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 Brilli, 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-03997939

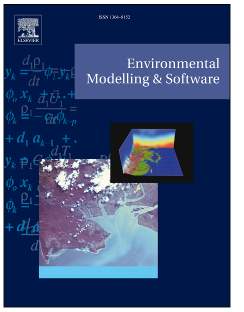
Submitted on 16 Mar 2023

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 Brilli, 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

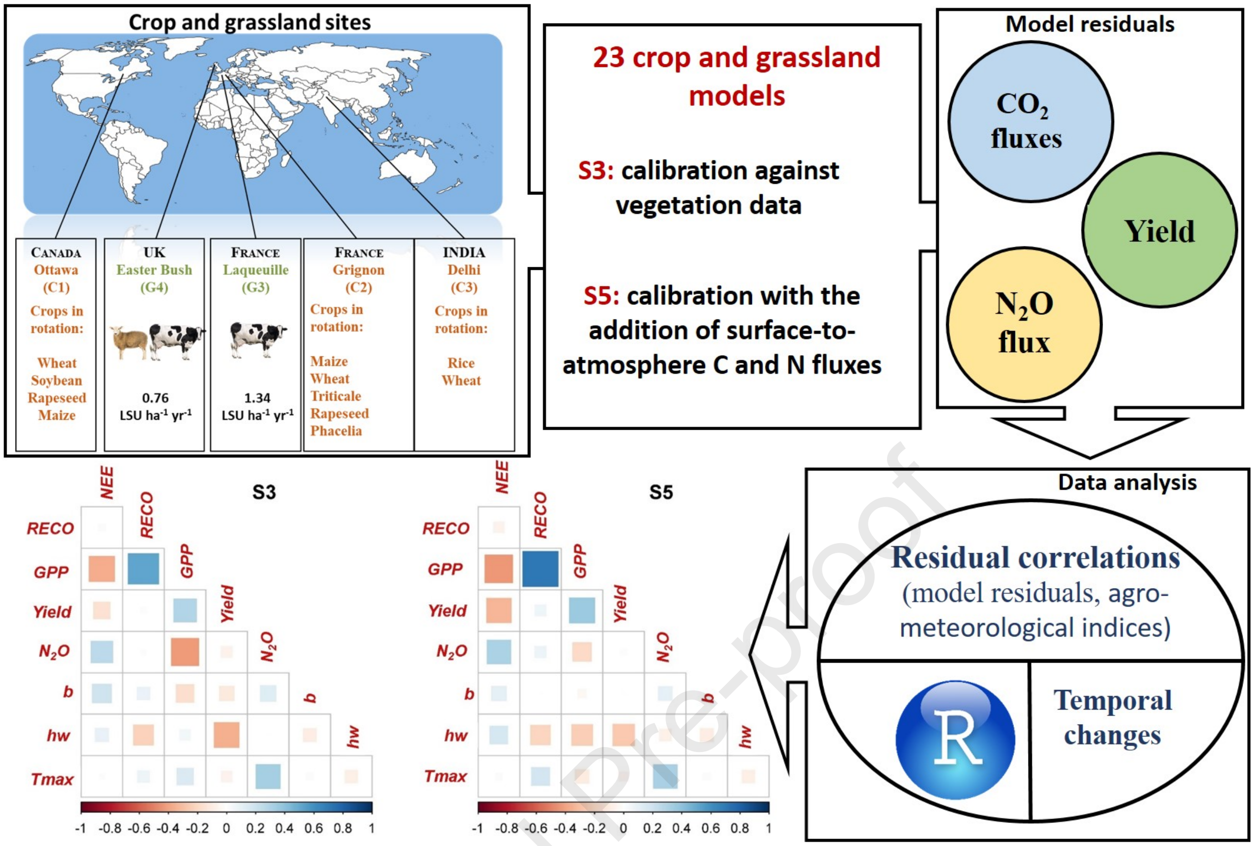
Revised Date: 30 October 2022

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., Brilli, 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.

© 2023 Published by Elsevier Ltd.



Journal Pre-proof

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 Basso[i], Arti Bhatia[j], Lorenzo Brilli[k,l], Jordi Doltra[m], Christopher D. Dorich[n], Luca Doro[o,p], Nuala Fitton[g], Brian Grant[q], Matthew Tom Harrison[r], 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]

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

[b]Agricultural Institute, ELKH CAR, 2462 Martonvásár, Hungary

[c]INRAE, CODIR, 75007 Paris, France

[d]RITTMO Agroenvironnement, Colmar, France

[e]Queensland University of Technology, Brisbane, Australia

[f]Université de Reims Champagne Ardenne, INRAE, FARE, 51097 Reims, France

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

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

[i]Department of Geological Sciences, Michigan State University, East Lansing MI, USA

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

[k]University of Florence, DAGRI, 50144 Florence, Italy

[l]IBE-CNR, 50145, Florence, Italy

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

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

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

[p]Texas A&M AgriLife Research, Blackland Research and Extension Center, Temple TX, USA

[q]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada

[r] University of Tasmania, Newnham Dr, Launceston, Tasmania, 7248, Australia [s]UK Centre for Ecology and Hydrology, Bush Estate, Penicuik, EH26 0QB, UK

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

[u]Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, 78850 Thiverval-Grignon, France

[v]INRAE, AgroImpact, 02000 Barenton-Bugny, France
[w]CSIRO, Agriculture Flagship, Black Mountain Laboratories, Canberra, Australia
[x]School of Geosciences, The University of Edinburgh, UK
[y]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany
[z]New Zealand Institute for Plant and Food Research, Christchurch, New Zealand
[aa]Sustainable Agriculture Systems, Rothamsted Research, North Wyke, Devon, UK
[ab]LAPC, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

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

Abstract

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

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

Journal Pre-proof

Journal Pre-proof

1. Introduction

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

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

Journal Pre-proof

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

With the aim of increasing reliability and confidence in the simulated results, this study explores patterns of simulated C-N and productivity responses with a multi-model ensemble approach. We included results from 23 crop and grassland models, used to simulate C-N and productivity outputs in five sites worldwide (three crop rotations with spring and winter cereals, soybean and rapeseed, and two temperate grasslands). This work builds on comprehensive foundations laid by Ehrhardt et al. (2018) for yield and nitrous oxide (N_2O) emissions, and Sándor et al. (2020) for C fluxes. Here, we analyse factors that may explain differences in simulated model responses. Viewing and interpreting a variety of modelled outputs is intended to lay ground for future model developments. We thus further explored the extent to which multi-model ensembles can be used to help identify deficiencies in model structures, which limit model performance in different situations. Specifically, we present an approach that uses a correlation matrix (with graphical representation) to correlate both the residuals of outputs from the ensemble against residuals of selected climate drivers. The estimation of uncertainty in simulation models is based on the assumption that model residuals (differences between model estimates and observations) are additive and independent. When the residuals of one model output are correlated with the residuals of other outputs, the different outputs would probably be the result of processes not included (or partially included) in the models. This suggests that interacting processes are sources of model-data mismatch and, in this case, non-negligible correlations between model residuals and external drivers might inspire a more detailed description of these same drivers to improve the models.

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

2. Materials and methods

2.1. Experimental sites and measurements

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

Table 1. Stages of model run (after Ehrhardt et al., 2018). The grey cells indicate the two stages (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 historical management and climate	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

For consistency, we have maintained the model and site identifiers specified by Ehrhardt et al. (2018). The variability of the multi-model simulation exercise across stages was documented by inspecting how the multi-model median (MMM) converged to the observations. Observational data were from two long-term (19 years in total), grazed experimental sites (G3, G4) and three cropland sites (C1, C2, C3), covering a variety of pedo-climatic conditions and agricultural practices from United Kingdom, France (two sites), Canada and India (Table 2). The selected cropping systems covered a range of climates, from continental (C1, Canada), oceanic (C2, France) and subtropical (C3, India). All cropland sites had rotations with at least one wheat crop (six growing seasons), while maize was present in C1 and C2 (three growing seasons), and rice was only grown in C3 (two growing seasons), for a total of 18 growing seasons (including fallow intercrops). The 23 models (Table A in the Supplementary material), and the model identifiers and outputs provided, encompass all but one of the 24 biogeochemical models described in Ehrhardt et al. (2018). Model M11 was not included in the analysis because it did not provide the C-flux related outputs. At cropland sites, we had: GPP from six models, NEE from seven models, RECO from 14 models, N_2O from 15 models, Yield from 15 models. At grassland sites, we had: GPP from 10 models, NEE from 10 models, RECO from 11 models, N_2O from nine models, Yield from nine models. The use of flux tower data allows the determination of NEE, which is partitioned into its (simulated) component fluxes - RECO and GPP – by flux partitioning methods. Separated from flux tower measurements of NEE, the estimated GPP provides information on the physiological processes that contribute to NEE, which is the balance between the C released by the RECO and the GPP (e.g. Raj et al., 2016). Climate data available at each site since 1980 were used to initialise the models (calibration stage S2).

Table 2. Cropland and grassland sites, and years of available data, for analysis on the following output variables from different models: GPP (g C m^{-2} yr^{-1}): gross primary production; RECO (g C m^{-2} yr^{-1}): ecosystem respiration; NEE (g C m^{-2} yr^{-1}): net ecosystem exchange of CO_2); N_2O (µg N_2O-N m^{-2} yr^{-1}): nitrous oxide emissions; Yield (kg DM m^{-2} yr^{-1}): annual grain yield for arable crops or annual above-ground net primary productivity for grasslands. Cropland sites used different crop rotations (Table B in the Supplementary material), including cereals (spring and winter wheat [W], triticale [T], maize [M] and rice [R]), legumes (soybean [S]), rapeseeds (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)

2.2. Agro-climatic metrics

Three metrics were selected to characterise the study-sites based on the extent to which they fulfil the need to report the response of models to water-limited and heat stressed conditions (Sándor et al., 2017, 2018; Farina et al., 2021). They are also important within a climate-change focus (Rivington et al., 2007, 2013; Matthews et al., 2008; Graux et al., 2013; Lardy et al., 2014, 2015; Eza et al., 2015). An increase in *Tmax* and frequency of *hw* is desirable if the two metrics are negatively correlated with model residuals. The aridity index (b) is defined in such a way (the higher it is, the lower the aridity) that, with a positive correlation, higher model residuals are expected in wetter conditions and, with a negative correlation, higher model residuals are expected in drier conditions. In fact, the De Martonne aridity index ($b \leq 100$) was 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 the total annual precipitation (mm), T_Y is the mean annual temperature (°C), p_a is the total precipitation of the driest month (mm), and t_a is the mean temperature of the driest month (°C). The possibility to discriminate between thermo-pluviometric conditions associated with aridity gradients is given by the range limits published by Diodato and Ceccarelli (2004): $b<5$: extreme 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; $b>59$: strong humidity. Adopting the definition of Confalonieri et al. (2010), after Barnett et al. (2006), for identifying the frequency of *hw* within a year in each site, we defined the heatwave event as the number of ≥7 consecutive days when *Tmax* was higher than the mean summer (northern hemisphere: June, July and August in the temperate sites; April, May and June in the monsoonal site) *Tmax* of all the available years (baseline) +3 °C. The range limits in this study were given by the minimum and the maximum numbers of the *hw* days of all sites: $hw \leq 14$: extremely moderate frequency; $14 < hw \leq 28$: very moderate frequency; $28 < hw \leq 42$: moderate frequency; $42 < hw \leq 56$: high frequency; $56 < hw \leq 70$: very high frequency; $hw > 70$: extremely high frequency. Fig. 1 displays the gradient of thermo-pluviometric conditions that are considered to analyse the response of the model residuals to climate drivers.

Journal Pre-proof

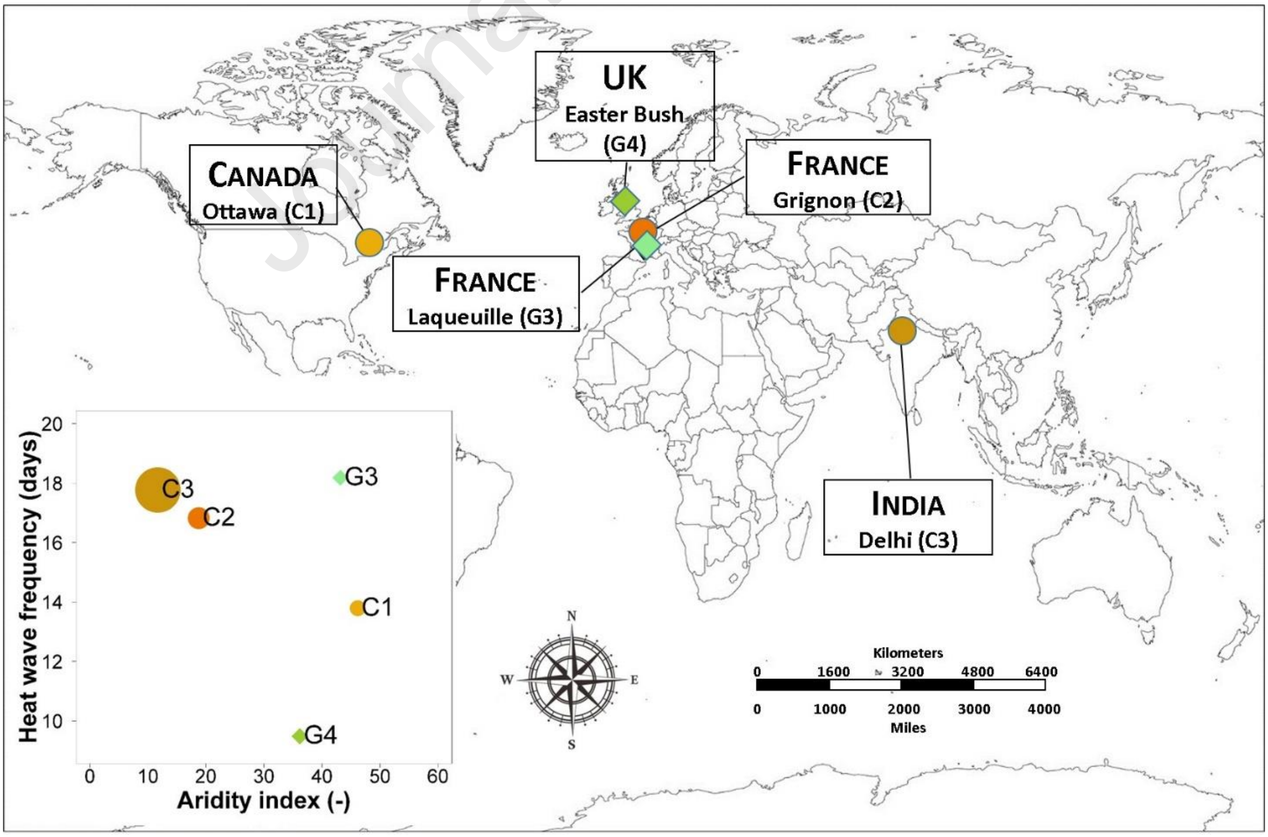


Fig. 1. Geographic location (diamonds: grassland sites; circles: cropland sites) and classification of study sites with respect to De Martonne-Gottmann aridity index and frequency

Journal Pre-proof

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

2.3. Residual scatterplot analyses

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

For both arable crops and grasslands, Ehrhardt et al. (2018) found that no model consistently outperformed the others in terms of both N_2O emissions and yield production. In particular, in the case of cereal crop yields, the MMM error decreased considerably from S1 (34%, 31% and 45% for wheat, maize and rice, respectively) to S3 (6.4%, 5.8% and 5.5% for wheat, maize and rice, respectively) and remained below 5% in S4 and S5. In the case of grassland yields, the MMM error decreased from 44% in S1 to 27% in S3 and finally increased to 46% in S5.

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

We thus quantified the correlations among standardised model residuals of GPP, RECO, NEE, N_2O and Yield (differences between ensemble MMM and mean of observations), based on the results from partially and fully calibrated simulations (stages S3 and S5). For both calibration stages, we also quantified the correlations between model residuals and three agro-climatic metrics (annual values) related to the occurrence of high temperature (mean maximum air temperature, *Tmax* and heatwave days, *hw*) and arid conditions (Figures A-E in the Supplementary material).

Arrays of pairwise scatterplots (scatterplot matrices) were generated with the panel plot option 'panel.smooth' (https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/panel.smooth.html) in the R language and environment for statistical computing (R Core Team, 2020). The function produces *x-y* scatterplots of each pair of variables below the diagonal (output residuals and agro-climatic metrics) and overlays a local non-parametric smoother curve (locally estimated scatterplot smoothing) on each plot to give some indication of trends without inferential characteristics (after Cleveland, 1979). For readability, the correlation between each variable and its significance (p value) is indicated in the lower triangular part of the matrices. The non-significant correlations ($p \geq 0.10$) are not discussed (e.g. Bellocchi et al., 2002). According to Sándor et al. (2017), we have selected an arbitrary (high enough) absolute minimum threshold, i.e. $r=|0.66|$, and identified the number of cases when the correlation coefficient equals or exceeds this minimum value. Correlations between external climate factors (mean maximum air temperature, aridity index and frequency of heatwave days) are reported but are not informative in the present context.

3. Results

3.1. Evaluation of output dynamics

In general, model results showed the largest spread with the S3 scenario, considering the C outputs such as NEE (Fig. 2), GPP and RECO (Appendix A), N_2O-N emissions (Fig. 2) and yield (Appendix A). In some years, the MMM of S3 and in some cases the S5 scenario also overestimated the amount of C respiration, e.g. at G4 site in 2002(S3: -0.05; S5: -0.04; 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}), while the N_2O-N emission was underestimated at this site. The MMM lines for all outputs were remarkably close to the observations at all sites, despite the wider range of S3 individual simulations (blue shaded area in Fig. 2 and Appendix A). The largest difference between the spread of S3 and S5 was found for the N_2O-N emissions.

Journal Pre-proof

Journal Pre-proof

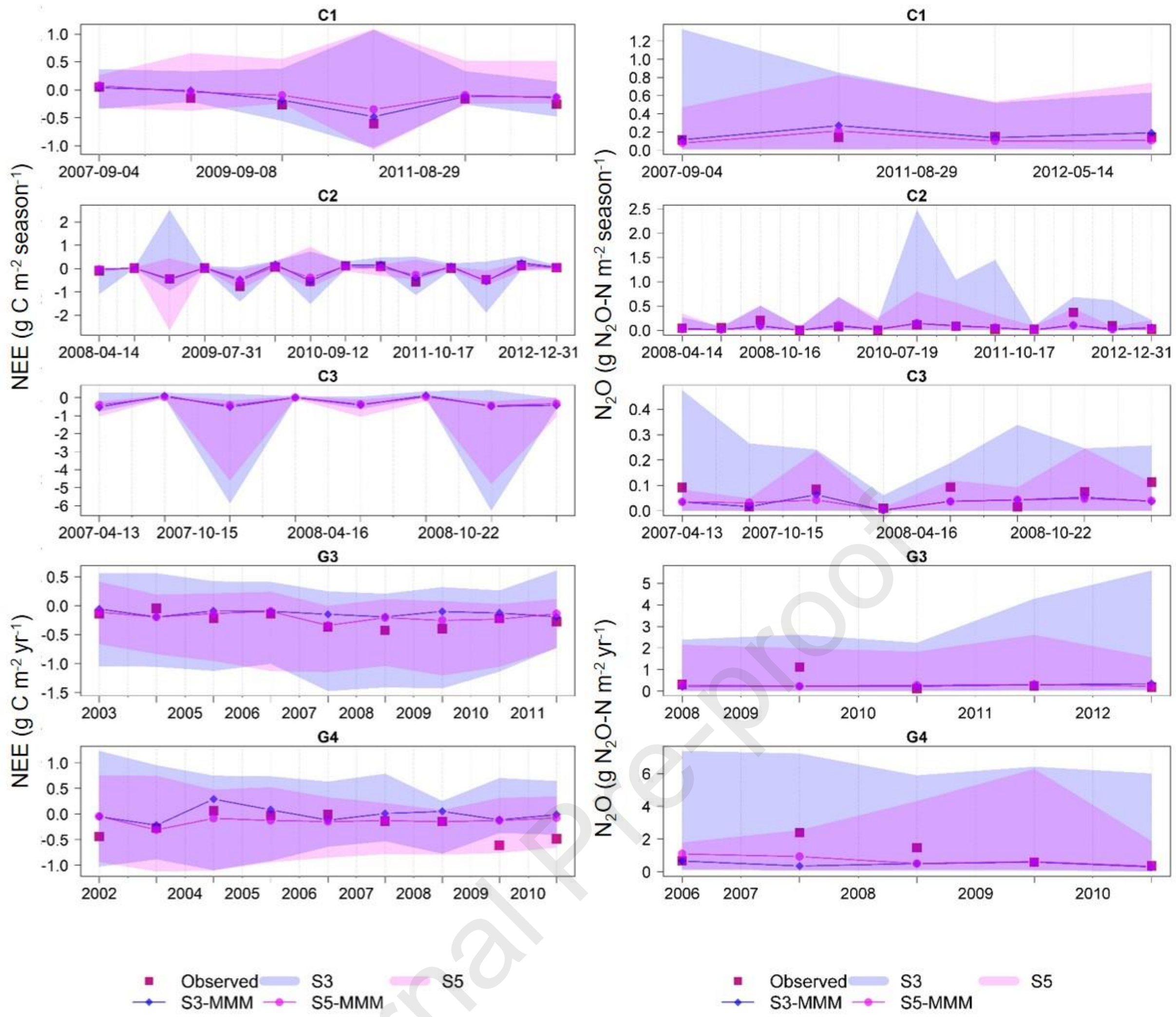


Fig. 2. Temporal changes of NEE (g C m^{-2} season^{-1} for crops and g C m^{-2} yr^{-1} for grasslands, left) and N_2O (g N_2O-N m^{-2} season^{-1} for crops and g N_2O-N m^{-2} yr^{-1} for grasslands, right) observations (Obs, red square) and simulations: S3 (stage 3, blue) and S5 (stage 5, pink) at 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 envelopes given by the edges of the most extreme model predictions (with the same colours as the lines). At cropland site C3, only modelled RECO data are reported.

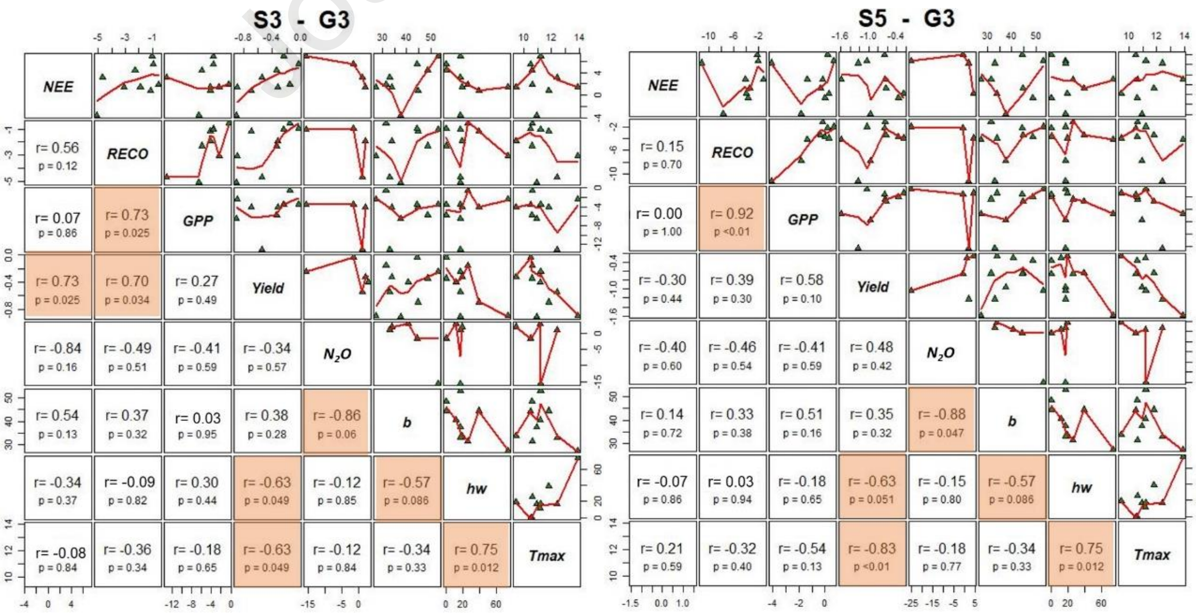
3.2. Residual analysis in grassland sites

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

Journal Pre-proof

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

Considering the climatic factors at the G3 site, aridity values (higher aridity index indicates wetter conditions) show a negative correlation with N_2O residuals ($r=-0.86$, $p=0.06$ and $r=-0.88$, $p=0.05$ at stages S3 and S5, respectively), with higher model residuals expected in drier conditions in the estimation of N_2O emissions. When *Tmax* is considered for both S3 and S5, the correlation with Yield residuals is significantly negative ($r=-0.63$, $p\sim0.05$ and $r=-0.83$, $p<0.01$, respectively). With S5, the days of heatwave are negatively correlated with Yield residuals ($r=-0.63$, $p=0.05$), with model outputs becoming less reliable at lower temperatures. This indicates that state-of-the-art models take into account the influence of climate factors, as periods of extreme heat and drought, or extremely wet conditions, tend to decrease or increase model errors. For instance, simulated N_2O emissions may show higher magnitude residuals under drier conditions, while yield and C-flux simulations may have lower magnitude residuals (e.g. models are more sensitive to wet G3 upland conditions).



Journal Pre-proof

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

Analysis of residual scatterplots at G4 (Easter Bush, United Kingdom) shows some similarities at both calibration stages (Fig. 4). The negative correlation between NEE and GPP residuals at S3 ($r=-0.29$, $p<0.01$) indicates that overestimation of NEE may be the result of underestimation of GPP. This is reflected in the negative correlation between NEE and Yield ($r=-0.42$ at S3, $p<0.01$). RECO and GPP residuals are significantly ($p<0.01$) positively correlated ($r=0.86$ at S3 and $r=0.57$ at S5). In addition, GPP and Yield residuals are positively correlated ($r=0.48$, $p<0.01$ and $r=0.25$, $p=0.02$ at S3 and S5, respectively). Overall, these correlations between C-fluxes and yield residuals are less important or less significant for the fully calibrated models (S5). However, N_2O residuals show significant correlations ($p<0.01$) with NEE residuals at both calibration stages ($r=0.37$ and $r=0.50$ at S3 and S5, respectively), while no significant correlations ($p>0.10$) were found with other C-flux residuals. Considering climatic factors, heatwaves do not have a significant impact on C-flux and Yield residuals in G4 (which is not exposed to severe heatwaves; Fig. 1). Interestingly, N_2O-emission residuals are significantly ($p<0.01$) positively correlated with heatwaves at both S3 ($r=0.36$) and S5 ($r=0.28$). Thus, increasingly long heatwaves may lead to greater model inaccuracy in simulating N_2O emissions, likely due to poor estimates of soil water content at higher temperatures or model limitations in appropriately reducing emission estimates at low soil water contents (Wang et al., 2021). The aridity index was negatively correlated ($p<0.05$) with NEE residuals for both S3 ($r=-0.22$) and S5 ($r=-0.24$), and was not correlated with N_2O, GPP, RECO and Yield residuals. These negative correlations indicate that simulations are generally more reliable under G4 humid conditions. Since *Tmax* is significantly negatively correlated with NEE at S3 ($r=-0.25$, $p<0.01$) and S5 ($r=-0.21$, $p<0.05$), the models are expected to give poorer C-flux simulations under colder conditions and better results at higher temperatures.

Journal Pre-proof

Journal Pre-proof

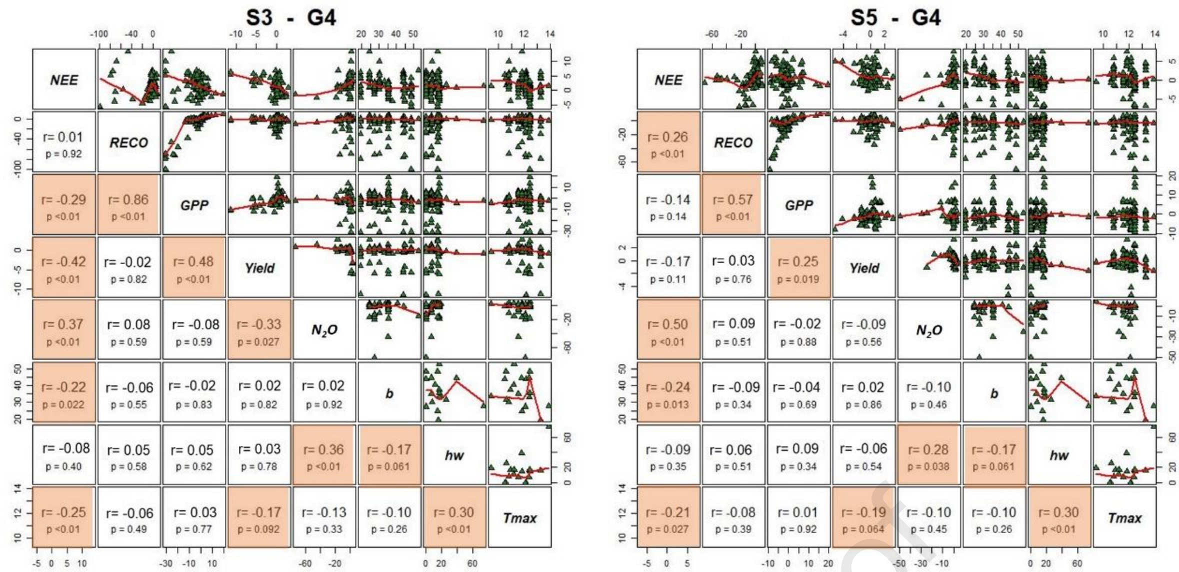


Fig. 4. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-model medians (MMM) for stages 3 (left) and 5 (right) at G4 grassland site, and the annual agro-climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature (*Tmax*). Overlaid (red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

3.3. Residual analysis in cropland sites

The results of the residual analysis differ among cropland sites, with the strongest differences occurring at the most humid study-site (Fig. 1), i.e. C1 (Ottawa, Canada), with seven significant correlations at S3 (Fig. 5, left), which reduce to four at S5 (Fig. 5, right). As with G4, the negative correlation between NEE and GPP residuals at S3 (r=-0.87, p<0.02) may indicate that an overestimation of NEE is likely to be the result of an underestimation of GPP, but this is not reflected in any other correlation between the model residuals (p>0.10). However, at C1, all model residuals in S3 are significantly correlated with either the aridity index (NEE, r=0.85, p=0.03; RECO, r=-0.92, p<0.01; N_2O, r=0.96, p=0.04), heatwaves (Yield, r=-0.82, p=0.05) or both (GPP: aridity, r=-0.75, p=0.08; heatwaves, r=-0.79, p=0.06). These correlations are less important with fully calibrated models. While the residuals of NEE and GPP at C1 are still negatively correlated in S5 (r=-0.99, p<0.01), among the environmental factors, it is essentially the aridity index that is positively (NEE, r=0.75, p=0.09) or negatively (GPP, r=-0.81, p=0.05; RECO, r=-0.88, p=0.02) correlated with C fluxes also after the full model calibration. The residuals of C and N fluxes are significantly correlated with aridity. GPP and Yield residuals are also negatively correlated with heatwaves.

Journal Pre-proof

Journal Pre-proof

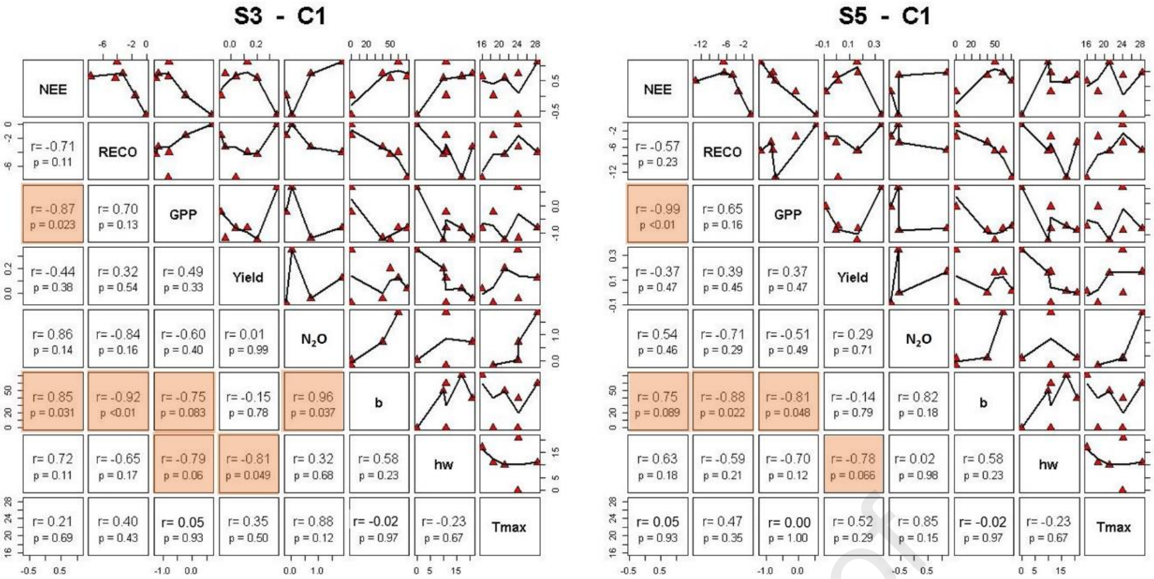


Fig. 5. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-model medians (MMM) for stages 3 (left) and 5 (right) at C1 cropland site, and the annual agro-climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature (*Tmax*). Overlaid (red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

At C2 (Grignon, France), there was some significant positive correlations, e.g. between NEE and N_2O residuals at S3 (Fig. 6; r=0.68, p=0.07) and between RECO and GPP at S5 (r=0.70, p=0.05). However, some significant correlations between GPP residuals and climatic factors (heatwaves: r=0.68, p=0.07; *Tmax*: r=-0.71, p=0.05) observed at S3 were no longer significant at S5 (p>0.10).

Journal Pre-proof

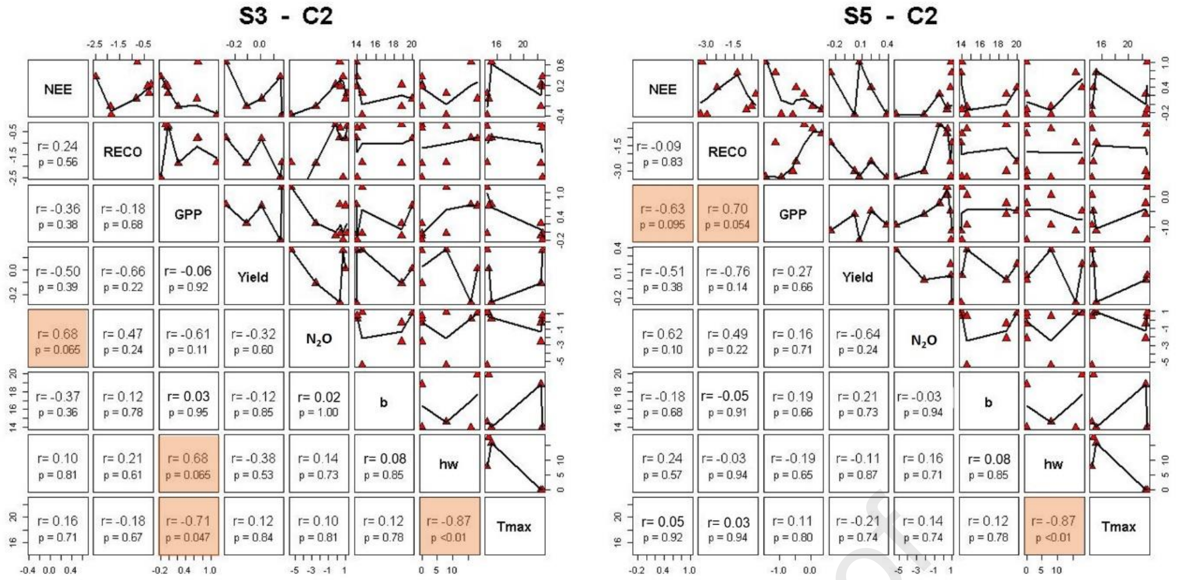


Fig. 6. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-model medians (MMM) for stages 3 (left) and 5 (right) at C2 cropland site, and the annual agro-climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature (*Tmax*). Overlaid (red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

At the Indian site of Delhi (C3), where NEE and GPP data are not available, it is relevant to note the significant positive correlation observed between RECO and N_2O residuals at S5 ($r=0.95$, $p=0.02$), not observed at S3 (Fig. 7). Then, there is a dependence of the simulation quality for these two fluxes on aridity (RECO: $r=0.92$, $p=0.03$) or *Tmax* (N_2O: $r=0.93$, $p=0.02$) at S3, or on *Tmax* only at S5 (RECO: $r=0.84$, $p=0.08$; N_2O: $r=0.89$, $p=0.04$).

Journal Pre-proof

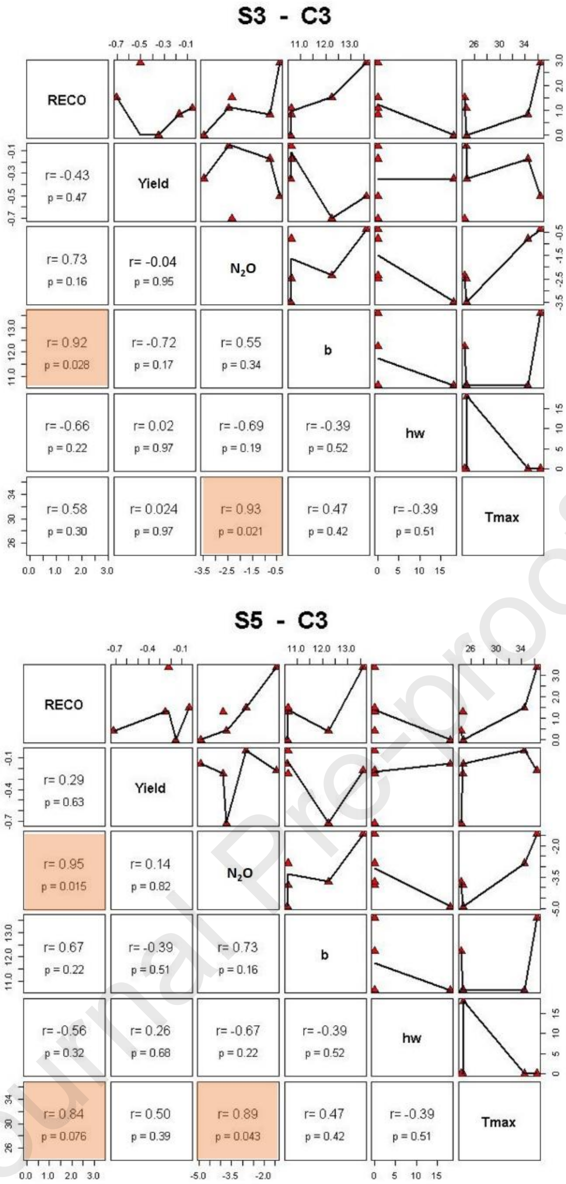


Fig. 7. Scatterplot correlation matrix of NEE, RECO, GPP and yield model residuals of multi-model medians (MMM) for stages 3 (left) and 5 (right) at C3 cropland site, and the annual agro-climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature (*Tmax*). Overlaid (red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

3.4. Geographical location, land use characteristics and calibration stages

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

Journal Pre-proof

Journal Pre-proof

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

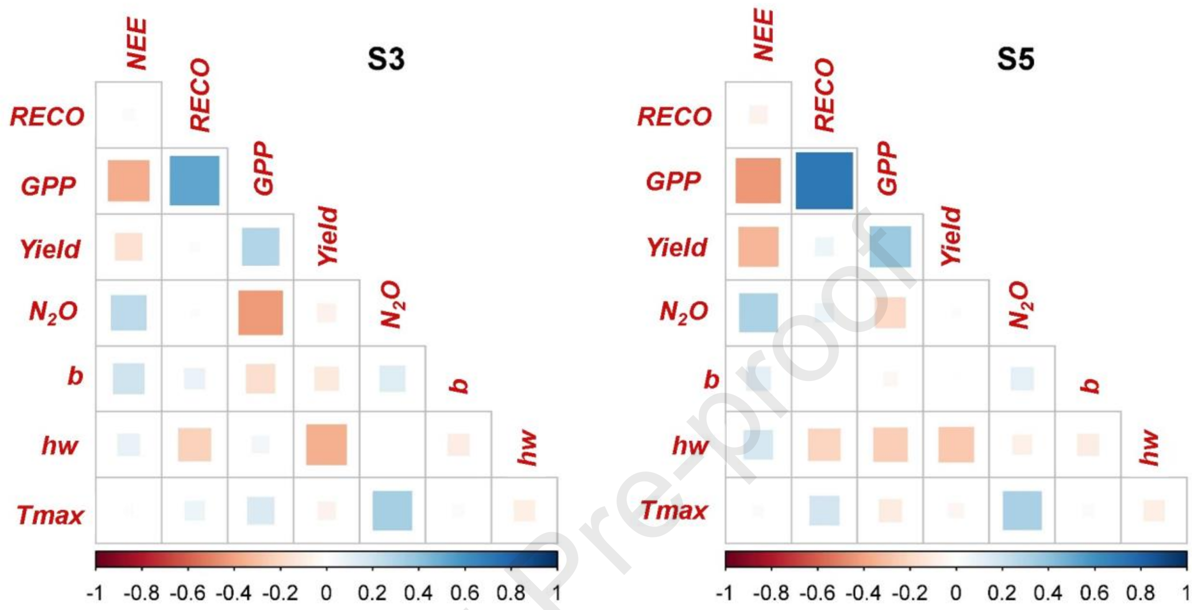


Fig. 8. Heatmap of mean correlation coefficients (r) between NEE, RECO, GPP and yield model residuals of multi-model medians (MMM) for stages 3 (left) and 5 (right) across sites/land uses, and the annual agro-climatic metrics aridity index (*b*), frequency of heatwaves (*hw*) and maximum temperature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

However, the multi-model simulations show complex patterns, illustrated by the analysis of land uses (grasslands, arable crops), study-sites (C1, C2, C3, G3 and G4) and calibration stages (S3 and S5) investigated, which show considerable differences in terms of correlation between model residuals, and between these residuals and weather metrics. Positive correlations were established between the RECO and GPP residuals at G3 (Fig. 3) and G4 (Fig. 4) in both calibration stages, and at C2 (Fig. 6) with fully calibrated models (along with a positive correlation between NEE and GPP residuals). At G4, positive correlations also characterise the relationships between GPP and Yield residuals (both calibration stages) and between RECO and NEE residuals (at S5). In addition, negative correlations were found at this site between NEE and GPP residuals (at S3), NEE and Yield residuals (at S3) and GPP and Yield residuals (at both calibration stages). At cropland site C1 (Fig. 5), NEE and GPP residuals are also negatively correlated (at both calibration stages). Overall, these results indicate that errors are likely to be propagated through C-flux (and yield) predictions, and full calibration with plant, soil and surface-to-atmosphere C-N fluxes does not always limit them. On the contrary, full calibration can also increase the propagation of errors through C fluxes, as obtained in G4 with RECO and NEE residuals (from r~0 at S3 to highly significant r=0.26 at S5). However, while many correlations between residuals are significant in G4, only the correlation between RECO and GPP residuals at S3 (r=0.86) is high in this site.

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

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

4. Discussion

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

4.1. Residual analysis and model quality

Journal Pre-proof

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

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

Journal Pre-proof

4.2. Effects of agro-climatic factors

While models estimating crop or pasture yields may not explicitly account for the impact of heatwaves on grain or biomass formation (e.g. Harrison et al., 2017; Mangani et al., 2019), the opposite impact of arid conditions on NEE (negative correlation) or RECO and GPP (positive correlations) residuals is somewhat unexpected, considering that one variable (NEE) is the difference of the two others. Considering that drought may be more effective in reducing CO_2 uptake by the plant than reducing ecosystem respiration (Gibelin et al., 2008; Nakano and Shinoda, 2015), better results are provided when simulating NEE with a multi-model ensemble (at C1 as at other sites, Fig. 2). This implies that there may be error compensation in the ensemble. Greater coverage of plant and soil processes is also likely when more models are used to simulate NEE than its basic components.

As far as N fluxes are concerned, N uptake by plants is computed by the models through a supply/demand scheme, with soil supply depending mainly on soil nitrate and ammonium concentrations and root length density (Lehuger et al., 2010). However, N_2O emissions are mostly controlled by soil properties and local climate conditions, and current soil N levels, and only to a lesser extent by the doses and types of N fertiliser applied (Butterbach-Bahl et al., 2013). For instance, increasing bulk density decreases soil porosity and thereby increases the likelihood of moisture conditions favourable to denitrification and N_2O emissions (Gabrielle et al., 2006). As well, the correlation between N_2O and NEE residuals may be due to soil processes because if heterotrophic respiration is too high there may be too many substrates (C and N) available for nitrate respiration and denitrification (e.g. Rajta et al., 2020). The high negative correlations ($r=-0.86$, $p=0.06$ and $r=-0.88$, $p=0.05$ at S3 and S5, respectively) between N_2O residuals and aridity index at grassland site G3 reflect the deficit of moisture occurring mostly in summer in central France (e.g. Klumpp et al., 2011), while in the wet climate of the United Kingdom (grassland site G4) most nitrate available for leaching may result in reduced N_2O emissions (e.g. Cardenas et al., 2013). In fact, grazed G4 grassland tends to have high N leaching rates (and corresponding limited N_2O emissions) due to added urinary N to the system and the non-uniform distribution of excreted organic N, which further enhances leaching due to N hotspot formation (Jones et al., 2017b). N_2O emissions are reported to increase with increasing temperature, which is attributed to an increase in the anaerobic volume fraction, caused by an increased respiratory oxygen sink (Smith et al., 2018). With a mean annual maximum annual temperature equal to 31.5 °C, N_2O residuals at the hot Indian cropland site

C3 are still positively correlated with *Tmax* with fully calibrated models (r=0.93, p=0.02 at S3; r=0.89, p=0.04 at S5).

5. Conclusions

Residuals from model-ensemble outputs tend to be less correlated when crop and grassland models are calibrated using soil and C-N fluxes together with vegetation data (compared to partial calibration with vegetation data alone). If full calibration can reduce the correlation between C- and N-flux residuals (e.g. between GPP and N_2O residuals), intense weather factors can also have significant effects on model residuals (e.g. N_2O residuals positively correlated with maximum air temperature at the hot Indian cropland site). However, complex multi-model simulation patterns indicate that full calibration does not always constrain the correlation between model residuals, and between these residuals and agro-climatic metrics. Our assessment, which remains limited to climate-related drivers calculated annually (and could then a future improvement be a seasonal climate analysis), holds potential for a wider analysis that surveys contextual soil and management factors, for which the current database was not designed. In that, we have proposed a somewhat *ad hoc* multi-output analysis that considers inter-dependencies in the model outputs, but there are challenges that require further work. These include how to quantitatively account for consistency with mechanistic viewpoints supported by alternative models of varying complexity as a further important requirement for model ensembles, as well as definitions of core concepts and metrics to provide a quantitative determination of the stability of simulation results under a variety of conditions. These challenges are interesting from a practical point of view because improving our understanding of these issues and finding better ways to deal with the plurality of models has the potential to increase the value of biogeochemical models in agriculture, where determining the robustness of results is a strategy to assess confidence in results. In the end, this may provide modellers with a clearer explanation of what they are doing in ensemble modelling (as well as how they are doing it), and stronger arguments as to when ensemble modelling can, or cannot, become a practical epistemic resource.

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

Journal Pre-proof

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

Acknowledgements

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

Journal Pre-proof

Journal Pre-proof

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

References

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

Allard, V., Soussana, J.-F., Falcimagne, R., Berbigier, P., Bonnefond, J.M., Ceschia, E., D'hour, P., Hénault, C., Laville, P., Martin, C., Pinarès-Patino, C., 2007. The role of grazing management for the net biome productivity and greenhouse gas budget (CO_2, N_2O and CH_4) of semi-natural grassland. Agriculture, Ecosystem & Environment 12, 47-58.

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

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

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

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

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

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

Journal Pre-proof

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

Butterbach-Bahl, K., Baggs, E.M., Dannenmann, M., Kiese, R., Zechmeister-Boltenstern, S., 2013. Nitrous oxide emissions from soils: how well do we understand the processes and their controls? Phylosophical Transactions of the Royal Society B 368:20130122.

Calanca, P., Deléglise, C., Martin, R., Carrère, P., Mosimann, E., 2016. Testing the ability of a simple grassland model to simulate the seasonal effects of drought on herbage growth. Field Crops Research 187, 12-23.

Calder, M., Craig, C., Culley, D., de Cani, R., Donnelly, C.A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., Hinds, D., Lane, D.C., Mitchell, D., Pavey, G., Robertson, D., Rosewell, B., Sherwin, S., Walport, M., Wilson, A., 2018. Computational modelling for decision-making: where, why, what, who and how. Royal Society Open Science 5:172096.

Cardenas, L.M., Gooday, R., Brown, L., Scholefield, D., Cuttle, S., Gilhespy, S., Matthews, R., Misselbrook, T., Wang, J., Li, C., Hughes, G., Lord, E., 2013. Towards an improved inventory of N_2O from agriculture: Model evaluation of N_2O emission factors and N fraction leached from different sources in UK agriculture. Atmospheric Environment 79, 340–348.

Cavalli, D., Bellocchi, G., Corti, M., Gallina, P.M., Bechini, L., 2019. Sensitivity analysis of C and N modules in biogeochemical crop and grassland models following manure addition to soil. European Journal of Soil Science 70, 833-846.

Challinor, A.J., Smith, M.S., Thornton, P., 2013. Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. Agricultural and Forest Meteorology 170, 2-7.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829-836.

Confalonieri, R., Bellocchi, G., Donatelli, M., 2010. A software component to compute agro-meteorological indicators. Environmental Modelling & Software 25, 1485-1486.

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

Dijkstra, T.K., 1988. On model uncertainty and its statistical implications. Springer Verlag, Berlin, Germany.

Diodato, N., Ceccarelli, M., 2004. Multivariate indicator Kriging approach using a GIS to classify soil degradation for Mediterranean agricultural lands. Ecological Indicators 4, 177-187.

Doro, L., Wang, X., Ammann, C., De Antoni Migliorati, M., Grünwald, T., Klumpp, K., Loubet, B., Pattey, E., Wohlfahrt, G., Williams, J.R., Norfleet, M.L., 2021. Improving the simulation of soil temperature within the EPIC model. Environmental Modelling & Software 144:105140.

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

Eza, U., Shtiliyanova, A., Borras, D., Bellocchi, G., Carrère, P., Martin, R., 2015. An open platform to assess vulnerabilities to climate change: An application to agricultural systems. Ecological Informatics 30, 389-396.

Farina, R., Sándor, R., Abdalla, M., Álvaro-Fuentes, J., Bechini, L., Bolinder, M.A., Brilli, L., Chenu, C., Clivot, H., De Antoni Migliorati, M., Di Bene, C., Dorich, C.D., Ehrhardt, F., Ferchaud, F., Fitton, N., Francaviglia, R., Franko, U., Giltrap, D.L., Grant, B.B., Guenet, B., Harrison, M.T., Kirschbaum, M.U.F., Kuka, K., Kulmala, L., Liski, J., McGrath, M.J., Meier, E., Menichetti, L., Moyano, F., Nendel, C., Recous, S., Reibold, N., Shepherd, A., Smith, W.N., Smith, P., Soussana, J.F., Stella, T., Taghizadeh-Toosi, A., Tsutskikh, E., Bellocchi, G., 2021. Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils. Global Change Biology 27, 904-928.

Gabrielle, B., Laville, P., Duval, O., Nicoullaud, B., Germon, J. C., H´enault, C., 2006. Process-based modeling of nitrous oxide emissions from wheat-cropped soils at the subregional scale. Global Biogeochemical Cycles 20: GB4018.

Gibelin, A.-L., Calvet, J.-C., Viovy, N., 2008. Modelling energy and CO_2 fluxes with an interactive vegetation land surface model - Evaluation at high and middle latitudes. Agricultural and Forest Meteorology 148, 1611-1628.

Falconnier, G.N., Corbeels, M., Boote, K.J., Affholder, F., Adam, M., MacCarthy, D.S., Ruane, A.C., Nendel, C., Whitbread, A.M., Justes, É., Ahuja, L.R., Akinseye, F.M., Alou, I.N., Amouzou, K.A., Anapalli, S.S., Baron, C., Basso, B., Baudron, F., Bertuzzi, P., Challinor, A.J., Chen, Y., Deryng, D., Elsayed, M.L., Faye, B., Gaiser, T., Galdos, M., Gayler, S., Gerardeaux, E., Giner, M., Grant, B., Hoogenboom, G., Ibrahim, E.S., Kamali, B., Kersebaum, K.C., Kim, S.-H., van der Laan, M., Leroux, L., Lizaso, J.I., Maestrini, B., Meier, E.A., Mequanint, F., Ndoli, A., Porter, C.H., Priesack, E., Ripoche, D., Sida, T.S., Singh, U., Smith, W.N., Srivastava, A., Sinha, S., Tao, F., Thorburn, P.J., Timlin, D., Traore, B., Twine, T., Webber, H., 2020. Modelling climate change impacts on maize yields under low nitrogen input conditions in sub-Saharan Africa. Global Change Biology 26, 5942-5964.

Graux, A.-I., Bellocchi, G., Lardy, R., Soussana, J.-F., 2013. Ensemble modelling of climate change risks and opportunities for managed grasslands in France. Agricultural and Forest Meteorology 170, 114-131.

Harrison, M.T., Cullen, B.R., Armstrong, D., 2017. Management options for dairy farms under climate change: Effects of intensification, adaptation and simplification on pastures, milk production and profitability. Agricultural Systems 155, 19-32.

Harrison, M.T., Cullen, B.R., Tomkins, N.W., McSweeney, C., Cohn, P., Eckard, R.J., 2016. The concordance between greenhouse gas emissions, livestock production and profitability of extensive beef farming systems. Animal Production Science 56, 370-384.

Harrison, M.T., Evans, J.R., Moore, A.D., 2012. Using a mathematical framework to examine physiological changes in winter wheat after livestock grazing: 1. Model derivation and coefficient calibration. Field Crops Research 136, 116-126.

Harrison, M.T., Roggero, P.P., Zavattaro, L., 2019. Simple, efficient and robust techniques for automatic multi-objective function parameterisation: Case studies of local and global optimisation using APSIM. Environmental Modelling & Software 117, 109-133.

Hidy, D., Barcza, Z., Marjanovič, H., Ostrogovič Sever, M.Z., Dobor, L., Gelybó, Gy., Fodor, N., Pintér, K., Churkina, G., Running, S.W. Thornton, P.E., Bellocchi, G., Haszpra, L., Horváth, F., Suyker, A., Nagy, Z., 2016. Terrestrial ecosystem process model Biome-BGCMuSo: summary of improvements and new modeling possibilities. Geoscientific Model Development 9, 4405-4437.

Jégo, G., Pattey, E., Liu, J., 2012. Using leaf area index, retrieved from optical imagery, in the STICS crop model for predicting yield and biomass of field crops. Field Crops Research 131, 63-74.

Jones, J.W., Antle, J.M., Basso, B.O., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Muñoz-Carpena, R., Porter, C.H.,

Journal Pre-proof

Rosenzweig, C., Wheeler, T.R., 2017a. Brief history of agricultural systems modelling. Agricultural Systems 155, 240-254.

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

Klumpp, K., Tallec, T., Guix, N., Soussana, J.-F., 2011. Long-term impacts of agricultural practices and climatic variability on carbon storage in a permanent pasture. Global Change Biology 17, 3534-3545.

Knutti, R., Baumberger, C., Hirsch Hadorn, G., 2019. Uncertainty quantification using multiple models - prospects and challenges. In: Beisbart C., Saam N.J. (eds.) Computer simulation validation: fundamental concepts, methodological frameworks, and philosophical perspectives. Springer: Cham, pp. 835–855.

Kollas, C., Kersebaum, K.C., Nendel, C., Manevski, K., Müller, C., Palosuo, T., Armas-Herrera, C.M., Beaudoin, N., Bindi, M., Charfeddine, M., Conradt, T., Constantin, J., Eitzinger, J., Ewert, F., Ferrise, R., Gaiser, T., Garcia de Cortazar-Atauri, I., Giglio, L., Hlavinka, P., Hoffmann, H., Hoffmann, M.P., Launay, M., Manderscheid, R., Mary, B., Mirschel, W., Moriondo, M., Olesen, J.E. Öztürk, I., Pacholski, A., Ripoche-Wachter, D., Roggero, P.P., Roncossek, S., Rötter, R.P., Ruget, F., Sharif, B., Trnkam, M., Ventrella, D., Waha, K., Wegehenkel, M., Weigel, H.-J., Wu, L., 2015. Crop rotation modelling - A European model intercomparison. European Journal of Agronomy 70, 98–111.

Lardy, R., Bachelet, B., Bellocchi, G., Hill, D.R.C., 2014. Towards vulnerability minimization of grassland soil organic matter using metamodels. Environmental Modelling & Software 52, 38-50.

Lardy, R., Bellocchi, G., Martin, R., 2015. Vuln-Indices: Software to assess vulnerability to climate change. Computers and Electronics in Agriculture 114, 53-57.

Laville, P., Lehuger, S., Loubet, B., Chaumartin, F., Cellier, P., 2011. Effect of management, climate and soil conditions on N_2O and NO emissions from an arable crop rotation using high temporal resolution measurements. Agricultural and Forest Meteorology 151, 228-240.

Lehuger, S., Gabrielle, B., Cellier, P., Loubet, B., Roche, R., Béziat, P., Ceschia, E., Wattenbach, M., 2010. Predicting the net carbon exchanges of crop rotations in Europe with an agro-ecosystem model. Agriculture, Ecosystems & Environment 139, 384-395.

Journal Pre-proof

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

Liebermann, R., Breuer, L., Houska, T., Kraus, D., Moser, G., Kraft, P., 2020. Simulating long-term development of greenhouse gas emissions, plant biomass, and soil moisture of a temperate grassland ecosystem under elevated atmospheric CO_2. Agronomy 10:50.

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

Mangani, R., Tesfamariam, E.H., Engelbrecht, C.J., Bellocchi, G., Hassen, A., Mangani, T., 2019. Potential impacts of extreme weather events in main maize (*Zea mays* L.) producing areas of South Africa under rainfed conditions. Regional Environmental Change 19, 1441–1452.

Martin, G., Allain, S., Bergez, J.-E., Burger-Leenhardt, D., Constantin, J., Duru, M., Hazard, L., Lacombe, C., Magda, D., Magne, M.-A., Ryschawy, J., Thénard, V., Tribouillois, H., Willaume, M., 2018. How to address the sustainability transition of farming systems? A conceptual framework to organize research. Sustainability 10:2083.

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

Matthews, K.B., Rivington, M., Buchan, K., Miller, D.G., Bellocchi, G., 2008. Characterising the agro-meteorological implications of climate change scenarios for land management stakeholders. Climate Research 37, 59-75.

Journal Pre-proof

Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validationof models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiology 25, 839–857.

Nakano, T., Shinoda, M., 2015. Modeling gross primary production and ecosystem respiration in a semiarid grassland of Mongolia. Soil Science and Plant Nutrition 61, 106-115.

Pattey, E., Edwards, G., Strachan, I.B., Desjardins, R.L., Kaharabata, S., Wagner, C., 2006. Towards standards for measuring greenhouse gas fluxes from agricultural fields using instrumented towers. Canadian Journal of Soil Science 86, 373-400.

R Core Team, 2020. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Raj, R., Hamm, N.A.S., van de Tol, C., Stein, A., 2006. Uncertainty analysis of gross primary production partitioned from net ecosystem exchange measurements. Biogeosciences 13, 1409-1422.

Rajta, A., Bhatia, R., Setia, H., Pathania, P., 2020. Role of heterotrophic aerobic denitrifying bacteria in nitrate removal from wastewater. Journal of Applied Microbiology 128, 1261-1278.

Riccio, G., Giunta, G., Galmarini, S., 2007. Seeking for the rational basis of the Median Model: the optimal combination of multi-model ensemble results. Atmospheric Chemistry and Physics 7, 6085-6098.

Riggers, C., Poeplau, C., Don, A., Bamminger, C., Höper, H., Dechow, R., 2019. Multi-model ensemble improved the prediction of trends in soil organic carbon stocks in German croplands. Geoderma 345, 17-30.

Rivington, M., Matthews, K.B., Bellocchi, G., Buchan, K., Stöckle, C.O., Donatelli, M., 2007. An integrated assessment approach to conduct analyses of climate change impacts on whole-farm systems. Environmental Modelling & Software 22, 202-210.

Rivington, M., Matthews, K.B., Buchan, K., Miller, D.G., Bellocchi, G., Russell, G., 2013. Climate change impacts and adaptation scope for agriculture indicated by agro-meteorological metrics. Agricultural Systems 114, 15-31.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid, E., Stehfest, E., Yang, H., Jones, J.W., 2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proc. Natl. Acad. Sci. USA 111, 3268-3273.

Journal Pre-proof

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

Ruane, A.C., Rosenzweig, C., Asseng, S., Boote, K.J., Elliott, J., Ewert, F., Jones, J.W., Martre, P., McDermid, S.P., Müller, C., Snyder, A., Thorburn, P.J., 2017. An AgMIP framework for improved agricultural representation in integrated assessment models. Environmental Research Letters 12: 125003.

Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E., Ma, S., Perego, A., Rolinski, S., Ruget, F., Sanna, M., Seddaiu, G., Wu, L., Bellocchi, G., 2017. Multi-model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean grasslands: Uncertainties and ensemble performance. European Journal of Agronomy 88, 22-40.

Sándor, R., Barcza, Z., Hidy, D., Lellei-Kovács, E., Ma, S., Bellocchi, G., 2016. Modelling of grassland fluxes in Europe: evaluation of two biogeochemical models. Agriculture, Ecosystem & Environment 215, 1-19.

Sándor, R., Ehrhardt, F., Brilli, L., Carozzi, M., Recous, S., Smith, P., Snow, V., Soussana, J.F., Dorich, C.D., Fuchs, K., Fitton, N., Gongadze, K., Klumpp, K., Liebig, M., Martin, R., Merbold, L., Newton, P.C.D., Rees, R.M., Rolinski, S., Bellocchi, G., 2018a. The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed grasslands. Science of the Total Environment 15, 292-306.

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.S., Moore, A., Myrgiotis, V., Pattey, E., Rolinski, R., Sharp, J., Skiba, U., Smith, W., Wu, L., Zhang, Q., Bellocchi, G., 2020. Ensemble modelling of carbon fluxes in grasslands and croplands. Field Crops Research 252: 107791.

Journal Pre-proof

Journal Pre-proof

Sándor, R., Picon-Cochard, C., Martin, R., Louault, F., Klumpp, K., Borras, D., Bellocchi, G., 2018b. Plant acclimation to temperature: Developments in the Pasture Simulation model. Field Crops Research 222, 238-255.

Sansoulet, J., Pattey, E., Kröbel, R., Grant, B., Smith, W., Jégo, G., Desjardins, R.L., Tremblay, N., Tremblay, G., 2014. Comparing the performance of the STICS, DNDC, and DayCent models for predicting N uptake and biomass of spring wheat in Eastern Canada. Field Crops Research 156, 135-150.

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

Smith, K.A., Ball, T., Conen, F., Dobbie, K.E., Massheder, J., Rey, A., 2018. Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes. European Journal of Soil Science 69, 10-20.

Snow, V., Rotz, C.A., Moore, A.D., Martin-Clouaire, R., Johnson, I.R., Hutchings, N.J., Eckard, R.J., 2014. The challenges - and some solutions - to process-based modelling of grazed agricultural systems. Environmental Modelling & Software 62, 420-436.

Spence, M.A., Blanchard, J.L., Rossberg, A.G., Heath, M.R., Heymans, J.J., Mackinson, S., Serpetti, N., Speirs, D., Thorpe, R.B., Blackwell, P.G., 2017. Multi-model ensembles for ecosystem prediction. arXiv: 1709.05189.

Van Oijen, M., Barcza Z., Confalonieri R., Korhonen P., Kröel-Dulay G., Lellei-Kovács E., Louarn G., Louault F., Martin R., Moulin T., Movedi E., Picon-Cochard C., Rolinski S., Viovy N., Wirth S.B., Bellocchi, G., 2020. Incorporating biodiversity into biogeochemistry models to improve prediction of ecosystem services in temperate grasslands: review and roadmap. Agronomy 10: 259.

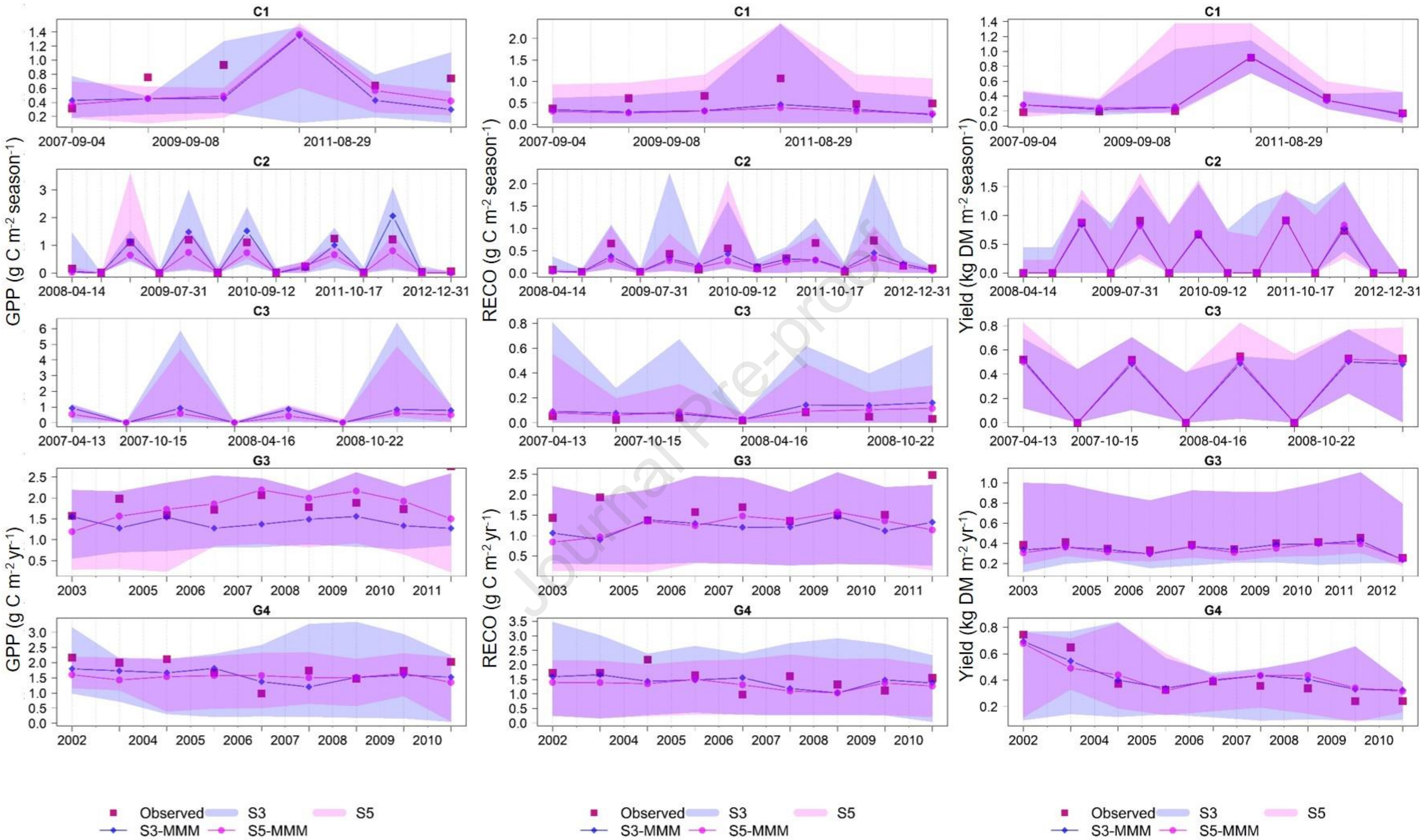
Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., van Ittersum, M., Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G., Dumont, B., Rezaei, E.E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom,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

Journal Pre-proof

Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hiremath, S., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M., Jansson, P.-E., Jing, Q., Justes, E., Kersebaum, K.C., Klosterhalfen, A., Launay, M., Lewan, E., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Padovan, G., Olesen, J.E., Poyda, A., Priesack, E., Pullens, J.W.M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A.K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T.K.D., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., Seidel, S.J., 2021. The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environmental Modelling & Software 145: 105206.

Wang, C., Amon, B., Schulz, K., Mehdi, B., 2021. Factors that influence nitrous oxide emissions from agricultural soils as well as their representation in simulation models: a review. Agronomy 11: 770.

Journal Pre-proof



Journal Pre-proof

Appendix A. Temporal changes of GPP (g C m^{-2} $season^{-1}$ for crops and g C m^{-2} yr^{-1} for grasslands, (left), RECO (g C m^{-2} $season^{-1}$ for crops and g C m^{-2} yr^{-1} for grasslands, middle) and Yield (kg DM m^{-2} $season^{-1}$ for crops and kg DM m^{-2} 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.

Journal Pre-proof

- 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

Journal Pre-proof

Authors declare no conflict of interest.

Journal Pre-proof