Supplement of Biogeosciences, 13, 4291–4313, 2016
http://www.biogeosciences.net/13/4291/2016/
doi:10.5194/bg-13-4291-2016-supplement

Biogeosciences Open Access EGU

Supplement of

Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms

Gianluca Tramontana et al.

Correspondence to: Gianluca Tramontana (g.tramontana@unitus.it)

S1 Eddy covariance study sites used for FLUXCOM experiment

Table S1: List of the La Thuile and CarboAfrica study sites used for this study. Elevation marked with * are filled by Google earth. Abbreviation of IGBP vegetation type are: CRO cropland, CSH closed shrubland, DBF deciduous broadleaf forest, EBF evergreen broadleaf forest, ENF evergreen needleleaf forest, GRA grassland, MF mixed forest, OSH open shrubland, SAV savannah, WET wetland, WSA woody savannah. Abbreviation for climate type are Arc arctic, Bor boreal, Dry dry climate arid and semiarid, Subtrop subtropical and mediterranean climate, Temp temperate climate, Temp/cont temperate continental climate, Temp/cont hot temperate continental climate with hot or warm summer, Trop is the tropical climate.

ID	Site Code	Lat (°N)	Long (°E)	Elevation (m)	VegType IGBP	Koeppen Climate class	Climate type
1	AT-Neu	47.12	11.32	970	GRA	Cfb	Temp
2	AU-Fog	-12.54	131.31	6*	WET	Aw	Trop
3	AU-How	-12.49	131.15	38*	WSA	Aw	Trop
4	AU-Tum	-35.66	148.15	1200	EBF	Cfb	Temp
5	AU-Wac	-37.43	145.19	545	EBF	Cfb	Temp
6	BE-Bra	51.31	4.52	16	MF	Cfb	Temp
7	BE-Jal	50.56	6.07	500	MF	Cfb	Temp
8	BE-Lon	50.55	4.74	167	CRO	Cfb	Temp
9	BE-Vie	50.31	6.00	450	MF	Cfb	Temp
10	BR-Ban	-9.82	-50.16	173*	EBF	Aw	Trop
11	BR-Ma2	-2.61	-60.21	120	EBF	Af	Trop
12	BR-Sa1	-2.86	-54.96	196*	EBF	Am	Trop
13	BR-Sa3	-3.02	-54.97	184*	EBF	Am	Trop
14	BR-Sp1	-21.62	-47.65	690	WSA	Aw	Trop
15	BW-Ghg	-21.51	21.74	1161*	SAV	BSh	Dry
16	BW-Ghm	-21.2	21.75	1149*	WSA	BSh	Dry
17	BW-Ma1	-19.92	23.56	950	WSA	BSh	Dry
18	CA-Ca1	49.87	-125.33	300	ENF	Cfb	Temp
19	CA-Ca2	49.87	-125.29	300	ENF	Cfb	Temp
20	CA-Ca3	49.53	-124.90	159*	ENF	Cfb	Temp
21	CA-Gro	48.22	-82.16	346*	MF	Dfb	Temp/cont hot
22	CA-Let	49.71	-112.94	960	GRA	Dfb	Temp/cont hot
23	CA-Man	55.88	-98.48	259	ENF	Dfc	Bor
24	CA-Mer	45.41	-75.52	70	WET	Dfb	Temp/cont hot
25	CA-NS1	55.88	-98.48	260	ENF	Dfc	Bor
26	CA-NS2	55.91	-98.52	260	ENF	Dfc	Bor
27	CA-NS3	55.91	-98.38	260	ENF	Dfc	Bor
28	CA-NS4	55.91	-98.38	260	ENF	Dfc	Bor

29	CA-NS5	55.86	-98.49	260	ENF	Dfc	Bor
30	CA-NS6	55.92	-98.96	259*	OSH	Dfc	Bor
31	CA-NS7	56.64	-99.95	271*	OSH	Dfc	Bor
32	CA-Oas	53.63	-106.20	530	DBF	Dfc	Bor
33	CA-Obs	53.99	-105.12	629	ENF	Dfc	Bor
34	CA-Ojp	53.92	-104.69	579	ENF	Dfc	Bor
35	CA-Qcu	49.27	-74.04	392	ENF	Dfc	Bor
36	CA-Qfo	49.69	-74.34	382	ENF	Dfc	Bor
37	CA-SF1	54.49	-105.82	536	ENF	Dfc	Bor
38	CA-SF2	54.25	-105.88	520	ENF	Dfc	Bor
39	CA-SF3	54.09	-106.00	540	ENF	Dfc	Bor
40	CA-SJ1	53.91	-104.66	580	ENF	Dfc	Bor
41	CA-SJ2	53.94	-104.65	580	ENF	Dfc	Bor
42	CA-SJ3	53.88	-104.64	495*	ENF	Dfc	Bor
43	CA-TP1	42.66	-80.56	265	ENF	Dfb	Temp/cont hot
44	CA-TP2	42.77	-80.46	212	ENF	Dfb	Temp/cont hot
45	CA-TP3	42.71	-80.35	184	ENF	Dfb	Temp/cont hot
46	CA-TP4	42.71	-80.36	184	ENF	Dfb	Temp/cont hot
47	CA-WP1	54.95	-112.47	540	MF	Dfc	Bor
48	CA-WP2	55.54	-112.33	789*	WET	Dfc	Bor
49	CA-WP3	54.47	-113.32	676*	WET	Dfc	Bor
50	CG-Hin	-4.68	12.00	118*	EBF	Aw	Trop
51	CG-Kis	-4.79	11.98	124*	EBF	Aw	Trop
52	CG-Tch	-4.29	11.66	83*	OSH	Aw	Trop
53	CH-Oe1	47.29	7.73	450	GRA	Cfb	Temp
54	CH-Oe2	47.29	7.73	452	CRO	Cfb	Temp
55	CN-Bed	39.53	116.25	30	EBF	Dwa	Temp/cont hot
56	CN-Cha	42.40	128.10	761	MF	Dwb	Temp/cont hot
57	CN-Do1	31.52	121.96	4	WET	Cfa	SubTrop
58	CN-Do2	31.58	121.90	4	WET	Cfa	SubTrop
59	CN-Do3	31.52	121.97	4	WET	Cfa	SubTrop
60	CN-Du1	42.05	116.67	1350	CRO	Dwb	Temp/cont hot
61	CN-Du2	42.05	116.28	1350	GRA	Dwb	Temp/cont hot
62	CN-HaM	37.37	101.18	3250	GRA	ET	Arc
63	CN-Ku1	40.54	108.69	1020	EBF	BSk	Dry
64	CN-Ku2	40.38	108.55	1160	OSH	BSk	Dry
65	CN-Xfs	44.13	116.33	1110*	GRA	BSk	Dry
66	CN-Xi1	43.55	116.68	1250	GRA	Dwb	Temp/cont hot
67	CN-Xi2	43.55	116.67	1250	GRA	Dwb	Temp/cont hot
68	CZ-BK1	49.50	18.54	908	ENF	Dfb	Temp/cont hot
69	CZ-wet	49.03	14.77	420	WET	Cfb	Temp
70	DE-Geb	51.10	10.91	162	CRO	Cfb	Temp
71	DE-Gri	50.95	13.51	385	GRA	Cfb	Temp
72	DE-Hai	51.08	10.45	430	DBF	Cfb	Temp
73	DE-Har	47.93	7.60	201	ENF	Cfb	Temp
74	DE-Kli	50.89	13.52	480	CRO	Cfb	Temp
75	DE-Meh	51.28	10.66	286	GRA	Cfb	Temp
76	DE-Tha	50.96	13.57	380	ENF	Cfb	Temp
77	DE-Wet	50.45	11.46	785	ENF	Cfb	Temp
78	DK-Fou	56.48	9.59	51	CRO	Cfb	Temp

79	DK-Lva	55.68	12.08	15	GRA	Cfb	Temp
80	DK-Ris	55.53	12.10	10	CRO	Cfb	Temp
81	DK-Sor	55.49	11.65	40	DBF	Cfb	Temp
82	ES-ES1	39.35	-0.32	10	ENF	Csa	SubTrop
83	ES-ES2	39.28	-0.32	10	CRO	Csa	SubTrop
84	ES-LJu	36.93	-2.75	1600	OSH	Csa	SubTrop
85	ES-LMa	39.94	-5.77	260	SAV	Csa	SubTrop
86	ES-VDA	42.15	1.45	1770	GRA	Cfb	Temp
87	FI-Hyy	61.85	24.29	181	ENF	Dfc	Bor
88	FI-Kaa	69.14	27.30	155	WET	Dfc	Bor
89	FI-Sii	61.83	24.19	169*	GRA	Dfc	Bor
90	FI-Sod	67.36	26.64	180	ENF	Dfc	Bor
91	FR-Aur	43.55	1.11	240*	CRO	Cfb	Temp
92	FR-Fon	48.48	2.78	90	DBF	Cfb	Temp
93	FR-Gri	48.84	1.95	125	CRO	Cfb	Temp
94	FR-Hes	48.67	7.06	300	DBF	Cfb	Temp
95	FR-Lam	43.49	1.24	182*	CRO	Cfb	Temp
96	FR-LBr	44.72	-0.77	61	ENF	Cfb	Temp
97	FR-Pue	43.74	3.60	270	EBF	Csa	SubTrop
98	GF-Guy	5.28	-52.93	35	EBF	Af	Trop
99	GH-Ank	5.27	-2.69	77*	EBF	Am	Trop
100	HU-Bug	46.69	19.60	140	GRA	Cfb	Temp
101	HU-Mat	47.85	19.73	350	GRA	Cfb	Temp
102	ID-Pag	-2.35	114.04	30	EBF	Af	Trop
103	IE-Ca1	52.86	-6.92	50	CRO	Cfb	Temp
104	IE-Dri	51.99	-8.75	187	GRA	Cfb	Temp
105	IL-Yat	31.34	35.05	650	ENF	BSh	Dry
106	IT-Amp	41.9	13.61	884	GRA	Cfa	SubTrop
107	IT-BCi	40.52	14.96	20	CRO	Csa	SubTrop
108	IT-Be2	46.00	13.03	62*	GRA	Cfb	Temp
109	IT-Cas	45.06	8.67	90*	CRO	Cfa	SubTrop
110	IT-Col	41.85	13.59	1550	DBF	Cfa	SubTrop
111	IT-Cpz	41.71	12.38	68	EBF	Csa	SubTrop
112	IT-Lav	45.96	11.28	1353	ENF	Cfb	Temp
113	IT-Lec	43.30	11.27	314	EBF	Cfa	SubTrop
114	IT-LMa	45.58	7.15	350	GRA	Cfb	Temp
115	IT-Mal	46.12	11.70	1730	GRA	Cfb	Temp
116	IT-MBo	46.02	11.05	1550	GRA	Cfb	Temp
117	IT-Noe	40.61	8.15	28	CSH	Csa	SubTrop
118	IT-Non	44.69	11.09	25	DBF	Cfa	SubTrop
119	IT-PT1	45.20	9.06	60	DBF	Cfa	SubTrop
120	IT-Ren	46.59	11.43	1730	ENF	Dfb	Temp
121	IT-Ro1	42.41	11.93	235	DBF	Csa	SubTrop
122	IT-Ro2	42.39	11.92	224	DBF	Csa	SubTrop
123	IT-SRo	43.73	10.28	4	ENF	Csa	SubTrop
124	IT-Vig	45.32	8.85	107*	DBF	Cfa	SubTrop
125	JP-Mas	36.05	140.03	12	CRO	Cfa	SubTrop
126	JP-Tom	42.74	141.51	140	MF	Dfb	Temp/cont hot
127	KR-Hnm	34.55	126.57	7*	CRO	Cfa	SubTrop
128	KR-Kw1	37.75	127.16	330	MF	Dwa	Temp/cont hot

129	ML-AgG	15.34	-1.48	286*	GRA	BWh	Dry
130	NL-Ca1	51.97	4.93	1	GRA	Cfb	Temp
131	NL-Haa	52.00	4.81	-2*	GRA	Cfb	Temp
132	NL-Hor	52.03	5.07	-2	GRA	Cfb	Temp
133	NL-Lan	51.95	4.90	-2*	CRO	Cfb	Temp
134	NL-Loo	52.17	5.74	25	ENF	Cfb	Temp
135	NL-Lut	53.40	6.36	0*	CRO	Cfb	Temp
136	PL-wet	52.76	16.31	54	WET	Cfb	Temp
137	PT-Esp	38.64	-8.60	95	EBF	Csa	SubTrop
138	PT-Mi1	38.54	-8.00	250	EBF	Csa	SubTrop
139	PT-Mi2	38.48	-8.02	190	GRA	Csa	SubTrop
140	RU-Che	68.61	161.34	3*	MF	Dfc	Bor
141	RU-Cok	70.62	147.88	23*	WET	Dfc	Bor
142	RU-Fyo	56.46	32.92	265	ENF	Dfb	Temp/cont hot
143	RU-Zot	60.80	89.35	90	ENF	Dfc	Bor
144	SD-Dem	13.28	30.48	542*	SAV	BWh	Dry
145	SE-Abi	68.36	18.79	361*	DBF	ET	Arc
146	SE-Deg	64.18	19.55	270	WET	Dfc	Bor
147	SE-Nor	60.09	17.48	43	ENF	Dfb	Temp/cont hot
148	SE-Sk1	60.13	17.92	42	ENF	Dfb	Temp/cont hot
149	SE-Sk2	60.13	17.84	55	ENF	Dfb	Temp/cont hot
150	SK-Tat	49.12	20.16	1050	ENF	Dfb	Temp/cont hot
151	UK-AMo	55.79	-3.24	270	WET	Cfb	Temp
152	UK-EBu	55.87	-3.21	190	GRA	Cfb	Temp
153	UK-ESa	55.91	-2.86	97	CRO	Cfb	Temp
154	UK-Ham	51.12	-0.86	80	DBF	Cfb	Temp
155	UK-Her	51.78	-0.48	140	CRO	Cfb	Temp
156	UK-PL3	51.45	-1.27	115	DBF	Cfb	Temp
157	UK-Tad	51.21	-2.83	3	GRA	Cfb	Temp
158	US-ARM	36.61	-97.49	314	CRO	Cfa	SubTrop
159	US-Atq	70.47	-157.41	15	WET	ET	Arc
160	US-Aud	31.59	-110.51	1469	GRA	BSk	Dry
161	US-Bar	44.06	-71.29	272	DBF	Dfb	Temp/cont hot
162	US-Bkg	44.35	-96.84	510	GRA	Dfa	Temp/cont hot
163	US-Blo	38.90	-120.63	1315	ENF	Csa	SubTrop
164	US-Bn1	63.92	-145.38	518	ENF	Dsc	Bor
165	US-Bn2	63.92	-145.38	410	DBF	Dsc	Bor
166	US-Bn3	63.92	-145.74	469	OSH	Dsc	Bor
167	US-Bo1	40.01	-88.29	219	CRO	Dfa	Temp/cont hot
168	US-Bo2	40.01	-88.29	219	CRO	Dfa	Temp/cont hot
169	US-Brw	71.32	-156.63	1	WET	ET	Arc
170	US-CaV	39.06	-79.42	994	GRA	Cfb	Temp
171	US-Dk1	35.97	-79.09	168	GRA	Cfa	SubTrop
172	US-Dk2	35.97	-79.10	168	DBF	Cfa	SubTrop
173	US-Dk3	35.98	-79.09	163	ENF	Cfa	SubTrop
174	US-Fmf	35.14	-111.73	2160	ENF	Csb	SubTrop
175	US-FPe	48.31	-105.10	634	GRA	BSk	Dry
176	US-FR2	29.95	-98.00	272	WSA	Cfa	SubTrop
177	US-Fuf	35.09	-111.76	2180	ENF	Csb	Temp/cont
178	US-Fwf	35.45	-111.77	2270	GRA	Csb	Temp/cont

179	US-Goo	34.25	-89.87	87	GRA	Cfa	SubTrop
180	US-Ha1	42.54	-72.17	340	DBF	Dfb	Temp/cont hot
181	US-Ho1	45.20	-68.74	60	ENF	Dfb	Temp/cont hot
182	US-IB1	41.86	-88.22	227	CRO	Dfa	Temp/cont hot
183	US-IB2	41.84	-88.24	227	GRA	Dfa	Temp/cont hot
184	US-Ivo	68.49	-155.75	674*	WET	ET	Arc
185	US-KS1	28.46	-80.67	2*	ENF	Cfa	SubTrop
186	US-KS2	28.61	-80.67	3	CSH	Cfa	SubTrop
187	US-Los	46.08	-89.98	480	WET	Dfb	Temp/cont hot
188	US-LPH	42.54	-72.18	378	DBF	Dfb	Temp/cont hot
189	US-Me1	44.58	-121.50	896	ENF	Dfb	Temp/cont hot
190	US-Me2	44.45	-121.56	1253	ENF	Dfb	Temp/cont hot
191	US-Me3	44.32	-121.61	1005	ENF	Dfb	Temp/cont hot
192	US-Me4	44.50	-121.62	922	ENF	Dfb	Temp/cont hot
193	US-MOz	38.74	-92.20	219	DBF	Cfa	SubTrop
194	US-NC1	35.81	-76.71	6	ENF	Cfa	SubTrop
195	US-NC2	35.80	-76.67	5	ENF	Cfa	SubTrop
196	US-Ne1	41.17	-96.48	361	CRO	Dfa	Temp/cont hot
197	US-Ne2	41.16	-96.47	362	CRO	Dfa	Temp/cont hot
198	US-Ne3	41.18	-96.44	363	CRO	Dfa	Temp/cont hot
199	US-NR1	40.03	-105.55	3050	ENF	Dfc	Bor
200	US-PFa	45.95	-90.27	470	MF	Dfb	Temp/cont hot
201	US-SO2	33.37	-116.62	1394	CSH	Csa	SubTrop
202	US-SO3	33.38	-116.62	1429	CSH	Csa	SubTrop
203	US-SO4	33.38	-116.64	1429	CSH	Csa	SubTrop
204	US-SP1	29.74	-82.22	50	ENF	Cfa	SubTrop
205	US-SP2	29.76	-82.24	50	ENF	Cfa	SubTrop
206	US-SP3	29.75	-82.16	50	ENF	Cfa	SubTrop
207	US-SRM	31.82	-110.87	1120	WSA	BSk	Dry
208	US-Syv	46.24	-89.35	540	MF	Dfb	Temp/cont hot
209	US-Ton	38.43	-120.97	177	WSA	Csa	SubTrop
210	US-UMB	45.56	-84.71	234	DBF	Dfb	Temp/cont hot
211	US-Var	38.41	-120.95	129	GRA	Csa	SubTrop
212	US-WCr	45.81	-90.08	520	DBF	Dfb	Temp/cont hot
213	US-Wi0	46.62	-91.08	340	ENF	Dfb	Temp/cont hot
214	US-Wi1	46.73	-91.23	342	DBF	Dfb	Temp/cont hot
215	US-Wi2	46.69	-91.15	381	ENF	Dfb	Temp/cont hot
216	US-Wi4	46.74	-91.17	377	ENF	Dfb	Temp/cont hot
217	US-Wi5	46.65	-91.09	369	ENF	Dfb	Temp/cont hot
218	US-Wi6	46.62	-91.3	357	OSH	Dfb	Temp/cont hot
219	US-Wi7	46.65	-91.07	345	ENF	Dfb	Temp/cont hot
220	US-Wi8	46.72	-91.25	389	DBF	Dfb	Temp/cont hot
221	US-Wi9	46.62	-91.08	341	ENF	Dfb	Temp/cont hot
222	US-Wkg	31.74	-109.94	1531	GRA	BSk	Dry
223	US-Wrc	45.82	-121.95	371	ENF	Csb	Temp
224	VU-Coc	-15.44	167.19	80	EBF	Af	Trop

S2 Description of additionally developed model

S2.1 MTE_M

The MTE_M algorithm grows several model trees with full extent until a small number of samples (2*number of regression variables) are in each leaf node. The splits are determined as described in Jung et al. (2009), but a certain fraction of data (default is one third) is randomly removed before the split is determined. After the split is found, the data previously hold-out are walked into the respective two children nodes as well. Within each children node a suitable multiple regression with variable selection, as described in Jung et al. (2009), is performed using a certain fraction of data (default is two third) and the remaining fraction of data to estimate the mean squared error of the multiple linear regression. The random local hold-out for both, split determination and regression, introduces instability in the tree induction algorithm and allows for generating an ensemble of model trees. The prediction of MTE_M is then the weighted average over all nodes (not only leaf nodes) of all trees where the conditions (split criteria) are applied. The weights are taken as the inverse of the mean squared error of each node. If the predicted value by a regression in one node is beyond the range of observed values for that node then the predicted value is truncated to the minimum or maximum of the respective observed values, and its weight is decreased by a factor of 1000.

The MTE_M algorithm is capable of making use of samples where some predictor variables are missing. In the initial model tree building phase all samples with missing values are removed. Afterwards, all samples with missing values are walked into those nodes where the missing predictors were not required either as split variable in the hierarchy above this node or as

regression variable. Then the multiple linear regressions and its mean squared error are recomputed for the respective 'updated' nodes.

S2.2 MTE_V

The MTE_V is an ensemble of m model trees (30 in FLUXCOM experiment). The model trees were created by the recursive partitioning of the training dataset (starting from the first node, named "root").

The splitting was carried out comparing the performance of a "reference regression" with a "splitted regression". More specifically the "reference regression" is the best multiple regression for the sample, emerging from a comparison of a user defined maximum number of regressions (10 in FLUXCOM experiment). The candidate regressions differing for the drivers that were randomly extracted (three drivers and their interactions were used in the FLUXCOM experiment). The metric of the regression's performance were the MEF and the RMSE, calculated from an X-fold cross comparison (five-fold in FLUXCOM experiment), by which the best reference multiple regression was selected.

The "splitted regressions" were established on the basis of splitting rules dividing the sample into two subsamples. Several splitting rules were extracted from an additional splitting dataset carrying both quantitative and categorical variables and then evaluated. The best multiple regression were established for each subsample (from the splitting rule) following the scheme adopted to estimate the "reference regressions". The splitting rule and associated regressions maximizing the accuracy of predictions was chosen as "splitted regression".

For the comparison between “splitted regressions” and “reference regression”, the performance were adjusted for the higher number of parameters into the “splitted regressions” (equations S2.2.1 and S2.2.2).

$$MEF_{adj} = MEF * \frac{n-p}{n} \quad \text{(S2.2.1)}$$

$$RMSE_{adj} = RMSE * \frac{n}{n-p} \quad \text{(S2.2.2)}$$

In eq. S2.2.1 and S2.2.2, n is the sample dimensions, and p the number of parameters. If the MEF_{adj} (and $RMSE_{adj}$) of the “splitted regressions” were greater (lower) than the “reference regression”, “splitted regressions” were accepted.

The subsample resulting from the splitting rules were used for another partitioning, and the regressions on the left and right “branches “used as “reference regression” for the next step. The development of a branch was stopped when the “reference regression” resulted better than any additional “splitted regressions”.

The variability among the model trees was determined by the random extraction of the candidate regressions and splitting rules.

The final output was the median estimates of the predictions across the m trees.

S2.3: Random Decision Forests and Gaussian Processes (RDF-GP)

The RDF-GP is a combination of Random Decision Forests (RDF) (Breiman et al., 2001) and Gaussian Process (GP) (Rasmussen et al., 2006).

A RDF is an ensemble method consisting of several decision trees. Decision trees are based on the hierarchical binary decision scheme: beginning from a root node, simple comparisons of attribute values with a threshold decide whether a data example is handed over to the left or the right child node of a currently processed node. In the last nodes of the trees there are regression models based on Gaussian Processes (GP).

In GP, the target (observed) variable (y_n) is modeled as the sum of some unknown latent function of the input *f(x)* plus constant power (homoscedastic) Gaussian noise e_n, i.e. $y_n=f(x)+e_n$.

Instead of proposing a parametric form for *f(x)* and learning its parameters in order to fit observed data well, GP regression proceeds in a Bayesian, non-parametric way. A zero mean GP prior is placed on the latent function *f(x)* and a Gaussian prior is used for each latent noise term *e*. Given the priors GP, samples drawn from f(x) at the set of training data points follow a joint multivariate Gaussian with zero mean and covariance matrix K, also known as kernel function. Computing the posterior distribution can be done analytically. Then, predictions for unseen points depend on the chosen kernel function measuring the similarity between training samples and unseen points.

The appropriate definition of the kernel is the bottleneck in any kernel method in general, and for GP in particular. Since we here deal with both real continuous and discrete features we introduce a composite kernel function as the sum of a kernel for continuous (K_c) and discrete data (K_d). For K_c we used the squared exponential (K_{SE}) kernel function (radial basis function), while for K_d we used the algorithms Overlap or Goodall4 described in Boriah et al. (2008).

For continuous data the K_{SE} kernel function computes the similarity between training (x) and unseen (x') points as:

$$K_{SE} = \sigma^2 \exp\left(-\frac{(x-x')^2}{2l}\right) \qquad \text{(S2.3.1)}$$

Were σ^2 and l are parameters which have to be optimized.

The Overlap measure returns 1 if the value for attribute d is equal and 0 otherwise. Goodall4 computes the similarity ($S_d(x_d,x_d')$) as:

$$S_d(x_d, x_d') = \begin{cases} \frac{f_d(x_d)(f_d(x_d)-1)}{n(n-1)} & \text{if } x_d = x_d' \\ 0 & \text{otherwise} \end{cases} \quad \text{(S2.3.2)}$$

where $f_d(x)$ is the frequency of how often the attribute d takes value x . GP are very powerful tools for the task of regression but they are often not applicable to large data sets directly because for the learning of the kernel the computational time is cubic in the number of training examples. In our experiments, we learn a GP with a rather small kernel using only the training examples which reached certain leafs of the RDF. Furthermore, the random extraction of variables and samples for the RDF training (Breimann et al., 2001) avoided the over-fitting for the training data.

S3 Description of indexes of soil water availability.

A simple soil water balance model is used to derive predictor variables aiming at capturing water stress effects in a better way than solely based on vapor pressure deficit, precipitation, or remotely sensed indices. Soil water storage (SWS) is treated as a bucket model with a defined plant available water storage capacity (AWC). In each daily time step (t), the soil water storage of the previous time step (t-1) is updated by water recharge (R(t)), and water loss by evapotranspiration (E(t)):

$$SWS(t) = SWS(t-1) + R(t) - E(t) \quad \text{(S3.1)}$$

Recharge is taken as the minimum of precipitation (P(t)) and the water deficit of the previous time step:

$$R(t) = \min[P(t), AWC - SWS(t-1)] \quad \text{(S3.2)}$$

Evapotranspiration is taken as the minimum of demand (E_{pot}) and water supply (E_{sup}) driven E:

$$E(t) = \min[E_{pot}(t), E_{sup}(t)] \quad \text{(S3.3)}$$

E_{pot} is calculated based on Priestley-Taylor ($E_{PT,}$ Priestley and Taylor, 1972) and scaled with the fraction of photosynthetic active radiation (fPAR), which is based on a smoothed mean seasonal cycle based on MODIS:

$$E_{pot}(t) = fPAR(t) * E_{PT}(t) \quad \text{(S3.4)}$$

Water supply limited evaporation follows Teuling et al. (2006) is modeled as a fraction (k) of SWS:

$$E_{sup}(t) = k * [SWS(t-1) + R(t)] \quad \text{(S3.5)}$$

An upper and a lower soil layer are realized by making the assumption that both recharge by precipitation and water loss by evaporation occur from top to bottom:

$$R_{upper}(t) = \min[R(t), AWC_{upper} - SWS_{upper}(t-1)]; \; R_{lower}(t) = R(t) - R_{upper}(t) \quad \text{(S3.6)}$$

$$E_{upper}(t) = \min[E(t), SWS_{upper}(t-1) + R_{upper}(t)]; \; E_{lower}(t) = E(t) - E_{upper}(t) \quad \text{(S3.7)}$$

The water availability index (WAI) is expressed as fractional available water:

$$WAI_{upper}(t) = SWS_{upper}(t) / AWC_{upper} \quad \text{(S3.8)}$$

$$WAI_{lower}(t) = SWS_{lower}(t) / AWC_{lower} \quad \text{(S3.9)}$$

An alternative index of water availability (IWA) is calculated analogously to evaporative fraction:

$$IWA(t) = E(t) / E_{pot}(t) \quad \text{(S3.10)}$$

The simple model requires the definition of the parameter k, and the storage capacities of the upper and lower soil layers. K was chosen to be 0.05**,** which corresponds to the median value of 19 analyzed site-years by Teuling et al. (2006). AWC_{upper} and AWC_{lower} were chosen heuristically as 15

mm and 100 mm, respectively. The model was run with the same parameters for all sites, a necessary requirement to use the derived predictor variables at global scale. AWC were initialized with full storage in 1989, and the necessary meteorological data are based on downscaled ERA-Interim reanalysis; those were replaced by measurements from the towers whenever available.

S4 List of the candidate predictors

The Table S4.1 presents a complete list of the candidate predictors. The predictors include time varying variables, mean seasonal cycle and its metrics (e.g. minimum, maximum, amplitude.) For further details see also paper Sect. 2.3.2.

Table S4.1: List of the candidate predictors

Name	Symbol	Units	Values	MSC	ANO
		Original Variables			
MODIS spectral reflectances BRDF adjusted bands 1 to 7	Reflectancebands1to7	None	RS	BOTH	RS
Daily land surface temperature	LSTday	°K	RS	BOTH	RS
Nightly land surface temperature	LSTnight	°K	RS	BOTH	RS
Enhanced vegetation index	EVI	None	RS	BOTH	RS
Normalized difference vegetation index	NDVI	None	RS	BOTH	RS
Land surface water index	LSWI	None	RS	BOTH	RS
Normalized difference water index	NDWI	None	RS	BOTH	RS
Fraction of absorbed Par	fpar	None	RS	BOTH	RS
Leaf Area index	LAI	None	RS	BOTH	RS
Aggregated Koeppen Climate	AggregatedKoeppenIds	None	RS+METEO		
Koeppen Climate	KoeppenIds	None	RS+METEO		
Relative humidity	Rh	None	RS+METEO	RS+METEO	RS+METEO
Daily mean air temperature	Tair	°C	RS+METEO	RS+METEO	RS+METEO
Daily maximum air temperature	Tmax	°C	RS+METEO	RS+METEO	RS+METEO
Daily minimum air temperature	Tmin	°C	RS+METEO	RS+METEO	RS+METEO
Vapor pressure deficit	VPD	KPa	RS+METEO	RS+METEO	RS+METEO
Precipitation	Precip	mm	RS+METEO	RS+METEO	RS+METEO
Index of water availability	IWA	None	RS+METEO	RS+METEO	RS+METEO
Water availability index upper	WAI_u	None	RS+METEO	RS+METEO	RS+METEO
Water availability index lower	WAI_l	None	RS+METEO	RS+METEO	RS+METEO
Global Radiation	Rg	MJ $m^{-2}d^{-1}$	BOTH	BOTH	BOTH
Plant Functional Type	PFT	None	BOTH		
Canopy height	Canopyheight	m	BOTH		
Potential Radiation	Rpot	MJ $m^{-2}d^{-1}$		BOTH	
Potential evapotranspiration	PET	mm		RS+METEO	RS+METEO
		Interactions			
Product between EVI and LST	EVI*LST	°K	RS	BOTH	RS
Product between EVI and Rg	EVI*Rg	MJ $m^{-2}d^{-1}$	RS	BOTH	RS
Product between EVI and Rpot	EVI*Rpot	MJ $m^{-2}d^{-1}$1	RS	BOTH	RS
Product between fPAR and LST	FPAR*LST	°K	RS	BOTH	RS
Product between fPAR and Rg	FPAR*Rg	MJ $m^{-2}d^{-1}$	RS	BOTH	RS

Product between fPAR and Rpot	FPAR*Rpot	MJ $m^{-2}d^{-1}$	RS	BOTH	RS
Product between NDVI and LST	NDVI*LST	°K	RS	BOTH	RS
Product between NDVI and Rg	NDVI*Rg	MJ $m^{-2}d^{-1}$	RS	BOTH	RS
Product between NDVI and Rpot	NDVI*Rpot	MJ $m^{-2}d^{-1}$	RS	BOTH	RS
Difference between daily and nightly LST	LSTday-LSTnight	°K	RS	BOTH	RS
Product between mean seasonal cycle of EVI and LST	MSC(EVI)*LST	°K	RS	BOTH	BOTH
Product between mean seasonal cycle of NDVI and LST	MSC(NDVI)*LST	°K	RS	BOTH	BOTH
Product between mean seasonal cycle of fPAR and LST	MSC(FPAR)*LST	°K	RS	BOTH	BOTH
Water balance (lag n days)	WB (lag 3 days)	mm	RS+METEO		
	WB (lag 5 days)	mm	RS+METEO		
	WB (lag 7 days)	mm	RS+METEO		
	WB (lag 9 days)	mm	RS+METEO		
	WB (lag 11 days)	mm	RS+METEO		
	WB (lag 13 days)	mm	RS+METEO		
	WB (lag 15 days)	mm	RS+METEO		
	WB (lag 17 days)	mm	RS+METEO		
	WB (lag 19 days)	mm	RS+METEO		
	WB (lag 21 days)	mm	RS+METEO		
	WB (lag 23 days)	mm	RS+METEO		
	WB (lag 25 days)	mm	RS+METEO		
	WB (lag 27 days)	mm	RS+METEO		
	WB (lag 29 days)	mm	RS+METEO		
Product among mean seasonal cycle of EVI, RG and IWA	MSC(EVI)*Rg*IWA	MJ $m^{-2}d^{-1}$	RS+METEO		
Product among mean seasonal cycle of fPAR, RG and IWA	MSC(FPAR)*Rg*IWA	MJ $m^{-2}d^{-1}$	RS+METEO		
Product among mean seasonal cycle of NDVI, RG and IWA	MSC(NDVI)*Rg*IWA	MJ $m^{-2}d^{-1}$	RS+METEO		
Product between mean seasonal cycle of EVI and Rg	MSC(EVI)*Rg	MJ $m^{-2}d^{-1}$	BOTH	BOTH	BOTH
Product between mean seasonal cycle of NDVI and Rg	MSC(NDVI)*Rg	MJ $m^{-2}d^{-1}$	BOTH	BOTH	BOTH
Product between mean seasonal cycle of fPAR and Rg	MSC(FPAR)*Rg	MJ $m^{-2}d^{-1}$	BOTH	BOTH	BOTH
Ratio between global and potential radiation	Rg/Rpot	none	BOTH		

S5 Description of the Guided hybrid genetic algorithm

GHGA is an optimization algorithm that combines a global search genetic algorithm tailored to variable selection problems, and a 'guided' procedure for local elimination of variables to speed-up the stochastic nature of the backward search of the GA (see Jung and Zscheischler (2013) for details). GHGA makes suggestions of variable sets, which are tested by a regression algorithm (e.g. RFs) and the resulting performance is quantified in a cost function. The cost function (c(v)) of the variable set v aims at identifying a compromise between performance (m) and number of variables (n(v)) and follows Jung and Zscheischler (2013):

$$c(v) = n(v) + 2^{\frac{M - m(v)}{\varepsilon}} \quad \text{(S5.1)}$$

Where m(v) is Nash-Sutcliff's modeling efficiency (MEF) for variable set v, M is the MEF identified so far during the search, and epsilon is a parameter that describes the accepted performance loss for retaining one variable less (set to 0.005).

The settings of GHGA were the recommended default values given in Jung and Zscheischler (2013). The training of RF was based on a randomly chosen half of FLUXNET sites, while the remaining half was used for validation (for which MEF was calculated). To minimize differences of MEF between different variable sets by chance, the stratification in training and validation sites and the bootstrap samples for growing the regression trees were always the same. The number of regression trees of each RF was set to 60 to limit the computational burden. The variable search stopped when no new global or local optimum were found within the last 1000 cost function evaluations, or when 10000 cost function evaluations were reached. The final selected variable set was the ones with smallest cost function values. The entire variable selection exercise required

nearly 100000 pairs of training and prediction of RF, (here used to make suggestion of variables with in total more than 5 million regression trees.

S6 Methods settings.

Machine learning methods need of free hyperparametres often related to the cost function used that one aims to minimize, the regularization terms that are in charge of controlling overfitting the training data, the shape and smoothness of the nonlinear functions. In Table S6.1 we presented the hyperparametres setting the applied machine learning applied for the training at eight daily time step by RS setup (see section 2.3.1). The variants for the training of the machine learning applied for the RS+METEO setup (daily time step) were shown in Table S6.2.

Table S6.1: Method settings adopted for the training of the eight daily time step for the RS setup. Acronym of methods are: RF Random Forest, MTE Model Tree Ensemble, SVM Support Vector Machine, KRR Kernel Ridge Regression, GPR Gaussian Processes, ANN feed forward Artificial Neural Networks, GMDH Group Method of Data Handling or polynomial neural networks, RDFGP Random Forest with Gaussian Processes in leafs node, MARS Multivariate Adaptive Regression Splines. NA Not Available

Name	Hyperparameters and settings	Scaling	Ensemble	Reference
	Tree methods			
RF	Minimum number of samples in leafs = 5. Fraction of variables to find split per node = 0.33. Surrogate splits activated to use samples with incomplete predictors.	none	200 Regression trees	Breimann (2001)
MTE	All continuous variables are used for split and regression. One two-fold cross-validation in leaf nodes to avoid overfitting.	none	25 Model Trees selected out of 2500	Jung et al. (2009)
MTE_M	All continuous variables are used for split and regression. Local hold-out fraction = 0.33.	none	50 Model Trees by randomly removing the hold-out	Supplementary material S2

			fraction locally	
MTE_V	Drivers for regressions in the leaf node = 3. Splitting rules and regressions randomly extracted choosing the best among 10 extractions. Five fold cross comparison to evaluate multiple regressions.	constrained between the minimum/maximum values of the domain of the regressions into the leaf node	Median ensemble of 30 trees.	Supplementary material S2
		Kernel methods		
KRR	Grid search of the squared exponential kernel lengthscale and the regularization parameter. Testing against a hold out of 50% of sites.	-1 to 1	NA	Shawe-Taylor and Cristianini (2004)
SVM	Grid search of the squared exponential kernel lengthscale. the epsilon-insensitivity zone for the cost function, and the regularization parameter to control errors penalization. Testing against a hold out of 50% of sites.	-1 to 1	NA	Vapnik et al. (1998)
GPR	Hyperparameters found by maximum likelihood of the marginal evidence.	Remove the mean, scaling all features between 0 to 1	NA	Rasmussen (2006)
RDF-GP	Hyperparameters found by maximum likelihood of the marginal evidence. Same initialization parameters as in GPR (for the prediction GPR model in the leaves of the RF). A minimum of 1000 sample in each leaf. Ensemble of 10 trees.	Remove the mean, scaling all features between 0 to 1	10	Fröhlich et al. (2012) Supplementary material S2
		Neural Network methods		
ANN	Feed forward network trained with the Levenberg-Marquardt learning algorithm. 5 initializations for each net; percentage of sites distributed among training, test and validation set 60, 20, 20 respectively. Net architecture with one or two layers, each one having from 5 to 12	0 to 1	10, randomly sampling sites for training, test and validation sets	Haykin (1999) Papale et al. (2003)

	neurons. The net with the best performance (on the validation set) and the simplest architecture was chosen. Maximum number of inputs for individual neurons = 3. Maximum number of neurons per layer equal to the number of predictor; Degree of polynomials in neurons = 3.			
GMDH		-1 to 1	20 (by randomly sampling sites for training and validation)	Ungaro et al. (2005)
	Multivariate Splines			
MARS	The maximal number of basis functions included in the forward model building phase = 21 (default). Generalized Cross-Validation (GCV) penalty per knot = 3 (default value). Maximum degree of interactions =2.	-1 to 1 for X Zscore for Y	20 (by bootstrapping)	Friedman et al. (1991)

Footnotes: MTE_M can handle samples with missing predictors; GMDH Neurons take input from preceding layer and from original input variables; MARS piecewise-cubic models no self interactions. In all methods, excepting the tree methods, the vegetation category was converted in Woody/non-woody dummy vector (1 for woody PFT and 0 for non-woody PFT).

Table S6.2: Method settings adopted for the training in the RS+METEO setup (daily time step). Acronym of the same for Table S6.1.

Name	Hyperparameters and settings	Scaling	Ensemble	Reference
	Tree methods			
RF	Minimum number of samples in leafs = 25. Fraction of variables to find split per node = 0.33. Surrogate splits not activated (not handle missing values)	none	200 Regression trees	Breimann (2001)
	Kernel methods			
KRR	Grid search of the squared exponential kernel lengthscale and the regularization parameter. Testing against a hold out of 50% of sites.	-1 to 1	20 models each one using a training set (10000 points) extracted by a stratified	Shawe-Taylor and Cristianini (2004)

			random sampling strategy.	
Neural Network methods				
ANN	Feed forward network trained with the Levenberg-Marquardt learning algorithm. 5 initializations for each net; percentage of sites distributed among training, test and validation set 60,20,20 respectively. Net architecture with one or two layers, each one having from 5 to 12 neurons. The net with the best performance (on the validation set) and the simplest architecture was choosen.	0 to 1	10, randomly sampling sites for training, test and validation sets	Haykin, (1999) Papale et al. (2003)
Multivariate Splines				
MARS*	The maximal number of basis functions included in the forward model building phase = 21 (default). Generalized Cross-Validation (GCV) penalty per knot = 3 (default value). Maximum degree of interactions =2.	-1 to 1 for X Zscore for Y	20 models; training set (10000 points) extracted by a stratified random sampling strategy.	Friedman et al. (1991)

Reference

Boriah S., Chandola V., and Kumar V.,: Similarity measures for categorical data: A comparative evaluation. Proceedings of the 2008 SIAM International Conference on Data Mining, Atlanta, 24-26 April, 243-254, 2008.

Breiman, L.: Random Forests, Mach. Learn., 45 (1), 5–32, doi:10.1023/A:1010933404324, 2001.

Friedman, J. H.: Multivariate Adaptive Regression Splines, Ann. Statist., 19, 1-67, doi:10.1214/aos/1176347963, 1991.

Fröhlich, B., Rodner, E., Kemmler, M. and Denzler, J.: Large-scale gaussian process classification using random decision forests, S. Mach. Perc., 22 (1), 113–120, DOI 10.1007/s00138-012-0480-y, 2012.

Haykin, S.: Neural Networks – A Comprehensive Foundation (2nd ed.)., Prentice Hall., 1999.

Jung, M., and Zscheischler, J.: A Guided Hybrid Genetic Algorithm for Feature Selection with Expensive Cost Functions, Procedia Computer Science, 18, 2337-2346, doi: 10.1016/j.procs.2013.05.405, 2013.

Jung, M., Reichstein, M. and Bondeau, A.: Towards global empirical upscaling of FLUXNET Eddy Covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001-2013, doi:10.5194/bg-6-2001-2009, 2009.

Papale, D. and Valentini, R.: A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization, Global Change Biol, 9, 525–535, doi: 10.1046/j.1365-2486.2003.00609.x, 2003.

Priestley, C.H.B. and Taylor, R.J.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. Mon Weather Rev 100, 81-92, 1972

Rasmussen C. E. and Williams C. K. I.: Gaussian Processes for Machine Learning, the MIT Press, ISBN 026218253X, 2006.

Shawe-Taylor, J. and Cristianini, N.: Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

Teuling A. J., Seneviratne S. I., Williams C. and Troch P. A.: Observed timescales of evapotranspiration response to soil moisture Geophysl res lett, 33, L23403. doi:10.1029/2006GL028178, 2006

Ungaro, F., Calzolari, C. and Busoni, E: Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padano–Veneta region of North Italy: water retention properties, Geoderma, 124, 293–317, doi:10.1016/j.geoderma.2004.05.007, 2005.

Vapnik, V., Golowich, S. and Smola, A.: Support vector method for function approximation, regression estimation, and signal processing, Adv Neur In, 9, 281–287, 1997.