Supplementary information to:

Partitioning soil organic carbon into its centennially active and stable fractions with machine-learning models based on Rock-Eval® thermal analysis ($PARTY_{SOC}$v2.0 and $PARTY_{SOC}v2.0_{EU}$)

Lauric Cécillon[1,2], François Baudin[3], Claire Chenu[4], Bent T. Christensen[5], Uwe Franko[6], Sabine Houot[4], Eva Kanari[2,3], Thomas Kätterer[7], Ines Merbach[8], Folkert van Oort[4], Christopher Poeplau[9], Juan Carlos Quezada[10,11,12], Florence Savignac[3], Laure N. Soucémarianadin[13], Pierre Barré[2]

[1]Normandie Univ., UNIROUEN, INRAE, ECODIV, Rouen, France
[2]Laboratoire de Géologie, École normale supérieure, CNRS, PSL Univ., IPSL, Paris, France
[3]Institut des Sciences de la Terre de Paris, Sorbonne Université, CNRS, Paris, 75005, France
[4]UMR 1402 ECOSYS, INRAE, AgroParisTech, Univ. Paris Saclay, Thiverval-Grignon, 78850, France
[5]Department of Agroecology, Aarhus University, AUFoulum, 8830 Tjele, Denmark
[6]Department of soil system science, Helmholtz Centre for Environmental Research, UFZ, 06120 Halle Germany
[7]Department of Ecology, Swedish University of Agricultural Sciences, 75007 Uppsala, Sweden
[8]Department Community Ecology, Helmholtz Centre for Environmental Research, UFZ, 06246 Bad Lauchstädt, Germany
[9]Thünen Institute of Climate-Smart Agriculture, 38116 Braunschweig, Germany
[10]Laboratory of Ecological Systems ECOS and Laboratory of Plant Ecology Research PERL, School of Architecture, Civil and Environmental Engineering ENAC, École Polytechnique Fédérale de Lausanne EPFL, 1015 Lausanne, Switzerland
[11]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 1015 Lausanne, Switzerland
[12]Ecosystem Management, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETHZ, 8092 Zürich, Switzerland
[13]ACTA – les instituts techniques agricoles, 75595 Paris, France

Correspondence to: Lauric Cécillon (lauric.cecillon@inrae.fr)

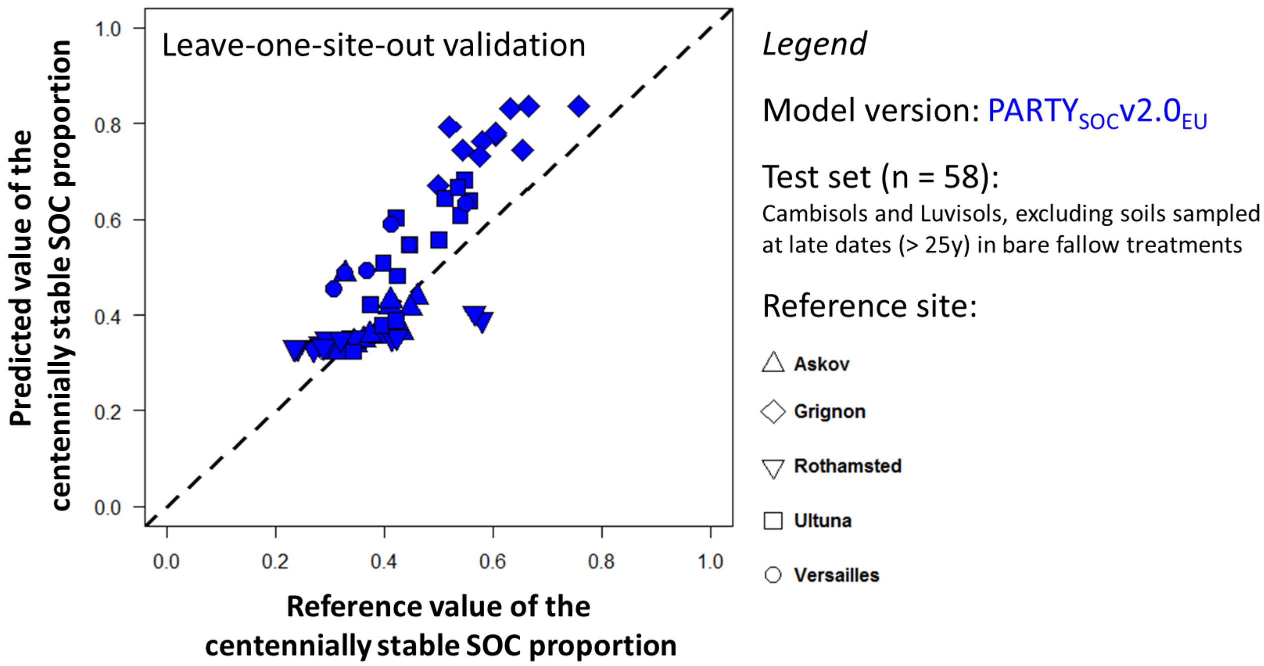

Figure S1: Sensitivity of the PARTY$_{SOC}$v2.0$_{EU}$ machine-learning model to the test set. Model sensitivity to the test set was assessed as its sensitivity to independent test samples (1) from a reference soil group not existing in the training set (*i.e.* excluding Chernozem soil samples from the test set); (2) that are unlikely to be encountered in agricultural soils (*i.e.* excluding from the test set soils sampled at late dates of bare fallow treatments, more than 25 years after the experiment onset, which cannot represent soils with regular carbon input). This figure is built with a test set composed exclusively of soils from independent Cambisols and Luvisols of Northwestern Europe. It should be compared to Fig. 2b (for PARTY$_{SOC}$v2.0$_{EU}$) that is built using the exact same training set and validation procedure (leave-one-site-out), but that uses a different test set including 32 additional independent test samples (15 Chernozem soils samples and 17 soil samples collected at late dates in bare fallow treatments).

Table S1: Main changes between the first version of $PARTY_{SOC}$ and the second version of the model ($PARTY_{SOC}$v2.0 and $PARTY_{SOC}$v2.0$_{EU}$). Abbreviations: SOC, soil organic carbon; SOC_{EA}: soil organic carbon content determined by elemental analysis; $TOC_{RE6,}$ soil organic carbon content determined by Rock-Eval® thermal analysis; MinC, soil inorganic carbon content determined by Rock-Eval® thermal analysis; LTBF, long-term bare fallow.

	First version of $PARTY_{SOC}$ (Cécillon et al., 2018)	**$PARTY_{SOC}$v2.0** (This study)	**$PARTY_{SOC}$v2.0$_{EU}$** (This study)
Number of reference sites	4	7	6
Method to estimate the centennially stable SOC content at the reference sites	Exclusively inferred from plant-free LTBF treatments; and calculated exclusively by modelling the decline with time of SOC initially present at the onset of the experiment	Inferred from plant-free LTBF treatments (6 sites) and from one vegetated site (La Cabaña); and estimated by modelling the decline with time of SOC initially present at the onset of the experiment, or by using a measured value of SOC content or C_4-plant derived SOC that is lower than the parameter *c* of Eq. (1)	Exclusively inferred from plant-free LTBF treatments; and estimated by modelling the decline with time of SOC initially present at the onset of the experiment, or by using a measured value of SOC content or C_4-plant derived SOC that is lower than the parameter *c* of Eq. (1)
Method to estimate the centennially stable SOC proportion in reference topsoil samples	By dividing the site-specific content of the centennially stable SOC content by the SOC_{EA} content of topsoil samples	By dividing the site-specific content of the centennially stable SOC content by the TOC_{RE6} or the TOC_{RE6}+MinC content of topsoil samples	By dividing the site-specific content of the centennially stable SOC content by the TOC_{RE6} or the TOC_{RE6}+MinC content of topsoil samples
Number of reference topsoil samples (per site)	118 (unequal number of samples per site)	105 (15 samples per reference site)	90 (15 samples per reference site)
Rock-Eval® predictor variables	30, not directly related to SOC content, not necessarily correlated to the proportion of the centennially stable SOC fraction	18, some of them are directly related to SOC content, all are highly correlated to the proportion of the centennially stable SOC fraction (Spearman's rho > 0.5)	18, some of them are directly related to SOC content, all are highly correlated to the proportion of the centennially stable SOC fraction (Spearman's rho > 0.5)
Criteria regarding the inclusion of topsoils in the	None (all available reference topsoil samples are used)	Exclusion of reference topsoil samples (1) from treatments experiencing repeated	Exclusion of reference topsoil samples (1) from treatments experiencing repeated

reference sample set of the model		applications of compost or manure; (2) having an organic carbon yield of Rock-Eval® analysis below 86% or above 116%. Then, selection of 15 reference topsoil samples per site (1) enhancing the range of the centennially stable SOC proportion at that site; (2) having the best organic carbon yield of Rock-Eval® analysis	applications of compost or manure; (2) having an organic carbon yield of Rock-Eval® analysis below 86% or above 116%. Then, selection of 15 reference topsoil samples per site (1) enhancing the range of the centennially stable SOC proportion; (2) having the best organic carbon yield of Rock-Eval® analysis
Validation procedure for the machine-learning model	Mostly "random splitting" procedure, "leave-one-site-out" procedure tested only for one site (Ultuna)	Fully independent "leave-one-site-out" procedure	Fully independent "leave-one-site-out" procedure

Table S2: Basic characteristics of the seven reference sites used as the training set of PARTY$_{SOC}$v2.0, and site-specific values of the parameter *c* in Eq. (1). Most data are taken from Barré et al. (2010), Cécillon et al. (2018), Franko and Merbach (2017) and Quezada et al. (2019). Abbreviations: SFE, static fertilization experiment (V120); FE, fallow experiment (V505a).

Site, Country	**Latitude, longitude** [° min s]	**Mean annual temperature** [°C]	**Mean annual precipitation** [mm]	**Soil type** [WRB; FAO, 2014]	**Land use before experiment**	**Sampling depth** [cm]	**pH** (in H_2O)	**Soil texture** [%] (clay/silt/sand)	**Parameter c in Eq. (1)** (g C kg^{-1})
Versailles, France	48°48’12.76”N; 2°05’09.95”E	10.7	628	Haplic Luvisol	grassland	0–25	5.6 to 6.4	17/57/26	6.22
Rothamsted, England	51°48’14.12”N; 0°21’41.39”E	9.5	712	Chromic Luvisol	grassland	0–23	5.2 to 6.3	25/62/13	10.46
Ultuna, Sweden	59°48’46.00”N; 17°39’01.75”E	5.5	533	Eutric Cambisol	arable	0–20	6.6	36/41/23	6.95
Grignon, France	48°51’01.35”N; 1°57’03.55”E	10.7	649	Calcaric Cambisol	grassland	0–25	8.0 to 8.3	16/54/30	7.12
Askov, Denmark	B3 55°28’12.89”N; 9°06’41.77”E B4 55°28’19.19”N; 9°07’00.02”E	7.8	862	Arenic Luvisol	arable	0–20	5.5 to 6.5	10/11/79	5.10
Bad Lauchstädt, Germany	SFE 51°23’24.34”N; 11°52’48.48”E FE 51°23’34.75”N; 11°52’39.61”E	8.9	484	Haplic Chernozem	arable	0–20 (SFE) 0–30 (FE)	6.9 to 7.4	21/68/11	16.22
La Cabaña, Colombia	4°16’N; 73°22’W	27	3400	Dystric Cambisol	grassland	0–10	4.0 to 4.6	30/34/36	5.12

Table S3: List of the 105 reference topsoil samples retained as the training set of PARTY$_{SOC}$v2.0, including information on their reference site, land cover, agronomical treatment, sampling year and values for the 40 Rock-Eval® parameters considered in this study. See the Section 2.2 for a description of the 40 Rock-Eval® parameters and their units. Supplementary Table S3 is provided as a separate csv file.

Table S4: Basic statistics of the 40 Rock-Eval® parameters, calculated on the 105 reference topsoil samples of the training set of PARTY$_{SOC}$v2.0. See the Section 2.2 for a description of the 40 Rock-Eval® parameters. The 18 Rock-Eval® parameters retained as predictor variables for the second version of PARTY$_{SOC}$ are shown in bold.

Rock-Eval® parameter (unit)		**Mean**	**Median**	**Minimum**	**Maximum**	**Standard deviation**
$T10_{HC\_PYR}$	(°C)	310	309	279	343	16
$T30_{HC\_PYR}$	(°C)	372	372	330	402	15
$T50_{HC\_PYR}$	(°C)	422	421	389	445	12
$T70_{HC\_PYR}$	**(°C)**	**465**	**462**	**446**	**484**	**10**
$T90_{HC\_PYR}$	**(°C)**	**523**	**522**	**495**	**547**	**15**
$T10_{CO\_PYR}$	(°C)	307	305	295	318	5
$T30_{CO\_PYR}$	(°C)	358	357	346	373	6
$T50_{CO\_PYR}$	(°C)	405	403	391	423	7
$T70_{CO\_PYR}$	(°C)	460	458	451	477	7
$T90_{CO\_PYR}$	(°C)	525	524	519	539	4
$T10_{CO2\_PYR}$	(°C)	286	284	276	300	7
$T30_{CO2\_PYR}$	**(°C)**	**343**	**339**	**329**	**365**	**10**
$T50_{CO2\_PYR}$	**(°C)**	**392**	**387**	**375**	**419**	**13**
$T70_{CO2\_PYR}$	**(°C)**	**445**	**441**	**427**	**469**	**13**
$T90_{CO2\_PYR}$	**(°C)**	**511**	**510**	**497**	**526**	**8**
$T10_{CO\_OX}$	(°C)	323	317	299	383	16
$T30_{CO\_OX}$	(°C)	374	364	331	495	32
$T50_{CO\_OX}$	(°C)	421	400	366	578	46
$T70_{CO\_OX}$	**(°C)**	**481**	**464**	**404**	**671**	**52**
$T90_{CO\_OX}$	(°C)	568	547	477	776	64
$T10_{CO2\_OX}$	(°C)	327	326	318	358	6
$T30_{CO2\_OX}$	(°C)	375	374	363	405	8
$T50_{CO2\_OX}$	**(°C)**	**417**	**413**	**398**	**455**	**14**
$T70_{CO2\_OX}$	**(°C)**	**469**	**465**	**437**	**509**	**19**
$T90_{CO2\_OX}$	**(°C)**	**540**	**538**	**511**	**568**	**12**
I-index	(unitless)	0.18	0.17	0.02	0.37	0.08
R-index	(unitless)	0.59	0.59	0.47	0.70	0.05
TLHC-index	(unitless)	0.63	0.64	0.53	0.72	0.05

HI	**(mg HC g^{-1} C)**	**156**	**143**	**69**	**284**	**58**
OI_{RE6}	(mg O_2 g^{-1} C)	183	184	146	246	21
TOC_{RE6}	**(g C kg^{-1})**	**16.0**	**14.3**	**5.6**	**41.5**	**7.3**
MinC	(g C kg^{-1})	2.7	1.7	0.6	12.5	2.8
PC	**(g C kg^{-1})**	**3.7**	**3**	**1.1**	**11**	**2.2**
S2	**(g C kg^{-1})**	**2.2**	**1.7**	**0.6**	**7.4**	**1.6**
PseudoS1	**(g C kg^{-1})**	**0.12**	**0.09**	**0.05**	**0.51**	**0.07**
PseudoS1/PC	(unitless)	0.04	0.03	0.02	0.11	0.02
PseudoS1/TOC_{RE6}	(unitless)	0.008	0.007	0.003	0.023	0.004
S2/PC	**(unitless)**	**0.56**	**0.54**	**0.42**	**0.76**	**0.09**
PC/TOC_{RE6}	**(unitless)**	**0.22**	**0.22**	**0.14**	**0.32**	**0.05**
HI/OI_{RE6}	**(mg HC mg^{-1} O_2)**	**0.87**	**0.74**	**0.43**	**1.92**	**0.39**

Table S5: Sensitivity of model performance to the reference sites included in the training set, using 15 topsoil samples from the sites of Grignon (a) or Versailles (b) as test sets. Site-specific performance statistics are calculated on a test set made of topsoil samples from Grignon or Versailles. Abbreviations are defined in Section 2.5.

a	**Grignon topsoil samples used as test set (n = 15)**				
Composition of the training set	**6 reference sites ($PARTY_{SOC}$v2.0, Grignon out)**	**5 reference sites ($PARTY_{SOC}$v2.0$_{EU}$, Grignon out)**	**4 reference sites (Ultuna, Versailles, Rothamsted, Askov)**	**3 reference sites (Ultuna, Versailles, Askov)**	**2 reference sites (Versailles, Askov)**
R^2_{OOB}	0.82	0.88	0.79	0.86	0.88
$RMSEP_{OOB}$	0.09	0.07	0.08	0.06	0.06
R^2	0.56	0.44	0.86	0.82	0.87
RMSEP	0.11	0.18	0.07	0.06	0.07
$_R$RMSEP	0.18	0.28	0.11	0.10	0.10
RPIQ	0.74	0.47	1.22	1.36	1.27
Bias	0.088	0.157	0.003	-0.035	-0.042

b	**Versailles topsoil samples used as test set (n = 15)**				
Composition of the training set	**6 reference sites ($PARTY_{SOC}$v2.0, Versailles out)**	**5 reference sites ($PARTY_{SOC}$v2.0$_{EU}$, Versailles out)**	**4 reference sites (Ultuna, Grignon, Askov, Rothamsted)**	**3 reference sites (Ultuna, Grignon, Askov)**	**2 reference sites (Ultuna, Grignon)**
R^2_{OOB}	0.85	0.89	0.83	0.89	0.85
$RMSEP_{OOB}$	0.08	0.07	0.06	0.04	0.05
R^2	0.48	0.66	0.79	0.83	0.79
RMSEP	0.17	0.14	0.13	0.10	0.11
$_R$RMSEP	0.28	0.24	0.23	0.17	0.19
RPIQ	1.77	2.08	2.20	2.90	2.67
Bias	0.036	0.018	0.052	0.006	0.025

Supplementary references

Barré, P., Eglin, T., Christensen, B. T., Ciais, P., Houot, S., Kätterer, T., van Oort, F., Peylin, P., Poulton, P. R., Romanenkov, V. and Chenu, C.: Quantifying and isolating stable soil organic carbon using long-term bare fallow experiments, Biogeosciences, 7(11), 3839–3850, https://doi.org/10.5194/bg-7-3839-2010, 2010.

Cécillon, L., Baudin, F., Chenu, C., Houot, S., Jolivet, R., Kätterer, T., Lutfalla, S., Macdonald, A., van Oort, F., Plante, A. F., Savignac, F., Soucémarianadin, L. N. and Barré, P.: A model based on Rock-Eval thermal analysis to quantify the size of the centennially persistent organic carbon pool in temperate soils, Biogeosciences, 15(9), 2835–2849, https://doi.org/10.5194/bg-15-2835-2018, 2018.

FAO: World reference base for soil resources 2014: international soil classification system for naming soils and creating legends for soil maps., FAO, Rome., 2014.

Franko, U. and Merbach, I.: Modelling soil organic matter dynamics on a bare fallow Chernozem soil in Central Germany, Geoderma, 303, 93–98, https://doi.org/10.1016/j.geoderma.2017.05.013, 2017.

Quezada, J. C., Etter, A., Ghazoul, J., Buttler, A. and Guillaume, T.: Carbon neutral expansion of oil palm plantations in the Neotropics, Sci. Adv., 5(11), eaaw4418, https://doi.org/10.1126/sciadv.aaw4418, 2019.