Learning Evaporative Fraction with Memory

Wenli Zhao[1, 2, 3], Alexander J. Winkler[1, 3], Markus Reichstein[1, 3], Rene Orth[1, 4], Pierre Gentine[2]

[1] Max Planck Institute for Biogeochemistry, Jena, Germany.

[2] Columbia University, New York, USA.

[3] ELLIS Jena Unit, Jena, Germany.

[4] Faculty of Environment and Natural Resources, University of Freiburg, Germany

Corresponding author: Wenli Zhao (dr.wenli.zhao.pku@gmail.com)

Contents of this file

Introduction

This supplementary information contains two Texts.

Text S1.

Data Preprocessing Procedures of the Eddy-Covariance Dataset

The original sampling frequency of the data is half hourly. The data filter procedure can be summarized as follows: First, to reduce the noise in nighttime measurements, the original data is filtered with sensible heat flux > 5 W/m^2 and shortwave incoming radiation > 50 W/m^2 to select the daytime data only. Then, the original data is averaged to daily scale value (precipitation is calculated as the daily sum). Secondly, we only keep days with a fraction of good quality data > 0.8. The gaps in the time series for input features were interpolated using established methods(Reichstein et al., 2005; Vuichard and Papale, 2015). We also visually checked site by site to ensure that the signal-to-noise ratio is acceptable. Note that all the half hourly *LE* data from the eddy covariance sites is corrected to achieve energy balance closure using the Bowen ratio method(Twine et al., 2000). Due to the data limitation, only the shallowest soil moisture measurements were used for comparison with the evaporative fraction prediction dynamics during the dry-down periods.

Text S2.

Model Interpretations – Expected Gradients (EG)

To interpret the contributions of input features to daily EF predictions, we employed the Expected Gradients (EG) method (Erion et al., 2021). EG is an extension of Integrated Gradients (IG), designed to overcome the sensitivity of IG to baseline input selection. Instead of requiring a single fixed baseline, EG samples baseline inputs from an underlying data distribution (e.g., the training dataset). The attribution score for the i-th feature is then obtained by integrating the gradients across all possible baseline inputs, weighted by the distribution density (Erion et al., 2021; Jiang et al., 2022; Sundararajan et al., 2017).

The EG method could unbox the LSTM-based machine learning model and trace back the specific contributions of the inputs and assign an importance score to each feature at each time prior to the predictions. A large positive EG score could indicate that the feature substantially increases the Evaporative Fraction predictions (e.g., that precipitation at the most proximate time may contribute more to current Evaporative Fraction projections than precipitation at an earlier time.). A large negative EG score indicates that the feature decreases the EF predictions. An EG score close to zero indicates little influence on the EF predictions. This way, our model could not only show the general feature importance but could also show the different feature importance at each time step prior the predictions. More specifically, it implies that temporal length of the input features will be considered for the EF predictions for different kinds of PFTs, in which hint the response of plant with different rooting depths during specific extreme events or environmental conditions, e.g., droughts with different severity level.

The IG score for the input feature x (e.g., the specific contribution of precipitation at the ith time step) is formulated as:

$$\emptyset_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

Where $\frac{\partial F(x'+\alpha\times(x-x'))}{\partial x_i}$ denotes the local gradient of the network F at a point interpolated from a baseline input (x' $when$ $\alpha = 0$), which is meant to represent the "absence" of the feature input, to the target input (x, when $\alpha = 1$). Note that the IG value is completeness and add up to the difference between the output of F at the target input x and the baseline input x', i.e., $\sum_i \emptyset_i(x) = F(x) - F(x')$. Therefore, the model output can be decomposed into the sum of features' individual contributions, and it enables us to examine the contribution of a group of features by summing up their individual IG scores.

Due to the integral, the original definition of IG is incalculable. Therefore, the implementation of the method in practice uses approximated value by replacing the integral with the summation:

$$\emptyset_i^{approx}(x) = (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

where m defines a number of interpolation steps.

Formally, given a baseline distribution D, the EG of the i-th feature can be calculated by integrating the gradients with all possible baseline inputs $x' \in D$ weighted by the density function p_D, which is expressed as:

$$\phi_i^{\mathrm{EG}}(f,x) = \int_{\mathrm{x}'} (\phi_i^{\mathrm{IG}}(f,x,\mathrm{x}') \times p_D(\mathrm{x}')d\mathrm{x}')$$

The formulation is therefore called expected gradients, and can be reformulated as

$$\phi_i^{\mathrm{EG}}(f,x) = E_{\mathrm{x}'\sim D,\alpha\sim U(0,1)}[(x_i - \mathrm{x}'_i) \times \frac{\partial f(\mathrm{x}' + \alpha(x - \mathrm{x}'))}{\partial \mathrm{x}'_i}]$$

where U(0,1) denotes the uniform distribution between 0 and 1.

In this study, we use the library captum (https://github.com/pytorch/captum) to obtain the EG scores for each feature at each time step.

References

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I.: Improving performance of deep learning models with axiomatic attribution priors and expected gradients, Nat Mach Intell, 3, 620–631, https://doi.org/10.1038/s42256-021-00343-w, 2021.

Jiang, S., Bevacqua, E., and Zscheischler, J.: River flooding mechanisms and their changes in Europe revealed by explainable machine learning, Hydrol Earth Syst Sc, 26, 6339–6359, https://doi.org/10.5194/hess-26-6339-2022, 2022.

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

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, http://arxiv.org/abs/1703.01365, 12 June 2017.

Twine, T. E., Kustas, W., Norman, J., Cook, D., Houser, Pr., Meyers, T., Prueger, J., Starks, P., and Wesely, M.: Correcting eddy-covariance flux underestimates over a grassland, Agricultural and forest meteorology, 103, 279–300, 2000.

Vuichard, N. and Papale, D.: Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis, Earth System Science Data, 7, 157–171, 2015.

Learning Evaporative Fraction with Memory

Wenli Zhao[1, 2, 3], Alexander J. Winkler[1, 3], Markus Reichstein[1, 3], Rene Orth[1, 4], Pierre Gentine[2]

[1] Max Planck Institute for Biogeochemistry, Jena, Germany.

[2] Columbia University, New York, USA.

[3] ELLIS Jena Unit, Jena, Germany.

[4] Faculty of Environment and Natural Resources, University of Freiburg, Germany

Corresponding author: Wenli Zhao (dr.wenli.zhao.pku@gmail.com)

Contents of this file

Figure S1 to S4

Introduction

This supplementary information contains one Text and Four Figures.

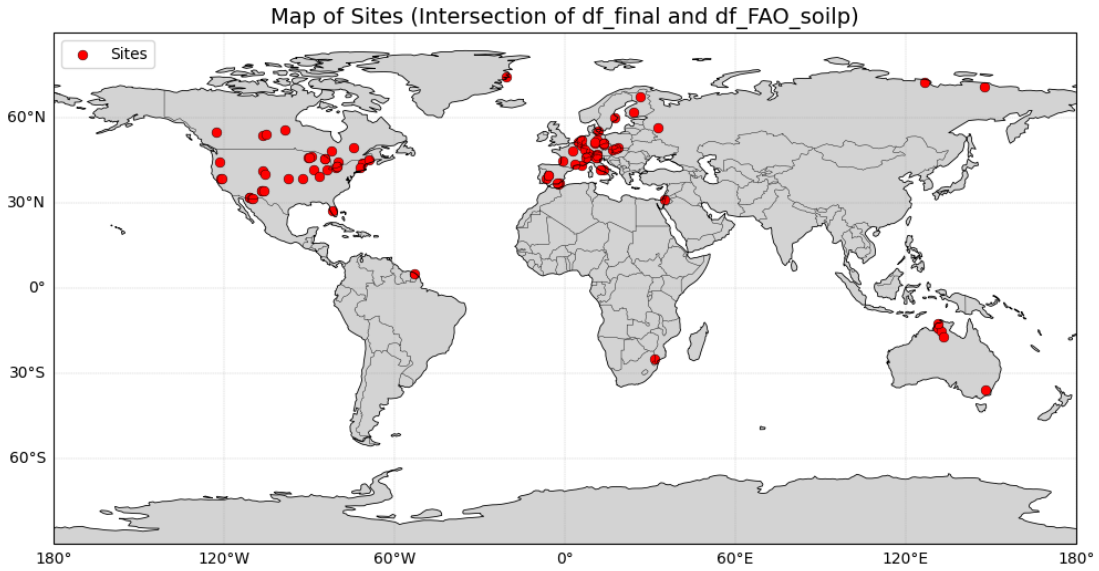

Figure S1. Map of Sites Distribution of the combined Fluxnet2015, AmeriFlux, ICOS networks. The sites which have more than 5 years' observations are used for model training.

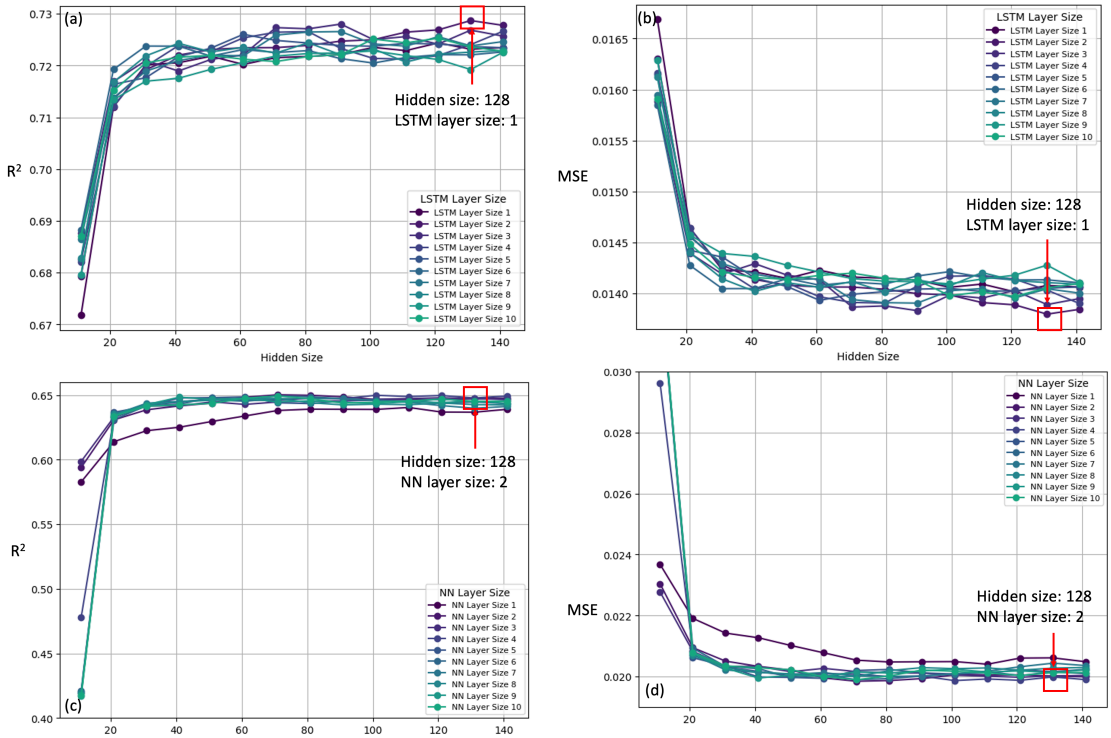

Figure S2. Optimal architecture of machine learning model. Note we sampled the entire dataset to speed up identifying the optimal model structure.

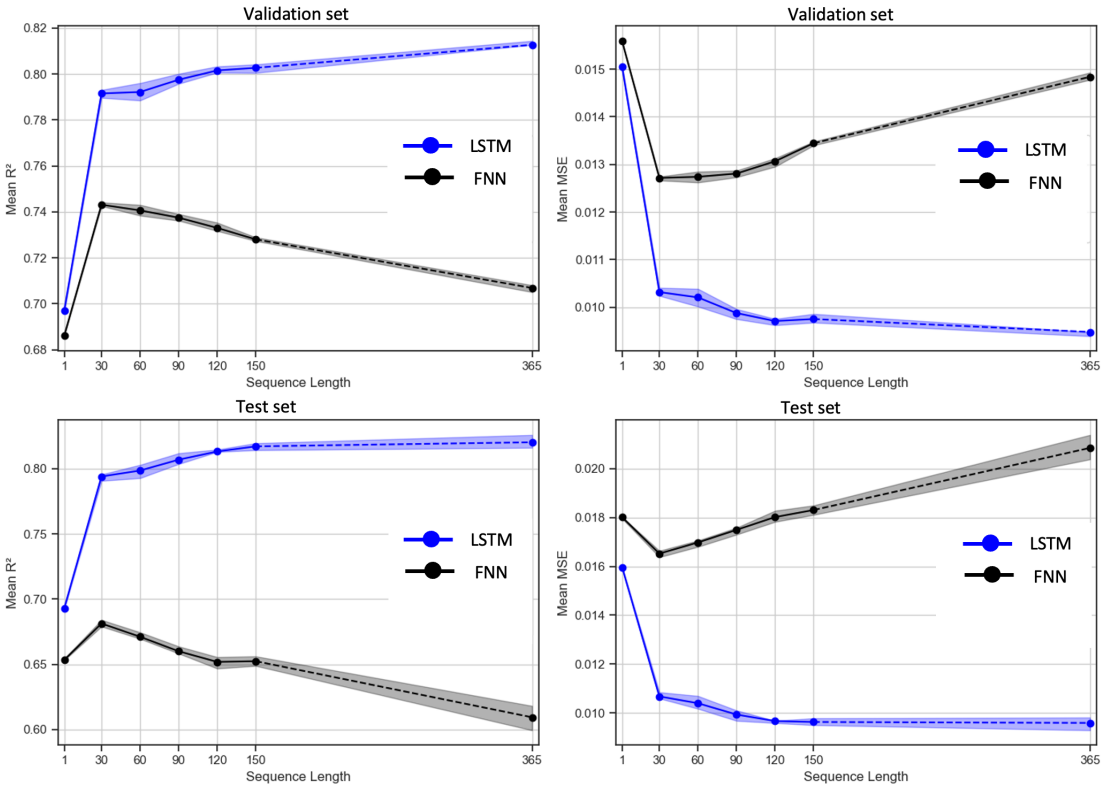

Figure S3. Optimal sequence length of machine learning model.

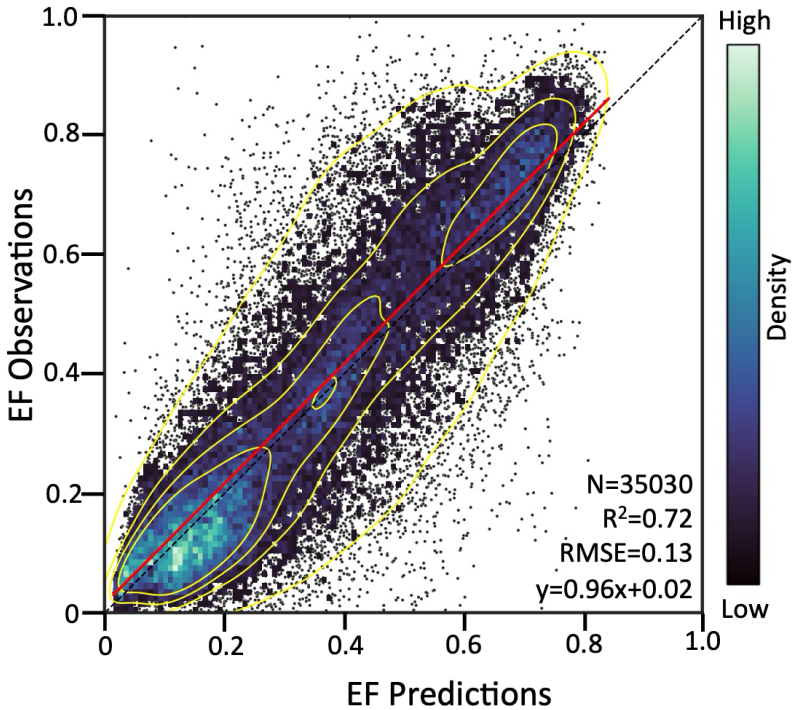

Figure S4. Model performance. The model performance with 365 days' time series data as input. The EF predictions shown here are based on the ensemble mean EF predictions of 20 repeated models. The density of scatter points is represented using shading colors. The diagonal black dashed line depicts the 1:1 line and the red solid line depicts the linear regression line. Note that N, R^2, and RMSE represent the number of points, coefficient of determination, and root mean square error, respectively. All the

metrics are calculated using the test set data. The dataset is split into training, validation, and test sets based on time periods: the last year of each site is used for testing, the second-to-last year for validation, and the remaining data for training.

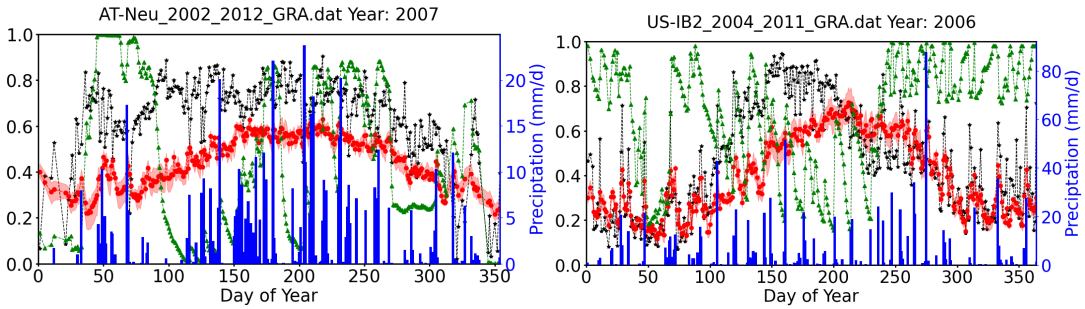

Figure S5. Grassland sites receiving sufficient precipitation. Blue bars show the observed daily sum precipitation (P), black curves show EF observations, red curves show EF predictions, green curves show soil moisture percentiles of the shallowest soil depth. The x-axis represents the day of year (DOY) of the whole year. The daily volumetric water content values are converted into percentiles, indicating the fraction of daily values lower than a specific value. Thus, the 100th percentile (or a percentile value of 1) represents the wettest soil conditions observed at a specific site throughout the study period, and the 0 percentile signifies the driest soil conditions. The memory length is set as 365 days for the machine learning model. The EF predictions here are using the ensemble mean EF predictions of 20 models with different initializations. Shaded areas represent regions of predictions uncertainty in the 25%-75% quartiles of these 20 repeat training models.

Learning Evaporative Fraction with Memory

Wenli Zhao[1, 2, 3], Alexander J. Winkler[1, 3], Markus Reichstein[1, 3], Rene Orth[1, 4], Pierre Gentine[2]

[1] Max Planck Institute for Biogeochemistry, Jena, Germany.

[2] Columbia University, New York, USA.

[3] ELLIS Jena Unit, Jena, Germany.

[4] Faculty of Environment and Natural Resources, University of Freiburg, Germany

Corresponding author: Wenli Zhao (dr.wenli.zhao.pku@gmail.com)

Contents of this file

Introduction

This supplementary information contains one Table.

Table S1. Information on the eddy covariance sites from the combined ICOS, AmeriFlux, FLUXNET 2015 tier 1 dataset

site	IGBP	latitude	longitude	period	DOI
AT-Neu	GRA	47,12	11,32	2002-2012	https://doi.org/10.18140/FLX/1440121
AU-DaP	GRA	-14,06	131,32	2007-2013	https://doi.org/10.18140/FLX/1440123
AU-DaS	SAV	-14,16	131,39	2008-2014	https://doi.org/10.18140/FLX/1440122
AU-Dry	SAV	-15,26	132,37	2008-2014	https://doi.org/10.18140/FLX/1440197
AU-How	WSA	-12,49	131,15	2001-2014	https://doi.org/10.18140/FLX/1440125
AU-Stp	GRA	-17,15	133,35	2008-2014	https://doi.org/10.18140/FLX/1440204
AU-Tum	EBF	-35,66	148,15	2001-2014	https://doi.org/10.18140/FLX/1440126
BE-Bra	MF	51,31	4,52	1996-2020	https://doi.org/10.18160/2G60-ZHAK
BE-Vie	MF	50,31	6	1996-2020	https://doi.org/10.18160/2G60-ZHAK
CA-Cbo	DBF	44,32	-79,93	1994-2020	https://doi.org/10.17190/AMF/1854365
CA-Gro	MF	48,22	-82,16	2003-2014	https://doi.org/10.18140/FLX/1440034
CA-LP1	ENF	55,11	-122,84	2007-2020	https://doi.org/10.17190/AMF/1832155
CA-Man	ENF	55,88	-98,48	1994-2008	https://doi.org/10.18140/FLX/1440035
CA-Oas	DBF	53,63	-106,2	1996-2010	https://doi.org/10.18140/FLX/1440043
CA-Obs	ENF	53,99	-105,12	1997-2010	https://doi.org/10.18140/FLX/1440044
CA-Qfo	ENF	49,69	-74,34	2003-2010	https://doi.org/10.18140/FLX/1440045
CA-TP1	ENF	42,66	-80,56	2002-2014	https://doi.org/10.18140/FLX/1440050
CA-TP3	ENF	42,71	-80,35	2002-2014	https://doi.org/10.18140/FLX/1440052
CA-TP4	ENF	42,71	-80,36	2002-2014	https://doi.org/10.18140/FLX/1440053
CH-Aws	GRA	46,58	9,79	2006-2020	https://doi.org/10.18160/2G60-ZHAK
CH-Cha	GRA	47,21	8,41	2005-2020	https://doi.org/10.18160/2G60-ZHAK
CH-Dav	ENF	46,82	9,86	1997-2020	https://doi.org/10.18160/2G60-ZHAK
CH-Fru	GRA	47,12	8,54	2005-2020	https://doi.org/10.18160/2G60-ZHAK
CH-Lae	MF	47,48	8,37	2004-2020	https://doi.org/10.18160/2G60-ZHAK
CZ-BK1	ENF	49,5	18,54	2004-2020	https://doi.org/10.18140/FLX/1440143
CZ-Lnz	MF	48,68	16,95	2015-2020	https://doi.org/10.18160/2G60-ZHAK
CZ-Stn	DBF	49,04	17,97	2010-2020	https://doi.org/10.18160/2G60-ZHAK
DE-Gri	GRA	50,95	13,51	2004-2020	https://doi.org/10.18160/2G60-ZHAK
DE-Hai	DBF	51,08	10,45	2000-2020	https://doi.org/10.18160/2G60-ZHAK
DE-HoH	DBF	52,09	11,22	2015-2020	https://doi.org/10.18160/2G60-ZHAK
DE-Hzd	DBF	50,96	13,49	2010-2020	https://doi.org/10.18160/2G60-ZHAK
DE-Lnf	DBF	51,33	10,37	2002-2012	https://doi.org/10.18140/FLX/1440150
DE-Obe	ENF	50,78	13,72	2008-2020	https://doi.org/10.18160/2G60-ZHAK
DE-Tha	ENF	50,96	13,57	1996-2020	https://doi.org/10.18160/2G60-ZHAK
DK-Sor	DBF	55,49	11,64	1996-2014	https://doi.org/10.18140/FLX/1440155
ES-Abr	SAV	38,7	-6,79	2015-2020	https://doi.org/10.18160/2G60-ZHAK
ES-Agu	OSH	36,94	-2,03	2006-2020	https://doi.org/10.18160/2G60-ZHAK
ES-Amo	OSH	36,83	-2,25	2007-2012	https://doi.org/10.18140/FLX/1440156
ES-LJu	OSH	36,93	-2,75	2004-2013	https://doi.org/10.18140/FLX/1440157
ES-LM1	SAV	39,94	-5,78	2014-2020	https://doi.org/10.18160/2G60-ZHAK
ES-LM2	SAV	39,93	-5,78	2014-2020	https://doi.org/10.18160/2G60-ZHAK
FI-Hyy	ENF	61,85	24,3	1996-2020	https://doi.org/10.18160/2G60-ZHAK

IGBP: The land cover classification defined by the International Geosphere Biosphere Programme (IGBP)

Continued Table S1. Information on the eddy covariance sites from the combined ICOS, AmeriFlux, FLUXNET 2015 tier 1 dataset

site	IGBP	latitude	longitude	period	DOI
FI-Sod	ENF	67,36	26,64	2001-2014	https://doi.org/10.18140/FLX/1440160
FR-FBn	MF	43,24	5,68	2008-2020	https://doi.org/10.18160/2G60-ZHAK
FR-Fon	DBF	48,48	2,78	2005-2020	https://doi.org/10.18160/2G60-ZHAK
FR-Hes	DBF	48,67	7,06	2014-2020	https://doi.org/10.18160/2G60-ZHAK
FR-LBr	ENF	44,72	-0,77	1996-2008	https://doi.org/10.18140/FLX/1440163
FR-Pue	EBF	43,74	3,6	2000-2014	https://doi.org/10.18140/FLX/1440164
GF-Guy	EBF	5,28	-52,93	2004-2020	https://doi.org/10.18160/2G60-ZHAK
GL-ZaH	GRA	74,47	-20,55	2000-2014	https://doi.org/10.18140/FLX/1440224
IL-Yat	ENF	31,34	35,05	2000-2020	https://doi.org/10.18160/2G60-ZHAK
IT-Col	DBF	41,85	13,59	1996-2014	https://doi.org/10.18140/FLX/1440167
IT-Cpz	EBF	41,71	12,38	1997-2009	https://doi.org/10.18140/FLX/1440233
IT-Lav	ENF	45,96	11,28	2003-2020	https://doi.org/10.18160/2G60-ZHAK
IT-MBo	GRA	46,01	11,05	2003-2020	https://doi.org/10.18160/2G60-ZHAK
IT-Ren	ENF	46,59	11,43	1999-2020	https://doi.org/10.18160/2G60-ZHAK
IT-Tor	GRA	45,84	7,58	2008-2020	https://doi.org/10.18160/2G60-ZHAK
NL-Loo	ENF	52,17	5,74	1996-2014	https://doi.org/10.18140/FLX/1440178
RU-Cok	OSH	70,83	147,49	2003-2014	https://doi.org/10.18140/FLX/1440182
RU-Fyo	ENF	56,46	32,92	1998-2020	https://doi.org/10.18160/2G60-ZHAK
RU-Sam	GRA	72,37	126,5	2002-2014	https://doi.org/10.18140/FLX/1440185
SE-Nor	ENF	60,09	17,48	2014-2020	https://doi.org/10.18160/2G60-ZHAK
US-Bar	DBF	44,06	-71,29	2004-2021	https://doi.org/10.17190/AMF/2006969
US-GLE	ENF	41,37	-106,24	2005-2020	https://doi.org/10.17190/AMF/1871136
US-Ha1	DBF	42,54	-72,17	1991-2020	https://doi.org/10.17190/AMF/1871137
US-Ho2	ENF	45,21	-68,75	1999-2020	https://doi.org/10.17190/AMF/1881581
US-IB2	GRA	41,84	-88,24	2004-2011	https://doi.org/10.18140/FLX/1440072
US-KLS	GRA	38,77	-97,57	2012-2019	https://doi.org/10.17190/AMF/1854367
US-MMS	DBF	39,32	-86,41	1999-2020	https://doi.org/10.17190/AMF/1854369
US-MOz	DBF	38,74	-92,2	2004-2019	https://doi.org/10.17190/AMF/1854370
US-Me2	ENF	44,45	-121,56	2002-2020	https://doi.org/10.17190/AMF/1854368
US-Mpj	WSA	34,44	-106,24	2008-2020	https://doi.org/10.17190/AMF/1832161
US-NR1	ENF	40,03	-105,55	1998-2016	https://doi.org/10.17190/AMF/1871141
US-ONA	GRA	27,38	-81,95	2015-2020	https://doi.org/10.17190/AMF/1832163
US-Oho	DBF	41,55	-83,84	2004-2013	https://doi.org/10.18140/FLX/1440088
US-PFa	MF	45,95	-90,27	1995-2014	https://doi.org/10.18140/FLX/1440089
US-SRG	GRA	31,79	-110,83	2008-2021	https://doi.org/10.17190/AMF/2204877
US-SRM	WSA	31,82	-110,87	2004-2014	https://doi.org/10.18140/FLX/1440090
US-Seg	GRA	34,36	-106,7	2007-2021	https://doi.org/10.17190/AMF/1984572
US-Ses	OSH	34,33	-106,74	2007-2021	https://doi.org/10.17190/AMF/1984573
US-Syv	MF	46,24	-89,35	2001-2014	https://doi.org/10.18140/FLX/1440091
US-Ton	WSA	38,43	-120,97	2001-2014	https://doi.org/10.18140/FLX/1440092
US-UMB	DBF	45,56	-84,71	2000-2014	https://doi.org/10.18140/FLX/1440093
US-UMd	DBF	45,56	-84,7	2007-2014	https://doi.org/10.18140/FLX/1440101

IGBP: The land cover classification defined by the International Geosphere Biosphere Programme (IGBP)

Continued Table S1. Information on the eddy covariance sites from the combined ICOS, AmeriFlux, FLUXNET 2015 tier 1 dataset

site	IGBP	latitude	longitude	period	DOI
US-Var	GRA	38,41	-120,95	2000-2021	https://doi.org/10.17190/AMF/1993904
US-WCr	DBF	45,81	-90,08	1999-2014	https://doi.org/10.18140/FLX/1440095
US-Whs	OSH	31,74	-110,05	2007-2020	https://doi.org/10.17190/AMF/1984574
US-Wjs	SAV	34,43	-105,86	2007-2021	https://doi.org/10.17190/AMF/1871146
US-Wkg	GRA	31,74	-109,94	2004-2021	https://doi.org/10.17190/AMF/1984575
ZA-Kru	SAV	-25,02	31,5	2000-2013	https://doi.org/10.18140/FLX/1440188

IGBP: The land cover classification defined by the International Geosphere Biosphere Programme (IGBP)