Ensemble modelling of carbon fluxes in grasslands and croplands

Renáta Sándor[1,2], Fiona Ehrhardt[3], Peter Grace[4], Sylvie Recous[5], Pete Smith[6], Val Snow[7], Jean-François Soussana[3], Bruno Basso[8], Arti Bhatia[9], Lorenzo Brilli[10,11], Jordi Doltra[12], Christopher D. Dorich[13], Luca Doro[14,15], Nuala Fitton[6], Brian Grant[16], Matthew Tom Harrison[17], Ute Skiba[18], Miko U.F. Kirschbaum[19], Katja Klumpp[1], Patricia Laville[20], Joel Léonard[21], Raphaël Martin[1], Raia Silvia Massad[20], Andrew Moore[22], Vasileios Myrgiotis[23], Elizabeth Pattey[16], Zhang Qing[24], Susanne Rolinski[25], Joanna Sharp[26], Ward Smith[16], Lianhai Wu[27], Gianni Bellocchi[1]

[1] UCA, INRAE, VetAgro Sup, Unité Mixte de Recherche sur Écosystème Prairial (UREP), 63000 Clermont-Ferrand, France

[2] Agricultural Institute, CAR HAS, 2462 Martonvásár, Hungary

[3] INRAE, CODIR, 75007 Paris, France

[4] Queensland University of Technology, Brisbane, Australia

[5] Université de Reims Champagne Ardenne, INRA, FARE, 51100 Reims, France

[6] Institute of Biological and Environmental Sciences, University of Aberdeen, UK

[7] AgResearch - Lincoln Research Centre, Private Bag 4749, Christchurch 8140, New Zealand

[8] Dept. Geological Sciences, Michigan State University, East Lansing MI, USA

[9] Indian Agricultural Research Institute, New Delhi, India

[10] University of Florence, DISPAA, 50144 Florence, Italy

[11] IBIMET-CNR, 50145, Florence, Italy

[12] Institute of Agrifood Research and Technology (IRTA-Mas Badia), La Tallada d'Empordà, Catalonia, Spain

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

[14] Desertification Research Group, University of Sassari, Sassari, Italy

[15]Texas A&M AgriLife Research, Blackland Research and Extension Center, Temple TX, USA
[16]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada
[17]Tasmanian Institute of Agriculture, 16-20 Mooreville Rd, Burnie, Tasmania 7320, Australia
[18]Centre for Ecology and Hydrology, Bush Estate, Penicuik, EH34 5DR, UK
[19]Landcare Research-Manaaki Whenua, Palmerston North, New Zealand
[20]AgroParisTech, INRA, ECOSYS, 78850 Thiverval-Grignon, France
[21]INRAE, AgroImpact, 02000 Barenton-Bugny, France
[22]CSIRO, Agriculture Flagship**,** Black Mountain Laboratories, Canberra, Australia
[23]School of Geosciences, The University of Edinburgh, UK
[24]LAPC, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
[25]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany
[26]New Zealand Institute for Plant and Food Research, Christchurch, New Zealand
[27]Sustainable Agriculture Systems, Rothamsted Research, North Wyke, Devon, UK

Abstract

Croplands and grasslands are agricultural systems that contribute to land–atmosphere exchanges of carbon (C). We evaluated and compared gross primary production (GPP), ecosystem respiration (RECO), net ecosystem exchange (NEE=RECO-GPP) of CO_2, and two derived outputs - C use efficiency (CUE=-NEE/GPP) and C emission intensity (Int_C= -NEE/Offtake [grazed or harvested biomass]). The outputs came from 23 models (11 crop-specific, eight grassland-specific, and four models covering both systems) at three cropping sites over several rotations with spring and winter cereals, soybean and rapeseed in Canada, France and India, and two temperate permanent grasslands in France and the United Kingdom. The models were run independently over multi-year simulation periods in five stages (S), either blind with no calibration and initialization data (S1), using historical management and climate for initialization (S2), calibrated against plant data (S3), plant and soil data together (S4), or with the addition of C and N fluxes (S5). Here, we provide a framework to address methodological uncertainties and contextualize results. Most of the models overestimated or underestimated the C fluxes observed during the growing seasons (or the whole years for grasslands), with substantial differences between models. For each simulated variable, changes in the multi-model median (MMM) from S1 to S5 was used as a descriptor of the ensemble performance. Overall, the greatest improvements (MMM approaching the mean of observations) were achieved at S3 or higher calibration stages. For instance, grassland GPP MMM was equal to 1632 g C m^{-2} yr^{-1} (S5) while the observed mean was equal to 1763 m^{-2} yr^{-1} (average for two sites). Nash-Sutcliffe modelling efficiency coefficients indicate that MMM outperformed individual models in 91.4% of cases (S3 and S5). Our study suggests a cautious use of large-scale, multi-model ensembles to estimate C fluxes in agricultural sites if some site-specific plant and soil observations are available for model calibration. The further development of crop/grassland ensemble modelling will hinge upon the interpretation of results in light of the way models represent the processes underlying C fluxes in complex agricultural systems (grassland and crop rotations including fallow periods).

Keywords: C fluxes; Croplands; Grasslands; Multi-model ensemble; Multi-model median (MMM)

1. Introduction

The global emissions of CO_2 in the atmosphere continue to increase together with impacts on climate (IPCC, 2013). With the global carbon (C) balance becoming an issue of great societal concern, process-based models are increasingly used to simulate biogeochemical processes (such as plant photosynthesis and ecosystem respiration) occurring in both natural and managed ecosystems, including agricultural systems (e.g. Brilli et al., 2017). These models use approaches that determine the allocation of C from atmospheric CO_2 into plant biomass down to the soil organic matter (van Oijen et al., 2014; Grosz et al., 2017; Kuhnert et al., 2017). Process-based crop and grassland models (hereafter 'models') are important tools in agricultural and environmental research to extrapolate local observations in time and space, and to assess the impact of climate and agricultural practices on the functioning of soil-plant-atmosphere systems (e.g. Jones et al., 2017a). They are largely used to represent current understanding of the impacts of soil physical conditions such as soil temperature and water content on soil processes such as net mineralisation and to estimate harvested phytomass (which is the output of major significance in agricultural production). Climate-change impact assessment studies have been conducted (at different places and scales) by forcing models with global-to-local scale projected climate data (e.g. Ludwig and Asseng, 2006; Tingem et al., 2008; Ruiz-Ramos and Mínguez, 2010; Graux et al., 2013; Vital et al., 2013; Zhang et al., 2017; Mangani et al., 2018), to determine the vulnerability of agricultural systems to a changing climate (e.g. Harrison et al., 2014, Lardy et al., 2014; Eza et al., 2015; Mangani et al., 2019). Extensively tested biogeochemical models (with sub-models describing C cycling, generally coupled to N cycling) are recognised as effective tools for studying the magnitude and spatial-temporal patterns of C fluxes (Chang et al., 2015; Ma et al., 2015). They also play a prominent role in testing the effect of specific changes in management, plant properties or environmental factors (e.g. Kirschbaum et al., 2017), and for designing policies specific to the soil, climate, and agricultural conditions of a location or region (e.g. Stocker et al., 2013). However, outputs from different crop/grassland models often differ (e.g. Palosuo et al., 2011; Sándor et al., 2016), thus leaving users with the question of deciding which model(s) they should use, and under which circumstances presenting a range of possible impacts and adaptation responses. This has led to a call for benchmarking actions at international level (Rosenzweig et al., 2013; Soussana et al., 2015), where an estimation of the uncertainties associated with models is done by running several models for the same system (ensemble modelling, e.g. Ehrhardt et al., 2018), which generate envelopes of uncertainty, and help to identify avenues for model improvement (Jones et al., 2017b; Challinor et al., 2018). Model inter-comparisons have been conducted using datasets collected worldwide, with the involvement of different modelling communities and the use of alternative simulation models (e.g. Martre et al., 2015; Sándor et al., 2017; Ehrhardt et al., 2018). These studies indicate that there are substantial differences between models. Many of the uncertainties regarding the simulation of crop and grassland processes can be attributed to differences in the structure of these models (Brilli et al., 2017). While there has been a range of published studies showing ensemble model simulation results for agricultural yield (e.g. Asseng et al., 2013; Bassu et al., 2014; Li et al., 2015), there are fewer studies targeting C dynamics (e.g. Smith et al., 1997; Kirschbaum et al., 2015; Basso et al., 2018; Puche et al., 2019), and we are not aware of any published model intercomparison specifically assessing C fluxes with multiple models across a range of different experimental sites. In this study, we extended the analysis of the ensemble modelling performed by Ehrhardt et al. (2018) on agricultural production and N_2O emissions via a multi-stage modelling protocol (from blind simulations to partial and full calibration) by including a focus on C fluxes. We used a set of 23 biogeochemical models (11 cropland and eight grassland models, plus four models simulating both crops and grasslands) and compared simulations with experimental data from five sites (three crop rotations with spring and winter cereals, soybean and rapeseed, and two temperate grasslands). Comparisons included gross primary production (GPP), ecosystem respiration (RECO), the carbon balance represented by net ecosystem exchange (NEE<0 indicating net C uptake by the system) and other derived outputs. The models were calibrated through different stages with access to different levels of site-specific information. They were evaluated as a multi-model ensemble, with the aim of quantifying model uncertainties in the simulation of C fluxes at different sites and with different land uses.

2. Materials and methods

2.1. Experimental sites and C measurements

Observational data were available from two long-term, grazed experimental grasslands and three cropland sites, covering a variety of pedo-climatic conditions and agricultural practices from Canada, France (two sites), India and United Kingdom (Table 1). For consistency, we have maintained the site identifiers from Ehrhardt et al. (2018).

Table 1. Crop and grassland sites for the modelling exercise, years of available data and evaluated variables. GPP: gross primary production; RECO: ecosystem respiration; NEE: net ecosystem exchange.

Sites, country (latitude, longitude, elevation)	**Years of available data**	**Evaluated variables**	**References**
C1: Ottawa, Canada (45.29, -75.77, 94 m a.s.l.)	2007-2012	GPP, RECO, NEE	Pattey et al. (2006); Jégo et al. (2012); Sansoulet et al. (2014)
C2: Grignon, France (48.85, 1.95, 125 m a.s.l.)	2008-2012	GPP, RECO, NEE	Laville et al. (2011); Loubet et al. (2011)
C3: Dehli, India (28.6, 78.22, 233 m a.s.l.)	2006-2009	RECO	Bhatia et al. (2012)
G3: Laqueuille, France (45.64, 2.74, 1040 m a.s.l.)	2003-2012	GPP, RECO, NEE	Allard et al. (2007); Klumpp et al. (2011)
G4: Easter Bush, United Kingdom (55.52, -3.33, 190 m a.s.l.)	2002-2010	GPP, RECO, NEE	Skiba et al. (2013), Jones et al. (2017c)

Cropland sites used different crop rotations (Table 2), including cereals (spring and winter wheat, triticale, maize and rice), legumes (soybean), rapeseeds (canola and mustard) and borages (phacelia). C-flux data were also observed and simulated for fallow periods, to better understand C fluxes due to ongoing soil processes and the decomposition of crop residues (e.g. Xiao et al., 2015), as well as the role of weeds in cultivated fields (e.g. Curtin et al., 2000).

Table 2. Details about crop rotations in each cropland site (as in Table 1).

Site	Crop	Sowing date	Harvesting date / end of crop	Length of the growing season (days)	Number of daily measurements (days)
C1	Spring wheat	2007-05-19	2007-09-04	109	109
	Soybean	2008-06-10	2008-10-15	128	128
	Rapeseed (canola)	2009-04-24	2009-09-08	138	138
	Maize	2010-05-12	2010-11-15	188	188
	Spring wheat	2011-05-10	2011-08-29	112	112
	Rapeseed (canola)	2012-05-15	2012-09-19	128	128
C2	Rapeseed (mustard)	2008-01-01	2008-04-14	104	104
	Maize	2008-04-27	2008-09-25	152	152
	Winter wheat	2008-10-17	2009-07-31	288	288
	Triticale	2009-10-13	2010-07-19	280	280
	Phacelia	2010-09-13	2011-04-19	219	219
	Maize	2011-04-20	2011-09-06	140	140
	Winter wheat	2011-10-18	2012-08-03	290	260
	Rapeseed (canola)	2012-10-25	2012-12-31	68	GPP:68; RECO: 66
C3	Winter wheat	2006-11-29	2007-04-13	136	32
	Rice	2007-07-14	2007-10-15	94	17
	Winter wheat	2007-12-01	2008-04-16	137	34
	Rice	2008-07-25	2008-10-22	90	3
	Winter wheat	2008-11-25	2009-04-22	149	0

These sites provided high quality, previously published data encompassing climate, soil, agricultural practices, and C and N fluxes. They were either equipped with an eddy covariance system to determine the net ecosystem exchange (NEE) of CO_2 or with closed chambers for measuring respiration fluxes and automated weather stations for recording climatic conditions. The NEE data were partitioned into two main fluxes: gross primary production (GPP), which is the photosynthetic plant production from atmospheric CO_2, and ecosystem respiration

(RECO), which is the total C respired by plants, soil organisms and (in the case of grasslands) grazing animals.

C-flux data were made available on a daily basis, for each day of year in grassland sites and for a varying number of days in crop sites (Table 2).

2.2. Models and simulation study

The 23 models (Table 3) and the model codes and outputs provided (Table 4) encompass all but one of the 24 biogeochemical models described in Ehrhardt et al. (2018). These models vary in their complexity (number of parameters, type of inputs and outputs) and in their constitutive processes (Ehrhardt et al., 2018, appendices S1 and S2). Model anonymity was maintained throughout the paper. The identities of models were kept anonymous by using model codes from M01 to M24 (the order of models being not identical with the one used in Table 3). Model M11 is not included here because it did not provide access to C-flux outputs. Modelling groups from 11 countries (Australia, Canada, China, France, Germany, India, Italy, New Zealand, Spain, United Kingdom and United States of America) were involved. Models were initialized and calibrated against vegetation, soil and atmospheric fluxes from the study sites as described in Ehrhardt et al (2018). During this exercise, modellers were given access to gradually more detailed data to run and evaluate their models (from uncalibrated to fully calibrated simulations), using a multi-stage protocol described in Ehrhardt et al. (2018). In short, model evaluation followed five ascending calibration levels including the use of: (S1) no data apart from site weather and management data for the simulation periods, i.e. a blind test without model calibration and initialization; (S2) additional historical climate and management data (for years preceding simulation periods for initialization purposes) and regional productivity; (S3) biomass production and phenology data; (S4) soil temperature, moisture and mineral N data;

(S5) N_2O emission and soil organic C and N flux data (the full suite of measurements taken at respective sites).

Nineteen models took part in stage S1. One of these stopped providing outputs at S2 and a second at S4. Four models entered the exercise at S2, and received feedback from these results and continued providing outputs until S5. Three modelling teams (M14, M16, M23) used automatic or semi-automatic techniques to calibrate the model parameters (i.e. Bayesian calibration, the Latin Hypercube Sampling method and a mixed manual/automatic method) while the others used a manual, informed *ad-hoc* approach.

Table 3. The 23 biogeochemical models used in the model intercomparison study.

Simulated system	Model name	Availability
Cropland	Agro-C v.1.0	On request to Yao Huang (huangy@mail.iap.ac.cn)
	APSIM v.7.5	http://www.apsim.info
	APSIM v.7.6	http://www.apsim.info
	CERES-EGC	https://www6.versailles-grignon.inra.fr/ecosys/Productions/Logiciels-Modeles/CERES-EGC
	DailyDayCent	On request to Brian Grant (Brian.Grant@canada.ca)
	DNDC	http://www.dndc.sr.unh.edu
	EPIC 810	http://epicapex.tamu.edu/model-executables
	FASSET v2.5	http://www.fasset.dk
	Infocrop	http://www.iari.res.in/?option=com_content&view=article&id=1334
	SALUS	On request to Bruno Basso (basso@msu.edu)
	STICS v.8.2	http://www6.paca.inra.fr/stics_eng
Grassland	APSIM-GRAZPLAN	http://www.apsim.info
	APSIM-SoilWater	http://www.apsim.info
	APSIM-SWIM v.7.7	http://www.apsim.info
	CenW v. 4.1	http://www.kirschbaum.id.au/Welcome_Page.htm
	DairyMod Ecomod v.5.3.1	http://www.imj.com.au/dm
	LPJmL v.3.5.3	https://www.pik-potsdam.de/research/projects/activities/biosphere-water-modelling/lpjml
	PaSim	https://www1.clermont.inra.fr/urep/modeles/pasim.htm
	SPACSYS v. 5.2	https://www.rothamsted.ac.uk/rothamsted-spacsys-model
	DayCent v4.5 2006[1]	http://www.nrel.colostate.edu/projects/daycent-downloads.html

Cropland and grassland	Daily DayCent 4.5 2010[1]	http://www.nrel.colostate.edu/projects/daycent-downloads.html
	DayCent v4.5 2013[1]	http://www.nrel.colostate.edu/projects/daycent-downloads.html
	Landscape DNDC v0.9.2	Under licence agreement with Institute of Meteorology and Climate Research, Germany (http://www.imk.kit.edu)

[1] Different versions of the model result in different parameter settings and a few variations in the model structure (Sándor et al., 2018): DayCent v4.5 2006 applies grazing on a daily basis as linear impact on aboveground biomass and root/shoot ratio, with aboveground biomass removed as a percentage of total aboveground biomass; DayCent v4.5 2010 and 2013 apply grazing on a daily basis with aboveground biomass removed as a percentage of total aboveground biomass rather than as continuous grazing.

Table 4. C-flux outputs (as in Table 1) provided by different models.

Model type	**Model code**	**Outputs**			**Calibration method[1]**
		GPP	**RECO**	**NEE**	
Crop models	M01	✓	✓	✓	Manual
	M02	NA	✓	NA	Manual
	M04	NA	✓	NA	Manual
	M09	✓	✓	✓	Manual
	M12	NA	✓	NA	Manual
	M13	NA	✓	NA	Manual
	M18	NA	✓	NA	Manual
	M19	✓	✓	✓	Manual
	M20	NA	NA	✓	Manual
	M25	NA	✓	NA	Manual
	M26	NA	✓	NA	Manual
Grassland models	M03	NA	✓	NA	Manual
	M06	✓	✓	✓	Manual
	M16	✓	✓	✓	Automatic
	M21	✓	✓	✓	Manual
	M22	✓	✓	✓	Manual
	M23	✓	✓	✓	Manual/ Automatic
	M24	✓	✓	✓	Manual
	M28	✓	✓	✓	Manual
Both systems	M05	✓	✓	✓	Manual

	M07	✓	✓	✓	Manual
	M08	NA	✓	NA	Manual
	M14	✓	✓	✓	Automatic

[1] With automatic methods, all the parameters were recalibrated at each calibration stage; with the manual methods, previously calibrated parameters were carried forward into the next calibration stage.

2.3. Data analysis

Three independent modelled C fluxes (GPP, RECO, NEE) were compared against observed values at each calibration stage. Modelled and measured outputs at the dates when measurements were made were aggregated and analysed by calendar year for grasslands (g C m^{-2} yr^{-1}), and by growing season (from sowing to harvest) for crops (g C m^{-2} $season^{-1}$). Fluxes from fallow periods (from harvest of one crop to the time of planting of the next crop) were considered separately. For crop rotations, data were aggregated by crop season, not by calendar year. To ensure consistency of results among the growing periods of different crops, a daily-based seasonal extrapolation of C fluxes ($C_{am(s)}$, g C m^{-2} $season^{-1}$) was obtained as a function of the number of measuring days in crop seasons (n_{meas}) and the length (number of days) of crop growing seasons (ns) as in Table 2:

$$C_{am(s)} = \frac{\sum_{i=1}^{n_{meas}} C_{am(d)}}{n_{meas}} \cdot n_s$$

where $C_{am(d)}$ is the daily amount of assimilated or emitted C (g C m^{-2} d^{-1}).

Two derived output variables were also analysed on seasonal basis for crops and on annual basis for grasslands, one representing C emission intensity and one C use efficiency. The potential to sustain or even increase crop/grassland yields is a desirable characteristic of any mitigation option both in terms of adoption of the technology by farmers (Vellinga et al., 2011) and its benefit in reducing GHG emissions per area of land and per unit of product, which is referred to as 'emission intensity' (van Groenigen et al. 2010). In this study, C emission intensity (Int_C) was calculated as the ratio between the amount of C emitted as CO_2 (C) and the total amount of C in harvested agricultural production, that is, grain yield for crops and the offtake (annual sum of animal intake and harvested aboveground biomass) for grasslands (after Ehrhardt et al., 2018). Carbon use efficiency (CUE) was obtained as the ratio between CO_2-C exchanged by the ecosystem and GPP (-NEE/GPP). A synthetic indicator such as the CUE is useful to inform about the ability to retain part of GPP and thus increase total C content in the agro-ecosystem (Sándor et al., 2016). The outputs analysed on seasonal/annual bases were also presented on a daily basis as a practical way to compare models across contrasting locations.

We documented the variability of the multi-model simulation exercise across different calibration stages, while inspecting how multi-model median (MMM) converged to the mean of observations. For each simulated variable, 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). MMM is the median value of simulated data, which was calculated on daily outputs for each stage. The advantage of using MMM was established on a theoretical basis and in practical studies in crop and grassland modelling (Wallach et al., 2018). The absolute bias (best, $0 \leq \text{ABIAS} < \infty$, worst) was calculated as an average of the absolute differences between MMM estimates and means of observations at each season or year. Scatterplots of simulated versus observed daily data and the modelling efficiency ($-\infty < \text{EF} \leq 1$, positive values indicating that model estimates are more accurate than the mean of the observed data; Nash and Sutcliffe, 1970) were also provided to compare individual models and the MMM. Then, to explore how MMM varied with the number of models in the ensemble we performed a calculation for each *z*-score transformed MMM, $z = \frac{MMM-\bar{O}}{sd_{obs}}$, obtained by dividing the multi-model data deviation from the mean of observations ($\bar{O}$) by the standard deviation of observations (sd_{obs}) (after Ehrhardt et al., 2018). We calculated *z*-scores on all possible combinations of sets of *k* out of *n*=15 models (*k*=2, … *n*). The minimum number of models providing plausible estimates at each site was that for which *z*-scores were comprised between -2 and +2 (approximating the 95% confidence limit of a normal distribution).

R software (https://cran.r-project.org) was used for statistical analysis and graphical visualization.

3. Results

The overall results are presented and discussed, with selected graphs, for grassland and cropland sites. At cropland sites, simulated C fluxes are also analysed for each individual crop. In this way, we addressed the models' ability to simulate different crops and environmental situations (beyond assessing C fluxes at different sites), where the ability to model C fluxes from one crop may not be the same as for another crop. Results from similar short cereals (triticale, winter and spring wheat) are grouped. Fallow C fluxes are associated with C fluxes from field crops because they cover their off-growing season period (i.e. between the harvest of one crop and the sowing of the next crop in a rotation).

3.1. Uncertainties and ensemble performance by land use

Fig. A in the Supplementary material and Table 5 show the multi-model uncertainties (spread of responses with different models) under different land uses (fallow, crop and grassland). Observed mean RECO varied between 32 g C m^{-2} yr^{-1} (fallow) and 1561 g C m^{-2} yr^{-1} (grassland) considering all calibration stages. The latter value is about three times higher than seasonal observed crop values, e.g. maize (674 g C m^{-2} season^{-1}) or triticale (553 g C m^{-2} season^{-1}), as in Table 5. Also, there is considerable difference between observed means and MMM RECO values, e.g. for S5, 1561 vs. 1123 g C m^{-2} yr^{-1} for grasslands, 674 vs. 375 g C m^{-2} season^{-1} for maize, 420 vs. 320 g C m^{-2} season^{-1} for spring wheat and 606 vs. 275 g C m^{-2} season^{-1} for soybean. Overall, observed RECO was underestimated by the MMM in all stages and land uses. The GPP MMM also showed high variability with winter wheat, triticale and maize (ranging from 745 to 1354 g CO_2-C m^{-2} $season^{-1}$ at S5 and S3, respectively), comparable with the variability of grasslands across calibration stages (1061-1568 g C m^{-2} yr^{-1}). There was also high variability in estimated NEE with winter cereals and maize (Fig. A in the Supplementary material and Table 5), but MMM generally approached observation means, e.g. maize mean observation and MMM were -539 (S3) and -544 g C m^{-2} $season^{-1}$ (S5), respectively. Model estimates for grasslands showed less variability in NEE predictions (from -157 at S5 to -99 at S1 compared to the observed mean of -219 g C m^{-2} yr^{-1}). Seasonal CUE values (presented on different scales for fallow, crop and grassland systems in Fig. A in the Supplementary material and Table 5) were generally positive, with the exception of phacelia. Models tend to show higher uncertainties towards negative values at early calibration stages, e.g. S1 of winter cereals, maize, phacelia and rice. A lower uncertainty is associated with Int_C values, mainly with grasslands. Some GPP (and Int_C) predictions were different from zero event under fallow conditions (some non-zero biomass production was also observed experimentally).

The absolute bias (ABIAS), calculated by comparing the MMM and observed mean of different output variables for different land uses, showed that we can expect an improvement of model performances after S3, when vegetation and yield data were provided for calibration (Fig. 1). For instance, GPP of maize, and RECO, GPP and NEE of spring wheat simulations show the best fit at S3, while triticale and winter wheat show greater improvement at S4 and S5.

3.1.1. Grassland systems

There was considerable variability in the simulated and observed GPP and RECO values (Fig. A in the Supplementary material and Table 5). On average, the annual mean of observed GPP values was 1763 g C m^{-2} yr^{-1}, but simulations underestimated it because MMM ranged from

1062 (S1) to 1568 (S5) g C m^{-2} yr^{-1}. Overall, RECO predictions had a wider range in grasslands than in crops (Fig. A in the Supplementary material and Table 5). Similar to GPP, models mostly underestimated mean of RECO (1561 g C m^{-2} yr^{-1}), as predictions varied from 969 (S2 MMM) to 1248 (S1 MMM) g C m^{-2} yr^{-1} (the latter was similar to S3 MMM=1235 g C m^{-2} yr^{-1}). On the other hand, NEE and Int$_{C}$ values were well estimated with MMM values lying within the range of observations (-610 to 66 g C m^{-2} yr^{-1} and -0.18 to 2.54 yr^{-1}, respectively). In addition, Int$_{C}$ was near zero in grasslands. The models tended to underestimate CUE and to slightly overestimate NEE. Best estimates (least difference between MMM and observation mean) were obtained at S5 for both: NEE: -157.4 versus -218.9 g C m^{-2} yr^{-1}; CUE: 0.07 yr^{-1} versus 0.11yr^{-1}.

3.1.2. Arable crops

The RECO MMM predictions varied between 58 (fallow, S1) and 512 (maize, S3) g C m^{-2} season^{-1} for the various crops (Fig. A in the Supplementary material and Table 5). The ABIAS values slightly reduced after S3 (Fig. 1). In general, there was high variability in RECO and GPP predictions, especially under maize, soybean and rice. On average, crops showed negative NEE predictions, with the exception of fallow and phacelia, which showed net C emission (NEE>0). Overall, CUE predictions and observations had similar patterns.

Table 5. Multi-model median values of ecosystem respiration (RECO), gross primary production (GPP), net ecosystem exchange (NEE), carbon use efficiency (CUE) and C intensity ($\mathrm{Int_C}$), calculated over multiple years at crop and grassland sites for two calibration stages (S3 and S5) and the observations (Obs).

Output / Land-use	**RECO**			**GPP**			**NEE**			**CUE**			**$\mathrm{Int_C}$**		
	S3	S5	Obs	S3	S5	Obs	S3	S5	Obs	S3	S5	Obs	S3	S5	Obs
Fallow	92.65	72.43	31.93	10.78	1.00	11.84	83.19	64.45	44.34	0.00	0.00	-3.81	0.00	0.00	0.00
Winter wheat	238.13	217.57	259.04	1353.82	745.43	1204.44	-561.93	-610.89	-622.59	0.48	0.55	0.52	0.91	0.88	0.75
Spring wheat	386.48	320.25	420.11	476.48	476.48	476.48	-62.42	-64.38	-56.37	0.16	0.23	0.05	0.13	0.15	0.08
Triticale	448.80	289.63	553.35	1517.53	1107.80	1107.80	-563.15	-501.46	-554.46	0.38	0.50	0.50	0.84	0.89	0.83
Maize	511.93	374.71	674.35	1338.48	1166.02	1241.43	-538.85	-544.12	-567.08	0.41	0.44	0.40	0.53	0.60	0.62
Soybean	287.96	274.96	605.62	453.97	453.97	753.79	-11.84	-37.48	-148.17	0.00	0.00	0.20	0.21	0.26	0.77
Rapeseed	199.61	164.45	296.25	255.78	256.10	450.97	-80.34	-81.93	-171.93	0.12	0.12	0.32	0.46	0.46	0.68
Phacelia	296.39	234.00	326.83	193.10	228.62	228.62	154.87	73.03	98.20	-0.43	0.00	-0.43	0.00	0.00	0.00
Rice	110.08	100.20	34.34	933.61	606.26	NA	-505.50	-405.69	NA	0.56	0.66	NA	1.12	1.02	NA
Grassland	1234.75	1123.23	1561.14	1456.27	1568.14	1762.74	-102.97	-157.40	-218.84	0.02	0.07	0.11	0.23	0.40	0.49

With wheat and triticale, simulations were lower than the measured RECO, whose means were about 259 g C m^{-2} season^{-1} for winter wheat, 420 g C m^{-2} season^{-1} for spring wheat and 553 g C m^{-2} season^{-1} for triticale (Fig. A in the Supplementary material and Table 5). Model performances improved after S3, with MMM of 238 for winter wheat, 386 for spring wheat and 449 g C m^{-2} season^{-1} for triticale. All three cereal crops had negative NEE values, especially under winter wheat and triticale. CUE was slightly overestimated under winter wheat. In contrast, CUE MMM of spring wheat and triticale were within the range of observed CUEs at higher calibration levels. CUE and Int$_C$ showed a similar pattern of model variability.

In rice, only the RECO values were measured, thus only simulated data are available for the other C outputs. The models tended to overestimate RECO in rice, where the observations ranged between 30 and 39 g C m^{-2} season^{-1}, whilst the S5 estimation was roughly three times higher (100 g C m^{-2} season^{-1}) (Fig. A in the Supplementary material and Table 5). Overall, there was high variability in model predictions for all calibration stages, but rice showed the greatest variability in GPP predictions. This was mostly evident at S4, when soil properties were included with plant measurements to perform calibration. The variability of NEE, CUE and Int$_C$, however, was similar to that of other crops.

All the models underestimated observed maize seasonal RECO and GPP values (661-1070 and 1102-1671 g C m^{-2} season^{-1}, respectively), but model variability was limited for NEE, CUE and Int$_C$ (Table 5 and Fig. A in the Supplementary material). Fig. 1 shows a complex pattern of ABIAS values, which were generally high at all calibration stages for RECO and GPP and even increased at S4 for GPP and Int$_C$, while simulations and observations were closer for NEE and CUE.

Overall, rapeseed was characterized by high variability in the observations: RECO: 69-660 g C m^{-2} season^{-1}, GPP: 59-930 g C m^{-2} season $^{-1}$, CUE: -0.73-0.57 season $^{-1}$, Int$_C$: 0-1.5 season $^{-1}$ (Fig. A in the Supplementary material). The models tended to underestimate RECO and GPP observations, and to overestimate NEE and CUE. Net C emissions were predicted against the net C uptake reflected in measurements. Variations of simulated MMM $\mathrm{Int_C}$ values were within the range of observations in spite of their high variability.

After maize and triticale, the simulations of soybean exhibited the highest variability within the investigated crops on seasonal aggregation (Fig. A in the Supplementary material, Table 5). RECO (606 g C m^{-2} $season^{-1}$) and GPP (754 g C m^{-2} $season^{-1}$) values were underestimated, with high model variability (Table 5), but ABIAS tended to decrease after S2 (Fig. 1). For NEE and CUE, observations and predictions were closer to each other, but there were large differences between observed and predicted NEE.

Among crops, phacelia showed the lowest uncertainty of RECO, GPP and NEE predictions. Simulated MMM RECO (234-296 g C m^{-2} $season^{-1}$) tended to underestimate the observed value (327 g C m^{-2} $season^{-1}$), in contrast to other outputs. With this crop, NEE was positive, which indicated net C emission. The observed mean was ~98 g C m^{-2} $season^{-1}$ and the MMM ranged between 35 (S5) to 209 (S1) g C m^{-2} $season^{-1}$.

With fallow, MMM RECO predictions were within the range of observations that ranged between 16 and 161 g C m^{-2} $season^{-1}$. Observed and simulated GPP values were close to zero and the simulations were within the range of variation of the measurements (5.8-31 g C m^{-2} $season^{-1}$). NEE values showed the second highest positive simulated and observed values after phacelia on a seasonal basis (MMM predictions were within the range of measurements: 15 and 130 g C m^{-2} $season^{-1}$, Fig. A in the Supplementary material). The observed CUE values were the lowest (Fig. A in the Supplementary material) while the ABIAS was the highest ($\mathrm{ABIAS_{CUE}}$=4.26 $season^{-1}$, Fig. 1). The observed variability, between -0.16 and -0.04 $season^{-1}$, was reflected in the model simulations.

3.2. Uncertainties and ensemble performance by site

Overall, RECO and GPP were underestimated at grassland sites (Fig. 2). Mean observed RECO was about 1650 g C m^{-2} yr^{-1} at G3 site and 1538 g C m^{-2} yr^{-1} at G4 site, while the MMM predictions varied from 716 to 1262 and from 1057 to 1457 g C m^{-2} yr^{-1}, respectively. Improvements were observed at S3, and best predictions were obtained at S5, especially at G4 site, e.g. 1457 g C m^{-2} yr^{-1} (S5 MMM) versus 1538 g C m^{-2} yr^{-1} (observed mean), Fig.2. For crop sites, we observed some considerable improvements after S2, e.g. with S3 showing the best estimates of RECO, where the MMM and observed mean were very similar 241 and 242 g C m^{-2} season^{-1} (average for C1, C2 and C3) (Fig. 2.).

3.2.1. C1

The mean of observed seasonal RECO (611 g C m^{-2} season^{-1}) was underestimated at all calibration stages, although there was an improvement after S3 (Fig. 2). The observed means of GPP (842 g C m^{-2} season^{-1}), CUE (0.21 season^{-1}) and Int$_C$ (0.74 season^{-1}) were well approached by the MMM predictions. The NEE values, which were lower than in C2 and C3, were generally underestimated. However, C fluxes excluded fallow periods, since data were not provided.

3.2.2. C2

Detailed C-flux data were available at this site for both cropped and fallow periods and showed large ranges of variability for all outputs. The MMM predictions were within these ranges. RECO and GPP were mostly overestimated (Fig. 2). The observed NEE (16 g C m^{-2} season^{-1})

of C2 was near zero. Model predictions tended to underestimate it but the simulations were still within the range of observations.

3.2.3. C3

Overall, C3 showed the lowest model variability. At this site, only RECO observations were available. The observed mean (42 g C m^{-2} season^{-1}) was overestimated, with the MMM ranging between 78 and 113 g C m^{-2} season^{-1}.

3.2.4. G3

GPP and RECO observations did not vary as much at this site as the model predictions. Fig. 2 shows that the accuracy of GPP predictions tended to increase through the calibration stages, with RECO showing best estimates at S3. For instance, at S5, MMM (1775 g C m^{-2} yr^{-1}) was close to the observed mean (1898 g C m^{-2} yr^{-1}). The MMM values of NEE, CUE and Int$_C$ showed slight differences for different calibration stages, but an improvement was observed at S5 (209 g C m^{-2} yr^{-1}, 0.11 yr^{-1}, 0.54 yr^{-1}, respectively) compared with the observations (-248 g C m^{-2} yr^{-1}, 0.13 yr^{-1}, 0.62 yr^{-1}, respectively). G3 showed a high C uptake (observed means NEE=-248 versus MMM at S5=-209 g C m^{-2} yr^{-1}).

3.2.5. G4

At this site, the ranges of variation of RECO and GPP observations were similar to G3. Observed GPP (1767 g C m^{-2} yr^{-1}) was generally underestimated by the models (ranging from 1255 to 1490 C m^{-2} yr^{-1}), but the MMM of RECO at S5 (1457 g C m^{-2} yr^{-1}) approached the mean of observations (1537 g C m^{-2} yr^{-1}). The MMM of NEE at S5 (-110 g C m^{-2} yr^{-1}) was also close to the observation mean (-148 g C m^{-2} yr^{-1}). For CUE, the positive values of both MMM (ranging from 0.03 to 0.13 yr^{-1}) and observation mean (0.12 yr^{-1}) reflected the C uptake at this grassland site.

3.3. Individual models versus multi-model ensemble

Daily comparisons were not straightforward in this study because discontinuous observations were tied to specific days, but the models did not have access to the diurnal pattern of the processes (e.g. timing of specific weather or management events). With this caveat in mind, for interpretation, we plotted simulated versus observed daily C fluxes as a visualisation tool to compare the model ensemble results with individual model results. The scatterplots of Figs. 3, 4 and 5 and Figs. B-J in the supplement, are examples for GPP, RECO and NEE at the S5 calibration stage for the G3 grassland site, of the comparison of the performances of individual models and MMM values. Consistent with the findings above, the MMM outperformed most of the individual models. Considering R^2 values and alignments with the 1:1 lines, this was the case for nine out of 10 models and seven out of 11 models simulating GPP and RECO. When, in a few cases, individual models provided relatively satisfactory results, this was generally true for one output but not for another. For example, M21 provided satisfactory results for GPP (Fig. 3) but not for RECO (Fig. 4). M16 (which was calibrated according to an automatic technique) was distinctly outperforming the MMM for both GPP (Fig. 3) and RECO (Fig. 4) estimates, but underperformed for other outputs, e.g. NEE (Fig. 5). Similar patterns of results were obtained at the S3 (Supplementary material, Figs. B-D) and other calibrations stages (data not shown), and for the G4 site (data not shown). Likewise for croplands, the MMM tended to outperform individual models, e.g. for GPP, RECO and NEE at C2 site (Figs. E to J in supplementary material for calibration stages S3 and S5).

Nash-Sutcliffe modelling efficiency coefficients (EF), calculated on daily data of GPP, RECO and NEE at the five sites for both S3 and S5 (Table 6), were not always positive with MMM (e.g. NEE at C1 and RECO at C3), but they indicate that MMM outperformed individual models in 215 out of 233 cases (that is, 92.3% of cases).

Table 6. Nash-Sutcliffe modelling efficiency (EF) values for C-flux outputs (as in Table 1) provided by different models (as in Table 4) at S3 and S5 calibration stages at cropland (C1, C2, C3) and grassland (G3, G4) sites. Grey cells indicate that output variables were neither measured nor simulated.

Model	Stage	Output	C1	C2	C3	G3	G4
M01	S3	RECO	-0.20	-0.29	-273.42		
		GPP	-0.72	0.22			
		NEE	-3.66	0.29			
	S5	RECO	-0.20	0.00	-273.42		
		GPP	-0.72	0.35			
		NEE	-3.67	0.36			
M02	S3	RECO	-0.04	-0.18	-9.87		
		GPP					
		NEE					
	S5	RECO	-0.42	-0.20	-14.69		
		GPP					
		NEE					
M03	S3	RECO					
		GPP					
		NEE					
	S5	RECO					
		GPP					
		NEE					
M04	S3	RECO	-1.39	-0.93	-1.26		
		GPP					
		NEE					
	S5	RECO	-1.39	-0.93	-1.38		
		GPP					
		NEE					
M05	S3	RECO	0.07	-1.30	-316.84	-0.44	-1.14
		GPP	0.52	-1.13		0.00	-0.65
		NEE	-0.07	-1.60		0.28	0.08
	S5	RECO	-0.34	-0.30	-17.97	-0.01	-0.79
		GPP	0.43	0.44		0.33	-0.25
		NEE	-0.02	0.48		0.11	0.30
M06	S3	RECO				-1.17	-0.41
		GPP				-0.38	0.16
		NEE				-0.38	0.24

	S5	RECO				-1.17	-0.31
		GPP				-0.38	0.22
		NEE				-0.38	0.29
M07	S3	RECO	-2.62	-3.96	-120.78	-0.47	0.09
		GPP	-0.69	-0.39		-0.47	-0.02
		NEE	-6.60	-21.28		-0.17	-0.07
	S5	RECO	-9.09	-3.80	-43.52	-0.88	-0.06
		GPP	-1.68	-17.22		-0.19	0.53
		NEE	-3.39	-21.02		-0.13	-0.66
M08	S3	RECO	-0.76	-0.83	-7.31	-1.39	-1.06
		GPP					
		NEE					
	S5	RECO	-0.75	-0.78	-15.17	-1.36	-1.01
		GPP					
		NEE					
M09	S3	RECO	0.02	-9.66	-193.83		
		GPP	-0.14	-1.07			
		NEE	-1.50	0.36			
	S5	RECO	-0.09	0.12	-260.92		
		GPP	-0.10	0.56			
		NEE	-1.33	0.32			
M12	S3	RECO	-0.57	-4.26	-21.31		
		GPP					
		NEE					
	S5	RECO	-0.73	-0.56	-13.32		
		GPP					
		NEE					
M13	S3	RECO	0.63	0.23	-9.05		
		GPP					
		NEE					
	S5	RECO	0.69	0.23	-9.05		
		GPP					
		NEE					
M14	S3	RECO	-3.25	-2.09	-2980.13	0.02	-1.36
		GPP	0.27	0.04		0.13	0.13
		NEE	-4.38	-1.04		-0.43	-1.85
	S5	RECO	-5.41	-3.60	-49.00	-0.47	-0.31
		GPP	0.50	0.08		-0.09	-0.06
		NEE	-6.37	-1.19		-0.10	-1.08
M16	S3	RECO				0.42	0.41
		GPP				0.20	0.26
		NEE				-0.73	-0.96
	S5	RECO				-0.11	0.41
		GPP				0.57	0.58
		NEE				-0.92	0.07
M18	S3	RECO	-0.01	-0.55	-34.56		
		GPP					
		NEE					
	S5	RECO	0.20	-0.72	-35.33		
		GPP					
		NEE					
M19	S3	RECO	-0.50	-2.25	-601.23		
		GPP	-0.93	-0.06			

		NEE	-1.43	-0.27			
	S5	RECO	-0.15	0.37	-100.46		
		GPP	0.11	0.47			
		NEE	0.15	0.26			
M20	S3	RECO					
		GPP					
		NEE	0.19	0.41			
	S5	RECO					
		GPP					
		NEE	0.34	0.49			
M21	S3	RECO				-0.52	-0.53
		GPP				0.07	0.48
		NEE				-3.66	-1.82
	S5	RECO				-0.55	-0.61
		GPP				0.18	0.39
		NEE				-4.26	-1.36
M22	S3	RECO				0.40	0.49
		GPP				0.16	0.58
		NEE				-0.49	0.27
	S5	RECO				0.40	0.49
		GPP				0.16	0.58
		NEE				-0.49	0.27
M23	S3	RECO				0.44	-0.61
		GPP				0.11	-0.21
		NEE				-0.76	-0.07
	S5	RECO				0.52	0.38
		GPP				0.47	0.40
		NEE				0.19	0.24
M24	S3	RECO				-0.63	-0.25
		GPP				-0.25	0.19
		NEE				-0.20	0.30
	S5	RECO				-0.63	-0.29
		GPP				-0.25	0.23
		NEE				-0.20	0.34
M25	S3	RECO	-0.03	-0.38	-4.78		
		GPP					
		NEE					
	S5	RECO	-0.03	-0.38	-4.78		
		GPP					
		NEE					
M26	S3	RECO	0.02	0.07	-15.64		
		GPP					
		NEE					
	S5	RECO	0.07	0.07	-15.52		
		GPP					
		NEE					
M28	S3	RECO				-0.89	-0.50
		GPP				0.22	-0.59
		NEE				-1.30	-0.45
	S5	RECO				-1.52	-0.72
		GPP				-0.52	-0.17
		NEE				-0.40	-0.05
MMM	S3	RECO	0.10	0.15	-6.12	0.21	0.38

	GPP	0.23	0.61		0.47	0.55
	NEE	0.12	0.57		0.29	0.30
S5	RECO	0.03	0.01	-3.53	0.17	0.25
	GPP	0.32	0.58		0.53	0.62
	NEE	0.22	0.55		0.30	0.45

3.4. Minimum ensemble size

We attempted to identify the minimum number of models required to obtain reliable results for stages S3 and S5, with focus on the three independent outputs (GPP, RECO, NEE) on both grassland and cropland sites (Figs. 6 and K-O in the supplement). For different sites, we observed that there could be large differences in the *z*-score results obtained with different ensemble sizes with different output variables. In general, grassland sites were characterized by greater *z*-score values than C1 and C2 crop sites. However, C3 (Indian crop site) showed the greatest deviation from observations (Fig. M in the supplement). For C1, our analysis suggests that the ensemble size could be reduced down to five models for RECO and even below for GPP, but for NEE only ensemble sizes of at least 13 models reduced *z*-score values within the range -2 and +2 (Fig. 6 and Fig. K in the Supplementary material). C2 resulted the easiest site to simulate, with *z*-scores mostly within the range -1 and +1 - (i.e. approximating the 68% confidence limit of a normal distribution) for any model ensemble at both S3 and S5 calibration stages for RECO, GPP and NEE (Fig. L in the Supplementary material). Compared to C1, the estimated minimum number at G3 varied less with output variables: 7 models for NEE, 9 models for GPP and 11 models with RECO (Fig. 6 and Fig. N in the Supplementary material). At G4, S5 calibration stage showed that the minimum number of models would be around nine for RECO, seven for GPP and six for NEE (Fig. O in the Supplementary material). Overall (considering all the sites), our analysis suggests that ensemble sizes below 13 models might not always guarantee sufficient accuracy in C-flux estimates. We note in particular the increasing variability of *z*-scores observed with RECO at C3 (up to about +15) as the ensemble size decreases (Fig. M in the Supplementary material).

4. Discussion

The results in this paper show that the suitability of a multi-model ensemble to simulate agricultural C fluxes depends on the variables being collected to calibrate models. With respect to emission-related processes, up to recently it has been considered that it is "premature to fully trust model outputs as representing reality" (Oertel et al., 2016, p. 344). In our exercise (which is the first on agricultural C fluxes), we provided an update on what we can reasonably expect from using an ensemble of biogeochemical models. These results reinforced the idea that at large-scale, multiple model ensembles could be a promising way to orient future modelling studies, with plant and soil observations as a minimum data requirement for model calibration (S3 and S4). Additional observations (such as C-N fluxes) might not be needed for a more detailed model calibration (e.g. results for S5 in Fig. 2). For instance, the use of N_2O emission data for calibration could increase the uncertainty of model estimates (e.g., Del Grosso et al., 2011; Hense et al., 2013), considering the high spatial and temporal variability associated with heterogeneous and intermittent N_2O emissions (e.g. Grant and Pattey, 2003). Unlikely our results have been affected by the different calibration techniques used. In fact, Wallach et al. (2019) showed that different calibration techniques do not seem to be primarily responsible for differences in model performance, and considering that most of the modelling teams derived parameter values based on a manual trial-and-error approach (Table 4). When several (differently packaged) models and complex datasets are mobilised in large-scale multi-model ensembles, the uncertainty in calibrated parameters tends to be confounded with the uncertainty in model structure (Wallach and Thorburn, 2017). Usually, calibration techniques are considered a lower priority in agricultural ensemble modelling, where the reduction of uncertainties is mostly limited by the limited quality of the calibration data (e.g. Angulo et al., 2003; Maiorano et al., 2017). However, each situation can be so unique (e.g. supplied data are incorrectly measured or are affected by unreported factors, such as pest damage) that generic lessons cannot be drawn from this whole exercise. During the course of this exercise, some modelling teams noticed model structural problems, which could later be resolved.

In our study, the improvements in C-flux estimates (and uncertainty reduction) obtained with the multi-stage calibration process showed that the use of additional data at S5 did not always lead to improved results compared to S3 and S4. In particular, the additional calibration performed with C and N fluxes (S5) produced some less accurate predictions of crop GPP than those obtained with soil properties and soil temperature and water dynamics (S4), which produced the best predictions in general. Then, we noticed some non-zero GPP values during fallow periods, when no growing plants are expected and GPP should be zero. In practice, some weeds may be present, giving some limited GPP. Models are not expected to confidently predict the occasional escape of some weeds from the attempts to control them but the way different models address site-, method- and weather-specific phenomena (which was not investigated in this study) could have produced some limited photosynthetic activity during fallow periods.

For an accurate estimate of GPP in grasslands, however, more detailed model calibration may be needed. C-flux estimates from grassland models are generally more uncertain than from crop models due to the inherent complexity of grassland systems (multi-species communities of grasses, legumes and forbs) and their management. The latter may include relatively simple grazing schemes, e.g. intensive grazing by heifer cows as in G3, and combinations of mowing and grazing with ewes, lambs, heifers and calves like in G4 (whose representation in models is not straightforward). S4 and S5 substantially improved some MMM predictions for both G3 and G4. Soil-based calibration (S4) improved the simulations but the full calibration (S5) provided the best fit. Future multi-model comparison studies should use mown grasslands (which are simpler management schemes than grazing) to try to resolve some of the differences between observations and modelled values.

The estimation of RECO was also more uncertain in grasslands than in crops. Big fluctuations of this variable in grasslands are likely due to the variability of grazing animals' respiration, which adds to the variability of plant and soil respiration fluxes (e.g. Kirschbaum et al., 2015; Cai et al., 2018). The envelope of inter-annual variability decreased after S2, which indicates that a calibration based on biomass growth and plant and leaf development is essential for reliable estimates of RECO.

Both observed and simulated NEE showed negative values (net C uptake), with the exception of fallow periods and the phacelia growing season. It is known that phacelia, as a cover crop, may increase soil CO_2 emission due to an enhanced input of organic residues (e.g. Bodner et al., 2018). The lowest values were associated with maize, winter wheat, rice and triticale crops. However, models could underestimate NEE values from the whole crop rotation system (e.g. C2 site), because they underestimate the release of CO_2 to the atmosphere from fallow periods. This means that models need to improve their simulations of bare soil processes during the intercrop period. However, improvements in model predictions were observed after the S3 calibration stage. The C uptake (NEE<0) observed and modelled in C1 and C3 crop rotations did not include fallow periods for which measurements were not made available, thus NEE values were only for the crop growing seasons. In both grassland sites G3 and G4, model results reflected the limited variability of NEE observations, which was roughly half those of RECO and GPP. Thus, with NEE, some performance gain was obtained from the uncertainty compensation.

Higher CUE promotes biomass accumulation and, indirectly, C stabilization in soil layers, while lower CUE favours respiration and C losses (Bradford and Crowther, 2013; Geyer et al., 2019). In our site–by-site analysis, CUE values were generally better estimated after S3 calibration stage. Among crops, phacelia and soybean showed the highest variability in their

MMM values, while fallow periods provided the worst estimates (Fig. A in the Supplementary material).

For Int$_C$, the provision of phenology and production data at S3 was effective in improving model predictions (Fig. A in the Supplementary material), which is expected considering that Int$_C$ is calculated on grain yield/grassland offtake.

Overall, the MMM provided more accurate simulations in most cases than individual models (as shown by the regression lines of Figs. 3-5 and Figs B-J in the Supplementary material). Even though some individual models were outperforming the MMM (e.g. M6, M16, M22, M23, M24) in certain cases (outputs/sites/calibration stages), that response was not general (e.g. Table 6). We confirm with this study 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 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 should be increased (at least 13 models) when model ensembles are implemented to simulate C fluxes at different climatic regions worldwide. Only in specific situations, e.g. C2 site, ~9 models could provide reliable C-flux estimates. With grasslands, the minimum ensemble size should include at least 11 models.

5. Summary and conclusions

This study presents a framework for interpretation of model performance and uncertainties obtained with a set of biogeochemical models (individually and in an ensemble) simulating C fluxes in cropping and grassland systems at a variety of distant and contrasted sites. There are multiple foci when designing multi-model studies of agricultural systems (such as crop rotations and grasslands) depending on the questions to be answered. Our study shows that we could not identify the best model(s) for crop and grassland C fluxes and no probability of success could be assigned to prove the suitability of using one biogeochemical model rather than another. We demonstrate the potential that a multi-model ensemble can have for jointly estimating different C fluxes (primary production, ecosystem respiration and net ecosystem exchanges) and production-scaled emissions (e.g. CO_2-C emission intensities and C use efficiencies).

We showed that reduced calibration datasets (vegetation data) could be adequate for providing sufficiently reliable outputs (e.g. to continue to progress towards updating the inventory of C databases, West et al., 2010), but additional biophysical and biogeochemical data can further improve results under certain circumstances. Further improvements of data sources, such as phenological observations, could help refine model estimates and form a baseline for screening agricultural practices and mitigation options at croplands and grasslands, as presented in Sándor et al. (2018). Moreover, there is a high uncertainty of modelled fluxes during fallow periods, which would need more accurate data.

These results paved the way for using model ensemble medians for field-scale estimation of C fluxes. Our results inform about the possible use of model ensembles for upscaling projections of C fluxes and derived outputs, from field scale to larger spatial units (e.g. gridded projections) as needed for Tier 3 national inventories (e.g. Folberth et al., 2016; Zscheischler et al., 2017).

However, model inter-comparisons have their limitations. Although our comparison was large compared to other studies (e.g. Sándor et al., 2016), there was a lack of case studies in this exercise from Africa, South America and Oceania, which would extend the geographical coverage. Our study-sites mostly targeted agricultural areas of the Northern hemisphere (four temperate and one tropical), as part of a broader study covering more agricultural areas in both hemispheres (Ehrhardt et al., 2018).

Moreover, the various model types and variants evaluated here did not cover all the modelling approaches used to simulate C fluxes from crop and grassland systems (e.g. the model used by Senapati et al., 2016). They reasonably represent current approaches (the basis of development and processes were scrutinized), but we think that crop and grassland model inter-comparisons with the inclusion of more models should be continued to assess and improve our ability to simulate biogeochemical processes with acceptable quality. Further analyses and better understanding of these multi-model ensembles are required to achieve key progress in crop and grassland modelling, by assessing more in-depth model responses and uncertainties against climate and management drivers.

Acknowledgements

This study was coordinated by the Integrative Research Group of the Global Research Alliance (GRA) on agricultural GHGs and was supported by five research projects (CN- MIP, Models4Pastures, MACSUR, COMET- Global and MAGGNET), which received funding by a multi-partner call on agricultural greenhouse gas research of the Joint Programming Initiative 'FACCE' through its national financing bodies.

RS received mobility funding from the French Embassy in Budapest (Hungary) by way of "Make Our Planet Great Again" programme. US acknowledges funding for the data collection through the EU projects GREENGRASS (EC EVK2-CT2001-00105), CarboEurope (GOCE-CT-2003-505572) and the NitroEurope Integrated Project (017841), and SRUC's contribution to compile the data (Stephanie K. Jones and Robert M. Rees). Giovanna Seddaiu from University of Sassari (Italy) is acknowledged for her support with EPIC simulations.

References

Allard, V., Soussana, J.-F., Falcimagne, R., Berbigier, P., Bonnefond, J.M., Ceschia, E., D'hour, P., Hénault, C., Laville, P., Martin, C., Pinarès-Patino, C., 2007. The role of grazing management for the net biome productivity and greenhouse gas budget (CO_2, N_2O and CH_4) of semi-natural grassland. Agr. Ecosyst. Environ. 12, 47-58. https://doi.org/10.1016/j.agee.2006.12.004.

Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S., Ewert, F., 2013. Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe. Agr. Forest Meteorol. 170, 32-46. https://doi.org/10.1016/j.agrformet.2012.11.017.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A., Boote, K.J., Thorburn, P., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J.E., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Travasso, M., Tao, F., Waha, K., Wallach, D., White, J.W., Wolf, J., 2013. Uncertainty in simulating wheat yields under climate change. Nat. Clim. Change 3, 827-832. https://doi.org/10.1038/nclimate1916.

Barnett, C., Hossel, J., Perry, M., Procter, C., Hughes, G., 2006. A handbook of climate trends across Scotland. Scotland and Northern Ireland Forum for Environmental Research, SNIFFER Project CC03, Edinburgh.

Basso, B., Dumont, B., Maestrini, B., Shcherbak, I., Robertson, G.P., Porter, J.R., Smith, P., Paustian, K., Grace, P.R., Asseng, S., Bassu, S., Biernath, C., Boote, K.J., Cammarano, D., De Sanctis, G., Durand, J.-L., Ewert, F., Gayler, S., Hyndman, D.W., Kent, J., Martre, P., Nendel, C., Priesack, E., Ripoche, D., Ruane, A.C., Sharp, J., Thorburn, P.J., Hatfield, J.L., Jones, J.W., Rosenzweig, C., 2018. Soil organic carbon and nitrogen feedbacks on crop yields under climate change. Agricultural and Environmental Letters 3: 180026. doi: 10.2134/ael2018.05.0026.Bassu, S., Brisson, N., Durand, J.L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.H., Kumar, N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? Global Change Biol. 20, 2301-2320. https://doi.org/10.1111/gcb.12520.

Bhatia, A., Pathak, H., Jain, N., Singh, P.K., Tomer, R., 2012. Greenhouse gas mitigation in rice-wheat system with leaf color chart-based urea application. Environ. Monit. Assess. 184, 3095-3107. https://doi.org/10.1007/s10661-011-2174-8.

Bodner, G., Mentler, A., Klik, A., Kaul, A.-P., Zechmeister-Boltenstern, S., 2018. Do cover crops enhance soil greenhouse gas losses during high emission moments under temperate Central Europe conditions? Journal of Land Management, Food and Environment 68, 171-187. https://doi.org/10.1515/boku-2017-0015.

Bradford, M.A., Crowther, T.W., 2013. Carbon use efficiency and storage in terrestrial ecosystems. New Phytol. 199, 7-9. https://doi.org/10.1111/nph.12334.

Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C.D., Doro, L., Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I., Klumpp, K., Léonard, J., Martin, R., Massad, R.S., Recous, S., Seddaiu, G., Sharp, J., Smith, P., Smith, W.N., Soussana, J.-F., Bellocchi, G., 2017. Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. Sci. Total Environ. 598, 445-470. https://doi.org/10.1016/j.scitotenv.2017.03.208.

Cai, Q., Yan, X., Li, Y., Wang, L., 2018. Global patterns of human and livestock respiration. Sci. Rep.-UK 8, 9278. https://doi.org/10.1038/s41598-018-27631-7.

Challinor, A.J., Müller, C., Asseng, S., Deva, C., Nicklin, K.J., Wallach, D., Vanuytrecht, E., Whitfield, S., Ramirez-Villega, J., Koehler, A.-K., 2018. Improving the use of crop models for risk assessment and climate change adaptation. Agric. Syst. 159, 296-306. https://doi.org/10.1016/j.agsy.2017.07.010.

Chang, J., Ciais, P., Viovy, N., Vuichard, N., Sultan, B., Soussana, J.-F., 2015. The greenhouse gas balance of European grasslands. Global Change Biol. 21, 3748-3761. https://doi.org/10.1111/gcb.12998.

Confalonieri, R., Bellocchi, G., Donatelli, M., 2010. A software component to compute agro-meteorological indicators. Environ. Modell. Softw. 25, 1485-1486. https://doi.org/10.1016/j.envsoft.2008.11.007.

Confalonieri, R., Bregaglio, S., Acutis, M., 2016. Quantifying uncertainty in crop model predictions due to the uncertainty in the observations used for calibration. Ecol. Model. 328, 72-77. https://doi.org/10.1016/j.ecolmodel.2016.02.013.

Curtin, D., Wang, H., Selles, F., Mcconkey, B.G., Campbell, C.A., 2000. Tillage effects on carbon fluxes in continuous wheat and fallow-wheat rotations. Soil Sci. Soc. Am. J. 64, 2080-2086. https://doi.org/10.2136/sssaj2000.6462080x.

De Martonne, E., 1942. Nouvelle carte mondiale de l'indice d'aridité. Annales de Géographie 51, 242–250. (in French)

Del Grosso, S.J., Wirth, T., Ogle, S.M., Parton, W.J., 2011. Estimating agricultural nitrous oxide emissions. Eos Transactions of the American Geophysical Union 89, 529-540.

Diodato, N., Ceccarelli, M., 2004. Multivariate indicator Kriging approach using a GIS to classify soil degradation for Mediterranean agricultural lands. Ecol. Indic. 4, 177–187. https://doi.org/10.1016/j.ecolind.2004.03.002.

Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., McAuliffe, R., Recous, S., Sándor, R., Smith, P., Snow, V., Migliorati, M.D.A., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Giacomini, S.J., Grant, B., Harrison, M.T., Jones, S.K., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Liebig, M., Lieffering, M., Martin, R., Massad, R.S., Meier, E., Merbold, L., Moore, A.D., Myrgiotis, V., Newton, P., Pattey, E., Rolinski, S., Sharp, J., Smith, W.N., Wu, L., Zhang, Q., 2018. Assessing uncertainties in crop and pasture ensemble model simulations of productivity and N_2O emissions. Global Change Biol. 24, e603-e616. https://doi.org/10.1111/gcb.13965.

Eza, E.H.U., Shtiliyanova, A., Borras, D., Bellocchi, G., Carrère, P., Martin, R., 2015. An open platform to assess vulnerabilities to climate change: An application to agricultural systems. Ecol. Inform. 30, 389-396. https://doi.org/10.1016/j.ecoinf.2015.10.009.

Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L.B., Obersteiner, M., van der Velde, M., 2015. Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. Nat. Commun. 7, 11872. https://doi.org/10.1038/ncomms11872.

Geyer, K.M., Dijkstra, P., Sinsabaugh, R., Frey, S.D., 2019. Clarifying the interpretation of carbon use efficiency in soil through methods comparison. Soil Biol. Biochem. 128, 79-88. https://doi.org/10.1016/j.soilbio.2018.09.036.

Grant, R., Pattey, E., 2003. Modelling variability in N_2O emissions from fertilized agricultural fields. Soil Biol. Biochem. 35, 225-243. https://doi.org/10.1016/S0038-0717(02)00256-0.

Graux, A.-I., Bellocchi, G., Lardy, R., Soussana, J.-F., 2013. Ensemble modelling of climate change risks and opportunities for managed grasslands in France. Agr. Forest Meteorol. 170, 114-131. https://doi.org/10.1016/j.agrformet.2012.06.010.

Grosz, B., Dechow, R., Gebbert, S., Hoffmann, H., Zhao, G., Constantin, J., Raynal, H., Wallach, D., Coucheney, E., Lewan, E., Eckersten, H., Specka, X., Kersebaum, K-C, Nendel, C., Kuhnert, M., Yeluripati, J., Haas, E., Teixeira, E., Bindi, M., Trombi, G., Moriondo, M., Doro, L., Roggero, PP., Zhao, Z., Wang, E., Tao, F., Roetter, R., Kassie, B., Cammarano, D., Asseng, S., Weihermueller, L., Siebert, S., Gaiser, T., Ewert, F., 2017. The implication of input data aggregation on up-scaling soil organic carbon changes. Environ. Modell. Softw. 96, 361-377. https://doi.org/10.1016/j.envsoft.2017.06.046.

Harrison, M.T., Roggero, P.P., Zavattaro, L., 2019. Simple, efficient and robust techniques for automatic multi-objective function parameterisation: Case studies of local and global optimisation using APSIM. Environ. Modell. Softw. 117, 109-133. https://doi.org/10.1016/j.envsoft.2019.03.010.

Harrison, M.T., Tardieu, F., Dong, Z., Messina, C.D., Hammer, G.L., 2014. Characterizing drought stress and trait influence on maize yield under current and future conditions. Global Change Biol. 20, 867-878. https://doi.org/10.1111/gcb.12381.

Hense, A., Skiba, U., Famulari, D., 2013. Low cost and state of the art methods to measure nitrous oxide emissions. Environ. Res. Lett. 8, 025022. https//doi.org/10.1088/1748-9326/8/2/025022.

IPCC (Intergovernmental Panel on Climate Change) (2013) IPCC 5th Assessment Report 'Climate Change 2013: the Physical Science Basis'. University Press, Cambridge. http://www.ipcc.ch/report/ar5/wg1/#.Uk7O1xBvCVq.

Jégo, G., Pattey, E., Liu, J., 2012. Using leaf area index, retrieved from optical imagery, in the STICS crop model for predicting yield and biomass of field crops. Field Crop. Res. 131, 63-74. https://doi.org/10.1016/j.fcr.2012.02.012.

Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017a. Brief history of agricultural systems modeling. Agr. Syst. 155, 240-254. https://doi.org/10.1016/j.agsy.2016.05.014.

Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017b. Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. Agr. Syst. 155, 269-288. https://doi.org/10.1016/j.agsy.2016.09.021.

Jones, S.K., Helfter, C., Anderson, M., Coyle, M., Campbell, C., Famulari, D., Di Marco, C., van Dijk, N., Topp, C.F.E., Kiese, R., Kindler, R., Siemens, J., Schrumpf, M., Kaiser, K., Nemitz, E., Levy, P., Rees, R.M., Sutton, M.A., Skiba, U.M., 2017c. The nitrogen, carbon and greenhouse gas budget of a grazed, cut and fertilised temperate grassland. Biogeosciences 14, 2069-2088. https://doi.org/10.5194/bg-14-2069-2017.

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

Kirschbaum, M.U.F., Schipper, L.A., Mudge, P.L., Rutledge, S., Puche, N.J.B., Campbell, D.I., 2017. The trade-offs between milk production and soil organic carbon storage in dairy systems under different management and environmental factors. Sci. Total Environ. 577, 61-72. https://doi.org/10.1016/j.scitotenv.2016.10.055.

Klumpp, K., Tallec, T., Guix, N., Soussana, J.-F., 2011. Long-term impacts of agricultural practices and climatic variability on carbon storage in a permanent pasture. Global Change Biol. 17, 3534-3545. https://doi.org/10.1111/j.1365-2486.2011.02490.x.

Kuhnert, M., Yeluripati, J., Smith, P., Hoffmann, H., van Oijen, M., Constantin, J., Coucheney, E., Dechow, R., Eckersten, H., Gaiser, T., Grosz, B., Haas, E., Kersebaum, K-C, Kiese, R., Klatt, S., Lewan, E., Nendel, C., Raynal, H., Sosa, C., Specka, X., Teixeira, E., Wang, E., Weihermüller, L., Zhao, G., Zhao, Z., Ogle, S., Ewert, F., 2017. Impact analysis of climate data aggregation at different spatial scales on simulated net primary productivity for croplands. Eur. J. Agron. 88, 41-52. https://doi.org/10.1016/j.eja.2016.06.005.

Lardy, R., Bachelet, B., Bellocchi, G., Hill, D., 2014. Towards vulnerability minimization of grassland soil organic matter using metamodels. Environ. Modell. Softw. 52, 38-50. https://doi.org/10.1016/j.envsoft.2013.10.015.

Laville, P., Lehuger, S., Loubet, B., Chaumartin, F., Cellier, P., 2011. Effect of management, climate and soil conditions on N_2O and NO emissions from an arable crop rotation using high temporal resolution measurements. Agr. Forest Meteorol. 151, 228-240. https://doi.org/10.1016/j.agrformet.2010.10.008.

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T., Gaydon, D., Marcaida III, M., Nakagawa, H., Oriol, P., Ruane, A.C., Ruget, F., Singh, B., Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z., Bouman, B., 2015. Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Global Change Biol. 21, 1328-1341. https://doi.org/10.1111/gcb.12758.

Loubet, B., Laville, P., Lehuger, S., Larmanou, E., Flechard, C., Mascher, N., Genermont, S., Roche, R., Ferrara, R. M., Stella, P., Personne, E., Durand, B., Decuq, C., Flura, D., Masson, S., Fanucci, O., Rampon, J.-N., Siemens, J., Kindler, R., Gabrielle, B., Schrumpf, M., Cellier, P., 2011. Carbon, nitrogen and greenhouse gases budgets over a four years crop rotation in northern France. Plant Soil 343, 109-137. https://doi.org/10.1007/s11104-011-0751-9.

Ludwig, F., Asseng, S., 2006. Climate change impacts on wheat production in a Mediterranean environment in Western Australia. Agr. Syst. 90, 159-179. https://doi.org/10.1016/j.agsy.2005.12.002.

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. Environ. Modell. Softw. 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., Ruane, A.C., Semenov, M.A., Wallach, D., Wang, E., Alderman, P.D., Kassie, B.T., Biernath, C., Basso, B., Cammarano, D., Challinor, A.J., Doltra, J., Dumont, B., Rezaei, E.E., Gayler, S., Kersebaum, K.C., Kimball, B.A., Koehler, A.-K., Liu, B., O'Leary, G., Olesen, J.E., Ottman, M.J., Priesack, E., Reynolds, M., Stratonovitch, P., Streck, T., Thorburn, P.J., Waha, K., Wall, G.W., White, J.W., Zhao, Z., Zhu, Y., 2017. Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. Field Crop. Res. 202, 5-20. https://doi.org/10.1016/j.fcr.2016.05.001.

Mangani, R., Tesfamariam, E., Bellocchi, G., Hassen, A., 2018. Modelled impacts of extreme heat and drought on maize yield in South Africa. Crop & Pasture Science 69, 703-716. https://doi.org/10.1071/CP18117.

Mangani, R., Tesfamariam, E., Engelbrecht, C.J., Bellocchi, G., Hassen, A., 2019. Potential impacts of extreme weather events in main maize (*Zea mays* L.) producing areas of South Africa under rainfed conditions. Reg. Environ. Change 19, 1441-1452. https://doi.org/10.1007/s10113-019-01486-8.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherback, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Global Change Biol. 21, 911-925. https://doi.org/10.1111/gcb.12768.

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

Oertel, C., Matschullat, J., Zurba, K., Zimmermann, F., 2016. Greenhouse gas emissions from soils - A review. Chem. Erde-Geochem. 76, 327-352. https://doi.org/10.1016/j.chemer.2016.04.002.

Palosuo, T., Kersebaum, K.C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J.O., Patil, R.H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F., Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B., Rötten, R., 2011. Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. Eur. J. Agron. 35, 103-114. https://doi.org/10.1016/j.eja.2011.05.001.

Pattey, E., Edwards, G., Strachan, I.B., Desjardins, R.L., Kaharabata, S., Wagner, C., 2006. Towards standards for measuring greenhouse gas fluxes from agricultural fields using instrumented towers. Can. J. Soil Sci. 86, 373-400. https://doi.org/10.4141/S05-100.

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.

Rodríguez, A., Ruiz-Ramos, M., Palosuo, T., Carter, T.R., Fronzek, S., Lorite, I.J., Ferrise, R., Pirttioja, N., Bindi, M., Baranowski, P., Buis, S., Cammarano, D., Chen, Y., Dumont, B., Ewert, F., Gaiser, T., Hlavinka, P., Hoffmann, H., Höhn, J.G., Jurecka, F., Kersebaum, K.C., Krzyszczak, J., Lana, M., Mechiche-Alami, A., Minet, J., Montesino, M., Nendel, C., Porter, J.R., Ruget, F., Semenov, M.A., Steinmetz, Z., Stratonovitch, P., Supit, I., Tao, F., Trnka, M., de Wit, A., Rötter, R.P., 2019. Implications of crop model ensemble size and composition for estimates of adaptation effects and agreement of recommendations. Agr. Forest Meteorol. 15, 351-362. https://doi.org/10.1016/j.agrformet.2018.09.018.

Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. Agr. Forest Meteorol. 170, 166-182. https://doi.org/10.1016/j.agrformet.2012.09.011.

Ruane, A.C., Hudson, N.I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., Boote, K.J., Thorburn, P.J., Aggarwal, P.K., Angulo, C., Basso, D., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Kumar, S.N., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Rötter, R.P., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Travasso, M., Waha, K., Wallach, D., White, J.W., Wolf, J., 2016. Multi-wheat model ensemble responses to interannual climate variability. Environ. Modell. Softw. 81, 86-101. https://doi.org/10.1016/j.envsoft.2016.03.008.

Ruiz-Ramos, M., Mínguez, M.I., 2010. Evaluating uncertainty in climate change impacts on crop productivity in the Iberian Peninsula. Clim. Res. 44, 69-82. https://doi.org/10.3354/cr00933.

Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E., Ma, S., Perego, A., Rolinski, S., Ruget, F., Sanna, M., Seddaiu, G., Wu, L., Bellocchi, G., 2017. Multi-model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean grasslands: uncertainties and ensemble performance. Eur. J. Agron. 88, 22-40. https://doi.org/10.1016/j.eja.2016.06.006.

Sándor, R., Barcza, Z., Hidy, D., Lellei-Kovács, E., Ma, S., Bellocchi, G., 2016. Modelling of grassland fluxes in Europe: Evaluation of two biogeochemical models. Agr. Ecosyst. Environ. 215, 1-19. https://doi.org/10.1016/j.agee.2015.09.001.

Sándor, R., Ehrhardt, F., Basso, B., Bellocchi, G., Bhatia, A., Brilli, L., De Antoni Migliorati, M., Doltra, J., Dorich, C., Doro, L., Fitton, N., Giacomini, S. J., Grace, P., Grant, B., Harrison, M. T., Jones, S., Kirschbaum M. U. F., Klumpp, K., Laville, P., Léonard, J., Liebig, M., Lieffering, M., Martin, R., McAuliffe, R., Meiser, E., Merbold, L., Moore, A., Myrgiotis, V., Newton, P., Pattey, E., Recous, S., Rolinski, S., Sharp, J., Massad, R. S., Smith, P., Smith, W., Snow, V., Wu, L., Zhang, Q., Soussana, J.-F., 2016. C and N models intercomparison - benchmark and ensemble crop and grassland model estimates for grassland production. Advances in Animal Biosciences 7, 227-228. https://doi.org/10.1017/S2040470016000297.

Sándor, R., Ehrhardt, F., Brilli, L., Carozzi, M., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Dorich, C.D., Fuchs, K., Fitton, N., Gongadze, K., Klumpp, K., Liebig, M., Martin, R., Merbold, L., Newton, P.C.D., Rees, R.M., Rolinski, R., Bellocchi, G., 2018. The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed grasslands. Sci. Total Environ. 642, 292-206. https://doi.org/10.1016/j.scitotenv.2018.06.020.

Sansoulet, J., Pattey, E., Kröbel, R., Grant, B., Smith, W., Jégo, G., Desjardins, R.L., Tremblay, N., Tremblay, G., 2014. Comparing the performance of the STICS, DNDC, and DayCent models for predicting N uptake and biomass of spring wheat in Eastern Canada. Field Crop. Res. 156, 135-150. https://doi.org/10.1016/j.fcr.2013.11.010.

Senapati, N., Jansson, P-E., Smith, P., Chabbi, A., 2016. Modelling heat, water and carbon fluxes in mown grassland under multi-objective and multi-criteria constraints. Environ. Modell. Softw. 80, 201-224. https://doi.org/10.1016/j.envsoft.2016.02.025.

Skiba, U., Jones, S. K., Drewer, J., Helfter, C., Anderson, M., Dinsmore, K., McKenzie, R., Nemitz, E., Sutton, M.A., 2013. Comparison of soil greenhouse gas fluxes from extensive and intensive grazing in a temperate maritime climate. Biogeosciences 10, 1231-1241. https://doi.org/10.5194/bg-10-1231-2013.

Smith, P., Smith, J.U., Powlson, D.S., McGill, W.B., Arah, R.M., Chertov, O.G., Coleman, K., Franko, U., Frolking, S., Jenkinson, D.S., Jensen, L.S., Kelly, R.H., Klein-Gunnewiek, H., Komarov, A.S., Li, C., Molina, J.A.E., Mueller, T., Parton, W.J., Thornley, J.H.M., 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.

Soussana, J.-F., Ehrhardt, F., Conant, R., Harrison, M., Lieffering, M., Bellocchi, G., Moore, A., Rolinski, S., Snow, V., Wu, L., Ruane, A., 2015. Projecting grassland sensitivity to climate change from an ensemble of models. Abstract Book of the conference ‘Our common future under climate change’, July 7–10, Paris, France, K-2223-02. http://pool7.kermeet.com/C/ewe/ewex/unesco/DOCS/CFCC_abstractBook.pdf

Stocker, B.D., Roth, R., Joos, F., Spahni, R., Steinacher, M., Zaehle, S., Bouwman, L., Ri, X., Prentice, I.C., 2013. Multiple greenhouse-gas feedbacks from the land biosphere under future climate change scenarios. Nat. Clim. Change 3, 666-672. https://doi.org/10.1038/nclimate1864.

Tingem, M., Rivington, M., Bellocchi, G., Azam-Ali, S., Colls, J., 2008. Effects of climate change on crop production in Cameroon. Clim. Res. 36, 65-77. https://doi.org/10.3354/cr00733.

van Groenigen, J.W., Velthof, G.L., Oenema, O., van Groenigen, K. J., Kessel, C.V., 2010. Towards an agronomic assessment of N_2O emissions: a case study for arable crops. European Journal of Soil Science 61, 903–913. https://doi.org/10.1111/j.1365-2389.2009.01217.x.van Oijen, M., Balkovi, J., Beer, C., Cameron, DR., Ciais, P., Cramer, W., Kato, T., Kuhnert, M., Martin, R., Mynemi, R., Ramming, A., Rolinski, S., Soussana, J-F, Thonicke, K., van der Velde, M., Xu, L., 2014. Impact of droughts on the carbon cycle in European vegetation: a probabilistic risk analysis using six vegetation models. Biogeosciences 11, 6357-6375. https://doi.org/10.5194/bg-11-6357-2014.

Vellinga, T.V., de Haan, M.H.A., Schils, R.L.M., Evers, A., van den Pol-van Dasselaar, A., 2011. Implementation of GHG mitigation on intensive dairy farms: Farmers' preferences and variation in cost effectiveness. Livestock Science 137, 185–195. https://doi.org/10.1016/j.livsci.2010.11.005.

Vital, J.-A., Gaurut, M., Lardy, R., Viovy, N., Soussana, J.-F., Bellocchi, G., Martin, R., 2013. High-performance computing for climate change impact studies with the Pasture Simulation model. Comput. Electron. Agr. 98, 131-135. https://doi.org/10.1016/j.compag.2013.08.004.

Wallach, D., Palosuo, T., Thorburn, P., Seidel, S.J., Gourdain, E., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hochman, Z., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M, Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Luo, Q., Maestrini, B., Moriondo, M., Nariman Zadeh, H., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T, Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., 2019. How well do crop models predict phenology, with emphasis on the effect of calibration? bioRxiv, https://doi.org/10.1101/708578.

Wallach, D., Thorburn, P.J., 2017. Estimating uncertainty in crop model predictions: Current situation and future prospects. Eur. J. Agron. 88, A1-A7. https://doi.org/10.1016/j.eja.2017.06.001.

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., van Ittersum, M., Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G., Dumont, B., Rezaei, E.E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C., Jones, C.D., Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A.-K., Maiorano, A., Minoli, S., Müller, C., Kumar, S.N., Nendel, C., O'Leary, G.J., Palosuo, T., Priesack, E., Ripoche, D., Rötten, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Fao, F., Wolf, J., Zhang, Z., 2018. Global Change Biology 24, 5072-5083.

Xiao, X., Kuang, X., Sauer, T.J., Heitman, J., Horton, R., 2015. Bare soil carbon dioxide fluxes with time and depth determined by high-resolution gradient-based measurements and surface chambers. Soil Sci. Soc. Am. J. 79, 1073-1083. https://doi.org/10.2136/sssaj2015.02.0079.

Zhang, W., Zhang, F., Qi, J., Hou, F., 2017. Modeling impacts of climate change and grazing effects on plant biomass and soil organic carbon in the Qinghai–Tibetan grasslands. Biogeosciences 14, 5455-5470. https://doi.org/10.5194/bg-14-5455-2017.

Zscheischler, J., Mahecha, M. D., Avitabile, V., Calle, L., Carvalhais, N., Ciais, P., 2017. Reviews and syntheses: An empirical spatiotemporal description of the global surface–atmosphere carbon fluxes: opportunities and data limitations. Biogeosciences 14, 3685-3703. https://doi.org/10.5194/bg-14-3685-2017.

Figure legends

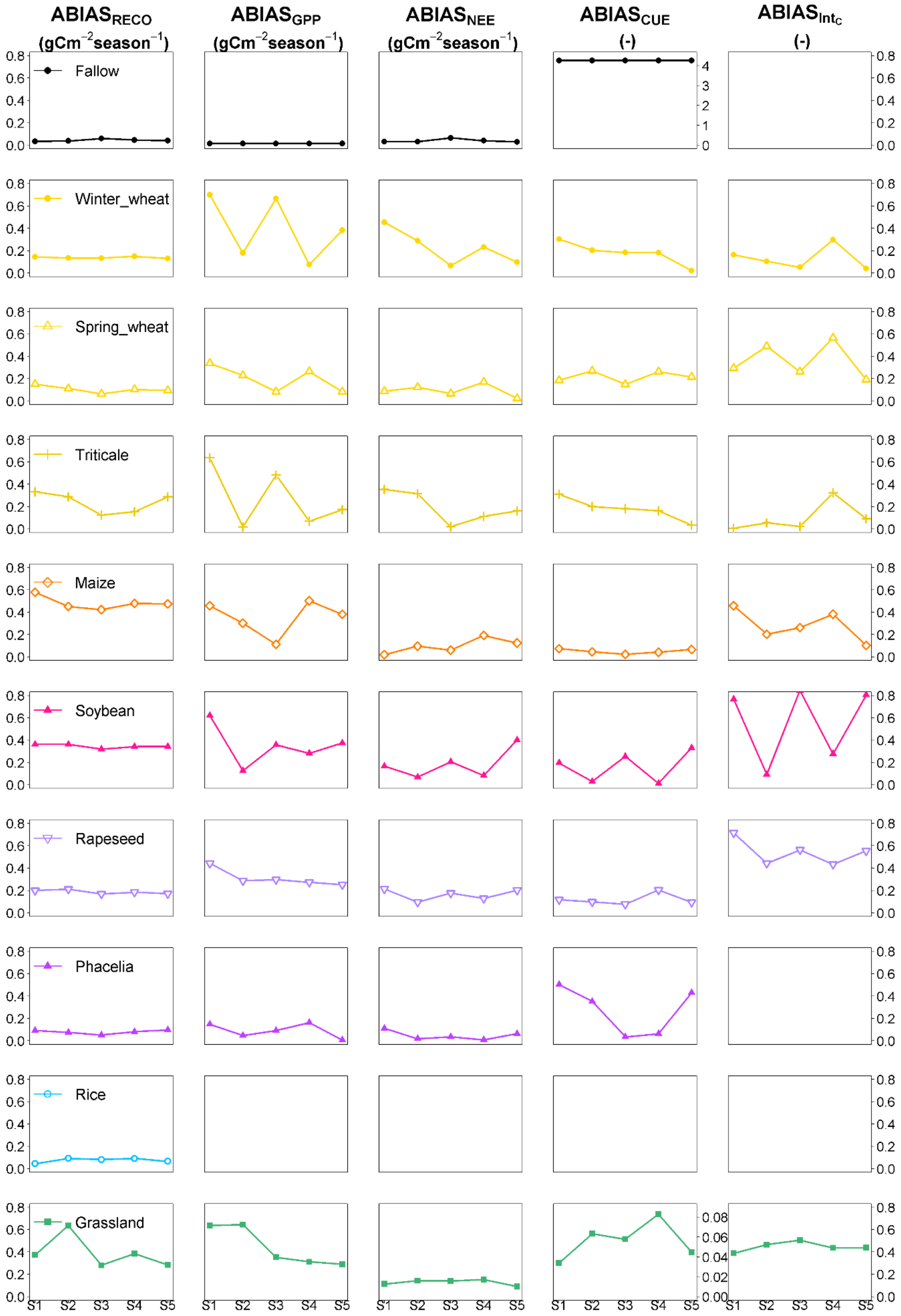
ABIASRECO
(gCm−2season−1)
ABIASGPP
(gCm−2season−1)
ABIASNEE
(gCm−2season−1)
ABIASCUE
(-)
ABIASIntc
(-)
Fallow
Winter_wheat
Spring_wheat
Triticale
Maize
Soybean
Rapeseed
Phacelia
Rice
Grassland
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
0.00
0.02
0.04
0.06
0.08
S1
S2
S3
S4
S5

Fig. 1. Variation of MMM absolute bias (ABIAS) values for ecosystem respiration (RECO), gross primary production (GPP), net ecosystem exchange (NEE), carbon use efficiency (CUE) and C intensity ($\mathrm{Int_C}$) calculated over multiple years at cropland (C1, C2 and C3) and grassland (G3 and G4) sites, for five calibration stages (S1-S5).

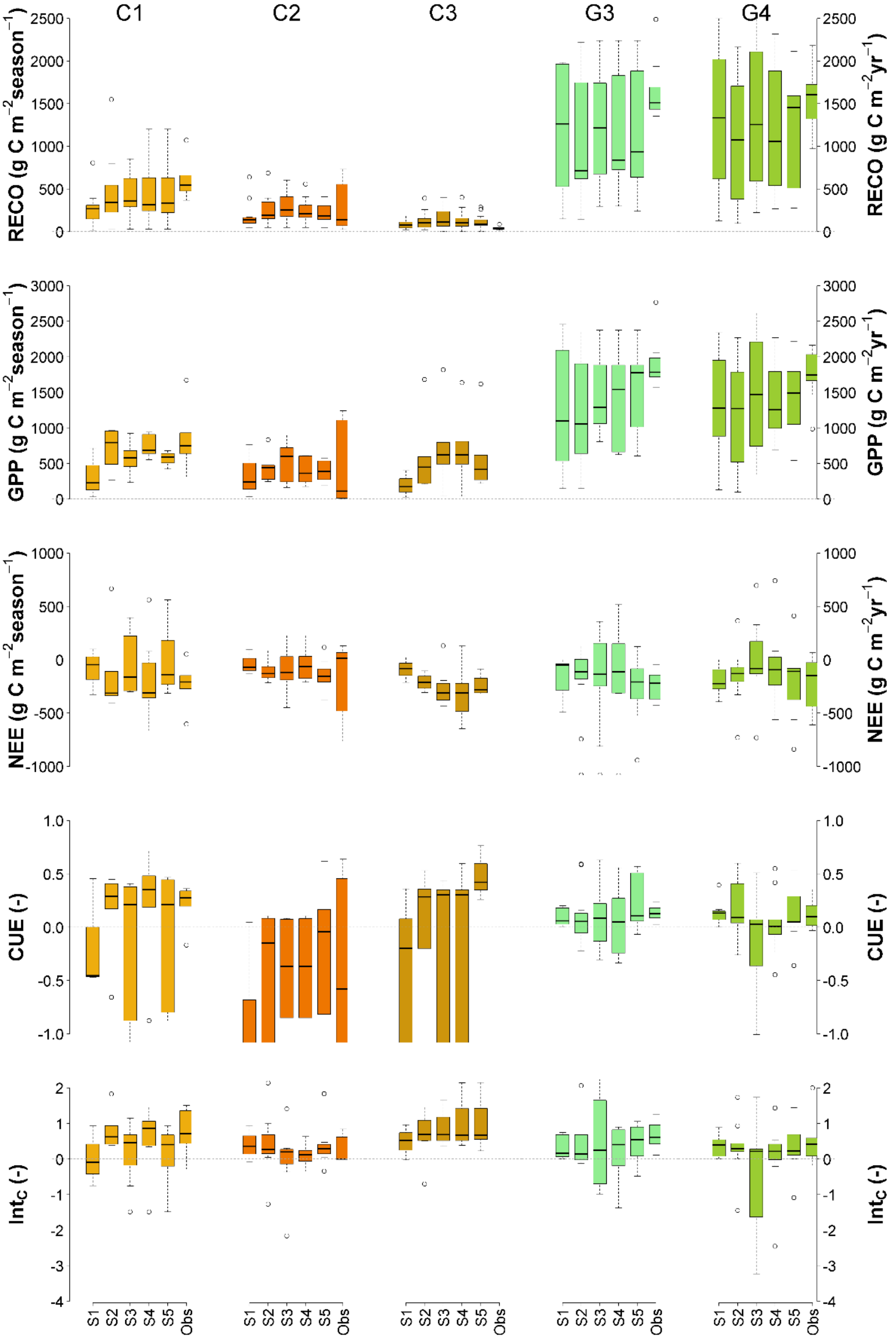

Fig. 2. Seasonal changes in ecosystem respiration (RECO), gross primary production (GPP), net ecosystem exchange (NEE), carbon use efficiency (CUE) and C intensity (Int$_C$) calculated over multiple years at C1, C2 and C3 crop, and G3 and G4 grassland sites, for five calibration stages

(S1 to S5) and the observation (Obs). Number of crop seasons/grassland years: soybean: 1; triticale: 1; phacelia: 1; spring wheat: 2; rice: 2; maize: 3; rapeseeds: 4; winter wheat: 5; fallow: 9; grasslands: 19. For each calibration stage, black lines show multi-model median. Boxes delimit the 25th and 75th percentiles. Whiskers are 10th and 90th percentiles. Circles indicate outliers. For Obs, black line shows the observed mean.

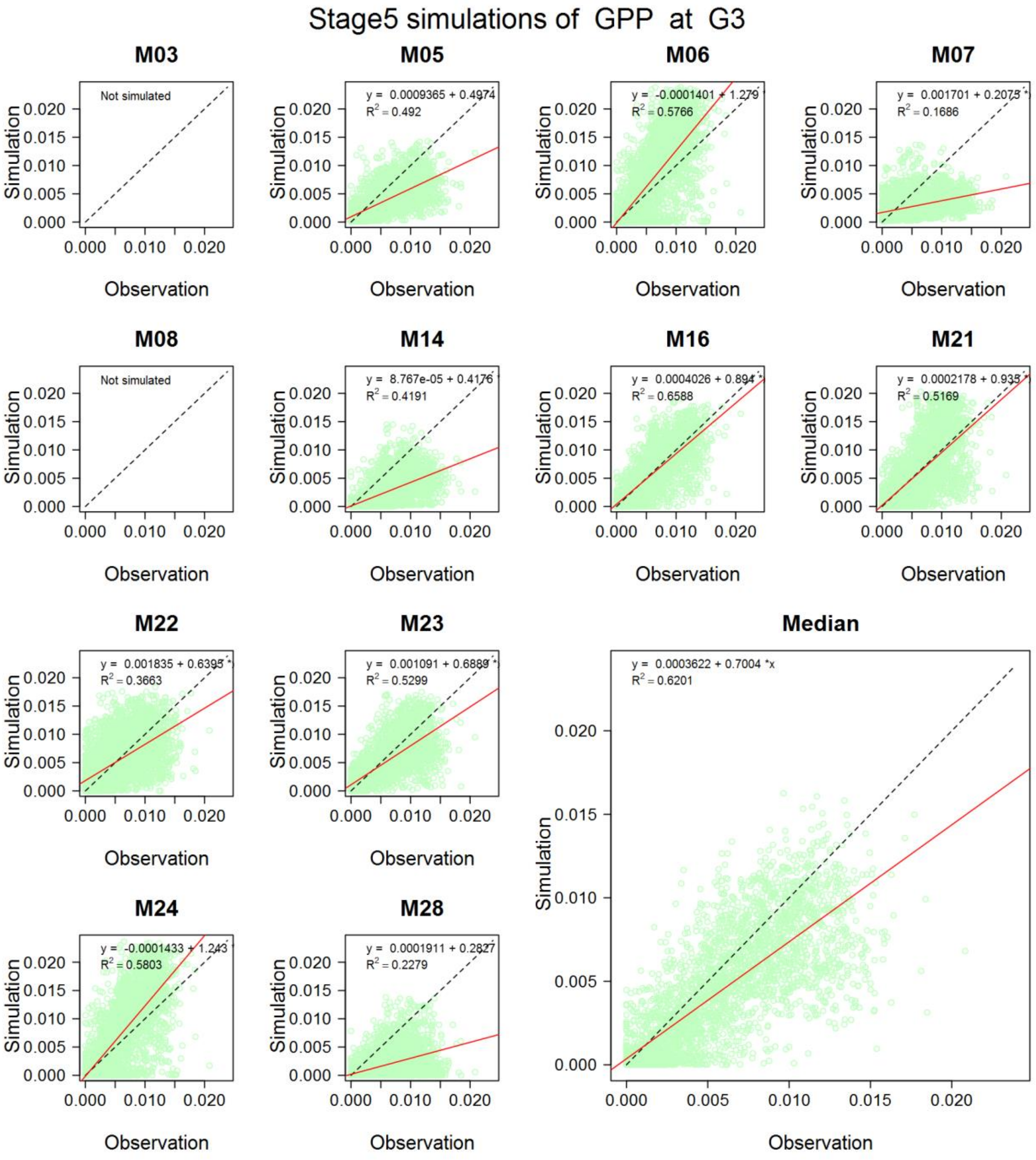

Fig. 3. S5 calibration stage: comparison of simulated (individual models and multi-model median) and observed daily gross primary production (GPP) data across multiple years at G3 site.

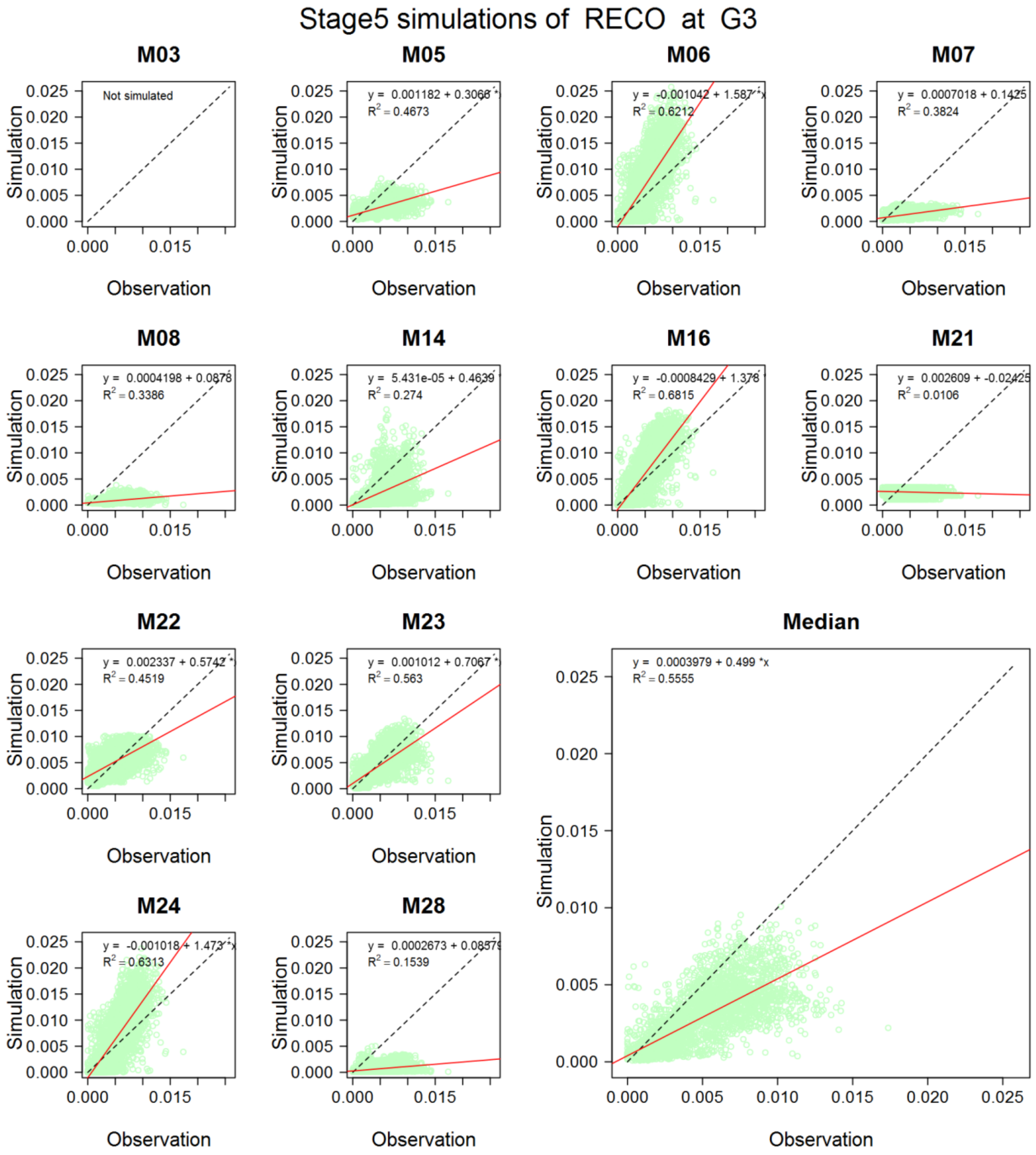

Fig. 4. S5 calibration stage: comparison of simulated (individual models and multi-model median) and observed daily ecosystem respiration (RECO) data across multiple years at G3 site.

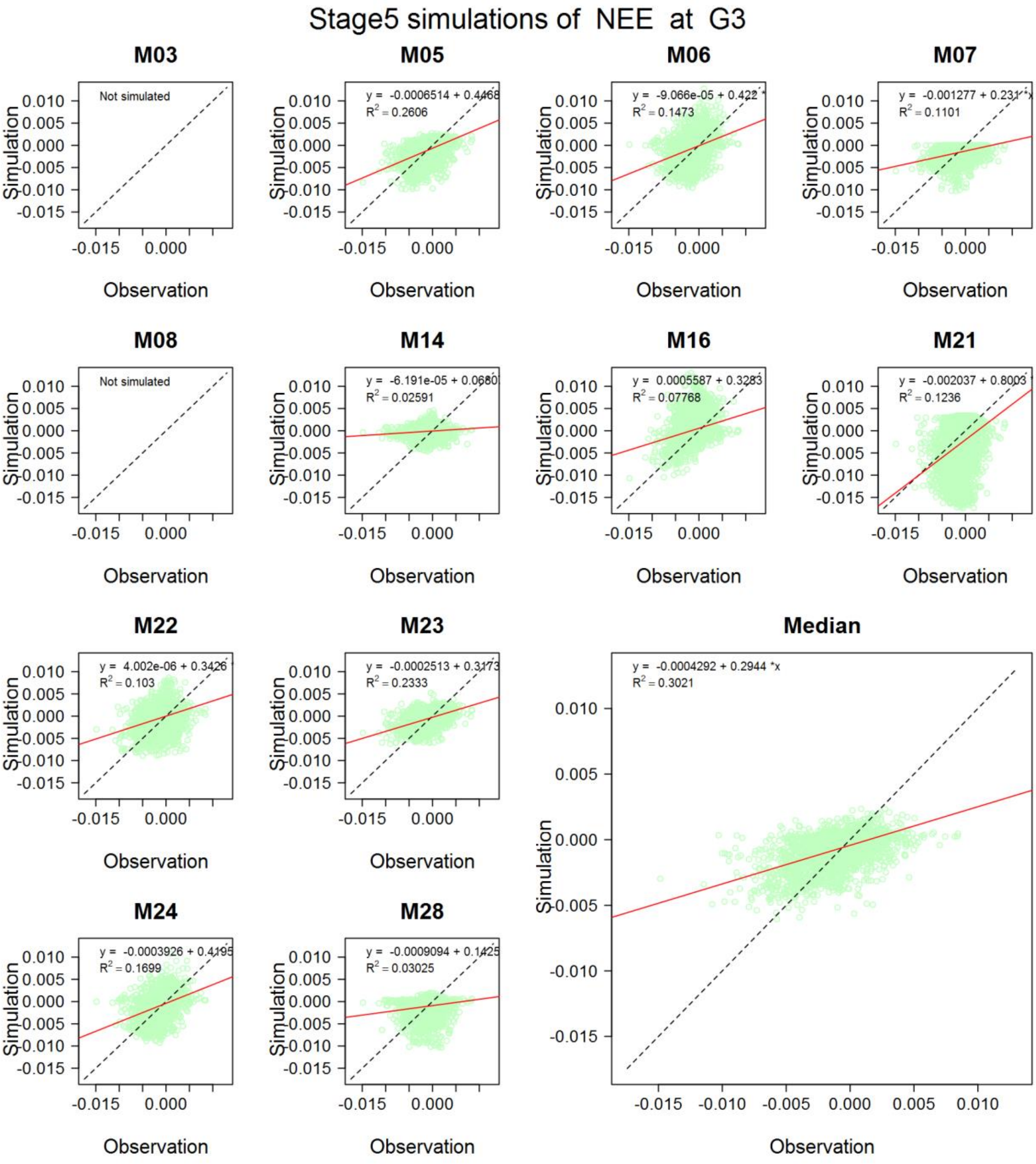

Fig. 5. S5 calibration stage: comparison of simulated (individual models and multi-model median) and observed daily net ecosystem exchange (NEE) data across multiple years at G3 site.

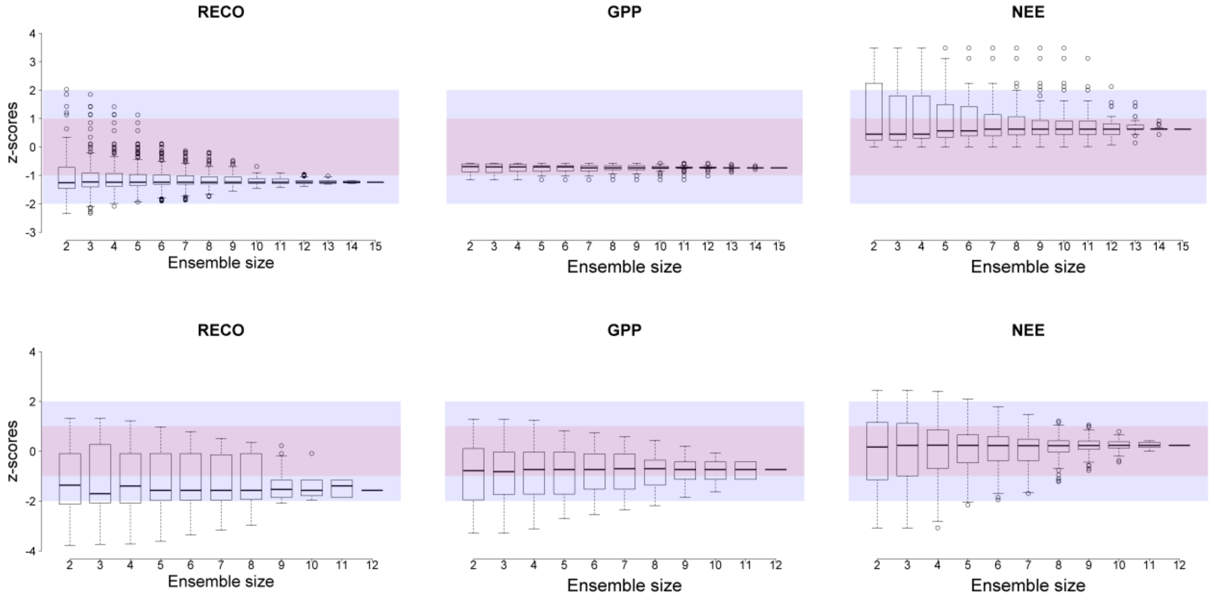

Fig. 6. *z*-scores for ecosystem respiration (RECO), gross primary production (GPP) and net ecosystem exchange (NEE) calculated with different ensemble sizes C1 crop site (top) and G3 grassland site (bottom), for calibration stage S5. Black lines show median values. Boxes delimit the 25th and 75th percentiles. Whiskers are 10th and 90th percentiles. Circles indicate outliers.