Canada.
1272 *Ecosystems* 14, 646–657. <https://doi.org/10.1007/s10021-011-9435-9>
1273
- 1274 Morin, T.H., Bohrer, G., Frasson, R.P. d. M., Naor-Azreli, L., Mesi, S., Stefanik, K.C.,
1275 Schäfer, K.V.R., 2014. Environmental drivers of methane fluxes from an urban
1276 temperate wetland park. *J. Geophys. Res. Biogeosci.* 119, 2188–2208.
1277 <https://doi.org/10.1002/2014jg002750>
1278
- 1279 Morin, T.H., Bohrer, G., Stefanik, K.C., Rey-Sanchez, A.C., Matheny, A.M., Mitsch, W.J.,
1280 2017. Combining eddy-covariance and chamber measurements to determine the
1281 methane budget from a small, heterogeneous urban floodplain wetland park. *Agric.*
1282 *For. Meteorol.* 237-238, 160–170. <https://doi.org/10.1016/j.agrformet.2017.01.022>

1283
1284 Muramatsu, K., Ono, K., Soyama, N., Thanyapraneedkul, J., Miyata, A., Mano, M., 2017.
1285 Determination of rice paddy parameters in the global gross primary production
1286 capacity estimation algorithm using 6 years of JP-MSE flux observation data. *Journal*
1287 *of Agricultural Meteorology* 73, 119–132. <https://doi.org/10.2480/agrmet.D-16-00017>
1288
1289 Nemitz, E., Mammarella, I., Ibrom, A., Aurela, M., Burba, G.G., Dengel, S., Gielen, B., Grelle,
1290 A., Heinesch, B., Herbst, M., Hörtnagl, L., Klemmedtsson, L., Lindroth, A., Lohila, A.,
1291 McDermitt, D.K., Meier, P., Merbold, L., Nelson, D., Nicolini, G., Nilsson, M.B.,
1292 Peltola, O., Rinne, J., Zahniser, M., 2018. Standardisation of eddy-covariance flux
1293 measurements of methane and nitrous oxide. *Int. Agrophys* 32, 517–549.
1294 <https://doi.org/10.1515/intag-2017-0042>
1295
1296 Neubauer, S.C., Megonigal, J.P., 2015. Moving beyond global warming potentials to quantify
1297 the climatic role of ecosystems. *Ecosystems* 18, 1000–1013.
1298 <https://doi.org/10.1007/s10021-015-9879-4>
1299
1300 Nilsson, M., Peichl, M., 2020. FLUXNET-CH4 SE-Deg Degero.
1301 <https://doi.org/10.18140/FLX/1669659>
1302
1303 Oikawa, P.Y., Sturtevant, C., Knox, S.H., Verfaillie, J., Huang, Y.W., Baldocchi, D.D., 2017.
1304 Revisiting the partitioning of net ecosystem exchange of CO₂ into photosynthesis and
1305 respiration with simultaneous flux measurements of ¹³CO₂ and CO₂, soil respiration
1306 and a biophysical model, CANVEG. *Agric. For. Meteorol.* 234-235, 149–163.
1307 <https://doi.org/10.1016/j.agrformet.2016.12.016>
1308
1309 Ooba, M., Hirano, T., Mogami, J.-I., Hirata, R., Fujinuma, Y., 2006. Comparisons of gap-
1310 filling methods for carbon flux dataset: A combination of a genetic algorithm and an
1311 artificial neural network. *Ecol. Modell.* 198, 473–486.
1312 <https://doi.org/10.1016/j.ecolmodel.2006.06.006>
1313
1314 Papale, D., 2020. Ideas and perspectives: enhancing the impact of the FLUXNET network of
1315 eddy covariance sites. *Biogeosciences* 17, 5587–5598. [https://doi.org/10.5194/bg-17-](https://doi.org/10.5194/bg-17-5587-2020)
1316 [5587-2020](https://doi.org/10.5194/bg-17-5587-2020)

1317
1318 Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter,
1319 C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van
1320 Ingen, C., Zhang, L., Amiro, B., Ammann, C., Arain, M.A., Ardö, J., Arkebauer, T.,
1321 Arndt, S.K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer,
1322 E., Marchesini, L.B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D.,
1323 Billesbach, D., Black, T.A., Blanken, P.D., Bohrer, G., Boike, J., Bolstad, P.V., Bonal,
1324 D., Bonnefond, J.-M., Bowling, D.R., Bracho, R., Brodeur, J., Brümmer, C.,
1325 Buchmann, N., Burban, B., Burns, S.P., Buysse, P., Cale, P., Cavagna, M., Cellier,
1326 P., Chen, S., Chini, I., Christensen, T.R., Cleverly, J., Collalti, A., Consalvo, C., Cook,
1327 B.D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P.S., D'Andrea, E., da Rocha,
1328 H., Dai, X., Davis, K.J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira,
1329 R.C., Delpierre, N., Desai, A.R., Di Bella, C.M., di Tommasi, P., Dolman, H.,
1330 Domingo, F., Dong, G., Dore, S., Duce, P., Dufrière, E., Dunn, A., Dušek, J., Eamus,
1331 D., Eichelmann, U., ElKhidir, H.A.M., Eugster, W., Ewenz, C.M., Ewers, B., Famulari,
1332 D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank,
1333 J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded,
1334 I., Goeckede, M., Goldstein, A.H., Gough, C.M., Goulden, M.L., Graf, A., Griebel, A.,
1335 Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B.U., Hanson,
1336 C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L.,
1337 Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W.,
1338 Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi,
1339 H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova,
1340 J., Kutsch, W.L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y.,
1341 Liddell, M., Limousin, J.-M., Lion, M., Liska, A.J., Lohila, A., López-Ballesteros, A.,
1342 López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S.,
1343 Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B.,
1344 Margolis, H.A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes,
1345 J.H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A.M.S., Merbold, L.,
1346 Meyer, W., Meyers, T., Miller, S.D., Minerbi, S., Moderow, U., Monson, R.K.,
1347 Montagnani, L., Moore, C.E., Moors, E., Moreaux, V., Moureaux, C., Munger, J.W.,
1348 Nakai, T., Neiryneck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto,
1349 M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J.E., Ourcival, J.-M., Papuga, S.A.,
1350 Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips,

1351 R.P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S.M., Rambal,
1352 S., Rannik, Ü., Raz-Yaseef, N., Reed, D., de Dios, V.R., Restrepo-Coupe, N.,
1353 Reverter, B.R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S.R., Sánchez-Cañete,
1354 E.P., Sanchez-Mejia, Z.M., Schmid, H.P., Schmidt, M., Schneider, K., Schrader, F.,
1355 Schroder, I., Scott, R.L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I.,
1356 Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens,
1357 R.M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N.,
1358 Thom, J., Tiedemann, F., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R.,
1359 van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J.,
1360 Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J.P., Walter-Shea, E.,
1361 Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S.,
1362 Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D.,
1363 Biraud, S., Torn, M., Papale, D., 2020. The FLUXNET2015 dataset and the ONEFlux
1364 processing pipeline for eddy covariance data. *Sci Data* 7, 225.
1365 <https://doi.org/10.1038/s41597-020-0534-3>
1366
1367 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
1368 Prettenhofer, P., Weiss, R., Dubourg, V., Others, 2011. Scikit-learn: Machine learning
1369 in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
1370
1371 Peltola, O., Vesala, T., Gao, Y., Rätty, O., Alekseychik, P., Aurela, M., Chojnicki, B., Desai,
1372 A.R., Dolman, A.J., Euskirchen, E.S., Friborg, T., Göckede, M., Helbig, M.,
1373 Humphreys, E., Jackson, R.B., Jocher, G., Joos, F., Klatt, J., Knox, S.H., Kowalska,
1374 N., Kutzbach, L., Lienert, S., Lohila, A., Mammarella, I., Nadeau, D.F., Nilsson, M.B.,
1375 Oechel, W.C., Peichl, M., Pypker, T., Quinton, W., Rinne, J., Sachs, T., Samson, M.,
1376 Schmid, H.P., Sonntag, O., Wille, C., Zona, D., Aalto, T., 2019. Monthly gridded
1377 data product of northern wetland methane emissions based on upscaling eddy
1378 covariance observations. *Earth Syst. Sci. Data* 11, 1263–1289.
1379 <https://doi.org/10.5194/essd-11-1263-2019>
1380
1381 Platt, J.C., 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to
1382 Regularized Likelihood Methods, in: *Advances in Large Margin Classifiers*.
1383

1384 Poffenbarger, H.J., Needelman, B.A., Megonigal, J.P., 2011. Salinity Influence on Methane
1385 Emissions from Tidal Marshes. *Wetlands* 31, 831–842.
1386 <https://doi.org/10.1007/s13157-011-0197-0>
1387

1388 Pohlert, T., 2014. The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR).
1389

1390 R Core Team, 2021. R: A Language and Environment for Statistical Computing.
1391

1392 Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C.,
1393 Buchmann, N., Gilmanov, T., Granier, A., Grunwald, T., Havrankova, K., Ilvesniemi,
1394 H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers,
1395 T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M.,
1396 Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., Valentini, R., 2005. On
1397 the separation of net ecosystem exchange into assimilation and ecosystem
1398 respiration: review and improved algorithm. *Glob. Chang. Biol.* 11, 1424–1439.
1399 <https://doi.org/10.1111/j.1365-2486.2005.001002.x>
1400

1401 Rey-Sanchez, A.C., Morin, T.H., Stefanik, K.C., Wrighton, K., Bohrer, G., 2018. Determining
1402 total emissions and environmental drivers of methane flux in a Lake Erie estuarine
1403 marsh. *Ecol. Eng.* 114, 7–15. <https://doi.org/10.1016/j.ecoleng.2017.06.042>
1404

1405 Richardson, A.D., Aubinet, M., Barr, A.G., Hollinger, D.Y., Ibrom, A., Lasslop, G., Reichstein,
1406 M., 2012. Uncertainty Quantification, in: Aubinet, M., Vesala, T., Papale, D. (Eds.),
1407 Eddy Covariance: A Practical Guide to Measurement and Data Analysis. Springer
1408 Netherlands, Dordrecht, pp. 173–209.
1409

1410 Richardson, A.D., Hollinger, D.Y., 2007. A method to estimate the additional uncertainty in
1411 gap-filled NEE resulting from long gaps in the CO₂ flux record. *Agric. For. Meteorol.*
1412 147, 199–208. <https://doi.org/10.1016/j.agrformet.2007.06.004>
1413

1414 Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S.,
1415 Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F.,
1416 Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial,

1417 hierarchical, or phylogenetic structure. *Ecography* 40, 913–929.
1418 <https://doi.org/10.1111/ecog.02881>
1419
1420 Rojas, R., 2013. *Neural Networks: A Systematic Introduction*. Springer Science & Business
1421 Media.
1422
1423 Rosentreter, J.A., Borges, A.V., Deemer, B.R., Holgerson, M.A., Liu, S., Song, C., Melack, J.,
1424 Raymond, P.A., Duarte, C.M., Allen, G.H., Olefeldt, D., Poulter, B., Battin, T.I., Eyre,
1425 B.D., 2021. Half of global methane emissions come from highly variable aquatic
1426 ecosystem sources. *Nat. Geosci.* 14, 225–230. [https://doi.org/10.1038/s41561-021-](https://doi.org/10.1038/s41561-021-00715-2)
1427 [00715-2](https://doi.org/10.1038/s41561-021-00715-2)
1428
1429 Runkle, B.R.K., Suvočarev, K., Reba, M.L., Reavis, C.W., Smith, S.F., Chiu, Y.-L., Fong, B.,
1430 2019. Methane Emission Reductions from the Alternate Wetting and Drying of Rice
1431 Fields Detected Using the Eddy Covariance Method. *Environ. Sci. Technol.* 53, 671–
1432 681. <https://doi.org/10.1021/acs.est.8b05535>
1433
1434 Russell, S.J., Norvig, P., 1995. *Artificial Intelligence: A Modern Approach*. Prentice Hall.
1435
1436 Saunois, M., Stavert, A.R., Poulter, B., Bousquet, P., Canadell, J.G., Jackson, R.B.,
1437 Raymond, P.A., Dlugokencky, E.J., Houweling, S., Patra, P.K., Ciais, P., Arora, V.K.,
1438 Bastviken, D., Bergamaschi, P., Blake, D.R., Brailsford, G., Bruhwiler, L., Carlson,
1439 K.M., Carrol, M., Castaldi, S., Chandra, N., Crevoisier, C., Crill, P.M., Covey, K.,
1440 Curry, C.L., Etiope, G., Frankenberg, C., Gedney, N., Hegglin, M.I., Höglund-
1441 Isaksson, L., Hugelius, G., Ishizawa, M., Ito, A., Janssens-Maenhout, G., Jensen,
1442 K.M., Joos, F., Kleinen, T., Krummel, P.B., Langenfelds, R.L., Laruelle, G.G., Liu, L.,
1443 Machida, T., Maksyutov, S., McDonald, K.C., McNorton, J., Miller, P.A., Melton, J.R.,
1444 Morino, I., Müller, J., Murguía-Flores, F., Naik, V., Niwa, Y., Noce, S., O'Doherty, S.,
1445 Parker, R.J., Peng, C., Peng, S., Peters, G.P., Prigent, C., Prinn, R., Ramonet, M.,
1446 Regnier, P., Riley, W.J., Rosentreter, J.A., Segers, A., Simpson, I.J., Shi, H., Smith,
1447 S.J., Steele, L.P., Thornton, B.F., Tian, H., Tohjima, Y., Tubiello, F.N., Tsuruta, A.,
1448 Viovy, N., Voulgarakis, A., Weber, T.S., van Weele, M., van der Werf, G.R., Weiss,
1449 R.F., Worthy, D., Wunch, D., Yin, Y., Yoshida, Y., Zhang, W., Zhang, Z., Zhao, Y.,
1450 Zheng, B., Zhu, Q., Zhu, Q., Zhuang, Q., 2020. The global methane budget 2000–

1451 2017. *Earth Syst. Sci. Data* 12, 1561–1623. <https://doi.org/10.5194/essd-12-1561->
1452 2020
1453
1454 Schuurmans, E.D., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *J.*
1455 *Mach. Learn. Res.* 7, 1–30.
1456
1457 Sonnentag, O., Helbig, M., 2020. FLUXNET-CH4 CA-SCB Scotty Creek Bog.
1458 <https://doi.org/10.18140/FLX/1669613>
1459
1460 Sturtevant, C., Ruddell, B.L., Knox, S.H., Verfaillie, J., Matthes, J.H., Oikawa, P.Y.,
1461 Baldocchi, D., 2016. Identifying scale-emergent, nonlinear, asynchronous processes
1462 of wetland methane exchange. *J. Geophys. Res. Biogeosci.* 121, 188–204.
1463 <https://doi.org/10.1002/2015jg003054>
1464
1465 Taoka, T., Iwata, H., Hirata, R., Takahashi, Y., Miyabara, Y., Itoh, M., 2020. Environmental
1466 controls of diffusive and ebullitive methane emissions at a subdaily time scale in the
1467 littoral zone of a midlatitude shallow lake. *J. Geophys. Res. Biogeosci.* 125.
1468 <https://doi.org/10.1029/2020jg005753>
1469
1470 Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram.
1471 *J. Geophys. Res.*, WMO TD-732 106, 7183–7192.
1472 <https://doi.org/10.1029/2000jd900719>
1473
1474 Taylor, R., 1990. Interpretation of the Correlation Coefficient: A Basic Review. *J. Diagn. Med.*
1475 *Sonogr.* 6, 35–39. <https://doi.org/10.1177/875647939000600106>
1476
1477 Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc.*
1478 *Series B Stat. Methodol.* 58, 267–288. <https://doi.org/10.1111/j.2517->
1479 [6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)
1480
1481 Tramontana, G., Jung, M., Schwalm, C.R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein,
1482 M., Arain, M.A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S.,
1483 Wolf, S., Papale, D., 2016. Predicting carbon dioxide and energy fluxes across global

1484 FLUXNET sites with regression algorithms. *Biogeosciences* 13, 4291–4313.
1485 <https://doi.org/10.5194/bg-13-4291-2016>
1486
1487 Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T.F., Camps-Valls, G.,
1488 Ogee, J., Verrelst, J., Papale, D., 2020. Partitioning net carbon dioxide fluxes into
1489 photosynthesis and respiration using neural networks. *Glob. Chang. Biol.* 26, 5235–
1490 5253. <https://doi.org/10.1111/gcb.15203>
1491
1492 Treat, C.C., Bloom, A.A., Marushchak, M.E., 2018. Nongrowing season methane emissions-
1493 a significant component of annual emissions across northern ecosystems. *Glob.*
1494 *Chang. Biol.* 24, 3331–3343. <https://doi.org/10.1111/gcb.14137>
1495
1496 Trifunovic, B., Vázquez-Lule, A., Capocci, M., Seyfferth, A.L., Moffat, C., Vargas, R., 2020.
1497 Carbon dioxide and methane emissions from temperate salt marsh tidal creek. *J.*
1498 *Geophys. Res. Biogeosci.*, NOAA National Estuarine Research Reserve, Central
1499 Data Management Office, Baruch Marine Laboratory, University of South Carolina
1500 125, 84. <https://doi.org/10.1029/2019jg005558>
1501
1502 Tuovinen, J.-P., Aurela, M., Hatakka, J., Räsänen, A., Virtanen, T., Mikola, J., Ivakhov, V.,
1503 Kondratyev, V., Laurila, T., 2019. Interpreting eddy covariance data from
1504 heterogeneous Siberian tundra: land-cover-specific methane fluxes and spatial
1505 representativeness. *Biogeosciences* 16, 255–274. [https://doi.org/10.5194/bg-16-255-](https://doi.org/10.5194/bg-16-255-2019)
1506 [2019](https://doi.org/10.5194/bg-16-255-2019)
1507
1508 Turetsky, M.R., Kotowska, A., Bubier, J., Dise, N.B., Crill, P., Hornibrook, E.R.C., Minkinen,
1509 K., Moore, T.R., Myers-Smith, I.H., Nykänen, H., Olefeldt, D., Rinne, J., Saarnio, S.,
1510 Shurpali, N., Tuittila, E.-S., Waddington, J.M., White, J.R., Wickland, K.P., Wilmking,
1511 M., 2014. A synthesis of methane emissions from 71 northern, temperate, and
1512 subtropical wetlands. *Glob. Chang. Biol.* 20, 2183–2197.
1513 <https://doi.org/10.1111/gcb.12580>
1514
1515 Ueyama, M., Hirano, T., Kominami, Y., 2020a. FLUXNET-CH4 JP-BBY Bibai bog.
1516 <https://doi.org/10.18140/FLX/1669646>
1517

1518 Ueyama, M., Yazaki, T., Hirano, T., Futakuchi, Y., Okamura, M., 2020b. Environmental
1519 controls on methane fluxes in a cool temperate bog. *Agric. For. Meteorol.* 281,
1520 107852. <https://doi.org/10.1016/j.agrformet.2019.107852>
1521

1522 Valach, A., Szutu, D., Eichelmann, E., Knox, S., Verfaillie, J., Baldocchi, D., 2020.
1523 FLUXNET-CH4 US-Tw1 Twitchell Wetland West Pond.
1524 <https://doi.org/10.18140/FLX/1669696>
1525

1526 Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual: (Python Documentation
1527 Manual Part 2). CreateSpace Independent Publishing Platform.
1528

1529 Vargas, R., Sánchez-Cañete P., E., Serrano-Ortiz, P., Curiel Yuste, J., Domingo, F., López-
1530 Ballesteros, A., Oyonarte, C., 2018. Hot-Moments of Soil CO₂ Efflux in a Water-
1531 Limited Grassland. *Soil Systems* 2, 47. <https://doi.org/10.3390/soilsystems2030047>
1532

1533 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł.U.,
1534 Polosukhin, I., 2017. Attention is All you Need, in: Guyon, I., Luxburg, U.V., Bengio,
1535 S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), *Advances in Neural*
1536 *Information Processing Systems* 30. Curran Associates, Inc., pp. 5998–6008.
1537

1538 Vázquez-Lule, A., Vargas, R., 2021. Biophysical drivers of net ecosystem and methane
1539 exchange across phenological phases in a tidal salt marsh. *Agric. For. Meteorol.* 300,
1540 108309. <https://doi.org/10.1016/j.agrformet.2020.108309>
1541

1542 Vesala, T., Tuittila, E.-S., Mammarella, I., Alekseychik, P., 2020a. FLUXNET-CH4 FI-Si2
1543 Siikaneva-2 Bog. <https://doi.org/10.18140/FLX/1669639>
1544

1545 Vesala, T., Tuittila, E.-S., Mammarella, I., Rinne, J., 2020b. FLUXNET-CH4 FI-Sii Siikaneva.
1546 <https://doi.org/10.18140/FLX/1669640>
1547

1548 Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D.,
1549 Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M.,
1550 Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E.,
1551 Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J.,

1552 Cimirman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro,
1553 A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0:
1554 fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272.
1555 <https://doi.org/10.1038/s41592-019-0686-2>
1556

1557 Vitale, D., Bilancia, M., Papale, D., 2019. Modelling random uncertainty of eddy covariance
1558 flux measurements. *Stoch. Environ. Res. Risk Assess.* 33, 725–746.
1559 <https://doi.org/10.1007/s00477-019-01664-4>
1560

1561 Vitale, D., Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF),
1562 University of Tuscia, via San Camillo de Lellis, 01100 Viterbo, Italy, Bilancia, M.,
1563 Papale, D., Ionian Department of Law, Economics and Environment, University of
1564 Bari Aldo Moro, Via Lago Maggiore angolo Via Ancona, 74121 Taranto, Italy,
1565 Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF),
1566 University of Tuscia, via San Camillo de Lellis, 01100 Viterbo, Italy, 2018. A Multiple
1567 Imputation Strategy for Eddy Covariance Data. *J. Environ. Inf.* 1–20.
1568 <https://doi.org/10.3808/jei.201800391>
1569

1570 Vourlitis, G., Dalmagro, H., de S. Nogueira, J., Johnson, M., Arruda, P., 2020. FLUXNET-
1571 CH4 BR-Npw Northern Pantanal Wetland. <https://doi.org/10.18140/FLX/1669368>
1572

1573 Vuichard, N., Papale, D., 2015. Filling the gaps in meteorological continuous data measured
1574 at FLUXNET sites with ERA-Interim reanalysis. *Earth Syst. Sci. Data* 7, 157–171.
1575 <https://doi.org/10.5194/essd-7-157-2015>
1576

1577 Wania, R., Melton, J.R., Hodson, E.L., Poulter, B., Ringeval, B., Spahni, R., Bohn, T., Avis,
1578 C.A., Chen, G., Eliseev, A.V., Hopcroft, P.O., Riley, W.J., Subin, Z.M., Tian, H., van
1579 Bodegom, P.M., Kleinen, T., Yu, Z.C., Singarayer, J.S., Zürcher, S., Lettenmaier,
1580 D.P., Beerling, D.J., Denisov, S.N., Prigent, C., Papa, F., Kaplan, J.O., 2013. Present
1581 state of global wetland extent and wetland methane modelling: methodology of a
1582 model inter-comparison project (WETCHIMP). *Geosci. Model Dev.* 6, 617–641.
1583 <https://doi.org/10.5194/gmd-6-617-2013>
1584

1585 Whiting, G.J., Chanton, J.P., 1993. Primary production control of methane emission from
1586 wetlands. *Nature* 364, 794–795. <https://doi.org/10.1038/364794a0>
1587

1588 Wutzler, T., Lucas-Moffat, A., Migliavacca, M., Knauer, J., Sickel, K., Šigut, L., Menzer, O.,
1589 Reichstein, M., 2018. Basic and extensible post-processing of eddy covariance flux
1590 data with REddyProc. *Biogeosciences* 15, 5015–5030. [https://doi.org/10.5194/bg-15-](https://doi.org/10.5194/bg-15-5015-2018)
1591 5015-2018
1592

1593 Yang, W.H., McNicol, G., Teh, Y.A., Estera-Molina, K., Wood, T.E., Silver, W.L., 2017.
1594 Evaluating the classical versus an emerging conceptual model of peatland methane
1595 dynamics: Peatland methane dynamics. *Global Biogeochem. Cycles* 31, 1435–1453.
1596 <https://doi.org/10.1002/2017gb005622>
1597

1598 Yvon-Durocher, G., Allen, A.P., Bastviken, D., Conrad, R., Gudas, C., St-Pierre, A., Thanh-
1599 Duc, N., del Giorgio, P.A., 2014. Methane fluxes show consistent temperature
1600 dependence across microbial to ecosystem scales. *Nature* 507, 488–491.
1601 <https://doi.org/10.1038/nature13164>
1602

1603 Zadrozny, B., Elkan, C., 2002. Transforming classifier scores into accurate multiclass
1604 probability estimates, in: *Proceedings of the Eighth ACM SIGKDD International*
1605 *Conference on Knowledge Discovery and Data Mining, KDD '02*. Association for
1606 Computing Machinery, New York, NY, USA, pp. 694–699.
1607