



**HAL**  
open science

## Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils

Roberta Farina, Renata Sándor, Mohamed Abdalla, Jorge Álvaro-fuentes,  
Luca Bechini, Martin Bolinder, Lorenzo Brillì, Claire Chenu, Hugues Clivot,  
Massimiliano de Antoni Migliorati, et al.

### ► To cite this version:

Roberta Farina, Renata Sándor, Mohamed Abdalla, Jorge Álvaro-fuentes, Luca Bechini, et al.. Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils. *Global Change Biology*, 2021, 27 (4), pp.904-928. 10.1111/gcb.15441 . hal-03081740

**HAL Id: hal-03081740**

**<https://hal.science/hal-03081740v1>**

Submitted on 11 Oct 2021

**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.

1                   **Ensemble modelling, uncertainty and robust predictions of organic**  
2                   **carbon in long-term bare-fallow soils**

3                   *Model inter-comparison of soil organic carbon*

4  
5 Farina, Roberta<sup>1,\*</sup>, Sándor, Renata<sup>2,3</sup>, Abdalla, Mohamed<sup>4</sup>, Álvaro-Fuentes, Jorge<sup>5</sup>, Bechini, Luca<sup>6</sup>,  
6 Bolinder, Martin A.<sup>7</sup>, Brilli, Lorenzo<sup>8</sup>, Chenu, Claire<sup>9</sup>, Clivot, Hugues<sup>10,11</sup>, De Antoni Migliorati,  
7 Massimiliano<sup>12</sup>, Di Bene, Claudia<sup>1</sup>, Dorich, Christopher D.<sup>13</sup>, Ehrhardt, Fiona<sup>14</sup>, Ferchaud,  
8 Fabien<sup>10</sup>, Fitton, Nuala<sup>4</sup>, Francaviglia, Rosa<sup>1</sup>, Franko, Uwe<sup>15</sup>, Giltrap, Donna L.<sup>16</sup>, Grant, Brian,  
9 B.<sup>17</sup>, Guenet, Bertrand<sup>18,19</sup>, Harrison, Matthew T.<sup>20</sup>, Kirschbaum, Miko U.F.<sup>16</sup>, Kuka, Katrin<sup>21</sup>,  
10 Kulmala, Liisa<sup>22</sup>, Liski, Jari<sup>22</sup>, McGrath, Matthew J.<sup>18</sup>, Meier, Elizabeth<sup>23</sup>, Menichetti, Lorenzo<sup>7</sup>,  
11 Moyano, Fernando<sup>24</sup>, Nendel, Claas<sup>25,29</sup>, Recous, Sylvie<sup>26</sup>, Reibold, Nils<sup>24</sup>, Shepherd, Anita<sup>4,27</sup>  
12 Smith, Ward N.<sup>17</sup>, Smith, Pete<sup>4</sup>, Soussana, Jean-François<sup>14</sup>, Stella, Tommaso<sup>25</sup>, Taghizadeh-Toosi,  
13 Arezoo.<sup>28</sup>, Tsutskikh, Elena<sup>25</sup>, Bellocchi, Gianni<sup>3</sup>

14  
15 <sup>1</sup> CREA - Council for Agricultural Research and Economics, Research Centre for Agriculture and  
16 Environment, Rome, Italy

17 <sup>2</sup> Agricultural Institute, Centre for Agricultural Research, Martonvásár, Hungary

18 <sup>3</sup> Université Clermont Auvergne, INRAE, VetAgro Sup, UREP, Clermont-Ferrand, France

19 <sup>4</sup> University of Aberdeen, UK

20 <sup>5</sup> Spanish National Research Council (CSIC), Zaragoza, Spain

21 <sup>6</sup> Università degli Studi di Milano, Italy

22 <sup>7</sup> Swedish University of Agricultural Sciences, Uppsala, Sweden

23 <sup>8</sup> CNR-IBE, Institute of Bioeconomy, Florence, Italy

24 <sup>9</sup> Université Paris Saclay, INRAE, AgroParisTech, Paris, France

25 <sup>10</sup> INRAE, BioEcoAgro, F-02000, Barenton-Bugny, France

- 26 <sup>11</sup> Université de Lorraine, INRAE, LAE, F-68000, Colmar, France
- 27 <sup>12</sup> Queensland University of Technology, Brisbane, Australia
- 28 <sup>13</sup> Colorado State University, Fort Collins CO, USA
- 29 <sup>14</sup> INRAE, CODIR, 75007 Paris, France
- 30 <sup>15</sup> Helmholtz Centre for Environmental Research, Halle, Germany
- 31 <sup>16</sup> Manaaki Whenua - Landcare Research, Palmerston North, New Zealand
- 32 <sup>17</sup> Ottawa Research and Development Centre, Agriculture and Agri-Food, Ottawa, Canada
- 33 <sup>18</sup> Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ,  
34 Université Paris-Saclay, 91191 Gif-sur-Yvette, France
- 35 <sup>19</sup> Laboratoire de Géologie de l'ENS, PSL Research University, Paris, France
- 36 <sup>20</sup> Tasmanian Institute of Agriculture, Australia
- 37 <sup>21</sup> JKI - Federal Research Centre for Cultivated Plants, Braunschweig, Germany
- 38 <sup>22</sup> Finnish Meteorological Institute, Helsinki, Finland
- 39 <sup>23</sup> CSIRO, Brisbane, Australia
- 40 <sup>24</sup> University of Gottingen, Germany
- 41 <sup>25</sup> Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany,
- 42 <sup>26</sup> Université de Reims Champagne Ardenne, INRAE, FARE, Reims, France
- 43 <sup>27</sup> formerly Rothamsted Research, North Wyke, Devon, UK
- 44 <sup>28</sup> Department of Agroecology, Aarhus University, Tjele, Denmark
- 45 <sup>29</sup> University of Potsdam, Germany
- 46
- 47
- 48 \*Corresponding author. Tel.: +39067005413; fax +39067005711
- 49 E-mail address: [roberta.farina@crea.gov.it](mailto:roberta.farina@crea.gov.it)

51

52 **Abstract**

53 Simulation models represent soil organic carbon (SOC) dynamics in global carbon (C) cycle  
54 scenarios to support climate-change studies. It is imperative to increase confidence in long-term  
55 predictions of SOC dynamics by reducing the uncertainty in model estimates. We evaluated SOC  
56 simulated from an ensemble of 26 process-based C models by comparing simulations to  
57 experimental data from seven long-term bare-fallow (vegetation-free) plots at six sites: Denmark  
58 (two sites), France, Russia, Sweden, the United Kingdom. The decay of SOC in these plots has  
59 been monitored for decades since the last inputs of plant material, providing the opportunity to test  
60 decomposition without the continuous input of new organic material. The models were run  
61 independently over multi-year simulation periods (from 28 to 80 years) in a blind test with no  
62 calibration (Bln) and with three calibration scenarios, each providing different levels of  
63 information and/or allowing different levels of model fitting: a) calibrating decomposition  
64 parameters separately at each experimental site (Spe); b) using a generic, knowledge-based,  
65 parameterisation applicable in the Central European region (Gen); and c) using a combination of  
66 both a) and b) strategies (Mix). We addressed uncertainties from different modelling approaches  
67 with or without spin-up initialisation of SOC. Changes in the multi-model median (MMM) of SOC  
68 were used as descriptors of the ensemble performance. On average across sites, Gen proved  
69 adequate in describing changes in SOC, with MMM equal to average SOC (and standard deviation)  
70 of  $39.2 (\pm 15.5)$  Mg C ha<sup>-1</sup> compared to the observed mean of  $36.0 (\pm 19.7)$  Mg C ha<sup>-1</sup> (last observed  
71 year), indicating sufficiently reliable SOC estimates. Moving to Mix ( $37.5 \pm 16.7$  Mg C ha<sup>-1</sup>) and  
72 Spe ( $36.8 \pm 19.8$  Mg C ha<sup>-1</sup>) provided only marginal gains in accuracy, but modellers would need  
73 to apply more knowledge and a greater calibration effort than in Gen, thereby limiting the wider  
74 applicability of models.

75

---

**LIST OF SYMBOLS AND ABBREVIATIONS**


---

<b>Symbol/abbreviation</b>	<b>Long version</b>	<b>Explanation</b>
<i>System variables</i>		
C	Carbon	Chemical element with atomic number 6
SOC	Soil organic carbon	Carbon stored in soil organic matter
SOM	Soil organic matter	The fraction of the soil that consists of plant, animal or microbial tissue in various stages of decomposition
N	Nitrogen	Chemical element with atomic number 7
<i>Experimentation</i>		
LTE	Long-term field experiment	Research facility providing data for monitoring trends and evaluating different agricultural management strategies over time
LTBF	Long-term bare-fallow experimental site	Research facility providing data for monitoring trends on bare-fallow soils
S1	Site 1	Askov (Denmark) – location 1
S2	Site 2	Askov (Denmark) – location 2
S3	Site 3	Grignon (France)
S4	Site 4	Kursk (Russia)
S5	Site 5	Rothamsted (United Kingdom)
S6	Site 6	Ultuna (Sweden)
S7	Site 7	Versailles (France)
<i>Modelling</i>		
M01, ..., M34	Model 01, ..., model 34	Simulation models (M) anonymously coded from 1 to 34
Bln	Blind	Uncalibrated simulations (blind test)
Gen	Generic	Generic simulation scenario
Mix	Mixed	Mixed simulation scenario
Spe	Specific	Specific simulation scenario
SP	Spin-up	Process of running the model from a set of conditions to initialise the state of C pools
NS	No spin-up	Any function (or analytical procedures) to make an initial partition of C pools (alternative to spin-up runs)
<i>Statistics</i>		
SD	Standard deviation	Variation amount of a set of data
MMM	Multi-model median	Median value of simulated data from different models
Obs	Observations	Observed data
RRMSE	Relative root mean square error	Aggregate magnitude of the errors in predictions relative to the mean of observations

---

EF	Modelling efficiency	Predictive power of a model with respect to the mean of observations
$R^2$	Coefficient of determination	Proportion of the variance in the modelled data that is predictable from the observations
r	Pearson's correlation coefficient	Degree to which predictions and observations are linearly related
P(t)	Paired Student t-test probability of I-type error	Probability to reject the true null hypothesis of equal means of two samples of paired data (i.e. predictions and observations)
d	Index of agreement	Ratio of the mean square error and the potential error represented by the largest value that the squared difference of each prediction/observation pair can attain
z	z-score transformation	Number of standard deviations by which the value of a raw score is above or below the mean value of the variable of interest
sd	Standard deviation	Standard deviation units expressing z-scores
$sd_{obs}$	Standard deviation of observations	Variation amount of a set of observed values
P	Predicted value	Value of a variable that is generated using a model
O	Observed value	Value of a variable that is actually observed
n	Number of predicted or observed values	Number of predicted/observed pairs
i	$i^{\text{th}}$ predicted or observed value	Subscript index of each predicted/observed pair
$\bar{O}$	Mean of observed values	Arithmetic mean of actually observed data
$\bar{P}$	Mean of predicted values	Arithmetic mean of actually observed data
$\bar{D}$	Mean difference	Arithmetic mean of the differences between predicted and observed values
$S_D$	Standard deviation of the differences	Variation amount of a set of differences between predictions and observations
p	Probability of I-type error	Probability to reject the true null hypothesis of null correlation between two variables
<hr/> <i>Agro-climatic metrics</i> <hr/>		
Tamp	Temperature amplitude	Difference between the highest and the lowest temperature in a year

---

Tmax	Maximum air temperature	Average of the highest daily temperatures in a year
Prec	Precipitation	Annual precipitation total
<i>b</i> <sup>a</sup>	De Martonne-Gottman aridity index	Indicator of aridity including both annual and monthly temperature and precipitation
<i>hw</i> <sup>a</sup>	Heatwave frequency	Number of at least seven consecutive days when the maximum air temperature is higher than the average summer (June, July and August) maximum temperature of a baseline value +3 °C

---

## 77 1. INTRODUCTION

78       The ability of soils to sequester and store large amounts of carbon (C) is well known (e.g.  
79 Lehmann and Kleber, 2015). Soil organic carbon (SOC) stocks are crucial for maintaining soil  
80 fertility and preventing erosion and desertification, and they positively influence the provision of  
81 ecosystem services at the local as well as the global scale (e.g. Lal, 2004, 2014). For these reasons,  
82 farmers aim to establish and maintain high organic C stocks in agricultural soils, which have often  
83 been depleted through historical land use practices (Fuchs et al., 2016; Gardi et al., 2016; Chenu et  
84 al., 2018). The continuing studies on SOC sources and biogeochemical processes in the soil  
85 environment provide key insights into climate-C feedbacks, and help prioritizing C sequestration  
86 initiatives (Gross and Harrison, 2019). In light of the climate change issue, the storage of C and  
87 additional sequestration of atmospheric C have received increasing attention recently (Rumpel et  
88 al., 2018; Whitehead et al., 2018; Lavallee et al., 2020), promoting land management, and agro-  
89 ecosystems in particular, as a key mitigation option (e.g. the ‘4 per mille Soils for Food Security  
90 and Climate’ initiative, Minasny et al., 2017; Soussana et al., 2017). However, the slow response  
91 of SOC to changes in management and environmental factors hampers our understanding of how  
92 SOC can be increased in a sustainable manner, especially under changing climatic conditions.  
93 Long-term field experiments (LTEs), in which SOC responses have been observed over several  
94 decades, provide this information and deliver reference data on SOC content for knowledge gain  
95 and model development (Johnston and Poulton, 2018). However, LTEs are costly to maintain, and  
96 it is generally difficult to extrapolate experimental results across space and time (Debreczeni and  
97 Körschens, 2003; Mirtl et al., 2018). Simulation models play a prominent role in SOC research  
98 because they provide a mathematical framework to integrate, examine and test the understanding  
99 of SOC dynamics (Campbell and Paustian, 2015). They can also be used to extrapolate from micro-  
100 (e.g. carbohydrate production during photosynthesis) to macro-scale dynamics (e.g. global C  
101 cycling) (e.g. Gottschalk et al., 2012; Sitch et al., 2003). In particular, complex agricultural and



102 environmental models incorporate a mechanistic view of processes and system interactions, in  
103 which the soil components are often represented by different, operationally defined, pools of  
104 different sizes and with different properties (e.g. Parton et al., 2015). The concept of multiple C-N  
105 pools represents C-N dynamics with an idealised description (Hill, 2003). The relative proportion  
106 of C and N (and sometimes lignin to N ratio) in the plant residue is the primary mode to divide  
107 plant inputs (from e.g. leaf litter and root exudates) into fresh litter pools, which then decompose  
108 into SOC (or SOM, i.e. soil organic matter) pools, each being modelled with different residence  
109 (or turnover) times, varying from months for labile products of microbial decomposition to  
110 hundreds to thousands of years for organic substances with firm organic-mineral bonds (e.g. Yadav  
111 and Malanson, 2007; Dungait et al., 2012). Plant material and animal manures are often modelled  
112 to enter the soil environment as either readily decomposable (carbohydrate-like) or resistant (lignin  
113 and cellulose-like) materials. A varying number of pools (often including inert and slow-  
114 decomposing organic matter, and microbial biomass) linked by first-order equations is usually  
115 simulating both C and N fluxes within and between each pool (Falloon and Smith, 2010). However,  
116 different models vary considerably in the underlying assumptions and C processes in current  
117 models, e.g. regarding number of pools, type of decomposition kinetics used and processes  
118 regulating SOC retention (Manzoni and Porporato, 2009; Cavalli et al., 2019).

119 Each model offers a distinctive synthesis of scientific knowledge (Brilli et al., 2017) and  
120 multi-model ensembles developed from several models may reduce uncertainties in biological and  
121 physical outputs that occur over large scales, such as regions and continents (e.g. Rötter et al.,  
122 2012; Asseng et al., 2013; Ehrhardt et al., 2018). The advantage of using ensemble estimates over  
123 individual models is that caused by compensation of errors across models, and a broader  
124 integration of model processes (Martre et al., 2015). It has been recommended to use model  
125 ensembles for reducing uncertainties in simulations of agricultural production (Asseng et al., 2013;  
126 Bassu et al., 2014; Challinor et al., 2014; Li et al., 2015; Ruane et al., 2016; Maiorano et al., 2017)

127 and other biophysical/biogeochemical outputs (Sándor et al., 2017, 2018a; Ehrhardt et al., 2018).  
128 However, after the pioneering study of Smith et al. (1997), who evaluated nine SOC models using  
129 12 datasets from seven LTEs, other modelling studies targeting SOC dynamics have often been  
130 limited in scope. Smith et al. (2012) used four models to assess the effect on SOC of crop residues'  
131 removal in 14 experiments in North America. Todd-Brown et al. (2013, 2014) performed global  
132 estimates of SOC changes with 11 Earth system models. Kirschbaum et al. (2015) used one  
133 simulation model and two years of eddy covariance measurements collected over an intensively  
134 grazed dairy pasture in New Zealand to better understand the drivers of changes in SOC stocks.  
135 Puche et al. (2019) performed a similar study in France. Using multi-model ensembles in scenario  
136 studies at eight sites worldwide, Basso et al. (2018) highlighted the importance of soil feedback  
137 effects (C and N) on the prediction of wheat and maize yield. We are not aware of any recent  
138 model inter-comparison studies specifically assessing soil C dynamics with several models across  
139 a range of experimental sites. This is a field where there is a need for standardised guidance to  
140 estimate C stocks at various spatial scales (Bispo et al., 2017). A difficulty in testing and comparing  
141 various models (and interpreting model outputs) lies in the interaction between soil and plant  
142 processes so that any of the model-data discrepancies could be due to errors in either component  
143 (e.g. Ehrmann and Ritz, 2014). A rigorous model testing and comparison would require different  
144 model components, e.g. plant and soil modules, to be assessed separately. Bare-fallow plots offer  
145 such an opportunity in that they are plots maintained for decades without any plant inputs. The  
146 changes in SOC stocks therefore result only from decomposition processes. To assess the function  
147 of soil-model components without interaction with plant processes, we conducted a model inter-  
148 comparison using a dataset from long-term bare-fallow experiments where plant inputs were zero.  
149 In this study, we refer to bare-fallow plots that were kept free of plants by manual and/or chemical  
150 means for several decades. We used seven bare-fallow treatments included in six long-term  
151 agricultural experiments (>25 years), all located in Europe (Denmark, France, Russia, Sweden and

152 United Kingdom). In these plots, the soils became progressively depleted in the more labile SOM  
153 components, as they decomposed, and relatively enriched in more stable SOM (Barré et al., 2010).  
154 The soil C concentrations determined at given years in these sites represented a unique opportunity  
155 to follow the decay of SOC from a multi-model ensemble perspective, without any interference  
156 from new plant C inputs, and conduct a multi-model ensemble comparison. The model inter-  
157 comparison included 26 process-based models from an international modelling community. Some  
158 models only accounted for soils and used C input from plants as an external input where others  
159 were full agro-ecosystem models that explicitly simulate plant growth and resulting C input into  
160 soils. These models all simulate interactions between the soil-atmosphere continuums in different  
161 ways, but for this comparison all models were run assuming no input of fresh plant-derived C,  
162 allowing the comparison of just the soil components of the models.

163 Here, we assess the models, by comparing multi-decadal simulations to experimental data  
164 from seven sites in Europe. The primary goal of this study was to assess the multi-model ensemble  
165 in simulating SOC dynamics across bare-fallow sites in Europe. To achieve this goal, model  
166 evaluation against actual measurements was performed before and after model calibration. In  
167 addition, deficient areas in models and their processes were identified, paving the road for future  
168 research directions.

169

## 170 **2. MATERIALS AND METHODS**

### 171 **2.1. Simulation models**

172 The ensemble of models consisted of 26 process-based models, mainly developed for crop or  
173 grassland ecosystems (or focussing just on soils) and covering a broad variety of approaches (Table  
174 1). While they are mostly based on first-order decay kinetics of multiple C pools (where C losses  
175 are proportional to SOC stocks with additional modifiers to represent the effects of other factors),  
176 ESOC1 simulates C fluxes with second-order kinetics equations based on concepts applied in

177 Schimel and Weintraub (2003) and reviewed in Wutzler and Reichstein (2008). In this case,  
178 organic matter decomposition includes reactions between SOC and decomposers (i.e. a microbial  
179 or enzyme pool). These different approaches depend mainly on alternative ways in which the C  
180 pools are linked. For instance, MONICA is one of the most complex models, considering three  
181 types of organic matter in six conceptual pools, viz. newly added organic matter, living soil  
182 microbial biomass and native non-living soil organic matter, each sub-divided into fast and slowly  
183 decomposing sub-pools. It simulates the turnover of C pools by applying first-order degradation  
184 to each pool due to microbial growth and maintenance respiration (after Abrahamsen and Hansen,  
185 2000). Then, like other models (e.g. CenW), MONICA also includes a coupled N-cycle and  
186 sophisticated temperature and water-balance calculations that act as modifiers of degradation and  
187 respiration rates. The decomposition rates of individual pools in such multi-pool SOC models are  
188 typically controlled by vastly different reaction coefficients that can result in highly nonlinear  
189 behaviour of the overall system (e.g. Caruso et al., 2018). The initial list included 34 models, but  
190 eight of them were excluded from further analysis because they showed severe limitations to run  
191 properly either under bare-fallow soils or under the given climate conditions. For all models,  
192 estimates of SOC were compared with measured SOC data.

193 Table 1. The process-based simulation models used. Model names were anonymised in the  
 194 reporting of simulation results using model codes from M01 to M34, from the initial list of 34  
 195 models, the order of models not being identical to that used in the table.  
 196

Model name	Version	C pools <sup>a</sup>	Spin-up	URL or contact for documentation/description	References
AMG	2	2 to 3	None	<a href="https://www6.hautsdefrance.inra.fr/agroimpact/Nos-dispositifs-outils/Modeles-et-outils-d-aide-a-la-decision/AMG-et-SIMEOS-AMG/AMG-model-description">https://www6.hautsdefrance.inra.fr/agroimpact/Nos-dispositifs-outils/Modeles-et-outils-d-aide-a-la-decision/AMG-et-SIMEOS-AMG/AMG-model-description</a>	Andriulo et al. (1999); Saffih-Hdadi and Mary (2008); Clivot et al. (2019)
-----					
APSIM	Apsim 7.9- r4044	3	Simulation from start of climate record (no additional simulation period)	<a href="http://www.apsim.info">http://www.apsim.info</a>	Keating et al. (2003); Holzworth et al. (2014)
-----					
	7.10 r4158		Yes		
-----					
CANDY_CIPS	1.0 (but always implemented in newest	4	None	<a href="https://www.ufz.de/export/data/2/95948_CANDY_MANUAL.pdf">https://www.ufz.de/export/data/2/95948_CANDY_MANUAL.pdf</a>	Kuka, (2005); Kuka et al. (2007)
-----					

	version of CANDY 29.06.2018					
CCB	2019.1.16	3	None	<a href="https://www.ufz.de/index.php?en=44046">https://www.ufz.de/index.php?en=44046</a>		Franko et al. (2011); Franko and Spiegel (2016); Franko and Merbach (2017)
Century	4.0	5 to 7	Yes	<a href="https://www2.nrel.colostate.edu/projects/century/MANUAL/html_manual/man96.html">https://www2.nrel.colostate.edu/projects/century/MANUAL/html_manual/man96.html</a>		Parton et al. (1987, 1994)
CenW	4.2	5	Uses an automatic spin-up routine to find equilibrium conditions under given environmental variables and specified system properties	<a href="http://www.kirschbaum.id.au/Welcome_Page.htm">http://www.kirschbaum.id.au/Welcome_Page.htm</a>		Kirschbaum (1999); Kirschbaum and Paul (2002)
C-TOOL	2014	3	None (can be run also with spin-up)	<a href="http://envs.au.dk/fileadmin/Resources/DMU/Luft/emission/SI_NKS/C-TOOL_Documentation_2015_.pdf">http://envs.au.dk/fileadmin/Resources/DMU/Luft/emission/SI_NKS/C-TOOL_Documentation_2015_.pdf</a>		Taghizadeh-Toosi and Olesen (2016); Taghizadeh-Toosi et al. (2014a, b, 2016)
Daily DayCent	4.5 2010 Daily DayCent 4.5 2013	5 to 9	Yes	<a href="http://www.nrel.colostate.edu/projects/daycent-home.html">http://www.nrel.colostate.edu/projects/daycent-home.html</a>		Parton et al. (1994, 1998); Del Grosso et al. (2001, 2002)

-----  
 Daily  
 DayCent  
 August 2014  
 -----

4.5 2013

DNDC	CAN	6	Yes (10 years recommended)	<a href="http://www.dndc.sr.unh.edu">http://www.dndc.sr.unh.edu</a>	Li et al. (2012); Smith et al. (2020)
DSSAT	...	5	Yes, 20 years prior to beginning of the experiment to estimate the proportions of carbon in each organic matter pool	<a href="http://dssat.net">http://dssat.net</a>	Jones et al. (2003); Porter et al. (2009); Gijssman et al. (2002); White et al. (2011); Thorp et al. (2012)
ECOSSE	5.0.1	5	None	<a href="https://www.abdn.ac.uk/staffpages/uploads/soi450/ECOSSE%20User%20manual%20310810.pdf">https://www.abdn.ac.uk/staffpages/uploads/soi450/ECOSSE%20User%20manual%20310810.pdf</a>	Smith et al. (2007, 2010a, b); Bell et al. (2010)
ESOC1	1.0	3	Yes	<a href="https://doi.org/10.5281/zenodo.3539484">https://doi.org/10.5281/zenodo.3539484</a>  <a href="mailto:fmoyano@uni-goettingen.de">fmoyano@uni-goettingen.de</a>	Moyano et al. (2018)

---



---

Exp		1	None	-	Lorenzo Menichetti ( <a href="mailto:lorenzo.menichetti@slu.se">lorenzo.menichetti@slu.se</a> )
Exp + inert		2	None	-	
ICBM	...	2	None	<a href="mailto:martin.bolinder@slu.se">martin.bolinder@slu.se</a>  <a href="https://www.slu.se">https://www.slu.se</a>	Andrén and Kätterer (1997); Andrén et al. (2008)
MONICA	2.0.2	7	None	<a href="http://monica.agrosystem-models.com">http://monica.agrosystem-models.com</a>	Nendel et al. (2011); Specka et al. (2016); Stella et al. (2019)
ORCHIDEE	2.0	3	Yes	<a href="https://vesg.ipsl.upmc.fr/thredds/fileServer/IPSLFS/orchidee/DOXYGEN/webdoc_2425/annotated.html">https://vesg.ipsl.upmc.fr/thredds/fileServer/IPSLFS/orchidee/DOXYGEN/webdoc_2425/annotated.html</a>	Krinner et al. (2005)
RothC	RothC10N ----- 26.3	4 to 5	None	<a href="https://www.rothamsted.ac.uk/rothamsted-carbon-model-rothc">https://www.rothamsted.ac.uk/rothamsted-carbon-model-rothc</a>	Coleman and Jenkinson (1999); Farina et al. (2013)
STICS	9.0	2 to 4	None	<a href="http://www6.paca.inra.fr/stics">http://www6.paca.inra.fr/stics</a>	Brisson et al. (1998, 2003, 2008); Coucheney et al. (2015)

---



---

YASSO15

15

5

Yes

<https://en.ilmatiiteenlaitos.fi/yasso>

Tuomi et al. (2009)

---

197 <sup>a</sup> Some models/model versions include options for varying C pools (this varying number may depend on the fact that the full

198 set of pools including fresh C can be optionally simplified in the case of bare-fallow treatments).

## 199 **2.2. Experimental sites**

200 We used data from a network of six long-term bare-fallow experimental sites (LTBF) in Europe  
201 (with two fields located in Askov, Denmark; Barré et al., 2010), to test the ability of the models to  
202 represent SOC dynamics. The sites were located at a range of latitudes between 48° to 59° North  
203 (Table 2; Fig. 1a), with experiments running for at least 28 years, which were used as a test bed  
204 for the models to represent SOC dynamics. Table 2 shows the main characteristics of each site and  
205 provides a brief description of the historical land use and management of the area (more details  
206 are given by Barré et al., 2010 and references therein). The documented history of the experimental  
207 sites referred to the presence of agricultural areas (grassland or cropland), without woodlands. Soil  
208 texture provides evidence of variability in soil physical properties, with a gradient of intermediate  
209 situations between the sandy loam of Askov (Denmark) and the clay loam of Ultuna (Sweden).  
210 Water relations (precipitation minus reference evapotranspiration) indicate positive climatic water  
211 balance for the two North Atlantic sites only (Askov in Denmark and Rothamsted in the United  
212 Kingdom). Mean annual temperatures vary from ~6 °C in the Sweden and Russian sites (Ultuna  
213 and Kursk, respectively) to near 11 °C in the two French sites (Grignon and Versailles). Annual  
214 air temperature amplitudes - from about 14 °C in Rothamsted to near 30 °C in Kursk - indicate  
215 that the study sites span a broad thermal gradient (Fig. 1b), which likely leads to different soil  
216 thermodynamics (e.g. Zhu et al., 2019). Two widely used metrics (aridity index and frequency of  
217 heatwaves; Sándor et al., 2017, 2018a, b) were also calculated to complete the climatic analysis of  
218 study sites (Fig. A, supplementary material).

219

220  
 221 Table 2. Long-term bare-fallow experimental sites. Table A in the supplementary material contains  
 222 the summary description of the experimental sites.

		Experimental sites (country)					
		S1, S2	S3	S4	S5	S6	S7
<b>General description</b>		Askov (Denmark)	Grignon (France)	Kursk (Russia)	Rothamsted (United Kingdom)	Ultuna (Sweden)	Versailles (France)
<b>Coordinates</b>	<b>Latitude</b>	55.28	48.51	51.73	51.82	59.49	48.48
	<b>Longitude</b>	9.07	1.55	36.19	0.35	17.38	2.08
<b>Soil</b>	<b>Sand/Silt/Clay (%)</b>	78/12/10 (sandy loam)	16/54/30 (silty clay loam)	5/65/30 (silty clay loam)	13/62/25 (silt loam)	23/41/36 (clay loam)	26/57/17 (silt loam)
	<b>Bulk density (Mg m<sup>-3</sup>)</b>	1.50	1.20	1.13	0.94	1.44	1.30
	<b>Experimental period</b>	<i>Bare-fallow years</i> 1956-1985	1959-2007	1965-2001	1959-2008	1956-2007	1929-2008
	<i>N. of data/replicates</i>	30/4, 29/4	11/6	6/0	14/4	18/4	9/6
	<b>Initial/final carbon stocks (Mg C ha<sup>-1</sup>)</b>	52.1/36.4	41.7/25.4	100.3/79.4	71.7/28.6	42.5/26.9	65.5/22.7
<b>Climate<sup>a</sup></b>	<b>Climate type<sup>b</sup></b>	Dfb (humid continental)	Cfb (oceanic)	Dfb (humid continental)	Cfb (oceanic)	Dfb (humid continental)	Cfb (oceanic)

	<b>Mean annual precipitation total (mm)</b>	890	584	482	723	457	608
	<b>Mean annual cumulative evaporation (mm)<sup>c</sup></b>	578	662	602	630	546	668
	<b>Mean annual air temperature (°C)</b>	7.4	10.7	6.2	9.4	6.0	10.7
	<b>Mean annual air temperature range (°C)<sup>d</sup></b>	17.6	16.8	29.8	14.4	22.8	16.7
<b>Vegetation (historical period)<sup>e</sup></b>	<b>ANPP (g C m<sup>-2</sup> yr<sup>-1</sup>)</b>	1.7	1.1	0.9	1.3	0.9	1.2
	<b>TNPP (g C m<sup>-2</sup> yr<sup>-1</sup>)</b>	3.3	2.2	1.7	2.5	1.7	2.2

223 <sup>a</sup> Climatic analysis was performed on longer periods than the experimental periods: 1956-1987/1929-2008/1944-  
224 2003/1856-2006/1956-1999/1929-2008.

225 <sup>b</sup> Köppen-Geiger climate classification (Kottek et al., 2006).

226 <sup>c</sup> Mean values over the bare-fallow period. Reference evaporation was estimated based on the Thornthwaite (1948)  
227 equation.

228 <sup>d</sup> Mean difference in temperature between the warmest and the coldest month of the year.

229 <sup>e</sup> Estimates of aboveground (ANPP) and total (TNPP) net primary productivity based on the precipitation levels of  
230 each site, as provided by Del Grosso et al. (2008) for non-tree dominated systems.

231

232

(Fig. 1 here)

233

### 235 **2.3. Study design**

236 Model simulations were carried out independently by each modelling team (which included model  
237 developers and users, and field experts of soil C dynamics) on commonly formatted data using  
238 their own approaches and technical background. Harmonising calibration techniques was out of  
239 scope of the inter-comparison exercise. The SOC outputs from each model were compared to data  
240 from the study sites before and after calibration. Calibration mostly focussed on parameters related  
241 to substrate use, C partitioning among pools and decomposition processes. However, rate  
242 equations for C pools often required the calibration of a large number of parameters, which are at  
243 the core of key processes responsible for differences among models in the understanding and  
244 interpretation of SOC processes (number of pools and type of decomposition kinetics used to  
245 represent C turnover). For the uncalibrated (blind test, Bln) simulations, the models were run for  
246 each site using the available data of weather, soil texture and bulk density (model inputs), and the  
247 initial SOC values, with no parameter adjustment other than initialisation based on historical  
248 management and land use. With this information, Bln reflects the ability of the models to simulate  
249 SOC decomposition after plant inputs has stopped, using the original parameter settings and  
250 calibration, simply by removing their components related to new C inputs. At this stage, default  
251 values were mostly used for all decomposition rates. C-pool fraction sizes were adjusted based  
252 only on C-input estimates from the information on land use prior to the establishment of the bare-  
253 fallow treatments.

254 After the blind simulations were completed, SOC measurements taken during the bare-  
255 fallow period were supplied to each modelling group for the calibration work. Details on  
256 management (tillage), which may have influenced the SOC dynamics before the bare-fallow  
257 treatment, were also provided to improve the initialisation process. It was requested that each  
258 modelling group adjust soil parameters to improve the simulations based on the observed data,  
259 using whatever techniques they normally use, and to document the changes. At this stage, models

260 were split into two categories: a) with spin-up (SP) and b) without spin-up (NS). Both SP and NS  
261 models require an initial estimate for SOC content and/or an adjustment of parameters towards  
262 balancing the split between soil C pools. The two classes of models work in the same way using  
263 information about plant residues and root growth that provide the C substrate for SOC dynamics  
264 simulations. NS-type models (e.g. DNDC and RothC) use the initial measured SOC value, where  
265 estimates of C inputs in the background of model runs are obtained with various methods (e.g.  
266 Keel et al., 2017) in order to initialise the SOC pools, which can sometimes be calculated  
267 analytically. In order to keep the legacy effect of previous land-use and past management practices,  
268 in SP models (e.g. DayCent) SOC pools are routinely initialised by running the models to achieve  
269 their own states of equilibrium, where change in C stocks is minimised (e.g. Lardy et al., 2011;  
270 Huntzinger et al., 2013). However, if soils are not at equilibrium (e.g. after a sudden disturbance),  
271 spin-up runs may not always be valid with the risk of starting simulations with biased initial values  
272 (e.g. Wutzler and Reichstein, 2007; Nemo et al., 2017) but a fuller discussion on the “spin-up  
273 problem” (Reynolds et al., 2007) is not within the scope of this paper. Carbon inputs are usually  
274 estimated through sub-models calculating total net primary production (TNPP). As it was not  
275 possible to derive TNPP data from local sources at each study-site, TNPP estimates were obtained  
276 at each site (Table 2) based on precipitation levels according to the approach of Del Grosso et al.  
277 (2008). In this way, the creation of the TNPP database used by modellers was based on an identical  
278 methodology, which is widely used worldwide, though the uncertainty in quantifying productivity  
279 across ecosystems is highlighted (e.g. Wieder et al., 2014).

280 The distinction between SP and NS models can appear somewhat arbitrary as virtually any  
281 model with more than one C pool could be spun-up or, alternatively, a function (or analytical  
282 procedures) can be used to make an initial pool partition. We refer here to common modelling  
283 practice, as performed by users within the constraints imposed by packaged (operational) solutions  
284 of SOC models (for which spin-up procedures may be operationally more difficult) or relying on

285 the procedure suggested by previous experience. For instance, although spin-up equilibrium runs  
286 are documented for RothC (e.g. Herbst et al., 2018), it is common practice to initialise three C  
287 pools for subsequent simulations through an internal routine over 10,000 years, with limited model  
288 inputs including clay fraction and weather, and a pre-defined ratio of decomposable over  
289 recalcitrant plant material (e.g. Xu et al., 2011; Weihermüller et al., 2013). Modellers were left to  
290 choose one option or the other when both were available for use in their models (e.g. C-TOOL).  
291 About 40% of the models (10 models) in the study did not use SP processes and set the initial SOC  
292 values manually (using the initial SOC observation).

293 For each model category (SP and NS), two main modelling approaches were identified: site-  
294 specific *versus* generic (single set of parameter values for all the sites). For the site-specific  
295 approach, at each site users informed models about historical management practices and land uses  
296 such as grassland or cropland (with both SP and NS models), SOC decomposition parameters (only  
297 for SP models) or the partitioning of C among different soil pools (only for NS models). With the  
298 generic (not site-specific) approach, model calibration was not applied separately for each  
299 experimental site but simultaneously on all available multi-location datasets to find for each model  
300 parameter values that would be applicable at regional scales. In this case, multi-location calibration  
301 was used to capture generic model parameter values so that the models could still perform well  
302 across a range of climate and management conditions in Europe (Dechow et al., 2019). Site-  
303 specific and non-site-specific approaches were variously combined with factors affecting model  
304 initialisation/parameterisation (Table 3) to create simulation scenarios Gen (generic), Mix (mixed)  
305 and Spe (specific).

306 Scenario Mix uses a site-specific approach for the initialisation of C pools with both SP and  
307 NS models and, for each model, a unique calibration of decomposition parameters. Fixed  
308 decomposition rate parameters (but not rate modifiers) were maintained at a constant value  
309 throughout all sites (e.g. the maximum passive pool decomposition rate in M25 was set to 0.003



310 yr<sup>-1</sup> at all sites), while site-specific climate and soil textural conditions provided supplementary  
 311 factors driving the actual decomposition curve (likely in the uncalibrated blind simulations as  
 312 well). In scenario Spe, decomposition rates could be changed separately at each experimental site,  
 313 which constrained the modelling to a fitting exercise, but made it possible to explore the spatial  
 314 variability of model parameters. Scenario Gen ignored base histories of each site: arable crops and  
 315 grasslands were not distinguished, past climate conditions were disregarded, and this translated  
 316 into discounting the variability in the TNPP levels among sites affecting the starting SOC level.  
 317

318 Table 3. Modelling approaches and simulation scenarios for spin-up and no spin-up models (Gen:  
 319 generic; Mix: mixed; Spe: specific).

Model category	Factors	Approaches	Calibration scenarios <sup>a</sup>		
			Gen	Mix	Spe
Spin-up (SP) based models	Historical management/land use	Site-specific		X	X
		Non-site-specific	X		
	Decomposition processes	Site-specific			X
		Non-site-specific	X	X	
No spin-up (NS) based models	Partitioning of C pools	Site-specific		X	X
		Non-site-specific	X		
	Decomposition processes	Site-specific			X
		Non-site-specific	X	X	

320 <sup>a</sup> The term ‘generic’, which refers to calibration, here means ‘ubiquitous’ or ‘universal’, since the aim of any model  
 321 is to work well under all conditions, without the need to adjust decomposition coefficients. In this case, the model  
 322 correctly represents the main processes and integrates the main factors to accurately simulate the C cycle. The  
 323 ‘specific’ calibration, which aims at improving the model performance, implicitly suggests an incomplete knowledge  
 324 of the SOC turnover. The ‘specific’ calibration allow exploring the spatial variability of model parameters, but this

325 amplitude (which is not discussed or reported here) may indicate the extend of degree of the knowledge gap in soil  
 326 processes (i.e. model parameters might need a huge adjustment across sites)

327  
 328 Twenty-six modelling teams participated in the blind test. At calibration stage, 17 teams  
 329 completed scenarios Spe and Mix, and 16 the scenario Gen. Some model packages are set to restrict  
 330 access to individual parameter values, which did not allow users to carry out some site-specific  
 331 scenarios (Mix and Spe). The same outputs were obtained with some models (e.g. RothC, DNDC),  
 332 which run blind and generic simulations with non-specific information like the previous land-use  
 333 type (arable crop or grassland) and the historical climate. When results from the blind test were  
 334 exactly equal to outputs from Gen scenario, they were not included for further analysis. Estimated  
 335 and observed SOC values ( $\text{Mg C ha}^{-1}$ ) were compared at blind test and for each calibration  
 336 scenario. The agreement between simulations and observations was evaluated by the inspection of  
 337 time series graphs and, numerically, through a set of performance metrics (Table 4) combining  
 338 difference- and correlation-based metrics (e.g. De Jager et al., 1994; Moriasi al., 2007;  
 339 Confalonieri et al., 2009; Bellocchi et al., 2002, 2010).

340  
 341 Table 4. Model performance metrics (P, predicted value; O, observed value; n, number of P/O  
 342 pairs; i, each of P/O pairs;  $\bar{O}$ , mean of observed values;  $\bar{D}$ , average of the differences between  
 343 predicted and observed values;  $S_D$ , standard deviation of the differences between estimated and  
 344 observed values).

Performance metric	Equation	Unit	Value range and purpose
RRMSE, relative root mean square error	$\text{RRMSE} = 100 \cdot \frac{\sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}}}{\bar{O}}$	%	0 (optimum) to positive infinity: the closer the values are to 0, the better the model performance



---

(Willmott and Wicks, 1980)	closer the values are to 1, the better the model
----------------------------	--

---

345

346 **2.4. Multi-model and ensemble assessment**

347 We first focussed on the quantification of model-data discrepancies and then assessed the

348 uncertainty of the individual models in comparison with the multi-model ensemble. The modelling

349 teams provided deterministic model simulation results according to the protocol established, which

350 meant that: 1) one run was provided for each site; 2) the spread of model results due to parameter

351 uncertainty was not specifically addressed. The latter would have dramatically increased the range

352 of model outputs used within the study and would have confounded the uncertainty in calibrated

353 parameters with the uncertainty in model structure (Wallach and Thorburn, 2017). While the

354 uncertainty in model predictions could be due to parameterisation, model calibration from different

355 users (i.e. ensemble of users within ensemble of models) cannot be regarded as the solution to

356 estimate uncertainty due to parameterization (Confalonieri et al., 2016). As well, different

357 calibration techniques do not seem to be primarily responsible for differences in model

358 performance (Wallach et al., 2020) and the contribution of the initialisation to the uncertainty in

359 SOC changes can be negligible compared to the uncertainty related to the model itself and

360 simulated systems characteristics (Dimassi et al., 2018). As uncertainty could not be associated

361 with any individual simulation, we focussed on the analysis of model residuals. We documented

362 the variability of the multi-model simulation exercise across two stages (blind test and alternative

363 calibration scenarios), while inspecting how the multi-model median (MMM) converged to the

364 observations. We used box-plots to compare the variability of estimates by different models (with

365 focus on multi-year averages) to the observed variability, and we represented model ensembles

366 with MMM, which has the advantage to exclude distinctly biased model members with a

367 disproportionate influence on the mean (Rodríguez et al., 2019). The advantage of using MMM

368 was established in practical studies in crop and grassland modelling but also on a theoretical basis  
369 (Wallach et al., 2018).

370 We also quantified the relationship among standardised model residuals of SOC, based on  
371 uncalibrated (BIn) and calibrated (Gen, Mix, Spe) simulations. Moreover, we quantified the  
372 relationship between residuals of agro-climatic metrics (annual values): temperature amplitude,  
373 mean maximum temperature and annual precipitation. Arrays of pairwise scatterplots (scatterplot  
374 matrices) were generated with the panel plot option in the R language and environment for  
375 statistical computing ('panel.smooth', [https://stat.ethz.ch/R-manual/R-  
376 devel/library/graphics/html/panel.smooth.html](https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/panel.smooth.html)), which also overlaid a local non-parametric  
377 smoother curve (locally estimated scatterplot smoothing) on each plot to give some indication of  
378 trends (after Cleveland, 1979).

379 To explore how MMM varied with the number of models in the ensemble, we performed a  
380 calculation for each  $z$ -score transformed MMM,  $z = \frac{MMM - \bar{O}}{sd_{obs}}$ , which was obtained by dividing the  
381 multi-model data deviation from the mean of observations ( $\bar{O}$ ) by the standard deviation of the  
382 observations ( $sd_{obs}$ ) (Sándor et al., 2020). A  $z$ -score can be placed on the normal distribution curve  
383 to indicate how much it deviates from the mean of the distribution. The units of a  $z$ -score are  $sd$   
384 units: zero equals the mean, positive  $z$ -scores exceed the mean, and negative  $z$ -scores are less than  
385 the mean. A  $z$ -score allows comparisons to be made between combinations of models with different  
386 distribution characteristics, i.e. different  $\bar{O}$  and  $sd_{obs}$  (used here as practical descriptors of time-  
387 series central tendency and spread). As illustrated in Fig. 2, different sites occupy distinct zones in  
388 the  $sd_{obs}$  versus  $\bar{O}$  space. Low variability and low mean SOC observations were found at Askov  
389 (S1, S2), Grignon (S3) and Utuna (S6). The variability was higher at Rothamsted (S5) and  
390 Versailles (S7), while the mean was the highest at Kursk (S4). None of the site occupies the upper  
391 right quadrant, i.e. high variability and high mean.

392

393 (Fig. 2 here)

394  
395 We calculated  $z$ -scores for all possible combinations of sets of  $k$  out of  $n=26$  models ( $k=2, \dots n$ ).  
396 The minimum number of models providing plausible estimates at each site was that for which the  
397  $z$ -scores lay within the ranges  $-1$  to  $+1$  or  $-2$  to  $+2$ . The arbitrary choice of these thresholds was  
398 due to a conventional rule, for which values falling within 1 and 2 times the standard deviation  
399 approximate the 68% ( $|z|=1$ ) and 95% ( $|z|=2$ ) confidence limits of a normal distribution,  
400 respectively (after Ehrhardt et al., 2018). R software (<https://cran.r-project.org>) was used for  
401 statistical analysis and graphical visualization.

402

### 403 **3. RESULTS**

#### 404 **3.1. Evaluation of SOC dynamics**

405 Fig. 3 show the range of model results (represented by the shaded area) for each scenario and the  
406 multi-model median (MMM hereinafter) together with the measured values. In general, the  
407 greatest spread of model results was found under the Bln scenario, followed by the Gen scenario.  
408 In some cases, the multi-model median of Bln and Gen scenarios overestimate observations (e.g.  
409 at S5, S6 and S7 sites). As expected, the tightest range of model results (simulation envelope) was  
410 found with site-specific simulations. MMM simulations of Spe came closest to the observations.  
411 All the MMM lines were remarkably close to the observations at sites S1, S2 and S3 (Fig. 3),  
412 despite the much wider spread of the individual simulations, while the MMM at other sites differed  
413 more substantially from the observations (e.g. S5, S6 and S7, Fig. 3). Overall, most of the  
414 simulations (Bln, Gen and Mix) tended to overestimate the amount of SOC (e.g. S5, S6 and S7,  
415 Fig. 3).

416 SOC stocks decreased under all bare-fallow sites during the investigated period. At S1, S2,  
417 S3, S4 and S6 (Fig. 3) sites, the decrease in SOC stock was from minimum to moderate whereas

418 at S5 and S7 (Fig. 3) SOC loss in the top 0.20 m was more rapid, with initial SOC halved during  
419 ~30 years. The decay tended to be more rapid in the first years and then the rate of loss decreased  
420 (e.g. at S7 site between 1929 and 1962, Fig. 3).

421 (Fig. 3 here)

422

### 423 **3.2. Ensemble performance by site**

424 Fig. 4 shows a high variability in the multi-model spread of responses at different sites. The results  
425 show that Kursk (S4) soil, which stored the highest amount of SOC, 91.8 Mg C ha<sup>-1</sup>, was  
426 approximated well by the models, mainly with calibration scenario Spe, with a MMM value of  
427 90.1 Mg C ha<sup>-1</sup>. For calibration scenario Gen, some underestimation is apparent (84.2 Mg C ha<sup>-1</sup>).  
428 Site S4 had the narrowest variability in the measured values, whilst the Bln simulation and  
429 calibration scenario Gen had the highest variability. Measured SOC was well estimated at S1, S2  
430 and S3, including with blind simulations, despite several outlying dots, mainly with Bln and Gen  
431 scenarios. The MMM tended to overestimate the measured SOC at S5 (42.5 Mg C ha<sup>-1</sup>) and S7  
432 (33.0 Mg C ha<sup>-1</sup>) with some scenarios: Bln, S5: 56.7 Mg C ha<sup>-1</sup>, S7: 44.49 Mg C ha<sup>-1</sup>; Mix scenario,  
433 S5: 50.0 Mg C ha<sup>-1</sup>, S7: 35.5 Mg C ha<sup>-1</sup>; Gen scenario, S5: 52.1 Mg C ha<sup>-1</sup>, S7: 40.0 Mg C ha<sup>-1</sup>.  
434 On the other hand, the MMM of Gen scenarios showed the closest values to the observed median  
435 at S5 and S7 (Fig. 4.).

436 Overall, with some exceptions, the MMM of calibrated runs were within the range of the  
437 25<sup>th</sup> and 75<sup>th</sup> percentiles of observations. The Spe scenario provided the best MMM estimation.

438 (Fig. 4 here)

### 439 **3.3. Individual models versus multi-model ensemble**

440 The scatterplot analysis for both each model and the MMM shows that SOC estimates were  
441 improved when moving from the Bln runs (Fig. 5) to the calibration Spe scenario (Fig. 6). Model  
442 performances for calibration Mix and Spe scenarios also showed better simulation results than the

443 Bln simulations (see also Appendix A and Appendix B). Considering all the sites and years, the  
444 predictions of some of the models (e.g. M02, M13, M22, M24 and MMM) were close to the  
445 observations even for the blind level simulations (correlation coefficient  $>0.9$ , Fig. 5). Simulations  
446 improved even further (correlation coefficient  $>0.98$  for half of the models, Fig. 6) under scenario  
447 Spe.

448 All the correlation coefficients of the simulations by other models also considerably improved with  
449 the site-specific data and got closer to the 1:1 line. For instance, for M31, the spread of simulation  
450 data in the blind simulations (Fig. 5) was mainly caused by incorrect initial SOC estimates for the  
451 different sites. When the model was re-run with correctly set initial SOC amounts (Fig. 6), the  
452 subsequent drawdown of SOC over the bare-fallow period was estimated fairly well.

453 Even with blind simulations, MMM gave results in agreement with the observations ( $R^2=0.94$ ).  
454 This level of agreement was only exceeded by M22 ( $R^2=0.95$ ) and approached by M02 ( $R^2=0.92$ )  
455 and M13 ( $R^2=0.90$ ). The MMM simulations continued to give the closest agreement with the  
456 observations even under the full site-specific calibrations ( $R^2=0.99$ ) with several other models  
457 performing equally well (i.e. M02, M05, M09, M13, M23, M26). Overall, with some specific  
458 information for model calibration, many models did remarkably well in reproducing the observed  
459 patterns of SOC loss over time.

460

461 (Fig. 5 here)

462

463 (Fig. 6 here)

464

### 465 **3.4. Analysis of model residuals**

466 The plots of the discrepancy between MMM and observations (Fig. 7) as a function of time shows  
467 a limited scatter (within  $\pm 1$ ) at each site. While Bln, Gen and Mix scenario overestimated the SOC



468 decomposition rate at Kursk (where the highest SOC content was measured), the standardized  
469 residuals were around zero at Grignon and both Askov sites during the whole of experimental  
470 period. However, the departure from observations may increase over time especially with Bln and  
471 Gen scenarios at some site (e.g. at Rothamsted, Ultuna, Versailles) indicating that models  
472 underestimate decomposition rates after a few years/decades.

473

474 (Fig. 7 here)

475

476 Model residuals displayed one versus the other can help establish relationships by exploring the  
477 correlation of residuals from different modelling scenarios, both among them and with external  
478 drivers. Residuals of blind test and calibration scenarios calculated from MMM (Fig. 8) and  
479 individual models (Figs. B1-26 in the supplementary material) were correlated with the mean  
480 annual climate indicators such as the precipitations, maximum temperatures and temperature  
481 amplitudes. When considering the MMM, residuals of Bln were strongly correlated with Gen  
482 ( $r=0.90$ ) and with Mix ( $r=0.59$ ) residuals, but less with Spe ( $r=0.25$ ) residuals, indicating a higher  
483 similarity of the first three approaches, while residuals of Spe were more correlated with those of  
484 Mix ( $r=0.65$ ) than of Gen ( $r=0.39$ ).

485 The most prominent effect of annual climate indicators was found at the blind test stage, whose  
486 residuals were negatively correlated with precipitation ( $r=-0.17$ ) and positively correlated with  
487 Tmax ( $r=0.41$ ). Combinations of high maximum air temperature and low precipitation values may  
488 thus generate greater errors in blind SOC simulations. Calibration scenario Gen did not show  
489 significant correlations to climate indicators. However, calibration scenario Spe and Gen had  
490 opposite correlations. The annual precipitation positively correlated with Spe residuals ( $r=0.26$ )  
491 and with scenario Mix ( $r=0.15$ ). Annual maximum temperature and scenario Spe negatively  
492 correlated ( $r=-0.10$ ). These correlations with climate indicators hint that the site-specific

493 calibration (scenario Spe) is more sensitive to precipitation than to maximum temperatures. On the  
494 contrary, Bln and Gen simulation residuals showed greater sensitivity to maximum temperatures.  
495 Residuals of individual models were approximately equally influenced by precipitation and  
496 temperature drivers, but with differences among models and scenarios (Figs. B1-26 in the  
497 supplementary material). In most of the cases, model residuals were positively correlated with  
498 annual maximum temperatures and negatively correlated with annual precipitation totals (e.g.  
499 M03, M09, M18, M22 for Bln). In some cases, e.g. M09 (Fig. B8 in the supplement), the  
500 correlations among SOC residuals for different scenarios were both positive and negative ( $r$  values  
501 ranged from -0.043 to 0.36), and even the effect of climate indicators were different (e.g. for  $T_{max}$ ,  
502  $r$  values ranged from -0.096 to 0.65). In other cases, e.g. M25 (Fig. B18 in the supplement), SOC  
503 residuals were more similar to each other ( $r$ -values 0.17-0.80) and the effect of precipitation and  
504 temperature drivers was often important (with  $r > 0.4$ ). It is interesting in this respect that the Spe  
505 residuals had near-zero correlations with climatic drivers, showing a lesser influence of these  
506 factors on model results with this scenario, whereas the Bln scenario showed some correlations  
507 with  $T_{amp}$  ( $r=0.13$ ),  $T_{max}$  ( $r=-0.44$ ) and precipitation ( $r=0.40$ ). For M25, Gen scenario residuals  
508 (Fig. B18 in the supplement) appeared unrelated with precipitation ( $r$ -value near zero), but not with  
509 temperature amplitude ( $r=0.50$ ) and maximum air temperature ( $r=-0.56$ ).

510

511 (Fig. 8 here) .

512

### 513 **3.5. Minimum ensemble size**

514 We attempted to identify the minimum number of models required to obtain reliable results for  
515 Bln and calibration scenarios Mix, Spe and Gen (Fig. 9 and Appendix C-E). We observed that  
516 there could be large differences in the  $z$ -scores obtained across sites with different ensemble sizes  
517 and scenarios. Overall, Bln is characterised by greater  $z$ -scores than the calibration scenarios. Our

518 analysis suggests that the ensemble size could be reduced to four models (or even fewer) at S3, S6  
519 and S7. For the other sites (e.g. S4), only ensemble sizes of at least 9-10 models reduced  $z$ -scores  
520 to within the range from -2 to +2, but this number should be raised to 20 or higher to comply with  
521 the most stringent criterion of  $z=|1|$ . A minimum ensemble size of 9-10 models was also identified  
522 with Gen at S4 (Fig. 9), while with Mix and Spe scenarios the number of models could be reduced  
523 down to 7 and 3, respectively (up to about 14 [Gen], 8 [Mix] and 4 [Spe] to comply with  $z=|1|$ )  
524 (Appendix C-E).

525

526 (Fig. 9 here)

527

## 528 4. DISCUSSION

### 529 4.1. Scenarios of ensemble SOC estimates

530 For Bln, Mix, Gen and Spe scenarios, the overall differences between the simulated and the  
531 observed first-year SOC values were  $-0.46$ ,  $+3.49$ ,  $+2.40$  and  $+1.92$  Mg C ha<sup>-1</sup>, respectively, for  
532 the NS models, and  $+0.58$ ,  $-0.29$ ,  $+0.95$  and  $-0.12$  Mg C ha<sup>-1</sup>, respectively, for the SP models.  
533 Despite manually setting the initial SOC values (magnitude of first SOC observation for the  
534 simulation period), the NS models mostly overestimated SOC content in the initial year of the  
535 model run. In first-year estimates of the calibrated (mainly with Spe and Mix scenarios), SP models  
536 deviated less from observations than NS models that overestimated SOC stocks for the first year  
537 with the exception of M25 ( $+8.4$  Mg C ha<sup>-1</sup> for Gen), M29 ( $+18.6$ ,  $+21.1$  and  $+23.7$  Mg C ha<sup>-1</sup> for  
538 Spe, Gen and Mix, respectively) and M31 ( $+25.2$  Mg C ha<sup>-1</sup> for Gen). In the case of M25, the  
539 model was run with a generic grassland spin-up (i.e. 7,000 years), which was applied to all sites.  
540 Thus, a generic history was simulated without considering the cropping history at each site. This  
541 spin-up protocol affected the simulated SOC, showing the poor ability of Gen scenario to produce  
542 results consistent with observations, which questions the practicality of spin-up processes under

543 generic calibration. With M31, there was a greater difference between simulated and observed  
544 SOC values in the initial simulation year and the model gave results that did not correspond to the  
545 observations at all sites (Appendix F), especially under the Bln and Gen scenarios. Though M31  
546 used the initial SOC observation as default parameter, it failed to reproduce the LTBF dynamics  
547 between sites because of large differences in C input to the soil from the former vegetation during  
548 the spin-up period. Consequently, the starting points of the LTBF simulations differed greatly from  
549 the observations, which were overestimated at S1, S2, S3 and S6, and underestimated at S4.  
550 Overall, Mix and Spe calibrations showed better performance indices than the Gen scenario  
551 (Appendix F). We note, however, that M13, for which the SOC pool sizes (humads and humus)  
552 were generically calibrated across sites, produced low RRMSE for Gen (5.7%).

553 The improved calibration knowledge obtained with the site-specific information also improved  
554 model accuracy. Moving from Bln (with knowledge of weather and soil texture, historical land use  
555 and management, and initial SOC; section 2.3) to the Gen scenario, we reproduced SOC data in a  
556 number of European bare-fallow experimental sites with a single set of calibrated, regional-scale  
557 parameter values (regardless of the possible soil, climate and past land-use dissimilarities between  
558 different sites). According to performance indicators in Appendix F, in the Bln simulations the NS  
559 models performed better than the SP models. For instance, average RRMSE and EF were 19.44%  
560 and 0.60, and 26.94% and 0.24, for NS and SP models, respectively. Compared to the Bln scenario,  
561 the discrepancy between the measured and estimated SOC values under the Gen scenario was  
562 slightly reduced with NS models and increased with SP models. Multi-site calibration can be  
563 characterised by lower uncertainty than site-specific calibration, because more data contribute to  
564 the calibration process (e.g. Minunno et al., 2014; Ma et al., 2015). The availability of a variety of  
565 detailed data from multiple sites thus offers the possibility of a genuine multi-location calibration  
566 of the model, assuming that a single calibration across sites is appropriate. The limit of the Gen  
567 scenario calibration was that it did not make it possible to explore the spatial variability of model

568 parameters. The latter was done with scenarios Mix and Spe, for which a basic requisite is that  
569 model parameters are not hard coded but configuration files are left open to the users. From Gen  
570 to Mix, parameters describing initial values of each pool were determined separately for each site.  
571 Moving from Mix to Spe, the decomposition parameters became site-specific. Hence, modellers  
572 needed to invest increasingly more knowledge (and more time-demanding calibration effort) than  
573 in Gen. Under these conditions, the improvement of simulations in SP models was evident (up to  
574 70% for some indicators, e.g. RRMSE and EF). On the contrary, NS models only had a slight  
575 improvement in accuracy of simulations from Bln (RRMSE=21.5%; EF=0.58) to Mix  
576 (RRMSE=18.6%, EF=0.55) or Gen (RRMSE=20.5%; EF=0.45). In our analysis, the two types of  
577 models (NS and SP) appear to be suitable for different sets of data. NS-type models, in most cases,  
578 can perform well even when data are limited to climate, initial C and historic land use, while SP  
579 models generally benefit from the availability of more detailed data. All metrics related to the  
580 performance of the SP models were improved with calibration. There were some differences in  
581 model performance among the sites, but site-specific soil or climatic conditions cannot easily  
582 explain such differences.

583 Overall, across the seven LTEs and using simulated and observed SOC data at the end of the  
584 experimental period we observe that the greatest and least differences from observations were  
585 approximately +14.3% with Bln and +2.2% with Spe (Fig. 10). The Gen scenario achieved almost  
586 half the error (+8.9%) of its closest competitor, i.e. the Bln scenario. More than one-third of the  
587 Bln-scenario error is achievable with the Mix scenario (+4.0%).

588

589

(Fig. 10 here)

590

591 This study has shown that it is difficult to define an *a priori* criterion that could be used to select  
592 a subset of models that would perform better than others would. In terms of the minimum number

593 of models required to obtain reliable results, our study indicates that the suggested minimum  
594 ensemble size (~10 models) proposed by Martre et al. (2015) for crop growth could be a reference  
595 also when model ensembles are implemented to blindly simulate SOC in bare-fallow soils, which  
596 can be reduced down to 3-4 models with a site-specific calibration. These sizes are lower than that  
597 found by Sándor et al. (2020) to provide reliable C-flux estimates in croplands and grasslands (i.e.  
598 ~13 models). While the current study applied the same methodology as Sándor et al. (2020), but  
599 as the present study focuses on one output variable only, SOC, evaluated in simplified systems  
600 (bare-fallow soils), its relative ease of simulation offers great advantages for scenario analyses in  
601 the absence of vegetation cover and plant residues, nor farming practices (only occasional tillage  
602 operations occurred at some sites and were considered by models which can simulate this option).  
603 This is reflected in the several z-scores within the range of -2 and +2, as obtained with a limited  
604 number of models, showing that reduced ensemble sizes can satisfactorily estimate the SOC  
605 content in bare-fallow systems, mainly when site-specific calibration is possible. However, our  
606 analysis of the Russian site (S4), which had low observed variability and high mean ( $sd_{obs}=6.9$ ,  $\bar{O}$   
607  $=91.8$  Mg C ha<sup>-1</sup>), is challenging because it showed that model ensembles that are too small might  
608 not always guarantee sufficient accuracy in SOC estimates of C-rich soils. An application to the  
609 peatlands located on the Mid-Russian Upland (e.g. Shumilovskikh et al., 2018) should thus be  
610 considered with caution.

611

#### 612 **4.2. Possibilities for model inaccuracies**

613 We presented an approach that uses a correlation matrix (with graphical representation) to account  
614 for possible correlations between Bln, Mix, Gen and Spe residuals and, additionally, climatic  
615 factors (mean air temperature amplitude, maximum air temperature and precipitation total). This  
616 residual analysis helps find correlations among alternative scenarios, which might indicate  
617 comparable scenarios in which error propagation within models is similar, though the way of error

618 propagation cannot be easily retrieved from the correlation matrix. This is the case of Bln, Gen  
619 and Mix, whose residuals are highly correlated, while the weak correlations between Spe and other  
620 scenarios highlight the distinct behaviour of the latter. This analysis can also help find correlations  
621 between the SOC output and external drivers, and thus suggest additional predictors that may need  
622 to be included in the models (e.g. Medlyn et al., 2005). This need emerged especially when specific  
623 models were run under Bln, Gen and Mix scenarios, for which some correlations ( $r > |0.4|$ ) were  
624 obtained between model residuals and drivers of thermal and moisture conditions. A weaker but  
625 significant correlation ( $r = 0.26$ ,  $p = 0.02$ ) was also obtained between Spe residuals and precipitation.  
626 These correlations indicate some limitations related to the response functions of SOC  
627 decomposition to soil temperature and soil moisture, though the relative uncertainties of our model  
628 ensemble are attenuated by the presence in the models of physical and chemical processes that  
629 explain the intra- and inter-annual variability of SOC. We add that such biophysical conditions  
630 affect the microbial activity (e.g. Blagodatskaya and Kuzyakov, 2008; Guenet et al., 2010; Wutzler  
631 and Reichstein, 2013), and care should be taken when extrapolating our results over long time  
632 frames (especially without locally calibrated models, Fig. 7) if no corroborating field evidence for  
633 long-term decay rates can be obtained (e.g. on how models are dealing such situations in which  
634 microbes become increasingly C limited as no new C input by plants occurs; Kuhry and Vitt,  
635 1996).

636

## 637 **5. CONCLUSIONS AND FUTURE DIRECTIONS**

638 This paper on SOC modelling offers a tentative answer to the questions about: (i) whether and to  
639 what extent an ensemble of models performs better than single models, (ii) the minimum ensemble  
640 size that is required to reduce the error below a given threshold, and (iii) the set of data required  
641 to prepare and substantiate ensemble estimates. This study presents a framework for interpretation  
642 of model performance and uncertainties obtained with a set of process-based biogeochemical

643 models (individually and in an ensemble) simulating soil C contents in bare-fallow experimental  
644 systems at a variety of European sites. One of the features of SOC modelling today is the huge  
645 amount and variety of models available. Although our analysis did not take into account all sources  
646 of uncertainty (e.g. the influence of the unique choices made by modellers), it enabled the  
647 integration of several modelling teams into an ensemble protocol. Classifying and comparing  
648 different approaches have revealed great model diversity, and is the basis for the development of  
649 dedicated ensemble protocols. In this model inter-comparison, the need to accommodate  
650 challenges experienced by modellers (including C pools of different nature, and optional  
651 initialisation and calibration procedures) was reflected in the co-creation (with modellers and data  
652 providers) of alternative calibration scenarios (Mix, Gen, Spe). As far as we are aware, no previous  
653 multi-model inter-comparison studies have examined differences in such calibration scenarios or  
654 differences between models with or without spin-up.

655 In our study, we did not aim to identify the best model(s) for simulating SOC dynamics for bare-  
656 fallows and no probability of success was assigned to prove the suitability of using one model  
657 rather than another. Overall, we showed that a calibration scenario with generic system knowledge  
658 was adequate for providing sufficiently reliable output, but additional site-specific knowledge can  
659 further improve results under certain circumstances. This is operationally relevant because the  
660 effort required to gather calibration data might no longer be feasible for modelling scenarios  
661 moving from single sites to increasingly larger spatial scales. Site-specific calibration could help  
662 refine model estimates. However, geographical locations have characteristics (e.g. soil and climate  
663 conditions, past history) that require specific model structures and local optimisation, and the  
664 application of models may be limited by the ability to provide representative parameter values.  
665 Soil-C model inter-comparisons including more models and experimental data from other regions  
666 should be continued to improve our ability to simulate biogeochemical processes with acceptable  
667 accuracy. Additional assessments are also recommended to complete the analysis of model



668 behaviour in the long term (like thousands of years) with constant inputs. While the various models  
669 evaluated here did not include all available modelling approaches used to simulate soil C  
670 dynamics, the present model inter-comparison was large compared to other studies. As such, it is  
671 a distinct improvement over previously published quantitative approaches because it represents a  
672 reasonable sub-population of common and current approaches. In this, we offer a method to allow  
673 a broad ensemble of models to be implemented using existing datasets and current modelling  
674 practices. Overall, this multi-model ensemble sets a precedent for key progress in soil C modelling  
675 because it provides essential information about SOC modelling and opens a path to a more in-  
676 depth analysis of the response of individual models and their uncertainties against soil and climate  
677 drivers. Now that we have examined SOC decomposition in-depth without the difficulties of C  
678 input uncertainties, a similar modelling study should be conducted on LTEs that examine both  
679 plant derived C inputs as well as C inputs from manures and other organic materials recycled in  
680 agroecosystems. In fact, under field conditions, the amount of C input is not only an important  
681 factor driving the changes in SOC stocks (including the changes due to tillage), but the amount of  
682 C input also drives the mineralization rate of the SOC (Mary et al., 2020). How simulation models  
683 compare under such conditions is important for improving our ability to evaluate and achieve  
684 climate C goals. With increasing availability of data and computational resources, there are many  
685 opportunities for the SOC modelling community to enrich its offering and to keep up with evolving  
686 methodologies, which would significantly increase transparency of the underpinning science and  
687 modelling practice. A number of recent actions are ongoing under the guidance of international  
688 initiatives such as the European Joint Programme (EJP) on Soil (<https://projects.au.dk/ejpsoil>).  
689 Started in 2020, the EJP-Soil is undertaking a detailed inventory of models and all available data  
690 sources (e.g. world soil maps, satellite images, downscaled weather data), and appears as an ideal  
691 arena to facilitate the exchange of information and to further explore SOC model developments  
692 and practice.

**693 ACKNOWLEDGEMENTS**

694 This study was supported by the project “C and N models inter-comparison and improvement to  
695 assess management options for GHG mitigation in agro-systems worldwide” (CN-MIP, 2014-  
696 2017), which received funding by a multi-partner call on agricultural greenhouse gas research of  
697 the Joint Programming Initiative ‘FACCE’ through national financing bodies. S. Recous, R.  
698 Farina, L. Brilli, G. Bellocchi and L. Bechini received mobility funding by way of the French-  
699 Italian GALILEO programme (CLIMSOC project). The authors acknowledge particularly the data  
700 holders for the Long Term Bare-Fallows, who made their data available and provided additional  
701 information on the sites: V. Romanenkov, B.T. Christensen, T. Kätterer, S. Houot, F. van Oort, A.  
702 Mc Donald, as well as P. Barré. The input of B. Guenet and C. Chenu contributes to the ANR  
703 “Investissements d’avenir” programme with the reference CLAND ANR-16-CONV-0003. The  
704 input of P. Smith and C. Chenu contributes to the CIRCASA project, which received funding from  
705 the European Union's Horizon 2020 Research and Innovation Programme under grant agreement  
706 no 774378 and the projects: DEVIL (NE/M021327/1) and Soils-R-GRREAT (NE/P019455/1).  
707 The input of B. Grant and W. Smith was funded by Science and Technology Branch, Agriculture  
708 and Agri-Food Canada, under the scope of project J-001793. The input of A. Taghizadeh-Toosi  
709 was funded by Ministry of Environment and Food of Denmark as part of the SINKS2 project. The  
710 input of M. Abdalla contributes to the SUPER-G project, which received funding from the  
711 European Union's Horizon 2020 Research and Innovation Programme under grant agreement no  
712 774124.

713

**714 AUTHOR CONTRIBUTIONS**

715 R. Farina, R. Sándor and G. Bellocchi coordinated the study, contributed to its design, conducted  
716 the analysis of data and produced the first draft of the manuscript. P. Smith, C. Chenu, F. Ehrhardt,  
717 M. A. Bolinder, C. Nendel and J.-F. Soussana contributed to the design of the study and the writing

718 of the manuscript. M. Abdalla, J. Álvaro-Fuentes, M. A. Bolinder, L. Brillì, H. Clivot, M. De  
719 Antoni, C. Di Bene, C. D. Dorich, F. Ferchaud, N. Fitton, R. Francaviglia, U. Franko, D. Giltrap,  
720 B. B. Grant, B. Guenet, M. T. Harrison, M. U. F. Kirschbaum, K. Kuka, L. Kulmala, J. Liski, M.  
721 J. McGrath, E. Meier, L. Menichetti, F. Moyano, N. Reibold, A. Shepherd, W. N. Smith, T. Stella,  
722 A. Taghizadeh-Toosi and E. Tsutskikh performed the model calibrations and runs.  
723 C. Dorich, L. Bechini, L. Menichetti, R. Francaviglia, S. Recous, W. Smith, F. Ferchaud, H. Clivot,  
724 M. A. Bolinder, W. Smith, A. Taghizadeh-Toosi, L. Brillì, R. Farina, G. Bellocchi, T. Stella and  
725 U. Franko discussed and decided upon the modelling scenarios at the CN-MIP final meeting  
726 (Rome, 6-7 June 2018). C. Dorich prepared a detailed protocol for second-stage simulations.  
727 Those interested in the details of the modelling process are encouraged to contact authors.

728

#### 729 **Data Availability Statement**

730 The data that support the findings of this study are available from the corresponding author upon  
731 reasonable request and permission of the third parties (i.e. the data holders for the Long Term  
732 Bare-Fallows, V. Romanenkov, B.T. Christensen, T. Kätterer, S. Houot, F. van Oort, A. Mc  
733 Donald, as well as P. Barré).

734

#### 735 **REFERENCES**

- 736 Abrahamsen, P., & Hansen, S. (2000). Daisy: an open soil-crop-atmosphere system model.  
737 *Environmental Modelling & Software*, **15**, 313-330. <https://doi.org/10.1016/S1364->  
738 [8152\(00\)00003-7](https://doi.org/10.1016/S1364-8152(00)00003-7)
- 739 Addiscott, T. M., & Whitmore, A. P. (1987). Computer simulation of changes in soil mineral  
740 nitrogen and crop nitrogen during autumn, winter and spring. *Journal of Agricultural Science*,  
741 **109**, 141-157. <https://doi.org/10.1017/S0021859600081089>

- 742 Andrén, O., & Kätterer, T. (1997). ICBM: The introductory carbon balance model for exploration  
743 of soil carbon balances. *Ecological Applications*, **7**, 1226-1236. <https://doi.org/10.1890/1051->  
744 [0761\(1997\)007\[1226:ITICBM\]2.0.CO;2](https://doi.org/10.1890/1051-0761(1997)007[1226:ITICBM]2.0.CO;2)
- 745 Andrén, O., Kätterer, T., Karlsson, T., & Eriksson, J. (2008). Soil C balances in Swedish  
746 agricultural soils 1990-2004, with preliminary projections. *Nutrient Cycling in Agroecosystems*,  
747 **81**, 129–144. <https://doi.org/10.1007/s10705-008-9177-z>
- 748 Andriulo, A., Mary, B., & Guerif, J. (1999). Modelling soil carbon dynamics with various cropping  
749 sequences on the rolling pampas. *Agronomie*, **19**, 365–377.  
750 <https://doi.org/10.1051/agro:19990504>
- 751 Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A., ... Wolf, J. (2013).  
752 Uncertainty in simulating wheat yields under climate change. *Nature Climate Change*, **3**, 827–  
753 832. <https://doi.org/10.1038/nclimate1916>
- 754 Barré, P., Eglin, T., Christensen, B. T., Ciais, P., Houot, S., Kätterer, T., ... Chenu, C. (2010).  
755 Quantifying and isolating stable soil organic carbon using long-term bare fallow experiments.  
756 *Biogeosciences*, **7**, 3839-3850. <https://doi.org/10.5194/bg-7-3839-2010>
- 757 Basso, B., Dumont, B., Maestrini, B., Shcherbak, I., Robertson, G. P., Porter, J. R., ... Rosenzweig,  
758 C. (2018). Soil organic carbon and nitrogen feedbacks on crop yields under climate change.  
759 *Agricultural and Environmental Letters*, **3**, 180026. <https://doi.org/10.2134/ael2018.05.0026>
- 760 Bassu, S., Brisson, N., Durand, J. L., Boote, K., Lizaso, J., Jones, J. W., ... Waha, K., 2014. How  
761 do various maize crop models vary in their responses to climate change factors? *Global Change*  
762 *Biology*, **20**, 2301–2320. <https://doi.org/10.1111/gcb.12520>
- 763 Bellocchi, G., Acutis, M., Fila, G., & Donatelli, M. (2002). An indicator of solar radiation model  
764 performance based on a fuzzy expert system. *Agronomy Journal*, **94**, 1222-1233.  
765 <https://doi.org/10.2134/agronj2002.1222>

- 766 Bellocchi, G., Rivington, M., Donatelli, M., & Acutis, M. (2010). Validation of biophysical  
767 models: issues and methodologies. A review. *Agronomy for Sustainable Development*, **30**, 109-  
768 130. <https://doi.org/10.1051/agro/2009001>
- 769 Bispo, A., Andersen, L., Angers, D. A., Bernoux, M., Brossard, M., Cécillon, L., ... Eglin, T.K.  
770 (2017). Accounting for carbon stocks in soils and measuring GHGs emission fluxes from soils:  
771 do we have the necessary standards? *Frontiers in Environmental Science*, **12 July 2017**.  
772 <https://doi.org/10.3389/fenvs.2017.00041>
- 773 Blagodatskaya, E., & Kuzyakov, Y. (2008). Mechanisms of real and apparent priming effects and  
774 their dependence on soil microbial biomass and community structure: critical review. *Biology  
775 and Fertility of Soils*, **45**, 115–131. <https://doi.org/10.1007/s00374-008-0334-y>
- 776 Brillì, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., ... Bellocchi, G. (2017).  
777 Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C  
778 and N fluxes. *Science of the Total Environment*, **598**, 445-470.  
779 <https://doi.org/10.1016/j.scitotenv.2017.03.208>
- 780 Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M. H., Ruget, F., Nicollaud, B., ... Delécolle, R.  
781 (1998). STICS: a generic model for the simulation of crops and their water and nitrogen  
782 balances. I. Theory and parameterization applied to wheat and corn. *Agronomie*, **18**, 311–346.  
783 <https://doi.org/10.1051/agro:19980501>
- 784 Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., ... Sinoquet, H. (2003). An  
785 overview of the crop model STICS. *European Journal of Agronomy*, **18**, 309-332.  
786 [https://doi.org/10.1016/S1161-0301\(02\)00110-7](https://doi.org/10.1016/S1161-0301(02)00110-7)
- 787 Brisson, N., Launay, M., Mary, B., & Baudoin, N. (2008). Conceptual basis, formalizations and  
788 parameterization of the STICS crop model. Paris (France): Editions Quae.

- 789 Campbell, E. E., & Paustian, K. (2015). Current developments in soil organic matter modeling and  
790 the expansion of model applications: a review. *Environmental Research Letters*, **10**, 123004.  
791 <https://doi.org/10.1088/1748-9326/10/12/123004>
- 792 Caruso, T., De Vries, F., Bardgett, R. D., & Lehmann, J. (2018). Soil organic carbon dynamics  
793 matching ecological equilibrium theory. *Ecology and Evolution*, **8**, 11169-11178.  
794 <https://doi.org/10.1002/ece3.4586>
- 795 Cavalli, D., Bellocchi, G., Corti, M., Gallina, P. M., & Bechini, L. (2019). Sensitivity analysis of  
796 C and N modules in biogeochemical crop and grassland models following manure addition to  
797 soil. *European Journal of Soil Science*, **70**, 833-846. <https://doi.org/10.1111/ejss.12793>
- 798 Challinor, A., Martre, P., Asseng, S., Thornton, P., & Ewert, F. (2014). Making the most of climate  
799 impacts ensembles. *Nature Climate Change*, **4**, 77-80. <https://doi.org/10.1038/nclimate2117>
- 800 Chenu, C., Angers, D. A., Barré, P., Derrien, D., Arrouays, D., & Balesdent, J. (2018). Increasing  
801 organic stocks in agricultural soils: Knowledge gaps and potential innovations. *Soil and Tillage  
802 Research*, **188**, 41-52. <https://doi.org/10.1016/j.still.2018.04.011>
- 803 Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am.  
804 Stat. Assoc.* **74**, 829-836. <https://doi.org/10.1080/01621459.1979.10481038>
- 805 Clivot, H., Mouny, J. C., Duparque, A., Dinh, J. L., Denoroy, P., Houot, S., ... Mary, B. (2019).  
806 Modeling soil organic carbon evolution in long-term arable experiments with AMG model.  
807 *Environmental Modelling & Software*, **118**, 99-113.  
808 <https://doi.org/10.1016/j.envsoft.2019.04.004>
- 809 Coleman, K., & Jenkinson, D.S. (1999). RothC-26.3 - A model for the turnover of carbon in soil:  
810 model description and Windows user guide. Harpenden (UK): Lawes Agricultural Trust.
- 811 Confalonieri, R., Acutis, M., Bellocchi, G., & Donatelli, M. (2009). Multi-metric evaluation of the  
812 models WARM, CropSyst, and WOFOST for rice. *Ecological Modelling*, **220**, 1395-1410.  
813 <https://doi.org/10.1016/j.ecolmodel.2009.02.017>

- 814 Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., ... Acutis, M.  
815 (2016). Uncertainty in crop model predictions: what is the role of users? *Environmental*  
816 *Modelling & Software*, **81**, 165-173. <https://doi.org/10.1016/j.envsoft.2016.04.009>
- 817 Coucheney, E., Buis, S., Launay, M., Constantin, J., Mary, B., García de Cortázar-Atauri, I., ...  
818 Léonard, J. (2015). Accuracy, robustness and behavior of the STICS soil–crop model for plant,  
819 water and nitrogen outputs: Evaluation over a wide range of agro-environmental conditions in  
820 France. *Environmental Modelling & Software*, **64**, 177-190.  
821 <https://doi.org/10.1016/j.envsoft.2014.11.024>
- 822 De Jager, J.M. (1994). Accuracy of vegetation evaporation ratio formulae for estimating final  
823 wheat yield. *Water SA*, **20**, 307-314. Retrieved from  
824 [https://journals.co.za/content/waters/20/4/AJA03784738\\_2194](https://journals.co.za/content/waters/20/4/AJA03784738_2194)
- 825 Debreczeni, K., & Körschens, M. (2003). Long-term field experiments of the world. *Archives of*  
826 *Agronomy and Soil Science*, **49**, 465-483. <https://doi.org/10.1080/03650340310001594754>
- 827 Dechow, R., Franko, U., Kätterer, T., & Kolbe, H. (2019). Evaluation of the RothC model as a  
828 prognostic tool for the prediction of SOC trends in response to management practices on arable  
829 land. *Geoderma*, **337**, 463-478. <https://doi.org/10.1016/j.geoderma.2018.10.001>
- 830 Del Grosso, S. J., Parton, W. J., Mosier, A. R., Hartman, M. D., Brenner, J., Ojima, D. S., &  
831 Schimel, D. S. (2001). Simulated interaction of carbon dynamics and nitrogen trace gas fluxes  
832 using the DayCent model. In M. J. Shaffer, L. Ma, & S. Hansen (Eds.), *Modeling carbon and*  
833 *nitrogen dynamics for soil management* (pp. 303-332). Boca Raton: CRC Press.
- 834 Del Grosso, S., Ojima, D., Parton, W., Mosier, A., Peterson, G., & Schimel, D. (2002). Simulated  
835 effects of dryland cropping intensification on soil organic matter and greenhouse gas exchanges  
836 using the DAYCENT ecosystem model. *Environmental Pollution*, **1**, S75-S83.  
837 [https://doi.org/10.1016/S0269-7491\(01\)00260-3](https://doi.org/10.1016/S0269-7491(01)00260-3)

- 838 Del Grosso, S., Parton, W., Stohlgren, T., Zheng, D., Bachelet, D., Prince, S., ... Olson, R. (2008).  
839 Global potential net primary production predicted from vegetation class, precipitation, and  
840 temperature. *Ecology*, **89**, 2117-2126. <https://doi.org/10.1890/07-0850.1>
- 841 Dimassi, B., Guenet, B., Saby, N. P. A., Munoz, F., Bardy, M., Millet, F., & Martin, M. P. (2018).  
842 The impacts of CENTURY model initialization scenarios on soil organic carbon dynamics  
843 simulation in French long-term experiments. *Geoderma*, **311**, 25-36.  
844 <https://doi.org/10.1016/j.geoderma.2017.09.038>
- 845 Dungait, J. A. J., Hopkins, D. W., Gregory, A. S., & Whitmore, A. P. (2012). Soil organic matter  
846 turnover is governed by accessibility not recalcitrance. *Global Change Biology*, **18**, 1781-1796.  
847 <https://doi.org/10.1111/j.1365-2486.2012.02665.x>
- 848 Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., Mcauliffe, R., Recous, S., ... Zhang, Q.  
849 (2018). Assessing uncertainties in crop and pasture ensemble model simulations of productivity  
850 and N<sub>2</sub>O emissions. *Global Change Biology*, **24**, e603-e616. <https://doi.org/10.1111/gcb.13965>
- 851 Ehrmann, J., & Ritz, K. (2014). Plant: soil interactions in temperate multi-cropping production  
852 systems. *Plant and Soil*, **376**, 1-29. <https://doi.org/10.1007/s11104-013-1921-8>
- 853 Falloon, P., & Smith, P. (2010). Modelling soil carbon dynamics. In W. L. Kutsch, M. Bahn, & A.  
854 Heinemeyer (Eds.), *Soil carbon dynamics: An integrated methodology* (pp. 221-244).  
855 Cambridge: Cambridge University Press.
- 856 Farina, R., Coleman, K., & Whitmore, A. P. (2013). Modification of the RothC model for  
857 simulations of soil organic C dynamics in dryland regions. *Geoderma*, **200-201**, 18-30.  
858 <https://doi.org/10.1016/j.geoderma.2013.01.021>
- 859 Franko, U., Kolbe, H., Thiel, E., & Liess, E. (2011). Multi-site validation of a soil organic matter  
860 model for arable fields based on generally available input data. *Geoderma*, **166**, 119-134.  
861 <https://doi.org/10.1016/j.geoderma.2011.07.019>



- 862 Franko, U., & Spiegel, H. (2016). Modeling soil organic carbon dynamics in an Austrian long-  
863 term tillage field experiment. *Soil and Tillage Research*, **156**, 83-90.
- 864 Franko, U., & Merbach, I. (2017). Modelling soil organic matter dynamics on a bare fallow  
865 Chernozem soil in Central Germany. *Geoderma*, **303**, 93-98.  
866 <https://doi.org/10.1016/j.geoderma.2017.05.013>
- 867 Fuchs, R., Schulp, C. J. E., Hengeveld, G. M., Verburg, P. H., Clevers, J. G. P. W., Schelhaas, M.-  
868 J., & Herold, M. (2016). Assessing the influence of historic net and gross land changes on the  
869 carbon fluxes of Europe. *Global Change Biology*, **22**, 2526-2539.  
870 <https://doi.org/10.1111/gcb.13191>
- 871 Gardi, C., Visioli, G., Conti, F. D., Scotti, M., Menta, C., & Bodini, A. (2016). High Nature Value  
872 Farmland: assessment of soil organic carbon in Europe. *Frontiers in Environmental Science*, 21  
873 June 2016. <https://doi.org/10.3389/fenvs.2016.00047>
- 874 Gijssman, A. J., Hoogenboom, G., Parton, W. J., & Kerridge, P. C. (2002). Modifying DSSAT crop  
875 models for low-input agricultural systems using a soil organic matter-residue module from  
876 CENTURY. *Agronomy Journal*, **94**, 462-474. <https://doi.org/10.2134/agronj2002.4620>
- 877 Gottschalk, P., Smith, J. U., Wattenbach, M., Bellarby, J., Stehfest, E., Arnell, N., ... Smith, P.  
878 (2012). How will organic carbon stocks in mineral soils evolve under future climate? Global  
879 projections using RothC for a range of climate change scenarios. *Biogeosciences*, **9**, 3151-3171.  
880 <https://doi.org/10.3390/soilsystems3020028>
- 881 Gross C. D., & Harrison, R. B. (2019). The case for digging deeper: soil organic carbon storage,  
882 dynamics, and controls in our changing world. *Soil Systems*, **3**, 28.  
883 <https://doi.org/10.3390/soilsystems3020028>
- 884 Guenet, B., Neill, C., Bardoux, G., & Abbadie, L. (2010). Is there a linear relationship between  
885 priming effect intensity and the amount of organic matter input? *Applied Soil Ecology*, **46**, 436-  
886 442. <https://doi.org/10.1016/j.apsoil.2010.09.006>

- 887 Herbst, M., Welp, G., Macdonald, A., Jate, M., Hädicke, A., Scherer, H., ... Vanderborght, J.  
888 (2018). Correspondence of measured soil carbon fractions and RothC pools for equilibrium and  
889 non-equilibrium states. *Geoderma*, **314**, 37-46.  
890 <https://doi.org/10.1016/j.geoderma.2017.10.047>
- 891 Hill, M. J. (2003). Generating generic response signals for scenario calculation of management  
892 effects on carbon sequestration in agriculture: approximation of main effects using CENTURY.  
893 *Environmental Modelling & Software*, **18**, 899-913. <https://doi.org/10.1016/S1364->  
894 [8152\(03\)00054-9](https://doi.org/10.1016/S1364-8152(03)00054-9)
- 895 Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., ...  
896 Keating, B. A. (2014). APSIM - Evolution towards a new generation of agricultural systems  
897 simulation. *Environmental Modelling & Software*, **62**, 327-350.  
898 <https://doi.org/10.1016/j.envsoft.2014.07.009>
- 899 Huntzinger, D. N., Schwalm, C., Michalak, A. M., Schaefer, K., King, A. W., Wei, Y., ... Zhu, Q.  
900 (2013). The North American Carbon Program Multi-scale synthesis and Terrestrial Model  
901 Intercomparison Project-Part 1: Overview and experimental design. *Geoscientific Model*  
902 *Development*, **6**, 2121-2133. <https://doi.org/10.5194/gmd-6-2121-2013>
- 903 Johnston, A. E., & Poulton, P. R. (2018). The importance of long-term experiments in agriculture:  
904 their management to ensure continued crop production and soil fertility; the Rothamsted  
905 experience. *European Journal of Soil Science*, **69**, 113-125. <https://doi.org/10.1111/ejss.12521>
- 906 Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., ...  
907 Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, **18**,  
908 235–265. [https://doi.org/10.1016/S1161-0301\(02\)00107-7](https://doi.org/10.1016/S1161-0301(02)00107-7)
- 909 Jørgensen, S. E., Kamp-Nielsen, L., Christensen, T., Windolf-Nielsen, J., & Westergaard, B.  
910 (1986). Validation of a prognosis based upon a eutrophication model. *Ecological Modelling*,  
911 **35**, 165-182. [https://doi.org/10.1016/0304-3800\(86\)90024-4](https://doi.org/10.1016/0304-3800(86)90024-4)

- 912 Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. L., Robertson, M. J., Holzworth, D.,  
913 ... Smith, C. J. (2003). An overview of APSIM, a model designed for farming systems  
914 simulation. *European Journal of Agronomy*, **18**, 267-288. [https://doi.org/10.1016/S1161-](https://doi.org/10.1016/S1161-0301(02)00108-9)  
915 [0301\(02\)00108-9](https://doi.org/10.1016/S1161-0301(02)00108-9)
- 916 Keel, S. G., Leifeld, J., Mayer, J., Taghizadeh-Toosi, A., and Olesen, J. E. (2017). Large  
917 uncertainty in soil carbon modelling related to method of calculation of plant carbon input in  
918 agricultural systems. *European Journal of Soil Science*, **68**, 953-863.  
919 <https://doi.org/10.1111/ejss.12454>
- 920 Kirschbaum, M.U.F. (1999). CenW, a forest growth model with linked carbon, energy, nutrient  
921 and water cycles. *Ecological Modelling*, **118**, 17–59. [https://doi.org/10.1016/S0304-](https://doi.org/10.1016/S0304-3800(99)00020-4)  
922 [3800\(99\)00020-4](https://doi.org/10.1016/S0304-3800(99)00020-4)
- 923 Kirschbaum, M. U. F., Rutledge, S., Kuijper, I. A., Mudge, P. L., Puche, N., Wall, A. M., ...  
924 Campbell, D. I. (2015). Modelling carbon and water exchange of a grazed pasture in New  
925 Zealand constrained by eddy covariance measurements. *Science of the Total Environment*, **512-**  
926 **513**, 273-286. <https://doi.org/10.1016/j.scitotenv.2015.01.045>
- 927 Kirschbaum, M. U. F., & Paul, K. I. (2002). Modelling carbon and nitrogen dynamics in forest  
928 soils with a modified version of the CENTURY model. *Soil Biology & Biochemistry*, **34**, 341-  
929 354. [https://doi.org/10.1016/S0038-0717\(01\)00189-4](https://doi.org/10.1016/S0038-0717(01)00189-4)
- 930 Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger  
931 climate classification updated. *Meteorologische Zeitschrift*, **15**, 259-263.  
932 <https://doi.org/10.1127/0941-2948/2006/0130>
- 933 Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., ... Colin  
934 Prentice, I. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-  
935 biosphere system. *Global Biogeochemical Cycles*, **19**, GB1015.  
936 <https://doi.org/10.1029/2003GB002199>

- 937 Kuhry, P., & Vitt, D.H. (1996). Fossil carbon/nitrogen ratios as a measure of peat decomposition.  
938 *Ecology*, **77**, 271–275. <https://doi.org/10.2307/2265676>
- 939 Kuka, K. (2005). Modellierung des Kohlenstoffhaushaltes in Ackerböden auf der Grundlage  
940 bodenstrukturabhängiger Umsatzprozesse. PhD thesis, Martin-Luther-University Halle-  
941 Wittenberg. Retrieved from  
942 <https://gepris.dfg.de/gepris/projekt/5247578?context=projekt&task=showDetail&id=5247578>  
943 & (in German)
- 944 Kuka, K., Franko, U., & Rühlmann, J. (2007) Modelling the impact of pore space distribution on  
945 carbon turnover. *Ecological Modelling*, **208**, 295–306.  
946 <https://doi.org/10.1016/j.ecolmodel.2007.06.002>
- 947 Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security.  
948 *Science*, **304**, 1623-1626. <https://doi.org/10.1126/science.1097396>
- 949 Lal, R. (2014). Soil conservation and ecosystem services. *International Soil and Water*  
950 *Conservation Research*, **2**, 36-47. [https://doi.org/10.1016/S2095-6339\(15\)30021-6](https://doi.org/10.1016/S2095-6339(15)30021-6)
- 951 Lardy, R., Bellocchi, G., & Soussana, J.-F. (2011). A new method to determine soil organic carbon  
952 equilibrium. *Environmental Modelling & Software*, **26**, 1759-1763.  
953 <https://doi.org/10.1016/j.envsoft.2011.05.016>
- 954 Lavallee, J. M., Soong, J. L., & Cotrufo, M. F. (2020). Conceptualizing soil organic matter into  
955 particulate and mineral-associated forms to address global change in the 21<sup>st</sup> century. *Global*  
956 *Change Biology*, **26**, 261-273. <https://doi.org/10.1111/gcb.14859>
- 957 Lehmann, J., & Kleber, M. (2015). The contentious nature of soil organic matter. *Nature*, **528**, 60-  
958 68. <https://doi.org/10.1038/nature16069>
- 959 Li, C., Salas, W., Zhang, R., Krauter, C., Rotz, A., & Mitloehner, F. (2012). Manure-DNDC: a  
960 biogeochemical process model for quantifying greenhouse gas and ammonia emissions from

- 961 livestock manure systems. *Nutrient Cycling in Agroecosystems*, **93**, 163-200.  
962 <https://doi.org/10.1007/s10705-012-9507-z>
- 963 Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., ... Bouman, B. (2015). Uncertainties  
964 in predicting rice yield by current crop models under a wide range of climatic conditions. *Global*  
965 *Change Biology*, **21**, 1328-1341. <https://doi.org/10.1111/gcb.12758>
- 966 Ma, S., Lardy, R., Graux, A.-I., Ben Touhami, H., Klumpp, K., Martin, R., Bellocchi, G. (2015).  
967 Regional-scale analysis of carbon and water cycles on managed grassland systems.  
968 *Environmental Modelling & Software*, **72**, 356-371.  
969 <https://doi.org/10.1016/j.envsoft.2015.03.007>
- 970 Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., ... Zhu, Y. (2017). Crop  
971 model improvement reduces the uncertainty of the response to temperature of multi-model  
972 ensembles. *Field Crops Research*, **202**, 5-20. <https://doi.org/10.1016/j.fcr.2016.05.001>
- 973 Manzoni, S., & Porporato, A. (2009). Soil carbon and nitrogen mineralization: Theory and models  
974 across scales. *Soil Biology & Biochemistry*, **41**, 1355-1379.  
975 <https://doi.org/10.1016/j.soilbio.2009.02.031>
- 976 Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., ... Wolf, J. (2015).  
977 Multimodel ensembles of wheat growth: Many models are better than one. *Global Change*  
978 *Biology*, **21**, 911-925. <https://doi.org/10.1111/gcb.12768>
- 979 Mary, B., Clivot, H., Blaszczyk, N., Labreuche, L., & Ferchaud, F. (2020). Soil carbon storage and  
980 mineralization rates are affected by carbon inputs rather than physical disturbance: Evidence  
981 from a 47-year tillage experiment. *Agriculture, Ecosystems & Environment*, **299**, 106972.  
982 <https://doi.org/10.1016/j.agee.2020.106972>
- 983 edlyn, B. E., Robinson, A. P., Clement, R., & McMurtrie, R. E. (2005). On the validation of models  
984 of forest CO<sub>2</sub> exchange using eddy covariance data: some perils and pitfalls. *Tree Physiology*,  
985 **25**, 839-857. <https://doi.org/10.1093/treephys/25.7.839>

- 986 Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., ...  
987 Winowiecki, L. (2017). Soil carbon 4 per mille. *Geoderma*, **292**, 59–86.  
988 <https://doi.org/10.1016/j.geoderma.2017.01.002>
- 989 Minunno, F., Peltoniemi, M., Launiainen, S., & Mäkelä, A. (2014). Integrating ecosystems  
990 measurements from multiple eddy-covariance sites to a simple model of ecosystem process -  
991 are there possibilities for a uniform model calibration? *Geophysical Research Abstracts*, **16**,  
992 EGU2014-10706-3. Retrieved from  
993 <https://meetingorganizer.copernicus.org/EGU2014/orals/14065>
- 994 Mirtl, M., Borer, E. T., Djukic, I., Forsius, M., Haubold, H., Hugo, W., Jourdane, J., ... Haase, P.  
995 (2018). Genesis, goals and achievements of long-term ecological research at the global scale: a  
996 critical review of ILTER and future directions. *Science of the Total Environment*, **626**, 1439-  
997 1462. <https://doi.org/10.1016/j.scitotenv.2017.12.001>
- 998 Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., & Veith, T. (2007). Model  
999 evaluation guidelines for systematic quantification of accuracy in watershed simulations.  
1000 *Transactions of the ASABE*, **50**, 885-900. <https://doi.org/10.13031/2013.23153>
- 1001 Moyano, F. E., Vasilyeva, N., & Menichetti, L. (2018). Diffusion limitations and Michaelis–  
1002 Menten kinetics as drivers of combined temperature and moisture effects on carbon fluxes of  
1003 mineral soils. *Biogeosciences*, **15**, 5031–5045. <https://doi.org/10.5194/bg-15-5031-2018>
- 1004 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - a  
1005 discussion of principles. *Journal of Hydrology*, **10**, 282-290. [https://doi.org/10.1016/0022-  
1006 1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- 1007 Nemo, R., Klumpp, K., Coleman, K., Dondini, M., Goulding, K., Hastings, A., ... Smith, P.  
1008 (2016). Soil organic carbon (SOC) equilibrium and model initialisation methods: an application  
1009 to the Rothamsted Carbon (RothC) model. *Environmental Modeling & Assessment*, **22**, 215-  
1010 229.

- 1011 Nendel, C., Berg, M., Kersebaum, K. C., Mirschel, W., Specka, X., Wegehenkel, M., ... Wieland,  
1012 R. (2011). The MONICA model: Testing predictability for crop growth, soil moisture and  
1013 nitrogen dynamics. *Ecological Modelling*, **222**, 1614–1625.  
1014 <https://doi.org/10.1016/j.ecolmodel.2011.02.018>
- 1015 Parton, W. J., Del Grosso, S., Plante, A. F., Adair, E. C., & Lutz, S. M. (2015). Modeling the  
1016 dynamics of soil organic matter and nutrient cycling. In E. A. Paul (Ed.), *Soil microbiology,*  
1017 *ecology and biochemistry, 4<sup>th</sup> edition* (pp. 505-537). Amsterdam: Elsevier Academic Press.
- 1018 Parton, W. J., Hartman, M., Ojima, D., & Schimel, D. (1998). DAYCENT and its land surface  
1019 submodel: description and testing. *Global and Planetary Change*, **19**, 35-48.  
1020 [https://doi.org/10.1016/S0921-8181\(98\)00040-X](https://doi.org/10.1016/S0921-8181(98)00040-X)
- 1021 Parton, W. J., Schimel, D. S., & Cole, C.V., & Ojima, D. S. (1987). Analysis of factors controlling  
1022 soil organic matter levels in Great Plains grasslands. *Soil Science Society of America Journal*,  
1023 **51**, 1173–1179. <https://doi.org/10.2136/sssaj1987.03615995005100050015x>
- 1024 Parton, W. J., Schimel, D. S., Ojima, D. S., & Cole, C. V. (1994). A general model for soil organic  
1025 matter dynamics: sensitivity to litter chemistry, texture and management. In R. B. Bryant & R.  
1026 W. Arnold (Eds.), *Quantitative modeling of soil forming processes* (pp. 147–167). Madison,  
1027 WI (USA): SSSA Spec. Pub. 39. ASA, CSSA and SSSA.
- 1028 Porter, C. H., Jones, J. W., Adiku, S., Gijssman, A. J., Gargiulo, O., & Naab, J. B. (2009). Modeling  
1029 organic carbon and carbon-mediated soil processes in DSSAT v4.5. *Operational Research*, **10**,  
1030 247-278. <https://doi.org/10.1007/s12351-009-0059-1>
- 1031 Puche, N. J. B., Senapati, N., Flechard, C. R., Klumpp, K., Kirschbaum, M. U. F, & Chabbi, A.  
1032 (2019). Modelling carbon and water fluxes of managed grasslands: comparing flux variability  
1033 and net carbon budgets between grazed and mowed systems. *Agronomy*, **9**, 183.  
1034 <https://doi.org/10.3390/agronomy9040183>

- 1035 Reynolds, K. M., Thomson, A. J., Köhl, M., Shannon, M. A., Ray, D., & Rennolls, K. (2007).  
1036 Sustainable forestry: from monitoring and modelling to knowledge management and policy  
1037 science. Wallingford: CAB International.
- 1038 Rodríguez, A., Ruiz-Ramos, M., Palosuo, T., Carter, T. R., Fronzek, S., Lorite, I. J., ... Rötter, R.  
1039 P. (2019). Implications of crop model ensemble size and composition for estimates of  
1040 adaptation effects and agreement of recommendations. *Agricultural and Forest Meteorology*,  
1041 **15**, 351-362. <https://doi.org/10.1016/j.agrformet.2018.09.018>
- 1042 Rötter, R. P., Palosuo, T., Kersebaum, K. C., Angulo, C., Bindi, M., Ewert, F., ... Trnka, M.  
1043 (2012). Simulation of spring barley yield in different climatic zones of Northern and Central  
1044 Europe – A comparison of nine crop models. *Field Crops Research*, **133**, 23–36.  
1045 <https://doi.org/10.1016/j.fcr.2012.03.016>
- 1046 Ruane, A. C., Hudson, N. I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., ... Wolf, J. (2016).  
1047 Multi-wheat-model ensemble responses to interannual climate variability. *Environmental*  
1048 *Modelling & Software*, **81**, 86-101. <https://doi.org/10.1016/j.envsoft.2016.03.008>
- 1049 Rumpel, C., Amiraslani, F., Koutika, L. S., Smith, P., Whitehead, D., & Wollenberg, E. (2018).  
1050 Put more carbon in soils to meet Paris climate pledges. *Nature*, **564**, 32-34.  
1051 <https://doi.org/10.1038/d41586-018-07587-4>
- 1052 Saffih-Hdadi, K., & Mary, B. (2008). Modeling consequences of straw residues export on soil  
1053 organic carbon. *Soil Biology & Biochemistry*, **40**, 594–607.  
1054 <https://doi.org/10.1016/j.soilbio.2007.08.022>
- 1055 Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., ... Bellocchi, G. (2017). Multi-  
1056 model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean  
1057 grasslands: Uncertainties and ensemble performance. *European Journal of Agronomy*, **88**, 22-  
1058 40. <https://doi.org/10.1016/j.eja.2016.06.006>



- 1059 Sándor, R., Ehrhardt, F., Brilli, L., Carozzi, M., Recous, S., Smith, P., ... Bellocchi, G. (2018a).  
1060 The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from  
1061 managed grasslands. *Science of the Total Environment*, **642**, 292-306.  
1062 <https://doi.org/10.1016/j.scitotenv.2018.06.020>
- 1063 Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., ... Bellocchi, G. (2020).  
1064 Ensemble modelling of carbon fluxes in grasslands and croplands. *Field Crops Research*, **252**,  
1065 107791. <https://doi.org/10.1016/j.fcr.2020.107791>
- 1066 Sándor, R., Picon-Cochard, C., Martin, R., Louault, F., Klumpp, K., Borrás, D., & Bellocchi, G.,  
1067 (2018b). Plant acclimation to temperature: Developments in the Pasture Simulation model.  
1068 *Field Crops Research*, **222**, 238-255. <https://doi.org/10.1016/j.fcr.2017.05.030>
- 1069 Schimel, J. P., & Weintraub, M. N. (2003). The implications of exoenzyme activity on microbial  
1070 carbon and nitrogen limitation in soil: a theoretical model. *Soil Biology & Biochemistry*, **35**,  
1071 549–563. [https://doi.org/10.1016/S0038-0717\(03\)00015-4](https://doi.org/10.1016/S0038-0717(03)00015-4)
- 1072 Shumilovskikh, L. S., Novenko, E., & Giesecke, T. (2018). Long-term dynamics of the East  
1073 European forest-steppe ecotone. *Journal of Vegetation Science*, **29**, 416-426.  
1074 <https://doi.org/10.1111/jvs.12585>
- 1075 Sitch, S., Smith, B., Prentice, I. C., Arneeth, A., Bondeau, A., Cramer, W., ... Venevsky, S. (2003).  
1076 Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ  
1077 dynamic global vegetation model. *Global Change Biology*, **9**, 161-185.  
1078 <https://doi.org/10.1046/j.1365-2486.2003.00569.x>
- 1079 Smith, J., Gottshalk, P., Bellarby, J., Chapman, S., Lilly, A., Towers, W., ... Smith, P. (2010a).  
1080 Estimating changes in national soil carbon stocks using ECOSSE – a new model that includes  
1081 upland organic soils. Part I. Model description and uncertainty in national scale simulations of  
1082 Scotland. *Climate Research*, **45**, 179-192. <https://doi.org/10.3354/cr00899>

- 1083 Smith, J., Gottschalk, P., Bellarby, J., Chapman, S., Lilly, A., Towers, W., ... Smith, P. (2010b).  
1084 Estimating changes in national soil carbon stocks using ECOSSE - a new model that includes  
1085 upland organic soils. Part II. Application in Scotland. *Climate Research*, **45**, 193-205.  
1086 <https://doi.org/10.3354/cr00902>
- 1087 Smith, P., Smith, J., Flynn, H., Killham, K., Rangel-Castro, I., Foereid, B., ... Falloon, P., 2007.  
1088 ECOSSE: Estimating Carbon in Organic Soils - Sequestration and Emissions. Final Report.  
1089 SEERAD Report, 166 pp. Retrieved from <http://nora.nerc.ac.uk/id/eprint/2233>
- 1090 Smith, P., Smith, J. U., Powlson, D. S., McGill, W. B., Arah, R. M., Chertov, O. G., ... Whitmore,  
1091 A. P. (1997). A comparison of the performance of nine soil organic matter models using datasets  
1092 from seven long-term experiments. *Geoderma*, **81**, 153-225. [https://doi.org/10.1016/S0016-](https://doi.org/10.1016/S0016-7061(97)00087-6)  
1093 [7061\(97\)00087-6](https://doi.org/10.1016/S0016-7061(97)00087-6)
- 1094 Smith, W. N., Grant, B. B., Campbell, C. A., McConkey, B. G., Desjardins, R. L., Kröbel, R. &  
1095 Malhi, S. S. (2012). Crop residue removal effects on soil carbon: Measured and inter-model  
1096 comparisons. *Agriculture, Ecosystems & Environment*, **161**, 27-38.  
1097 <https://doi.org/10.1016/j.agee.2012.07.024>
- 1098 Smith, W. N., Grant, B., Qi, Z., He, W., VanderZaag, A., Drury, C. F., & Helmers, M. (2020).  
1099 Development of the DNDC model to improve soil hydrology and incorporate mechanistic tile  
1100 drainage: A comparative analysis with RZWQM2. *Environmental Modelling & Software*, **123**,  
1101 104577. <https://doi.org/10.1016/j.envsoft.2019.104577>
- 1102 Soussana, J.-F., Lutfalla, S., Ehrhardt, F., Rosenstock, T. S., Lamanna, C., Havlik, P., ... Lal, R.  
1103 (2017). Matching policy and science: Rationale for the '4 per 1000 - soils for food security and  
1104 climate' initiative. *Soil and Tillage Research*, **188**, 3-15.  
1105 <https://doi.org/10.1016/j.still.2017.12.002>
- 1106 Specka, X., Nendel, C., Hagemann, U., Pohl, M., Hoffmann, M., Barkusky, D., ... van Oost, K.  
1107 (2016). Reproducing CO<sub>2</sub> exchange rates o a crop rotation at contrasting terrain positions using

- 1108 two different modelling approaches. *Soil and Tillage Research*, **156**, 219–229.  
1109 <https://doi.org/10.1016/j.still.2015.05.007>
- 1110 Stella, T., Mouratiadou, I., Gaiser, T., Berg-Mohnicke, M., Wallor, E., Ewert, F., & Nendel, C.  
1111 (2019). Estimating the contribution of crop residues to soil organic carbon conservation.  
1112 *Environmental Research Letters* 14, 094008. <https://doi.org/10.1088/1748-9326/ab395c>
- 1113 Taghizadeh-Toosi, A., Christensen, B. T., Hutchings, N. J., Vejlin, J., Kätterer, T., Glendining,  
1114 M., & Olesen, J. E. (2014a). C-TOOL: A simple model for simulating whole-profile carbon  
1115 storage in temperate agricultural soils. *Ecological Modelling*, **292**, 11–25.  
1116 <https://doi.org/10.1016/j.ecolmodel.2014.08.016>
- 1117 Taghizadeh-Toosi, A., Olesen, J. E., Kristensen, K., Elsgaard, L., Østergaard, H. S., Lægdsmand,  
1118 M., ... Christensen, B. T. (2014b). Changes in carbon stocks of Danish agricultural mineral  
1119 soils between 1986 and 2009. *European Journal of Soil Science*, **65**, 730–740.  
1120 <https://doi.org/10.1111/ejss.12169>
- 1121 Taghizadeh-Toosi, A., & Olesen, J. E. (2016). Modelling soil organic carbon in Danish agricultural  
1122 soils suggests low potential for future carbon sequestration. *Agricultural Systems*, **145**, 83–89.  
1123 <https://doi.org/10.1016/j.agsy.2016.03.004>
- 1124 Taghizadeh-Toosi, A., Christensen, B. T., Glendining, M., & Olesen, J. E. (2016). Consolidating  
1125 soil carbon turnover models by improved estimates of belowground carbon input. *Scientific  
1126 Reports*, **6**, 32568. <https://doi.org/10.1038/srep32568>
- 1127 Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical  
1128 Review*, **38**, 55–94. <https://doi.org/10.2307/210739>
- 1129 Thorp, K. R., White, J. W., Porter, C. H., Hoogenboom, G., Nearing, G. S., & French, A. N. (2012).  
1130 Methodology to evaluate the performance of simulation models for alternative compiler and  
1131 operating system configurations. *Computers and Electronics in Agriculture*, **81**, 62–71.  
1132 <https://doi.org/10.1016/j.compag.2011.11.008>

- 1133 Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E.  
1134 A. G., & Allison, S. D. (2013). Causes of variation in soil carbon simulations from CMIP5  
1135 Earth system models and comparison with observations. *Biogeosciences*, **10**, 1717–1736.  
1136 <https://doi.org/10.5194/bg-10-1717-2013>
- 1137 Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., ... Allison,  
1138 S. D. (2014). Changes in soil organic carbon storage predicted by Earth system models during  
1139 the 21st century. *Biogeosciences*, **11**, 2341–2356. <https://doi.org/10.5194/bg-11-2341-2014>
- 1140 Tuomi, M., Thum, T., Järvinen, H., Fronzek, S., Berg, B., Harmon, M., ... Liski, J. (2009). Leaf  
1141 litter decomposition - Estimates of global variability based on Yasso07 model. *Ecological*  
1142 *Modelling*, **220**, 3362–3371. <https://doi.org/10.1016/j.ecolmodel.2009.05.016>
- 1143 Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., ... Zhang, Z. (2018). Multi-  
1144 model ensembles improve predictions of crop-environment-management interactions. *Global*  
1145 *Change Biology*, **24**, 5072–5083. <https://doi.org/10.1111/gcb.14411>
- 1146 Wallach, D., Palosuo, T., Thorburn, P., Seidel, S. J., Gourdain, E., Asseng, S., ... Zhu, Y. (2020).  
1147 How well do crop models predict phenology, with emphasis on the effect of calibration?  
1148 *bioRxiv*, March 30, 2020. <https://doi.org/10.1101/708578>
- 1149 Wallach, D., & Thorburn, P. J. (2017). Estimating uncertainty in crop model predictions: Current  
1150 situation and future prospects. *European Journal of Agronomy*, **88**, A1–A7.  
1151 <https://doi.org/10.1016/j.eja.2017.06.001>
- 1152 Weihermüller, L., Graf, A., Herbst, M., & Vereecken, H. (2013). Simple pedotransfer functions to  
1153 initialize reactive carbon pools of the RothC model. *European Journal of Soil Science*, **64**, 567-  
1154 575. <https://doi.org/10.1111/ejss.12036>
- 1155 White, J. W., Hoogenboom, G., Kimball, B. A., & Wall, G. W. (2011). Methodologies for  
1156 simulating impacts of climate change on crop production. *Field Crops Research*, **124**, 357–368.  
1157 <https://doi.org/10.1016/j.fcr.2011.07.001>

- 1158 Whitehead, D., Schipper, L. A., Pronger, J., Moinet, G. Y., Mudge, P. L., Pereira, R. C., ... Camps-  
1159 Arbestain, M. (2018). Management practices to reduce losses or increase soil carbon stocks in  
1160 temperate grazed grasslands: New Zealand as a case study. *Agriculture, Ecosystems &*  
1161 *Environment*, **265**, 432-443. <https://doi.org/10.1016/j.agee.2018.06.022>
- 1162 Wieder, W. R., Boehnert, J., & Bonan, G. B. (2014). Evaluating soil biogeochemistry  
1163 parameterizations in Earth system models with observations. *Global Biogeochemical Cycles*,  
1164 **28**, 211-222. <https://doi.org/10.1002/2013GB004665>
- 1165 Willmott, C. J., & Wicks, D. E. (1980). An empirical method for the spatial interpolation of  
1166 monthly precipitation within California. *Physical Geography*, **1**, 59-73.  
1167 <https://doi.org/10.1080/02723646.1980.10642189>
- 1168 Wutzler, T., & Reichstein, M. (2007). Soils apart from equilibrium - consequences for soil carbon  
1169 balance modelling. *Biogeosciences*, **4**, 125-136. <https://doi.org/10.5194/bg-4-125-2007>
- 1170 Wutzler, T., & Reichstein, M. (2008). Colimitation of decomposition by substrate and  
1171 decomposers - a comparison of model formulations. *Biogeosciences*, **5**, 749-759.  
1172 <https://doi.org/10.5194/bg-5-749-2008>
- 1173 Wutzler, T., & Reichstein, M. (2013). Priming and substrate quality interactions in soil organic  
1174 matter models. *Biogeosciences*, **10**, 2089-2103. <https://doi.org/10.5194/bg-10-2089-2013>
- 1175 Xu, X., Wen L., & Kiely, G. (2011). Modeling the change in soil organic carbon of grassland in  
1176 response to climate change: Effects of measured versus modelled carbon pools for initializing  
1177 the Rothamsted Carbon model. *Agriculture, Ecosystems & Environment*, **140**, 372-381.  
1178 <https://doi.org/10.1016/j.agee.2010.12.018>
- 1179 Yadav, V., & Malanson, G. (2007). Progress in soil organic matter research: litter decomposition,  
1180 modelling, monitoring and sequestration. *Progress in Physical Geography*, **31**, 131-154.  
1181 <https://doi.org/10.1177/0309133307076478>
- 1182 Zhu, D., Ciais, P., Krinner, G., Maignan, F., Puig,  
A.J., & Hugelius, G. (2019). Controls of soil organic matter on soil thermal dynamics in the

1183 northern high latitudes. *Nature Communications*, **10**, 3172. <https://doi.org/10.1038/s41467->

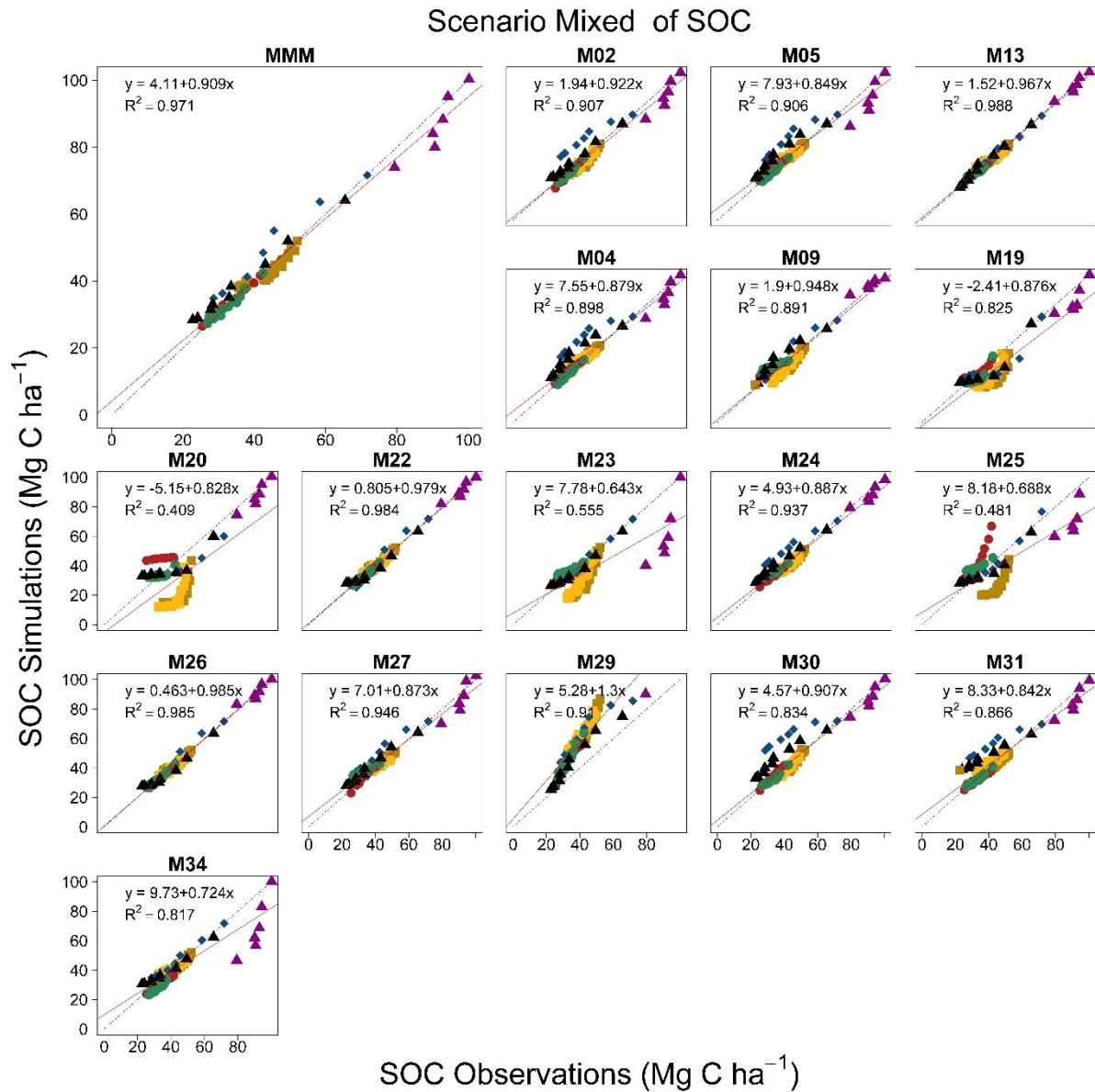
1184 [019-11103-1](https://doi.org/10.1038/s41467-019-11103-1)

1185

1186 **Appendix A**

1187 Multi-year, multi-site comparison of individual model simulation of SOC ( $\text{Mg C ha}^{-1}$ ): multi-  
 1188 model medians (MMM) from Mix scenario simulations (17 models) versus observations.  
 1189 (coloured symbols represent sites as in Fig. 1).

1190

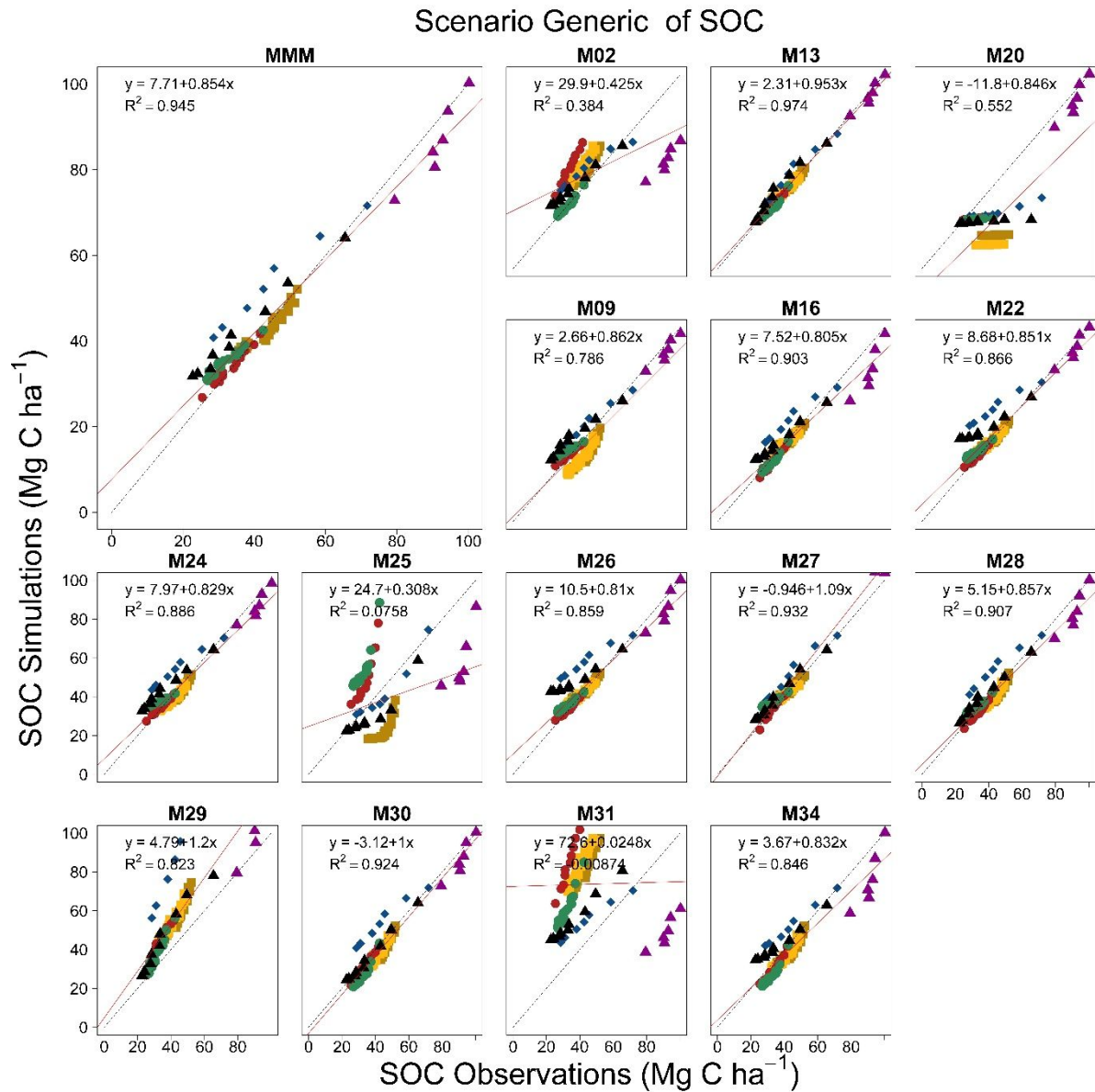


1191

1192 **Appendix B**

1193 Multi-year, multi-site comparison of individual model simulation of SOC ( $\text{Mg C ha}^{-1}$ ): multi-  
 1194 model medians (MMM) from Gen scenario simulations (16 models) versus observations.  
 1195 (coloured symbols represent sites as in Fig. 1).

1196



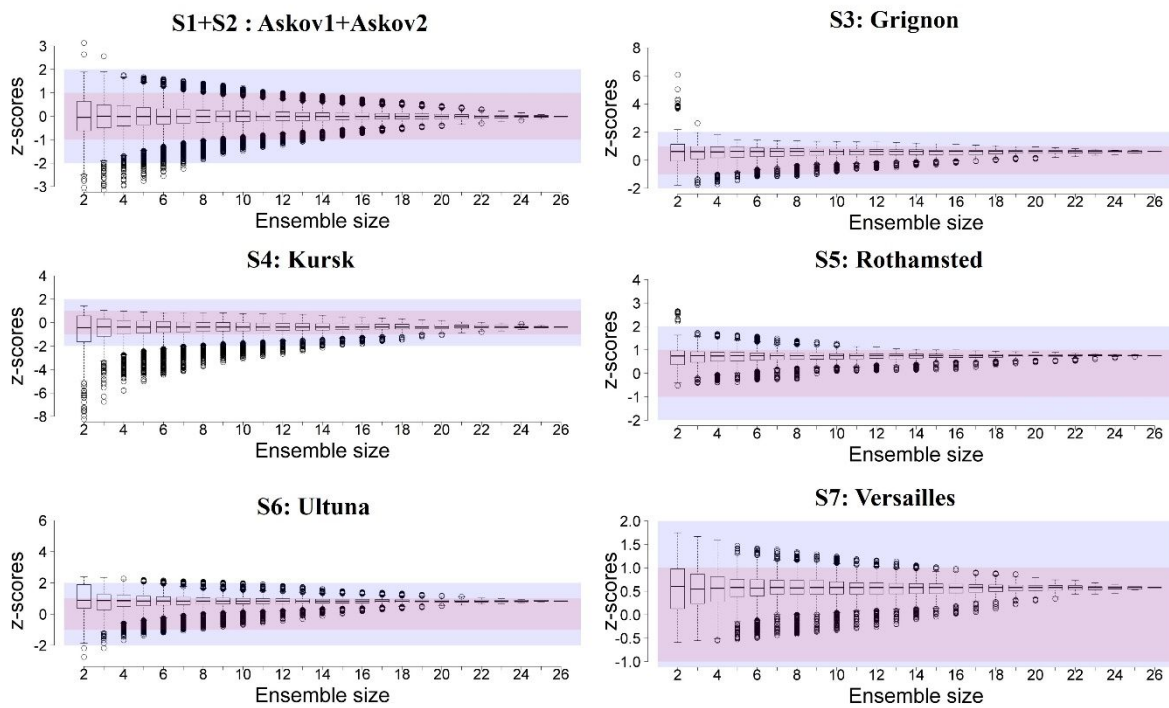
1197

1198



1199 **Appendix C**

1200 z-scores calculated with different ensemble sizes for SOC estimates obtained with Bln scenario at  
 1201 different experimental sites. Black lines show median values. Boxes delimit the 25<sup>th</sup> and 75<sup>th</sup>  
 1202 percentiles. Whiskers are 10<sup>th</sup> and 90<sup>th</sup> percentiles. Circles indicate outliers. Coloured bands mark  
 1203 two critical values:  $z=|1|$  (light purple) and  $z=|2|$  (light blue).

**Blind scenarios**

1204

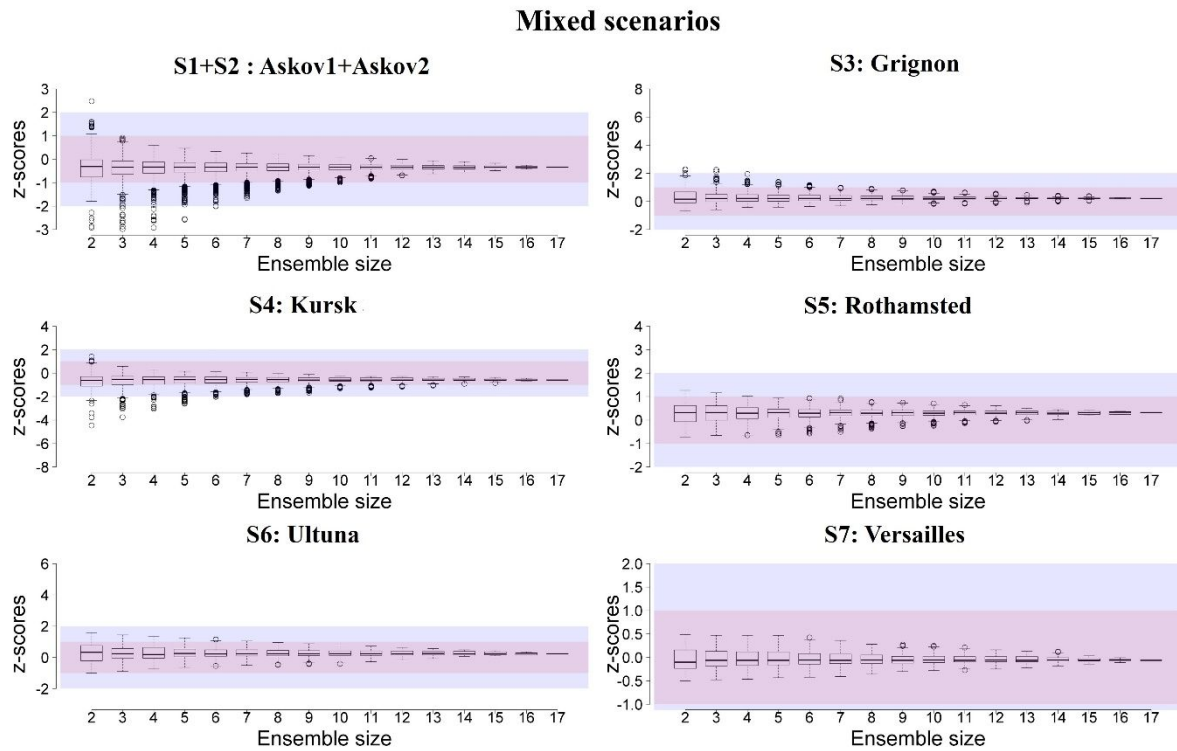
1205

1206

1207

1208 **Appendix D**

1209  $z$ -scores calculated with different ensemble sizes for SOC estimates obtained with Mix scenario at  
 1210 different experimental sites. Black lines show median values. Boxes delimit the 25<sup>th</sup> and 75<sup>th</sup>  
 1211 percentiles. Whiskers are 10<sup>th</sup> and 90<sup>th</sup> percentiles. Circles indicate outliers. Coloured bands mark  
 1212 two critical values:  $z=|1|$  (light purple) and  $z=|2|$  (light blue).

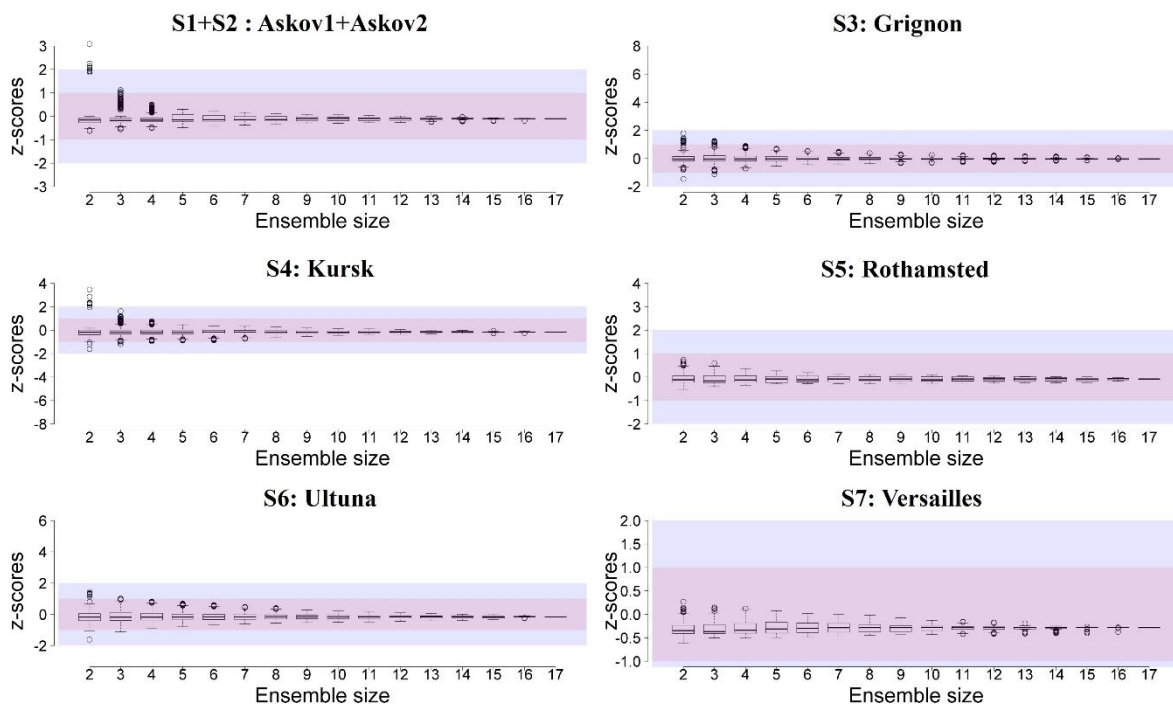


1213

1214

1215 **Appendix E**

1216  $z$ -scores calculated with different ensemble sizes for SOC estimates obtained with Spe scenario at  
 1217 different experimental sites. Black lines show median values. Boxes delimit the 25<sup>th</sup> and 75<sup>th</sup>  
 1218 percentiles. Whiskers are 10<sup>th</sup> and 90<sup>th</sup> percentiles. Circles indicate outliers. Coloured bands mark  
 1219 two critical values:  $z=|1|$  (light purple) and  $z=|2|$  (light blue).

**Specific scenarios**

1220

1221

1222

1223 **Appendix F**

1224 Individual and multi-model ensemble (MMM) performance metrics (as in Table 4) for blind (Bln)

1225 and calibration scenarios (Mix, Spe and Gen as in Table 3) across sites. Red (*italic*) and blue (**bold**)

1226 numbers indicate the worst and best performances by metric, respectively.

Performance metric Scenario	Model																									MMM	
	M01	M02	M03	M04	M05	M06	M07	M09	M12	M13	M16	M18	M19	M20	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32		M34
R <sup>2</sup>																											
Bln	0.73	0.92	0.67	0.83	0.79	0.86	0.76	0.89	0.83	0.90	0.33	0.81	0.69	0.63	<b>0.95</b>	0.76	0.92	0.41	0.86	0.76	0.92	<i>0.21</i>	0.82	0.35	0.57	0.80	0.94
Gen	NA	0.39	NA	NA	NA	NA	NA	0.79	NA	<b>0.97</b>	0.90	NA	NA	0.56	0.87	NA	0.89	0.09	0.86	0.93	0.91	0.82	0.93	<i>-0.00</i>	NA	0.85	0.95
Mix	NA	0.91	NA	0.90	0.91	NA	NA	0.89	NA	<b>0.99</b>	NA	NA	0.83	<i>0.41</i>	0.98	0.56	0.94	0.49	<b>0.99</b>	0.95	NA	0.91	0.84	0.87	NA	0.82	0.97
Spe	0.97	<b>0.99</b>	NA	0.98	<b>0.99</b>	NA	NA	<b>0.99</b>	NA	<b>0.99</b>	0.96	NA	NA	0.96	0.98	<b>0.99</b>	NA	0.91	<b>0.99</b>	0.97	<i>0.88</i>	0.93	0.98	0.94	NA	NA	<b>0.99</b>

BIn	0.88	0.97	0.84	0.93	0.90	0.89	0.90	0.97	0.93	0.97	0.71	0.94	0.85	0.79	<b>0.99</b>	0.89	0.97	0.73	0.95	0.91	0.95	0.59	0.95	<i>0.52</i>	0.85	0.93	0.98
Gen	NA	0.71	NA	NA	NA	NA	NA	0.93	NA	<b>0.99</b>	0.97	NA	NA	0.66	0.96	NA	0.97	0.53	0.95	0.97	0.97	0.81	0.97	<i>0.23</i>	NA	0.94	0.98
Mix	NA	0.97	NA	0.96	0.97	NA	NA	0.97	NA	<b>-1.00</b>	NA	NA	0.89	<i>0.69</i>	<b>-1.00</b>	0.79	0.98	0.81	<b>-1.00</b>	0.98	NA	0.76	0.96	0.96	NA	0.93	0.99
Spe	0.99	<b>-1.00</b>	NA	<b>-1.00</b>	<b>-1.00</b>	NA	NA	<b>-1.00</b>	NA	<b>-1.00</b>	0.99	NA	NA	0.99	<b>-1.00</b>	0.99	NA	0.97	<b>-1.00</b>	0.99	0.95	<i>0.76</i>	0.99	0.98	NA	NA	<b>-1.00</b>

RRMSE (%)

BIn	24.1	10.9	28.0	18.6	21.9	21.9	23.1	12.5	17.7	11.8	28.6	15.5	27.2	33.1	<b>7.9</b>	25.4	11.0	36.6	14.0	24.0	14.4	48.4	16.3	<i>69.1</i>	27.7	16.3	10.4
Gen	NA	30.8	NA	NA	NA	NA	NA	17.9	NA	<b>5.7</b>	11.5	NA	NA	51.3	14.0	NA	12.1	49.4	14.5	12.7	10.9	37.9	12.4	<i>92.1</i>	NA	15.8	10.6
Mix	NA	11.0	NA	12.6	11.5	NA	NA	11.7	NA	<b>3.8</b>	NA	NA	23.3	45.6	4.4	29.0	8.9	33.0	4.2	9.4	NA	<i>46.5</i>	14.4	13.4	NA	15.9	7.2

Spe	6.5	3.4	NA	5.0	<b>3.2</b>	NA	NA	3.8	NA	3.8	8.2	NA	NA	6.7	4.4	5.0	NA	14.5	4.1	6.2	14.9	<b>46.2</b>	5.5	8.7	NA	NA	<b>3.2</b>
-----																											
Bln	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	<b>0.64</b>	0.02	<i>-0.00</i>	<i>-0.00</i>	0.31	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	0.45	0.05	<i>-0.00</i>	<i>-0.00</i>	0.13	<i>-0.00</i>	<i>-0.00</i>	0.01	<i>-0.00</i>
Gen	NA	<i>-0.00</i>	NA	NA	NA	NA	NA	<i>-0.00</i>	NA	0.13	<b>0.17</b>	NA	NA	<i>-0.00</i>	<i>-0.00</i>	NA	0.08	0.04	<i>-0.00</i>	<i>-0.00</i>	0.06	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	NA	<i>-0.00</i>	<i>-0.00</i>
Mix	NA	<i>-0.00</i>	NA	<i>-0.00</i>	<i>-0.00</i>	NA	NA	0.55	NA	0.31	NA	NA	<i>-0.00</i>	<i>-0.00</i>	<b>0.76</b>	<i>-0.00</i>	0.54	<i>-0.00</i>	0.31	<i>-0.00</i>	NA	<i>-0.00</i>	0.24	<i>-0.00</i>	NA	<i>-0.00</i>	0.49
Spe	0.46	0.99	NA	0.06	0.03	NA	NA	0.85	NA	0.34	<i>-0.00</i>	NA	NA	0.12	0.93	<i>-0.00</i>	NA	<i>-0.00</i>	<b>-1.00</b>	0.29	<i>-0.00</i>	<i>-0.00</i>	<i>-0.00</i>	0.68	NA	NA	0.83
-----																											
Bln	0.52	0.90	0.49	0.72	0.60	0.60	0.56	0.87	0.74	0.88	0.33	0.80	0.39	0.09	<b>0.95</b>	0.47	0.90	-0.11	0.84	0.53	0.83	-0.93	0.78	<b>-2.95</b>	0.37	0.78	0.91
Gen	NA	0.22	NA	NA	NA	NA	NA	0.73	NA	0.97	0.89	NA	NA	-1.17	0.84	NA	0.88	-0.49	0.83	0.87	0.90	-0.19	0.87	<b>-6.00</b>	NA	0.79	<b>0.93</b>

Mix	NA	0.90	NA	0.87	0.89	NA	NA	0.89	NA	<b>0.99</b>	NA	NA	0.55	-0.72	0.98	0.31	0.93	0.34	<b>0.99</b>	0.93	NA	<b>-0.78</b>	0.83	0.85	NA	0.79	0.97
Spe	0.97	<b>0.99</b>	NA	0.98	<b>0.99</b>	NA	NA	<b>0.99</b>	NA	<b>0.99</b>	0.94	NA	NA	0.96	0.98	0.98	NA	0.87	<b>0.99</b>	0.97	0.82	<b>-0.76</b>	0.97	0.94	NA	NA	<b>0.99</b>

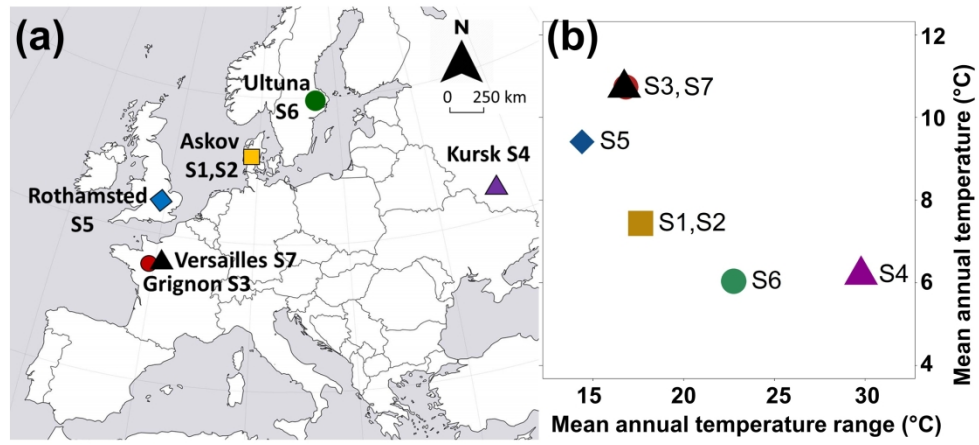


Fig. 1. Location (a) and characterisation of the study sites (b) with respect to mean annual temperature (°C) and mean annual temperature range (°C). Details about study sites are in Table 2.

558x254mm (150 x 150 DPI)



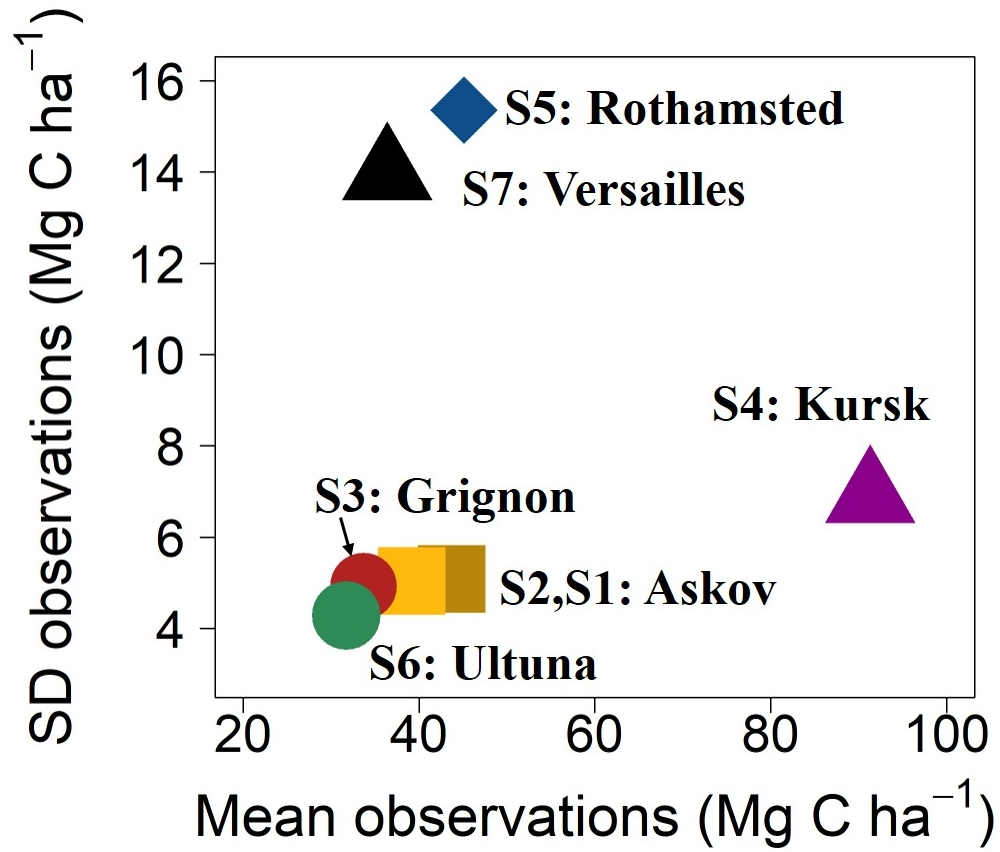


Fig. 2. Standard deviation (SD) and mean of SOC observations at the study sites (details are in Table 2).

169x169mm (150 x 150 DPI)

## Unable to Convert Image

The dimensions of this image (in pixels) are too large to be converted. For this image to convert, the total number of pixels (height x width) must be less than 40,000,000 (40 megapixels).

Fig. 3. Temporal changes of soil organic carbon (SOC, Mg C ha<sup>-1</sup>) observations (Observed, purple square) and simulations: blind (Blind, blue) simulations (26 models); three calibration scenarios, Generic (16 models, pink), Mixed (17 models, green) and Specific (17 models, grey) at all sites ( as in Table 2). Lines represent the multi-model median (MMM) of the simulations and shaded area represents the simulation envelope

## Unable to Convert Image

The dimensions of this image (in pixels) are too large to be converted. For this image to convert, the total number of pixels (height x width) must be less than 40,000,000 (40 megapixels).

Fig. 4. Soil organic carbon (SOC, Mg C ha<sup>-1</sup>) at each site (as in Table 2), for blind simulations (Blind, (26 models), three calibration scenarios (Mixed, 17 models; Specific and Generic, 16 models) and observations (Observed). For each boxplot, black horizontal lines show the multi-model median. Boxes delimit the 25th and 75th percentiles. Whiskers are 10th and 90th percentiles. Dots indicate outliers.

## Unable to Convert Image

The dimensions of this image (in pixels) are too large to be converted. For this image to convert, the total number of pixels (height x width) must be less than 40,000,000 (40 megapixels).

Fig. 5. Multi-year, multi-site comparison of individual model simulation of SOC (Mg C ha<sup>-1</sup>): multi-model medians (MMM) from blind simulations (26 models as in Table 1) versus observations (coloured symbols represent sites as in Fig. 1).

## Unable to Convert Image

The dimensions of this image (in pixels) are too large to be converted. For this image to convert, the total number of pixels (height x width) must be less than 40,000,000 (40 megapixels).

Fig. 6. Multi-year, multi-site comparison of individual model simulation of SOC (Mg C ha<sup>-1</sup>): multi-model medians (MMM) from Specific scenario simulations (17 models as in Table 1) versus observations (coloured symbols represent sites as in Fig. 1).

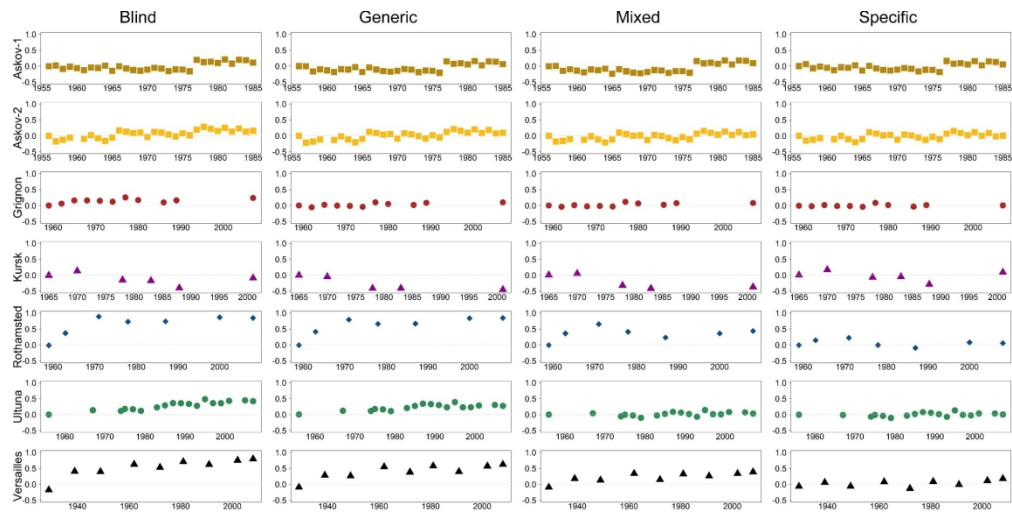


Fig. 7. Standardized model residuals (  $(MMM-O) / \sqrt{sd\_obs}$  ) over time for blind (Blind) simulations and calibration scenarios Mixed, Specific and Generic at each site.

381x190mm (150 x 150 DPI)

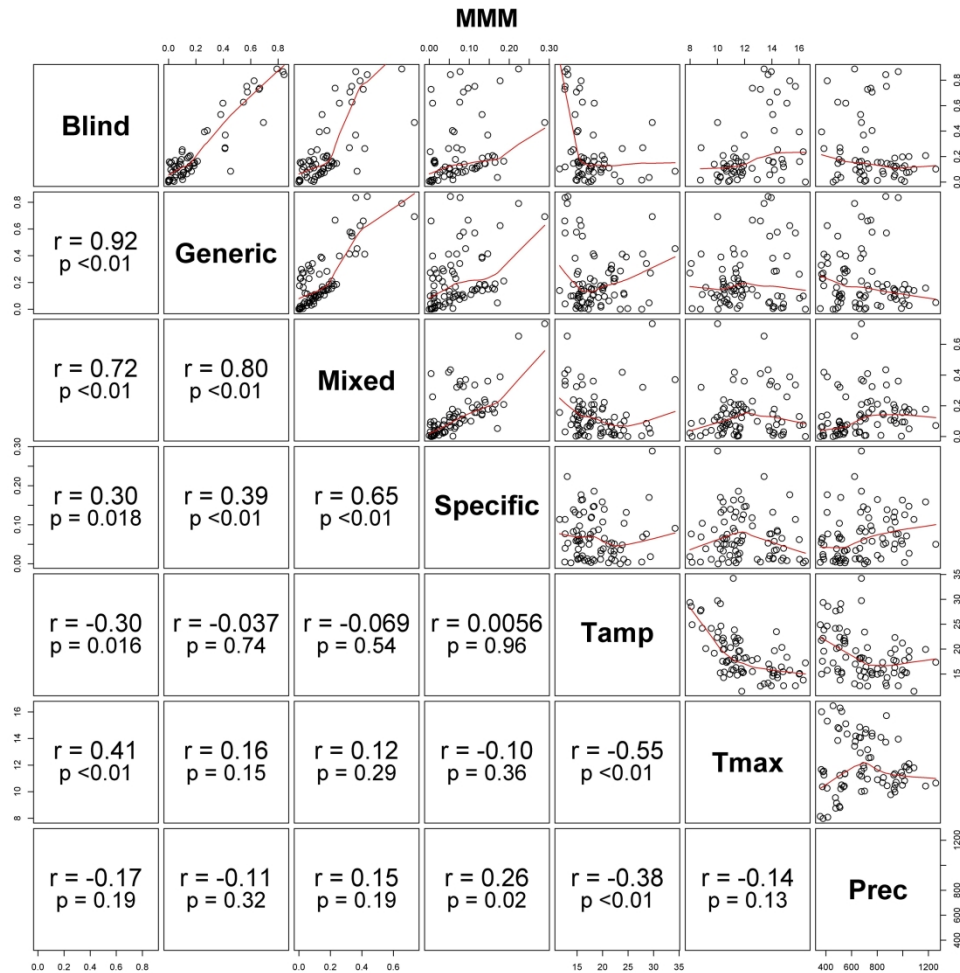


Fig. 8. Scatterplot correlation matrix of SOC (Mg C ha<sup>-1</sup>) model residuals of multi-model medians (MMM) for blind simulations (Blind) and calibrations scenarios (Generic, Mixed and Specific as in Table 3), and the annual climate metrics maximum temperature (Tmax), mean temperature amplitude (Tamp) and precipitation (Prec). Overlaid (red line) is a local non-parametric smoother curve.

304x304mm (300 x 300 DPI)

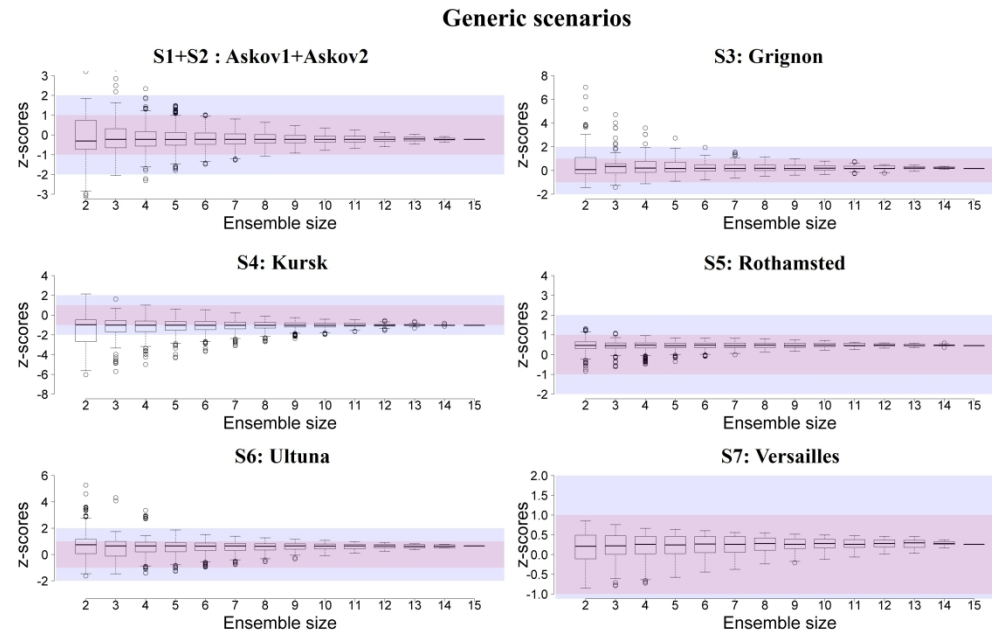


Fig. 9. z-scores calculated with different ensemble sizes for SOC estimates obtained with Generic scenario at different experimental sites. Black lines show median values. Boxes delimit the 25th and 75th percentiles. Whiskers are 10th and 90th percentiles. Circles indicate outliers. Coloured bands mark two critical values:  $z=|1|$  (light purple) and  $z=|2|$  (light blue).

677x438mm (150 x 150 DPI)



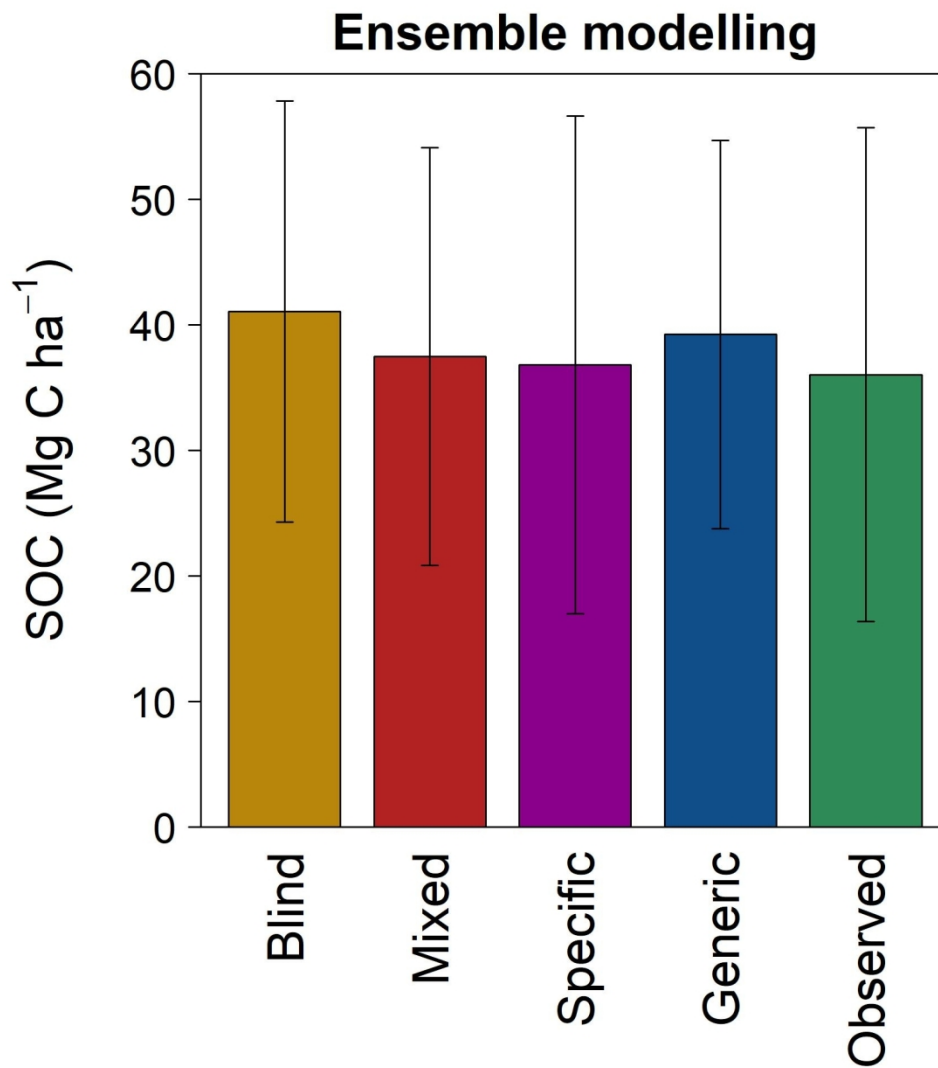


Fig. 10. Multi-site averages (vertical bars) and standard deviations (vertical lines) of observed and estimated (ensemble multi-model median) values of SOC ( $\text{Mg C ha}^{-1}$ ) in the last year of the experimental period. The ensemble modelling was applied with blind simulations (Blind) and calibration scenarios (Mixed, Specific and Generic as in Table 3).

152x169mm (300 x 300 DPI)