MS. FIONA EHRHARDT (Orcid ID : 0000-0002-8116-1804)

PROF. PETE SMITH (Orcid ID : 0000-0002-3784-1124)

DR. VAL SNOW (Orcid ID : 0000-0002-6911-8184)

DR. MATTHEW TOM HARRISON (Orcid ID : 0000-0001-7425-452X)

DR. MIKO UWE F KIRSCHBAUM (Orcid ID : 0000-0002-5451-116X)

DR. ELIZABETH MEIER (Orcid ID : 0000-0003-2394-8120)

DR. PAUL CD NEWTON (Orcid ID : 0000-0001-6346-5399)

Article type : Primary Research Articles

Assessing uncertainties in crop and pasture ensemble model simulations of productivity and N_2O emissions.

Running head: an international model intercomparison

Fiona Ehrhardt[1], Jean-François Soussana[1*], Gianni Bellocchi[2], Peter Grace[3], Russel McAuliffe[4], Sylvie Recous[5], Renáta Sándor[2, 6], Pete Smith[7], Val Snow[4], Massimiliano D. A. Migliorati[3], Bruno Basso[8], Arti Bhatia[9], Lorenzo Brilli[10], Jordi Doltra[11], Christopher D. Dorich[12], Luca Doro[13], Nuala Fitton[7], Sandro J. Giacomini[14], Brian Grant[15], Matthew T. Harrison[16], Stephanie K. Jones[17], Miko U. F. Kirschbaum[18], Katja Klumpp[2], Patricia Laville[19], Joël Léonard[20], Mark Liebig[21], Mark Lieffering[22], Raphaël Martin[2], Raia Silvia Massad[19], Elizabeth Meier[23], Lutz Merbold[24,25], Andrew D. Moore[26], Vasileios Myrgiotis[17], Paul Newton[22], Elizabeth Pattey[15], Susanne Rolinski[27], Joanna Sharp[28], Ward N. Smith[15], Lianhai Wu[29], Qing Zhang[30]

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/gcb.13965

[1]INRA, Paris, France,

[2]INRA, UMR Ecosystème Prairial, VetAgroSup, 63000 Clermont-Ferrand, France,

[3]Queensland University of Technology, Brisbane, Australia,

[4]AgResearch, Lincoln Research Centre, Lincoln, New Zealand,

[5]INRA, UMR FARE, Reims, France,

[6]HAS, CAR, Agricultural Institute, Martonvásár, Hungary,

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

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

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

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

[11]Cantabrian Agricultural Research and Training Center (CIFA), Muriedas, Spain,

[12]NREL, Colorado State University, Fort Collins, USA,

[13]Desertification Research Centre, University of Sassari, Sassari, Italy,

[14]Federal University of Santa Maria (UFSM), Soil Department, Santa Maria, Brazil,

[15]Agriculture and Agri-Food Canada, Ottawa Research and Development Center, Ottawa, Canada,

[16]Tasmanian Institute of Agriculture, P.O. Box 3523, Burnie, Tasmania, Australia, 7320,

[17]SRUC, West Mains Rd, Edinburgh, UK, EH9 3JG,

[18]Landcare Research, Palmerston North, New Zealand,

[19]INRA, UMR ECOSYS, Université Paris-Saclay, Thiverval-Grignon, France,

[20]INRA, UR AgroImpact, Laon, France,

[21]USDA Agricultural Research Service, Mandan, USA,

[22]AgResearch, Grasslands Research Centre, Palmerton North, New Zealand,

[23]CSIRO Agriculture and Food, St Lucia, Australia,

[24]ETH Zurich, Institute of Agricultural Sciences, 8092 Zurich, Switzerland,

[25]International Livestock Research Institute (ILRI), Mazingira Centre, P.O. Box 30709, 00100 Nairobi, Kenya,

[26]CSIRO, Agriculture & Food, Black Mountain Science and Innovation Precinct, GPO Box 1700, Canberra, Australia,

[27]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany,

[28]New Zealand Institute for Plant and Food Research, Christchurch, New Zealand,

[29]Sustainable Soils and Grassland Systems, Rothamsted Research, North Wyke, Devon, UK,

[30]LAPC, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China.

*Corresponding author: jean-francois.soussana@inra.fr; +33 (0)1 42 75 92 50

Key words: *greenhouse gases; climate change; agriculture; benchmarking; biogeochemical models; nitrous oxide; yield; soil*

Type of paper: *Primary Research Article*

Abstract

Simulation models are extensively used to predict agricultural productivity and greenhouse gas (GHG) emissions. However, the uncertainties of (reduced) model ensemble simulations have not been assessed systematically for variables affecting food security and climate change mitigation, within multispecies agricultural contexts. We report an international model comparison and benchmarking exercise, showing the potential of multimodel ensembles to predict productivity and nitrous oxide (N_2O) emissions for wheat, maize, rice and temperate grasslands. Using a multistage modelling protocol, from blind simulations (stage 1) to partial (stages 2-4) and full calibration (stage 5), 24 process-based biogeochemical models were assessed individually or as an ensemble against long-term

experimental data from four temperate grassland and five arable crop rotation sites spanning four continents. Comparisons were performed by reference to the experimental uncertainties of observed yields and N_2O emissions. Results showed that across sites and crop/grassland types, 23 to 40% of the uncalibrated individual models were within two standard deviations (s.d.) of observed yields, while 42 (rice) to 96% (grasslands) of the models were within one s.d. of observed N_2O emissions. At stage 1, ensembles formed by the three lowest prediction model errors (*RRMSE*) predicted both yields and N_2O emissions within experimental uncertainties for 44 and 33% of the crop and grassland growth cycles, respectively. Partial model calibration (stages 2 to 4) markedly reduced prediction errors of the full model ensemble E-median for crop grain yields (from 36% at stage 1 down to 4% on average) and grassland productivity (from 44 to 27%) and to a lesser and more variable extent for N_2O emissions. Yield-scaled N_2O emissions (N_2O emissions divided by crop yields) were ranked accurately by 3-model ensembles across crop species and field sites. The potential of using process-based model ensembles to predict jointly productivity and N_2O emissions at field scale is discussed.

Introduction

The need to mitigate climate change requires the abatement of greenhouse gas (GHG) emissions and the sequestration of organic carbon (C) in cropland and grassland soils. However, this must be accomplished while increasing agricultural productivity under climate change to keep up with global increasing demand and improve food and nutritional security (Smith *et al.*, 2008; Smith *et al.*, 2014; FAO, 2016). In order to meet the joint goals of reducing the impact of agriculture on climate change (UN Sustainable Development Goal, SDG 13) and delivering zero hunger (SDG 2), it is necessary to find solutions that reduce GHG emissions and that do not compromise food production. A measure that reduces GHG

emissions but that reduces productivity would be of limited use, as would a measure that increases production but that also increases GHG emissions. The concepts of 'yield-scaled emissions' as defined by Van Groenigen et al (2010), or emissions intensity (emissions per unit product), provide relevant indicators for food security and climate change (Venterea et al, 2011; Valin et al., 2013). It is therefore essential to compare both outputs (agricultural productivity and N_2O emissions) simultaneously with experimental data and simulation models.

Field experiments are essential to develop reference data on agricultural productivity, GHG emissions and mitigation options (Liebig *et al.,* 2016). However, they incur large costs, take many years to produce useful results, and it is generally difficult to extrapolate experimental results across space and time. Since the 1990s, the international scientific community has developed a number of simulation models that estimate GHG emissions and the dynamics of C and nitrogen (N) in agricultural (cropland and managed grassland) soils (Challinor *et al.*, 2013; Moore *et al.*, 2014; Jones *et al.*, 2016a). These models simulate interactions between the soil-plant-atmosphere continuum and agricultural management, enabling computation of transport and transformations of C and N in crop and pasture systems and subsequent responses of trace gas fluxes, such as N_2O emissions (Chen *et al.*, 2008) to agricultural practices. Sub-models are designed to interact with each other to describe cycles of water, C and N; thus any change in the management and environmental factors collectively affects a group of physical and biogeochemical processes either directly or indirectly via flow-on effects. Each of these process-based models offers a distinctive synthesis of scientific knowledge (Brilli *et al*., 2017) and has been calibrated and assessed by comparison with both field and laboratory experiments.

It has recently been shown that an ensemble of models may reduce the uncertainties of crop yield simulations across contrasting soil and climate conditions in comparison with single models (Asseng *et al.*, 2013; Challinor *et al.*, 2014; Li *et al.*, 2015; Maiorano *et al.*, 2016). The advantage of using ensemble predictors over individual models is due to compensation of errors across models, and a broader integration of model processes (Martre *et al.*, 2015). The use of model ensembles for reducing uncertainties at the national and international scales in simulations of agricultural production, such as grain and biomass production has therefore been recommended, noting however the benefits of using reduced-size model ensembles to limit the cost and complexity of multiple model simulations (Ruane *et al.*, 2016; Wallach *et al.*, 2016a). While there has been a range of published studies showing ensemble model simulation results for crop yields (e.g. Asseng *et al.*, 2013), we are not aware of any published model intercomparison assessing multiple models across experimental sites for N_2O emissions apart from the early study by Frolking et *al.* (1998), which investigated four individual models across three sites but did, however, not consider the median or mean of this model ensemble. Moreover, to our knowledge, no published study so far has assessed model ensembles across experimental sites for both N_2O and yields.

In previous ensemble studies, soil properties (e.g. soil N, soil organic C and soil moisture) that can affect crop simulations have been reset at the start of each growing season, thereby neglecting year-to-year plant-soil interactions that could potentially have large cumulative effects on yields, GHG emissions and soil organic C stocks (Basso *et al.*, 2015; Kollas *et al.*, 2015; Paustian *et al.*, 2016). Model ensembles were used for yield predictions with annual crop monocultures (e.g. maize: Bassu *et al.*, 2014; rice: Li *et al.*, 2015; wheat: Ruane *et al.*, 2016), but to a much lesser extent for crop rotations (Kollas *et al.*, 2015) and grasslands (Sándor *et al.*, 2016).

Here, we assess and report the results of 24 process-based integrated C&N models (16 cropland and 12 grassland models), by comparing multi-year (1 to 11 years) simulations to experimental data from nine sites (four temperate grasslands and five arable crop rotations with wheat, maize and rice) spanning four continents. The aim of this study was firstly to quantify the uncertainties of single models and model ensemble simulations; secondly, to assess, for the first time, the potential of model ensembles for predicting agricultural productivity and N_2O emissions, jointly, at field scale.

Materials & Methods

Experimental sites

The experimental sites were selected from those volunteered through an open call using research networks. The potential list was shortened to four permanent temperate grassland sites and five arable crop rotation sites covering geographically-diverse locations. These sites provided high-quality and previously published data (Table 1) encompassing climate, soil, agricultural practices, yields, crop and pasture development, N_2O emissions and, to the extent of possible, changes in soil organic C stocks. The main characteristics of the sites and the corresponding agricultural practices are summarized in Supplementary Materials (Tables S1 and S2). The experimental sites were also selected to cover a wide range of temperatures (annual means between 6 and 25 °C for croplands and between 6 and 13 °C for temperate grasslands) and precipitation amounts (annual totals in the range 630 - 1,800 mm and 430 - 1,100 mm at cropland and grassland sites, respectively) (Fig. 1).

The selected cropping systems covered a range of climates, from continental (C1, Canada), oceanic (C2, France), subtropical & semi-arid (C3, India; C4, Australia), to subtropical (C5, Brazil). All sites were in cultivated rotations and among the variety of crops within the crop sequence (detailed in Table S1), the most common crop types were simulated by models, i.e.

wheat, maize and rice. Sites included at least one wheat crop within the crop rotation, while maize was present at C1, C2, C4 and C5 locations, and rice was only cultivated at C3. The study compared, in total, 17 growing seasons with a total of nine, six and two crop growth cycles for wheat, maize and rice, respectively. At each crop site, the simulation periods of one year (C4), two years (C5), three years (C3) and five years (C1 and C2).

International collaborations have enabled the pooling and sharing of experimental data for temperate grasslands, including one site from the MAGGnet project (Liebig *et al.*, 2016) situated in the United States (G1), one Free-Air CO_2-Enriched experiment located in New Zealand (G2; only the ambient CO_2 treatment was considered here), and two European experimental sites (G3, France; G4, United Kingdom). G1 was a rangeland situated in the Northern Great Plains of the USA with a humid continental climate, while G2, G3 (semi-natural upland) and G4 (semi-natural moorland, intensively managed) were in the oceanic climate zone. All selected pastures were grazed with varying animal types: yearling steers (G1), non-lactating sheep (G2), heifers (G3) and, ewes, lambs and heifers-in-calf (G4). In addition, G4 had cutting events to harvest pasture for silage as well as supplementary feeding. Simulation periods, defined by the availability of experimental datasets, were 4 (G1), 9 (G4), 10 (G3) and 11 years (G2). Grassland yields were calculated as the Above-ground Net Primary Productivity (ANPP), which was determined at all sites either with grazing exclusion cages (at G3 and G4) with different herbage cutting heights (4 and 5 cm respectively), by the clipping method (at G1) or by 'difference' method (between herbage mass pre-grazing and post-grazing) at G2 (Table S3).

Nitrous oxide (N_2O) emissions were measured at all sites except G2. At one site (C1), measurements were performed by eddy flux covariance (Pattey *et al.*, 2006). At the remainder of the sites, N_2O emissions were measured by chambers, either using manual (C3,

C5, G1, G4) or automated (C2, C4, G3) measurements (Table S4). Other data of relevance to climate change mitigation (Table S5) were also obtained from the sites but are not reported here.

Contributing models

Modelling groups contributed to the study in response to an open call through the Global Research Alliance (GRA) on agricultural greenhouse gases, FACCE-JPI projects and other research networks, resulting in a set of representative coupled C-N models that are commonly used. The 24 published models selected (Table 1) simulate plant-soil-atmosphere interactions based on processes that are influenced by agricultural practices and that are designed to predict crop and/or pasture production, N_2O emissions (for 21 models) and changes in soil organic C stocks. A complete description of the contributing models is provided in Appendices S1 and S2, showing that these models vary in their complexity (number of parameters, type of inputs and outputs) and in their constitutive processes (Moore *et al.*, 2014; Brilli *et al.*, 2017). A total of 24 modelling teams from 11 countries contributed with 16 and 12 models to arable crop and grassland simulations, respectively, with four models contributing to both ensembles. The majority of the simulation models were run by a single modelling group. Nevertheless, five variants of APSIM and four variants of DayCent, each run by a different team, contributed to the simulations. Model anonymity was maintained throughout the process and model results are presented without attributing them to specific models or modelling teams.

A multistage protocol to compare and benchmark an ensemble of models

To ensure that model results would not be influenced by prior knowledge of the experimental data, a blind procedure was initially adopted, i.e. with no prior access to site-specific data concerning the simulated output variables (e.g. productivity and N_2O emissions). Moreover,

modelers were not provided with site name nor with the exact location, since sites were labelled at random (from C1 to C5 for crops and from G1 to G4 for grasslands).

Site-specific model calibration was performed at each modelling stage, with gradual access to site data from stage 2 onwards, to inform and parameterize the models. The protocol was organized in five stages (Fig. 2), from blind (stage 1) to partial (stages 2-4) and full (stage 5) calibration, by providing: (1) only basic data covering the experimental measurement period for model initialization (such as climate, soil initial properties and basic site management information, including description of crop rotation/grazing settings, fertilization and irrigation); (2) historical site-specific data for climate (Ruane et al., 2015) and management enabling long-term initialization periods, and regional statistics for wheat yield and pasture productivity from expert estimates; (3) site-specific phenology data, crop/pasture vegetation development (e.g. leaf area index), observed grain yields, monthly estimated grassland offtake (biomass removed by cuts or animal intake); (4) dynamic soil process data (temperature, moisture, mineral N) during the experiment; (5) observed data against which model outputs were compared, i.e. agricultural productivity (grain yields or ANPP together with daily changes in live weights of livestock categories and daily grassland offtake), GHG emissions and soil organic C stock changes (Fig. 2, Table S5). This final step opens possibilities for testing a set of mitigation options at the sites with the ensemble of fully calibrated models.

The modelling teams carried out their work independently and simultaneously for each of the five successive modelling stages. Access to additional experimental data was only allowed when the results of the previous stage had been submitted by all groups. Continuous multi-year simulations (i.e. without model re-initialization of dynamic soil variables) were required in all cases. Generic (site-independent) parameter values were kept constant, while a set of site-specific parameters were iteratively adjusted, based on the combination of experimental

data provided in each stage and on modeler's judgement. Modelers were requested to limit the number of adjusted parameters to those most influential whilst maintaining parameterization settings throughout the exercise. The results were submitted by each modelling team using common reporting templates, ensuring consistency in the units and a common definition of variables. The operations of data provision upload of model results and archives were all centralized within a common IT system.

Data analysis

At each stage and for each site, model outputs were compared with means of replicated field measurements and their standard deviation (s.d.) over the experimental period. In order to account for carry-over effects in the simulated responses, annual grain yields from the same crop type (wheat, maize, rice) grown within a rotation were averaged (Table S3). For pastures, annual ANPP was calculated based on measurements during the growing season (Table S3) and averaged by calendar year over the experimental period. N_2O emissions were calculated as means of daily emission values over days in which measurements were performed. Replicates were available for sites equipped with chambers (C2, C3, C4, C5, G1, G3, G4) while measurements with flux towers (C1) were performed with high temporal resolution but not replicated. In the latter case, the uncertainty in N_2O flux data was estimated from the literature (Kroon *et al.*, 2010). Daily N_2O flux datasets included, at some sites, a small number of negative values, suggesting an uptake of N_2O by the soil (Ammann *et al.*, 2009; Chapuis-Lardy *et al.*, 2007). The reliability of negative N_2O measurements is still questioned in the literature (e.g. Chapuis-Lardy et al., 2007; Cowan et al., 2014; Myrgiotis et al., 2016). Negative values were considered as negligible as they represented 8.3% of N_2O values at grasslands sites, 3.2% in wheat crops, 2.8% in maize crops and were not present in rice crops. The average difference between series with vs. without negative values was

between +0.1 to +12 µg N_2O-N m^{-2} d^{-1} across all sites. All this considered, model outputs were not compared to negative values. Means of daily N_2O fluxes (for observed days with non-negative values) were calculated with their associated s.d., over a crop cycle (from seeding to harvest) or over a calendar year for pastures.

Individual models and model ensemble accuracy compared to experimental uncertainties

The median of the multi-model ensemble (E-median) was taken as an indicator of the central tendency of the models. The relative average prediction error of the individual models and of the E-median was firstly assessed by using the relative root mean square error (*RRMSE*) (Bennett *et al.*, 2013):

$$RRMSE = 100 * \frac{1}{\bar{O}} * \sqrt{\frac{\sum_i^n (P_i - O_i)^2}{n}} \qquad (1)$$

where O_i and P_i are the ith observed and simulated values respectively, n is the number of O, P pairs, $\bar{O}$ is the mean of the observations. The individual model *RRMSE*s were initially calculated at each modelling stage across sites, with O, P pairs corresponding to the mean of observed and simulated seasonal or annual values, respectively, for grain yield of wheat, maize, rice and for grasslands yield (i.e. ANPP). For N_2O emissions, the O, P pairs correspond to the mean of observed and simulated daily fluxes, respectively, for days with measurements. The relative average prediction error of the E-median ($RRMSE_{E\text{-}median}$) was calculated from the median of individual model simulated values.

Secondly, model performances were assessed by reference to the variability in the experimental data, using centered and reduced model data deviation ($Z_{m,i}$), calculated for model m and observation i as:

$$Z_{m,i} = \frac{S_{m,i} - \bar{O}_i}{\sigma_{obs,i}} \qquad (2)$$

where, for the ith observation, $Z_{m,i}$ is the model (or E-median) data deviation, $S_{m,i}$ is the model (or E-median) simulated value, $\bar{O}_i$ is the observed value and $\sigma_{obs,i}$ is the standard deviation (s.d.) of observations. When the absolute value of $Z_{m,i}$ is lower or equal to x, the model (or the E-median) is within x s.d. of the observation mean. The number of models providing plausible estimates simultaneously for yields and N_2O at each site, was calculated by selecting models for which $Z_{m,i}$ was comprised between -2 and +2 for yields (x=2) and between -1 and +1 for N_2O emissions (x=1). The arbitrary choice of these thresholds was due to a conventional rule in a normal distribution, for which about 68% and 95% of the values fall within 1 s.d. and 2 s.d. of the mean, respectively. Thus, the threshold defined for N_2O is more stringent than the one for yields.

To rank models based on their mean prediction error for the two variables considered simultaneously (i.e. yields and N_2O emissions), we calculated a combined $RRMSE_c$ index as:

$$RRMSE_c = RRMSE_{yield} + RRMSE_{N2O} \qquad (3)$$

where, $RRMSE_{yield}$ and $RRMSE_{N2O}$ are the *RRMSE* for yield and for N_2O emissions, respectively. *RRMSEc* allowed ranking individual models with least average prediction errors across sites. Based on this ranking, ensembles formed by the three models with least *RRMSEc* were selected with the three crop species (wheat, maize, rice) and with grasslands and their median (3-median) values were calculated both for yields and for N_2O emissions. Finally, N_2O emission intensity (g N_2O-N kg^{-1} DM) was calculated by dividing N_2O emissions by crop grain DM production. The statistical package Sigmaplot v12.5 (Systat software) was used for statistical analysis.

Results

Observed crop and pasture productivity and N_2O emissions

The observed inter-annual means of grain yields for wheat, maize and rice were calculated for each site. Crop grain yields ranged between 0.25 and 0.82 kg DM m^{-2} $season^{-1}$ for wheat (five sites), 0.67 and 0.92 kg DM m^{-2} $season^{-1}$ for maize (four sites) and was 0.52 kg DM m^{-2} $season^{-1}$ for rice (single site) (Table S3). For grasslands, the ANPP estimated from replicated measurements over 3 to 11 years presented large contrasts, from 0.08 (G1) up to 1.27 kg DM m^{-2} yr^{-1} (G4) (Table S3). The mean yield coefficient of variation (CV) was 8.7% for wheat, 11.4% for maize and 9.3% for rice across years and all sites considered together, while with grassland ANPP, the mean CV across all grasslands sites was 17.5% (data not shown).

Daily means of soil N_2O emissions in cropping systems ranged from 300 to 1,200 µg N_2O-N m^{-2} d^{-1} for wheat (C1, C2, C3, C4, C5), from 360 to 1,300 µg N_2O-N m^{-2} d^{-1} for maize (C2, C4, C5) and reached 860 µg N_2O-N m^{-2} d^{-1} for rice (C3). N_2O fluxes had high CVs, with day-to-day variation ranging between 20 and 176% for wheat, 74 and 259% for maize and about 22% for rice (Table S4). At site C1, N_2O flux measurements were only provided for the wheat crop cycle and the subsequent snowmelt. The daily N_2O emissions from grasslands varied between 380 and 3,500 µg N_2O-N m^{-2} $d^{-1,}$ with CV values comprised between 101 and 186% for sites with chamber measurements (G1, G3, G4) (Table S4). At the site equipped with a flux tower (C1), the CV was assumed to be 50% at daily timescale following (Kroon *et al.*, 2010). No N_2O measurements were available at G2 site.

Prediction error from individual and ensemble models

Both with arable crops and with grasslands, no single model consistently outperformed other models by having the lowest *RRMSE* value both for N_2O and for yields (Tables S6, S7 and S8). With the E-median for crop grain yields, the *RRMSE* declined sharply from stage 1 (34,

31 and 45% for wheat, maize and rice, respectively) to stage 3 (6.4, 5.8 and 5.5% for wheat, maize and rice, respectively) and remained below 5% at stages 4 and 5 (Fig. 3a). For grasslands yield (i.e. ANPP), the $RRMSE_{E\text{-}median}$ declined from 44% at stage 1 to 27% at stage 3 and finally increased up to 46% at stage 5 (Fig. 3a).

For N_2O emissions of wheat and maize, the $RRMSE_{E\text{-}median}$ (Fig. 3b) was relatively stable throughout the calibration process and comprised between 42-47% and 50-55%, respectively. For rice N_2O emissions, the $RRMSE_{E\text{-}median}$ declined from 32% at stage 1 to 3% at stage 4, but increased up to 18% at stage 5. For grasslands N_2O emissions, the $RRMSE_{E\text{-}median}$ varied between 67% (at stage 1) and 96% (at stage 2).

Performances and uncertainties in model ensemble estimates and in observations

Yield estimates by individual models were considered to be plausible when they were within two standard deviations of the observed site mean (i.e. $Z_{m,i}$ between -2 and +2). At stages 1 and 2 (Fig. 4a, b, c), the E-median relative deviation with observed means was usually negative, showing an underestimation of yields by models with wheat (except at C1), rice and maize (except at C4). At further calibration stages (i.e. 3, 4 and 5), the E-median values were within two s.d. of the observed mean for all crops. At all stages, grassland yield (i.e. ANPP) was overestimated at G1 and under-estimated at G4 by the E-median (Fig. 4d). At G2 and G3, E-medians were within two s.d. of observed means at stages 2 to 4 and at stages 2 to 5, respectively.

At all stages and for all sites, the E-median of N_2O emissions was within one s.d. of the observed mean both for crops and grasslands, except at C3 for wheat where N_2O emissions were underestimated (Fig. 4 e, f, g, h). The detailed model relative deviations ($Z_{m,i}$) for yields and N_2O emissions according to sites and modelling stages are shown in Supplementary Materials (Figs. S1 and S2).

Finally, Table 2 compares full size model ensembles (E-median) and reduced size model ensembles (3-median, for the ensemble of three models with least average *RRMSEc*). Site specific E-medians and 3-medians were considered as plausible when they were within two and one s.d. of observed means for yields and for N_2O emissions, respectively. With uncalibrated models (stage 1) and for the prediction of both variables together, the 3-median provided plausible estimates at two wheat sites out of five (C2, C4), one maize site out of three (C5), at the single rice site (C3) and at one grassland site out of three (G1). The full size model ensemble E-median did not perform better, since it was a plausible estimator of both yield and N_2O emissions at two wheat sites out of five and one maize site out of three, while failing to predict in a plausible way the rice site or any of the grassland sites.

Using Spearman's rank correlations with reduced model ensembles (3-Median) (Fig.S4), we show a significant correlation between simulated and observed N_2O emission intensities (g N_2O-N kg^{-1} DM) across sites and crops since stage 1 (ρ=0.72, p=0.025). This correlation becomes highly significant after provision of phenology data at stage 3 (ρ=0.82, p<0.005) and further increases at stage 5 (ρ=0.93, p<0.0001).

Proportion of contributing models with plausible estimates

At stage 1 and all sites taken together (C1-C5), plausible estimates (i.e. $Z_{m,i}$ between -2 and +2) were found for 26, 40 and 23% of the contributing models for wheat, maize and rice grain yields, respectively. At stage 2, this proportion decreased slightly for maize and increased slightly for wheat and rice. At stages 3, 4 and 5, the percentage of plausible models increased, reaching at the final stage 60, 70 and 60% for wheat, maize and rice, respectively. For grassland yield (i.e. ANPP), the mean percentage of plausible models (G1-G4) was in the same range than with grain yields (26 and 37% at stage 1 and stage 2, respectively) and decreased in subsequent stages down to 25% at final stage (Fig. 5a).

N_2O emission estimates by individual models were considered to be plausible when they were within one s.d. of the observed mean (i.e. $Z_{m,i}$ between -1 and +1). Taken as a mean of all sites (C1-C5 and G1-G4), up to 84 and 96 % of individual model estimates were found to be plausible for maize (at stage 2) and for grasslands (at stage 1). In contrast, the percentage of plausible models did not exceed 60 and 42% for wheat (at stage 5) and for rice (at stage 1) (Fig. 5b).

The percentage of individual models with plausible estimates both for yields and for N_2O emissions reached up to 39 and 49% for wheat (stage 5) and for maize (stage 4) respectively, while it did not exceed 20% for rice (from stage 2) and 23% for grasslands (at stage 2) (Fig. 5c).

Discussion

This study provides the first assessment of process-based simulation models used for simultaneous estimates of crop and pasture productivity and of N_2O emissions in response to climate, soil and management conditions. The statistical approach of model error adopted in this study is based on predictions averaged over space (means of replicate measurements) and time (seasonal and annual means) (Wallach & Thorburn, 2014). Compared to Willmott *et al.* (2012), where model performance metric (index of agreement, d_r) ranges from 0 to 1, our dimensionless indicator ($Z_{m,i}$) scales the model performance by considering the uncertainties in the measurements and allows for assessing model estimates on an observed s.d. basis. We have compared simulation results from multiple model structures (i.e. model ensembles), multiple input vectors (i.e. site comparison) and multiple parameter vectors by allowing for improved calibration of each model during successive modelling stages as recommended by Wallach *et al.* (2016b). Modelers' expertise and knowledge still remains a non-negligible source of uncertainties, as described by Confalonieri *et al.* (2016) and an investigation on

how model experts used the information gradually released in our exercise would be of great interest to understand how and why models results were improved.

Grain yields

Compared to previous studies (Asseng *et al.*, 2013; Bassu *et al.*, 2014), grain yields were estimated here by models able to simulate full crop rotations, including fallows, without resetting soil states and thereby provided estimates resulting from integrated C and N cycles at field scale. Therefore, the model ensemble used in our study differs substantially from ensembles used in previous studies, e.g. only eight models in the present study were in common with the 27 models reported by Asseng *et al*. (2013).

Without site-specific information (stage 1), the $RRMSE_{E\text{-}median}$ was approximately three times larger than the s.d. of the observations in the case of wheat and maize yields. Providing measured phenology and grain yield values at stage 3 allowed for improved model calibration corresponding to a strong reduction in the model ensemble prediction error ($RRMSE_{E\text{-}median}$ reduced down to 6% for wheat, maize and rice yields). These results are in line with those reported by Asseng et al. (2013) for wheat with uncalibrated and calibrated models (23 and 5% respectively), by Bassu *et al.* (2014) for maize (7% for fully calibrated models) and by Li *et al.* (2015) for rice grain yields. Compared to these reports, where flowering dates were used to run uncalibrated models, we provided only sowing and harvest dates at stage 1 which resulted in larger prediction errors. In the same way for wheat, Palosuo *et al.* (2011) noted that, in spite of phenological observations (emergence, flowering, ripening and harvest dates) being provided to models, simulated dates of flowering and maturity are highly variable across models, and that model simulations are greatly improved by accessing such phenological data. Another factor that impacts model simulations may be the dynamics of available N contributing to the grain filling of wheat. N mineralization rate as a function of

soil temperature and moisture (Salo *et al.*, 2016) is often not well captured by models which may explain the absence of model improvements at stage 4 (i.e. after provision of physico-chemical soil data to the modelling teams).

In our study, the E-median estimates for wheat, maize and rice grain yields were as good as those presented in previous multi-model studies with simplified modelling methodologies, thus confirming the reliability of using model ensembles for realistic field conditions (multi-year crop rotations and grazed pastures) and reinforcing the conclusions by Basso *et al*. (2015) and Kollas *et al*. (2015).

Grassland productivity (ANPP)

In grazed pastures, herbage offtake by domestic herbivores is a function of the grazing pressure (driven by animal stocking density and liveweight) thus can be directly estimated from variables provided at stage 1, and is therefore not useful for model benchmarking purposes. In order to keep a strict blind test, the ANPP was used to benchmark simulated grassland productivity. Modelling the ANPP of temperate pastures has often been found to be difficult, given the large variability in vegetation composition and structure (Snow *et al.*, 2014). Indeed at stage 1, grassland ANPP was poorly predicted by the ensemble of models (E-median prediction error of 44% with only 22% of plausible models). At further stages (2-5), only few improvement were obtained and systematic trends in the E-median data-deviation was observed with minimum value at stage 3 for all sites (except G1), while estimated monthly biomass removal (i.e. biomass cut and grazed, the latter calculated from information about the animal stock liveweight and density), leaf area index and flowering dates were provided. Such discrepancies between simulated and observed values can be caused both by data and by model limitations. Indeed, methods for measuring grassland ANPP were not standardized across sites (i.e. varying cutting heights within grazing

exclosure cages, number of replicates and sampling frequencies, Table S3) causing likely substantial bias in productivity estimates at some sites (Smit *et al.*, 2008). Moreover, model overestimation compared to measurements could be explained by several factors: i) models include all aboveground compartments in ANPP calculations, while measurements only include shoots above the cutting height without plant residues (i.e. stubble); ii) most models do not account for effects of spatial heterogeneity (i.e. trampling, vegetation, dung and urine patches) on pasture productivity (e.g. Snow *et al.*, 2017); iii) calibration methods in response to grazing offtake vary across models. Such differences cause limitations to the use of large model ensembles for grasslands ANPP estimates and we highlighted improved performances of reduced-size model ensemble.

Crop and pasture N_2O emissions

To account for the large variability across replicated N_2O emission measurements (Table S4), a more stringent criterion for model plausibility was adopted, i.e. within one s.d. of the observed mean. Already at stage 1, the E-median was plausible for N_2O emissions however, in contrast to grain yield, prediction errors of E-medians ($RRMSE_{E\text{-}median}$) for N_2O emissions did not show a large decline through the calibration stages, and were ranged between 67 and 96% for grasslands, 42 and 55% for wheat and maize and between 3 and 32% for rice. These values are somewhat lower than with previous reports, since Frolking *et al.* (1998) reported simulated N_2O fluxes within a factor of about two of the observed annual fluxes. With fully calibrated models for a highly fertilized winter wheat-summer maize rotation system, Zhang *et al.* (2015) obtained a lower average *RRMSE* (27%) but an overestimation of N_2O emissions with three models. These authors suggested that a model ensemble would perform better than single models, but they did not show a reduction in prediction errors by using model medians.

In addition, it has been suggested by Frolking *et al.* (1998) and Abdalla *et al.* (2009) that soil moisture content and water filled pore space are key requirements for reliable simulations of N_2O emissions. Similarly, Saggar *et al.* (2013) have underlined the role of soil temperature and nitrate availability to impact nitrification and denitrification processes. However, providing seasonal values of soil temperature, moisture and mineral N did not significantly reduce the $RRMSE_{E\text{-}median}$ in our study, with the exception of rice at stage 4 ($RRMSE_{E\text{-}median}$ down to 24%). Differences in calibration methods (number of soil and plant parameters being adjusted, use of automated or manual calibration routines) may explain the overall lack of improvement in accuracy with model calibration for wheat, maize and grasslands N_2O estimates. It should also be noted that not all the events and management activities causing N_2O emissions (e.g. freeze/thaw cycles, or water management in rice) were recorded at the experimental sites.

Implications for field estimates of agricultural productivity and N_2O emissions

Our study has allowed model assessment and calibration at distant and contrasted sites, thereby potentially overcoming bias caused by model inter-comparison in a local specific context. Both for agricultural productivity and for N_2O emissions, we show that reduced complexity model ensembles, obtained by selecting uncalibrated models with least average error, can perform as well, or even better, than full model ensembles. This result paves the way to the use of small model ensemble medians for field scale estimation of yields and soil based GHG emissions. Nevertheless, the three model ensembles selected in our study differ across crop species and are not the same with grasslands compared to arable crops. Further improvements of data sources (e.g. phenological observations) could help defining best model ensembles that could be used for screening agricultural practices and mitigation options at international crop and grassland sites.

For the first time, our results show the potential of multi-model ensembles for estimating jointly agricultural productivity and N_2O emissions. Yield-scaled emissions (i.e. N_2O emission intensities) are relevant for two policy dimensions: food security and climate change. With arable crops, our results show that the median of 3-model ensembles predicts significantly the ranking of observed N_2O emission intensities (i.e. N_2O emitted per unit of grain production, $gN_2O\text{-}N.kg^{-1}$ DM) (Figure S4). Therefore, cropping systems could be simulated and ranked by N_2O emission intensity, in order to test options for improving agricultural productivity while reducing GHG emissions.

Finally, our results question the use of model ensembles for upscaling projections of agricultural productivity and N_2O emissions from field scale to larger spatial units (e.g. gridded projections) as needed for Tier 3 national inventories. Such ensemble projections have recently been used for global simulations of climate change impacts on wheat and maize yields (Rosenzweig *et al.*, 2014; Elliott *et al.*, 2015), neglecting however soil spatial variability which is likely to reduce the accuracy of yield projections (Folberth *et al.*, 2016). The establishment of a global network of carefully standardized and long-term field experiments measuring GHG emissions, soil organic C stocks, and crop and grassland yields, would provide an essential foundation to further reduce uncertainties of model ensemble estimates both at field and regional scales, and to test the impacts of mitigation practices and of climate change. International modelling efforts should converge to work on complementary scales (from local to global), since global estimates (such as grids) are essential to determine major trends, while field scale simulations help in refining agricultural practices or selecting new/other cultivars adapted to existing local agronomic contexts. Such efforts could be fueled by the emergence of new generation technologies, especially collaborative online platforms facilitating the sharing of data and modelling tools and supporting decision making, as well as Tier 3 methods for national GHG inventories.

Author contributions

JFS, FE, GB, PG, SR, VS and PS conceived and designed the research. EP, PL, RSM, AB, MDAM, SJG, PN, ML, KK, SKJ and LM provided experimental data. BB, AB, LB, JD, CD, LD, NF, BG, MTH, MUFK, JL, RM, RSM, EM, AMM, VM, SR, JS, ML, WNS, LW and QZ performed simulations. FE, RMA, RS assisted in the collation of simulation results. FE performed statistical analysis. FE and JFS wrote the manuscript. GB, PG, SR, VS, PS firstly, and JD, BG, MTH, KK, MUFK, AMM, LM, WNS secondly, commented on and revised the manuscript.

Acknowledgements

FE acknowledges support through a grant from ADEME (n° 12-60-C0023). This study was coordinated by the Integrative Research Group of the Global Research Alliance (GRA) on agricultural greenhouse gases and was supported by five research projects (CN-MIP, Models4Pastures, MACSUR, COMET-Global and MAGGNET) funded by a multi-partner call on agricultural GHGs with support of FACCE JPI. LM was supported by the Swiss National Science Foundation under the 40FA40_154245 / 1 grant agreement; JD participated in the framework of Red REMEDIA. The authors wish to thank Dr. Alex Ruane (NASA GISS) for provision of AgMERRA weather data; Benjamin Loubet (INRA) and Kathrin Fuchs (ETH-Zürich) for their contribution to the consolidation of experimental datasets; Laura Cardenas (Rothamsted Research), Marco Carozzi (INRA) and DairyMod team (Karen Christie, Brendan Cullen, Rachel Meyer, Richard Eckard, Richard Rawnsley) for their modelling efforts; Marco Bindi (University of Florence), Rich Conant (Colorado State University), Heinrike Mielenz (Julius Kühn Institute) and Kairsty Topp (SRUC), for their help as supervisors.

Supplementary information

Supplementary Materials and Methods, Supplementary Results and Appendices on models can be found in the Supplementary Material file.

References

Abdalla M, Wattenbach M, Smith P, Ambus P, Jones M, Williams M (2009) Application of the DNDC model to predict emissions of N_2O from Irish agriculture. *Geoderma*, **151**, 327–337.

Aita C, Gonzatto R, Miola ECC *et al*. (2014) Injection of dicyandiamide-treated pig slurry reduced ammonia volatilization without enhancing soil nitrous oxide emissions from no-till corn in Southern Brazil. *Journal of Environment Quality*, **43**, 789.

Allard V, Soussana J-F, Falcimagne R *et al.* (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, Ecosystems & Environment*, **121**, 47–58.

Ammann C, Spirig C, Leifeld J, Neftel A (2009) Assessment of the nitrogen and carbon budget of two managed temperate grassland fields. *Agriculture, Ecosystems & Environment*, **133**, 150–162.

Asseng S, Ewert F, Rosenzweig C *et al*. (2013) Quantifying uncertainties in simulating wheat yields under climate change. *Nature Climate Change*, **33**, pp.827-832.

Basso B, Hyndman DW, Kendall AD, Grace PR, Robertson GP (2015) Can impacts of climate change and agricultural adaptation strategies be accurately quantified if crop models are annually re-initialized? PloS one, **10**(6), e0127333.

Bassu S, Brisson N, Durand JL *et al.* (2014) How do various maize crop models vary in their responses to climate change factors? *Global Change Biology*, **20**, 2301–2320.

Bennett ND, Croke BFW, Guariso G *et al*. (2013) Characterising performance of environmental models. *Environmental Modelling & Software*, **40**, 1–20.

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

Brilli L, Bechini L, Bindi *et al.* (2017) Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. *Science of the Total Environment*, in press. 10.1016/j.scitotenv.2017.03.208.

Challinor A, Martre P, Asseng S, Thornton P, Ewert F (2014) Making the most of climate impacts ensembles. *Nature Climate Change*, **4**, 77–80.

Challinor AJ, Smith MS, Thornton P (2013) Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. *Agricultural and Forest Meteorology*, **170**, 2–7.

Chapuis-Lardy L , Wrage N, Metay A, Chotte JL, Bernoux M (2007) Soils, a sink for N_2O? A review. *Global Change Biology*, **13**, 1–17.

Chen D, Li Y, Grace P *et al.* (2008) N_2O emissions from agricultural lands: A synthesis of simulation approaches. *Plant and Soil*, **309**, 169–189.

Confalonieri R, Orlando F, Paleari L *et al*. (2016) Uncertainty in crop model predictions: What is the role of users? *Environmental Modelling & Software*, **81**, 165–173.

Cowan NJ, Famulari D, Levy PE, Anderson M, Reay DS, Skiba UM (2014) Investigating uptake of N_2O in agricultural soils using a high-precision dynamic chamber method. *Atmospheric Measurement Techniques*, **7**, 8125-8147.

De Antoni Migliorati M, Scheer C, Grace PR, Rowlings DW, Bell M, McGree J (2014) Influence of different nitrogen rates and DMPP nitrification inhibitor on annual N_2O emissions from a subtropical wheat–maize cropping system. *Agriculture, Ecosystems & Environment*, **186**, 33–43.

Del Grosso SJ, Parton WJ, Mosier AR, Walsh MK, Ojima DS, Thornton PE (2006) DAYCENT national-scale simulations of nitrous oxide emissions from cropped soils in the United States. *Journal of Environmental Quality*, **35**, 1451–1460.

Elliott J, Müller C, Deryng D *et al.* (2015) The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0). *Geoscientific Model Development*, **8**, 261–277.

FAO (ed.) (2016) *Climate change, agriculture and food security*. FAO, Rome, 173 pp.

Folberth C, Skalský R, Moltchanova E, Balkovič J, Azevedo LB, Obersteiner M, van der Velde M (2016) Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. *Nature Communications*, **7**, 11872.

Frolking SE, Mosier AR, Ojima DS *et al.* (1998) Comparison of N_2O emissions from soils at three temperate agricultural sites: simulations of year-round measurements by four models. *Nutrient Cycling in Agroecosystems*, **52**, 77–105.

Gerber PJ, Steinfeld H, Henderson B *et al.* (2013) Tackling climate change through livestock – A global assessment of emissions and mitigation opportunities. Food and Agriculture Organization of the United Nations (FAO), Rome.

GRA, 2017. Global Research Alliance on agricultural GHG, https://globalresearchalliance.org/about/; accessed 09/25/17.

IPCC (2006) IPCC Guidelines for national greenhouse gas inventories. Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge.

Jones JW, Antle JM, Basso BO *et al*. (2016a) Brief history of agricultural systems modeling, *Agricultural Systems*, **155**, 240-254.

Jones SK, Helfter C, Anderson M *et al*. (2016b) The nitrogen, carbon and greenhouse gas budget of a grazed, cut and fertilised temperate grassland. *Biogeosciences Discuss*., **14,** 2069-2088.

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

Kollas C, Kersebaum KC, Nendel C *et al.* (2015) Crop rotation modelling—A European model intercomparison. *European Journal of Agronomy*, **70**, 98–111.

Kroon PS, Hensen A, Jonker HJJ, Ouwersloot HG, Vermeulen AT, Bosveld FC (2010) Uncertainties in eddy covariance flux measurements assessed from CH4 and N_2O observations. *Agricultural and Forest Meteorology*, **150**, 806–816.

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.

Li T, Hasegawa T, Yin X *et al.* (2015) Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. *Global Change Biology*, **21**, 1328–1341.

Liebig MA, Franzluebbers AJ, Alvarez C *et al*. (2016) MAGGnet: An international network to foster mitigation of agricultural greenhouse gases. *Carbon Management*, **7,** 243-248.

Liebig MA, Gross JR, Kronberg SL, Hanson JD, Frank AB, Phillips RL (2006) Soil response to long-term grazing in the northern Great Plains of North America. *Agriculture, Ecosystems & Environment*, **115**, 270–276.

Liebig MA, Gross JR, Kronberg SL, Phillips RL (2010) Grazing management contributions to net global warming potential: a long-term evaluation in the Northern Great Plains. *Journal of Environment Quality*, **39**, 799.

Liebig MA, Kronberg SL, Hendrickson JR, Dong X, Gross JR (2013) Carbon dioxide efflux from long-term grazing management systems in a semiarid region. *Agriculture, Ecosystems & Environment*, **164**, 137–144.

Loubet B, Laville P, Lehuger S *et al.* (2011) Carbon, nitrogen and greenhouse gases budgets over a four years crop rotation in northern France. *Plant and Soil*, **343**, 109–137.

Maiorano A, Martre P, Asseng S *et al.* (2016) Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research*, **202**, 5-20.

Martre P, Wallach D, Asseng S *et al.* (2015) Multimodel ensembles of wheat growth: Many models are better than one. *Global Change Biology*, **21**, 911–925.

Moore AD, Holzworth DP, Herrmann NI *et al.* (2014) Modelling the manager: Representing rule-based management in farming systems simulation models. *Environmental Modelling & Software*, **62**, 399–410.

Myrgiotis V, Williams M, Rees RM, Smith KE, Thorman RE, Topp CFE (2016) Model evaluation in relation to soil N_2O emissions: An algorithmic method which accounts for variability in measurements and possible time lags. *Environmental Modelling & Software* **84**, 251-262.

Newton PCD, Lieffering M, Bowatte WMSD, Brock SC, Hunt CL, Theobald PW, Ross DJ (2010) The rate of progression and stability of progressive nitrogen limitation at elevated atmospheric CO_2 in a grazed grassland over 11 years of Free Air CO_2 enrichment. *Plant and Soil*, **336**, 433–441.

Newton PCD, Lieffering M, Parsons AJ *et al*. (2014) Selective grazing modifies previously anticipated responses of plant community composition to elevated CO_2 in a temperate grassland. *Global Change Biology*, **20**, 158–169.

Ogle SM, Jay Breidt F, Easter M, Williams S, Killian K, Paustian K (2010) Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. *Global Change Biology*, **16**, 810–822.

O'Leary GJ, Liu DL, Ma Y *et al.* (2016) Modelling soil organic carbon 1. Performance of APSIM crop and pasture modules against long-term experimental data. *Geoderma*, **264**, 227–237.

Palosuo T, Kersebaum KC, Angulo C *et al.* (2011) Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. *European Journal of Agronomy*, **35**, 103–114.

Pattey E, Edwards G, Strachan IB *et al.* (2006) Towards standards for measuring greenhouse gas fluxes from agricultural fields using instrumented towers. *Canadian Journal of Soil Science*, **86**, 373-400.

Paustian K, Lehmann J, Ogle S, Reay D, Robertson GP, Smith P (2016) Climate-smart soils. *Nature*, **532**, 49–57.

Rosenzweig C, Elliott J, Deryng D *et al.* (2014) Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences*, **111**, pp 3268-3273.

Ruane AC, Goldberg R, Chryssanthacopoulos J (2015) Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation. *Agricultural and Forest Meteorology*, **200**, 233–248.

Ruane AC, Hudson NI, Asseng S *et al.* (2016) Multi-wheat-model ensemble responses to interannual climate variability. *Environmental Modelling and Software*, **81**, 86–101.

Saggar S, Jha N, Deslippe J *et al.* (2013) Denitrification and N_2O: N_2 production in temperate grasslands: Processes, measurements, modelling and mitigating negative impacts. *Science of the Total Environment*, **465**, 173–195.

Salo TJ, Palosuo T, Kersebaum KC *et al.* (2016) Comparing the performance of 11 crop simulation models in predicting yield response to nitrogen fertilization. *The Journal of Agricultural Science*, **154**, 1218–1240.

Sándor R, Barcza Z, Acutis M *et al.* (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.

Sansoulet J, Pattey E, Kröbel R *et al.* (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 SK, Drewer J *et al.* (2013) Comparison of soil greenhouse gas fluxes from extensive and intensive grazing in a temperate maritime climate. *Biogeosciences*, **10**, 1231–1241.

Smit HJ, Metzger MJ, Ewert F (2008) Spatial distribution of grassland productivity and land use in Europe. *Agricultural Systems*, **98**, 208–219.

Smith P, Clark H, Dong H *et al.* (2014) Chapter 11 - Agriculture, forestry and other land use (AFOLU). In: Climate Change 2014: Mitigation of Climate Change. IPCC Working Group III Contribution to AR5. Cambridge University Press.

Smith P, Martino D, Cai Z *et al.* (2008) Greenhouse gas mitigation in agriculture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 789–813.

Smith P, Smith JU, Powlson DS *et al.* (1997) A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments: evaluation and comparison of soil organic matter models. *Geoderma*, **81**, 153–225.

Snow VO, Cichota R, McAuliffe RJ, Hutchings NJ, Vejlin J (2017) Increasing the spatial scale of process-based agricultural systems models by representing heterogeneity: The case of urine patches in grazed pastures. *Environmental Modelling & Software*, **90**, 89–106.

Snow VO, Rotz CA, Moore AD, Martin-Clouaire R, Johnson IR, Hutchings NJ, Eckard RJ (2014) The challenges - and some solutions - to process-based modelling of grazed agricultural systems. *Environmental Modelling and Software*, **62**, 420–436.

Valin H, Havlik P, Mosnier, A, Herrero M, Schmid E, Obersteiner M (2013) Agricultural productivity and greenhouse gas emissions: trade-offs or synergies between mitigation and food security? *Environmental Research Letters*, **8**, 035019.

Van Groenigen JW, Velthof GL, Oenema O, Van Groenigen KJ, Van Kessel C (2010) Towards an agronomic assessment of N_2O emissions: a case study for arable crops. *European Journal of Soil Science*, **61**, 903-913.

Venterea RT, Bijesh M, Dolan MS (2011) Fertilizer source and tillage effects on yield-scaled nitrous oxide emissions in a corn cropping system. *Journal of Environmental Quality*, **40**, 1521-1531.

Wallach, D, Mearns, LO, Ruane, AC, *et al.* (2016a) Lessons from climate modeling on the design and use of ensembles for crop modeling. *Climatic Change*, **139** (3-4), 551-564.

Wallach D, Thorburn PJ (2014) The error in agricultural systems model prediction depends on the variable being predicted. *Environmental Modelling and Software*, **62**, 487–494.

Wallach D, Thorburn P, Asseng S *et al.* (2016b) Estimating model prediction error: Should you treat predictions as fixed or random? *Environmental Modelling and Software*, **84**, 529–539.

Willmott CJ, Robeson SM, Matsuura K (2012) A refined index of model performance. *International Journal of Climatology*, **32**, 2088–2094.

Yin X, Kersebaum KC, Kollas C *et al.* (2017) Performance of process-based models for simulation of grain N in crop rotations across Europe. *Agricultural Systems*, **154**, 63–77.

Zhang W, Liu C, Zheng X *et al.* (2015) Comparison of the DNDC, LandscapeDNDC and IAP-N-GAS models for simulating nitrous oxide and nitric oxide emissions from the winter wheat-summer maize rotation system. *Agricultural Systems*, **140**, 1–10.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Tables

Table S1. Field experiments, climate characteristics and crop management of five selected crop rotations used for model comparison.

Table S2. Field experiments, climate characteristics and grassland management of four selected grassland sites used for model comparison.

Table S3. Experimental set-up for yield measurements at the ten selected sites (C1-C5 and G1-G4) with mean, standard deviation (s.d.) and coefficient of variation (CV) of grain yields (wheat, maize and rice) and ANPP observations.

Table S4. Experimental set-up for N_2O measurements at the ten selected sites (C1-C5 and G1-G4) with overall mean, standard deviation (s.d.) and coefficient of variation (CV) of daily N_2O emissions.

Table S5. Input data provided to models according to the multistage approach.

Table S6. Individual model *RRMSE* according to modelling stages for the estimation of wheat, maize and rice grain yields.

Table S7. Individual model *RRMSE* according to modelling stages for the estimation of wheat, maize and rice N_2O emissions.

Table S8. Individual model *RRMSE* according to modelling stages for the estimation of grasslands ANPP and N_2O emissions.

Figures

Figure S1. Model-data relative deviation (expressed in s.d. of observation) for wheat, maize and rice grain yields (a, b, c) and for grasslands ANPP (d) according to sites and modelling stages 1 to 5.

Figure S2. Model-data relative deviation (expressed in s.d. of observation) for wheat, maize, rice and grasslands N_2O emissions (a, b, c, d) according to sites and modelling stages 1 to 5.

Figure S3. Percentage of single models providing plausible estimates simultaneously for yields and for N_2O emissions by site (C1-C5; G1-G4) and by modelling stage (1-5) for wheat, maize, rice and grasslands.

Figure S4. Simulated vs. observed N_2O emission intensity for wheat, maize and rice at stages 1, 3 and 5.

Appendices

Appendix S1. Description of models.

Appendix S2. References for model description.

Tables

Table 1 References of experimental sites and of models contributing to the model benchmarking. C, cropland site; G, grassland site. Sites were arbitrarily numbered from 1 to 5 for crop sites and from 1 to 4 for grasslands. A detailed description of the models and literature references is provided in Supplementary Materials (Appendices S1 and S2).

Site	Country	Main references
C1	Canada	Pattey *et al.* (2006); Sansoulet *et al.* (2014)
C2	France	Laville *et al.* (2011); Loubet *et al.* (2011)
C3	India	Bhatia *et al.* (2012)
C4	Australia	De Antoni Migliorati *et al.* (2014)
C5	Brazil	Aita *et al.* (2014)
G1	USA	Liebig *et al.* (2006, 2010, 2013)
G2	New Zealand	Newton *et al.* (2010, 2014)
G3	France	Allard *et al.* (2007); Klumpp *et al.* (2011)
G4	UK	Skiba *et al.* (2013); Jones *et al.* (2016b)

Model	Version	System	Web address
Agro-C	1.0	C	
APSIM	7.5	C	
APSIM	7.6	C	
APSIM	7.6 Grazplan	G	http://www.apsim.info
APSIM	7.7 SoilWat	G	
APSIM	7.7 SWIM	G	
CenW	4.1	G	http://www.kirschbaum.id.au/Welcome_Page.htm
CERES-EGC		C	https://www6.versailles-grignon.inra.fr/ecosys/Productions/Logiciels-Modeles/CERES-EGC
DairyMod/SGS	4	G	http://www.imj.com.au/dm
DayCent	4.5 2006	C; G	
DayCent	4.5 2013	C; G	
Daily DayCent	4.5 2010	C; G	http://www.nrel.colostate.edu/projects/daycent-downloads.html
Daily Daycent	4.5 2013	C	
DNDC	CAN	C	http://www.dndc.sr.unh.edu
			http://gramp.org.uk/models/104
DSSAT	GHG	C	http://dssat.net
EPIC	810	C	http://epicapex.tamu.edu/model-executables
FASSET	2.5	C	http://www.fasset.dk
INFOCROP	2.1	C	http://www.iari.res.in/?option=com_content&view=article&id=1334
Landscape-DNDC	0.9.2	C; G	Under licence agreement with Institute of Meteorology and Climate Research, Germany
LPJmL	3.5.003	C	Precursor model (LPJ): https://www.pik-potsdam.de/research/projects/activities/biosphere-water-modelling/lpjml. Current version (LPJmL v4), available by December 2016
PaSim		G	https://www1.clermont.inra.fr/urep/modeles/pasim.htm Request to raphael.martin@inra.fr
SALUS		C	Request to basso@msu.edu
SPACSYS	5.0	G	Request to lianai.wu@rothamsted.ac.uk
STICS	831	C	http://www6.paca.inra.fr/stics_eng

Table 2 Summary of uncalibrated (stage 1) model ensembles assessment for the accuracy of yield and N_2O emission predictions. E-median and 3-median correspond to full (up to 15 for crops and to 9 for grasslands) and three ensemble of models, respectively (see Materials and Methods); 1-var perf. and 2-var perf., are the number of sites with plausible medians out of the total number of sites for one and two variables, respectively; Black cell represent plausible estimate by the median (within 2 and 1 s.d. of observed means for yields and for N_2O emissions, respectively); grey cell, non-available experimental data; white cell, median outside the plausibility range.

Stage 1

Crop rotations		**Wheat**		**Maize**		**Rice**		**Permanent grasslands**		
		Yield	N_2O	Yield	N_2O	Yield	N_2O		ANPP	N_2O
	Site	[-2; 2]	[-1; 1]	[-2; 2]	[-1; 1]	[-2; 2]	[-1; 1]	Site	[-2; 2]	[-1; 1]
E-median	C1							G1		
	C2							G2		
	C3							G3		
	C4							G4		
	C5							1-var perf.	0/4	3/4
	1-var perf.	3/5	3/5	2/4	3/3	0/1	0/1	2-var perf.	0/4	
	2-var perf.	2/5		1/3		0/1				
3-median	C1							G1		
	C2							G2		
	C3							G3		
	C4							G4		
	C5							1-var perf.	2/4	3/3
	1-var perf.	2/5	4/5	2/4	3/3	1/1	1/1	2-var perf.	1/3	
	2-var perf.	2/5		1/3		1/1				
Top models		M13, M20, M09		M09, M25, M13		M09, M13, M26		Top models	M05, M24, M03	

Figure captions

Figure 1 Location of the experimental sites (a) and their distribution depending on annual precipitation and temperature (b) of cropping (circles) and grasslands locations (triangles). C refers to cropping systems - C1: Canada, C2: France, C3: India, C4: Australia, C5: Brazil; G

refers to grassland systems; G refers to grasslands - G1: USA, G2: New-Zealand, G3: France, G4: UK. Source of the background map: J. Foley, University of Minnesota; http://www.nationalgeographic.com/foodfeatures/feeding-9-billion/

Figure 2 Chart of the five-stage protocol adopted for model comparison and benchmarking. A detailed list of input data provided to models according to modelling stages is shown in Supplementary Material (Table S3).

Figure 3 Relative average prediction errors of E-medians for yields (a) and for soil N_2O emissions (b) from modelling stages 1 to 5. Data are Relative Root Mean Square Error of E-medians ($RRMSE_{E\text{-}median}$) ± s.e. (standard error based on individual models *RRMSE*) for wheat, maize, rice and for grasslands.

Figure 4 E-median relative deviation to observed means of yields (a to d) and of N_2O emissions (e to h) for wheat, maize, rice and grassland sites over modelling stages 1 to 5. The shaded area shows the range within two standard deviations (2 s.d.) of the experimental mean for grain yield and grassland ANPP, and within one standard deviation (1 s.d.) of the experimental mean for N_2O emissions.

Figure 5 Percentage of single models providing plausible estimates for yields (grain yields at crop sites and ANPP at grassland sites) (a), N_2O emissions (b) and for both variables combined (c), over modelling stages 1 to 5. Model estimates were considered plausible when within two and one s.d. of the observed mean for yields and N_2O emissions, respectively.