



HAL
open science

Residual correlation and ensemble modelling to improve crop and grassland models

Renáta Sándor, Fiona Ehrhardt, Peter Grace, Sylvie Recous, Pete Smith, Val Snow, Jean-François Soussana, Arti Bhatia, Lorenzo Brillì, Jordi Doltra, et al.

► To cite this version:

Renáta Sándor, Fiona Ehrhardt, Peter Grace, Sylvie Recous, Pete Smith, et al.. Residual correlation and ensemble modelling to improve crop and grassland models. *Environmental Modelling and Software*, 2023, 161, pp.105625. 10.1016/j.envsoft.2023.105625 . hal-03997939

HAL Id: hal-03997939

<https://hal.inrae.fr/hal-03997939v1>

Submitted on 16 Mar 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.

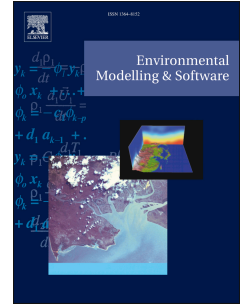


Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Journal Pre-proof

Residual correlation and ensemble modelling to improve crop and grassland models

Renáta Sándor, Fiona Ehrhardt, Peter Grace, Sylvie Recous, Pete Smith, Val Snow, Jean-François Soussana, Bruno Basso, Arti Bhatia, Lorenzo Brillì, Jordi Doltra, Christopher D. Dorich, Luca Doro, Nuala Fitton, Brian Grant, Matthew Tom Harrison, Ute Skiba, Miko U.F. Kirschbaum, Katja Klumpp, Patricia Laville, Joel Léonard, Raphaël Martin, Raia Silvia Massad, Andrew Moore, Vasileios Myrgiotis, Elizabeth Pattey, Susanne Rolinski, Joanna Sharp, Ward Smith, Lianhai Wu, Qing Zhang, Gianni Bellocchi



PII: S1364-8152(23)00011-7

DOI: <https://doi.org/10.1016/j.envsoft.2023.105625>

Reference: ENSO 105625

To appear in: *Environmental Modelling and Software*

Received Date: 7 April 2022

Revised Date: 30 October 2022

Accepted Date: 10 January 2023

Please cite this article as: Sándor, Rená., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, Jean.-Franç., Basso, B., Bhatia, A., Brillì, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Skiba, U., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, Raphaë., Massad, R.S., Moore, A., Myrgiotis, V., Pattey, E., Rolinski, S., Sharp, J., Smith, W., Wu, L., Zhang, Q., Bellocchi, G., Residual correlation and ensemble modelling to improve crop and grassland models, *Environmental Modelling and Software* (2023), doi: <https://doi.org/10.1016/j.envsoft.2023.105625>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Residual correlation and ensemble modelling to improve crop and grassland models

Renáta Sándor^{a,b}, Fiona Ehrhardt^{c,d}, Peter Grace^e, Sylvie Recous^f, Pete Smith^g, Val Snow^h, Jean-François Soussana^c, Bruno Bassoⁱ, Arti Bhatia^j, Lorenzo Brillik^{k,l}, Jordi Doltra^m, Christopher D. Dorichⁿ, Luca Doro^{o,p}, Nuala Fitton^g, Brian Grant^q, Matthew Tom Harrison^f, Ute Skiba^s, Miko U.F. Kirschbaum^t, Katja Klumpp^a, Patricia Laville^u, Joel Léonard^v, Raphaël Martin^a, Raia Silvia Massad^u, Andrew Moore^w, Vasileios Myrgiotis^x, Elizabeth Pattey^q, , Susanne Rolinski^y, Joanna Sharp^z, Ward Smith^q, Lianhai Wu^{aa}, Qing Zhang^{ab}, Gianni Bellocchi^a

^aUCA, INRAE, VetAgro Sup, Unité Mixte de Recherche sur Écosystème Prairial (UREP), 63000 Clermont-Ferrand, France

^bAgricultural Institute, ELKH CAR, 2462 Martonvásár, Hungary

^cINRAE, CODIR, 75007 Paris, France

^dRITMO Agroenvironnement, Colmar, France

^eQueensland University of Technology, Brisbane, Australia

^fUniversité de Reims Champagne Ardenne, INRAE, FARE, 51097 Reims, France

^gInstitute of Biological and Environmental Sciences, University of Aberdeen, UK

^hAgResearch - Lincoln Research Centre, Private Bag 4749, Christchurch 8140, New Zealand

ⁱDepartment of Geological Sciences, Michigan State University, East Lansing MI, USA

^jIndian Agricultural Research Institute, New Delhi, India

^kUniversity of Florence, DAGRI, 50144 Florence, Italy

^lIBE-CNR, 50145, Florence, Italy

^mInstitute of Agrifood Research and Technology (IRTA-Mas Badia), La Tallada d'Empordà, Catalonia, Spain

ⁿNREL, Colorado State University, Fort Collins CO, USA

^oDesertification Research Group, University of Sassari, Sassari, Italy

^pTexas A&M AgriLife Research, Blackland Research and Extension Center, Temple TX, USA

^qAgriculture and Agri-Food Canada, Ottawa, Ontario, Canada

^rUniversity of Tasmania, Newnham Dr, Launceston, Tasmania, 7248, Australia ^sUK Centre for Ecology and Hydrology, Bush Estate, Penicuik, EH26 0QB, UK

^tLandcare Research-Manaaki Whenua, Palmerston North, New Zealand

^uUniversité Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, 78850 Thiverval-Grignon, France

34 ^vINRAE, AgroImpact, 02000 Barenton-Bugny, France

35 ^wCSIRO, Agriculture Flagship, Black Mountain Laboratories, Canberra, Australia

36 ^xSchool of Geosciences, The University of Edinburgh, UK

37 ^yPotsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association,
38 Potsdam, Germany

39 ^zNew Zealand Institute for Plant and Food Research, Christchurch, New Zealand

40 ^{aa}Sustainable Agriculture Systems, Rothamsted Research, North Wyke, Devon, UK

41 ^{ab}LAPC, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

42

43 Corresponding author. Agricultural Institute, ELKH CAR, 2462 Martonvásár, Hungary.
44 sandor.rencsi@gmail.com

45

46 **Abstract**

47 Multi-model ensembles are becoming increasingly accepted for the estimation of agricultural
48 carbon-nitrogen fluxes, productivity and sustainability. There is mounting evidence that with
49 some site-specific observations available for model calibration (with vegetation data as a
50 minimum requirement), median outputs assimilated from biogeochemical models (multi-model
51 medians) provide more accurate simulations than individual models. Here, we evaluate
52 potential deficiencies in how model ensembles represent (in relation to climatic factors) the
53 processes underlying biogeochemical outputs in complex agricultural systems such as grassland
54 and crop rotations including fallow periods. We do that by exploring the correlation of model
55 residuals. We restricted the distinction between partial and full calibration to the two most
56 relevant calibration stages, i.e. with plant data only (partial) and with a combination of plant,
57 soil physical and biogeochemical data (full). It introduces and evaluates the trade-off between
58 (1) what is practical to apply for model users and beneficiaries, and (2) what constitutes best
59 modelling practice. The lower correlations obtained overall with fully calibrated models
60 highlight the centrality of the full calibration scenario for identifying areas of model structures
61 that require further development.

62

63 **Keywords:** biogeochemical models; correlation matrices; ensemble modelling; model
64 calibration; residual plot analysis

65

66 1. Introduction

67 The development of a robust modelling capacity is needed to carry out assessments of
68 agricultural carbon (C) and nitrogen (N) fluxes (productivity, leaching and export) and to
69 quantify the outcomes of agricultural management and policy decisions, as it supports
70 participatory frameworks, as well as sensitivity and uncertainty analyses of model outputs (e.g.
71 Martin et al., 2018; Harrison et al., 2019). Several biogeochemical models are available for
72 estimating variables of agronomic, environmental and ecological interest in croplands and
73 grasslands (see a summary in Brilli et al., 2017). Owing to insufficient knowledge,
74 approximations, inaccurate parameterisations and/or lack of biological and physical
75 representations, each crop or grassland model is an imperfect representation of the biophysical
76 and biogeochemical processes in the vegetation, soil and atmosphere that are critical to
77 ecosystem functioning (e.g. Challinor et al., 2013; Snow et al., 2014; Calanca et al., 2016; Jones
78 et al., 2017a). Thus, each model represents a balance between parsimony and excessive
79 complexity (Harrison et al., 2012). Models may give different answers to the same scientific
80 question, not just in terms of the estimated magnitude of output, but also in the direction of
81 change under climate or management scenarios (Brilli et al., 2017; Bilotto et al., 2021).
82 Comparing and contrasting different models for their fit, precision, scope, validity and
83 reliability may lead to choosing the one model that is optimal for the intended purposes (e.g.
84 Bellocchi et al., 2010). However, relying on a single model deemed to be the best, ignores the
85 uncertainty associated with alternative model structures and underestimates the possible effects
86 of inaccurate estimates, especially when models are used in contexts outside the original
87 development area (e.g. Riccio et al., 2007). Many authors have recognised the drawbacks of
88 ignoring model uncertainties (e.g. papers cited by Dijkstra, 1988). Due to a lack of knowledge
89 about whether any model is an appropriate representation of the target system/output in
90 question, epistemic uncertainties, in particular, contribute to model spread. This is realised by
91 a range of responses in a model ensemble (e.g. Knutti et al., 2019).

92 Ensemble modelling is an emerging method that involves running several related (but different)
93 modelling solutions and then combining their results into a single result (or comparing them),
94 which creates a consensus on the predictions obtained with multiple models (Spence et al.,
95 2017; Calder et al., 2018). In addition, a smaller selection of models can approximate the
96 median of a larger ensemble once all models are verified (e.g. Ehrhardt et al., 2018). Multi-
97 model ensembles aim to reduce uncertainties in the prediction because ensemble estimates
98 include multiple alternative representations of the same biophysical and biogeochemical
99 processes in agricultural systems. They also provide more reliable information on the

100 uncertainties of the outputs predicted by the diversity amongst ensemble members, as
101 highlighted in crop/grassland modelling exercises (e.g. Bassu et al., 2014; Rosenzweig et al.,
102 2014; Kollas et al., 2015; Li et al., 2015; Ruane et al., 2016, 2017; Sándor et al., 2017). The
103 assumption underlying the use of multiple models is that a measure of central tendency of the
104 results of different models reduces uncertainties by balancing the errors of the individual
105 models and thus results in a better fit (e.g. Riggers et al., 2019). In many cases, the median
106 value of multi-model predictions was shown to be able to outperform any single deterministic
107 model in reproducing observational data at different locations (as explained by Martre et al.,
108 2015 and, on a theoretical basis, by Wallach et al., 2018). In particular, model simulations are
109 less accurate in situations of limited inputs and below-potential yield situations, where soil
110 processes need to be adequately simulated, and model ensembles offer higher accuracy than
111 randomly taken models (Falconnier et al., 2020). For this reason, ensemble modelling is a
112 proposed means of reducing some of the uncertainties in model estimates of productivity and
113 other C and N fluxes in croplands and grasslands (Ehrhardt et al., 2018; Sándor et al., 2020).
114 Intrinsic differences between models may also become a useful asset to be exploited for more
115 informed decision-making support, e.g. towards alternative farming practices to reduce net
116 greenhouse gas emissions (Alcock et al., 2015; Harrison et al., 2016; Sándor et al., 2018a). As
117 a corollary to reducing ensemble uncertainties, running more models can highlight model
118 shortcomings, as it is unlikely that all models represent each physical phenomenon in the same
119 way (e.g. Sándor et al., 2016). Thus, the envelope of possible model outputs can be narrowed
120 as our understanding of key processes improves, or with the inclusion of a particular process
121 not previously considered, or to save time in scaling up.

122 With the aim of increasing reliability and confidence in the simulated results, this study explores
123 patterns of simulated C-N and productivity responses with a multi-model ensemble approach.
124 We included results from 23 crop and grassland models, used to simulate C-N and productivity
125 outputs in five sites worldwide (three crop rotations with spring and winter cereals, soybean
126 and rapeseed, and two temperate grasslands). This work builds on comprehensive foundations
127 laid by Ehrhardt et al. (2018) for yield and nitrous oxide (N₂O) emissions, and Sándor et al.
128 (2020) for C fluxes. Here, we analyse factors that may explain differences in simulated model
129 responses. Viewing and interpreting a variety of modelled outputs is intended to lay ground for
130 future model developments. We thus further explored the extent to which multi-model
131 ensembles can be used to help identify deficiencies in model structures, which limit model
132 performance in different situations. Specifically, we present an approach that uses a correlation
133 matrix (with graphical representation) to correlate both the residuals of outputs from the

134 ensemble against residuals of selected climate drivers. The estimation of uncertainty in
 135 simulation models is based on the assumption that model residuals (differences between model
 136 estimates and observations) are additive and independent. When the residuals of one model
 137 output are correlated with the residuals of other outputs, the different outputs would probably
 138 be the result of processes not included (or partially included) in the models. This suggests that
 139 interacting processes are sources of model-data mismatch and, in this case, non-negligible
 140 correlations between model residuals and external drivers might inspire a more detailed
 141 description of these same drivers to improve the models.

142 Focusing on the correlation among model residuals, the central assumption of this study is that
 143 an ensemble of partially or fully calibrated models can produce uncorrelated residuals which
 144 would validate the assumptions of error independence. Using the median of the outputs of
 145 several models as a metric of the multi-model ensemble, the aim was to compare the
 146 standardised residuals of the different outputs of an ensemble of models run with limited
 147 calibration datasets (partial calibration desirable for users and beneficiaries) and rich datasets
 148 (full calibration more suitable for scientists).

149

150 2. Materials and methods

151 2.1. Experimental sites and measurements

152 We adopt multi-year model outputs, obtained from 23 crop and grassland simulation models at
 153 five agricultural sites worldwide (Sándor et al., 2020). The approach was based on a multi-
 154 model study, in which all participating teams received the same data and were asked to return
 155 simulated outputs for the same conditions using their usual calibration techniques (for a
 156 discussion on the validity of calibration practices for good modelling, see Wallach et al., 2021).
 157 The models were run independently in five stages (S), as shown in Table 1, from blind
 158 modelling (S1) to partial (S2 to S4) and full (S5) calibrations. In particular, site-specific model
 159 parameterisation was performed at each modelling stage, with gradual access to site data from
 160 S2 onwards, to inform and parameterise the models.

161

162 Table 1. Stages of model run (after Ehrhardt et al., 2018). The grey cells indicate the two stages
 163 (S3: partial calibration; S5: full calibration) on which this study focuses.

	Modelling stage	Description
S1	blind with no calibration and initialisation data	Basic data covering the simulation period of experimental measurements (climate, initial soil properties and site management information, crop

			rotation/grazing configuration, fertilisation and irrigation)
S2	initialisation with management and climate	historical	Historical site-specific data for climate and management allowing for long-term initialisation periods, and regional statistics for crop yields and pasture productivity from expert estimates
S3	calibration against vegetation data		Site-specific phenology data, crop/pasture vegetation development (e.g. leaf area index), observed grain yields, monthly estimated grassland offtake (biomass removed by mowing or animal intake)
S4	calibration against vegetation and soil data together		Dynamic soil process data (temperature, moisture, mineral N dynamics)
S5	calibration with the addition of surface-to-atmosphere C and N fluxes		C-N emissions and soil organic C stock changes

164

165 For consistency, we have maintained the model and site identifiers specified by Ehrhardt et al.
166 (2018). The variability of the multi-model simulation exercise across stages was documented
167 by inspecting how the multi-model median (MMM) converged to the observations.
168 Observational data were from two long-term (19 years in total), grazed experimental sites (G3,
169 G4) and three cropland sites (C1, C2, C3), covering a variety of pedo-climatic conditions and
170 agricultural practices from United Kingdom, France (two sites), Canada and India (Table 2).
171 The selected cropping systems covered a range of climates, from continental (C1, Canada),
172 oceanic (C2, France) and subtropical (C3, India). All cropland sites had rotations with at least
173 one wheat crop (six growing seasons), while maize was present in C1 and C2 (three growing
174 seasons), and rice was only grown in C3 (two growing seasons), for a total of 18 growing
175 seasons (including fallow intercrops). The 23 models (Table A in the Supplementary material),
176 and the model identifiers and outputs provided, encompass all but one of the 24 biogeochemical
177 models described in Ehrhardt et al. (2018). Model M11 was not included in the analysis because
178 it did not provide the C-flux related outputs. At cropland sites, we had: GPP from six models,
179 NEE from seven models, RECO from 14 models, N₂O from 15 models, Yield from 15 models.
180 At grassland sites, we had: GPP from 10 models, NEE from 10 models, RECO from 11 models,
181 N₂O from nine models, Yield from nine models. The use of flux tower data allows the
182 determination of NEE, which is partitioned into its (simulated) component fluxes - RECO and
183 GPP – by flux partitioning methods. Separated from flux tower measurements of NEE, the
184 estimated GPP provides information on the physiological processes that contribute to NEE,
185 which is the balance between the C released by the RECO and the GPP (e.g. Raj et al., 2016).
186 Climate data available at each site since 1980 were used to initialise the models (calibration
187 stage S2).

188

189 Table 2. Cropland and grassland sites, and years of available data, for analysis on the following
 190 output variables from different models: GPP ($\text{g C m}^{-2} \text{ yr}^{-1}$): gross primary production; RECO
 191 ($\text{g C m}^{-2} \text{ yr}^{-1}$): ecosystem respiration; NEE ($\text{g C m}^{-2} \text{ yr}^{-1}$): net ecosystem exchange of CO_2); N_2O
 192 ($\mu\text{g N}_2\text{O-N m}^{-2} \text{ yr}^{-1}$): nitrous oxide emissions; Yield ($\text{kg DM m}^{-2} \text{ yr}^{-1}$): annual grain yield for
 193 arable crops or annual above-ground net primary productivity for grasslands. Cropland sites
 194 used different crop rotations (Table B in the Supplementary material), including cereals (spring
 195 and winter wheat [W], triticale [T], maize [M] and rice [R]), legumes (soybean [S]), rapeseeds
 196 (canola and mustard [C]), borages (phacelia, F) and fallow intercrop periods [I].

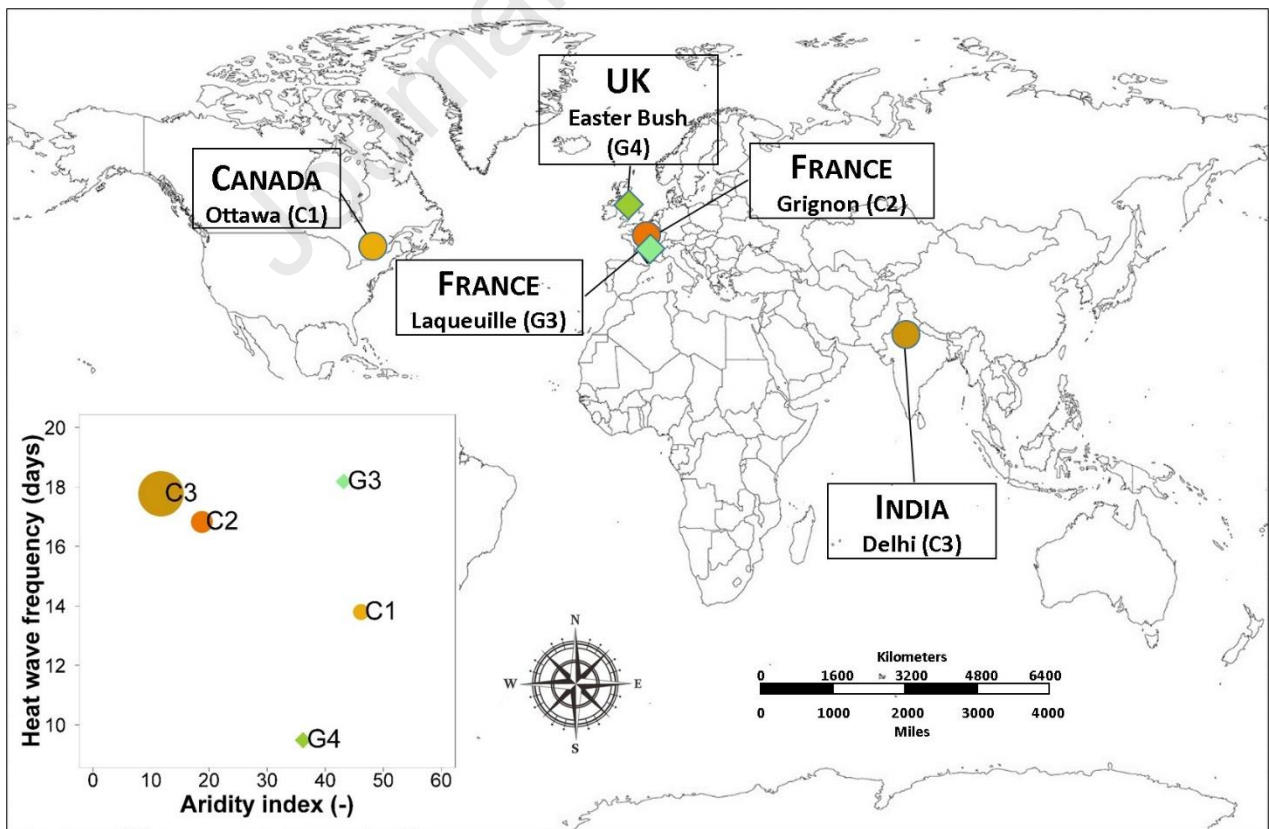
Sites, country (latitude, longitude, elevation)	Years of available data (simulation period)	Land use	References
C1: Ottawa, Canada (45.29, -75.77, 94 m a.s.l.)	2007-2012	W/S/C/M/W/C	Pattey et al. (2006); Jégo et al. (2012); Sansoulet et al. (2014)
C2: Grignon, France (48.85, 1.95, 125 m a.s.l.)	2008-2012	C/M/W/T/P/M/W/I	Laville et al. (2011); Loubet et al. (2011)
C3: Delhi, India (28.60, 78.22, 233 m a.s.l.)	2006-2009	W/R/W/R/W	Bhatia et al. (2012)
G3: Laqueuille, France (45.64, 2.74, 1040 m a.s.l.)	2003-2012	Permanent grassland	Allard et al. (2007); Klumpp et al. (2011)
G4: Easter Bush, United Kingdom (55.52, -3.33, 190 m a.s.l.)	2002-2010	Permanent grassland	Skiba et al. (2013); Jones et al. (2017b)

197

198 2.2. Agro-climatic metrics

199 Three metrics were selected to characterise the study-sites based on the extent to which they
 200 fulfil the need to report the response of models to water-limited and heat stressed conditions
 201 (Sándor et al., 2017, 2018; Farina et al., 2021). They are also important within a climate-change
 202 focus (Rivington et al., 2007, 2013; Matthews et al., 2008; Graux et al., 2013; Lardy et al.,
 203 2014, 2015; Eza et al., 2015). An increase in T_{max} and frequency of hw is desirable if the two
 204 metrics are negatively correlated with model residuals. The aridity index (b) is defined in such
 205 a way (the higher it is, the lower the aridity) that, with a positive correlation, higher model
 206 residuals are expected in wetter conditions and, with a negative correlation, higher model
 207 residuals are expected in drier conditions. In fact, the De Martonne aridity index ($b \leq 100$) was

208 derived following Gottmann (De Martonne, 1942), as $b = \frac{1}{2} \cdot \left(\frac{P_Y}{T_Y+10} + 12 \cdot \frac{p_a}{t_a+10} \right)$, where P_Y is
 209 the total annual precipitation (mm), T_Y is the mean annual temperature ($^{\circ}\text{C}$), p_a is the total
 210 precipitation of the driest month (mm), and t_a is the mean temperature of the driest month ($^{\circ}\text{C}$).
 211 The possibility to discriminate between thermo-pluviometric conditions associated with aridity
 212 gradients is given by the range limits published by Diodato and Ceccarelli (2004): $b < 5$: extreme
 213 aridity; $5 \leq b \leq 14$: aridity; $15 \leq b \leq 19$: semi-aridity; $20 \leq b \leq 29$: sub-humidity; $30 \leq b \leq 59$: humidity;
 214 $b > 59$: strong humidity. Adopting the definition of Confalonieri et al. (2010), after Barnett et al.
 215 (2006), for identifying the frequency of hw within a year in each site, we defined the heatwave
 216 event as the number of ≥ 7 consecutive days when T_{max} was higher than the mean summer
 217 (northern hemisphere: June, July and August in the temperate sites; April, May and June in the
 218 monsoonal site) T_{max} of all the available years (baseline) $+3^{\circ}\text{C}$. The range limits in this study
 219 were given by the minimum and the maximum numbers of the hw days of all sites: $hw \leq 14$:
 220 extremely moderate frequency; $14 < hw \leq 28$: very moderate frequency; $28 < hw \leq 42$: moderate
 221 frequency; $42 < hw \leq 56$: high frequency; $56 < hw \leq 70$: very high frequency; $hw > 70$: extremely
 222 high frequency. Fig. 1 displays the gradient of thermo-pluviometric conditions that are
 223 considered to analyse the response of the model residuals to climate drivers.
 224



225 Fig. 1. Geographic location (diamonds: grassland sites; circles: cropland sites) and
 226 classification of study sites with respect to De Martonne-Gottmann aridity index and frequency
 227

228 of heatwave days (left-bottom graph). The area of the circles and diamonds in the left-bottom
229 graph is proportional to the mean maximum air temperature of each site.

230

231 2.3. Residual scatterplot analyses

232 According to Ehrhardt et al. (2018) and Sándor et al. (2020), although detailed observations
233 (i.e. C-N fluxes) to support full model calibration (S5) may be desirable, multiple model
234 ensembles with plant observations as a minimum data requirement (S3) could be a promising
235 way to guide modelling applications.

236 For both arable crops and grasslands, Ehrhardt et al. (2018) found that no model consistently
237 outperformed the others in terms of both N₂O emissions and yield production. In particular, in
238 the case of cereal crop yields, the MMM error decreased considerably from S1 (34%, 31% and
239 45% for wheat, maize and rice, respectively) to S3 (6.4%, 5.8% and 5.5% for wheat, maize and
240 rice, respectively) and remained below 5% in S4 and S5. In the case of grassland yields, the
241 MMM error decreased from 44% in S1 to 27% in S3 and finally increased to 46% in S5.

242 Sándor et al. (2020) reported that the MMM outperformed the individual models in 92.3% of
243 the cases and, in general, they obtained the greatest improvements (MMM close to the mean of
244 the observations) at calibration stages S3 or higher. For instance, the best cropland RECO
245 estimates were obtained with S3, where the MMM and the observed mean were similar: 241
246 and 242 g C m⁻² season⁻¹, respectively (mean of sites C1, C2 and C3). For the GPP of grasslands,
247 the best estimates were obtained with S5, where the MMM was equal to 1632 g C m⁻² yr⁻¹ and
248 the observed mean was equal to 1763 g C m⁻² yr⁻¹ (mean of sites G3 and G4).

249 We thus quantified the correlations among standardised model residuals of GPP, RECO, NEE,
250 N₂O and Yield (differences between ensemble MMM and mean of observations), based on the
251 results from partially and fully calibrated simulations (stages S3 and S5). For both calibration
252 stages, we also quantified the correlations between model residuals and three agro-climatic
253 metrics (annual values) related to the occurrence of high temperature (mean maximum air
254 temperature, *T_{max}* and heatwave days, *hw*) and arid conditions (Figures A-E in the
255 Supplementary material).

256 Arrays of pairwise scatterplots (scatterplot matrices) were generated with the panel plot option
257 'panel.smooth' ([https://stat.ethz.ch/R-manual/R-](https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/panel.smooth.html)
258 [devel/library/graphics/html/panel.smooth.html](https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/panel.smooth.html)) in the R language and environment for
259 statistical computing (R Core Team, 2020). The function produces *x*-*y* scatterplots of each pair
260 of variables below the diagonal (output residuals and agro-climatic metrics) and overlays a local
261 non-parametric smoother curve (locally estimated scatterplot smoothing) on each plot to give

262 some indication of trends without inferential characteristics (after Cleveland, 1979). For
263 readability, the correlation between each variable and its significance (p value) is indicated in
264 the lower triangular part of the matrices. The non-significant correlations ($p \geq 0.10$) are not
265 discussed (e.g. Bellocchi et al., 2002). According to Sándor et al. (2017), we have selected an
266 arbitrary (high enough) absolute minimum threshold, i.e. $r = |0.66|$, and identified the number of
267 cases when the correlation coefficient equals or exceeds this minimum value. Correlations
268 between external climate factors (mean maximum air temperature, aridity index and frequency
269 of heatwave days) are reported but are not informative in the present context.

270

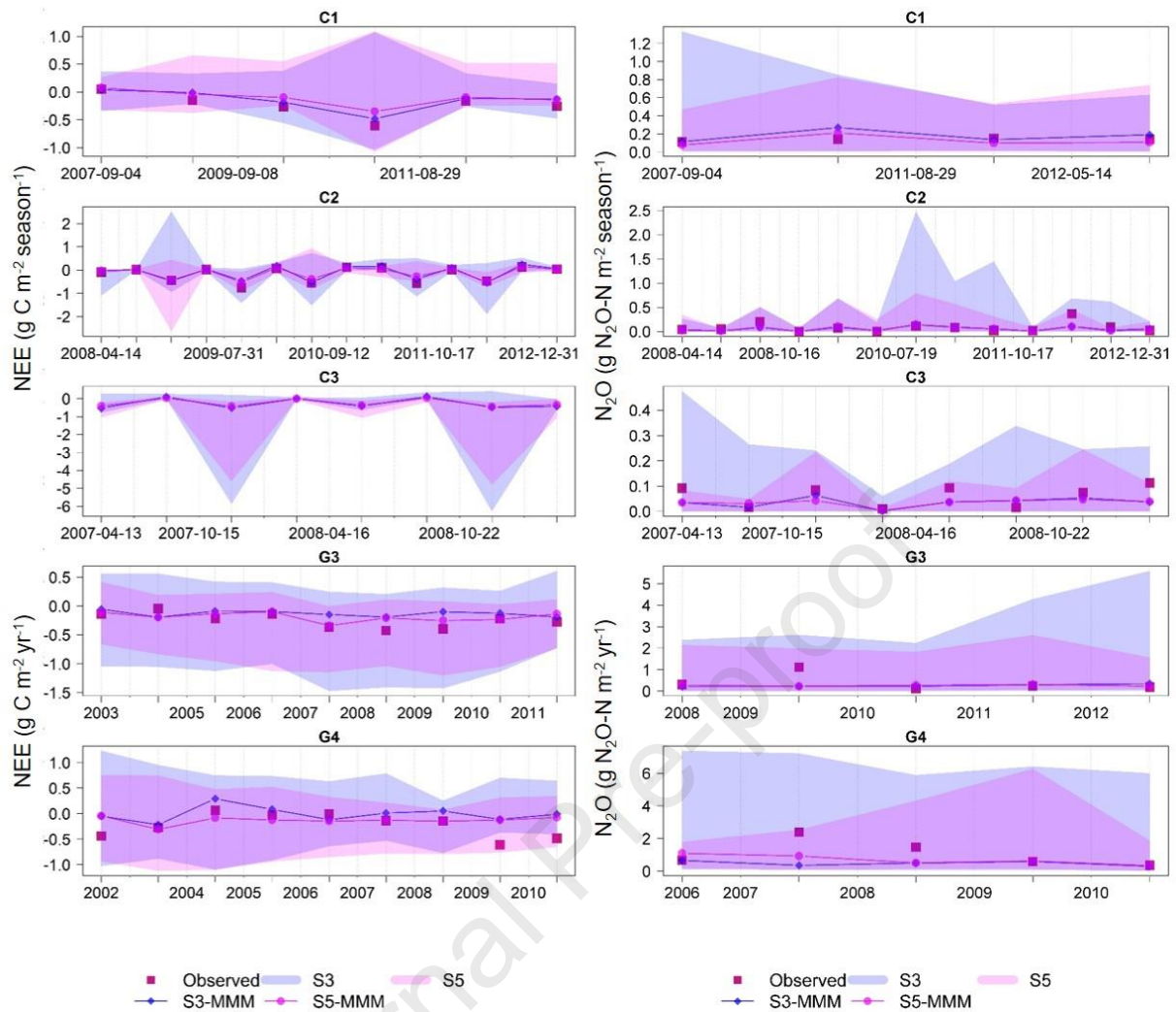
271 **3. Results**

272 *3.1. Evaluation of output dynamics*

273 In general, model results showed the largest spread with the S3 scenario, considering the C
274 outputs such as NEE (Fig. 2), GPP and RECO (Appendix A), N₂O-N emissions (Fig. 2) and
275 yield (Appendix A). In some years, the MMM of S3 and in some cases the S5 scenario also
276 overestimated the amount of C respiration, e.g. at G4 site in 2002 (S3: -0.05; S5: -0.04;
277 observed: -0.44 g C m⁻² yr⁻¹) and 2010 (S3: -0.01; S5: -0.07; observed: -0.48 g C m⁻² yr⁻¹),
278 while the N₂O-N emission was underestimated at this site. The MMM lines for all outputs were
279 remarkably close to the observations at all sites, despite the wider range of S3 individual
280 simulations (blue shaded area in Fig. 2 and Appendix A). The largest difference between the
281 spread of S3 and S5 was found for the N₂O-N emissions.

282

283



284
 285 Fig. 2. Temporal changes of NEE ($\text{g C m}^{-2} \text{ season}^{-1}$ for crops and $\text{g C m}^{-2} \text{ yr}^{-1}$ for grasslands,
 286 left) and N_2O ($\text{g N}_2\text{O-N m}^{-2} \text{ season}^{-1}$ for crops and $\text{g N}_2\text{O-N m}^{-2} \text{ yr}^{-1}$ for grasslands, right)
 287 observations (Obs, red square) and simulations: S3 (stage 3, blue) and S5 (stage 5, pink) at all
 288 sites (site codes as in Fig. 1). Lines represent the multi-model median (MMM) of the S3 and S5
 289 simulations, and shaded areas represent the simulation envelopes given by the edges of the most
 290 extreme model predictions (with the same colours as the lines). At cropland site C3, only
 291 modelled RECO data are reported.

292

293 3.2. Residual analysis in grassland sites

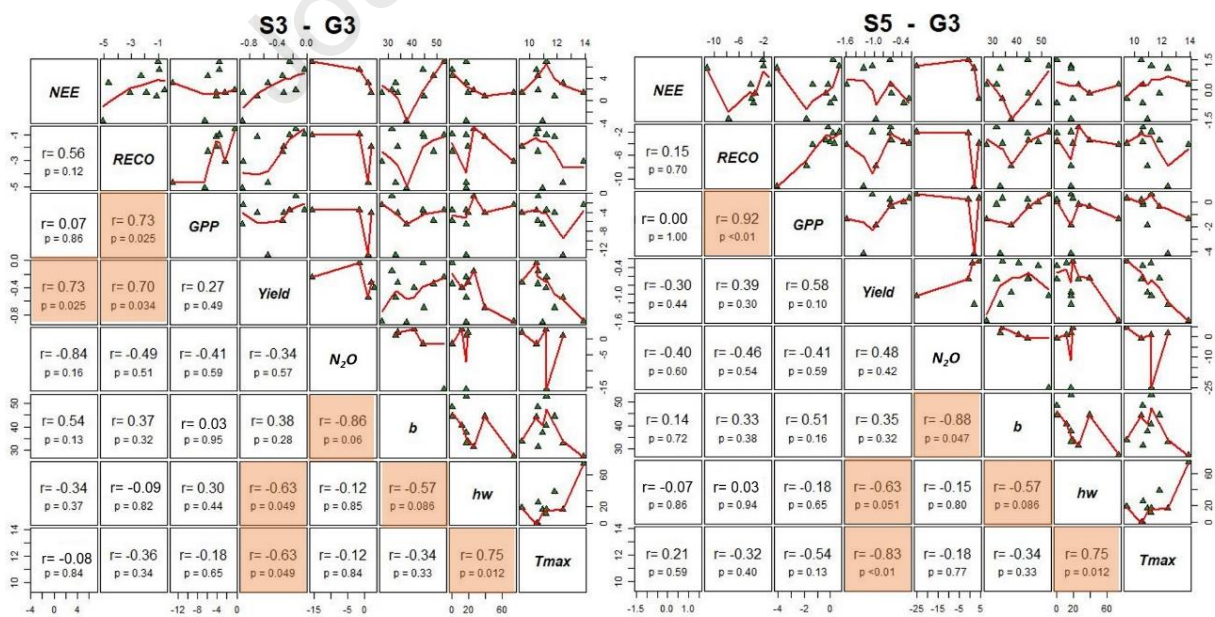
294 The MMM analysis of residual scatterplot clouds at G3 (Laqueuille, France) shows some
 295 similarities between the S3 (Fig. 3, left) and S5 (Fig. 3, right) calibration stages. The values of
 296 RECO and GPP residuals are positively correlated ($r=0.73$, $p=0.03$ and $r=0.92$, $p<0.01$ for S3
 297 and S5, respectively), so any overestimation in RECO could also lead to an overestimation of
 298 GPP. However, since there is no effective correlation between NEE and GPP ($r\sim 0$ at both

299 calibration stages), over- or underestimation of GPP would not be responsible for over- or
 300 underestimation of NEE. In S3 stage (i.e. when only plant data like yield biomass and leaf area
 301 index were used for calibration), Yield residuals positively correlated with NEE and RECO
 302 residuals ($r=0.73$, $p=0.03$ and $r=0.70$, $p<0.01$, respectively), so overestimation of yield biomass
 303 tended to be associated with overestimated C-flux simulations (e.g. overestimated yield would
 304 lead to underestimation of NEE values). At S5, Yield residuals do not show a significant
 305 correlation ($p>0.10$) with C residuals.

306 Considering the climatic factors at the G3 site, aridity values (higher aridity index indicates
 307 wetter conditions) show a negative correlation with N₂O residuals ($r=-0.86$, $p=0.06$ and $r=-0.88$,
 308 $p=0.05$ at stages S3 and S5, respectively), with higher model residuals expected in drier
 309 conditions in the estimation of N₂O emissions. When *Tmax* is considered for both S3 and S5,
 310 the correlation with Yield residuals is significantly negative ($r=-0.63$, $p=0.05$ and $r=-0.83$,
 311 $p<0.01$, respectively). With S5, the days of heatwave are negatively correlated with Yield
 312 residuals ($r=-0.63$, $p=0.05$), with model outputs becoming less reliable at lower temperatures.
 313 This indicates that state-of-the-art models take into account the influence of climate factors, as
 314 periods of extreme heat and drought, or extremely wet conditions, tend to decrease or increase
 315 model errors. For instance, simulated N₂O emissions may show higher magnitude residuals
 316 under drier conditions, while yield and C-flux simulations may have lower magnitude residuals
 317 (e.g. models are more sensitive to wet G3 upland conditions).

318

319



320

321

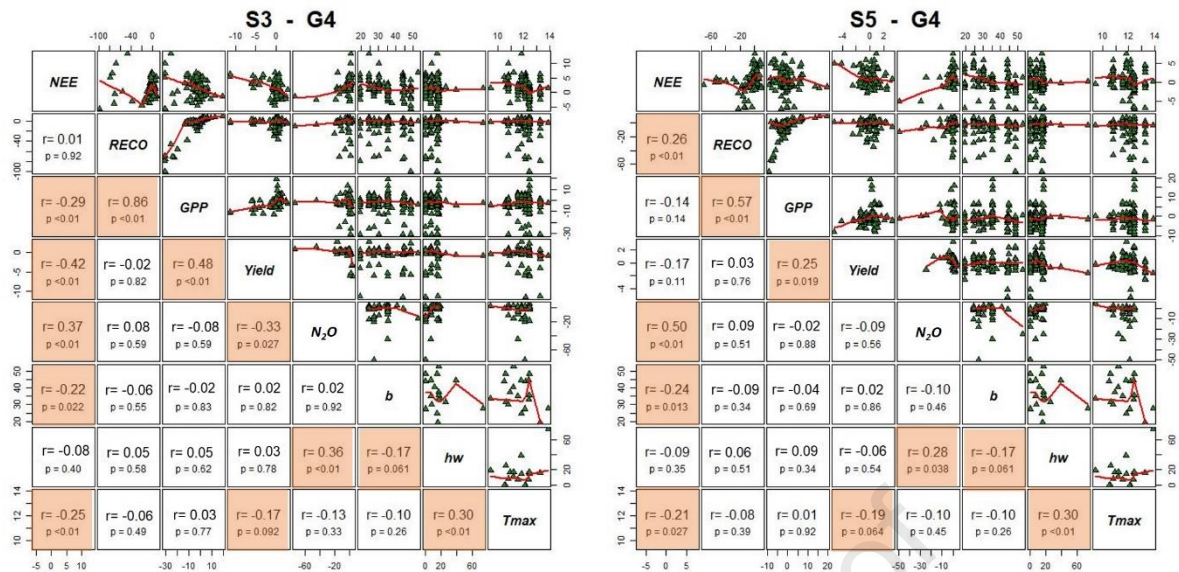
322 Fig. 3. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-
323 model medians (MMM) for stages 3 (left) and 5 (right) at G3 grassland site, and the annual
324 agro-climatic metrics aridity index (*b*), heatwave frequency (*hw*) and maximum temperature
325 (*Tmax*). Overlaid (red line) is a local non-parametric smoother curve. Coloured areas indicate
326 significant correlations ($p < 0.10$).

327

328 Analysis of residual scatterplots at G4 (Easter Bush, United Kingdom) shows some similarities
329 at both calibration stages (Fig. 4). The negative correlation between NEE and GPP residuals at
330 S3 ($r = -0.29$, $p < 0.01$) indicates that overestimation of NEE may be the result of underestimation
331 of GPP. This is reflected in the negative correlation between NEE and Yield ($r = -0.42$ at S3,
332 $p < 0.01$). RECO and GPP residuals are significantly ($p < 0.01$) positively correlated ($r = 0.86$ at
333 S3 and $r = 0.57$ at S5). In addition, GPP and Yield residuals are positively correlated ($r = 0.48$,
334 $p < 0.01$ and $r = 0.25$, $p = 0.02$ at S3 and S5, respectively). Overall, these correlations between C-
335 fluxes and yield residuals are less important or less significant for the fully calibrated models
336 (S5). However, N₂O residuals show significant correlations ($p < 0.01$) with NEE residuals at
337 both calibration stages ($r = 0.37$ and $r = 0.50$ at S3 and S5, respectively), while no significant
338 correlations ($p > 0.10$) were found with other C-flux residuals. Considering climatic factors,
339 heatwaves do not have a significant impact on C-flux and Yield residuals in G4 (which is not
340 exposed to severe heatwaves; Fig. 1). Interestingly, N₂O-emission residuals are significantly
341 ($p < 0.01$) positively correlated with heatwaves at both S3 ($r = 0.36$) and S5 ($r = 0.28$). Thus,
342 increasingly long heatwaves may lead to greater model inaccuracy in simulating N₂O
343 emissions, likely due to poor estimates of soil water content at higher temperatures or model
344 limitations in appropriately reducing emission estimates at low soil water contents (Wang et al.,
345 2021). The aridity index was negatively correlated ($p < 0.05$) with NEE residuals for both S3
346 ($r = -0.22$) and S5 ($r = -0.24$), and was not correlated with N₂O, GPP, RECO and Yield residuals.
347 These negative correlations indicate that simulations are generally more reliable under G4
348 humid conditions. Since *Tmax* is significantly negatively correlated with NEE at S3 ($r = -0.25$,
349 $p < 0.01$) and S5 ($r = -0.21$, $p < 0.05$), the models are expected to give poorer C-flux simulations
350 under colder conditions and better results at higher temperatures.

351

352



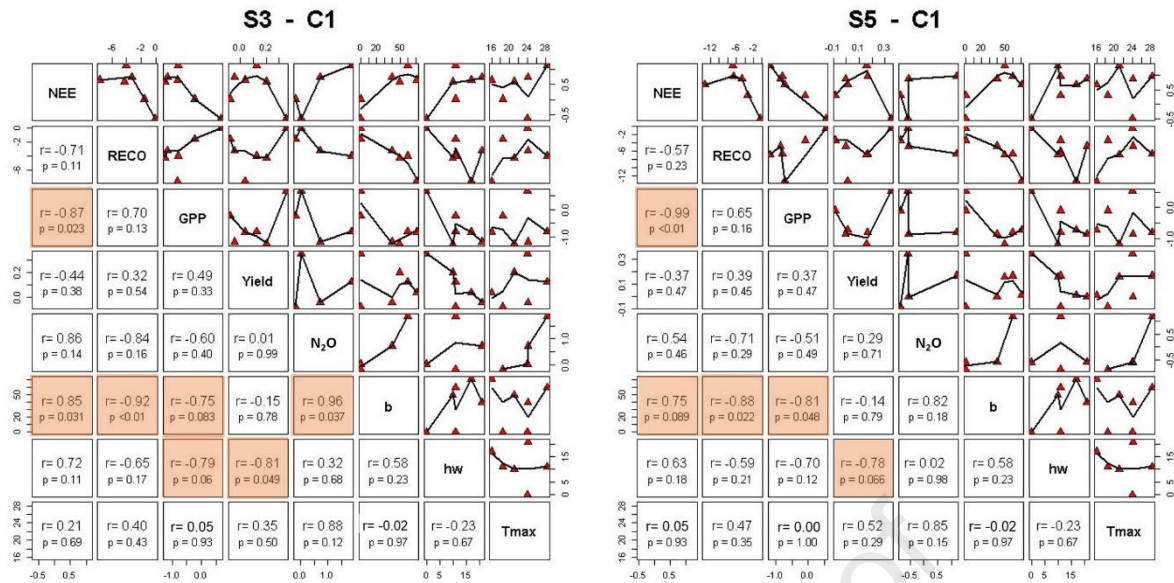
353
 354 Fig. 4. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-
 355 model medians (MMM) for stages 3 (left) and 5 (right) at G4 grassland site, and the annual
 356 agro-climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature
 357 (*Tmax*). Overlaid (red line) is a local non-parametric smoother curve. Coloured areas indicate
 358 significant correlations ($p < 0.10$).
 359

359

360 3.3. Residual analysis in cropland sites

361 The results of the residual analysis differ among cropland sites, with the strongest differences
 362 occurring at the most humid study-site (Fig. 1), i.e. C1 (Ottawa, Canada), with seven significant
 363 correlations at S3 (Fig. 5, left), which reduce to four at S5 (Fig. 5, right). As with G4, the
 364 negative correlation between NEE and GPP residuals at S3 ($r = -0.87$, $p < 0.02$) may indicate that
 365 an overestimation of NEE is likely to be the result of an underestimation of GPP, but this is not
 366 reflected in any other correlation between the model residuals ($p > 0.10$). However, at C1, all
 367 model residuals in S3 are significantly correlated with either the aridity index (NEE, $r = 0.85$,
 368 $p = 0.03$; RECO, $r = -0.92$, $p < 0.01$; N₂O, $r = 0.96$, $p = 0.04$), heatwaves (Yield, $r = -0.82$,
 369 $p = 0.05$) or both (GPP: aridity, $r = -0.75$, $p = 0.08$; heatwaves, $r = -0.79$, $p = 0.06$). These correlations are less
 370 important with fully calibrated models. While the residuals of NEE and GPP at C1 are still
 371 negatively correlated in S5 ($r = -0.99$, $p < 0.01$), among the environmental factors, it is essentially
 372 the aridity index that is positively (NEE, $r = 0.75$, $p = 0.09$) or negatively (GPP, $r = -0.81$,
 373 $p = 0.05$; RECO, $r = -0.88$, $p = 0.02$) correlated with C fluxes also after the full model calibration. The
 374 residuals of C and N fluxes are significantly correlated with aridity. GPP and Yield residuals
 375 are also negatively correlated with heatwaves.





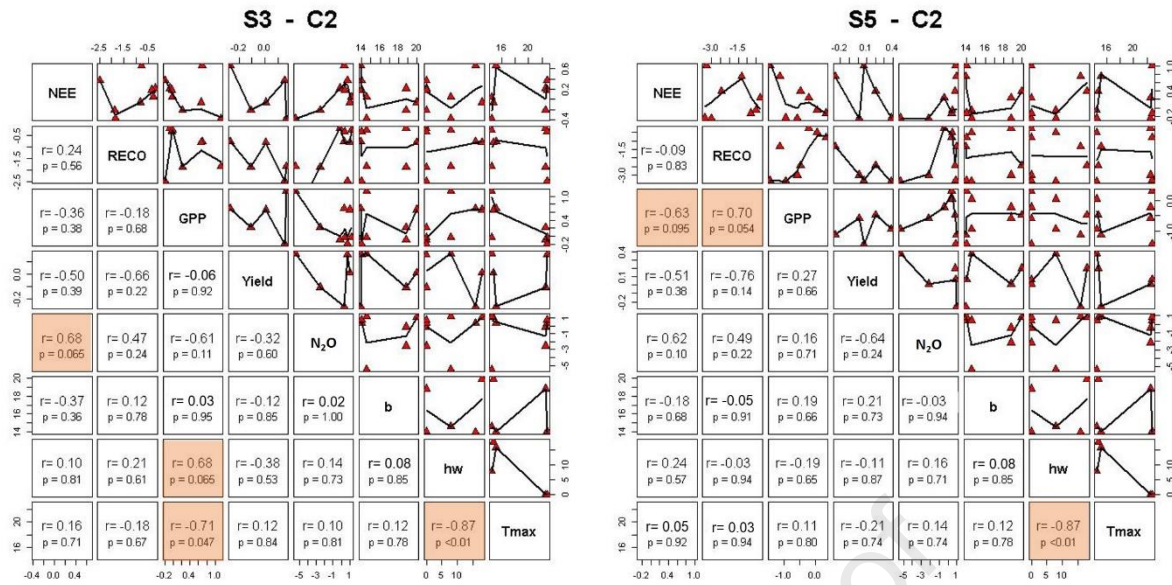
386

400

401 Fig. 5. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-
 402 model medians (MMM) for stages 3 (left) and 5 (right) at C1 cropland site, and the annual agro-
 403 climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature
 404 (*Tmax*). Overlaid (red line) is a local non-parametric smoother curve. Coloured areas indicate
 405 significant correlations ($p < 0.10$).

406

407 At C2 (Grignon, France), there was some significant positive correlations, e.g. between NEE
 408 and N₂O residuals at S3 (Fig. 6; $r = 0.68$, $p = 0.07$) and between RECO and GPP at S5 ($r = 0.70$,
 409 $p = 0.05$). However, some significant correlations between GPP residuals and climatic factors
 410 (heatwaves: $r = 0.68$, $p = 0.07$; *Tmax*: $r = -0.71$, $p = 0.05$) observed at S3 were no longer significant
 411 at S5 ($p > 0.10$).



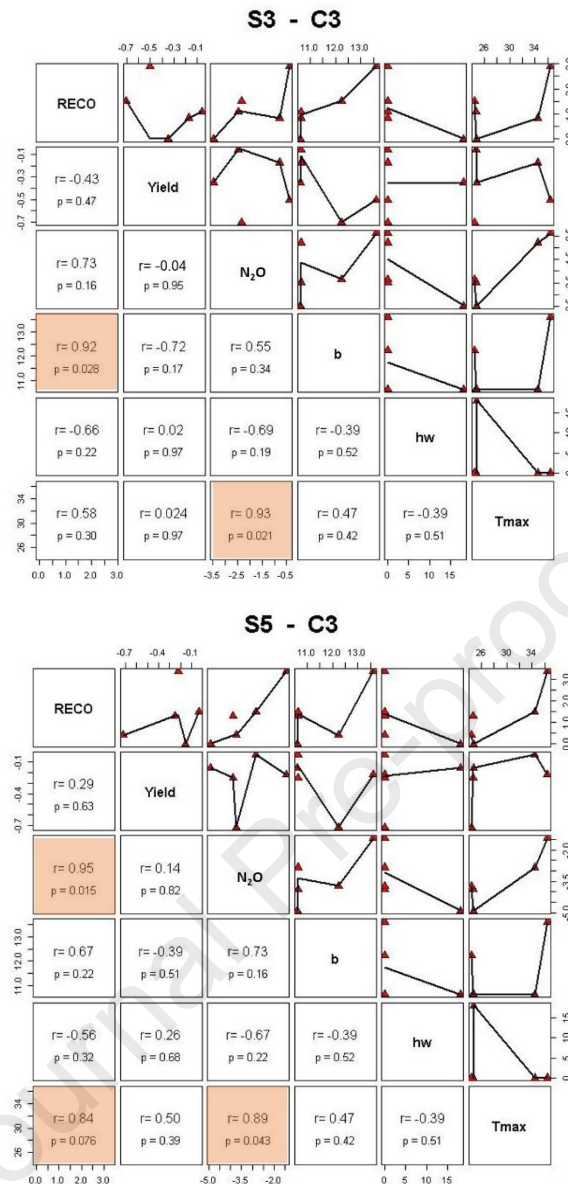
412

436 Fig. 6. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-
 437 model medians (MMM) for stages 3 (left) and 5 (right) at C2 cropland site, and the annual agro-
 438 climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature
 439 (*Tmax*). Overlaid (red line) is a local non-parametric smoother curve. Coloured areas indicate
 440 significant correlations ($p < 0.10$).

441

442 At the Indian site of Delhi (C3), where NEE and GPP data are not available, it is relevant to
 443 note the significant positive correlation observed between RECO and N₂O residuals at S5
 444 ($r = 0.95$, $p = 0.02$), not observed at S3 (Fig. 7). Then, there is a dependence of the simulation
 445 quality for these two fluxes on aridity (RECO: $r = 0.92$, $p = 0.03$) or *Tmax* (N₂O: $r = 0.93$, $p = 0.02$)
 446 at S3, or on *Tmax* only at S5 (RECO: $r = 0.84$, $p = 0.08$; N₂O: $r = 0.89$, $p = 0.04$).

447



481

484

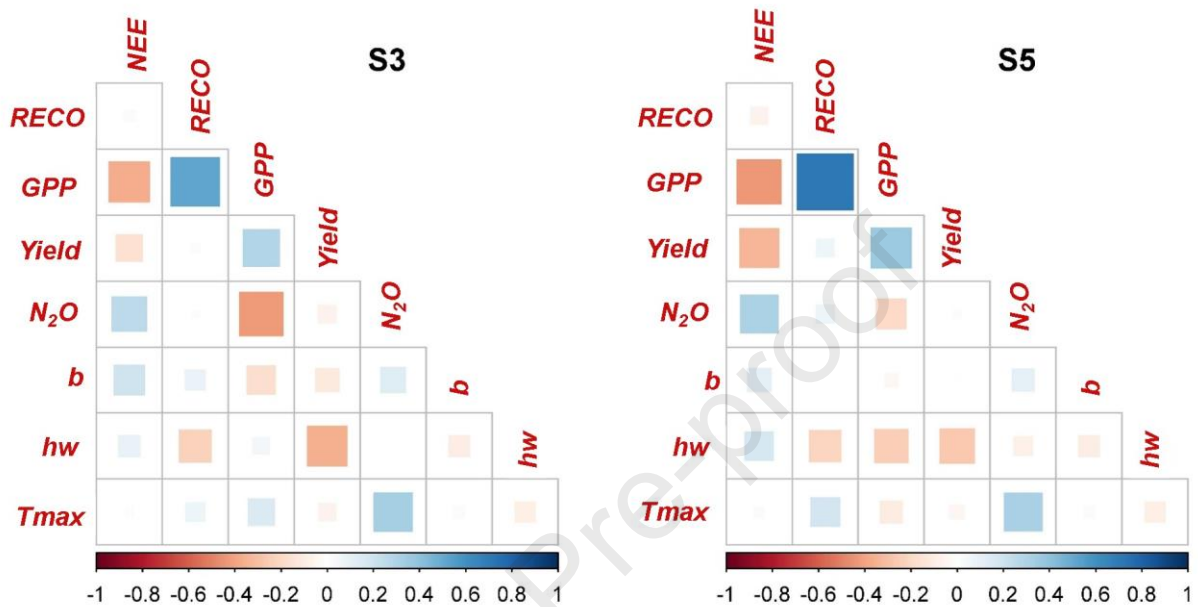
485 Fig. 7. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-
 486 model medians (MMM) for stages 3 (left) and 5 (right) at C3 cropland site, and the annual agro-
 487 climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature
 488 (*Tmax*). Overlaid (red line) is a local non-parametric smoother curve. Coloured areas indicate
 489 significant correlations ($p < 0.10$).

490

491 3.4. Geographical location, land use characteristics and calibration stages

492 Fig. 8 is a summary plot (correlogram) that averages the changes between partial (S3) and full
 493 (S5) calibration for each of the model residuals and weather metrics investigated. The heatmap
 494 values show mean correlation coefficients between model output residuals and weather drivers
 495 across all study-sites and land uses with partial and full calibration. Overall, there are quite
 496 strong positive correlations (on a gradient of $r \sim 0.5$ and $r \sim 0.7$) between GPP and RECO

497 residuals, and GPP residuals are negatively correlated with NEE residuals ($r \sim -0.4$). Although
 498 these correlations do not decrease with full calibration, we note that S5 markedly reduces the
 499 negative correlation between GPP and N_2O residuals ($r \sim -0.2$ from $r \sim -0.4$ at S3). At S5, we also
 500 observe near-zero correlations between yield and C-flux residuals and aridity conditions.
 501



502
 503 Fig. 8. Heatmap of mean correlation coefficients (r) between NEE, RECO, GPP and yield model
 504 residuals of multi-model medians (MMM) for stages 3 (left) and 5 (right) across sites/land uses,
 505 and the annual agro-climatic metrics aridity index (b), frequency of heatwaves (hw) and
 506 maximum temperature ($Tmax$).

507
 508 However, the multi-model simulations show complex patterns, illustrated by the analysis of
 509 land uses (grasslands, arable crops), study-sites (C1, C2, C3, G3 and G4) and calibration stages
 510 (S3 and S5) investigated, which show considerable differences in terms of correlation between
 511 model residuals, and between these residuals and weather metrics. Positive correlations were
 512 established between the RECO and GPP residuals at G3 (Fig. 3) and G4 (Fig. 4) in both
 513 calibration stages, and at C2 (Fig. 6) with fully calibrated models (along with a positive
 514 correlation between NEE and GPP residuals). At G4, positive correlations also characterise the
 515 relationships between GPP and Yield residuals (both calibration stages) and between RECO
 516 and NEE residuals (at S5). In addition, negative correlations were found at this site between
 517 NEE and GPP residuals (at S3), NEE and Yield residuals (at S3) and GPP and Yield residuals
 518 (at both calibration stages). At cropland site C1 (Fig. 5), NEE and GPP residuals are also
 519 negatively correlated (at both calibration stages). Overall, these results indicate that errors are

520 likely to be propagated through C-flux (and yield) predictions, and full calibration with plant,
521 soil and surface-to-atmosphere C-N fluxes does not always limit them. On the contrary, full
522 calibration can also increase the propagation of errors through C fluxes, as obtained in G4 with
523 RECO and NEE residuals (from $r \sim 0$ at S3 to highly significant $r=0.26$ at S5). However, while
524 many correlations between residuals are significant in G4, only the correlation between RECO
525 and GPP residuals at S3 ($r=0.86$) is high in this site.

526 The occurrence of intense weather factors such as high temperatures and arid conditions also
527 had significant effects on the model residuals. At cropland site C1, high negative correlations
528 between NEE and GPP residuals ($r=-0.87$ at S3 and $r=-0.99$ at S5) are accompanied by positive
529 high correlations between NEE residuals and the aridity index ($r=0.85$ at S3 and $r=0.75$ at S5),
530 while other negative correlations occurring between residuals and aridity (RECO and GPP) or
531 heatwaves (Yield) indicate higher residuals under more arid and hotter conditions.

532 In the Indian site (cropland site C3; Fig. 7), which is the most arid site investigated here (Fig.
533 1), we cannot explore the full correlation pattern of C-flux residuals because GPP and NEE
534 outputs are missing. However, we see that RECO residuals are positively correlated with the
535 aridity index at S3 ($r=0.92$, $p=0.03$), likely associated with the irrigation regime adopted in this
536 site (~ 250 mm yr^{-1} for spring wheat and >1000 mm yr^{-1} for rice), which may limit model
537 capacity in the presence of soil water saturation. Under these conditions, it appears that the
538 introduction of biogeochemical data in the calibration procedure (stage S5) becomes essential
539 to improve C-flux estimates (RECO residuals-aridity index $r=0.67$, $p=0.22$).

540

541 **4. Discussion**

542 This study provides a tentative answer to the question of whether, and to what extent, the results
543 of an ensemble of models can give insights into the limitations of the ensemble and offers
544 suggestions for model improvement. In particular, residual correlation matrices were used to
545 illustrate some of the main (and not unique) challenges of the emerging multi-model ensemble
546 approach in agricultural modelling to evaluate whether the overall pattern of model outputs can
547 help make progress in crop and grassland modelling by assessing model responses and
548 uncertainties against climatic factors. Focusing on the results of the ensemble, no attempt was
549 made to identify the best model(s) for crop and grassland C and N fluxes, and no probability of
550 success was assigned for the relevance of including or excluding one biogeochemical model
551 over another in the ensemble exercise.

552

553 *4.1. Residual analysis and model quality*

554 Residual analysis can help to find relationships between certain output variables, and between
555 output variables and external factors (and thus help to find additional variables that may need
556 to be included in the models as predictors, e.g. Medlyn et al., 2005). This analysis can indicate
557 the dependence of errors in case of error propagation in a model, although the mode of error
558 propagation cannot be attributed to a particular process using a correlation matrix. For instance,
559 overestimation of crop yields can lead to overestimation of shading of the soil surface by
560 (overestimated) plant biomass, which interferes with the simulation of soil heat and water
561 balances. Parallel to that, plant residues, senescent roots and the application of organic manure
562 feed the fresh organic matter pool of soil and are slowly decomposed after incorporation in soil.
563 Thus, biases in heat and water balances can interact with soil respiration, affecting the RECO
564 estimates and hence the C-budget estimates (i.e. NEE estimates). In this regard, it is notable
565 that significant correlations between NEE and Yield residuals were only observed in grassland
566 sites (at S3), where aboveground biomass and vegetation cover are continuously reduced by
567 grazing and recover after grazing cessation. In contrast, croplands are generally characterised
568 by alternating episodes of high C uptake or loss during the crop-growing season, directly related
569 to farmers' management practices like mineral fertilisation, grain and straw removal rates,
570 fallowing and tillage (Lehuger et al., 2010).

571 The net fixation of C being directly related to global solar radiation levels up to the saturation
572 point can lead to irregular patterns of net photosynthesis. Thus, while inaccurate simulations of
573 the soil water balance may affect plant biomass, e.g. due to an incorrect representation of the
574 effect of drought, it is also possible that inaccurate estimates of plant biomass (e.g. GPP) lead
575 to incorrect simulations of the water cycle due to an altered representation of evapotranspiration
576 or other water-related processes. Ensemble techniques are certainly a feasible method to
577 simulate biogeochemical processes in crops and grasslands, but model development is a must
578 to improve the multi-model approach (e.g. Hidy et al., 2016 for processes related to soil moisture
579 and N balance; Sándor et al., 2018b for the acclimation of grassland vegetation to temperature;
580 Liebermann et al., 2020 for feedbacks between different landscape compartments; Doro et al.,
581 2021 for soil heat transfer). In general, C fluxes (and interlinked N fluxes) remain difficult to
582 estimate in croplands and grasslands, likely due to incomplete representation of key functions
583 in models. For instance, rhizosphere-soil organic matter interactions, which include enzyme
584 production, maintenance and overflow metabolism, are mostly not represented (Cavalli et al.,
585 2019). Specifically, for grassland models, the simulation of biogeochemical cycles is generally
586 not coupled with simulation of plant species dynamics, which leads to considerable uncertainty
587 in the quality of estimates (van Oijen et al., 2020).

588

589 *4.2. Effects of agro-climatic factors*

590 While models estimating crop or pasture yields may not explicitly account for the impact of
591 heatwaves on grain or biomass formation (e.g. Harrison et al., 2017; Mangani et al., 2019), the
592 opposite impact of arid conditions on NEE (negative correlation) or RECO and GPP (positive
593 correlations) residuals is somewhat unexpected, considering that one variable (NEE) is the
594 difference of the two others. Considering that drought may be more effective in reducing CO₂
595 uptake by the plant than reducing ecosystem respiration (Gibelin et al., 2008; Nakano and
596 Shinoda, 2015), better results are provided when simulating NEE with a multi-model ensemble
597 (at C1 as at other sites, Fig. 2). This implies that there may be error compensation in the
598 ensemble. Greater coverage of plant and soil processes is also likely when more models are
599 used to simulate NEE than its basic components.

600 As far as N fluxes are concerned, N uptake by plants is computed by the models through a
601 supply/demand scheme, with soil supply depending mainly on soil nitrate and ammonium
602 concentrations and root length density (Lehuger et al., 2010). However, N₂O emissions are
603 mostly controlled by soil properties and local climate conditions, and current soil N levels, and
604 only to a lesser extent by the doses and types of N fertiliser applied (Butterbach-Bahl et al.,
605 2013). For instance, increasing bulk density decreases soil porosity and thereby increases the
606 likelihood of moisture conditions favourable to denitrification and N₂O emissions (Gabrielle et
607 al., 2006). As well, the correlation between N₂O and NEE residuals may be due to soil processes
608 because if heterotrophic respiration is too high there may be too many substrates (C and N)
609 available for nitrate respiration and denitrification (e.g. Rajta et al., 2020). The high negative
610 correlations ($r=-0.86$, $p=0.06$ and $r=-0.88$, $p=0.05$ at S3 and S5, respectively) between N₂O
611 residuals and aridity index at grassland site G3 reflect the deficit of moisture occurring mostly
612 in summer in central France (e.g. Klumpp et al., 2011), while in the wet climate of the United
613 Kingdom (grassland site G4) most nitrate available for leaching may result in reduced N₂O
614 emissions (e.g. Cardenas et al., 2013). In fact, grazed G4 grassland tends to have high N
615 leaching rates (and corresponding limited N₂O emissions) due to added urinary N to the system
616 and the non-uniform distribution of excreted organic N, which further enhances leaching due
617 to N hotspot formation (Jones et al., 2017b). N₂O emissions are reported to increase with
618 increasing temperature, which is attributed to an increase in the anaerobic volume fraction,
619 caused by an increased respiratory oxygen sink (Smith et al., 2018). With a mean annual
620 maximum annual temperature equal to 31.5 °C, N₂O residuals at the hot Indian cropland site

621 C3 are still positively correlated with T_{max} with fully calibrated models ($r=0.93$, $p=0.02$ at S3;
622 $r=0.89$, $p=0.04$ at S5).

623

624 5. Conclusions

625 Residuals from model-ensemble outputs tend to be less correlated when crop and grassland
626 models are calibrated using soil and C-N fluxes together with vegetation data (compared to
627 partial calibration with vegetation data alone). If full calibration can reduce the correlation
628 between C- and N-flux residuals (e.g. between GPP and N₂O residuals), intense weather factors
629 can also have significant effects on model residuals (e.g. N₂O residuals positively correlated
630 with maximum air temperature at the hot Indian cropland site). However, complex multi-model
631 simulation patterns indicate that full calibration does not always constrain the correlation
632 between model residuals, and between these residuals and agro-climatic metrics. Our
633 assessment, which remains limited to climate-related drivers calculated annually (and could
634 then a future improvement be a seasonal climate analysis), holds potential for a wider analysis
635 that surveys contextual soil and management factors, for which the current database was not
636 designed. In that, we have proposed a somewhat *ad hoc* multi-output analysis that considers
637 inter-dependencies in the model outputs, but there are challenges that require further work.
638 These include how to quantitatively account for consistency with mechanistic viewpoints
639 supported by alternative models of varying complexity as a further important requirement for
640 model ensembles, as well as definitions of core concepts and metrics to provide a quantitative
641 determination of the stability of simulation results under a variety of conditions. These
642 challenges are interesting from a practical point of view because improving our understanding
643 of these issues and finding better ways to deal with the plurality of models has the potential to
644 increase the value of biogeochemical models in agriculture, where determining the robustness
645 of results is a strategy to assess confidence in results. In the end, this may provide modellers
646 with a clearer explanation of what they are doing in ensemble modelling (as well as how they
647 are doing it), and stronger arguments as to when ensemble modelling can, or cannot, become a
648 practical epistemic resource.

649 One of the features of C-N modelling today is the huge quantity and variety of models available.
650 Our analysis, which did not consider all sources of uncertainty (e.g. the influence of the unique
651 choices made by modellers), relied on the integration of several modelling teams into an
652 ensemble protocol. Comparing different approaches have revealed great model diversity and
653 the need to accommodate challenges experienced by modellers (including initialization and
654 calibration procedures), as reflected in the co-creation (with modellers and data providers) of

655 alternative calibration scenarios. The distinction between partial and full calibration, limited
656 here to the two most relevant calibration stages, i.e. with plant data only (S3) and with plant,
657 soil physical and biogeochemical data (S5), introduced and formalised a dialectical perspective
658 (or compromise approach) between what is practical to implement for the users and
659 beneficiaries of models (S3) and what constitutes (scientifically) the best modelling practice
660 (S5). In fact, with overall lower or less significant correlations obtained with the fully calibrated
661 models, the centrality of the S5 calibration scenario emerges overall if not for the practical
662 implementation of model ensembles (which requires simplified datasets), for the identification
663 of areas of model structures requiring further development. All this considered, this study on
664 ensemble results presents important elements that can lead individual modelling teams to
665 identify a spectrum of actions for model (and modelling practice) improvement.

666

667 **Acknowledgements**

668 This study was coordinated by the Integrative Research Group of the Global Research Alliance
669 (GRA) on agricultural GHGs and was supported by five research projects (CN-MIP,
670 Models4Pastures, MACSUR, COMET-Global and MAGNET), which received funding by a
671 multi-partner call on agricultural greenhouse gas research of the Joint Programming Initiative
672 'FACCE' through its national financing bodies. It falls within the thematic area of the French
673 government IDEX-ISITE initiative (reference: 16-IDEX-0001; project CAP 20-25). We
674 acknowledge funding for the data collection through the EU projects GREENGRASS (EC
675 EVK2-CT2001-00105), CarboEurope (GOCE-CT-2003-505572) and NitroEurope (017841).
676 US acknowledges SRUC's contribution (Stephanie K. Jones and Robert M. Rees) to compile
677 the data of the C4 grassland site (Easter Bush, UK). The research in support of C1(Ottawa, ON,
678 Canada) site data acquisition was conducted with the financial support of Agriculture and Agri-
679 Food Canada A-base funding. Data for the C2 cropland site (Grignon, France) were obtained
680 from the Fr-Gri ecosystem site ICOS (Integrated Carbon Observation System;
681 <https://www.icos-cp.eu>), for which we thank Pauline Buysse and Benjamin Loubet (INRAE,
682 Grignon) for access. Data for the G3 grassland site (Laqueuille, France) were obtained from the
683 FR-Lq1 SOERE-ACBB (*Système D'observation Et D'expérimentation Sur Le Long Terme Pour*
684 *La Recherche En Environnement - Agro-Écosystème, Cycle Bio-Géochimique Et Biodiversité*;
685 <https://www.soere-acbb.com>) ecosystem site (ICOS) financed by French National Agency for
686 Research (ANAAE-F, ANR-11-INBS-0001). SR (PIK) acknowledges financial support from
687 the BMBF (Federal Ministry of Education and Research of Germany) for funding of the projects
688 MACMIT (grant 01LN1317A) and Climasteppe (grant 01DJ18012). RS and GB received

689 mobility funding from the French-Hungarian bilateral partnership through the BALATON (N°
690 44703TF)/TéT (2019-2.1.11-TÉT-2019-00031) programme.

691

692 **References**

693 Alcock, D.J., Harrison, M.T., Rawnsley, R.P., Eckard, R.J., 2015. Can animal genetics and
694 flock management be used to reduce greenhouse gas emissions but also maintain
695 productivity of wool-producing enterprises? *Agricultural Systems* 132, 25-34.

696 Allard, V., Soussana, J.-F., Falcimagne, R., Berbigier, P., Bonnefond, J.M., Ceschia, E.,
697 D'hour, P., Hénault, C., Laville, P., Martin, C., Pinarès-Patino, C., 2007. The role of grazing
698 management for the net biome productivity and greenhouse gas budget (CO₂, N₂O and CH₄)
699 of semi-natural grassland. *Agriculture, Ecosystem & Environment* 12, 47-58.

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

703 Bassu, S., Brisson, N., Durand, J.L., Boote, K.J., Lizaso, J., Jones, J.W., Rosenzweig, C., Adam,
704 M., Basso, B., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M.,
705 Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C.,
706 Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.-H., Kumar, N.S., Makowski,
707 D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F.,
708 Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop models vary in their
709 responses to climate change factors? *Global Change Biology* 20, 2301-2320.

710 Bellocchi, G., Acutis, M., Fila, G., Donatelli, M., 2002. An indicator of solar radiation model
711 performance based on a fuzzy expert system. *Agronomy Journal* 94, 1222-1233.

712 Bellocchi, G., Rivington, M., Donatelli, M., Matthews, K., 2010. Validation of biophysical
713 models: issues and methodologies. A review. *Agronomy for Sustainable Development* 30,
714 109-113.

715 Bhatia, A., Pathak, H., Jain, N., Singh, P.K., Tomer, R., 2012. Greenhouse gas mitigation in
716 rice-wheat system with leaf color chart-based urea application. *Environmental Monitoring
717 and Assessment* 184, 3095-3107.

718 Bilotto, F., Harrison, M.T., Migliorati, M.D.A., Christie, K.M., Rowlings, D.W., Grace, P.R.,
719 Smith, A.P., Rawnsley, R.P., Thorburn, P.J., Eckard, R.J., 2021. Can seasonal soil N
720 mineralisation trends be leveraged to enhance pasture growth? *Science of the Total
721 Environment* 772: 145031.

- 722 Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C.D., Doro, L.,
723 Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I.,
724 Klumpp, K., Léonard, J., Martin, R., Massad, R.S., Recous, S., Seddaiu, G., Sharp, J., Smith,
725 P., Smith, W.N., Soussana, J-F., Bellocchi, G., 2017. Review and analysis of strengths and
726 weaknesses of agro-ecosystem models for simulating C and N fluxes. *Sci. Total Environ.*
727 598, 445-470.
- 728 Butterbach-Bahl, K., Baggs, E.M., Dannenmann, M., Kiese, R., Zechmeister-Boltenstern, S.,
729 2013. Nitrous oxide emissions from soils: how well do we understand the processes and their
730 controls? *Philosophical Transactions of the Royal Society B* 368:20130122.
- 731 Calanca, P., Deléglise, C., Martin, R., Carrère, P., Mosimann, E., 2016. Testing the ability of a
732 simple grassland model to simulate the seasonal effects of drought on herbage growth. *Field*
733 *Crops Research* 187, 12-23.
- 734 Calder, M., Craig, C., Culley, D., de Cani, R., Donnelly, C.A., Douglas, R., Edmonds, B.,
735 Gascoigne, J., Gilbert, N., Hargrove, C., Hinds, D., Lane, D.C., Mitchell, D., Pavey, G.,
736 Robertson, D., Rosewell, B., Sherwin, S., Walport, M., Wilson, A., 2018. Computational
737 modelling for decision-making: where, why, what, who and how. *Royal Society Open*
738 *Science* 5:172096.
- 739 Cardenas, L.M., Gooday, R., Brown, L., Scholefield, D., Cuttle, S., Gilhespy, S., Matthews, R.,
740 Misselbrook, T., Wang, J., Li, C., Hughes, G., Lord, E., 2013. Towards an improved
741 inventory of N₂O from agriculture: Model evaluation of N₂O emission factors and N fraction
742 leached from different sources in UK agriculture. *Atmospheric Environment* 79, 340–348.
- 743 Cavalli, D., Bellocchi, G., Corti, M., Gallina, P.M., Bechini, L., 2019. Sensitivity analysis of C
744 and N modules in biogeochemical crop and grassland models following manure addition to
745 soil. *European Journal of Soil Science* 70, 833-846.
- 746 Challinor, A.J., Smith, M.S., Thornton, P., 2013. Use of agro-climate ensembles for quantifying
747 uncertainty and informing adaptation. *Agricultural and Forest Meteorology* 170, 2-7.
- 748 Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. *Journal*
749 *of the American Statistical Association* 74, 829-836.
- 750 Confalonieri, R., Bellocchi, G., Donatelli, M., 2010. A software component to compute agro-
751 meteorological indicators. *Environmental Modelling & Software* 25, 1485-1486.
- 752 De Martonne, E., 1942. Nouvelle carte mondiale de l'indice d'aridité. *Annales de Géographie*
753 51, 242-250. (in French)
- 754 Dijkstra, T.K., 1988. On model uncertainty and its statistical implications. Springer Verlag,
755 Berlin, Germany.

- 756 Diodato, N., Ceccarelli, M., 2004. Multivariate indicator Kriging approach using a GIS to
757 classify soil degradation for Mediterranean agricultural lands. *Ecological Indicators* 4, 177-
758 187.
- 759 Doro, L., Wang, X., Ammann, C., De Antoni Migliorati, M., Grünwald, T., Klumpp, K.,
760 Loubet, B., Pattey, E., Wohlfahrt, G., Williams, J.R., Norfleet, M.L., 2021. Improving the
761 simulation of soil temperature within the EPIC model. *Environmental Modelling & Software*
762 144:105140.
- 763 Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., McAuliffe, R., Recous, S., Sándor, R.,
764 Smith, P., Snow, V., Migliorati, M.D.A., Basso, B., Bhatia, A., Brillì, L., Doltra, J., Dorich,
765 C.D., Doro, L., Fitton, N., Giacomini, S.J., Grant, B., Harrison, M.T., Jones, S.K.,
766 Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Liebig, M., Lieffering, M.,
767 Martin, R., Massad, R.S., Meier, E., Merbold, L., Moore, A.D., Myrriotis, V., Newton, P.,
768 Pattey, E., Rolinski, S., Sharp, J., Smith, W.N., Wu, L., Zhang, Q., 2018. Assessing
769 uncertainties in crop and pasture ensemble model simulations of productivity and N₂O
770 emissions. *Global Change Biology* 24, e603-e616.
- 771 Eza, U., Shtiliyanova, A., Borrás, D., Bellocchi, G., Carrère, P., Martin, R., 2015. An open
772 platform to assess vulnerabilities to climate change: An application to agricultural systems.
773 *Ecological Informatics* 30, 389-396.
- 774 Farina, R., Sándor, R., Abdalla, M., Álvaro-Fuentes, J., Bechini, L., Bolinder, M.A., Brillì, L.,
775 Chenu, C., Clivot, H., De Antoni Migliorati, M., Di Bene, C., Dorich, C.D., Ehrhardt, F.,
776 Ferchaut, F., Fitton, N., Francaviglia, R., Franko, U., Giltrap, D.L., Grant, B.B., Guenet, B.,
777 Harrison, M.T., Kirschbaum, M.U.F., Kuka, K., Kulmala, L., Liski, J., McGrath, M.J.,
778 Meier, E., Menichetti, L., Moyano, F., Nendel, C., Recous, S., Reibold, N., Shepherd, A.,
779 Smith, W.N., Smith, P., Soussana, J.F., Stella, T., Taghizadeh-Toosi, A., Tsutskikh, E.,
780 Bellocchi, G., 2021. Ensemble modelling, uncertainty and robust predictions of organic
781 carbon in long-term bare-fallow soils. *Global Change Biology* 27, 904-928.
- 782 Gabrielle, B., Laville, P., Duval, O., Nicoullaud, B., Germon, J. C., Hénault, C., 2006. Process-
783 based modeling of nitrous oxide emissions from wheat-cropped soils at the subregional
784 scale. *Global Biogeochemical Cycles* 20: GB4018.
- 785 Gibelin, A.-L., Calvet, J.-C., Viovy, N., 2008. Modelling energy and CO₂ fluxes with an
786 interactive vegetation land surface model - Evaluation at high and middle latitudes.
787 *Agricultural and Forest Meteorology* 148, 1611-1628.
- 788 Falconnier, G.N., Corbeels, M., Boote, K.J., Affholder, F., Adam, M., MacCarthy, D.S., Ruane,
789 A.C., Nendel, C., Whitbread, A.M., Justes, É., Ahuja, L.R., Akinseye, F.M., Alou, I.N.,

- 790 Amouzou, K.A., Anapalli, S.S., Baron, C., Basso, B., Baudron, F., Bertuzzi, P., Challinor,
791 A.J., Chen, Y., Deryng, D., Elsayed, M.L., Faye, B., Gaiser, T., Galdos, M., Gayler, S.,
792 Gerardeaux, E., Giner, M., Grant, B., Hoogenboom, G., Ibrahim, E.S., Kamali, B.,
793 Kersebaum, K.C., Kim, S.-H., van der Laan, M., Leroux, L., Lizaso, J.I., Maestrini, B.,
794 Meier, E.A., Mequanint, F., Ndoli, A., Porter, C.H., Priesack, E., Ripoche, D., Sida, T.S.,
795 Singh, U., Smith, W.N., Srivastava, A., Sinha, S., Tao, F., Thorburn, P.J., Timlin, D., Traore,
796 B., Twine, T., Webber, H., 2020. Modelling climate change impacts on maize yields under
797 low nitrogen input conditions in sub-Saharan Africa. *Global Change Biology* 26, 5942-5964.
- 798 Graux, A.-I., Bellocchi, G., Lardy, R., Soussana, J.-F., 2013. Ensemble modelling of climate
799 change risks and opportunities for managed grasslands in France. *Agricultural and Forest*
800 *Meteorology* 170, 114-131.
- 801 Harrison, M.T., Cullen, B.R., Armstrong, D., 2017. Management options for dairy farms under
802 climate change: Effects of intensification, adaptation and simplification on pastures, milk
803 production and profitability. *Agricultural Systems* 155, 19-32.
- 804 Harrison, M.T., Cullen, B.R., Tomkins, N.W., McSweeney, C., Cohn, P., Eckard, R.J., 2016.
805 The concordance between greenhouse gas emissions, livestock production and profitability
806 of extensive beef farming systems. *Animal Production Science* 56, 370-384.
- 807 Harrison, M.T., Evans, J.R., Moore, A.D., 2012. Using a mathematical framework to examine
808 physiological changes in winter wheat after livestock grazing: 1. Model derivation and
809 coefficient calibration. *Field Crops Research* 136, 116-126.
- 810 Harrison, M.T., Roggero, P.P., Zavattaro, L., 2019. Simple, efficient and robust techniques for
811 automatic multi-objective function parameterisation: Case studies of local and global
812 optimisation using APSIM. *Environmental Modelling & Software* 117, 109-133.
- 813 Hidy, D., Barcza, Z., Marjanovič, H., Ostrogovič Sever, M.Z., Dobor, L., Gelybó, Gy., Fodor,
814 N., Pintér, K., Churkina, G., Running, S.W. Thornton, P.E., Bellocchi, G., Haszpra, L.,
815 Horváth, F., Suyker, A., Nagy, Z., 2016. Terrestrial ecosystem process model Biome-
816 BGCMuSo: summary of improvements and new modeling possibilities. *Geoscientific Model*
817 *Development* 9, 4405-4437.
- 818 Jégo, G., Pattey, E., Liu, J., 2012. Using leaf area index, retrieved from optical imagery, in the
819 STICS crop model for predicting yield and biomass of field crops. *Field Crops Research*
820 131, 63-74.
- 821 Jones, J.W., Antle, J.M., Basso, B.O., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J.,
822 Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Muñoz-Carpena, R., Porter, C.H.,

- 823 Rosenzweig, C., Wheeler, T.R., 2017a. Brief history of agricultural systems modelling.
824 Agricultural Systems 155, 240-254.
- 825 Jones, S.K., Helfter, C., Anderson, M., Coyle, M., Campbell, C., Famulari, D., Di Marco, C.,
826 van Dijk, N., Topp, C.F.E., Kiese, R., Kindler, R., Siemens, J., Schrumpf, M., Kaiser, K.,
827 Nemitz, E., Levy, P., Rees, R.M., Sutton, M.A., Skiba, U.M., 2017b. The nitrogen, carbon
828 and greenhouse gas budget of a grazed, cut and fertilised temperate grassland.
829 Biogeosciences 14, 2069-2088.
- 830 Klumpp, K., Tallec, T., Guix, N., Soussana, J.-F., 2011. Long-term impacts of agricultural
831 practices and climatic variability on carbon storage in a permanent pasture. Global Change
832 Biology 17, 3534-3545.
- 833 Knutti, R., Baumberger, C., Hirsch Hadorn, G., 2019. Uncertainty quantification using multiple
834 models - prospects and challenges. In: Beisbart C., Saam N.J. (eds.) Computer simulation
835 validation: fundamental concepts, methodological frameworks, and philosophical
836 perspectives. Springer: Cham, pp. 835–855.
- 837 Kollas, C., Kersebaum, K.C., Nendel, C., Manevski, K., Müller, C., Palosuo, T., Armas-
838 Herrera, C.M., Beaudoin, N., Bindi, M., Charfeddine, M., Conradt, T., Constantin, J.,
839 Eitzinger, J., Ewert, F., Ferrise, R., Gaiser, T., Garcia de Cortazar-Atauri, I., Giglio, L.,
840 Hlavinka, P., Hoffmann, H., Hoffmann, M.P., Launay, M., Manderscheid, R., Mary, B.,
841 Mirschel, W., Moriondo, M., Olesen, J.E. Öztürk, I., Pacholski, A., Ripoche-Wachter, D.,
842 Roggero, P.P., Roncossek, S., Rötter, R.P., Ruget, F., Sharif, B., Trnkam, M., Ventrella, D.,
843 Waha, K., Wegehenkel, M., Weigel, H.-J., Wu, L., 2015. Crop rotation modelling - A
844 European model intercomparison. European Journal of Agronomy 70, 98–111.
- 845 Lardy, R., Bachelet, B., Bellocchi, G., Hill, D.R.C., 2014. Towards vulnerability minimization
846 of grassland soil organic matter using metamodels. Environmental Modelling & Software
847 52, 38-50.
- 848 Lardy, R., Bellocchi, G., Martin, R., 2015. Vuln-Indices: Software to assess vulnerability to
849 climate change. Computers and Electronics in Agriculture 114, 53-57.
- 850 Laville, P., Lehuger, S., Loubet, B., Chaumartin, F., Cellier, P., 2011. Effect of management,
851 climate and soil conditions on N₂O and NO emissions from an arable crop rotation using
852 high temporal resolution measurements. Agricultural and Forest Meteorology 151, 228-240.
- 853 Lehuger, S., Gabrielle, B., Cellier, P., Loubet, B., Roche, R., Béziat, P., Ceschia, E.,
854 Wattenbach, M., 2010. Predicting the net carbon exchanges of crop rotations in Europe with
855 an agro-ecosystem model. Agriculture, Ecosystems & Environment 139, 384-395.

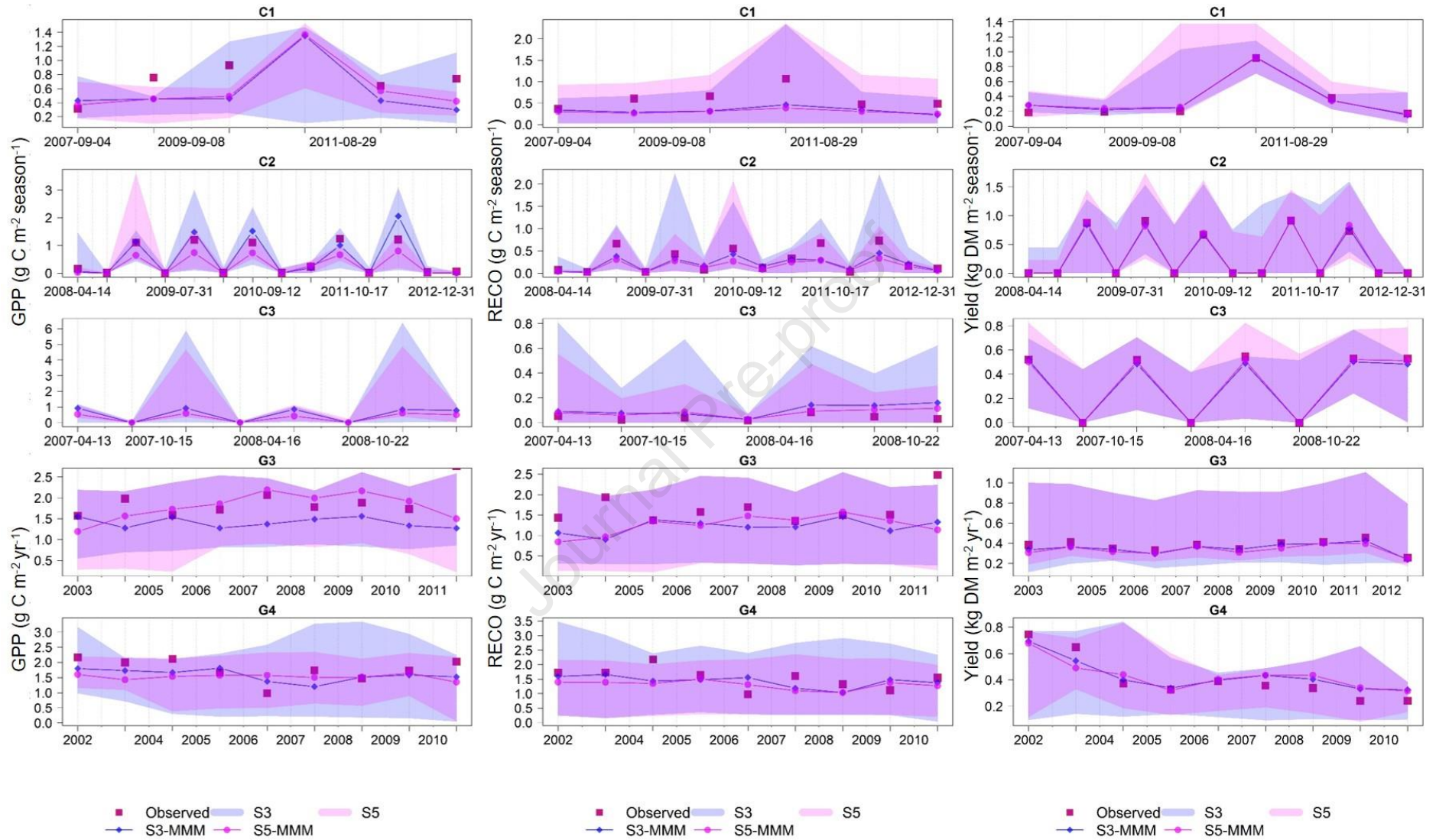
- 856 Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S.,
857 Confalonieri, R., Fumoto T., Gaydon, D., Marcaida III, M., Nakagawa, H., Oriol, P., Ruane,
858 A.C., Ruget, F., Balwinder -Singh, B., Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida,
859 H., Zhang, Z., Bouman, B., 2015. Uncertainties in predicting rice yield by current crop
860 models under a wide range of climatic conditions. *Global Change Biology* 21, 1328–1341.
- 861 Liebermann, R., Breuer, L., Houska, T., Kraus, D., Moser, G., Kraft, P., 2020. Simulating long-
862 term development of greenhouse gas emissions, plant biomass, and soil moisture of a
863 temperate grassland ecosystem under elevated atmospheric CO₂. *Agronomy* 10:50.
- 864 Loubet, B., Laville, P., Lehuger, S., Larmanou, E., Flechard, C., Mascher, N., Genermont, S.,
865 Roche, R., Ferrara, R. M., Stella, P., Personne, E., Durand, B., Decuq, C., Flura, D., Masson,
866 S., Fanucci, O., Rampon, J.-N., Siemens, J., Kindler, R., Gabrielle, B., Schrumpf, M.,
867 Cellier, P., 2011. Carbon, nitrogen and greenhouse gases budgets over a four years crop
868 rotation in northern France. *Plant and Soil* 343, 109-137.
- 869 Mangani, R., Tesfamariam, E.H., Engelbrecht, C.J., Bellocchi, G., Hassen, A., Mangani, T.,
870 2019. Potential impacts of extreme weather events in main maize (*Zea mays* L.) producing
871 areas of South Africa under rainfed conditions. *Regional Environmental Change* 19, 1441–
872 1452.
- 873 Martin, G., Allain, S., Bergez, J.-E., Burger-Leenhardt, D., Constantin, J., Duru, M., Hazard,
874 L., Lacombe, C., Magda, D., Magne, M.-A., Ryschawy, J., Thénard, V., Tribouillois, H.,
875 Willaume, M., 2018. How to address the sustainability transition of farming systems? A
876 conceptual framework to organize research. *Sustainability* 10:2083.
- 877 Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., Boote, K.J., Ruane,
878 A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K.,
879 Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J.,
880 Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J.,
881 Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O’leary, G.,
882 Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A.,
883 Shcherback, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F.,
884 Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat
885 growth: many models are better than one. *Global Change Biology* 21, 911-925.
- 886 Matthews, K.B., Rivington, M., Buchan, K., Miller, D.G., Bellocchi, G., 2008. Characterising
887 the agro-meteorological implications of climate change scenarios for land management
888 stakeholders. *Climate Research* 37, 59-75.

- 889 Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models
890 of forest CO₂ exchange using eddy covariance data: some perils and pitfalls. *Tree Physiology*
891 25, 839–857.
- 892 Nakano, T., Shinoda, M., 2015. Modeling gross primary production and ecosystem respiration
893 in a semiarid grassland of Mongolia. *Soil Science and Plant Nutrition* 61, 106-115.
- 894 Pattey, E., Edwards, G., Strachan, I.B., Desjardins, R.L., Kaharabata, S., Wagner, C., 2006.
895 Towards standards for measuring greenhouse gas fluxes from agricultural fields using
896 instrumented towers. *Canadian Journal of Soil Science* 86, 373-400.
- 897 R Core Team, 2020. A language and environment for statistical computing. R Foundation for
898 Statistical Computing, Vienna, Austria. <https://www.R-project.org>
- 899 Raj, R., Hamm, N.A.S., van de Tol, C., Stein, A., 2006. Uncertainty analysis of gross primary
900 production partitioned from net ecosystem exchange measurements. *Biogeosciences* 13,
901 1409-1422.
- 902 Rajta, A., Bhatia, R., Setia, H., Pathania, P., 2020. Role of heterotrophic aerobic denitrifying
903 bacteria in nitrate removal from wastewater. *Journal of Applied Microbiology* 128, 1261-
904 1278.
- 905 Riccio, G., Giunta, G., Galmarini, S., 2007. Seeking for the rational basis of the Median Model:
906 the optimal combination of multi-model ensemble results. *Atmospheric Chemistry and*
907 *Physics* 7, 6085-6098.
- 908 Riggers, C., Poeplau, C., Don, A., Bamminger, C., Höper, H., Dechow, R., 2019. Multi-model
909 ensemble improved the prediction of trends in soil organic carbon stocks in German
910 croplands. *Geoderma* 345, 17-30.
- 911 Rivington, M., Matthews, K.B., Bellocchi, G., Buchan, K., Stöckle, C.O., Donatelli, M., 2007.
912 An integrated assessment approach to conduct analyses of climate change impacts on whole-
913 farm systems. *Environmental Modelling & Software* 22, 202-210.
- 914 Rivington, M., Matthews, K.B., Buchan, K., Miller, D.G., Bellocchi, G., Russell, G., 2013.
915 Climate change impacts and adaptation scope for agriculture indicated by agro-
916 meteorological metrics. *Agricultural Systems* 114, 15-31.
- 917 Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J.,
918 Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid,
919 E., Stehfest, E., Yang, H., Jones, J.W., 2014. Assessing agricultural risks of climate change
920 in the 21st century in a global gridded crop model intercomparison. *Proc. Natl. Acad. Sci.*
921 *USA* 111, 3268-3273.

- 922 Ruane, A.C., Hudson, N.I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., Boote, K.J.,
923 Thorburn, P.J., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson,
924 N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J.,
925 Hunt, L.A., Ingwersen, J., Izaurrealde, R.C., Kersebaum, K.C., Kumar, S.N., Müller, C.,
926 Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D.,
927 Rötter, R.P., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P.,
928 Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J.W., Wolf, J.,
929 2016. Multi-wheat-model ensemble responses to interannual climate variability.
930 *Environmental Modelling & Software* 81, 86-101.
- 931 Ruane, A.C., Rosenzweig, C., Asseng, S., Boote, K.J., Elliott, J., Ewert, F., Jones, J.W., Martre,
932 P., McDermid, S.P., Müller, C., Snyder, A., Thorburn, P.J., 2017. An AgMIP framework for
933 improved agricultural representation in integrated assessment models. *Environmental*
934 *Research Letters* 12: 125003.
- 935 Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E.,
936 Ma, S., Perego, A., Rolinski, S., Ruget, F., Sanna, M., Seddaiu, G., Wu, L., Bellocchi, G.,
937 2017. Multi-model simulation of soil temperature, soil water content and biomass in Euro-
938 Mediterranean grasslands: Uncertainties and ensemble performance. *European Journal of*
939 *Agronomy* 88, 22-40.
- 940 Sándor, R., Barcza, Z., Hidy, D., Lellei-Kovács, E., Ma, S., Bellocchi, G., 2016. Modelling of
941 grassland fluxes in Europe: evaluation of two biogeochemical models. *Agriculture,*
942 *Ecosystem & Environment* 215, 1-19.
- 943 Sándor, R., Ehrhardt, F., Brillì, L., Carozzi, M., Recous, S., Smith, P., Snow, V., Soussana, J.F.,
944 Dorich, C.D., Fuchs, K., Fitton, N., Gongadze, K., Klumpp, K., Liebig, M., Martin, R.,
945 Merbold, L., Newton, P.C.D., Rees, R.M., Rolinski, S., Bellocchi, G., 2018a. The use of
946 biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed
947 grasslands. *Science of the Total Environment* 15, 292-306.
- 948 Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B.,
949 Bhatia, A., Brillì, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T.,
950 Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.S., Moore,
951 A., Myrriotis, V., Pattey, E., Rolinski, R., Sharp, J., Skiba, U., Smith, W., Wu, L., Zhang,
952 Q., Bellocchi, G., 2020. Ensemble modelling of carbon fluxes in grasslands and croplands.
953 *Field Crops Research* 252: 107791.

- 954 Sándor, R., Picon-Cochard, C., Martin, R., Louault, F., Klumpp, K., Borrás, D., Bellocchi, G.,
955 2018b. Plant acclimation to temperature: Developments in the Pasture Simulation model.
956 *Field Crops Research* 222, 238-255.
- 957 Sansoulet, J., Pattey, E., Kröbel, R., Grant, B., Smith, W., Jégo, G., Desjardins, R.L., Tremblay,
958 N., Tremblay, G., 2014. Comparing the performance of the STICS, DNDC, and DayCent
959 models for predicting N uptake and biomass of spring wheat in Eastern Canada. *Field Crops*
960 *Research* 156, 135-150.
- 961 Skiba, U., Jones, S.K., Drewer, J., Helfter, C., Anderson, M., Dinsmore, K., McKenzie, R.,
962 Nemitz, E., Sutton, M.A., 2013. Comparison of soil greenhouse gas fluxes from extensive
963 and intensive grazing in a temperate maritime climate. *Biogeosciences* 10, 1231-1241.
- 964 Smith, K.A., Ball, T., Conen, F., Dobbie, K.E., Massheder, J., Rey, A., 2018. Exchange of
965 greenhouse gases between soil and atmosphere: interactions of soil physical factors and
966 biological processes. *European Journal of Soil Science* 69, 10-20.
- 967 Snow, V., Rotz, C.A., Moore, A.D., Martin-Clouaire, R., Johnson, I.R., Hutchings, N.J.,
968 Eckard, R.J., 2014. The challenges - and some solutions - to process-based modelling of
969 grazed agricultural systems. *Environmental Modelling & Software* 62, 420-436.
- 970 Spence, M.A., Blanchard, J.L., Rossberg, A.G., Heath, M.R., Heymans, J.J., Mackinson, S.,
971 Serpetti, N., Speirs, D., Thorpe, R.B., Blackwell, P.G., 2017. Multi-model ensembles for
972 ecosystem prediction. arXiv: 1709.05189.
- 973 Van Oijen, M., Barcza Z., Confalonieri R., Korhonen P., Kröel-Dulay G., Lellei-Kovács E.,
974 Louarn G., Louault F., Martin R., Moulin T., Movedi E., Picon-Cochard C., Rolinski S.,
975 Viovy N., Wirth S.B., Bellocchi, G., 2020. Incorporating biodiversity into biogeochemistry
976 models to improve prediction of ecosystem services in temperate grasslands: review and
977 roadmap. *Agronomy* 10: 259.
- 978 Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., van Ittersum, M.,
979 Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De
980 Sanctis, G., Dumont, B., Rezaei, E.E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila,
981 M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C., Jones, C.D.,
982 Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A.-K., Maiorano, A., Minoli, S., Müller,
983 C., Kumar, S.N., Nendel, C., O'Leary, G.J., Palosuo, T., Priesack, E., Ripoche, D., Rötten,
984 R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Fao, F., Wolf, J.,
985 Zhang, Z., 2018. Multi-model ensembles improve predictions of crop-environment-
986 management interactions. *Global Change Biology* 24, 5072-5083.

- 987 Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng,
988 S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C.,
989 Gayler, S., Ghahramani, A., Hiremath, S., Hoek, S., Horan, H., Hoogenboom, G., Huang,
990 M., Jabloun, M., Jansson, P.-E., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A.,
991 Launay, M., Lewan, E., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh,
992 H., Padovan, G., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze,
993 N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G.,
994 Wallor, E., Wang, J., Weber, T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L.,
995 Zhao, C., Zhu, Y., Seidel, S.J., 2021. The chaos in calibrating crop models: Lessons learned
996 from a multi-model calibration exercise. *Environmental Modelling & Software* 145: 105206.
- 997 Wang, C., Amon, B., Schulz, K., Mehdi, B., 2021. Factors that influence nitrous oxide
998 emissions from agricultural soils as well as their representation in simulation models: a
999 review. *Agronomy* 11: 770.
- 1000



Appendix A. Temporal changes of GPP ($\text{g C m}^{-2} \text{ season}^{-1}$ for crops and $\text{g C m}^{-2} \text{ yr}^{-1}$ for grasslands, (left), RECO ($\text{g C m}^{-2} \text{ season}^{-1}$ for crops and $\text{g C m}^{-2} \text{ yr}^{-1}$ for grasslands, middle) and Yield ($\text{kg DM m}^{-2} \text{ season}^{-1}$ for crops and $\text{kg DM m}^{-2} \text{ yr}^{-1}$ for grasslands, right) observations (Obs, red square) and simulations: S3 (stage 3, blue) and S5 (stage 5, pink) at all sites (site codes as in Fig. 1). Lines represent the multi-model median (MMM) of the S3 and S5 simulations, and shaded areas represent the simulation envelope (with the same colours as the lines). At cropland site C3, only modelled GPP and RECO data are reported.

- We investigate multi-model performance in simulating C and N fluxes in agriculture.
- Correlated model residuals hinder reliable C-N flux estimates.
- Residual correlation analysis is applied to ensemble crop and grassland models.
- Partially calibrated models can be practical for implementing model ensembles.
- Fully calibrated models are key to model development.

Journal Pre-proof

Authors declare no conflict of interest.

Journal Pre-proof