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

Model inter-comparison of soil organic carbon

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

[1] CREA - Council for Agricultural Research and Economics, Research Centre for Agriculture and Environment, Rome, Italy

[2] Agricultural Institute, Centre for Agricultural Research, Martonvásár, Hungary

[3] Université Clermont Auvergne, INRAE, VetAgro Sup, UREP, Clermont-Ferrand, France

[4] University of Aberdeen, UK

[5] Spanish National Research Council (CSIC), Zaragoza, Spain

[6] Università degli Studi di Milano, Italy

[7] Swedish University of Agricultural Sciences, Uppsala, Sweden

[8] CNR-IBE, Institute of Bioeconomy, Florence, Italy

[9] Université Paris Saclay, INRAE, AgroParisTech, Paris, France

[10] INRAE, BioEcoAgro, F-02000, Barenton-Bugny, France

[11] Université de Lorraine, INRAE, LAE, F-68000, Colmar, France

[12] Queensland University of Technology, Brisbane, Australia

[13] Colorado State University, Fort Collins CO, USA

[14] INRAE, CODIR, 75007 Paris, France

[15] Helmholtz Centre for Environmental Research, Halle, Germany

[16] Manaaki Whenua - Landcare Research, Palmerston North, New Zealand

[17] Ottawa Research and Development Centre, Agriculture and Agri-Food, Ottawa, Canada

[18] Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

[19] Laboratoire de Géologie de l'ENS, PSL Research University, Paris, France

[20] Tasmanian Institute of Agriculture, Australia

[21] JKI - Federal Research Centre for Cultivated Plants, Braunschweig, Germany

[22] Finnish Meteorological Institute, Helsinki, Finland

[23] CSIRO, Brisbane, Australia

[24] University of Gottingen, Germany

[25] Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany,

[26] Université de Reims Champagne Ardenne, INRAE, FARE, Reims, France

[27] formerly Rothamsted Research, North Wyke, Devon, UK

[28] Department of Agroecology, Aarhus University, Tjele, Denmark

[29] University of Potsdam, Germany

*Corresponding author. Tel.: +39067005413; fax +39067005711

E-mail address: roberta.farina@crea.gov.it

Abstract

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

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol/abbreviation	Long version	Explanation
System variables		
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
Experimentation		
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)
Modelling		
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)
Statistics		
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	ith 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
Agro-climatic metrics		
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
b[a]	De Martonne-Gottman aridity index	Indicator of aridity including both annual and monthly temperature and precipitation
hw[a]	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

[a] Supplementary material.

1. INTRODUCTION

The ability of soils to sequester and store large amounts of carbon (C) is well known (e.g. Lehmann and Kleber, 2015). Soil organic carbon (SOC) stocks are crucial for maintaining soil fertility and preventing erosion and desertification, and they positively influence the provision of ecosystem services at the local as well as the global scale (e.g. Lal, 2004, 2014). For these reasons, farmers aim to establish and maintain high organic C stocks in agricultural soils, which have often been depleted trough historical land use practices (Fuchs et al., 2016; Gardi et al., 2016; Chenu et al., 2018). The continuing studies on SOC sources and biogeochemical processes in the soil environment provide key insights into climate-C feedbacks, and help prioritizing C sequestration initiatives (Gross and Harrison, 2019). In light of the climate change issue, the storage of C and additional sequestration of atmospheric C have received increasing attention recently (Rumpel et al., 2018; Whitehead et al., 2018; Lavallee et al., 2020), promoting land management, and agro-ecosystems in particular, as a key mitigation option (e.g. the '4 per mille Soils for Food Security and Climate' initiative, Minasny et al., 2017; Soussana et al., 2017). However, the slow response of SOC to changes in management and environmental factors hampers our understanding of how SOC can be increased in a sustainable manner, especially under changing climatic conditions. Long-term field experiments (LTEs), in which SOC responses have been observed over several decades, provide this information and deliver reference data on SOC content for knowledge gain and model development (Johnston and Poulton, 2018). However, LTEs are costly to maintain, and it is generally difficult to extrapolate experimental results across space and time (Debreczeni and Körschens, 2003; Mirtl et al., 2018). Simulation models play a prominent role in SOC research because they provide a mathematical framework to integrate, examine and test the understanding of SOC dynamics (Campbell and Paustian, 2015). They can also be used to extrapolate from micro- (e.g. carbohydrate production during photosynthesis) to macro-scale dynamics (e.g. global C cycling) (e.g. Gottschalk et al., 2012; Sitch et al., 2003). In particular, complex agricultural and environmental models incorporate a mechanistic view of processes and system interactions, in which the soil components are often represented by different, operationally defined, pools of different sizes and with different properties (e.g. Parton et al., 2015). The concept of multiple C-N pools represents C-N dynamics with an idealised description (Hill, 2003). The relative proportion of C and N (and sometimes lignin to N ratio) in the plant residue is the primary mode to divide plant inputs (from e.g. leaf litter and root exudates) into fresh litter pools, which then decompose into SOC (or SOM, i.e. soil organic matter) pools, each being modelled with different residence (or turnover) times, varying from months for labile products of microbial decomposition to hundreds to thousands of years for organic substances with firm organic-mineral bonds (e.g. Yadav and Malanson, 2007; Dungait et al., 2012). Plant material and animal manures are often modelled to enter the soil environment as either readily decomposable (carbohydrate-like) or resistant (lignin and cellulose-like) materials. A varying number of pools (often including inert and slow-decomposing organic matter, and microbial biomass) linked by first-order equations is usually simulating both C and N fluxes within and between each pool (Falloon and Smith, 2010). However, different models vary considerably in the underlying assumptions and C processes in current models, e.g. regarding number of pools, type of decomposition kinetics used and processes regulating SOC retention (Manzoni and Porporato, 2009; Cavalli et al., 2019).

Each model offers a distinctive synthesis of scientific knowledge (Brilli et al., 2017) and multi-model ensembles developed from several models may reduce uncertainties in biological and physical outputs that occur over large scales, such as regions and continents (e.g. Rötter et al., 2012; Asseng et al., 2013; Ehrhardt et al., 2018). The advantage of using ensemble estimates over individual models is that caused by compensation of errors across models, and a broader integration of model processes (Martre et al., 2015). It has been recommended to use model ensembles for reducing uncertainties in simulations of agricultural production (Asseng et al., 2013; Bassu et al., 2014; Challinor et al., 2014; Li et al., 2015; Ruane et al., 2016; Maiorano et al., 2017) and other biophysical/biogeochemical outputs (Sándor et al., 2017, 2018a; Ehrhardt et al., 2018). However, after the pioneering study of Smith et al. (1997), who evaluated nine SOC models using 12 datasets from seven LTEs, other modelling studies targeting SOC dynamics have often been limited in scope. Smith et al. (2012) used four models to assess the effect on SOC of crop residues' removal in 14 experiments in North America. Todd-Brown et al. (2013, 2014) performed global estimates of SOC changes with 11 Earth system models. Kirschbaum et al. (2015) used one simulation model and two years of eddy covariance measurements collected over an intensively grazed dairy pasture in New Zealand to better understand the drivers of changes in SOC stocks. Puche et al. (2019) performed a similar study in France. Using multi-model ensembles in scenario studies at eight sites worldwide, Basso et al. (2018) highlighted the importance of soil feedback effects (C and N) on the prediction of wheat and maize yield. We are not aware of any recent model inter-comparison studies specifically assessing soil C dynamics with several models across a range of experimental sites. This is a field where there is a need for standardised guidance to estimate C stocks at various spatial scales (Bispo et al., 2017). A difficulty in testing and comparing various models (and interpreting model outputs) lies in the interaction between soil and plant processes so that any of the model-data discrepancies could be due to errors in either component (e.g. Ehrmann and Ritz, 2014). A rigorous model testing and comparison would require different model components, e.g. plant and soil modules, to be assessed separately. Bare-fallow plots offer such an opportunity in that they are plots maintained for decades without any plant inputs. The changes in SOC stocks therefore result only from decomposition processes. To assess the function of soil-model components without interaction with plant processes, we conducted a model inter-comparison using a dataset from long-term bare-fallow experiments where plant inputs were zero. In this study, we refer to bare-fallow plots that were kept free of plants by manual and/or chemical means for several decades. We used seven bare-fallow treatments included in six long-term agricultural experiments (>25 years), all located in Europe (Denmark, France, Russia, Sweden and

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

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

2. MATERIALS AND METHODS

2.1. Simulation models

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

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

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

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

	version of CANDY 29.06.2018				
CCB	2019.1.16	3	None	https://www.ufz.de/index.php?en=44046	Franko et al. (2011); Franko and Spiegel (2016); Franko and Merbach (2017)
Century	4.0	5 to 7	Yes	https://www2.nrel.colostate.edu/projects/century/MANUAL/html_manual/man96.html	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	http://www.kirschbaum.id.au/Welcome_Page.htm	Kirschbaum (1999); Kirschbaum and Paul (2002)
C-TOOL	2014	3	None (can be run also with spin-up)	http://envs.au.dk/fileadmin/Resources/DMU/Luft/emission/SINKS/C-TOOL_Documentation__2015_.pdf	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	http://www.nrel.colostate.edu/projects/daycent-home.html	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)	http://www.dndc.sr.unh.edu	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	http://dssat.net	Jones et al. (2003); Porter et al. (2009); Gijsman et al. (2002); White et al. (2011); Thorp et al. (2012)
ECOSSE	5.0.1	5	None	https://www.abdn.ac.uk/staffpages/uploads/soi450/ECOSSE%20User%20manual%20310810.pdf	Smith et al. (2007, 2010a, b); Bell et al. (2010)
ESOC1	1.0	3	Yes	https://doi.org/10.5281/zenodo.3539484 fmoyano@uni-goettingen.de	Moyano et al. (2018)

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

YASSO15	15	5	Yes	https://en.ilmatieteenlaitos.fi/yasso	Tuomi et al. (2009)

[a] Some models/model versions include options for varying C pools (this varying number may depend on the fact that the full set of pools including fresh C can be optionally simplified in the case of bare-fallow treatments).

2.2. Experimental sites

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

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

	General description		Experimental sites (country)					
			S1, S2 Askov (Denmark)	S3 Grignon (France)	S4 Kursk (Russia)	S5 Rothamsted (United Kingdom)	S6 Ultuna (Sweden)	S7 Versailles (France)
Coordinates	**Latitude**		55.28	48.51	51.73	51.82	59.49	48.48
	Longitude		9.07	1.55	36.19	0.35	17.38	2.08
Soil	**Sand/Silt/Clay (%)**		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)
	Bulk density (Mg m^{-3})		1.50	1.20	1.13	0.94	1.44	1.30
	Experimental period	***Bare-fallow years***	1956-1985	1959-2007	1965-2001	1959-2008	1956-2007	1929-2008
		N. of data/replicates	30/4, 29/4	11/6	6/0	14/4	18/4	9/6
	Initial/final carbon stocks (Mg C ha^{-1})		52.1/36.4	41.7/25.4	100.3/79.4	71.7/28.6	42.5/26.9	65.5/22.7
Climate[a]	**Climate type**[b]		Dfb (humid continental)	Cfb (oceanic)	Dfb (humid continental	Cfb (oceanic)	Dfb (humid continental	Cfb (oceanic)

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

[a] Climatic analysis was performed on longer periods than the experimental periods: 1956-1987/1929-2008/1944-2003/1856-2006/1956-1999/1929-2008.

[b] Köppen-Geiger climate classification (Kottek et al., 2006).

[c] Mean values over the bare-fallow period. Reference evaporation was estimated based on the Thornthwaite (1948) equation.

[d] Mean difference in temperature between the warmest and the coldest month of the year.

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

(Fig. 1 here)

2.3. Study design

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

After the blind simulations were completed, SOC measurements taken during the bare-fallow period were supplied to each modelling group for the calibration work. Details on management (tillage), which may have influenced the SOC dynamics before the bare-fallow treatment, were also provided to improve the initialisation process. It was requested that each modelling group adjust soil parameters to improve the simulations based on the observed data, using whatever techniques they normally use, and to document the changes. At this stage, models were split into two categories: a) with spin-up (SP) and b) without spin-up (NS). Both SP and NS models require an initial estimate for SOC content and/or an adjustment of parameters towards balancing the split between soil C pools. The two classes of models work in the same way using information about plant residues and root growth that provide the C substrate for SOC dynamics simulations. NS-type models (e.g. DNDC and RothC) use the initial measured SOC value, where estimates of C inputs in the background of model runs are obtained with various methods (e.g. Keel et al., 2017) in order to initialise the SOC pools, which can sometimes be calculated analytically. In order to keep the legacy effect of previous land-use and past management practices, in SP models (e.g. DayCent) SOC pools are routinely initialised by running the models to achieve their own states of equilibrium, where change in C stocks is minimised (e.g. Lardy et al., 2011; Huntzinger et al., 2013). However, if soils are not at equilibrium (e.g. after a sudden disturbance), spin-up runs may not always be valid with the risk of starting simulations with biased initial values (e.g. Wutzler and Reichstein, 2007; Nemo et al., 2017) but a fuller discussion on the "spin-up problem" (Reynolds et al., 2007) is not within the scope of this paper. Carbon inputs are usually estimated through sub-models calculating total net primary production (TNPP). As it was not possible to derive TNPP data from local sources at each study-site, TNPP estimates were obtained at each site (Table 2) based on precipitation levels according to the approach of Del Grosso et al. (2008). In this way, the creation of the TNPP database used by modellers was based on an identical methodology, which is widely used worldwide, though the uncertainty in quantifying productivity across ecosystems is highlighted (e.g. Wieder et al., 2014).

The distinction between SP and NS models can appear somewhat arbitrary as virtually any model with more than one C pool could be spun-up or, alternatively, a function (or analytical procedures) can be used to make an initial pool partition. We refer here to common modelling practice, as performed by users within the constraints imposed by packaged (operational) solutions of SOC models (for which spin-up procedures may be operationally more difficult) or relying on the procedure suggested by previous experience. For instance, although spin-up equilibrium runs are documented for RothC (e.g. Herbst et al., 2018), it is common practice to initialise three C pools for subsequent simulations through an internal routine over 10,000 years, with limited model inputs including clay fraction and weather, and a pre-defined ratio of decomposable over recalcitrant plant material (e.g. Xu et al., 2011; Weihermüller et al., 2013). Modellers were left to choose one option or the other when both were available for use in their models (e.g. C-TOOL). About 40% of the models (10 models) in the study did not use SP processes and set the initial SOC values manually (using the initial SOC observation).

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

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

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

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

Model category	**Factors**	**Approaches**	**Calibration scenarios**[a]		
			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	

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

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

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

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

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

(Jørgensen et al., 1986)			
EF, modelling efficiency (Nash and Sutcliffe, 1970)	$\mathrm{EF} = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$	-	negative infinity to 1 (optimum): the closer the values are to 1, the better the model
Coefficient of determination (R²) of the linear regression estimates versus measurements / r, Pearson's correlation coefficient of the estimates versus measurements (Addiscott and Whitmore, 1987)	$R^2 = \frac{\sum_{i=1}^{n}(P_i - O_i)\cdot(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2\cdot\sum_{i=1}^{n}(O_i - \bar{O})^2}}$ - $r = \sqrt{R^2}$	-	0 (absence of fit of the regression line) to 1 (perfect fit of the regression line): the closer the values are to 1, the better the model -1 (full negative correlation) to 1 (full positive correlation): the closer the values are to 1, the better the model
P(t), Paired Student t-test probability of means being equal	$\mathrm{P(t)} = \mathrm{Probability}\left(\frac{\bar{D}}{\frac{S_D}{\sqrt{n}}}\right)$	-	0 (absence of agreement) to 1 (perfect agreement): the closer the values are to 1, the better the model
d, index of agreement	$d = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}$	-	0 (absence of agreement) to 1 (perfect agreement): the

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

2.4. Multi-model and ensemble assessment

We first focussed on the quantification of model-data discrepancies and then assessed the uncertainty of the individual models in comparison with the multi-model ensemble. The modelling teams provided deterministic model simulation results according to the protocol established, which meant that: 1) one run was provided for each site; 2) the spread of model results due to parameter uncertainty was not specifically addressed. The latter would have dramatically increased the range of model outputs used within the study and would have confounded the uncertainty in calibrated parameters with the uncertainty in model structure (Wallach and Thorburn, 2017). While the uncertainty in model predictions could be due to parameterisation, model calibration from different users (i.e. ensemble of users within ensemble of models) cannot be regarded as the solution to estimate uncertainty due to parameterization (Confalonieri et al., 2016). As well, different calibration techniques do not seem to be primarily responsible for differences in model performance (Wallach et al., 2020) and the contribution of the initialisation to the uncertainty in SOC changes can be negligible compared to the uncertainty related to the model itself and simulated systems characteristics (Dimassi et al., 2018). As uncertainty could not be associated with any individual simulation, we focussed on the analysis of model residuals. We documented the variability of the multi-model simulation exercise across two stages (blind test and alternative calibration scenarios), while inspecting how the multi-model median (MMM) converged to the observations. We used box-plots to compare the variability of estimates by different models (with focus on multi-year averages) to the observed variability, and we represented model ensembles with MMM, which has the advantage to exclude distinctly biased model members with a disproportionate influence on the mean (Rodríguez et al., 2019). The advantage of using MMM was established in practical studies in crop and grassland modelling but also on a theoretical basis (Wallach et al., 2018).

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

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

(Fig. 2 here)

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

3. RESULTS

3.1. Evaluation of SOC dynamics

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

SOC stocks decreased under all bare-fallow sites during the investigated period. At S1, S2, S3, S4 and S6 (Fig. 3) sites, the decrease in SOC stock was from minimum to moderate whereas at S5 and S7 (Fig. 3) SOC loss in the top 0.20 m was more rapid, with initial SOC halved during ~30 years. The decay tended to be more rapid in the first years and then the rate of loss decreased (e.g. at S7 site between 1929 and 1962, Fig. 3).

(Fig. 3 here)

3.2. Ensemble performance by site

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

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

(Fig. 4 here)

3.3. Individual models versus multi-model ensemble

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

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

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

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

(Fig. 5 here)

(Fig. 6 here)

3.4. Analysis of model residuals

The plots of the discrepancy between MMM and observations (Fig. 7) as a function of time shows a limited scatter (within ±1) at each site. While Bln, Gen and Mix scenario overestimated the SOC decomposition rate at Kursk (where the highest SOC content was measured), the standardized residuals were around zero at Grignon and both Askov sites during the whole of experimental period. However, the departure from observations may increase over time especially with Bln and Gen scenarios at some site (e.g. at Rothamsted, Ultuna, Versailles) indicating that models underestimate decomposition rates after a few years/decades.

(Fig. 7 here)

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

The most prominent effect of annual climate indicators was found at the blind test stage, whose residuals were negatively correlated with precipitation (r=-0.17) and positively correlated with Tmax (r=0.41). Combinations of high maximum air temperature and low precipitation values may thus generate greater errors in blind SOC simulations. Calibration scenario Gen did not show significant correlations to climate indicators. However, calibration scenario Spe and Gen had opposite correlations. The annual precipitation positively correlated with Spe residuals (r=0.26) and with scenario Mix (r=0.15). Annual maximum temperature and scenario Spe negatively correlated (r=-0.10). These correlations with climate indicators hint that the site-specific calibration (scenario Spe) is more sensitive to precipitation than to maximum temperatures. On the contrary, Bln and Gen simulation residuals showed greater sensitivity to maximum temperatures. Residuals of individual models were approximately equally influenced by precipitation and temperature drivers, but with differences among models and scenarios (Figs. B1-26 in the supplementary material). In most of the cases, model residuals were positively correlated with annual maximum temperatures and negatively correlated with annual precipitation totals (e.g. M03, M09, M18, M22 for Bln). In some cases, e.g. M09 (Fig. B8 in the supplement), the correlations among SOC residuals for different scenarios were both positive and negative (r values ranged from -0.043 to 0.36), and even the effect of climate indicators were different (e.g. for Tmax, r values ranged from -0.096 to 0.65). In other cases, e.g. M25 (Fig. B18 in the supplement), SOC residuals were more similar to each other (r-values 0.17-0.80) and the effect of precipitation and temperature drivers was often important (with r>0.4). It is interesting in this respect that the Spe residuals had near-zero correlations with climatic drivers, showing a lesser influence of these factors on model results with this scenario, whereas the Bln scenario showed some correlations with Tamp (r=0.13), Tmax (r=-0.44) and precipitation (r=0.40). For M25, Gen scenario residuals (Fig. B18 in the supplement) appeared unrelated with precipitation (r-value near zero), but not with temperature amplitude (r=0.50) and maximum air temperature (r=-0.56).

(Fig. 8 here) .

3.5. Minimum ensemble size

We attempted to identify the minimum number of models required to obtain reliable results for Bln and calibration scenarios Mix, Spe and Gen (Fig. 9 and Appendix C-E). We observed that there could be large differences in the *z*-scores obtained across sites with different ensemble sizes and scenarios. Overall, Bln is characterised by greater *z*-scores than the calibration scenarios. Our analysis suggests that the ensemble size could be reduced to four models (or even fewer) at S3, S6 and S7. For the other sites (e.g. S4), only ensemble sizes of at least 9-10 models reduced *z*-scores to within the range from -2 to +2, but this number should be raised to 20 or higher to comply with the most stringent criterion of *z*=|1|. A minimum ensemble size of 9-10 models was also identified with Gen at S4 (Fig. 9), while with Mix and Spe scenarios the number of models could be reduced down to 7 and 3, respectively (up to about 14 [Gen], 8 [Mix] and 4 [Spe] to comply with *z*=|1|) (Appendix C-E).

(Fig. 9 here)

4. DISCUSSION

4.1. Scenarios of ensemble SOC estimates

For Bln, Mix, Gen and Spe scenarios, the overall differences between the simulated and the observed first-year SOC values were −0.46, +3.49, +2.40 and +1.92 Mg C ha^{-1}, respectively, for the NS models, and +0.58, -0.29, +0.95 and -0.12 Mg C ha^{-1}, respectively, for the SP models. Despite manually setting the initial SOC values (magnitude of first SOC observation for the simulation period), the NS models mostly overestimated SOC content in the initial year of the model run. In first-year estimates of the calibrated (mainly with Spe and Mix scenarios), SP models deviated less from observations than NS models that overestimated SOC stocks for the first year with the exception of M25 (+8.4 Mg C ha^{-1} for Gen), M29 (+18.6, +21.1 and +23.7 Mg C ha^{-1} for Spe, Gen and Mix, respectively) and M31 (+25.2 Mg C ha^{-1} for Gen). In the case of M25, the model was run with a generic grassland spin-up (i.e. 7,000 years), which was applied to all sites. Thus, a generic history was simulated without considering the cropping history at each site. This spin-up protocol affected the simulated SOC, showing the poor ability of Gen scenario to produce results consistent with observations, which questions the practicality of spin-up processes under generic calibration. With M31, there was a greater difference between simulated and observed SOC values in the initial simulation year and the model gave results that did not correspond to the observations at all sites (Appendix F), especially under the Bln and Gen scenarios. Though M31 used the initial SOC observation as default parameter, it failed to reproduce the LTBF dynamics between sites because of large differences in C input to the soil from the former vegetation during the spin-up period. Consequently, the starting points of the LTBF simulations differed greatly from the observations, which were overestimated at S1, S2, S3 and S6, and underestimated at S4. Overall, Mix and Spe calibrations showed better performance indices than the Gen scenario (Appendix F). We note, however, that M13, for which the SOC pool sizes (humads and humus) were generically calibrated across sites, produced low RRMSE for Gen (5.7%).

The improved calibration knowledge obtained with the site-specific information also improved model accuracy. Moving from Bln (with knowledge of weather and soil texture, historical land use and management, and initial SOC; section 2.3) to the Gen scenario, we reproduced SOC data in a number of European bare-fallow experimental sites with a single set of calibrated, regional-scale parameter values (regardless of the possible soil, climate and past land-use dissimilarities between different sites). According to performance indicators in Appendix F, in the Bln simulations the NS models performed better than the SP models. For instance, average RRMSE and EF were 19.44% and 0.60, and 26.94% and 0.24, for NS and SP models, respectively. Compared to the Bln scenario, the discrepancy between the measured and estimated SOC values under the Gen scenario was slightly reduced with NS models and increased with SP models. Multi-site calibration can be characterised by lower uncertainty than site-specific calibration, because more data contribute to the calibration process (e.g. Minunno et al., 2014; Ma et al., 2015). The availability of a variety of detailed data from multiple sites thus offers the possibility of a genuine multi-location calibration of the model, assuming that a single calibration across sites is appropriate. The limit of the Gen scenario calibration was that it did not make it possible to explore the spatial variability of model parameters. The latter was done with scenarios Mix and Spe, for which a basic requisite is that model parameters are not hard coded but configuration files are left open to the users. From Gen to Mix, parameters describing initial values of each pool were determined separately for each site. Moving from Mix to Spe, the decomposition parameters became site-specific. Hence, modellers needed to invest increasingly more knowledge (and more time-demanding calibration effort) than in Gen. Under these conditions, the improvement of simulations in SP models was evident (up to 70% for some indicators, e.g. RRMSE and EF). On the contrary, NS models only had a slight improvement in accuracy of simulations from Bln (RRMSE=21.5%; EF=0.58) to Mix (RRMSE=18.6%, EF=0.55) or Gen (RRMSE=20.5%; EF=0.45). In our analysis, the two types of models (NS and SP) appear to be suitable for different sets of data. NS-type models, in most cases, can perform well even when data are limited to climate, initial C and historic land use, while SP models generally benefit from the availability of more detailed data. All metrics related to the performance of the SP models were improved with calibration. There were some differences in model performance among the sites, but site-specific soil or climatic conditions cannot easily explain such differences.

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

(Fig. 10 here)

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

4.2. Possibilities for model inaccuracies

We presented an approach that uses a correlation matrix (with graphical representation) to account for possible correlations between Bln, Mix, Gen and Spe residuals and, additionally, climatic factors (mean air temperature amplitude, maximum air temperature and precipitation total). This residual analysis helps find correlations among alternative scenarios, which might indicate comparable scenarios in which error propagation within models is similar, though the way of error propagation cannot be easily retrieved from the correlation matrix. This is the case of Bln, Gen and Mix, whose residuals are highly correlated, while the weak correlations between Spe and other scenarios highlight the distinct behaviour of the latter. This analysis can also help find correlations between the SOC output and external drivers, and thus suggest additional predictors that may need to be included in the models (e.g. Medlyn et al., 2005). This need emerged especially when specific models were run under Bln, Gen and Mix scenarios, for which some correlations ($r>|0.4|$) were obtained between model residuals and drivers of thermal and moisture conditions. A weaker but significant correlation ($r=0.26$, $p=0.02$) was also obtained between Spe residuals and precipitation. These correlations indicate some limitations related to the response functions of SOC decomposition to soil temperature and soil moisture, though the relative uncertainties of our model ensemble are attenuated by the presence in the models of physical and chemical processes that explain the intra- and inter-annual variability of SOC. We add that such biophysical conditions affect the microbial activity (e.g. Blagodatskaya and Kuzyakov, 2008; Guenet et al., 2010; Wutzler and Reichstein, 2013), and care should be taken when extrapolating our results over long time frames (especially without locally calibrated models, Fig. 7) if no corroborating field evidence for long-term decay rates can be obtained (e.g. on how models are dealing such situations in which microbes become increasingly C limited as no new C input by plants occurs; Kuhry and Vitt, 1996).

5. CONCLUSIONS AND FUTURE DIRECTIONS

This paper on SOC modelling offers a tentative answer to the questions about: (i) whether and to what extent an ensemble of models performs better than single models, (ii) the minimum ensemble size that is required to reduce the error below a given threshold, and (iii) the set of data required to prepare and substantiate ensemble estimates. This study presents a framework for interpretation of model performance and uncertainties obtained with a set of process-based biogeochemical models (individually and in an ensemble) simulating soil C contents in bare-fallow experimental systems at a variety of European sites. One of the features of SOC modelling today is the huge amount and variety of models available. Although our analysis did not take into account all sources of uncertainty (e.g. the influence of the unique choices made by modellers), it enabled the integration of several modelling teams into an ensemble protocol. Classifying and comparing different approaches have revealed great model diversity, and is the basis for the development of dedicated ensemble protocols. In this model inter-comparison, the need to accommodate challenges experienced by modellers (including C pools of different nature, and optional initialisation and calibration procedures) was reflected in the co-creation (with modellers and data providers) of alternative calibration scenarios (Mix, Gen, Spe). As far as we are aware, no previous multi-model inter-comparison studies have examined differences in such calibration scenarios or differences between models with or without spin-up.

In our study, we did not aim to identify the best model(s) for simulating SOC dynamics for bare-fallows and no probability of success was assigned to prove the suitability of using one model rather than another. Overall, we showed that a calibration scenario with generic system knowledge was adequate for providing sufficiently reliable output, but additional site-specific knowledge can further improve results under certain circumstances. This is operationally relevant because the effort required to gather calibration data might no longer be feasible for modelling scenarios moving from single sites to increasingly larger spatial scales. Site-specific calibration could help refine model estimates. However, geographical locations have characteristics (e.g. soil and climate conditions, past history) that require specific model structures and local optimisation, and the application of models may be limited by the ability to provide representative parameter values. Soil-C model inter-comparisons including more models and experimental data from other regions should be continued to improve our ability to simulate biogeochemical processes with acceptable accuracy. Additional assessments are also recommended to complete the analysis of model behaviour in the long term (like thousands of years) with constant inputs. While the various models evaluated here did not include all available modelling approaches used to simulate soil C dynamics, the present model inter-comparison was large compared to other studies. As such, it is a distinct improvement over previously published quantitative approaches because it represents a reasonable sub-population of common and current approaches. In this, we offer a method to allow a broad ensemble of models to be implemented using existing datasets and current modelling practices. Overall, this multi-model ensemble sets a precedent for key progress in soil C modelling because it provides essential information about SOC modelling and opens a path to a more in-depth analysis of the response of individual models and their uncertainties against soil and climate drivers. Now that we have examined SOC decomposition in-depth without the difficulties of C input uncertainties, a similar modelling study should be conducted on LTEs that examine both plant derived C inputs as well as C inputs from manures and other organic materials recycled in agroecosystems. In fact, under field conditions, the amount of C input is not only an important factor driving the changes in SOC stocks (including the changes due to tillage), but the amount of C input also drives the mineralization rate of the SOC (Mary et al., 2020). How simulation models compare under such conditions is important for improving our ability to evaluate and achieve climate C goals. With increasing availability of data and computational resources, there are many opportunities for the SOC modelling community to enrich its offering and to keep up with evolving methodologies, which would significantly increase transparency of the underpinning science and modelling practice. A number of recent actions are ongoing under the guidance of international initiatives such as the European Joint Programme (EJP) on Soil (https://projects.au.dk/ejpsoil). Started in 2020, the EJP-Soil is undertaking a detailed inventory of models and all available data sources (e.g. world soil maps, satellite images, downscaled weather data), and appears as an ideal arena to facilitate the exchange of information and to further explore SOC model developments and practice.

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

R. Farina, R. Sándor and G. Bellocchi coordinated the study, contributed to its design, conducted the analysis of data and produced the first draft of the manuscript. P. Smith, C. Chenu, F. Ehrhardt, M. A. Bolinder, C. Nendel and J.-F. Soussana contributed to the design of the study and the writing of the manuscript. M. Abdalla, J. Álvaro-Fuentes, M. A. Bolinder, L. Brilli, H. Clivot, M. De Antoni, C. Di Bene, C. D. Dorich, F. Ferchaud, N. Fitton, R. Francaviglia, U. Franko, D. Giltrap, B. B. Grant, B. Guenet, M. T. Harrison, M. U. F. Kirschbaum, K. Kuka, L. Kulmala, J. Liski, M. J. McGrath, E. Meier, L. Menichetti, F. Moyano, N, Reibold, A. Shepherd, W. N. Smith, T. Stella, A. Taghizadeh-Toosi and E. Tsutskikh performed the model calibrations and runs.

C. Dorich, L. Bechini, L. Menichetti, R. Francaviglia, S. Recous, W. Smith, F. Ferchaud, H. Clivot, M. A. Bolinder, W. Smith, A. Taghizadeh-Toosi, L. Brilli, R. Farina, G. Bellocchi, T. Stella and U. Franko discussed and decided upon the modelling scenarios at the CN-MIP final meeting (Rome, 6-7 June 2018). C. Dorich prepared a detailed protocol for second-stage simulations.

Those interested in the details of the modelling process are encouraged to contact authors.

Data Availability Statement

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

REFERENCES

Abrahamsen, P., & Hansen, S. (2000). Daisy: an open soil-crop-atmosphere system model. *Environmental Modelling & Software*, **15**, 313-330. https://doi.org/10.1016/S1364-8152(00)00003-7

Addiscott, T. M., & Whitmore, A. P. (1987). Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring. *Journal of Agricultural Science*, **109**, 141-157. https://doi.org/10.1017/S0021859600081089

Andrén, O., & Kätterer, T. (1997). ICBM: The introductory carbon balance model for exploration of soil carbon balances. *Ecological Applications*, **7**, 1226-1236. https://doi.org/10.1890/1051-0761(1997)007[1226:ITICBM]2.0.CO;2

Andrén, O., Kätterer, T., Karlsson, T., & Eriksson, J. (2008). Soil C balances in Swedish agricultural soils 1990-2004, with preliminary projections. *Nutrient Cycling in Agroecosystems*, **81**, 129–144. https://doi.org/10.1007/s10705-008-9177-z

Andriulo, A., Mary, B., & Guerif, J. (1999). Modelling soil carbon dynamics with various cropping sequences on the rolling pampas. *Agronomie*, **19**, 365–377. https://doi.org/10.1051/agro:19990504

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A., … Wolf, J. (2013). Uncertainty in simulating wheat yields under climate change. *Nature Climate Change*, **3**, 827–832. https://doi.org/10.1038/nclimate1916

Barré, P., Eglin, T., Christensen, B. T., Ciais, P., Houot, S., Kätterer, T., ... Chenu, C. (2010). Quantifying and isolating stable soil organic carbon using long-term bare fallow experiments. *Biogeosciences*, **7**, 3839-3850. https://doi.org/10.5194/bg-7-3839-2010

Basso, B., Dumont, B., Maestrini, B., Shcherbak, I., Robertson, G. P., Porter, J. R., … Rosenzweig, C. (2018). Soil organic carbon and nitrogen feedbacks on crop yields under climate change. *Agricultural and Environmental Letters*, **3**, 180026. https://doi.org/10.2134/ael2018.05.0026

Bassu, S., Brisson, N., Durand, J. L., Boote, K., Lizaso, J., Jones, J. W., … Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? *Global Change Biology*, **20**, 2301–2320. https://doi.org/10.1111/gcb.12520

Bellocchi, G., Acutis, M., Fila, G., & Donatelli, M. (2002). An indicator of solar radiation model performance based on a fuzzy expert system. *Agronomy Journal*, **94**, 1222-1233. https://doi.org/10.2134/agronj2002.1222

Bellocchi, G., Rivington, M., Donatelli, M., & Acutis, M. (2010). Validation of biophysical models: issues and methodologies. A review. *Agronomy for Sustainable Development*, **30**, 109-130. https://doi.org/10.1051/agro/2009001

Bispo, A., Andersen, L., Angers, D. A., Bernoux, M., Brossard, M., Cécillon, L., … Eglin, T.K. (2017). Accounting for carbon stocks in soils and measuring GHGs emission fluxes from soils: do we have the necessary standards? Frontiers in Environmental Science, **12 July 2017**. https://doi.org/10.3389/fenvs.2017.00041

Blagodatskaya, E., & Kuzyakov, Y. (2008). Mechanisms of real and apparent priming effects and their dependence on soil microbial biomass and community structure: critical review. *Biology and Fertility of Soils*, **45**, 115–131. https://doi.org/10.1007/s00374-008-0334-y

Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., … Bellocchi, G. (2017). Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. *Science of the Total Environment*, **598**, 445-470. https://doi.org/10.1016/j.scitotenv.2017.03.208

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M. H., Ruget, F., Nicollaud, B., … Delécolle, R. (1998). STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. *Agronomie*, **18**, 311–346. https://doi.org/10.1051/agro:19980501

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., … Sinoquet, H. (2003). An overview of the crop model STICS. *European Journal of Agronomy*, **18**, 309-332. https://doi.org/10.1016/S1161-0301(02)00110-7

Brisson, N., Launay, M., Mary, B., & Baudoin, N. (2008). Conceptual basis, formalizations and parameterization of the STICS crop model. Paris (France): Editions Quae.

Campbell, E. E., & Paustian, K. (2015). Current developments in soil organic matter modeling and the expansion of model applications: a review. *Environmental Research Letters*, **10**, 123004. https://doi.org/10.1088/1748-9326/10/12/123004

Caruso, T., De Vries, F., Bardgett, R. D., & Lehmann, J. (2018). Soil organic carbon dynamics matching ecological equilibrium theory. *Ecology and Evolution*, **8**, 11169-11178. https://doi.org/10.1002/ece3.4586

Cavalli, D., Bellocchi, G., Corti, M., Gallina, P. M., & Bechini, L. (2019). Sensitivity analysis of C and N modules in biogeochemical crop and grassland models following manure addition to soil. *European Journal of Soil Science*, **70**, 833-846. https://doi.org/10.1111/ejss.12793

Challinor, A., Martre, P., Asseng, S., Thornton, P., & Ewert, F. (2014). Making the most of climate impacts ensembles. *Nature Climate Change*, **4**, 77-80. https://doi.org/10.1038/nclimate2117

Chenu, C., Angers, D. A., Barré, P., Derrien, D., Arrouays, D., & Balesdent, J. (2018). Increasing organic stocks in agricultural soils: Knowledge gaps and potential innovations. *Soil and Tillage Research*, **188**, 41-52. https://doi.org/10.1016/j.still.2018.04.011

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74, 829-836. https://doi.org/10.1080/01621459.1979.10481038

Clivot, H., Mouny, J. C., Duparque, A., Dinh, J. L., Denoroy, P., Houot, S., … Mary, B. (2019). Modeling soil organic carbon evolution in long-term arable experiments with AMG model. *Environmental Modelling & Software*, **118**, 99-113. https://doi.org/10.1016/j.envsoft.2019.04.004

Coleman, K., & Jenkinson, D.S. (1999). RothC-26.3 - A model for the turnover of carbon in soil: model description and Windows user guide. Harpenden (UK): Lawes Agricultural Trust.

Confalonieri, R., Acutis, M., Bellocchi, G., & Donatelli, M. (2009). Multi-metric evaluation of the models WARM, CropSyst, and WOFOST for rice. *Ecological Modelling*, **220**, 1395-1410. https://doi.org/10.1016/j.ecolmodel.2009.02.017

Confalonieri, R., Orlando, F., Paleari, L., Stella, T., Gilardelli, C., Movedi, E., ... Acutis, M. (2016). Uncertainty in crop model predictions: what is the role of users? *Environmental Modelling & Software*, **81**, 165-173. https://doi.org/10.1016/j.envsoft.2016.04.009

Coucheney, E., Buis, S., Launay, M., Constantin, J., Mary, B., García de Cortázar-Atauri, I., … Léonard, J. (2015). Accuracy, robustness and behavior of the STICS soil–crop model for plant, water and nitrogen outputs: Evaluation over a wide range of agro-environmental conditions in France. *Environmental Modelling & Software*, **64**, 177-190. https://doi.org/10.1016/j.envsoft.2014.11.024

De Jager, J.M. (1994). Accuracy of vegetation evaporation ratio formulae for estimating final wheat yield. *Water SA*, **20**, 307-314. Retrieved from https://journals.co.za/content/waters/20/4/AJA03784738_2194

Debreczeni, K., & Körschens, M. (2003). Long-term field experiments of the world. *Archives of Agronomy and Soil Science*, **49**, 465-483. https://doi.org/10.1080/03650340310001594754

Dechow, R., Franko, U., Kätterer, T., & Kolbe, H. (2019). Evaluation of the RothC model as a prognostic tool for the prediction of SOC trends in response to management practices on arable land. *Geoderma*, **337**, 463-478. https://doi.org/10.1016/j.geoderma.2018.10.001

Del Grosso, S. J., Parton, W. J., Mosier, A. R., Hartman, M. D., Brenner, J., Ojima, D. S., & Schimel, D. S. (2001). Simulated interaction of carbon dynamics and nitrogen trace gas fluxes using the DayCent model. In M. J. Shaffer, L. Ma, & S. Hansen (Eds.), *Modeling carbon and nitrogen dynamics for soil management* (pp. 303-332). Boca Raton: CRC Press.

Del Grosso, S., Ojima, D., Parton, W., Mosier, A., Peterson, G., & Schimel, D. (2002). Simulated effects of dryland cropping intensification on soil organic matter and greenhouse gas exchanges using the DAYCENT ecosystem model. *Environmental Pollution*, **1**, S75-S83. https://doi.org/10.1016/S0269-7491(01)00260-3

Del Grosso, S., Parton, W., Stohlgren, T., Zheng, D., Bachelet, D., Prince, S., … Olson, R. (2008). Global potential net primary production predicted from vegetation class, precipitation, and temperature. *Ecology*, **89**, 2117-2126. https://doi.org/10.1890/07-0850.1

Dimassi, B., Guenet, B., Saby, N. P. A., Munoz, F., Bardy, M., Millet, F., & Martin, M. P. (2018). The impacts of CENTURY model initialization scenarios on soil organic carbon dynamics simulation in French long-term experiments. *Geoderma*, **311**, 25-36. https://doi.org/10.1016/j.geoderma.2017.09.038

Dungait, J. A. J., Hopkins, D. W., Gregory, A. S., & Whitmore, A. P. (2012). Soil organic matter turnover is governed by accessibility not recalcitrance. *Global Change Biology*, **18**, 1781-1796. https://doi.org/10.1111/j.1365-2486.2012.02665.x

Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., Mcauliffe, R., Recous, S., … Zhang, Q. (2018). Assessing uncertainties in crop and pasture ensemble model simulations of productivity and N_2O emissions. *Global Change Biology*, **24**, e603-e616. https://doi.org/10.1111/gcb.13965

Ehrmann, J., & Ritz, K. (2014). Plant: soil interactions in temperate multi-cropping production systems. *Plant and Soil*, **376**, 1-29. https://doi.org/10.1007/s11104-013-1921-8

Falloon, P., & Smith, P. (2010). Modelling soil carbon dynamics. In W. L. Kutsch, M. Bahn, & A. Heinemeyer (Eds.), *Soil carbon dynamics: An integrated methodology* (pp. 221-244). Cambridge: Cambridge University Press.

Farina, R., Coleman, K., & Whitmore, A. P. (2013). Modification of the RothC model for simulations of soil organic C dynamics in dryland regions. *Geoderma*, **200-201**, 18-30. https://doi.org/10.1016/j.geoderma.2013.01.021

Franko, U., Kolbe, H., Thiel, E., & Liess, E. (2011). Multi-site validation of a soil organic matter model for arable fields based on generally available input data. *Geoderma*, **166**, 119-134. https://doi.org/10.1016/j.geoderma.2011.07.019

Franko, U., & Spiegel, H. (2016). Modeling soil organic carbon dynamics in an Austrian long-term tillage field experiment. *Soil and Tillage Research*, **156**, 83-90.

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

Fuchs, R., Schulp, C. J. E., Hengeveld, G. M., Verburg, P. H., Clevers, J. G. P. W., Schelhaas, M.-J., & Herold, M. (2016). Assessing the influence of historic net and gross land changes on the carbon fluxes of Europe. *Global Change Biology*, **22**, 2526-2539. https://doi.org/10.1111/gcb.13191

Gardi, C., Visioli, G., Conti, F. D., Scotti, M., Menta, C., & Bodini, A. (2016). High Nature Value Farmland: assessment of soil organic carbon in Europe. Frontiers in Environmental Science, 21 June 2016. https://doi.org/10.3389/fenvs.2016.00047

Gijsman, A. J., Hoogenboom, G., Parton, W. J., & Kerridge, P. C. (2002). Modifying DSSAT crop models for low-input agricultural systems using a soil organic matter-residue module from CENTURY. *Agronomy Journal*, **94**, 462-474. https://doi.org/10.2134/agronj2002.4620

Gottschalk, P., Smith, J. U., Wattenbach, M., Bellarby, J., Stehfest, E., Arnell, N., … Smith, P. (2012). How will organic carbon stocks in mineral soils evolve under future climate? Global projections using RothC for a range of climate change scenarios. *Biogeosciences*, **9**, 3151-3171. https://doi.org/10.3390/soilsystems3020028

Gross C. D., & Harrison, R. B. (2019). The case for digging deeper: soil organic carbon storage, dynamics, and controls in our changing world. *Soil Systems*, **3**, 28. https://doi.org/10.3390/soilsystems3020028

Guenet, B., Neill, C., Bardoux, G., & Abbadie, L. (2010). Is there a linear relationship between priming effect intensity and the amount of organic matter input? *Applied Soil Ecology*, **46**, 436–442. https://doi.org/10.1016/j.apsoil.2010.09.006

Herbst, M., Welp, G., Macdonald, A., Jate, M., Hädicke, A., Scherer, H., … Vanderborght, J. (2018). Correspondence of measured soil carbon fractions and RothC pools for equilibrium and non-equilibrium states. *Geoderma*, **314**, 37-46. https://doi.org/10.1016/j.geoderma.2017.10.047

Hill, M. J. (2003). Generating generic response signals for scenario calculation of management effects on carbon sequestration in agriculture: approximation of main effects using CENTURY. *Environmental Modelling & Software*, **18**, 899-913. https://doi.org/10.1016/S1364-8152(03)00054-9

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., … Keating, B. A. (2014). APSIM - Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, **62**, 327-350. https://doi.org/10.1016/j.envsoft.2014.07.009

Huntzinger, D. N., Schwalm, C., Michalak, A. M., Schaefer, K., King, A. W., Wei, Y., … Zhu, Q. (2013). The North American Carbon Program Multi-scale synthesis and Terrestrial Model Intercomparison Project-Part 1: Overview and experimental design. *Geoscientific Model Development*, **6**, 2121-2133. https://doi.org/10.5194/gmd-6-2121-2013

Johnston, A. E., & Poulton, P. R. (2018). The importance of long-term experiments in agriculture: their management to ensure continued crop production and soil fertility; the Rothamsted experience. *European Journal of Soil Science*, **69**, 113-125. https://doi.org/10.1111/ejss.12521

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., … Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, **18**, 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7

Jørgensen, S. E., Kamp-Nielsen, L., Christensen, T., Windolf-Nielsen, J., & Westergaard, B. (1986). Validation of a prognosis based upon a eutrophication model. Ecological Modelling, **35**, 165-182. https://doi.org/10.1016/0304-3800(86)90024-4

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. L., Robertson, M. J., Holzworth, D., … Smith, C. J. (2003). An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, **18**, 267-288. https://doi.org/10.1016/S1161-0301(02)00108-9

Keel, S. G., Leifeld, J., Mayer, J., Taghizadeh-Toosi, A., and Olesen, J. E. (2017). Large uncertainty in soil carbon modelling related to method of calculation of plant carbon input in agricultural systems. *European Journal of Soil Science*, **68**, 953-863. https://doi.org/10.1111/ejss.12454

Kirschbaum, M.U.F. (1999). CenW, a forest growth model with linked carbon, energy, nutrient and water cycles. *Ecological Modelling*, **118**, 17–59. https://doi.org/10.1016/S0304-3800(99)00020-4

Kirschbaum, M. U. F., Rutledge, S., Kuijper, I. A., Mudge, P. L., Puche, N., Wall, A. M., … Campbell, D. I. (2015). Modelling carbon and water exchange of a grazed pasture in New Zealand constrained by eddy covariance measurements. *Science of the Total Environment*, **512-513**, 273-286. https://doi.org/10.1016/j.scitotenv.2015.01.045

Kirschbaum, M. U. F., & Paul, K. I. (2002). Modelling carbon and nitrogen dynamics in forest soils with a modified version of the CENTURY model. *Soil Biology & Biochemistry*, **34**, 341-354. https://doi.org/10.1016/S0038-0717(01)00189-4

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, **15**, 259-263. https://doi.org/10.1127/0941-2948/2006/0130

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., … Colin Prentice, I. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, **19**, GB1015. https://doi.org/10.1029/2003GB002199

Kuhry, P., & Vitt, D.H. (1996). Fossil carbon/nitrogen ratios as a measure of peat decomposition. *Ecology*, **77**, 271–275. https://doi.org/10.2307/2265676

Kuka, K. (2005). Modellierung des Kohlenstoffhaushaltes in Ackerböden auf der Grundlage bodenstrukturabhängiger Umsatzprozesse. PhD thesis, Martin-Luther-University Halle-Wittenberg. Retrieved from https://gepris.dfg.de/gepris/projekt/5247578?context=projekt&task=showDetail&id=5247578 & (in German)

Kuka, K., Franko, U., & Rühlmann, J. (2007) Modelling the impact of pore space distribution on carbon turnover. *Ecological Modelling*, **208**, 295–306. https://doi.org/10.1016/j.ecolmodel.2007.06.002

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, **304**, 1623-1626. https://doi.org/10.1126/science.1097396

Lal, R. (2014). Soil conservation and ecosystem services. *International Soil and Water Conservation Research*, **2**, 36-47. https://doi.org/10.1016/S2095-6339(15)30021-6

Lardy, R., Bellocchi, G., & Soussana, J.-F. (2011). A new method to determine soil organic carbon equilibrium. *Environmental Modelling & Software*, **26**, 1759-1763. https://doi.org/10.1016/j.envsoft.2011.05.016

Lavallee, J. M., Soong, J. L., & Cotrufo, M. F. (2020). Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. *Global Change Biology*, **26**, 261-273. https://doi.org/10.1111/gcb.14859

Lehmann, J., & Kleber, M. (2015). The contentious nature of soil organic matter. *Nature*, **528**, 60-68. https://doi.org/10.1038/nature16069

Li, C., Salas, W., Zhang, R., Krauter, C., Rotz, A., & Mitloehner, F. (2012). Manure-DNDC: a biogeochemical process model for quantifying greenhouse gas and ammonia emissions from livestock manure systems. *Nutrient Cycling in Agroecosystems*, **93**, 163-200. https://doi.org/10.1007/s10705-012-9507-z

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., … Bouman, B. (2015). Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. *Global Change Biology*, **21**, 1328-1341. https://doi.org/10.1111/gcb.12758

Ma, S., Lardy, R., Graux, A.-I., Ben Touhami, H., Klumpp, K., Martin, R., Bellocchi, G. (2015). Regional-scale analysis of carbon and water cycles on managed grassland systems. *Environmental Modelling & Software*, **72**, 356-371. https://doi.org/10.1016/j.envsoft.2015.03.007

Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., … Zhu, Y. (2017). Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research*, **202**, 5-20. https://doi.org/10.1016/j.fcr.2016.05.001

Manzoni, S., & Porporato, A. (2009). Soil carbon and nitrogen mineralization: Theory and models across scales. *Soil Biology & Biochemistry*, **41**, 1355-1379. https://doi.org/10.1016/j.soilbio.2009.02.031

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., … Wolf, J. (2015). Multimodel ensembles of wheat growth: Many models are better than one. *Global Change Biology*, **21**, 911-925. https://doi.org/10.1111/gcb.12768

Mary, B., Clivot, H., Blaszczyk, N., Labreuche, L., & Ferchaud, F. (2020). Soil carbon storage and mineralization rates are affected by carbon inputs rather than physical disturbance: Evidence from a 47-year tillage experiment. *Agriculture, Ecosystems & Environment*, **299**, 106972. https://doi.org/10.1016/j.agee.2020.106972

edlyn, B. E., Robinson, A. P., Clement, R., & McMurtrie, R. E. (2005). On the validation of models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls. *Tree Physiology*, **25**, 839-857. https://doi.org/10.1093/treephys/25.7.839

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., … Winowiecki, L. (2017). Soil carbon 4 per mille. *Geoderma*, **292**, 59–86. https://doi.org/10.1016/j.geoderma.2017.01.002

Minunno, F., Peltoniemi, M., Launiainen, S., & Mäkelä, A. (2014). Integrating ecosystems measurements from multiple eddy-covariance sites to a simple model of ecosystem process - are there possibilities for a uniform model calibration? *Geophysical Research Abstracts*, **16**, EGU2014-10706-3. Retrieved from https://meetingorganizer.copernicus.org/EGU2014/orals/14065

Mirtl, M., Borer, E. T., Djukic, I., Forsius, M., Haubold, H., Hugo, W., Jourdane, J., … Haase, P. (2018). Genesis, goals and achievements of long-term ecological research at the global scale: a critical review of ILTER and future directions. *Science of the Total Environment*, **626**, 1439-1462. https://doi.org/10.1016/j.scitotenv.2017.12.001

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., & Veith, T. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, **50**, 885-900. https://doi.org/10.13031/2013.23153

Moyano, F. E., Vasilyeva, N., & Menichetti, L. (2018). Diffusion limitations and Michaelis–Menten kinetics as drivers of combined temperature and moisture effects on carbon fluxes of mineral soils. *Biogeosciences*, **15**, 5031–5045. https://doi.org/10.5194/bg-15-5031-2018

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - a discussion of principles. *Journal of Hydrology*, **10**, 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Nemo, R., Klumpp, K., Coleman, K., Dondini, M., Goulding, K., Hastings, A., … Smith, P. (2016). Soil organic carbon (SOC) equilibrium and model initialisation methods: an application to the Rothamsted Carbon (RothC) model. *Environmental Modeling & Assessment*, **22**, 215-229.

Nendel, C., Berg, M., Kersebaum, K. C., Mirschel, W., Specka, X., Wegehenkel, M., … Wieland, R. (2011). The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. *Ecological Modelling*, **222**, 1614–1625. https://doi.org/10.1016/j.ecolmodel.2011.02.018

Parton, W. J., Del Grosso, S., Plante, A. F., Adair, E. C., & Lutz, S. M. (2015). Modeling the dynamics of soil organic matter and nutrient cycling. In E. A. Paul (Ed.), *Soil microbiology, ecology and biochemistry, 4th edition* (pp. 505-537). Amsterdam: Elsevier Academic Press.

Parton, W. J., Hartman, M., Ojima, D., & Schimel, D. (1998). DAYCENT and its land surface submodel: description and testing. *Global and Planetary Change*, **19**, 35-48. https://doi.org/10.1016/S0921-8181(98)00040-X

Parton, W. J., Schimel, D. S., & Cole, C.V., & Ojima, D. S. (1987). Analysis of factors controlling soil organic matter levels in Great Plains grasslands. Soil Science Society of America Journal, **51**, 1173–1179. https://doi.org/10.2136/sssaj1987.03615995005100050015x

Parton, W. J., Schimel, D. S., Ojima, D. S., & Cole, C. V. (1994). A general model for soil organic matter dynamics: sensitivity to litter chemistry, texture and management. In R. B. Bryant & R. W. Arnold (Eds.), *Quantitative modeling of soil forming processes* (pp. 147–167). Madison, WI (USA): SSSA Spec. Pub. 39. ASA, CSSA and SSSA.

Porter, C. H., Jones, J. W., Adiku, S., Gijsman, A. J., Gargiulo, O., & Naab, J. B. (2009). Modeling organic carbon and carbon-mediated soil processes in DSSAT v4.5. *Operational Research*, **10**, 247-278. https://doi.org/10.1007/s12351-009-0059-1

Puche, N. J. B., Senapati, N., Flechard, C. R., Klumpp, K., Kirschbaum, M. U. F, & Chabbi, A. (2019). Modelling carbon and water fluxes of managed grasslands: comparing flux variability and net carbon budgets between grazed and mowed systems. *Agronomy*, **9**, 183. https://doi.org/10.3390/agronomy9040183

Reynolds, K. M., Thomson, A. J., Köhl, M., Shannon, M. A., Ray, D., & Rennolls, K. (2007). Sustainable forestry: from monitoring and modelling to knowledge management and policy science. Wallingford: CAB International.

Rodríguez, A., Ruiz-Ramos, M., Palosuo, T., Carter, T. R., Fronzek, S., Lorite, I. J., … Rötter, R. P. (2019). Implications of crop model ensemble size and composition for estimates of adaptation effects and agreement of recommendations. *Agricultural and Forest Meteorology*, **15**, 351-362. https://doi.org/10.1016/j.agrformet.2018.09.018

Rötter, R. P., Palosuo, T., Kersebaum, K. C., Angulo, C., Bindi, M., Ewert, F., … Trnka, M. (2012). Simulation of spring barley yield in different climatic zones of Northern and Central Europe – A comparison of nine crop models. *Field Crops Research*, **133**, 23–36. https://doi.org/10.1016/j.fcr.2012.03.016

Ruane, A. C., Hudson, N. I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., … Wolf, J. (2016). Multi-wheat-model ensemble responses to interannual climate variability. *Environmental Modelling & Software*, **81**, 86-101. https://doi.org/10.1016/j.envsoft.2016.03.008

Rumpel, C., Amiraslani, F., Koutika, L. S., Smith, P., Whitehead, D., & Wollenberg, E. (2018). Put more carbon in soils to meet Paris climate pledges. *Nature*, 564, 32-34. https://doi.org/10.1038/d41586-018-07587-4

Saffih-Hdadi, K., & Mary, B. (2008). Modeling consequences of straw residues export on soil organic carbon. *Soil Biology & Biochemistry*, **40**, 594–607. https://doi.org/10.1016/j.soilbio.2007.08.022

Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., … Bellocchi, G. (2017). Multi-model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean grasslands: Uncertainties and ensemble performance. *European Journal of Agronomy*, **88**, 22-40. https://doi.org/10.1016/j.eja.2016.06.006

Sándor, R., Ehrhardt, F., Brilli, L., Carozzi, M., Recous, S., Smith, P., … Bellocchi, G. (2018a). The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed grasslands. *Science of the Total Environment*, **642**, 292-306. https://doi.org/10.1016/j.scitotenv.2018.06.020

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., … Bellocchi, G. (2020). Ensemble modelling of carbon fluxes in grasslands and croplands. *Field Crops Research*, **252**, 107791. https://doi.org/10.1016/j.fcr.2020.107791

Sándor, R., Picon-Cochard, C., Martin, R., Louault, F., Klumpp, K., Borras, D., & Bellocchi, G., (2018b). Plant acclimation to temperature: Developments in the Pasture Simulation model. *Field Crops Research*, **222**, 238-255. https://doi.org/10.1016/j.fcr.2017.05.030

Schimel, J. P., & Weintraub, M. N. (2003). The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model. *Soil Biology & Biochemistry*, **35**, 549–563. https://doi.org/10.1016/S0038-0717(03)00015-4

Shumilovskikh, L. S., Novenko, E., & Giesecke, T. (2018). Long-term dynamics of the East European forest-steppe ecotone. *Journal of Vegetation Science*, **29**, 416-426. https://doi.org/10.1111/jvs.12585

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., … Venevsky, S. (2003). Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, **9**, 161-185. https://doi.org/10.1046/j.1365-2486.2003.00569.x

Smith, J., Gottshcalk, P., Bellarby, J., Chapman, S., Lilly, A., Towers, W., … Smith, P. (2010a). Estimating changes in national soil carbon stocks using ECOSSE – a new model that includes upland organic soils. Part I. Model description and uncertainty in national scale simulations of Scotland. *Climate Research*, **45**, 179-192. https://doi.org/10.3354/cr00899

Smith, J., Gottschalk, P., Bellarby, J., Chapman, S., Lilly, A., Towers, W., … Smith, P. (2010b). Estimating changes in national soil carbon stocks using ECOSSE - a new model that includes upland organic soils. Part II. Application in Scotland. *Climate Research*, **45**, 193-205. https://doi.org/10.3354/cr00902

Smith, P., Smith, J., Flynn, H., Killham, K., Rangel-Castro, I., Foereid, B., … Falloon, P., 2007. ECOSSE: Estimating Carbon in Organic Soils - Sequestration and Emissions. Final Report. SEERAD Report, 166 pp. Retrieved from http://nora.nerc.ac.uk/id/eprint/2233

Smith, P., Smith, J. U., Powlson, D. S., McGill, W. B., Arah, R. M., Chertov, O. G., … Whitmore, A. P. (1997). A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. *Geoderma*, **81**, 153-225. https://doi.org/10.1016/S0016-7061(97)00087-6

Smith, W. N., Grant, B. B., Campbell, C. A., McConkey, B. G., Desjardins, R. L., Kröbel, R. & Malhi, S. S. (2012). Crop residue removal effects on soil carbon: Measured and inter-model comparisons. *Agriculture, Ecosystems & Environment*, **161**, 27-38. https://doi.org/10.1016/j.agee.2012.07.024

Smith, W. N., Grant, B., Qi, Z., He, W., VanderZaag, A., Drury, C. F., & Helmers, M. (2020). Development of the DNDC model to improve soil hydrology and incorporate mechanistic tile drainage: A comparative analysis with RZWQM2. *Environmental Modelling & Software*, **123**, 104577. https://doi.org/10.1016/j.envsoft.2019.104577

Soussana, J.-F., Lutfalla, S., Ehrhardt, F., Rosenstock, T. S., Lamanna, C., Havlik, P., … Lal, R. (2017). Matching policy and science: Rationale for the '4 per 1000 - soils for food security and climate' initiative. *Soil and Tillage Research*, **188**, 3-15. https://doi.org/10.1016/j.still.2017.12.002

Specka, X., Nendel, C., Hagemann, U., Pohl, M., Hoffmann, M., Barkusky, D., … van Oost, K. (2016). Reproducing CO_2 exchange rates o a crop rotation at contrasting terrain positions using two different modelling approaches. *Soil and Tillage Research*, **156**, 219–229. https://doi.org/10.1016/j.still.2015.05.007

Stella, T., Mouratiadou, I., Gaiser, T., Berg-Mohnicke, M., Wallor, E., Ewert, F., & Nendel, C. (2019). Estimating the contribution of crop residues to soil organic carbon conservation. Environmental Research Letters 14, 094008. https://doi.org/10.1088/1748-9326/ab395c

Taghizadeh–Toosi, A., Christensen, B. T., Hutchings, N. J., Vejlin, J., Kätterer, T., Glendining, M., & Olesen, J. E. (2014a). C-TOOL: A simple model for simulating whole-profile carbon storage in temperate agricultural soils. *Ecological Modelling*, **292**, 11-25. https://doi.org/10.1016/j.ecolmodel.2014.08.016

Taghizadeh-Toosi, A., Olesen, J. E., Kristensen, K., Elsgaard, L., Østergaard, H. S., Lægdsmand, M., … Christensen, B. T. (2014b). Changes in carbon stocks of Danish agricultural mineral soils between 1986 and 2009. *European Journal of Soil Science*, **65**, 730-740. https://doi.org/10.1111/ejss.12169

Taghizadeh-Toosi, A., & Olesen, J. E. (2016). Modelling soil organic carbon in Danish agricultural soils suggests low potential for future carbon sequestration. *Agricultural Systems*, **145**, 83-89. https://doi.org/10.1016/j.agsy.2016.03.004

Taghizadeh-Toosi, A., Christensen, B. T., Glendining, M., & Olesen, J. E. (2016). Consolidating soil carbon turnover models by improved estimates of belowground carbon input. *Scientific Reports*, **6**, 32568. https://doi.org/10.1038/srep32568

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review*, **38**, 55-94. https://doi.org/10.2307/210739

Thorp, K. R., White, J. W., Porter, C. H., Hoogenboom, G., Nearing, G. S., & French, A. N. (2012). Methodology to evaluate the performance of simulation models for alternative compiler and operating system configurations. *Computers and Electronics in Agriculture*, **81**, 62-71. https://doi.org/10.1016/j.compag.2011.11.008

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., & Allison, S. D. (2013). Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*, **10**, 1717–1736. https://doi.org/10.5194/bg-10-1717-2013

Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., … Allison, S. D. (2014). Changes in soil organic carbon storage predicted by Earth system models during the 21st century. *Biogeosciences*, **11**, 2341–2356. https://doi.org/10.5194/bg-11-2341-2014

Tuomi, M., Thum, T., Järvinen, H., Fronzek, S., Berg, B., Harmon, M., … Liski, J. (2009). Leaf litter decomposition - Estimates of global variability based on Yasso07 model. *Ecological Modelling*, **220**, 3362-3371. https://doi.org/10.1016/j.ecolmodel.2009.05.016

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., … Zhang, Z. (2018). Multi-model ensembles improve predictions of crop-environment-management interactions. *Global Change Biology*, **24**, 5072-5083. https://doi.org/10.1111/gcb.14411

Wallach, D., Palosuo, T., Thorburn, P., Seidel, S. J., Gourdain, E., Asseng, S., … Zhu, Y. (2020). How well do crop models predict phenology, with emphasis on the effect of calibration? *bioRxiv*, March 30, 2020. https://doi.org/10.1101/708578

Wallach, D., & Thorburn, P. J. (2017). Estimating uncertainty in crop model predictions: Current situation and future prospects. *European Journal of Agronomy*, **88**, A1-A7. https://doi.org/10.1016/j.eja.2017.06.001

Weihermüller, L., Graf, A., Herbst, M., & Vereecken, H. (2013). Simple pedotransfer functions to initialize reactive carbon pools of the RothC model. *European Journal of Soil Science*, **64**, 567-575. https://doi.org/10.1111/ejss.12036

White, J. W., Hoogenboom, G., Kimball, B. A., & Wall, G. W. (2011). Methodologies for simulating impacts of climate change on crop production. *Field Crops Research*, **124**, 357-368. https://doi.org/10.1016/j.fcr.2011.07.001

Whitehead, D., Schipper, L. A., Pronger, J., Moinet, G. Y., Mudge, P. L., Pereira, R. C., … Camps-Arbestain, M. (2018). Management practices to reduce losses or increase soil carbon stocks in temperate grazed grasslands: New Zealand as a case study. *Agriculture, Ecosystems & Environment*, **265**, 432-443. https://doi.org/10.1016/j.agee.2018.06.022

Wieder, W. R., Boehnert, J., & Bonan, G. B. (2014). Evaluating soil biogeochemistry parameterizations in Earth system models with observations. *Global Biogeochemical Cycles*, **28**, 211-222. https://doi.org/10.1002/2013GB004665

Willmott, C. J., & Wicks, D. E. (1980). An empirical method for the spatial interpolation of monthly precipitation within California. *Physical Geography*, **1**, 59-73. https://doi.org/10.1080/02723646.1980.10642189

Wutzler, T., & Reichstein, M. (2007). Soils apart from equilibrium - consequences for soil carbon balance modelling. *Biogeosciences*, **4**, 125-136. https://doi.org/10.5194/bg-4-125-2007

Wutzler, T., & Reichstein, M. (2008). Colimitation of decomposition by substrate and decomposers - a comparison of model formulations. *Biogeosciences*, **5**, 749–759. https://doi.org/10.5194/bg-5-749-2008

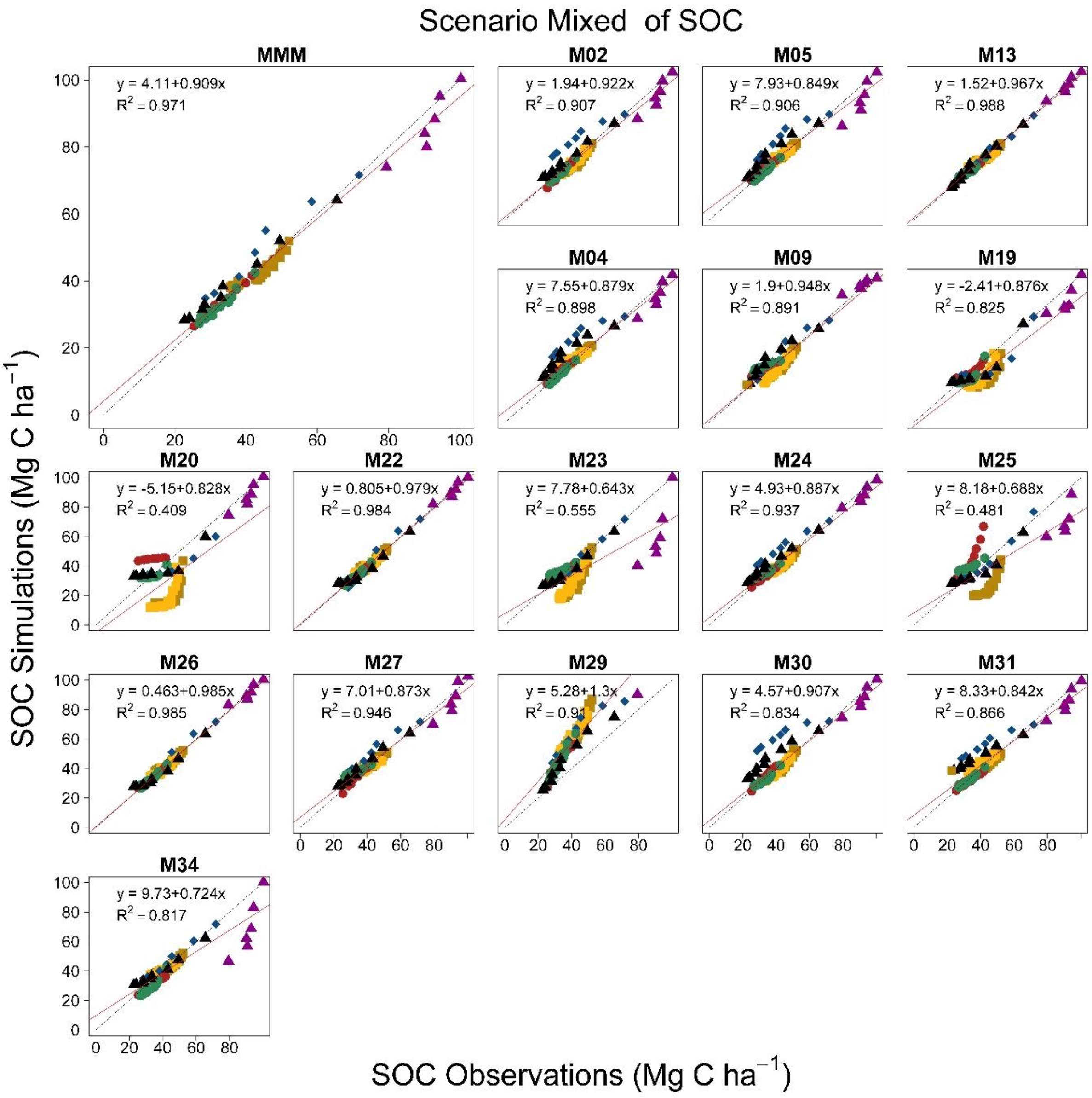
Wutzler, T., & Reichstein, M. (2013). Priming and substrate quality interactions in soil organic matter models. *Biogeosciences*, **10**, 2089–2103. https://doi.org/10.5194/bg-10-2089-2013

Xu, X., Wen L., & Kiely, G. (2011). Modeling the change in soil organic carbon of grassland in response to climate change: Effects of measured versus modelled carbon pools for initializing the Rothamsted Carbon model. *Agriculture, Ecosystems & Environment*, **140**, 372-381. https://doi.org/10.1016/j.agee.2010.12.018

Yadav, V., & Malanson, G. (2007). Progress in soil organic matter research: litter decomposition, modelling, monitoring and sequestration. *Progress in Physical Geography*, **31**, 131-154. https://doi.org/10.1177/0309133307076478Zhu, D., Ciais, P., Krinner, G., Maignan, F., Puig, A.J., & Hugelius, G. (2019). Controls of soil organic matter on soil thermal dynamics in the northern high latitudes. *Nature Communications*, **10**, 3172. https://doi.org/10.1038/s41467-019-11103-1

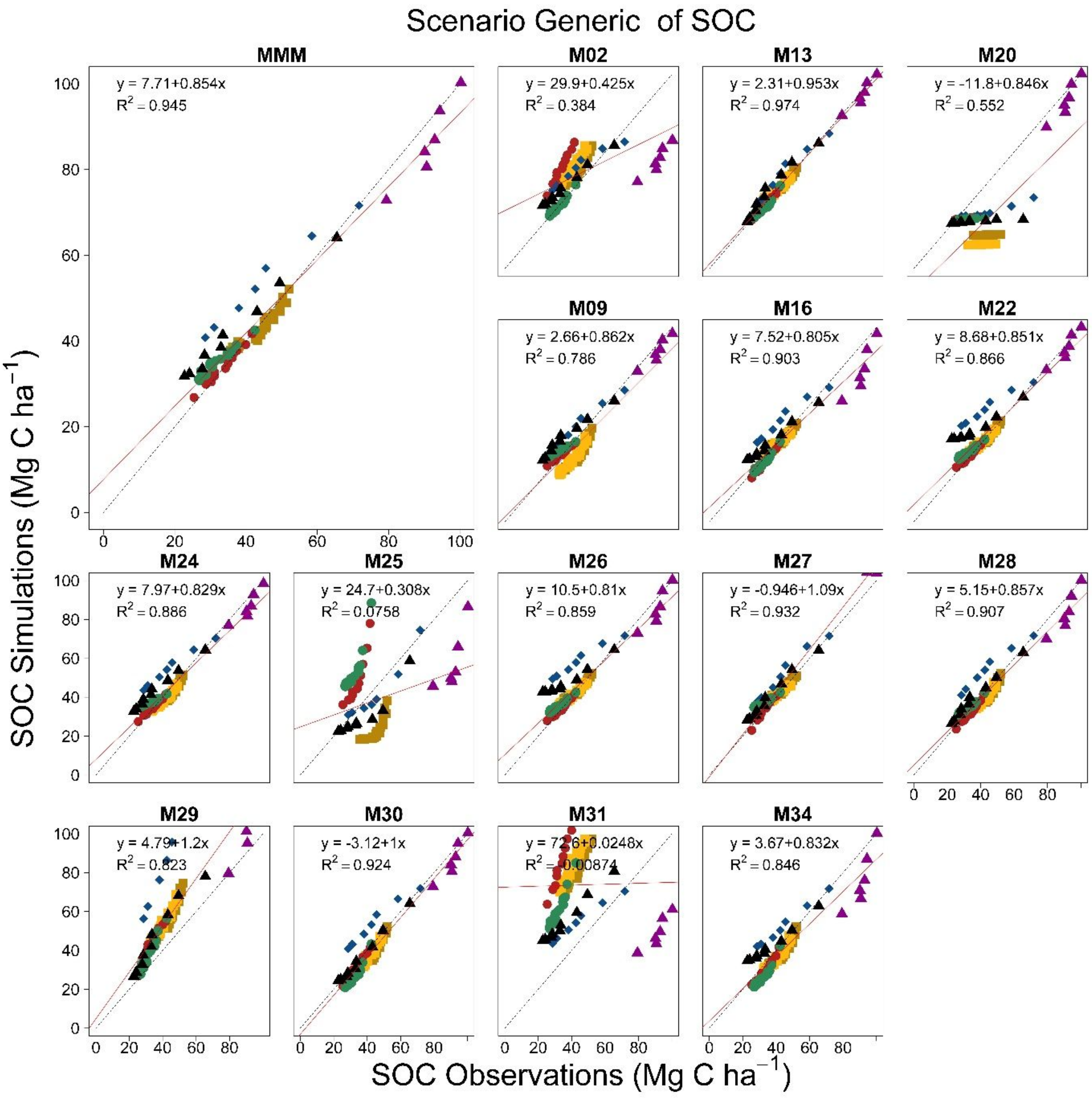
Appendix A

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



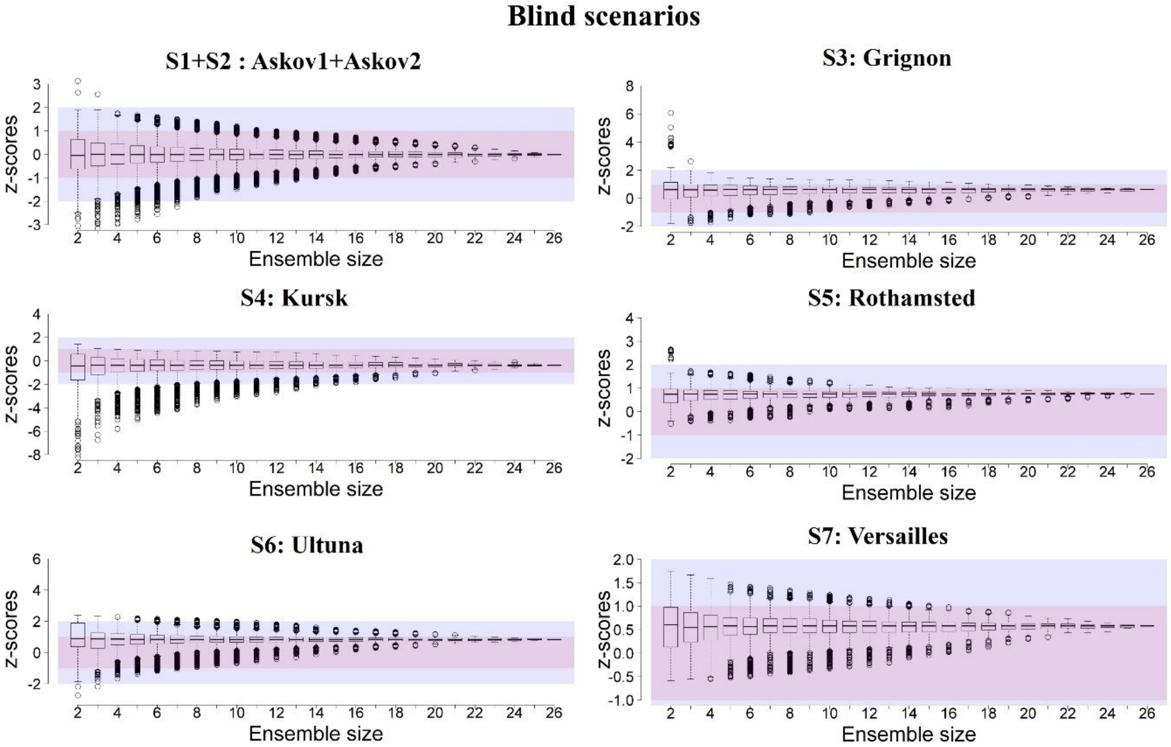
Appendix B

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



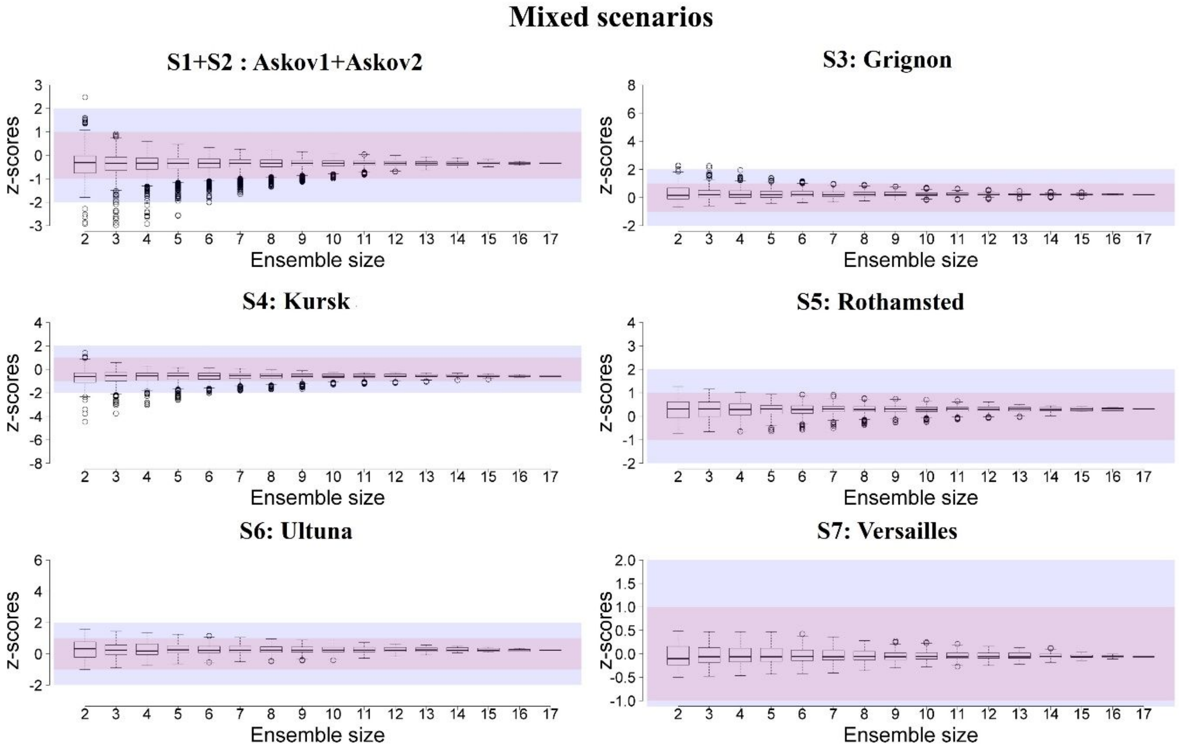
Appendix C

z-scores calculated with different ensemble sizes for SOC estimates obtained with Bln 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).



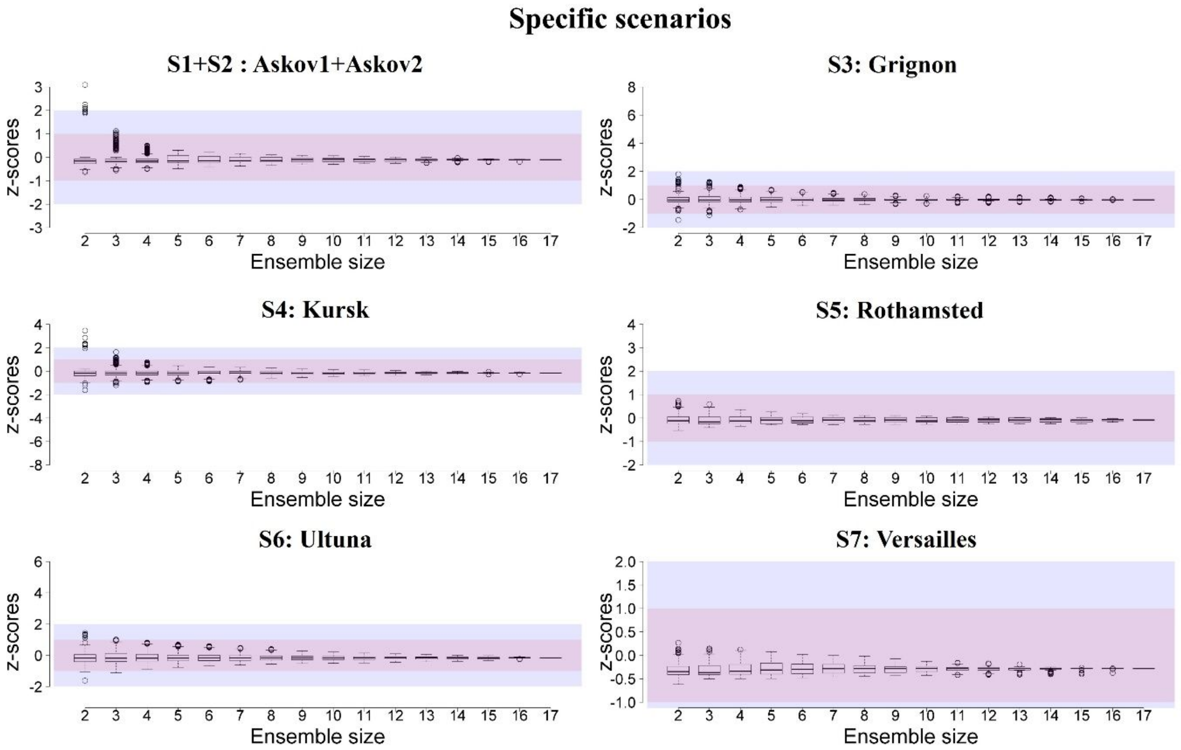
Appendix D

z-scores calculated with different ensemble sizes for SOC estimates obtained with Mix 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).



Appendix E

z-scores calculated with different ensemble sizes for SOC estimates obtained with Spe 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).



Appendix F

Individual and multi-model ensemble (MMM) performance metrics (as in Table 4) for blind (Bln) and calibration scenarios (Mix, Spe and Gen as in Table 3) across sites. Red (italic) and blue (bold) 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^2																												
	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	**0.95**	0.76	0.92	0.41	0.86	0.76	0.92	*0.21*	0.82	0.35	0.57	0.80	0.94
	Gen	NA	0.39	NA	NA	NA	NA	NA	0.79	NA	**0.97**	0.90	NA	NA	0.56	0.87	NA	0.89	0.09	0.86	0.93	0.91	0.82	0.93	*~0.00*	NA	0.85	0.95
	Mix	NA	0.91	NA	0.90	0.91	NA	NA	0.89	NA	**0.99**	NA	NA	0.83	*0.41*	0.98	0.56	0.94	0.49	**0.99**	0.95	NA	0.91	0.84	0.87	NA	0.82	0.97
	Spe	0.97	**0.99**	NA	0.98	**0.99**	NA	NA	**0.99**	NA	**0.99**	0.96	NA	NA	0.96	0.98	**0.99**	NA	0.91	**0.99**	0.97	*0.88*	0.93	0.98	0.94	NA	NA	**0.99**
d																												

Bln	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	**0.99**	0.89	0.97	0.73	0.95	0.91	0.95	0.59	0.95	*0.52*	0.85	0.93	0.98
Gen	NA	0.71	NA	NA	NA	NA	NA	0.93	NA	**0.99**	0.97	NA	NA	0.66	0.96	NA	0.97	0.53	0.95	0.97	0.97	0.81	0.97	*0.23*	NA	0.94	0.98
Mix	NA	0.97	NA	0.96	0.97	NA	NA	0.97	NA	**~1.00**	NA	NA	0.89	*0.69*	**~1.00**	0.79	0.98	0.81	**~1.00**	0.98	NA	0.76	0.96	0.96	NA	0.93	0.99
Spe	0.99	**~1.00**	NA	**~1.00**	**~1.00**	NA	NA	**~1.00**	NA	**~1.00**	0.99	NA	NA	0.99	**~1.00**	0.99	NA	0.97	**~1.00**	0.99	0.95	*0.76*	0.99	0.98	NA	NA	**~1.00**
RRMSE (%)																											
Bln	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	**7.9**	25.4	11.0	36.6	14.0	24.0	14.4	48.4	16.3	*69.1*	27.7	16.3	10.4
Gen	NA	30.8	NA	NA	NA	NA	NA	17.9	NA	**5.7**	11.5	NA	NA	51.3	14.0	NA	12.1	49.4	14.5	12.7	10.9	37.9	12.4	*92.1*	NA	15.8	10.6
Mix	NA	11.0	NA	12.6	11.5	NA	NA	11.7	NA	**3.8**	NA	NA	23.3	45.6	4.4	29.0	8.9	33.0	4.2	9.4	NA	*46.5*	14.4	13.4	NA	15.9	7.2

	Spe	6.5	3.4	NA	5.0	**3.2**	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	*46.2*	5.5	8.7	NA	NA	**3.2**
P(t)	Bln	*~0.00*	*~0.00*	*~0.00*	*~0.00*	*~0.00*	*~0.00*	*~0.00*	*~0.00*	*~0.00*	*~0.00*	**0.64**	0.02	*~0.00*	*~0.00*	0.31	*~0.00*	*~0.00*	*~0.00*	0.45	0.05	*~0.00*	*~0.00*	0.13	*~0.00*	*~0.00*	0.01	*~0.00*
	Gen	NA	*~0.00*	NA	NA	NA	NA	NA	*~0.00*	NA	0.13	**0.17**	NA	NA	*~0.00*	*~0.00*	NA	0.08	0.04	*~0.00*	*~0.00*	0.06	*~0.00*	*~0.00*	*~0.00*	NA	*~0.00*	*~0.00*
	Mix	NA	*~0.00*	NA	*~0.00*	*~0.00*	NA	NA	0.55	NA	0.31	NA	NA	*~0.00*	*~0.00*	**0.76**	*~0.00*	0.54	*~0.00*	0.31	*~0.00*	NA	*~0.00*	0.24	*~0.00*	NA	*~0.00*	0.49
	Spe	0.46	0.99	NA	0.06	0.03	NA	NA	0.85	NA	0.34	*~0.00*	NA	NA	0.12	0.93	*~0.00*	NA	*~0.00*	**~1.00**	0.29	*~0.00*	*~0.00*	*~0.00*	0.68	NA	NA	0.83
EF	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	**0.95**	0.47	0.90	-0.11	0.84	0.53	0.83	-0.93	0.78	*-2.95*	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	*-6.00*	NA	0.79	**0.93**

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

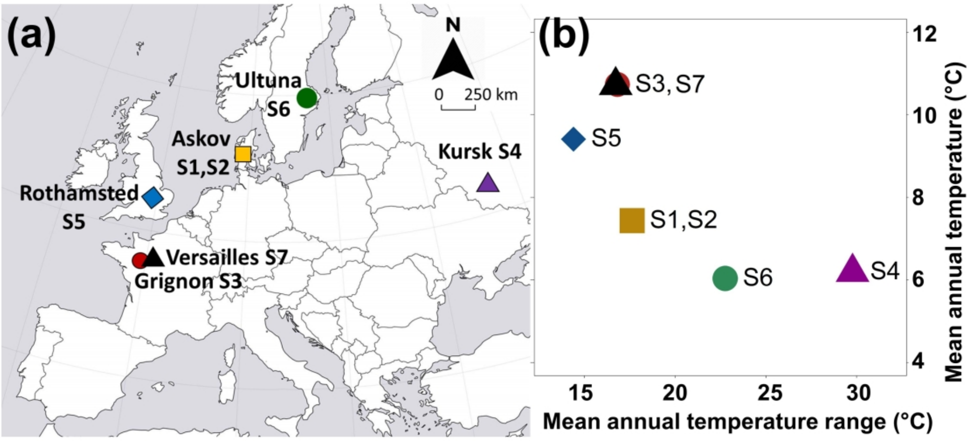


Fig. 1. Location (a) and characterisation of the study sites (b) with respect to mean annual temperature (°C) and mean annual temperature range (°C). Details about study sites are in Table 2.

558x254mm (150 x 150 DPI)

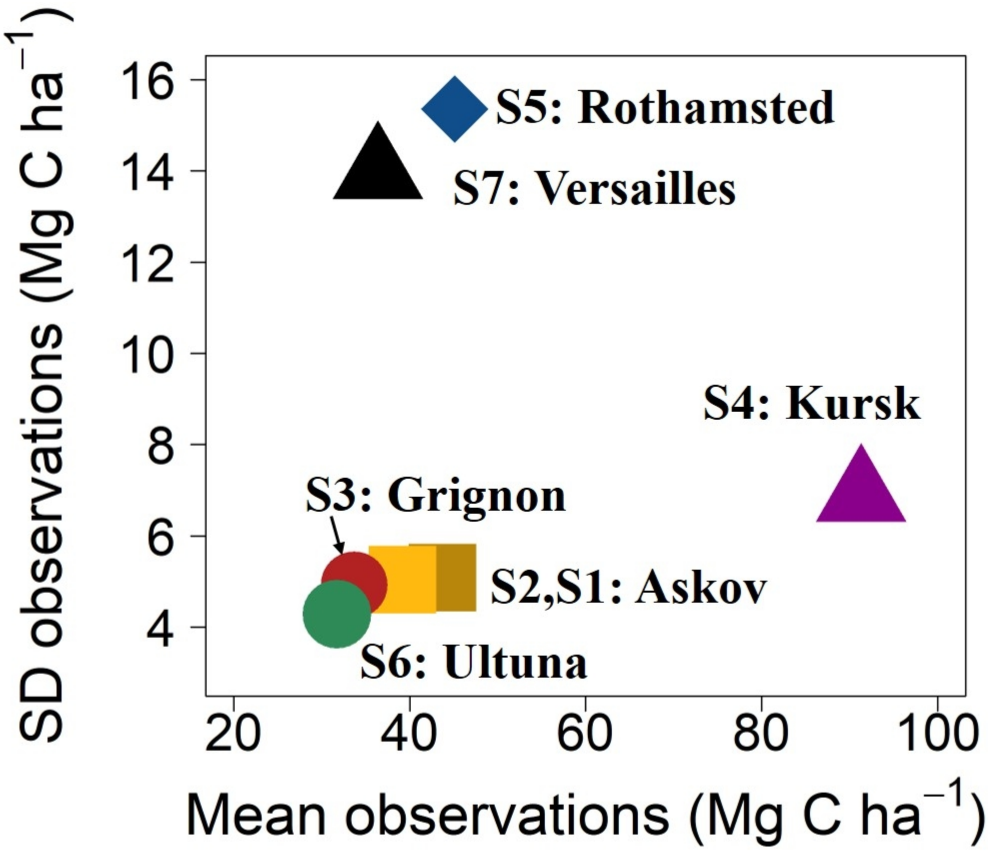
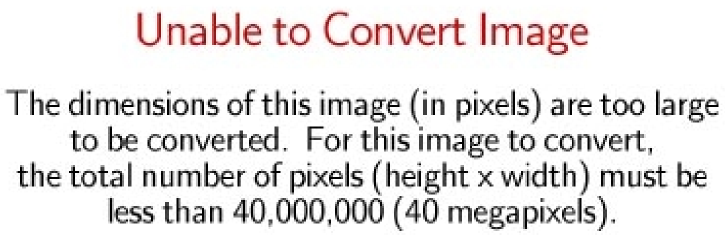


Fig. 2. Standard deviation (SD) and mean of SOC observations at the study sites (details are in Table 2).

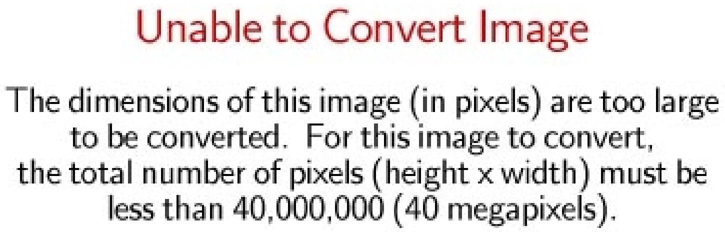
169x169mm (150 x 150 DPI)



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).

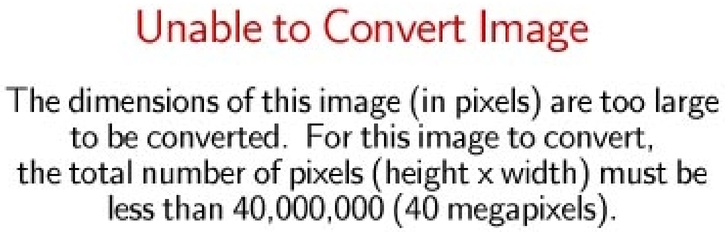
Fig. 3. Temporal changes of soil organic carbon (SOC, Mg C ha-1) observations (Observed, purple square) and simulations: blind (Blind, blue) simulations (26 models); three calibration scenarios, Generic (16 models, pink), Mixed (17 models, green) and Specific (17 models, grey) at all sites (as in Table 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).

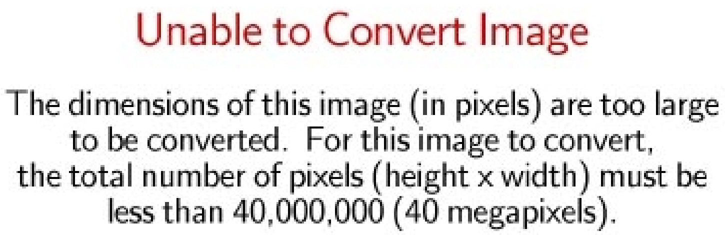
Fig. 4. Soil organic carbon (SOC, Mg C ha-1) at each site (as in Table 2), for blind simulations (Blind, (26 models), three calibration scenarios (Mixed, 17 models; Specific and Generic, 16 models) and observations (Observed). For each boxplot, black horizontal lines show the multi-model median. Boxes 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-1): 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-1): multi-model medians (MMM) from Specific scenario simulations (17 models as in Table 1) versus observations (coloured symbols represent sites as in Fig. 1).

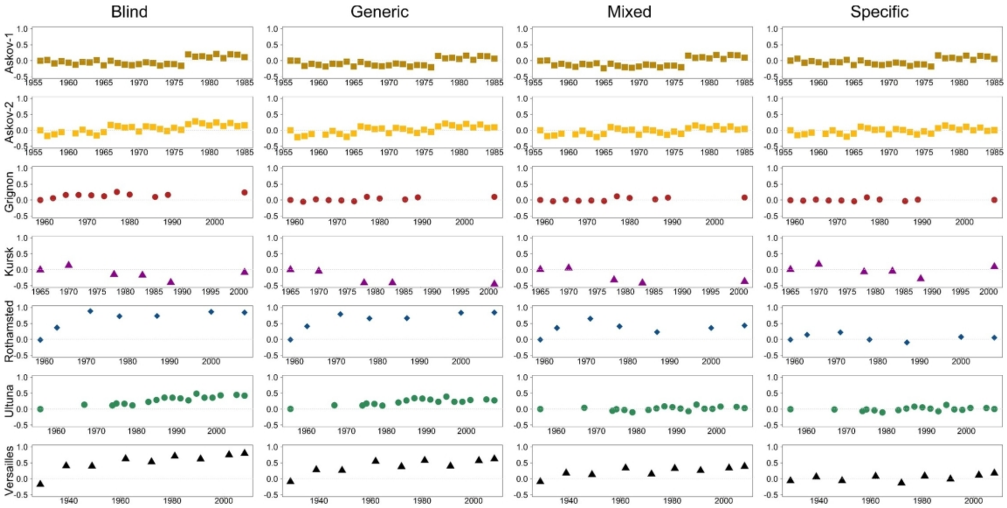


Fig. 7. Standardized model residuals ((MMM-O)/〖sd〗_obs) over time for blind (Blind) simulations and calibration scenarios Mixed, Specific and Generic at each site.

381x190mm (150 x 150 DPI)

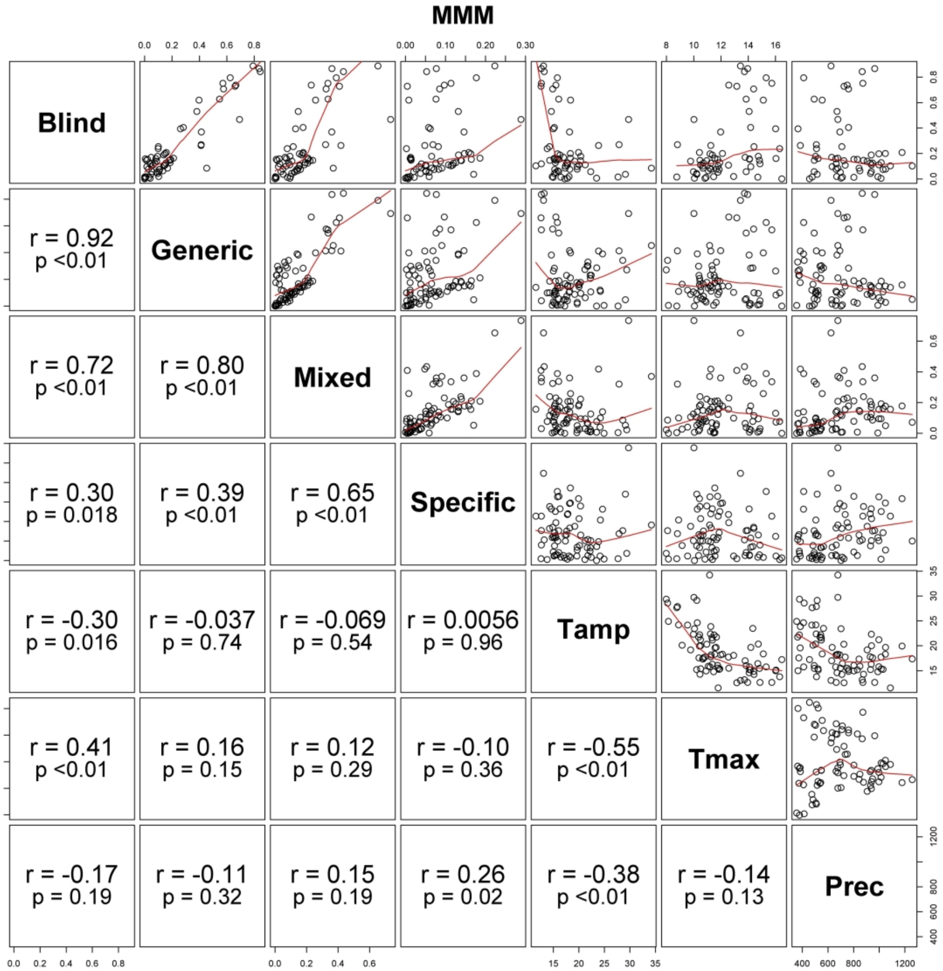


Fig. 8. Scatterplot correlation matrix of SOC (Mg C ha-1) model residuals of multi-model medians (MMM) for blind simulations (Blind) and calibrations scenarios (Generic, Mixed and Specific as in Table 3), and the annual climate metrics maximum temperature (Tmax), mean temperature amplitude (Tamp) and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

304x304mm (300 x 300 DPI)

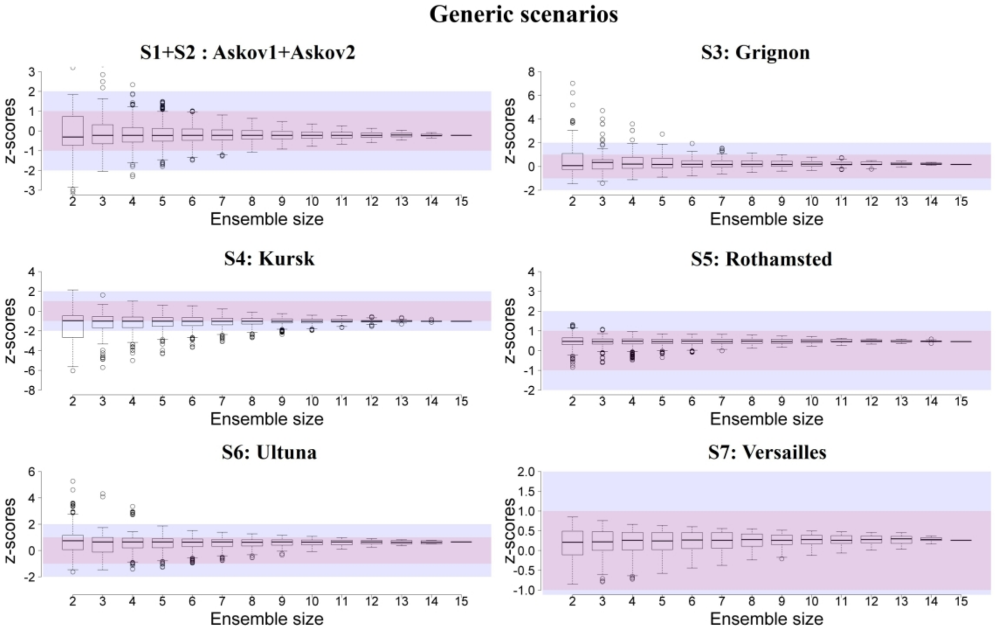


Fig. 9. z-scores calculated with different ensemble sizes for SOC estimates obtained with Generic scenario at different experimental sites. Black lines show median values. Boxes delimit the 25th and 75th percentiles. Whiskers are 10th and 90th percentiles. Circles indicate outliers. Coloured bands mark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

677x438mm (150 x 150 DPI)

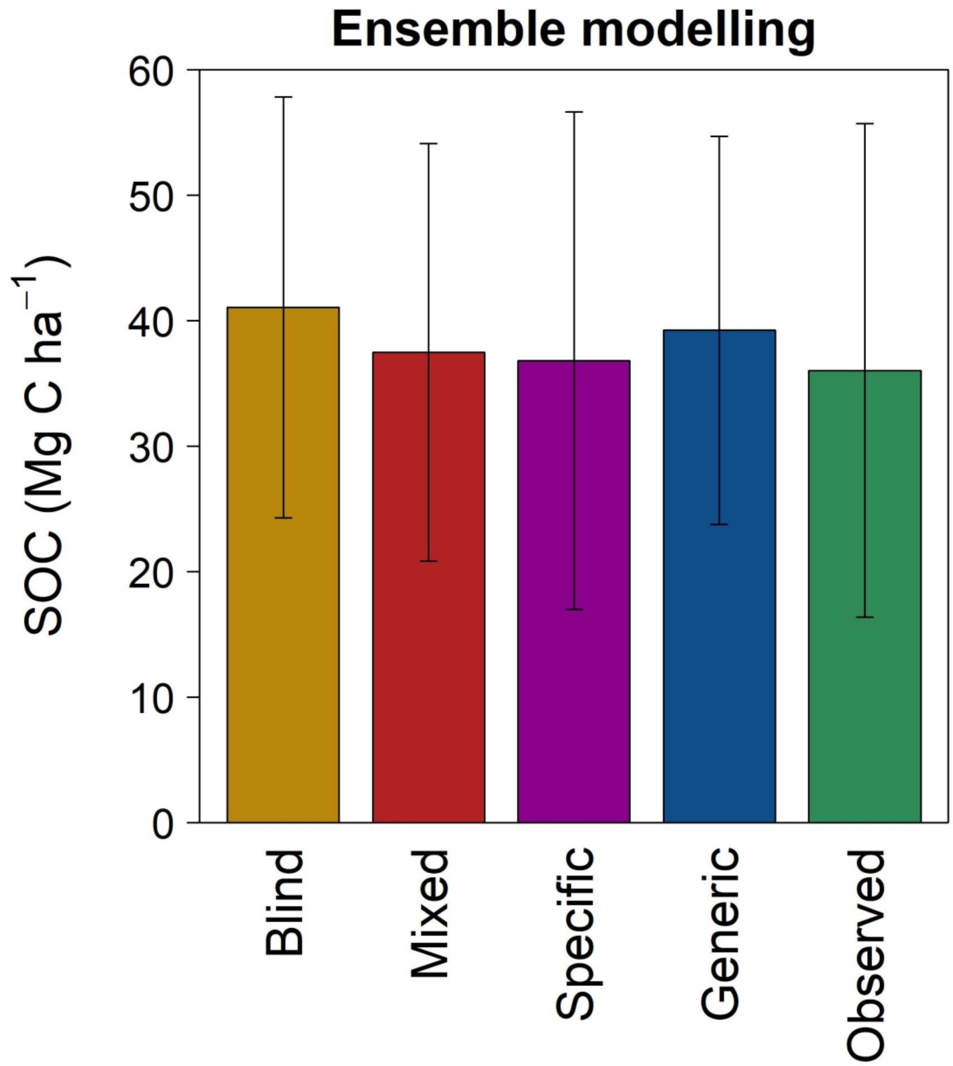


Fig. 10. Multi-site averages (vertical bars) and standard deviations (vertical lines) of observed and estimated (ensemble multi-model median) values of SOC (Mg C ha-1) in the last year of the experimental period. The ensemble modelling was applied with blind simulations (Blind) and calibration scenarios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

152x169mm (300 x 300 DPI)