



HAL
open science

Ensemble modelling of carbon fluxes in grasslands and croplands

Renata Sandor, Fiona Ehrhardt, Peter Grace, Sylvie Recous, Pete Smith, Val Snow, Jean-François Soussana, Bruno Basso, Arti Bhatia, Lorenzo Brillì, et al.

► **To cite this version:**

Renata Sandor, Fiona Ehrhardt, Peter Grace, Sylvie Recous, Pete Smith, et al.. Ensemble modelling of carbon fluxes in grasslands and croplands. *Field Crops Research*, 2020, 10.1016/j.fcr.2020.107791 . hal-02905731

HAL Id: hal-02905731

<https://hal.inrae.fr/hal-02905731>

Submitted on 7 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



This document is a postprint version of an article published in Field Crops Research© Elsevier after peer review. To access the final edited and published work see <https://doi.org/10.1016/j.fcr.2020.107791>

Document downloaded from:



Please cite as:

Sandor et al. (2020) Ensemble modelling of carbon fluxes in grasslands and croplands. *Field Crops Research*. Accepted for publication 28 March 2020.

1

2 **Ensemble modelling of carbon fluxes in grasslands and croplands**

3 Renáta Sándor^{1,2}, Fiona Ehrhardt³, Peter Grace⁴, Sylvie Recous⁵, Pete Smith⁶, Val Snow⁷, Jean-
4 François Soussana³, Bruno Basso⁸, Arti Bhatia⁹, Lorenzo Brilli^{10,11}, Jordi Doltra¹², Christopher
5 D. Dorich¹³, Luca Doro^{14,15}, Nuala Fitton⁶, Brian Grant¹⁶, Matthew Tom Harrison¹⁷, Ute
6 Skiba¹⁸, Miko U.F. Kirschbaum¹⁹, Katja Klumpp¹, Patricia Laville²⁰, Joel Léonard²¹, Raphaël
7 Martin¹, Raia Silvia Massad²⁰, Andrew Moore²², Vasileios Myrgiotis²³, Elizabeth Pattey¹⁶,
8 Zhang Qing²⁴, Susanne Rolinski²⁵, Joanna Sharp²⁶, Ward Smith¹⁶, Lianhai Wu²⁷, Gianni
9 Bellocchi¹

10

11 ¹UCA, INRAE, VetAgro Sup, Unité Mixte de Recherche sur Écosystème Prairial (UREP),
12 63000 Clermont-Ferrand, France

13 ²Agricultural Institute, CAR HAS, 2462 Martonvásár, Hungary

14 ³INRAE, CODIR, 75007 Paris, France

15 ⁴Queensland University of Technology, Brisbane, Australia

16 ⁵Université de Reims Champagne Ardenne, INRA, FARE, 51100 Reims, France

17 ⁶Institute of Biological and Environmental Sciences, University of Aberdeen, UK

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

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

20 ⁹Indian Agricultural Research Institute, New Delhi, India

21 ¹⁰University of Florence, DISPAA, 50144 Florence, Italy

22 ¹¹IBIMET-CNR, 50145, Florence, Italy

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

25 ¹³NREL, Colorado State University, Fort Collins CO, USA

26 ¹⁴Desertification Research Group, University of Sassari, Sassari, Italy

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

41

42 **Abstract**

43 Croplands and grasslands are agricultural systems that contribute to land–atmosphere
44 exchanges of carbon (C). We evaluated and compared gross primary production (GPP),
45 ecosystem respiration (RECO), net ecosystem exchange ($NEE=RECO-GPP$) of CO_2 , and two
46 derived outputs - C use efficiency ($CUE=-NEE/GPP$) and C emission intensity ($Int_c=$ -
47 $NEE/Offtake$ [grazed or harvested biomass]). The outputs came from 23 models (11 crop-
48 specific, eight grassland-specific, and four models covering both systems) at three cropping
49 sites over several rotations with spring and winter cereals, soybean and rapeseed in Canada,
50 France and India, and two temperate permanent grasslands in France and the United Kingdom.
51 The models were run independently over multi-year simulation periods in five stages (S), either
52 blind with no calibration and initialization data (S1), using historical management and climate
53 for initialization (S2), calibrated against plant data (S3), plant and soil data together (S4), or
54 with the addition of C and N fluxes (S5). Here, we provide a framework to address
55 methodological uncertainties and contextualize results. Most of the models overestimated or
56 underestimated the C fluxes observed during the growing seasons (or the whole years for
57 grasslands), with substantial differences between models. For each simulated variable, changes
58 in the multi-model median (MMM) from S1 to S5 was used as a descriptor of the ensemble
59 performance. Overall, the greatest improvements (MMM approaching the mean of
60 observations) were achieved at S3 or higher calibration stages. For instance, grassland GPP
61 MMM was equal to $1632\text{ g C m}^{-2}\text{ yr}^{-1}$ (S5) while the observed mean was equal to $1763\text{ m}^{-2}\text{ yr}^{-1}$
62 (average for two sites). Nash-Sutcliffe modelling efficiency coefficients indicate that MMM
63 outperformed individual models in 91.4% of cases (S3 and S5). Our study suggests a cautious
64 use of large-scale, multi-model ensembles to estimate C fluxes in agricultural sites if some site-
65 specific plant and soil observations are available for model calibration. The further development
66 of crop/grassland ensemble modelling will hinge upon the interpretation of results in light of

67 the way models represent the processes underlying C fluxes in complex agricultural systems
68 (grassland and crop rotations including fallow periods).

69

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

72

73 **1. Introduction**

74 The global emissions of CO₂ in the atmosphere continue to increase together with impacts on
75 climate (IPCC, 2013). With the global carbon (C) balance becoming an issue of great societal
76 concern, process-based models are increasingly used to simulate biogeochemical processes
77 (such as plant photosynthesis and ecosystem respiration) occurring in both natural and managed
78 ecosystems, including agricultural systems (e.g. Brillì et al., 2017). These models use
79 approaches that determine the allocation of C from atmospheric CO₂ into plant biomass down
80 to the soil organic matter (van Oijen et al., 2014; Grosz et al., 2017; Kuhnert et al., 2017).
81 Process-based crop and grassland models (hereafter ‘models’) are important tools in
82 agricultural and environmental research to extrapolate local observations in time and space, and
83 to assess the impact of climate and agricultural practices on the functioning of soil-plant-
84 atmosphere systems (e.g. Jones et al., 2017a). They are largely used to represent current
85 understanding of the impacts of soil physical conditions such as soil temperature and water
86 content on soil processes such as net mineralisation and to estimate harvested phytomass (which
87 is the output of major significance in agricultural production). Climate-change impact
88 assessment studies have been conducted (at different places and scales) by forcing models with
89 global-to-local scale projected climate data (e.g. Ludwig and Asseng, 2006; Tingem et al., 2008;
90 Ruiz-Ramos and Mínguez, 2010; Graux et al., 2013; Vital et al., 2013; Zhang et al., 2017;
91 Mangani et al., 2018), to determine the vulnerability of agricultural systems to a changing
92 climate (e.g. Harrison et al., 2014, Lardy et al., 2014; Eza et al., 2015; Mangani et al., 2019).
93 Extensively tested biogeochemical models (with sub-models describing C cycling, generally
94 coupled to N cycling) are recognised as effective tools for studying the magnitude and spatial-
95 temporal patterns of C fluxes (Chang et al., 2015; Ma et al., 2015). They also play a prominent
96 role in testing the effect of specific changes in management, plant properties or environmental
97 factors (e.g. Kirschbaum et al., 2017), and for designing policies specific to the soil, climate,

98 and agricultural conditions of a location or region (e.g. Stocker et al., 2013). However, outputs
99 from different crop/grassland models often differ (e.g. Palosuo et al., 2011; Sándor et al., 2016),
100 thus leaving users with the question of deciding which model(s) they should use, and under
101 which circumstances presenting a range of possible impacts and adaptation responses. This has
102 led to a call for benchmarking actions at international level (Rosenzweig et al., 2013; Soussana
103 et al., 2015), where an estimation of the uncertainties associated with models is done by running
104 several models for the same system (ensemble modelling, e.g. Ehrhardt et al., 2018), which
105 generate envelopes of uncertainty, and help to identify avenues for model improvement (Jones
106 et al., 2017b; Challinor et al., 2018). Model inter-comparisons have been conducted using
107 datasets collected worldwide, with the involvement of different modelling communities and the
108 use of alternative simulation models (e.g. Martre et al., 2015; Sándor et al., 2017; Ehrhardt et
109 al., 2018). These studies indicate that there are substantial differences between models. Many
110 of the uncertainties regarding the simulation of crop and grassland processes can be attributed
111 to differences in the structure of these models (Brilli et al., 2017). While there has been a range
112 of published studies showing ensemble model simulation results for agricultural yield (e.g.
113 Asseng et al., 2013; Bassu et al., 2014; Li et al., 2015), there are fewer studies targeting C
114 dynamics (e.g. Smith et al., 1997; Kirschbaum et al., 2015; Basso et al., 2018; Puche et al.,
115 2019), and we are not aware of any published model intercomparison specifically assessing C
116 fluxes with multiple models across a range of different experimental sites. In this study, we
117 extended the analysis of the ensemble modelling performed by Ehrhardt et al. (2018) on
118 agricultural production and N₂O emissions via a multi-stage modelling protocol (from blind
119 simulations to partial and full calibration) by including a focus on C fluxes. We used a set of
120 23 biogeochemical models (11 cropland and eight grassland models, plus four models
121 simulating both crops and grasslands) and compared simulations with experimental data from
122 five sites (three crop rotations with spring and winter cereals, soybean and rapeseed, and two

123 temperate grasslands). Comparisons included gross primary production (GPP), ecosystem
 124 respiration (RECO), the carbon balance represented by net ecosystem exchange (NEE<0
 125 indicating net C uptake by the system) and other derived outputs. The models were calibrated
 126 through different stages with access to different levels of site-specific information. They were
 127 evaluated as a multi-model ensemble, with the aim of quantifying model uncertainties in the
 128 simulation of C fluxes at different sites and with different land uses.

129

130 **2. Materials and methods**

131 *2.1. Experimental sites and C measurements*

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

136

137 Table 1. Crop and grassland sites for the modelling exercise, years of available data and
 138 evaluated variables. GPP: gross primary production; RECO: ecosystem respiration; NEE: net
 139 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)

140

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

146

147 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

148

149 These sites provided high quality, previously published data encompassing climate, soil,
 150 agricultural practices, and C and N fluxes. They were either equipped with an eddy covariance
 151 system to determine the net ecosystem exchange (NEE) of CO₂ or with closed chambers for
 152 measuring respiration fluxes and automated weather stations for recording climatic conditions.
 153 The NEE data were partitioned into two main fluxes: gross primary production (GPP), which
 154 is the photosynthetic plant production from atmospheric CO₂, and ecosystem respiration

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

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

159

160 2.2. Models and simulation study

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

179 (S5) N₂O emission and soil organic C and N flux data (the full suite of measurements taken at
 180 respective sites).

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

187

188 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	http://www.imj.com.au/dm
	Ecomod v.5.3.1	https://www.pik-potsdam.de/research/projects/activities/biosphere-water-modelling/lpjml
	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)

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

194

195 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

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

198

199 2.3. Data analysis

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

$$210 \quad C_{am(s)} = \frac{\sum_{i=1}^{n_{meas}} C_{am(d)}}{n_{meas}} \cdot n_s$$

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

212 Two derived output variables were also analysed on seasonal basis for crops and on annual
 213 basis for grasslands, one representing C emission intensity and one C use efficiency. The
 214 potential to sustain or even increase crop/grassland yields is a desirable characteristic of any
 215 mitigation option both in terms of adoption of the technology by farmers (Vellinga et al., 2011)
 216 and its benefit in reducing GHG emissions per area of land and per unit of product, which is
 217 referred to as ‘emission intensity’ (van Groenigen et al. 2010). In this study, C emission
 218 intensity (Int_c) was calculated as the ratio between the amount of C emitted as CO₂ (C) and the

219 total amount of C in harvested agricultural production, that is, grain yield for crops and the
220 offtake (annual sum of animal intake and harvested aboveground biomass) for grasslands (after
221 Ehrhardt et al., 2018). Carbon use efficiency (CUE) was obtained as the ratio between CO₂-C
222 exchanged by the ecosystem and GPP (-NEE/GPP). A synthetic indicator such as the CUE is
223 useful to inform about the ability to retain part of GPP and thus increase total C content in the
224 agro-ecosystem (Sándor et al., 2016). The outputs analysed on seasonal/annual bases were also
225 presented on a daily basis as a practical way to compare models across contrasting locations.

226 We documented the variability of the multi-model simulation exercise across different
227 calibration stages, while inspecting how multi-model median (MMM) converged to the mean
228 of observations. For each simulated variable, we used box-plots to compare the variability of
229 estimates by different models (with focus on multi-year averages) to the observed variability,
230 and we represented model ensembles with MMM, which has the advantage to exclude distinctly
231 biased model members with a disproportionate influence on the mean (Rodríguez et al., 2019).

232 MMM is the median value of simulated data, which was calculated on daily outputs for each
233 stage. The advantage of using MMM was established on a theoretical basis and in practical
234 studies in crop and grassland modelling (Wallach et al., 2018). The absolute bias (best,
235 $0 \leq \text{ABIAS} < \infty$, worst) was calculated as an average of the absolute differences between MMM
236 estimates and means of observations at each season or year. Scatterplots of simulated versus
237 observed daily data and the modelling efficiency ($-\infty < \text{EF} \leq 1$, positive values indicating that
238 model estimates are more accurate than the mean of the observed data; Nash and Sutcliffe,
239 1970) were also provided to compare individual models and the MMM. Then, to explore how
240 MMM varied with the number of models in the ensemble we performed a calculation for each
241 z -score transformed MMM, $z = \frac{\text{MMM} - \bar{O}}{sd_{obs}}$, obtained by dividing the multi-model data deviation
242 from the mean of observations (\bar{O}) by the standard deviation of observations (sd_{obs}) (after
243 Ehrhardt et al., 2018). We calculated z -scores on all possible combinations of sets of k out of

244 $n=15$ models ($k=2, \dots n$). The minimum number of models providing plausible estimates at
245 each site was that for which z -scores were comprised between -2 and +2 (approximating the
246 95% confidence limit of a normal distribution).

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

249

250 **3. Results**

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

259

260 *3.1. Uncertainties and ensemble performance by land use*

261 Fig. A in the Supplementary material and Table 5 show the multi-model uncertainties (spread
262 of responses with different models) under different land uses (fallow, crop and grassland).
263 Observed mean RECO varied between $32 \text{ g C m}^{-2} \text{ yr}^{-1}$ (fallow) and $1561 \text{ g C m}^{-2} \text{ yr}^{-1}$ (grassland)
264 considering all calibration stages. The latter value is about three times higher than seasonal
265 observed crop values, e.g. maize ($674 \text{ g C m}^{-2} \text{ season}^{-1}$) or triticale ($553 \text{ g C m}^{-2} \text{ season}^{-1}$), as in
266 Table 5. Also, there is considerable difference between observed means and MMM RECO
267 values, e.g. for S5, 1561 vs. $1123 \text{ g C m}^{-2} \text{ yr}^{-1}$ for grasslands, 674 vs. $375 \text{ g C m}^{-2} \text{ season}^{-1}$ for
268 maize, 420 vs. $320 \text{ g C m}^{-2} \text{ season}^{-1}$ for spring wheat and 606 vs. $275 \text{ g C m}^{-2} \text{ season}^{-1}$ for

269 soybean. Overall, observed RECO was underestimated by the MMM in all stages and land uses.
270 The GPP MMM also showed high variability with winter wheat, triticale and maize (ranging
271 from 745 to 1354 g CO₂-C m⁻² season⁻¹ at S5 and S3, respectively), comparable with the
272 variability of grasslands across calibration stages (1061-1568 g C m⁻² yr⁻¹). There was also high
273 variability in estimated NEE with winter cereals and maize (Fig. A in the Supplementary
274 material and Table 5), but MMM generally approached observation means, e.g. maize mean
275 observation and MMM were -539 (S3) and -544 g C m⁻² season⁻¹ (S5), respectively. Model
276 estimates for grasslands showed less variability in NEE predictions (from -157 at S5 to -99 at
277 S1 compared to the observed mean of -219 g C m⁻² yr⁻¹). Seasonal CUE values (presented on
278 different scales for fallow, crop and grassland systems in Fig. A in the Supplementary material
279 and Table 5) were generally positive, with the exception of phacelia. Models tend to show
280 higher uncertainties towards negative values at early calibration stages, e.g. S1 of winter
281 cereals, maize, phacelia and rice. A lower uncertainty is associated with Int_C values, mainly
282 with grasslands. Some GPP (and Int_C) predictions were different from zero event under fallow
283 conditions (some non-zero biomass production was also observed experimentally).
284 The absolute bias (ABIAS), calculated by comparing the MMM and observed mean of different
285 output variables for different land uses, showed that we can expect an improvement of model
286 performances after S3, when vegetation and yield data were provided for calibration (Fig. 1).
287 For instance, GPP of maize, and RECO, GPP and NEE of spring wheat simulations show the
288 best fit at S3, while triticale and winter wheat show greater improvement at S4 and S5.

289

290 *3.1.1. Grassland systems*

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

294 1062 (S1) to 1568 (S5) $\text{g C m}^{-2} \text{ yr}^{-1}$. Overall, RECO predictions had a wider range in grasslands
295 than in crops (Fig. A in the Supplementary material and Table 5). Similar to GPP, models
296 mostly underestimated mean of RECO ($1561 \text{ g C m}^{-2} \text{ yr}^{-1}$), as predictions varied from 969 (S2
297 MMM) to 1248 (S1 MMM) $\text{g C m}^{-2} \text{ yr}^{-1}$ (the latter was similar to S3 MMM= $1235 \text{ g C m}^{-2} \text{ yr}^{-1}$
298 ¹). On the other hand, NEE and Int_C values were well estimated with MMM values lying within
299 the range of observations (-610 to $66 \text{ g C m}^{-2} \text{ yr}^{-1}$ and -0.18 to 2.54 yr^{-1} , respectively). In
300 addition, Int_C was near zero in grasslands. The models tended to underestimate CUE and to
301 slightly overestimate NEE. Best estimates (least difference between MMM and observation
302 mean) were obtained at S5 for both: NEE: -157.4 versus $-218.9 \text{ g C m}^{-2} \text{ yr}^{-1}$; CUE: 0.07 yr^{-1}
303 versus 0.11 yr^{-1} .

304

305 *3.1.2. Arable crops*

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

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

Output / Land-use	RECO			GPP			NEE			CUE			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

315

316 With wheat and triticale, simulations were lower than the measured RECO, whose means were
317 about 259 g C m⁻² season⁻¹ for winter wheat, 420 g C m⁻² season⁻¹ for spring wheat and 553 g
318 C m⁻² season⁻¹ for triticale (Fig. A in the Supplementary material and Table 5). Model
319 performances improved after S3, with MMM of 238 for winter wheat, 386 for spring wheat and
320 449 g C m⁻² season⁻¹ for triticale. All three cereal crops had negative NEE values, especially
321 under winter wheat and triticale. CUE was slightly overestimated under winter wheat. In
322 contrast, CUE MMM of spring wheat and triticale were within the range of observed CUEs at
323 higher calibration levels. CUE and Int_C showed a similar pattern of model variability.

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

332 All the models underestimated observed maize seasonal RECO and GPP values (661-1070 and
333 1102-1671 g C m⁻² season⁻¹, respectively), but model variability was limited for NEE, CUE and
334 Int_C (Table 5 and Fig. A in the Supplementary material). Fig. 1 shows a complex pattern of
335 ABIAS values, which were generally high at all calibration stages for RECO and GPP and even
336 increased at S4 for GPP and Int_C, while simulations and observations were closer for NEE and
337 CUE.

338 Overall, rapeseed was characterized by high variability in the observations: RECO: 69-660 g C
339 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
340 (Fig. A in the Supplementary material). The models tended to underestimate RECO and GPP

341 observations, and to overestimate NEE and CUE. Net C emissions were predicted against the
342 net C uptake reflected in measurements. Variations of simulated MMM Int_C values were within
343 the range of observations in spite of their high variability.

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

350 Among crops, phacelia showed the lowest uncertainty of RECO, GPP and NEE predictions.
351 Simulated MMM RECO (234-296 g C m⁻² season⁻¹) tended to underestimate the observed value
352 (327 g C m⁻² season⁻¹), in contrast to other outputs. With this crop, NEE was positive, which
353 indicated net C emission. The observed mean was ~98 g C m⁻² season⁻¹ and the MMM ranged
354 between 35 (S5) to 209 (S1) g C m⁻² season⁻¹.

355 With fallow, MMM RECO predictions were within the range of observations that ranged
356 between 16 and 161 g C m⁻² season⁻¹. Observed and simulated GPP values were close to zero
357 and the simulations were within the range of variation of the measurements (5.8-31 g C m⁻²
358 season⁻¹). NEE values showed the second highest positive simulated and observed values after
359 phacelia on a seasonal basis (MMM predictions were within the range of measurements: 15 and
360 130 g C m⁻² season⁻¹, Fig. A in the Supplementary material). The observed CUE values were
361 the lowest (Fig. A in the Supplementary material) while the ABIAS was the highest
362 (ABIAS_{CUE}=4.26 season⁻¹, Fig. 1). The observed variability, between -0.16 and -0.04 season⁻¹,
363 was reflected in the model simulations.

364

365 *3.2. Uncertainties and ensemble performance by site*

366 Overall, RECO and GPP were underestimated at grassland sites (Fig. 2). Mean observed RECO
367 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
368 predictions varied from 716 to 1262 and from 1057 to 1457 g C m⁻² yr⁻¹, respectively.
369 Improvements were observed at S3, and best predictions were obtained at S5, especially at G4
370 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
371 crop sites, we observed some considerable improvements after S2, e.g. with S3 showing the
372 best estimates of RECO, where the MMM and observed mean were very similar 241 and 242
373 g C m⁻² season⁻¹ (average for C1, C2 and C3) (Fig. 2.).

374

375 3.2.1. C1

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

382

383 3.2.2. C2

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

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

389

390 3.2.3. C3

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

394

395 3.2.4. G3

396 GPP and RECO observations did not vary as much at this site as the model predictions. Fig. 2
397 shows that the accuracy of GPP predictions tended to increase through the calibration stages,
398 with RECO showing best estimates at S3. For instance, at S5, MMM ($1775 \text{ g C m}^{-2} \text{ yr}^{-1}$) was
399 close to the observed mean ($1898 \text{ g C m}^{-2} \text{ yr}^{-1}$). The MMM values of NEE, CUE and Intc
400 showed slight differences for different calibration stages, but an improvement was observed at
401 S5 ($209 \text{ g C m}^{-2} \text{ yr}^{-1}$, 0.11 yr^{-1} , 0.54 yr^{-1} , respectively) compared with the observations (-248 g
402 $\text{C m}^{-2} \text{ yr}^{-1}$, 0.13 yr^{-1} , 0.62 yr^{-1} , respectively). G3 showed a high C uptake (observed means
403 $\text{NEE}=-248$ versus MMM at S5= $-209 \text{ g C m}^{-2} \text{ yr}^{-1}$).

404

405 3.2.5. G4

406 At this site, the ranges of variation of RECO and GPP observations were similar to G3.
407 Observed GPP ($1767 \text{ g C m}^{-2} \text{ yr}^{-1}$) was generally underestimated by the models (ranging from
408 1255 to $1490 \text{ g C m}^{-2} \text{ yr}^{-1}$), but the MMM of RECO at S5 ($1457 \text{ g C m}^{-2} \text{ yr}^{-1}$) approached the
409 mean of observations ($1537 \text{ g C m}^{-2} \text{ yr}^{-1}$). The MMM of NEE at S5 ($-110 \text{ g C m}^{-2} \text{ yr}^{-1}$) was also
410 close to the observation mean ($-148 \text{ g C m}^{-2} \text{ yr}^{-1}$). For CUE, the positive values of both MMM

411 (ranging from 0.03 to 0.13 yr⁻¹) and observation mean (0.12 yr⁻¹) reflected the C uptake at this
412 grassland site.

413

414 |3.3. Individual models versus multi-model ensemble

415 Daily comparisons were not straightforward in this study because discontinuous observations
416 were tied to specific days, but the models did not have access to the diurnal pattern of the
417 processes (e.g. timing of specific weather or management events). With this caveat in mind, for
418 interpretation, we plotted simulated versus observed daily C fluxes as a visualisation tool to
419 compare the model ensemble results with individual model results. The scatterplots of Figs. 3,
420 4 and 5 and Figs. B-J in the supplement, are examples for GPP, RECO and NEE at the S5
421 calibration stage for the G3 grassland site, of the comparison of the performances of individual
422 models and MMM values. Consistent with the findings above, the MMM outperformed most
423 of the individual models. Considering R² values and alignments with the 1:1 lines, this was the
424 case for nine out of 10 models and seven out of 11 models simulating GPP and RECO. When,
425 in a few cases, individual models provided relatively satisfactory results, this was generally true
426 for one output but not for another. For example, M21 provided satisfactory results for GPP (Fig.
427 3) but not for RECO (Fig. 4). M16 (which was calibrated according to an automatic technique)
428 was distinctly outperforming the MMM for both GPP (Fig. 3) and RECO (Fig. 4) estimates, but
429 underperformed for other outputs, e.g. NEE (Fig. 5). Similar patterns of results were obtained
430 at the S3 (Supplementary material, Figs. B-D) and other calibrations stages (data not shown),
431 and for the G4 site (data not shown). Likewise for croplands, the MMM tended to outperform
432 individual models, e.g. for GPP, RECO and NEE at C2 site (Figs. E to J in supplementary
433 material for calibration stages S3 and S5).

434

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

439

440 Table 6. Nash-Sutcliffe modelling efficiency (EF) values for C-flux outputs (as in Table 1)
 441 provided by different models (as in Table 4) at S3 and S5 calibration stages at cropland (C1,
 442 C2, C3) and grassland (G3, G4) sites. Grey cells indicate that output variables were neither
 443 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
	RECO	0.03	0.01	-3.53	0.17	0.25
S5	GPP	0.32	0.58		0.53	0.62
	NEE	0.22	0.55		0.30	0.45

444

445 3.4. Minimum ensemble size

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

467

468 **4. Discussion**

469 The results in this paper show that the suitability of a multi-model ensemble to simulate
470 agricultural C fluxes depends on the variables being collected to calibrate models. With respect
471 to emission-related processes, up to recently it has been considered that it is “premature to fully
472 trust model outputs as representing reality” (Oertel et al., 2016, p. 344). In our exercise (which
473 is the first on agricultural C fluxes), we provided an update on what we can reasonably expect
474 from using an ensemble of biogeochemical models. These results reinforced the idea that at
475 large-scale, multiple model ensembles could be a promising way to orient future modelling
476 studies, with plant and soil observations as a minimum data requirement for model calibration
477 (S3 and S4). Additional observations (such as C-N fluxes) might not be needed for a more
478 detailed model calibration (e.g. results for S5 in Fig. 2). For instance, the use of N₂O emission
479 data for calibration could increase the uncertainty of model estimates (e.g., Del Grosso et al.,
480 2011; Hense et al., 2013), considering the high spatial and temporal variability associated with
481 heterogeneous and intermittent N₂O emissions (e.g. Grant and Pattey, 2003). Unlikely our
482 results have been affected by the different calibration techniques used. In fact, Wallach et al.
483 (2019) showed that different calibration techniques do not seem to be primarily responsible for
484 differences in model performance, and considering that most of the modelling teams derived
485 parameter values based on a manual trial-and-error approach (Table 4). When several
486 (differently packaged) models and complex datasets are mobilised in large-scale multi-model
487 ensembles, the uncertainty in calibrated parameters tends to be confounded with the uncertainty
488 in model structure (Wallach and Thorburn, 2017). Usually, calibration techniques are
489 considered a lower priority in agricultural ensemble modelling, where the reduction of
490 uncertainties is mostly limited by the limited quality of the calibration data (e.g. Angulo et al.,
491 2003; Maiorano et al., 2017). However, each situation can be so unique (e.g. supplied data are

492 incorrectly measured or are affected by unreported factors, such as pest damage) that generic
493 lessons cannot be drawn from this whole exercise. During the course of this exercise, some
494 modelling teams noticed model structural problems, which could later be resolved.

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

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

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

523 Both observed and simulated NEE showed negative values (net C uptake), with the exception
524 of fallow periods and the phacelia growing season. It is known that phacelia, as a cover crop,
525 may increase soil CO₂ emission due to an enhanced input of organic residues (e.g. Bodner et
526 al., 2018). The lowest values were associated with maize, winter wheat, rice and triticale crops.
527 However, models could underestimate NEE values from the whole crop rotation system (e.g.
528 C2 site), because they underestimate the release of CO₂ to the atmosphere from fallow periods.
529 This means that models need to improve their simulations of bare soil processes during the
530 intercrop period. However, improvements in model predictions were observed after the S3
531 calibration stage. The C uptake (NEE<0) observed and modelled in C1 and C3 crop rotations
532 did not include fallow periods for which measurements were not made available, thus NEE
533 values were only for the crop growing seasons. In both grassland sites G3 and G4, model results
534 reflected the limited variability of NEE observations, which was roughly half those of RECO
535 and GPP. Thus, with NEE, some performance gain was obtained from the uncertainty
536 compensation.

537 Higher CUE promotes biomass accumulation and, indirectly, C stabilization in soil layers,
538 while lower CUE favours respiration and C losses (Bradford and Crowther, 2013; Geyer et al.,
539 2019). In our site-by-site analysis, CUE values were generally better estimated after S3
540 calibration stage. Among crops, phacelia and soybean showed the highest variability in their

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

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

546 Overall, the MMM provided more accurate simulations in most cases than individual models
547 (as shown by the regression lines of Figs. 3-5 and Figs B-J in the Supplementary material).
548 Even though some individual models were outperforming the MMM (e.g. M6, M16, M22, M23,
549 M24) in certain cases (outputs/sites/calibration stages), that response was not general (e.g.
550 Table 6). We confirm with this study that it is difficult to define an *a priori* criterion that could
551 be used to select a subset of models that would perform better than others would. In terms of
552 minimum number of models required to obtain reliable results, our study indicates that the
553 suggested minimum ensemble size (~10 models) proposed by Martre et al. (2015) for crop
554 growth should be increased (at least 13 models) when model ensembles are implemented to
555 simulate C fluxes at different climatic regions worldwide. Only in specific situations, e.g. C2
556 site, ~9 models could provide reliable C-flux estimates. With grasslands, the minimum
557 ensemble size should include at least 11 models.

558 **5. Summary and conclusions**

559 This study presents a framework for interpretation of model performance and uncertainties
560 obtained with a set of biogeochemical models (individually and in an ensemble) simulating C
561 fluxes in cropping and grassland systems at a variety of distant and contrasted sites. There are
562 multiple foci when designing multi-model studies of agricultural systems (such as crop rotations
563 and grasslands) depending on the questions to be answered. Our study shows that we could not
564 identify the best model(s) for crop and grassland C fluxes and no probability of success could
565 be assigned to prove the suitability of using one biogeochemical model rather than another. We

566 demonstrate the potential that a multi-model ensemble can have for jointly estimating different
567 C fluxes (primary production, ecosystem respiration and net ecosystem exchanges) and
568 production-scaled emissions (e.g. CO₂-C emission intensities and C use efficiencies).
569 We showed that reduced calibration datasets (vegetation data) could be adequate for providing
570 sufficiently reliable outputs (e.g. to continue to progress towards updating the inventory of C
571 databases, West et al., 2010), but additional biophysical and biogeochemical data can further
572 improve results under certain circumstances. Further improvements of data sources, such as
573 phenological observations, could help refine model estimates and form a baseline for screening
574 agricultural practices and mitigation options at croplands and grasslands, as presented in Sándor
575 et al. (2018). Moreover, there is a high uncertainty of modelled fluxes during fallow periods,
576 which would need more accurate data.

577 These results paved the way for using model ensemble medians for field-scale estimation of C
578 fluxes. Our results inform about the possible use of model ensembles for upscaling projections
579 of C fluxes and derived outputs, from field scale to larger spatial units (e.g. gridded projections)
580 as needed for Tier 3 national inventories (e.g. Folberth et al., 2016; Zscheischler et al., 2017).
581 However, model inter-comparisons have their limitations. Although our comparison was large
582 compared to other studies (e.g. Sándor et al., 2016), there was a lack of case studies in this
583 exercise from Africa, South America and Oceania, which would extend the geographical
584 coverage. Our study-sites mostly targeted agricultural areas of the Northern hemisphere (four
585 temperate and one tropical), as part of a broader study covering more agricultural areas in both
586 hemispheres (Ehrhardt et al., 2018).

587 Moreover, the various model types and variants evaluated here did not cover all the modelling
588 approaches used to simulate C fluxes from crop and grassland systems (e.g. the model used by
589 Senapati et al., 2016). They reasonably represent current approaches (the basis of development
590 and processes were scrutinized), but we think that crop and grassland model inter-comparisons

591 with the inclusion of more models should be continued to assess and improve our ability to
592 simulate biogeochemical processes with acceptable quality. Further analyses and better
593 understanding of these multi-model ensembles are required to achieve key progress in crop and
594 grassland modelling, by assessing more in-depth model responses and uncertainties against
595 climate and management drivers.

596

597 **Acknowledgements**

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

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

609

610 **References**

611 Allard, V., Soussana, J.-F., Falcimagne, R., Berbigier, P., Bonnefond, J.M., Ceschia, E.,
612 D’hour, P., Hénault, C., Laville, P., Martin, C., Pinarès-Patino, C., 2007. The role of grazing
613 management for the net biome productivity and greenhouse gas budget (CO₂, N₂O and CH₄)
614 of semi-natural grassland. *Agr. Ecosyst. Environ.* 12, 47-58.
615 <https://doi.org/10.1016/j.agee.2006.12.004>.

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

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

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

631 Basso, B., Dumont, B., Maestrini, B., Shcherbak, I., Robertson, G.P., Porter, J.R., Smith, P.,
632 Paustian, K., Grace, P.R., Asseng, S., Bassu, S., Biernath, C., Boote, K.J., Cammarano, D.,
633 De Sanctis, G., Durand, J.-L., Ewert, F., Gayler, S., Hyndman, D.W., Kent, J., Martre, P.,
634 Nendel, C., Priesack, E., Ripoche, D., Ruane, A.C., Sharp, J., Thorburn, P.J., Hatfield, J.L.,
635 Jones, J.W., Rosenzweig, C., 2018. Soil organic carbon and nitrogen feedbacks on crop
636 yields under climate change. *Agricultural and Environmental Letters* 3: 180026. doi:
637 10.2134/ael2018.05.0026. Bassu, S., Brisson, N., Durand, J.L., Boote, K., Lizaso, J., Jones,
638 J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard,
639 H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield,
640 J., Hoek, S., Izaurrealde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.H.,

641 Kumar, N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F.,
642 Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop
643 models vary in their responses to climate change factors? *Global Change Biol.* 20, 2301-
644 2320. <https://doi.org/10.1111/gcb.12520>.

645 Bhatia, A., Pathak, H., Jain, N., Singh, P.K., Tomer, R., 2012. Greenhouse gas mitigation in
646 rice-wheat system with leaf color chart-based urea application. *Environ. Monit. Assess.* 184,
647 3095-3107. <https://doi.org/10.1007/s10661-011-2174-8>.

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

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

654 Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C.D., Doro, L.,
655 Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I.,
656 Klumpp, K., Léonard, J., Martin, R., Massad, R.S., Recous, S., Seddaiu, G., Sharp, J., Smith,
657 P., Smith, W.N., Soussana, J.-F., Bellocchi, G., 2017. Review and analysis of strengths and
658 weaknesses of agro-ecosystem models for simulating C and N fluxes. *Sci. Total Environ.*
659 598, 445-470. <https://doi.org/10.1016/j.scitotenv.2017.03.208>.

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

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

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

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

672 Confalonieri, R., Bregaglio, S., Acutis, M., 2016. Quantifying uncertainty in crop model
673 predictions due to the uncertainty in the observations used for calibration. *Ecol. Model.* 328,
674 72-77. <https://doi.org/10.1016/j.ecolmodel.2016.02.013>.

675 Curtin, D., Wang, H., Selles, F., Mcconkey, B.G., Campbell, C.A., 2000. Tillage effects on
676 carbon fluxes in continuous wheat and fallow-wheat rotations. *Soil Sci. Soc. Am. J.* 64,
677 2080-2086. <https://doi.org/10.2136/sssaj2000.6462080x>.

678 De Martonne, E., 1942. Nouvelle carte mondiale de l'indice d'aridité. *Annales de Géographie*
679 51, 242–250. (in French)

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

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

685 Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., McAuliffe, R., Recous, S., Sándor, R.,
686 Smith, P., Snow, V., Migliorati, M.D.A., Basso, B., Bhatia, A., Brillì, L., Doltra, J., Dorich,
687 C.D., Doro, L., Fitton, N., Giacomini, S.J., Grant, B., Harrison, M.T., Jones, S.K.,
688 Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Liebìg, M., Lieffering, M.,
689 Martin, R., Massad, R.S., Meier, E., Merbold, L., Moore, A.D., Myrghiotis, V., Newton, P.,
690 Pattey, E., Rolinski, S., Sharp, J., Smith, W.N., Wu, L., Zhang, Q., 2018. Assessing

691 uncertainties in crop and pasture ensemble model simulations of productivity and N₂O
692 emissions. *Global Change Biol.* 24, e603-e616. <https://doi.org/10.1111/gcb.13965>.

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

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

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

702 Grant, R., Pattey, E., 2003. Modelling variability in N₂O emissions from fertilized agricultural
703 fields. *Soil Biol. Biochem.* 35, 225-243. [https://doi.org/10.1016/S0038-0717\(02\)00256-0](https://doi.org/10.1016/S0038-0717(02)00256-0).

704 Graux, A.-I., Bellocchi, G., Lardy, R., Soussana, J.-F., 2013. Ensemble modelling of climate
705 change risks and opportunities for managed grasslands in France. *Agr. Forest Meteorol.* 170,
706 114-131. <https://doi.org/10.1016/j.agrformet.2012.06.010>.

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

714

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

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

722 Hense, A., Skiba, U., Famulari, D., 2013. Low cost and state of the art methods to measure
723 nitrous oxide emissions. *Environ. Res. Lett.* 8, 025022. [https://doi.org/10.1088/1748-](https://doi.org/10.1088/1748-9326/8/2/025022)
724 [9326/8/2/025022](https://doi.org/10.1088/1748-9326/8/2/025022).

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

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

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

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

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

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

749 Kirschbaum, M.U.F., Schipper, L.A., Mudge, P.L., Rutledge, S., Puche, N.J.B., Campbell, D.I.,
750 2017. The trade-offs between milk production and soil organic carbon storage in dairy
751 systems under different management and environmental factors. *Sci. Total Environ.* 577, 61-
752 72. <https://doi.org/10.1016/j.scitotenv.2016.10.055>.

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

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

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

765 Laville, P., Lehuger, S., Loubet, B., Chaumartin, F., Cellier, P., 2011. Effect of management,
766 climate and soil conditions on N₂O and NO emissions from an arable crop rotation using
767 high temporal resolution measurements. *Agr. Forest Meteorol.* 151, 228-240.
768 <https://doi.org/10.1016/j.agrformet.2010.10.008>.

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

775 Loubet, B., Laville, P., Lehuger, S., Larmanou, E., Flechard, C., Mascher, N., Genermont, S.,
776 Roche, R., Ferrara, R. M., Stella, P., Personne, E., Durand, B., Decuq, C., Flura, D., Masson,
777 S., Fanucci, O., Rampon, J.-N., Siemens, J., Kindler, R., Gabrielle, B., Schrumpf, M.,
778 Cellier, P., 2011. Carbon, nitrogen and greenhouse gases budgets over a four years crop
779 rotation in northern France. *Plant Soil* 343, 109-137. [https://doi.org/10.1007/s11104-011-](https://doi.org/10.1007/s11104-011-0751-9)
780 [0751-9](https://doi.org/10.1007/s11104-011-0751-9).

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

784 Ma, S., Lardy, R., Graux, A.-I., Ben Touhami, H., Klumpp, K., Martin, R., Bellocchi, G., 2015.
785 Regional-scale analysis of carbon and water cycles on managed grassland systems. *Environ.*
786 *Modell. Softw.* 72, 356-371. <https://doi.org/10.1016/j.envsoft.2015.03.007>.

787 Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R.P., Ruane, A.C., Semenov,
788 M.A., Wallach, D., Wang, E., Alderman, P.D., Kassie, B.T., Biernath, C., Basso, B.,
789 Cammarano, D., Challinor, A.J., Doltra, J., Dumont, B., Rezaei, E.E., Gayler, S.,

790 Kersebaum, K.C., Kimball, B.A., Koehler, A.-K., Liu, B., O'Leary, G., Olesen, J.E., Ottman,
791 M.J., Priesack, E., Reynolds, M., Stratonovitch, P., Streck, T., Thorburn, P.J., Waha, K.,
792 Wall, G.W., White, J.W., Zhao, Z., Zhu, Y., 2017. Crop model improvement reduces the
793 uncertainty of the response to temperature of multi-model ensembles. *Field Crop. Res.* 202,
794 5-20. <https://doi.org/10.1016/j.fcr.2016.05.001>.

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

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

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

812 Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - A
813 discussion of principles. *J. Hydrol.* 10, 282-290. [https://doi.org/10.1016/0022-](https://doi.org/10.1016/0022-1694(70)90255-6)
814 [1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).

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

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

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

826 Puche, N.J.B., Senapati, N., Flechard, C.R., Klumpp, K., Kirschbaum, M.U.F, Chabbi, A.,
827 2019. Modelling carbon and water fluxes of managed grasslands: comparing flux variability
828 and net carbon budgets between grazed and mowed systems. *Agronomy* 9, 183.
829 <https://doi.org/10.3390/agronomy9040183>.

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

838 Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M.,
839 Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D.,

840 Baigorria, G., Winter, J.M., 2013. The agricultural model intercomparison and improvement
841 project (AgMIP): protocols and pilot studies. *Agr. Forest Meteorol.* 170, 166-182.
842 <https://doi.org/10.1016/j.agrformet.2012.09.011>.

843 Ruane, A.C., Hudson, N.I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., Boote, K.J.,
844 Thorburn, P.J., Aggarwal, P.K., Angulo, C., Basso, D., Bertuzzi, P., Biernath, C., Brisson,
845 N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J.,
846 Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Kumar, S.N., Nendel, C.,
847 O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Rötter,
848 R.P., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T.,
849 Supit, I., Travasso, M., Waha, K., Wallach, D., White, J.W., Wolf, J., 2016. Multi-wheat
850 model ensemble responses to interannual climate variability. *Environ. Modell. Softw.* 81,
851 86-101. <https://doi.org/10.1016/j.envsoft.2016.03.008>.

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

855 Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E.,
856 Ma, S., Perego, A., Rolinski, S., Ruget, F., Sanna, M., Seddaiu, G., Wu, L., Bellocchi, G.,
857 2017. Multi-model simulation of soil temperature, soil water content and biomass in Euro-
858 Mediterranean grasslands: uncertainties and ensemble performance. *Eur. J. Agron.* 88, 22-
859 40. <https://doi.org/10.1016/j.eja.2016.06.006>.

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

863 Sándor, R., Ehrhardt, F., Basso, B., Bellocchi, G., Bhatia, A., Brilli, L., De Antoni Migliorati,
864 M., Doltra, J., Dorich, C., Doro, L., Fitton, N., Giacomini, S. J., Grace, P., Grant, B.,

865 Harrison, M. T., Jones, S., Kirschbaum M. U. F., Klumpp, K., Laville, P., Léonard, J., Liebig,
866 M., Lieffering, M., Martin, R., McAuliffe, R., Meiser, E., Merbold, L., Moore, A., Myrghiotis,
867 V., Newton, P., Pattey, E., Recous, S., Rolinski, S., Sharp, J., Massad, R. S., Smith, P., Smith,
868 W., Snow, V., Wu, L., Zhang, Q., Soussana, J.-F., 2016. C and N models intercomparison -
869 benchmark and ensemble crop and grassland model estimates for grassland production.
870 *Advances in Animal Biosciences* 7, 227-228. <https://doi.org/10.1017/S2040470016000297>.

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

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

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

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

888 Smith, P., Smith, J.U., Powlson, D.S., McGill, W.B., Arah, R.M., Chertov, O.G., Coleman, K.,
889 Franko, U., Frohling, S., Jenkinson, D.S., Jensen, L.S., Kelly, R.H., Klein-Gunnewiek, H.,

890 Komarov, A.S., Li, C., Molina, J.A.E., Mueller, T., Parton, W.J., Thornley, J.H.M.,
891 Whitmore, A.P., 1997. A comparison of the performance of nine soil organic matter models
892 using datasets from seven long-term experiments. *Geoderma* 81, 153-225.
893 [https://doi.org/10.1016/S0016-7061\(97\)00087-6](https://doi.org/10.1016/S0016-7061(97)00087-6).

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

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

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

906

907 van Groenigen, J.W., Velthof, G.L., Oenema, O., van Groenigen, K. J., Kessel, C.V., 2010.
908 Towards an agronomic assessment of N₂O emissions: a case study for arable crops.
909 *European Journal of Soil Science* 61, 903–913. [https://doi.org/10.1111/j.1365-](https://doi.org/10.1111/j.1365-2389.2009.01217.x)
910 [2389.2009.01217.x](https://doi.org/10.1111/j.1365-2389.2009.01217.x).van Oijen, M., Balkovi, J., Beer, C., Cameron, DR., Ciais, P., Cramer,
911 W., Kato, T., Kuhnert, M., Martin, R., Mynemi, R., Ramming, A., Rolinski, S., Soussana, J-
912 F, Thonicke, K., van der Velde, M., Xu, L., 2014. Impact of droughts on the carbon cycle in
913 European vegetation: a probabilistic risk analysis using six vegetation models.
914 *Biogeosciences* 11, 6357-6375. <https://doi.org/10.5194/bg-11-6357-2014>.

915 Vellinga, T.V., de Haan, M.H.A., Schils, R.L.M., Evers, A., van den Pol-van Dasselaar, A.,
916 2011. Implementation of GHG mitigation on intensive dairy farms: Farmers' preferences
917 and variation in cost effectiveness. *Livestock Science* 137, 185–195.
918 <https://doi.org/10.1016/j.livsci.2010.11.005>.

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

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

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

934 Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., van Ittersum, M.,
935 Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De
936 Sanctis, G., Dumont, B., Rezaei, E.E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila,
937 M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C., Jones, C.D.,
938 Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A.-K., Maiorano, A., Minoli, S., Müller,
939 C., Kumar, S.N., Nendel, C., O'Leary, G.J., Palosuo, T., Priesack, E., Ripoche, D., Rötten,

940 R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Fao, F., Wolf, J.,
941 Zhang, Z., 2018. *Global Change Biology* 24, 5072-5083.

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

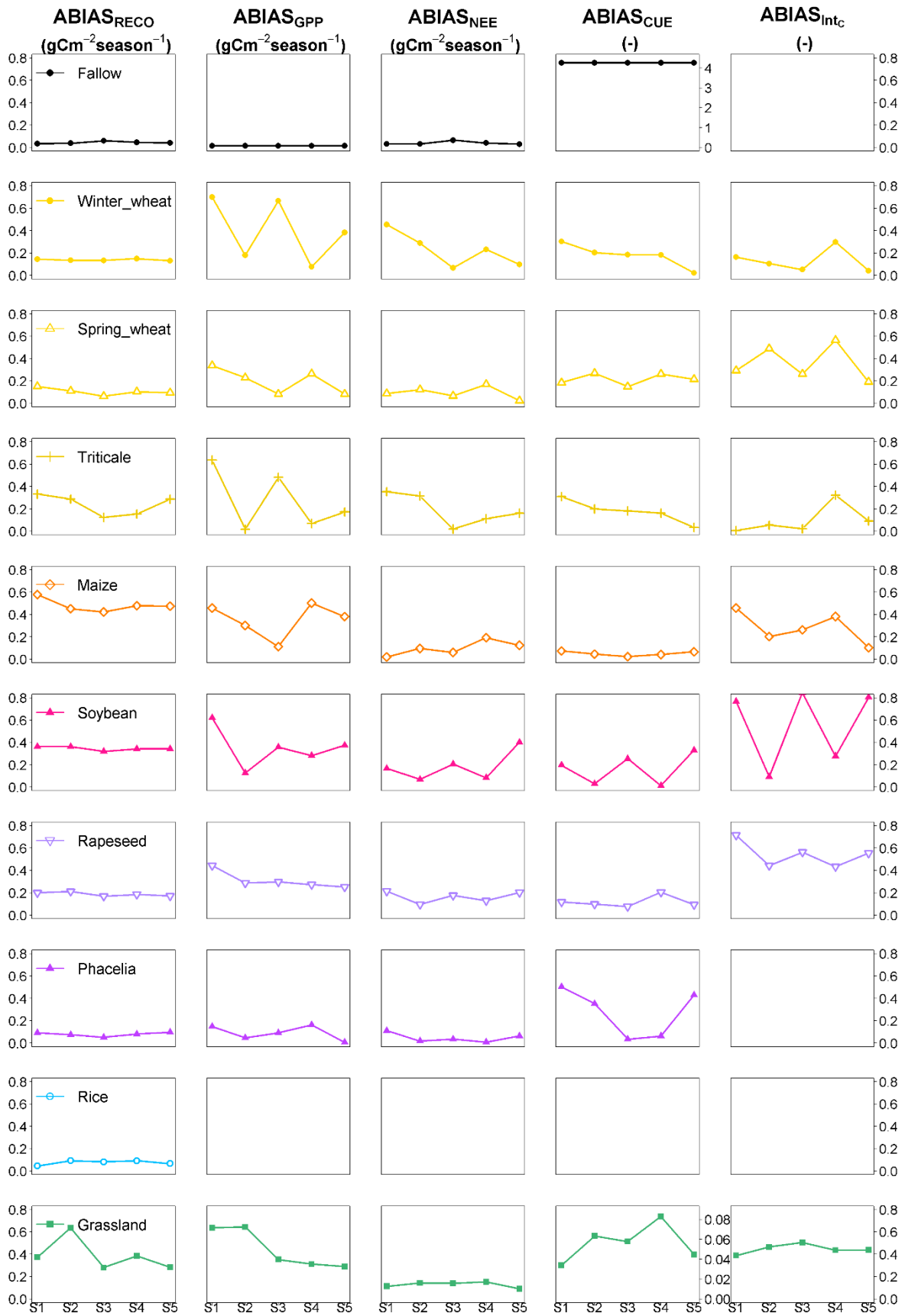
946 Zhang, W., Zhang, F., Qi, J., Hou, F., 2017. Modeling impacts of climate change and grazing
947 effects on plant biomass and soil organic carbon in the Qinghai–Tibetan grasslands.
948 *Biogeosciences* 14, 5455-5470. <https://doi.org/10.5194/bg-14-5455-2017>.

949 Zscheischler, J., Mahecha, M. D., Avitabile, V., Calle, L., Carvalhais, N., Ciais, P., 2017.
950 Reviews and syntheses: An empirical spatiotemporal description of the global surface–
951 atmosphere carbon fluxes: opportunities and data limitations. *Biogeosciences* 14, 3685-
952 3703. <https://doi.org/10.5194/bg-14-3685-2017>.

953

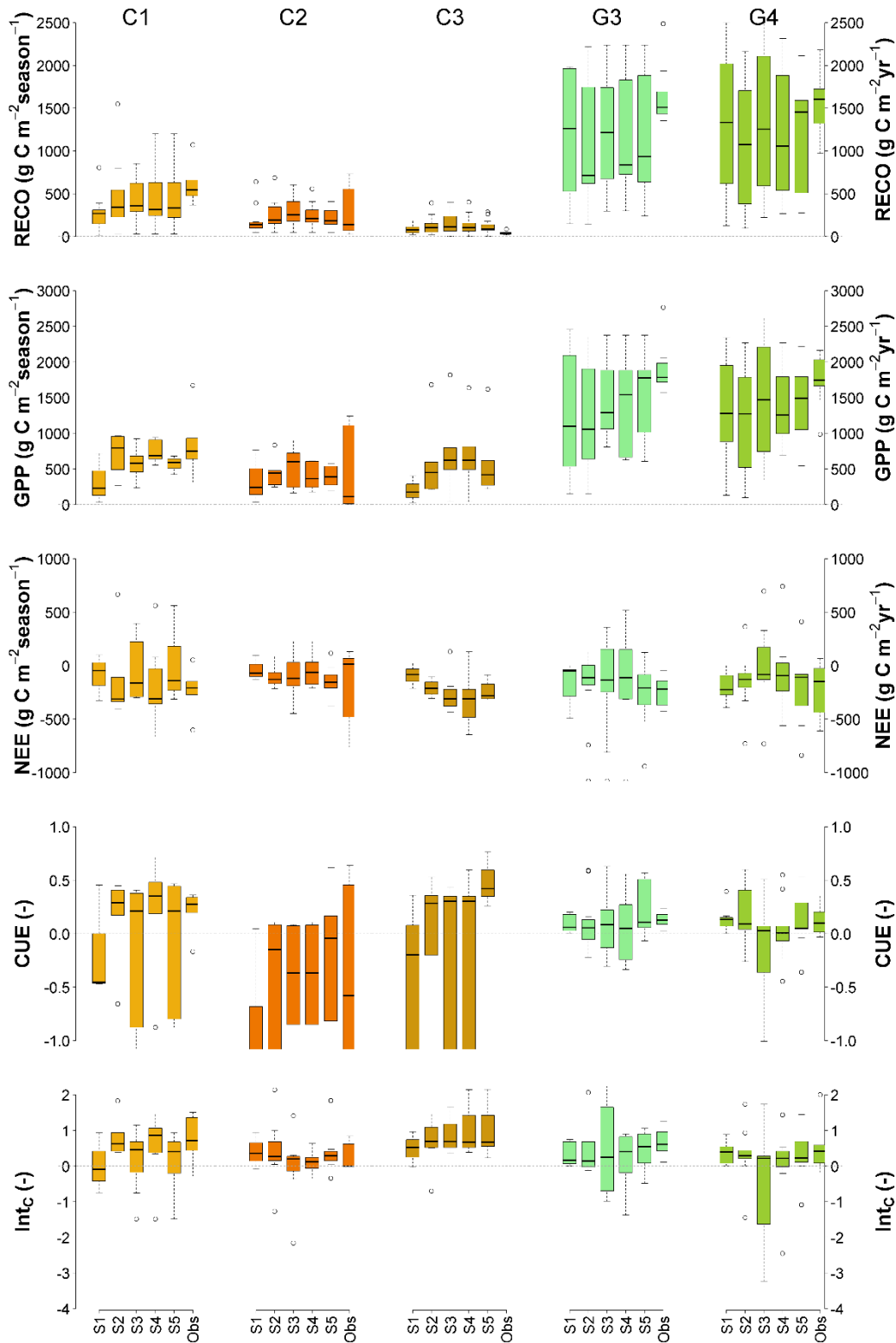
954 **Figure legends**

955



956

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



961

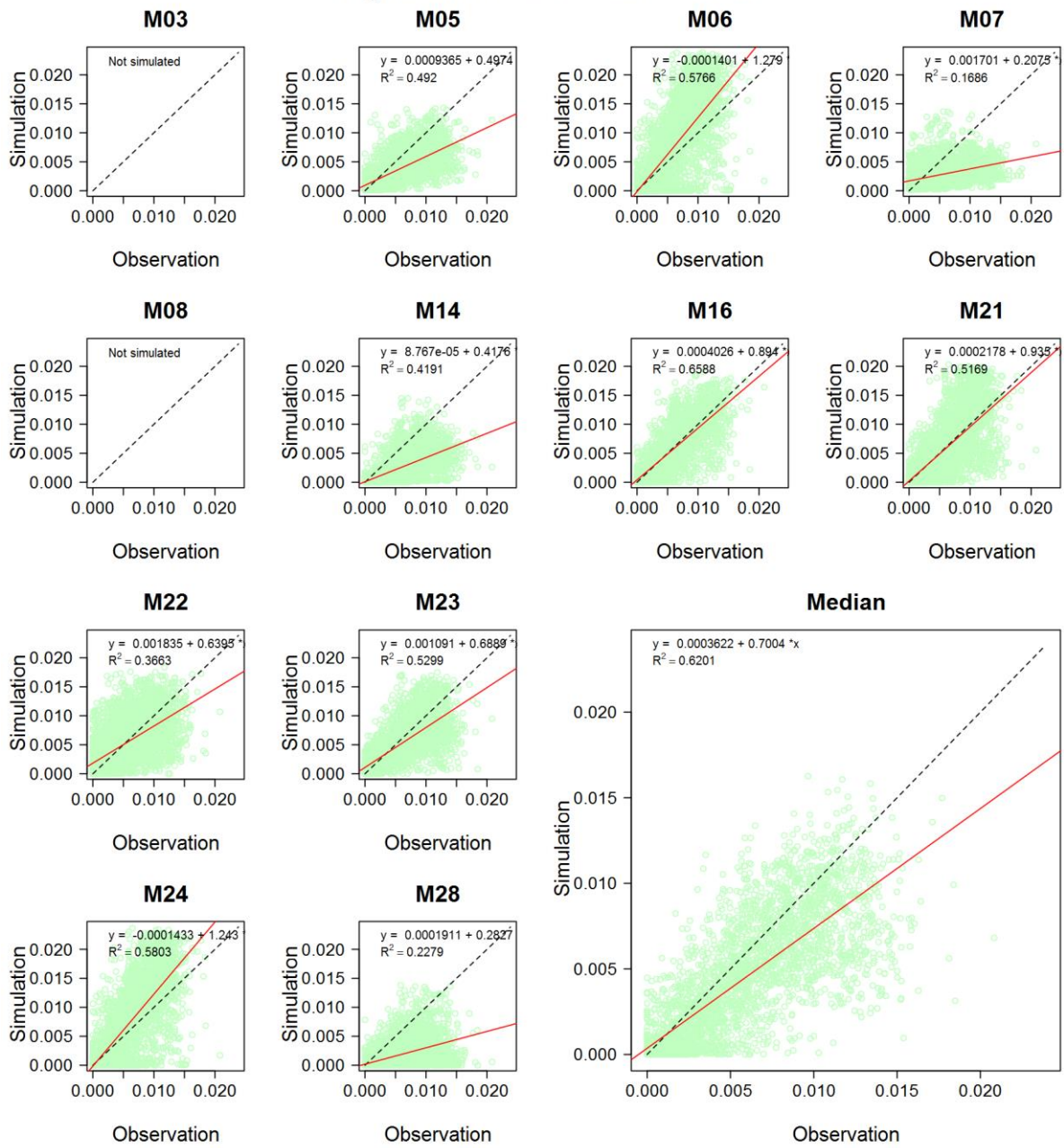
962 Fig. 2. Seasonal changes in ecosystem respiration (RECO), gross primary production (GPP), net
 963 ecosystem exchange (NEE), carbon use efficiency (CUE) and C intensity (Int_c) calculated over
 964 multiple years at C1, C2 and C3 crop, and G3 and G4 grassland sites, for five calibration stages

965 (S1 to S5) and the observation (Obs). Number of crop seasons/grassland years: soybean: 1;
966 triticale: 1; phacelia: 1; spring wheat: 2; rice: 2; maize: 3; rapeseeds: 4; winter wheat: 5; fallow:
967 9; grasslands: 19. For each calibration stage, black lines show multi-model median. Boxes
968 delimit the 25th and 75th percentiles. Whiskers are 10th and 90th percentiles. Circles indicate
969 outliers. For Obs, black line shows the observed mean.

970

971

Stage5 simulations of GPP at G3



972

973 Fig. 3. S5 calibration stage: comparison of simulated (individual models and multi-model

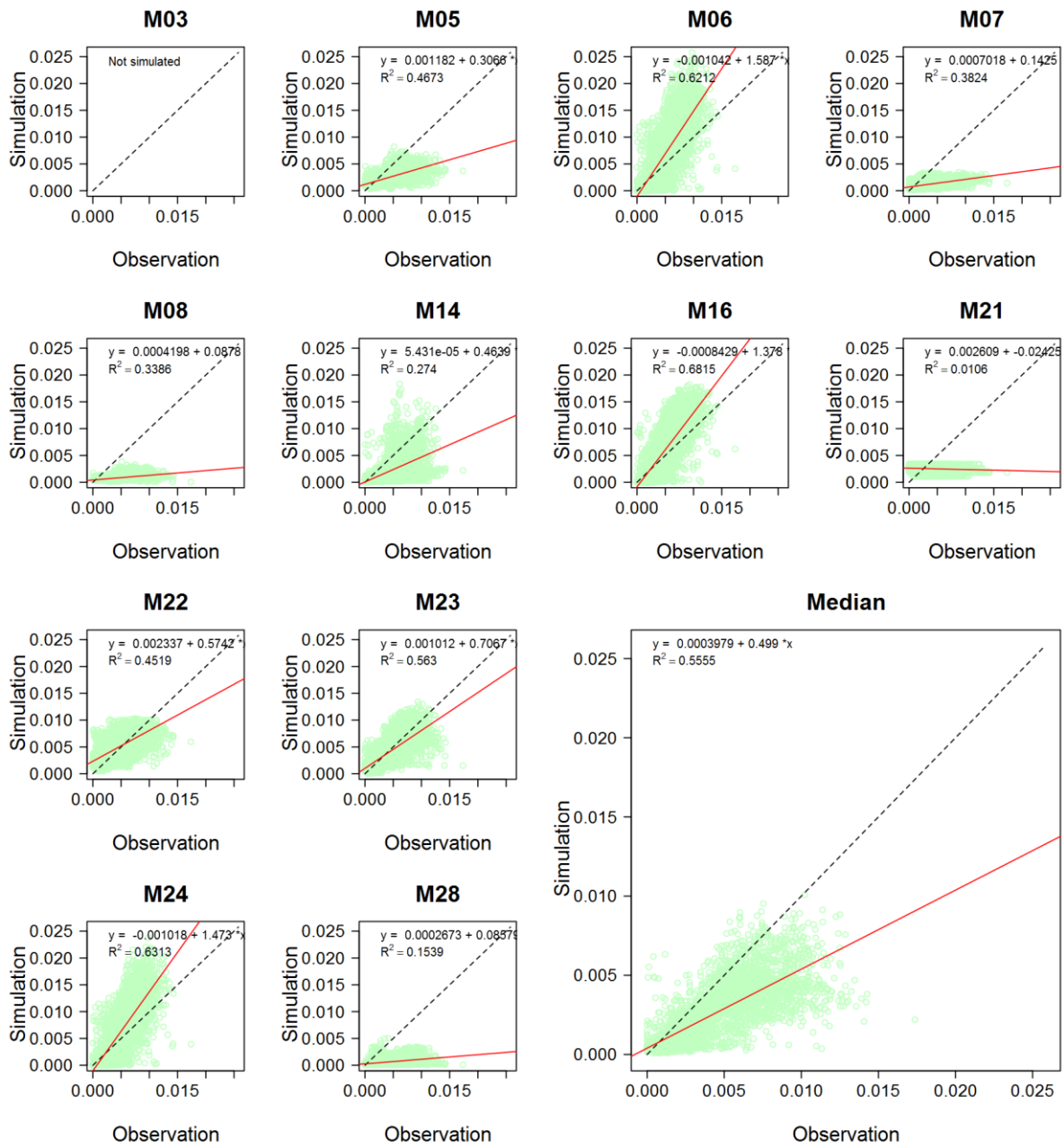
974 median) and observed daily gross primary production (GPP) data across multiple years at G3

975 site.

976

977

Stage5 simulations of RECO at G3



978

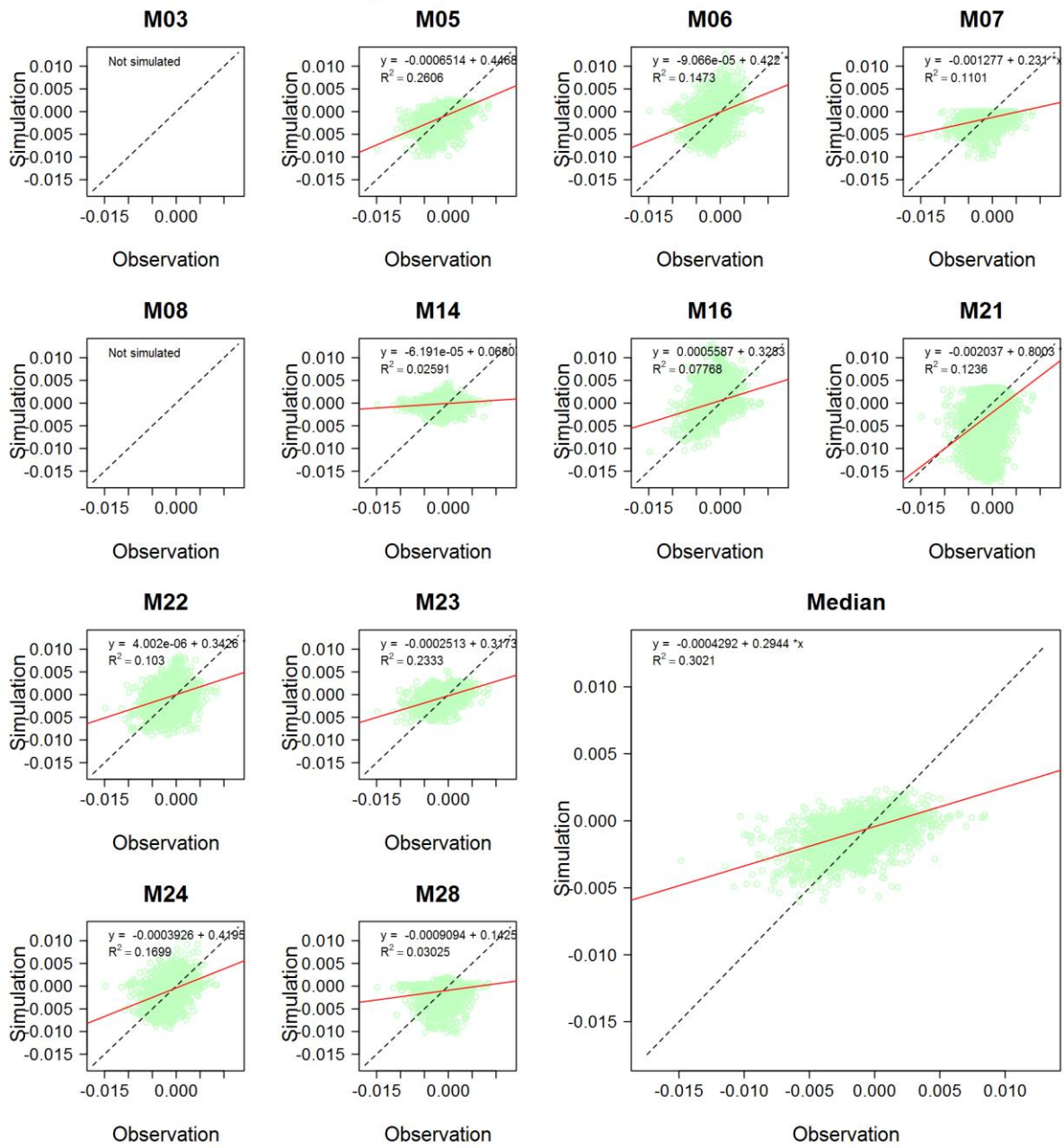
979 Fig. 4. S5 calibration stage: comparison of simulated (individual models and multi-model

980 median) and observed daily ecosystem respiration (RECO) data across multiple years at G3

981 site.

982

Stage5 simulations of NEE at G3



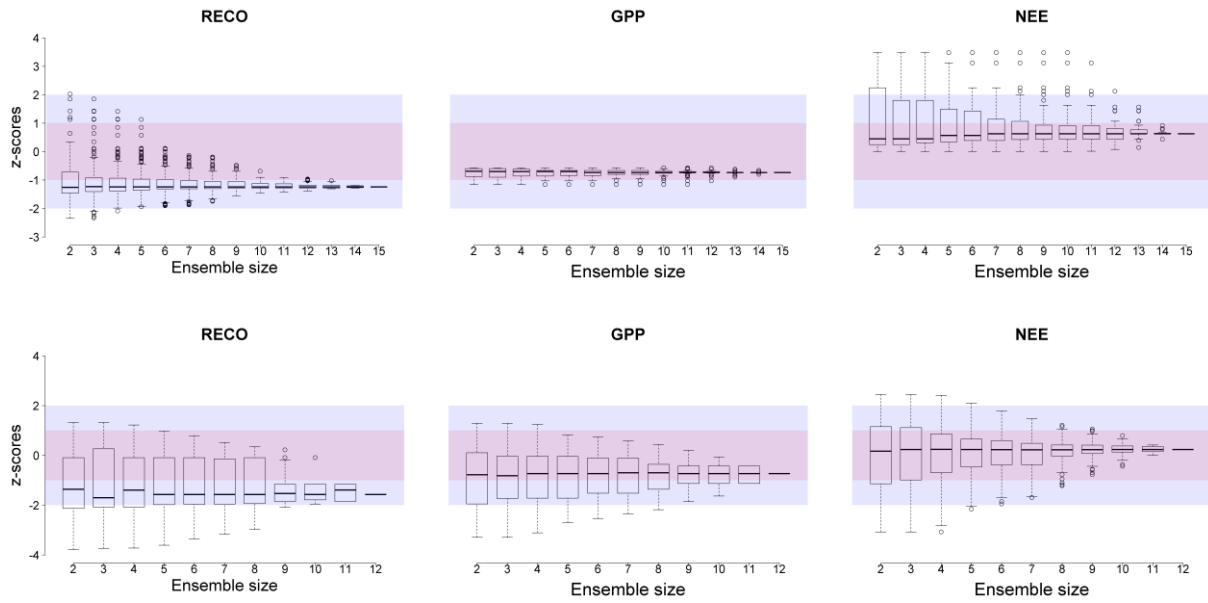
983

984 Fig. 5. S5 calibration stage: comparison of simulated (individual models and multi-model

985 median) and observed daily net ecosystem exchange (NEE) data across multiple years at G3

986 site.

987



988

989

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

994