How modellers model: the overlooked social and human dimensions in model intercomparison studies

Fabrizio Albanito[1], David McBey[1], Matthew Tom Harrison[2], Pete Smith[1], Fiona Ehrhardt[3,4], Arti Bhatia[5], Gianni Bellocchi[6], Lorenzo Brilli[7], Marco Carozzi[8], Karen Christie[2], Jordi Doltra[9], Christopher D. Dorich[10], Luca Doro[11,12], Peter Grace[13], Brian Grant[14], Joël Léonard[15], Mark Liebig[16], Cameron Ludemann[17], Raphael Martin[6], Elizabeth Meier[18], Rachelle Meyer[19], Massimiliano De Antoni Migliorati[13, 20], Vasileios Myrgiotis[21], Sylvie Recous[22], Renáta Sándor[23], Val Snow[24], Jean-François Soussana[3], Ward N. Smith[14] and Nuala Fitton[1,25].

Affiliations:

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

[2]Tasmanian Institute of Agriculture, University of Tasmania. 16-20 Mooreville Rd, Burnie, TAS, 7320 Australia

[3]INRAE, CODIR, Paris, France

[4]RITTMO AgroEnvironnement, Colmar, France

[5]ICAR-Indian Agricultural Research Institute, New Delhi, India

[6]Université Clermont Auvergne, INRAE, VetAgro Sup, UREP, 63000 Clermont-Ferrand, France

[7]CNR-IBE, National Research Council Institute for the BioEconomy, Via Caproni 8, 50145, Florence, Italy

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

[9]Institute of Agrifood Research and Technology, IRTA Mas Badia, 17134 La Tallada d'Empordà, Girona, Spain

[10]Natural Resource Ecology Lab, Colorado State University, Fort Collins, CO, 80521 USA

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

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

[13]Queensland University of Technology, Brisbane, Australia

[14]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON, Canada

[15]BioEcoAgro Joint Research Unit, INRAE, Barenton-Bugny, France

[16]USDA-ARS Northern Great Plains Research Laboratory, Mandan, ND USA

[17]Cameron Ludemann Consulting, Arnhem, The Netherlands.

[18]CSIRO Agriculture and Food, St Lucia, Qld, Australia

[19]University of Melbourne, Faculty of Veterinary & Agricultural Sciences, Parkville, VIC 3010, Australia

[20]Department of Environment and Science, Brisbane, Australia

[21]School of GeoSciences, University of Edinburgh, Edinburgh EH9 3JN, UK

[22]Université de Reims Champagne-Ardenne, INRAE, FARE laboratory, Reims, France

[23]Agricultural Institute, Centre for Agricultural Research, ELKH, Martonvásár, Hungary

[24]AgResearch, PB 4749, Christchurch 8140, New Zealand

[25] South Pole, Technoparkstr. 1, Zurich, ZH 8005, CH

Abstract

There is a growing realisation that the complexity of model ensemble studies depends not only the models used, but also on the experience and approach used by modellers to calibrate and validate results, which remain a source of uncertainty. Here, we applied a multi-criteria decision-making method to investigate the rationale applied by modellers in a model ensemble study where twelve process-based different biogeochemical model types were compared across five successive calibration stages. The modellers shared a common level of agreement about the importance of the variables used to initialise their models for calibration. However, we found inconsistency among modellers when judging the importance of input variables across the different calibration stages. The level of subjective weighting attributed by modellers to calibration data decreased sequentially as the extent and number of variables

provided increased. In this context, the perceived importance attributed to variables such as fertilisation rate, irrigation regime, soil texture, pH, and initial levels of soil organic carbon and nitrogen stocks were statistically different when classified according to model types. The importance attributed to input variables such as experimental duration, gross primary production, net ecosystem exchange, varied significantly according to the length of the modeller's experience. We argue that the gradual access to input data across the five calibration stages negatively influenced the consistency of the interpretations made by the modellers, with cognitive bias in "trial-and-error" calibration routines. Our study highlights that overlooked human and social attributes is critical in the outcomes of modelling and model intercomparison studies. While complexity of the processes captured in the model algorithms and parameterisation are important, we contend that (1) the modeller's assumptions on the extent to which parameters should be altered, and (2) modeller perceptions of the importance of model parameters, are just as critical in obtaining a quality model calibration as numerical or analytical details.

Keywords: Model ensembles, biogeochemical models, multi-criteria decision-making, modelcalibration,model intercomparison, climate change, greenhouse gases, soil carbon., AgMIP

Synopsis: This study outlines subjective inconsistencies in the prioritization of variables used in model calibration, with implication in the outcomes of modelling and model intercomparison studies.

Introduction

Multi-model ensemble comparisons are becoming increasingly common in contemporary research using agricultural simulation models to understand the impacts of weather variability[1], climate change[2], greenhouse gas (GHG) emissions from agriculture[3,4] and carbon stock[5,6] and the development of mitigation options.[7,8] Ensemble modelling has long been used by climate modellers to overcome uncertainty in understanding processes, but it is a relatively new concept in the domain of agricultural systems modelling.[9] Running multiple biogeochemical models and model versions, in combination with different sets of site conditions, helps to distil uncertainty derived from individual model simulations.[2] It is generally accepted by the modelling community that - provided models are diverse and independent – the prediction error decreases when using the ensemble approach.[10] A number of questions, however,

continue to prompt discussion and debate of what model ensemble studies tell us about the uncertainty surrounding the impact of the future climate on agriculture, and the effectiveness of climate mitigation strategies in agriculture under different emission scenarios.[3,11,12] As well, the use of multiple models generally increases the range of results, increases the workload, and requires more diverse skillsets to be successful.[13,14] The answers to these questions are relevant beyond the bound of agricultural science, as climate mitigation and adaptation decisions may be influenced by what is learned from multi-model ensemble studies.

Terrestrial biogeochemical and eco-physiological models typically comprise sets of mathematical equations simulating a continuum of interlinked atmosphere-plant-soil processes (e.g. plant photosynthesis, organic matter decomposition, ammonia volatilisation, nitrification and denitrification), enabling the simulation of spatial-temporal patterns of carbon (C) and nitrogen (N) cycles in crop and grassland systems, and subsequent responses of GHG emissions to agricultural practices.[3,15,16,17] As a result of their fixed, semi-empirical and nonlinear model structure, biogeochemical models were often described as black-box models.[18,19] They often have many parameters (e.g. 100-1000) that have no intuitive meaning[20,21] and/or cannot be measured and must be inferred from the data. Consequently, one of the main challenges in biogeochemical modelling is that bulk observations of C and N cycling or GHG emissions rarely contain sufficient information to reliably estimate model parameters.[12]

Agricultural model intercomparison studies are becoming increasingly common. To date, a number of studies have discussed the complexity and limitations characterising agro-ecosystems from multi-model ensemble studies.[3,22,23,24,25] In model ensemble studies, there is not only uncertainty about the structural limitations of the model from which the contribution of agricultural systems should be generated.[26] There is also uncertainty about how the initial conditions (i.e. input data) in the model simulations should be interpreted[28]; uncertainty in model internal coefficients that cannot be altered by the users; and further uncertainty concerning which processes are included in the model by the developer.[20,21] This gives rise to a branch of studies examining automatic multi-objective parameterisation of several model parameters simultaneously.[13] Ensemble studies include and compare results from models that have varying development histories, funding support, as well as varying priorities of developers, including their perceived importance of processes and parameters. Depending on the intent to which a model was

built, some models include representations of agricultural processes that other models do not, include and based on model structure, each model may require different input data and calibration strategies. Accordingly, there may be substantial variability between model outputs when different modellers are using the same calibration data, even when all are using the same model and version.[3,27,28]

There is a growing realisation that the complexity of model ensemble studies arises not only due to the models used, but that the human dimension also has a prominent role to play, considering the experience, perceptions, expectations and approaches brought forth by modellers to calibrate parameters and validate results. The human dimension remains a key but often recalcitrant source of uncertainty.[23] In this context, there is a little information on the social and psychological aspects of model calibration or intercomparison, including how parameters are chosen for calibration, how parameters are calibrated or weighted against available data, and how model are technically verified and outputs validated against observed data.[29] To address this gap, we surveyed and interviewed several modellers who contributed to a model ensemble study that aimed to simulate productivity and nitrous oxide (N_2O) emissions from cropland and grassland sites spanning four continents.[3] These modellers varied in nationality, experience, gender, and discipline, giving us an ideal cross-section of geographical and disciplinary expertise. We analysed the rationale used by these modellers in a multi-stage model ensemble study where different model types were compared across five successive stages (i.e. from blind parameterisation to partial and full calibration) to benchmark their performance in relation to the input data provided at each stage.[3] The objectives are to describe: (i) the heterogeneity in modellers' prioritisation of different variables in modelling decision contexts, (ii) the perceived importance of the variables across the five stages of the modelling protocol, (iii) the perceived variable structure and interrelationships, and (iv) a process through which surveys of modellers' insights can be used to improve model intercomparison guidelines.

Materials and Methods

The model ensemble study described in Ehrhardt et al.[3] was based on the contribution of 24 modellers from 11 countries, reporting the results of 24 process-based integrated C-N models by comparing multi-year (1-11 years) simulations with experimental data from nine sites (four temperate permanent

grassland sites and five arable crop rotations with wheat, maize, rice and other crops). Following the multi-stage modelling protocol of Ehrhardt et al.[3], here we implemented a Multi-Criteria Decision Making (MCDM) method that collected and analysed information on the modelling experience, priorities, and decisions made by the modellers who contributed to the model ensemble study.

Multi-stage modelling protocol

The model ensemble protocol described in Ehrhardt et al.[3] included 55 input variables clustered into seven categories that were released to the modellers in successive stages (Figure 1). In Stage 1, input data used for initial model testing included information on experimental farm site conditions (such as general site information (SI), climate during the experiment (CL), management practices during the experiment (MPDE) and soil information (SOI). Stage 2 provided long-term (i.e. historical) site-specific data on climate (LTCL) and management practices (LTMP) for the long-term model calibration period.[3] Stage 3 provided part of the experimental data from site (EDS) describing plant phenology, crop/grassland vegetation development (e.g. leaf area index), and grain yields or monthly grassland offtake (biomass removed by haying or animal intake determined monthly). In Stage 4, modellers accessed additional EDS data on the dynamic trends of soil temperature, moisture, and mineral N during the experiment. Finally, Stage 5 included the remaining EDS information against which model outputs were compared, such as agricultural productivity (ANPP together with daily changes in live weights of livestock and daily grassland offtake), GHG emissions and soil organic C (SOC) stock changes. In the five modelling stages, modellers were free to choose a calibration procedure of their choice based on their own subjective knowledge, the model type used and the agricultural system targeted.

Framework of the survey

This study was introduced during a meeting of the Global Research Alliance on Agricultural Greenhouse Gases, hosted by former INRA (currently INRAe) in Paris (France), on 13-15 December 2017. In this workshop, the modellers discussed the objectives of the survey in relation to the work performed in previous multi-stage model ensemble studies. Following this meeting, the modellers were invited to participate in the survey, which included a consent form and a background questionnaire to be completed prior to receiving the questionnaire (see S.1 and S.2 in the Supplementary Information).

In particular, the background questionnaire collected general information such as gender, education level, academic rank, modelling experience, location, institution, general features of the model/model version used and the calibration method adopted.

A second invitation was sent to the modellers who agreed to participate in the survey, which included a participant instruction document explaining the methodology used in the survey, a demonstration video accompanied by a video help script describing how to complete the pairwise questionnaire (see S.3 and S.4 in the Supplementary Information). The pairwise questionnaire included a number of pairwise comparison matrices (PCMs) grouped by variable categories, where the modellers assessed the relative importance and influence (i.e. relationship) that each input variable had against each other. In particular, we asked the modellers to use pre-defined rating scales to rank the data based on the steps followed during the stages of the model inter-comparison study (see S.5 in Supplementary Information).

After completing of the pairwise questionnaire, the participants received a third invitation for an interview. The interviews were conducted by telephone or videoconference and were 'semi-structured' into a list of open-ended questions (see S.6 in the Supplementary Information) that allowed participants to fully express their opinions on the questionnaire.[30] Broad topics discussed with each participant included (1) feedback on the study, (2) problems encountered during the pairwise process, and (3) discussion of the pairwise results with the possibility to change any response.

Multi-Criteria Decision-Making questionnaire

The 12 model types used in the ensemble study encompassed biogeochemical processes (e.g. plant growth, organic matter decomposition, atmospheric processes, ammonia volatilisation, nitrification, denitrification and other carbon and nitrogen processes) designed to interact with each other to describe the cycling of water, C and N for the target ecosystems.[26] As such, across the five modelling stages, each modeller subjectively decided how to select and prioritise the parameters that should be calibrated using the input data provided, and how their model outputs should be validated against specific observed data. In particular, each modeller selected the parameters that they deemed to be the most important in contributing to high model performance (i.e., the quality of fit of several output variables to the provided data). To deal with the complexity, we applied an MCDM process (Figure 2) that combined the Decision

Making Trial and Evaluation Laboratory (DEMATEL)[31] and the Analytic Network Process (ANP) method.[32] Using DEMATEL, we visualised the complex interrelationships between the different variable categories, outlining the degree of influence imparted by each category, as envisaged by the modellers. In ANP, the strength of relationships outlined in DEMATEL were integrated into a network of dependencies and feedbacks to determine the relative importance of each input variable across the five stages of the modelling protocol (see S.7 in Supplementary Information).

Data analysis

To assess the level of agreement between the modellers, Kendall's concordance coefficient (K_W)[33] was applied to the importance scores for the variable categories and input variables included in the pairwise questionnaires (Eq. 1):

$$K_w = \frac{12\ SS}{m^2(n^3-n)-mF} \quad (1)$$

where *SS* is the sum-of-squares from sums of rank scores a_{ij} (see Eq. 9 in S.7 of the Supplementary Information), *n* is the number of elements in the *PCMs*, *m* is the number of modellers that participated to the survey, and *F* is a correction factor for tied ranks.[34] The null hypothesis of K_w is that the modellers provided independent ranking scores for each input variable and category (i.e. the modellers were not in agreement with each other). Perfect agreement is indicated by K_w values of 1, while no agreement is indicated by values of 0. When the null hypothesis was rejected, we tested significant effects ($p < 0.05$) against the null hypothesis that there is no agreement between the modellers.

A one-way multivariate analysis of variance was applied using SPSS statistical software (IBM SPSS v.25) to determine whether there were differences in the ratings (i.e. dependent variables) given by the modellers in the pairwise questionnaires, based on the 12 model types used and their modelling experience ranging from <5 to >20 years. Wilks' lambda test was utilised to determine whether there were significant differences ($p < 0.05$) between the mean scores of the modellers across the combination of dependent variables.

Data analysis included the correlation between the MCDM results (i.e. modelling priorities) and the ensemble modelling prediction errors described in Ehrhardt et al.[3] Model prediction error, in particular, was represented by the root mean square error normalised by the mean of the observed data (*RRMSE*) of the individual models across the five stages for simulations of N_2O emissions from arable and grassland systems, maize, wheat and rice crop yields, and ANPP in grasslands.[3]

The relationship between *RRMSE* and modelling priorities across stages was investigated as:

$$MER = \frac{RRMSE}{\sum P_i} \quad (2)$$

where, $\sum P_i$ represents the cumulative modelling importance of the input variable (see Eq. 10 in S.7 of the Supplementary Information) across the five stages of the model ensemble protocol, and *MER* is the model error rate corresponding to the change in *RRMSE* per unit of importance given to the input variable accessed across the five stages.

RESULTS AND DISCUSSION

Characteristics of participating modellers

Table 1 shows an overview of the information gathered in the background questionnaire and during the interviews with the modellers who participated in the survey. Overall, the 20 modellers that participated in the study were aged between 25 to 64 years, the majority were male (54%), 68% held a PhD degree, 58% were employed under fixed-term contracts, and 84% had > 5 years modelling experience. Modellers within the 35-44 and 45-54 age category generally used, and had published, information from a larger number of models (Table 1). The 20 modellers interviewed used 12 different models types:

i) APSIM (The Agricultural Production Systems sIMulator)[35] (Holzworth, 2014)
ii) CERES-EGC (Crop Environment REsource Synthesis - Environnement et Grandes Cultures)[36]
iii) DayCent and Daily DayCent[37]
iv) DNDC (DeNitrification-DeComposition)[38,39]
v) Landscape DNDC[40]
vi) DSSAT (Decision Support System For Agro-technology Transfer)[41,42,43]

vii) EPIC (Environmental Policy Integrated Climate)[44]

viii) PaSim (Pasture Simulation model)[45]

ix) DairyMod/SGS[46]

x) FASSET[47]

xi) STICS[48] (Brisson et al, 1998)

xii) INFOCROP[49] (Aggarwal et al., 2006)

Further details are provided in the Supplementary Information of Ehrhardt et al.[3], Appendix S1.

Modellers' prioritisation and uncertainties in the variables provided

During the interviews, the modellers discussed their systematic approach across the five stages of the modelling protocol, as well as the uncertainties they encountered when answering the pairwise questionnaire. Here we summarise and explain some of the uncertainties discussed with the modellers in relation to the modelling decision contexts.

In the model ensemble study, the modellers were given a set of choices about how many parameters should be calibrated against the available input data, and how the models should be evaluated when the model outputs are validated against the observed data. Based on the information gathered from the interviews, in the first two stages of the modelling protocol the modellers based their model calibration on their own experience and knowledge of the expected outcomes. In the last three stages, most modellers adopted the "trial-and-error" calibration routine, with only one modeller consistently applying Bayesian calibration. It is plausible that the gradual access to input data across the five stages negatively influenced the logic applied by the modellers in the calibration and validation processes, employing inconsistent modelling decisions between each stage (i.e. cognitive biases[50]).

The results of the pairwise questionnaires confirmed that all modellers showed some level of inconsistency in judging the relative importance of the input variables. The consistency of the modeller's judgements was assessed though the consistency ratio (CR), which outlines the degree of bias in the pairwise judgments related to the rank order and mutual preference of alternative input data within each input category (Table 2). In this context, the responses from one modeller were excluded

from the analysis due to high inconsistency (CR >30%), above the 10% cut-off threshold. The remaining 19 modellers completed the questionnaire with a consistency ratio of 7±1% (mean±standanrd deviation). Where the CR was above 10%, an in-person review was undertaken with the modellers to address the source of inconsistencies and find possible corrections. CR was above 10% for 37% of the modellers when ranking the variables in SOI, 21% for the scores given to EDS, 11% for the variables listed in MPDE and LTMP, and 5% when ranking the variables in SI and LTCL. Behavioural science could help to further address these findings. The pairwise judgements expressed by the modellers may have been affected by systematic biases in judgements, which reduced the complex tasks of determining the importance and influence of several input variables within each categories to simpler judgmental operations related to the modelling approach. Some of these biases may be mediated by "heuristics principles" in judgements under uncertainties, overconfidence, neglect of base-rate information, and overestimates of frequency of events that are easy to recall [51].

Importance of (and interactions between) different calibration variables perceived by modellers

The use of DEMATEL and ANP allowed visualization of the perceived importance and the relationship between the input data across the five stages of the modelling protocol. Overall, in the ensemble study Stage 1 included more than 50% of the input variables used in the simulations (i.e. 28 input variables) (Figure 1), and accounted for 67% of importance in the model ensemble framework (Table 3). In contrast, the cumulative importance of the inputs released in Stage 2 was 11%, 6% for Stage 3, 5% for Stage 4, and 11% for Stage 5. We found a common agreement between modellers about the importance of the data used in Stage 1 to initialise the models for calibration, which comprised data included in the categories SI, CL, SOI and MPDE (Table 2). The high importance of MPDE may reflect the fact that the models involved in the ensemble study required information about farming practices such as harvesting, mowing, fertilisation, tillage and irrigation.[26] Whereas the low level of agreement for the priority attributed to MPDE may reflect differences in the simulations of cropland and grassland systems, as well as model characteristics, rather than disagreement between modellers on the relative importance of the input variables in MPDE. However, the importance of input variables such as fertilisation rate, irrigation regime, soil texture, field capacity and/or water-filled pore space, pH, SOC

and soil organic nitrogen (SON) stocks, and atmospheric CO_2 concentration were statistically different when classified according to model types (Table 2).

The input data given in Stage 1 in the categories CL, LTCL and SI were considered net influencers in the modelling protocol (Figure 3). This means that 60% of the relationship within the climate variables (CL and LTCL) was directed towards other input variables (i.e. a positive relationship). In contrast, the categories EDS, MPDE, LTMP and SOI, which spread the data across the five modelling stages, were considered net receivers, with >50% of their relationship based on the influence received from other variable categories (i.e. a negative relationship). In particular, the category EDS used in Stage 3, 4 and 5 (Table 3), included important in-season and end-of-season experimental data used to validate model outputs, such as site-specific experimental data on crop phenology, grassland offtake, dynamic soil processes, crop yields, ANPP, GHG emissions and SOC stock changes. The low level of agreement between the modellers about the priorities given to EDS may reflect the heterogeneity in modellers' knowledge on the use of experimental data for model calibration. In the model intercomparison study, the models APSIM, DairyMod, and DayCent were used by more than one modeller or modelling team. For these model types, the opinion about variables included in the categories MPDE, SOI, and EDS was characterised by low level of agreements between modellers. The modellers that used APSIM, and DairyMod, in particular, prioritised information on yield and dynamic vegetation. While, for the modellers that used DayCent, the importance of EDS was focussed on parameters related to the components of the ecosystem GHG budget (such as N_2O and CH_4 emissions) or gross primary production (GPP), net ecosystem production (NEP), net ecosystem exchange (NEE), and ecosystem respiration (Reco) (see Table S.8 in Supplementary Information).

Overall, the importance given to input variables such as experimental duration, GPP, NEP, NEE, Reco and soil temperature was statistically different among modellers with different experience (Table 2). This is an important result, as the trial-and-error manual calibration routines applied in the final stage of the modelling protocol depend not only on users' knowledge and expertise of the model structure, but also on their understanding of the variables measured in the targeted agroecosystems.[52] The analysis of the influence given and received between the variables showed contradictory results for EDS, which had a negligible influence on the value of variables included in CL, LTCL, MPDE and SI (Figure 3).

The SI category, in particular, was perceived as a net influencer, and included a relatively high incoming influence in the system. Further investigation would be needed to understand whether these results are due to biases related to: (i) specific features of the model structure, (ii) physical or biogeochemical processes characterising agricultural systems, (iii) the complexity of the multi-stage modelling protocol in answering the pairwise questionnaires, or iv) the uncertainty and variability implicit to the measured input data. In addition to the MCDM analysis, we used qualitative interviews to better understand how modellers' attitudes (e.g., best practices), the influence of outside actors (e.g., fellow researchers, literature), and other factors (e.g., data quality, time constraints) impact their approach to modelling (manuscript in preparation).

Relationship between modelling decisions and uncertainty of the ensemble outcomes.

Overall the patterns of uncertainty between single models and model ensemble simulations suggest that the modeller's choices were governed by general rational rules. However, across the five modelling stages modellers may have come across significant challenge, particularly when the same numerical result could be arrived at in multiple ways (ie. the right answer for the wrong reasons). In the context of decision-making, the modeller's decision could have be restricted by "narrow framing"[53], limited "accessibility" which is a technical term for the ease with which mental contents come to mind[54], and "decision bracketing"[55]. The choices that the modellers faced arose one at a time, and the problems were considered as they arose. This means that in each modelling stage the problem at hand and the immediate consequences of the choices made were far more accessible than all other considerations, and as a result the overall modelling problem was framed far more narrowly than the rational modelling assumes. In that respect, we found that the gradual access to additional input data across the five stages did not show a clear benefit in reducing the model ensemble uncertainty (Figure 4). Across the five stages, the mean *RRMSE* of the model simulations was 99% for N_2O emission, 81% for ANPP and 31% for crop yield (Figure 4). It is plausible that the gap between high model complexity and limited data availability in the initial stages of modelling generated uncertainties related to parameter equifinality or non-identifiability and ill-defined problems.[12,13,56,57,58] In particular, equifinality or non-identifiability arises when different combinations of parameter values give the same results. Such results have been shown to be sensitive to the inclusion of extreme events, such as very wet and dry seasons, in the

calibration.[59] Ill-posed problems occur when the number of parameters to be optimised is greater than the boundary conditions and the number of measured data points used in the model calibration.[13,20,21]

The number of input data and their perceived importance was clustered in the first two stages of the modelling study (Table 3). This limited the possibility to extract detailed information about the incremental effect of the different variable categories on the ensemble simulations. The change in model prediction errors per unit of dataset importance given by the modellers (MER) showed that in the crop productivity simulation, the input variables used in the first two stages (i.e. 78% of overall dataset importance) were sufficient to calibrate the models and obtain plausible results. The ensemble simulations of N_2O emissions and ANPP, however, showed that only after receiving approximately 90% of all input data of the modelling protocol, the modellers were able to achieve the highest accuracy of the ensemble simulations. In particular, the use of historical data on climate and management practices in Stage 2 reduced the MER by 25% for the ensemble prediction of N_2O emissions in Stage 1. However, in Stage 3 the additional access of experimental information on vegetation data such as LAI, plant phenology and extracted yields (i.e. 6% of the relative modelling importance) increased the MER for N_2O emission simulation by 18%. Only with access to additional experimental data in Stage 4 (dynamic measurement of soil moisture, temperature and mineral N) did the simulation of N_2O emissions improve, with a mean reduction in MER of 50% compared to Stage 1. The ANPP predictions showed a similar trend in MEP as the N_2O emissions. In this case, however, the ANPP predictions of ANPP benefitted only marginally from access to site-specific experimental data in Stages 3, 4 and 5 (Figure 4).

The development of generic guidelines including information about how to characterise the data required for agroecosystem modelling, with complementary and clear protocols for estimating model parameters and validation model results, remains a major challenge of agroecosystem model studies. Here, we used a multi-model ensemble study to highlight the psychology of modellers in ranking and interpreting the variables used in the simulations.

Two major conclusions can be drawn from our analysis. First, modellers perceive variables such as general site information, climate conditions and management practices as being of vital importance for

modelling cropland and grassland systems. The perceived importance of these variables was related to the calibration of processes in the first two stages of the modelling protocol, requiring information such as precipitation, air temperature, crop yield, fertilisation rate, irrigation regime, soil texture, field capacity and water-filled pore space. However, these input variables were not sufficient to obtain satisfactory ensemble simulations of crop production and GHG emissions. In this respect, the intercomparison study here showed that the crop yield simulations achieved plausible results after accessing the crop phenology and yield values, which corresponded to 84% of the variables given in the whole modelling protocol. These findings agree with[23], who identified minimum input data requirements for crop model intercomparisons including weather, soil and crop management data, as well as some site-specific measurements of crop responses to test a given comparison.

Second, the framework for multi-model intercomparison studies needs to pay more attention to the structure of the models, the understanding of the interrelationships between the different processes and the experience of the modellers. The models used in the ensemble study included numerous biogeochemical processes (e.g. plant growth, organic matter decomposition, atmospheric processes, ammonia volatilisation, nitrification and denitrification) designed to interact with each other to describe the water, C and N cycles for the target ecosystems.[28] In this context, we visualised the relationship between the different variables used in a multi-stage modelling protocol, partitioning these between into the categories of net influencers and net receivers. Although general site information and climate data only represent 30% of the input data used in the ensemble protocol, the modellers' opinions on the importance and level of influence of these variables, used to initialise the model calibrations, depended by the model type used. In addition, the ensemble simulations of N_2O emissions and grassland above ground biomass required more than 90% of the input data used in the modelling protocol (i.e. four out of five stages) to obtain plausible results. In this context, Ehrhardt et al.[3] outlined several limitations in the calibration methods and model structure that could explain the discrepancies between simulated and observed data. The opinion of the modellers, however, was that fundamental parameters such as crop management, soil characteristics and experimental data from sites were net receivers in the framework of the modelling protocol. Importantly, the ranking of the most important input data, such as experimental length and season, irrigation, SOC stock, soil temperature, GPP, NEP, NEE and Reco,

varied according to the experience of the modellers. We argue that it is likely that, among the limitations explaining the uncertainty of the ensemble study, the interpretation made in the "trial-and-error" calibration routines, and the structure of the modelling protocol itself, also lead to uncertainty in the simulations. We argue that it is likely that, among the limitations explaining the uncertainty of the ensemble study, the interpretation made in the "trial-and-error" calibration routines, and the structure of the modelling protocol itself, also lead to uncertainty in the simulations. What is natural and intuitive in a given modelling situation is not the same for everyone: different experiences favour different modelling intuitions about the meaning of input variables, and modelling behaviours become intuitive as skills are acquired[51]. In the Ehrhardt, et al.[3] study only one modelling team used the automatic calibration method. It is plausible that in automatic calibration methods, the selection of parametrisation algorithm or software is one such human decision factor among many that could have a large bearing on the validity of calibration and consequential model performance. Thus the experience and skills of the modellers again influences model outputs via their initial capability, knowledge and confidence in using a given approach for calibration.

Moving forward, ensemble studies should include in their guidelines an understanding of how data interpretations and model structures influence the calibration and validation strategies and collect information on this. This study would have been particularly helpful if it had been carried out before and during the model ensemble study, as the information obtained could have contributed to the guidelines for the ensemble study. The structure of the multi-stage benchmarking protocol was a major limitation of our analysis. Firstly, the model intercomparison study involved 20 modellers that used 12 distinct model types. This means that in our study only for three model types we had the possibility to sample more than modeller. Secondly, the first two stages of the protocol comprised the majority of the input data used by the modellers, corresponding to 78% of the variables considered by the modellers to be the most important. In this context, a release of data across the stages in line with modelling priorities and model structure could have helped to organize the five stages of the ensemble study to understand the relative contribution between data interpretation, model calibration methods, model structures and site-specific variability of observations to the uncertainty of the ensemble simulation.

AUTHOR INFORMATION

Corresponding Authors

*Fabrizio Albanito - Institute of Biological and Environmental Science, School of Biological Science, University of Aberdeen, 23 St Machar Drive, Aberdeen, AB24 3UU, UK. E-mail: f.albanif@abdn.ac.uk

*David McBey - Institute of Biological and Environmental Science, School of Biological Science, University of Aberdeen, 23 St Machar Drive, Aberdeen, AB24 3UU, UK. E-mail: d.mcbey@abdn.ac.uk

Author Contributions

F.A., N.F. and D.M. designed the research; F.A. performed the data analysis; F.A. wrote the manuscript with contributions from MTH, P.S., N.F., D.M.; V.S, G.B., and M.L reviewed it; P.S., P.G., M.L., S.R., V.S., and J-F.S. served on the steering committee of the C and N Models Inter-comparison and Improvement project (CN MIP); N.F. and F.E. coordinated the modellers' survey; A.B., L.B., M.C., K.C., J.D., C.D.D., L.D., B.G., M.T.H., J.L., M.L., C.L., R.M., E.M., R.M., M.D.A.M., V.M., V.S., W.S. and N.F. participated to the multi-criteria decision survey; G.B. and R.S. were members of the consortium CN-MIP. All authors commented on manuscript drafts.

ACKNOWLEDGMENT

Fabrizio Albanito gratefully acknowledges funding from RETINA project (NERC, NE/V003259/1) and the FACCE-JPI projects: CN-MIP, Models4Farmers and MACSUR. 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. The study falls within the thematic area of the French government IDEX-ISITE initiative (reference: 16-IDEX-0001; project CAP 20-25)

Figure captions:

Figure 1: Framework of the variables, and partition between input categories and variables used in the five stages of the model ensemble protocol described in Ehrhardt et al. (2018).

Figure 2: Steps of the Multi-criteria Decision Method process combining the DEMATEL (Decision Making Trial and Evaluation Laboratory) and ANP (Analytic Network Process) methods. Through DEMATEL we visualise the perceived relationship between different categories of variables. While in ANP, the strength of the relationships outlined in DEMATEL is integrated into a network of dependencies and feedbacks among input variables to determine their relative importance across the five stage of the modelling protocol.

Figure 3: The table reports the DEMATEL total relation matrix summarising the level (mean±standard deviation) of direct and indirect influence given (G) and received (R) in each input category, the net influence (G-R), and the total level of influence (or dominance) (G+R) of the model input category used in the model ensemble study. Categories with a positive G-R have a net influence towards the value of other variable categories and are denoted as "influential" categories. The circular diagram outlines the causal relationship in the model ensemble protocol between General site information (SI), Climate during the experiment (CL), Long-term climate (LTCL), Management practices during the experiment (MPDE), long-term management practices (LTMP), Environmental data from site (EDS), and Soil information (SOI). The arrows in the diagram show the direction and level of influence that each input category gives and receives from the other categories. The coloured arrows highlight the three categories of variable that resulted to be net influencers in the model ensemble protocol (i.e. positive G-R). Radial bar numbers represent the total level of R+C influence, and the relative percentage of the casual relationship within each input category.

Figure 4: The table summarises the Relative Root Mean Square Error (RRMSE), averaged across 19 models, for the ensemble simulations of soil N_2O emissions from arable and grassland systems, crop yields of annual crop monocultures such as maize, wheat and rice, and above-ground net primary productivity in grassland (ANPP). Pi corresponds to the cumulative modelling importance of the input variables accessed in the five stages of the model ensemble framework. MER represents the model simulation error rate for N_2O, yield and ANPP per unit of modelling importance in each stage. The bar chart below the table outlines the trend of MER across the five stages of the model ensemble protocol.

Table 1: Background information reported by age class (AC) of modellers who participated in the multistage intercomparison protocol and the MCDM survey. N = number of modellers, F = proportion of modellers identified as female, PhD = proportion of modellers holding a PhD degree, FTC = proportion of modellers with a fix-term contract, MU = knowledge on number of models, MP = number of models published in peer-reviewed articles and MT = type of model used.

Table 2: Summary of the importance of the input variables and their categories, the consistency ratio of the modellers' judgments in the pairwise comparison matrix of each input category, and the level of Kendal concordance between the modellers. Ranking scores (i.e., importance) with the letters *m* and *e*

are significantly different between model groups and modeller experience groups, respectively. * indicates the level of significant concordance within each category of variables ($p < 0.05$).

Table 3: Cumulative importance of the five stages of the model ensemble protocol. In each stage, the ranking of the input variables shown in Table 2 was normalised to the importance score of their corresponding categories.

References

(1) Ruane, A.C.; Hudson, N.I.; Asseng, S. et al. Multi-wheat-model ensemble responses to interannual climate variability. *Environmental Modelling and Software.* **2016**, 81, 86-101.

(2) Asseng, S.; Ewert, F.; Rosenzweig, C.; Jones, J.W.; Hatfield, J.L.; Ruane, A.C.; Boote, K.J.; Thorburn, P.J.; Rotter, R.P.; Cammarano, D.; Brisson, N.; Basso, B.; Martre, P.; Aggarwal, P.K.; Angulo, C.; Bertuzzi, P.; Biernath, C.; Challinor, A.J.; Doltra, J.; Gayler, S.; Goldberg, R.; Grant, R.; Heng, L.; Hooker, J.; Hunt, L.A.; Ingwersen, J.; Izaurralde, R.C.; Kersebaum, K.C.; Muller, C.; Naresh Kumar, S.; Nendel, C.; O'Leary, G.; Olesen, J.E.; Osborne, T.M.; Palosuo, T.; Priesack, E.; Ripoche, D.; Semenov, M.A.; Shcherbak, I.; Steduto, P.; Stockle, C.; Stratonovitch, P.; Streck, T.; Supit, I.; Tao, F.; Travasso, M.; Waha, K.; Wallach, D.; White, J.W.; Williams, J.R. and Wolf, J. Uncertainty in simulating wheat yields under climate change. *Nature Climate Change.* **2013**, 3, 827-832.

(3) Ehrhardt, F.; Soussana, J. F.; Bellocchi, G.; Grace, P. et al. Assessing uncertainties in crop and pasture ensemble model simulations of productivity and N2O emissions. *Global Change Biology*. **2018**, 24 (2), e603-e616.

(4) Sándor, R.; Ehrhardt, F.; Grace, P. et al.. Ensemble modelling of carbon fluxes in grasslands and croplands. *Field Crops Research.* **2020**, 252, 107791.

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

(6) Farina, R.; Sándor, R.; Abdalla, M. et al. Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils. *Global Change Biology*. **2021**, 27, 904-928.

(7) Fuchs, K. et al. Multimodel evaluation of nitrous oxide emissions from an intensively managed grassland. *Journal of Geophysical Research: Biogeosciences*. **2020**, 125, 1. e2019JG005261

(8) Sándor, R.; Ehrhardt, F.; Brilli, L.; et al. The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed grasslands. *Science of the Total Environment*. **2018**, 15, 292-306.

(9) Holzworth, D.P.; Snow, V.O.; Janssen, S.; Athanasiadis, I.N.; Donatelli, M.; Hoogenboom, G.; White, J.W. and Thorburn, P.J. Agricultural production systems modelling and software: Current status and future prospects. *Environmental Modelling & Software*. **2015**. 72, 276–286.

(10) Martre, P.; Wallach, D.;, Asseng, S.; Ewert, F.; Jones, J.W.; Rötter, 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.; Shcherbak, I.; Steduto, P.; Stöckle, C.O.; Stratonovitch, P.; Streck, T.; Supit, I.; Tao, F.; Travasso, M.; Waha, K.; White, J.W. and Wolf, J. Multimodel ensembles of wheat growth: many models are better than one. *Global Change Biology*. **2015**, 21(2), 911-25. doi: 10.1111/gcb.12768.

(11) Wiebe et al. Climate change impacts on agriculture in 2050 under a range of plausible socioeconomic and emissions scenarios. *Environmental Research Letters*. **2015**, 10, 085010.

(12) Marschmann, G.L., Pagel, H., Kügler, P., Streck, T. Equifinality, sloppiness, and emergent structures of mechanistic soil biogeochemical models. *Environ. Modelling and Software*. **2019**, 122, 104518. 10.1016/j.envsoft.2019.104518

(13) Harrison, M.T.; Roggero, P.P and Zavattaro L. Simple, efficient and robust techniques for automatic multi-objective function parameterisation: Case studies of local and global optimisation using APSIM. *Environmental Modelling & Software*. **2019**, 117, 109-33

(14) Harrison, M.T.; Cullen, B.R.; Mayberry, D.E.; Cowie, A.L.; Bilotto, F.; Badgery, W.B.; Liu, K.; Davison, T.; Christie, K.M.; Muleke, A. and Eckard, R.J. Carbon myopia: The urgent need for integrated social, economic and environmental action in the livestock sector. *Global Change Biology*. **2021**, 27, 5726-61.

(15) Harrison, M.T.; Cullen, B.R.; Tomkins, N.W.; McSweeney, C.; Cohn P. and Eckard, R.J. The concordance between greenhouse gas emissions, livestock production and profitability of extensive beef farming systems. *Animal Production Science*. **2016**, 56 370-84.

(16) Christie, K.M.; Smith, A.P.; Rawnsley, R.P.; Harrison, M.T.; Eckard, R.J. Simulated seasonal responses of grazed dairy pastures to nitrogen fertilizer in SE Australia: Pasture production. *Agricultural Systems*. **2018**, 166, 36-47.

(17) Christie, K.M.; Smith, A.P.; Rawnsley, R.P.; Harrison, M.T.; Eckard, R.J. Simulated seasonal responses of grazed dairy pastures to nitrogen fertilizer in SE Australia: N loss and recovery. *Agricultural Systems*. **2020**, 182, 102847.

(18) Verghese, G. Getting to the gray box: Some challenges for model reduction. In: American Control Conference. IEEE, 5–6. **2009**.

(19) Transtrum, M.K. and Qiu, P. Bridging mechanistic and phenomenological models of complex biological systems. *PLoS Computational Biology*. **2016**, 12 (5), e1004915.

(20) Harrison, M.T.; Evans, J.R. and Moore, A.D. Using a mathematical framework to examine physiological changes in winter wheat after livestock grazing: 1. Model derivation and coefficient calibration. *Field Crops Research*. **2012a**, 136, 116-26.

(21) Harrison, M.T.; Evans, J.R. and Moore, A.D. Using a mathematical framework to examine physiological changes in winter wheat after livestock grazing: 2. Model validation and effects of grazing management. *Field Crops Research*. **2012b**, 136, 127-37.

(22) Boote, K.J.; Porter, C.; Jones, J.W.; Thorburn, P.J.; Kersebaum, K.C.; Hoogenboom, G.; White, J.W. and Hatfield, J.L. Sentinel Site Data for Crop Model Improvement—Definition and Characterization. In: Improving Modeling Tools to Assess Climate Change Effects on Crop Response. **2016**. DOI: 10.2134/advagricsystmodel7.2014.0019

(23) Confalonieri, R.; Orlando, F.; Paleari, L.; Stella, T.; Gilardelli, C.; Movedi, E.; Pagani, V.; Cappelli, G.; Vertemara, A.; Alberti, L.; Alberti, P.; Atanassiu, S.; Bonaiti, M.; Cappelletti, G.; Ceruti, M.; Confalonieri, A.; Corgatelli, G.; Corti, P.; Dell'Oro, M.; Ghidoni, A.; Lamarta, A.; Maghini, A.; Mambretti, M.; Manchia, A.; Massoni, G.; Mutti, P.; Pariani, S.; Pasini, D.; Pesenti, A.; Pizzamiglio, G.; Ravasio, A.; Rea, A.; Santorsola, D.; Serafini, G.; Slavazza, M.; Acutis, M. Uncertainty in crop model predictions: What is the role of users? *Environmental Modelling & Software*. **2016**, 81, 165-173; https://doi.org/10.1016/j.envsoft.2016.04.009.

(24) Gaillard, R. K.; Jones, C. D.; Ingraham, P.; Collier, S.; Izaurralde, R. C.; Jokela, W.; Osterholz, W.; Salas, W.; Vadas, P.; and Ruark, M. Underestimation of N_2O emissions in a comparison of the DayCent, DNDC, and EPIC models. *Ecological Applications*. **2018**, 28(3), 694-708.

(25) Fitton, N., et al. Modelling biological N fixation and grass-legume dynamics with process-based biogeochemical models of varying complexity. *European Journal of Agronomy*. **2019**, 106, 58-66.

(26) Brilli, L., et al. Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. Science of the Total Environment. **2017**, 598, 445-470. https://doi.org/10.1016/j.scitotenv.2017.03.208

(27) Wallach, D.; Palosuo, T.; Thorburn, P.; Hochman, Z.; Gourdain, E.; Andrianasolo, F.; Asseng, S.; Basso, B.; Buis, S.; Crout, N.; et al. The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. Environmental Modelling & Software. **2021**, 145, 105206

(28) Brown, C.D.; Baer, U.; Günther, P.; Trevisan, M.; Walker, A. Ring test with the models LEACHP, PRZM-2 and VARLEACH: variability between model users in prediction of pesticide leaching using a standard data set. *Pesticide Science*. **1996**, 47, 249-258.

(29) Seidel, S.J.; Palosuo, T.; Thorburn, P.; Wallach D. Towards improved calibration of crop models – where are we now and where should we go? *European Journal of Agronomy*. **2018**, 94, 25-35.

(30) Harvey-Jordan, S.; and Long, S. The process and the pitfalls of semi-structured interviews. Community Practitioner. **2001**, 74(6), 219.

(31) Falatoonitoosi, E.; Leman, Z.; Sorooshian, S.; Salimi, M. Decision-Making Trial and Evaluation Laboratory. *Research Journal of Applied Sciences, Engineering and Technology*. **2013**, 5, 3476–3480.

(32) Saaty, T.L. Theory and Applications ofthe Analytic Network Process: Decision Making with Benefits, Opportunities, Costs, and Risks; RWS Publications: Pittsburgh, PA, USA. **2005**.

(33) Kendall, M.G. A Million Random Digits with 100,000 Normal Deviates. *Economica*. **1955**, 22 (88), p.365-366

(34) Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill. 1956.

(35) Holzworth, D.P.; Huth, N.I.; deVoil, P.G.; et al. APSIM – Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*. **2014**, 62, 327-350.

(36) Gabrielle, B.; Menasseri, S.; Houot, S. Analysis and field evaluation of the CERES models water balance component. *Soil Science Society of American Journal*. **1995**, 59 (5), 1403–1412. http://dx.doi.org/10.2136/sssaj1995.03615995005900050029x.

(37) Parton, W.J., et al. A general model for soil organic matter dynamics: sensitivity to litter chemistry, texture and management. p. 147–167. Quantitative Modeling of Soil Forming Processes, SSSA Spec. Public. No. 39. Madison, WI, USA. **1994**.

(38) Li, C.; Frolking, S.; Frolking, T.A. A model of nitrous oxide evolution from soil driven by rainfall events: 2. Model applications. *Journal of Geophysical Research*. **1992a**, 97, 9777–9783. http://dx. doi.org/10.1029/92JD00509.

(39) Smith, W.; Grant, B.; Qi, Z.; He, W.; Vander Zaag, A.; Drury, C.F.; Helmers, M. Development of the DNDC model to improve soil hydrology and incorporate mechanistic tile drainage: A comparative analysis with RZWQM2. *Environmental Modelling and Software*. **2020**, 123, 104577. doi.org/10.1016/j.envsoft.2019.104577

(40) Haas, E.; Klatt, S.; Fröhlich, A. et al. LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. *Landscape Ecology*. **2013**, 28, 615–636. https://doi.org/10.1007/s10980-012-9772-x

(41) Tsuji, G.Y. Network management and information dissemination for agrotechnology transfer. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Understanding Options for Agricultural Production. Kluwer Academic Publishers, Dordrecht, The Netherlands, 367–381. **1998**.

(42) Uehara, G. Synthesis. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Understanding options for agricultural production. Kluwer Academic Publishers, Dordrecht, The Netherlands, 389–392. **1998**.

(43) Jones, J.W., et al. Decision support system for agrotechnology transfer; DSSAT v3. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Understanding Options for Agricultural Production. Kluwer Academic Publishers, Dordrecht, the Netherlands, 157–177. **1998**.

(44) Williams, J.R. The EPIC model. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO, USA, 909–1000. **1995**.

(45) Riedo, M., et al. A Pasture Simulation Model for dry matter production, and fluxes of carbon, nitrogen, water and energy. *Ecological Modelling*. **1998**, 105, 141–183. http://dx.doi.org/10.1016/S0304-3800(97)00110-5.

(46) Johnson I.R. DairyMod and the SGS Pasture Model: a mathematical description of the biophysical model structure. **2016**, IMJ Consultants, Dorrigo, NSW, Australia.

(47) Berntsen, J.; Petersen, B.M.; Olesen, J.E. and Hutchings, N. FASSET - An Integrated Economic and Environmental Farm Simulation Model. - Proceedings of the International Symposium Modelling Cropping Systems. **1999**, Lleida, 21-23 June, Catalonia, Spain.

(48) Brisson, N., et al. STICS: a generic model for the simulation of crops and their water and nitrogen balance. I. Theory and parameterization applied to wheat and corn. *Agronomie*. **1998**, 18, 311–346.

(49) Aggarwal, P.K.; Kalra, N.; Chander, S.; and Pathak, H. InfoCrop: A dynamic simulation model for the assessment of crop yields, losses due to pests, and environmental impact of agro-ecosystems in tropical environments. I. Model description. *Agricultural Systems*. **2006**, 89 (1), 1-25. https://doi.org/10.1016/j.agsy.2005.08.001.

(50) Parker, W.S. Ensemble modeling, uncertainty and robust predictions. *WIREs Climate Change*. **2013**, 4, 213–223.

(51) Kahnemann, D. "Bounded Rationality Maps: Psychology for Behavioral Economics." *American Economic Review*. **2003**, 93 (5), 1449-1475. DOI: 10.1257/000282803322655392

(52) Madsen, H. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology*. **2000**, 235, 276–288.

(53) Kahneman, D. and Lovallo, D. Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking. *Management Science*. **1993**, 39 (1), 17–31.

(54) Thaler, R. H. Mental Accounting Matters. *Journal of Behavioral Decision Making*. **1999**, 12 (3), pp. 183–206.

(55) Read, D. Loewenstein, G. and Rabin, M. Choice Bracketing. *Journal of Risk and Uncertainty*. **1999**, 19 (1–3), 171–97.

(56) Beven, K.; Freer, J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*. **2001**, 249, 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

(57) Guillaume, J.H.A.; Jakeman, J.D.; Marsili-Libelli, S.; Asher, M., Brunner, P.; Croke, B.; Hill, M.C.; Jakeman, A.J.; Keesman, K.J.; Razavi, S.; Stigter, J.D. Introductory overview of identifiability analysis: A guide to evaluating whether you have the right type of data for your modeling purpose. *Environmental Modelling & Software*. **2019**, 119, 418-432.

(58) Machta, B.B.; Chachra, R.; Transtrum, M.K.; Sethna, J.P. Parameter space compression underlies emergent theories and predictive models. *Science*. **2013**, 342 (6158), 604–607.

(59) Her, Y.; and Chaubey I. Impact of the numbers of observations and calibration parameters on equifinality, model performance, and output and parameter uncertainty. *Hydrological Processes*. **2015**, 29, 4220-4237. 10.1002/hyp.10487.

Figure 1: Framework of the variable partition between input categories and variables used in the five stages of the model ensemble protocol described in Ehrhardt et al. (2018).

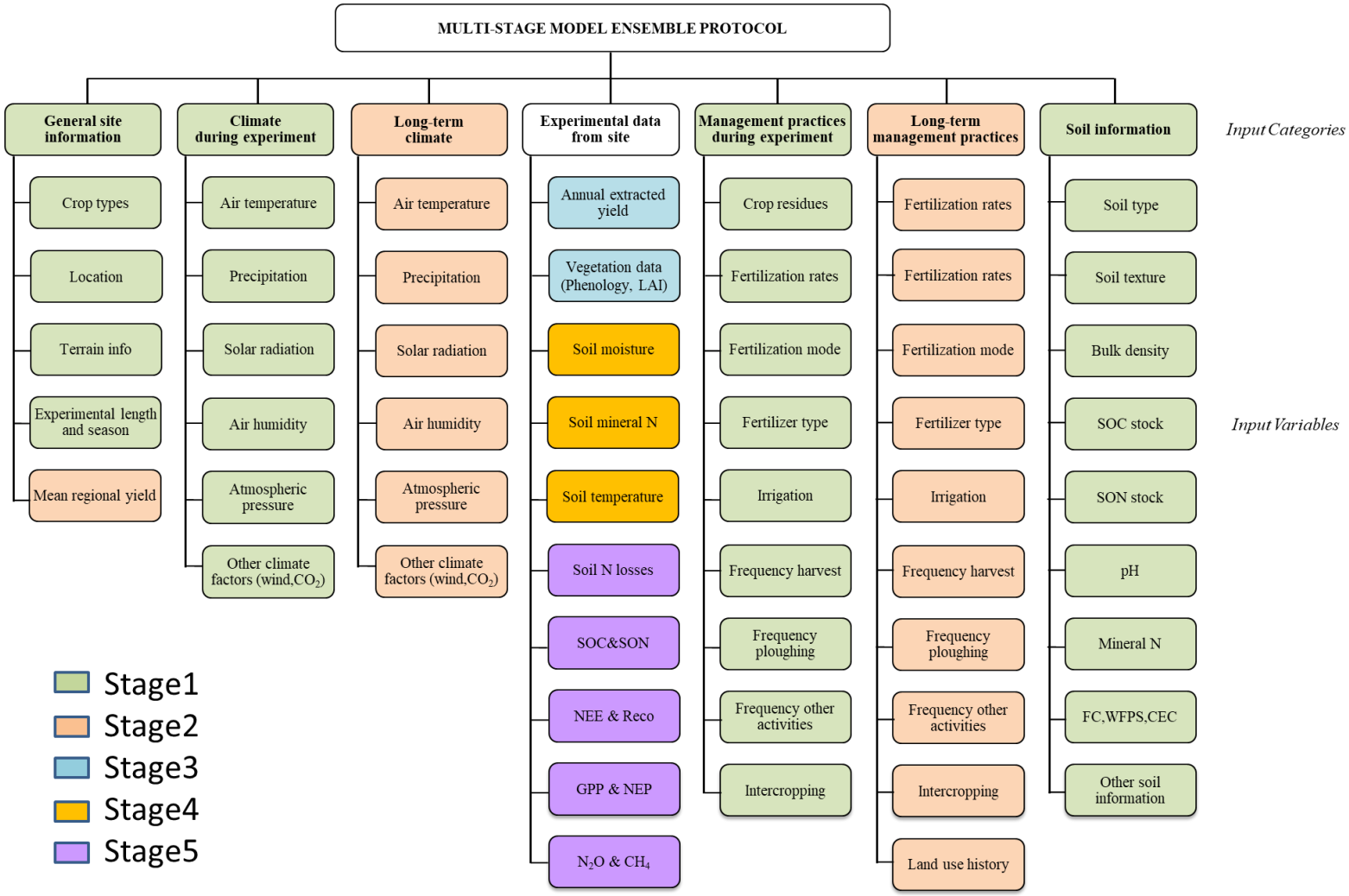


Figure 2: Steps of the Multi-criteria Decision Method process combining the Decision Making Trial and Evaluation Laboratory (DEMATEL) and the Analytic Network Process (ANP) methods. Through DEMATEL we visualize the perceived relationship existing between different variable categories. While in ANP, the strength of the relationships outlined in DEMATEL are integrated in a network of dependencies and feedbacks among input variables to determine their relative importance across the five stages of the modelling protocol.

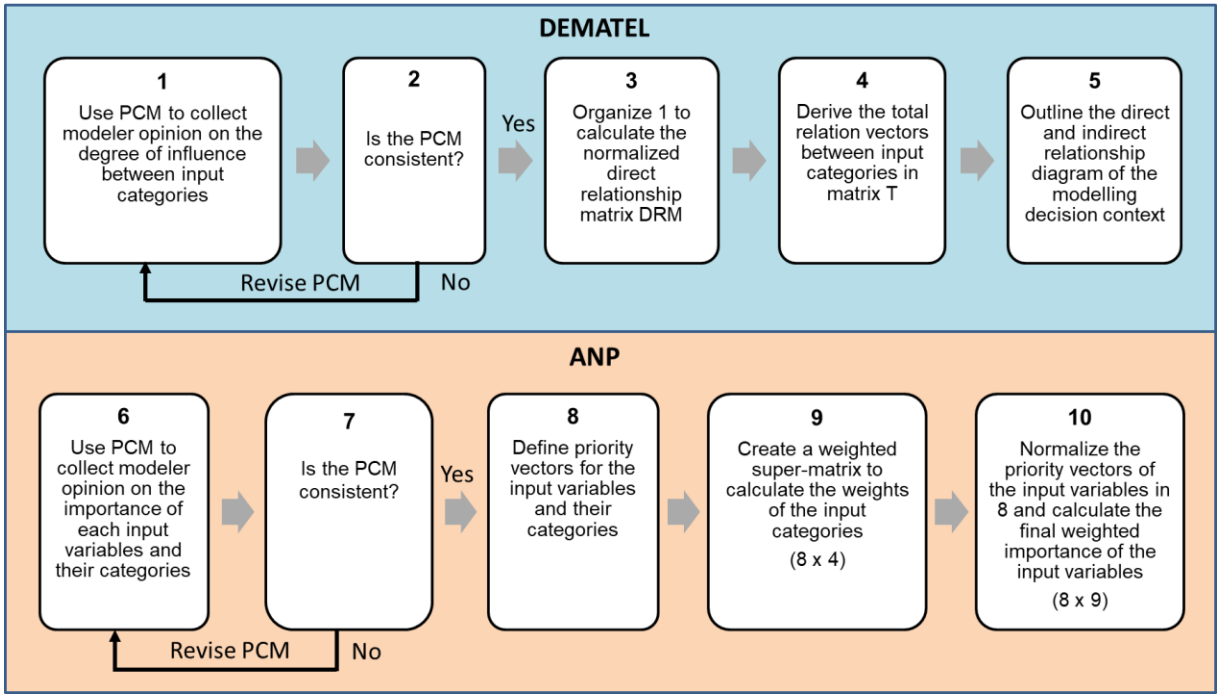
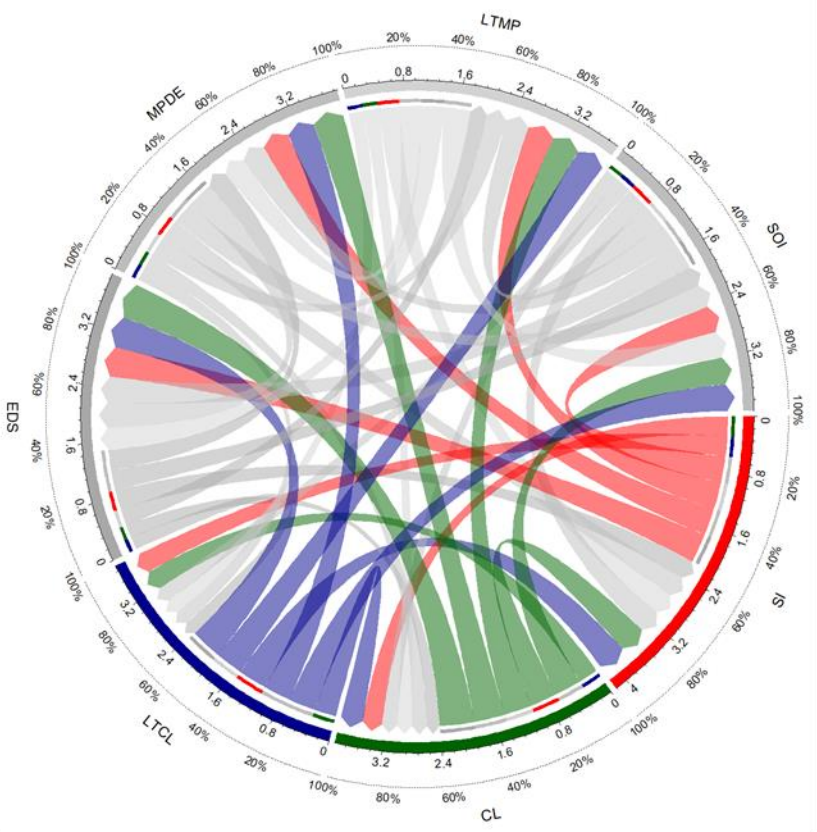


Figure 3: The table reports the total relation matrix of DEMATEL summarising the level (mean±sd) of direct and indirect influence given (G) and received (R) in each input category, the net influence (G-R), and the total level of influence (or dominance) (G+R) of the model input category used in the model ensemble study. Categories with positive G-R have a net influence towards the value of other variable categories and are denoted as "influential" categories. The circular diagram outlines the causal relationship in the model ensemble protocol between General site information (SI; red lines), Climate during experiment (CL; green lines), Long-term climate (LTCL; purple lines), Management practices during experiment (MPDE), long-term management practices (LTMP), Environmental data from site (EDS), and Soil information (SOI). The arrows in the diagram show the direction and the level of influence that each input category gives and receives from other categories. The coloured arrows highlight the three variable categories that resulted to be net influencers in the model ensemble protocol (i.e. positive G-R in Table 4). Radial bar numbers represent the total level of influence R+C, and the relative percentage of the casual relationship within each input category.



	SI	CL	LTCL	EDS	MPDE	LTMP	SOI	Given influence (G)	Net influence (G-R)	Total influence (G+R)
SI	0.31 ± 0.14	0.28 ± 0.15	0.29 ± 0.15	0.42 ± 0.19	0.39 ± 0.14	0.39 ± 0.14	0.38 ± 0.16	2.47 ± 0.85	0.14	4.80
CL	0.39 ± 0.18	0.25 ± 0.10	0.28 ± 0.12	0.50 ± 0.20	0.45 ± 0.17	0.40 ± 0.16	0.39 ± 0.16	2.65 ± 0.79	0.98	4.32
LTCL	0.41 ± 0.19	0.31 ± 0.15	0.24 ± 0.11	0.44 ± 0.19	0.40 ± 0.15	0.41 ± 0.15	0.39 ± 0.15	2.61 ± 0.83	0.97	4.25
EDS	0.28 ± 0.14	0.19 ± 0.09	0.19 ± 0.10	0.26 ± 0.11	0.29 ± 0.12	0.24 ± 0.10	0.29 ± 0.13	1.75 ± 0.60	-0.95	4.44
MPDE	0.30 ± 0.17	0.21 ± 0.13	0.20 ± 0.12	0.36 ± 0.19	0.27 ± 0.11	0.29 ± 0.13	0.33 ± 0.16	1.96 ± 0.86	-0.51	4.42
LTMP	0.31 ± 0.16	0.21 ± 0.13	0.21 ± 0.13	0.32 ± 0.18	0.31 ± 0.16	0.26 ± 0.11	0.38 ± 0.17	2.02 ± 0.86	-0.32	4.36
SOI	0.33 ± 0.16	0.22 ± 0.11	0.22 ± 0.11	0.38 ± 0.20	0.35 ± 0.15	0.34 ± 0.14	0.29 ± 0.13	2.12 ± 0.82	-0.32	4.56
Received influence(R)	2.33 ± 1.07	1.67 ± 0.81	1.64 ± 0.79	2.70 ± 1.18	2.46 ± 0.91	2.34 ± 0.86	2.44 ± 0.99			

Figure 4: The table summarises the Relative Root Mean Square Error (RRMSE) averaged across 19 models for the ensemble simulations of soil N_2O emissions from arable and grassland systems, crop yields of annual crop monocultures such as maize, wheat and rice, and above-ground net primary productivity in grassland (ANPP). Pi corresponds to the cumulative modelling importance of the input variable accessed in the five stages of the model ensemble framework. MER represents the rate of model simulation error for yield, N_2O, and ANPP per unit of modelling importance in each stage. The bar chart below the table outlines the trend of MER across the five stages of the model ensemble protocol.

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
RRMSE crop yield	0.36±0.28	0.47±0.23	0.24±0.15	0.24±0.16	0.25±0.19
RRMSE N_2O	1.11±0.95	0.96±0.83	1.22±0.95	0.97±0.53	0.68±0.36
RRMSE ANPP	1.01±0.40	0.70±0.25	0.77±0.28	0.76±0.18	0.81±0.19
Pi	0.67±0.02	0.11±0.01	0.06±0.03	0.05±0.01	0.11±0.02
MER crop yields	0.53	0.61	0.29	0.27	0.25
MER N_2O emissions	1.65	1.24	1.45	1.09	0.68
MER ANPP	1.50	0.90	0.91	0.85	0.81

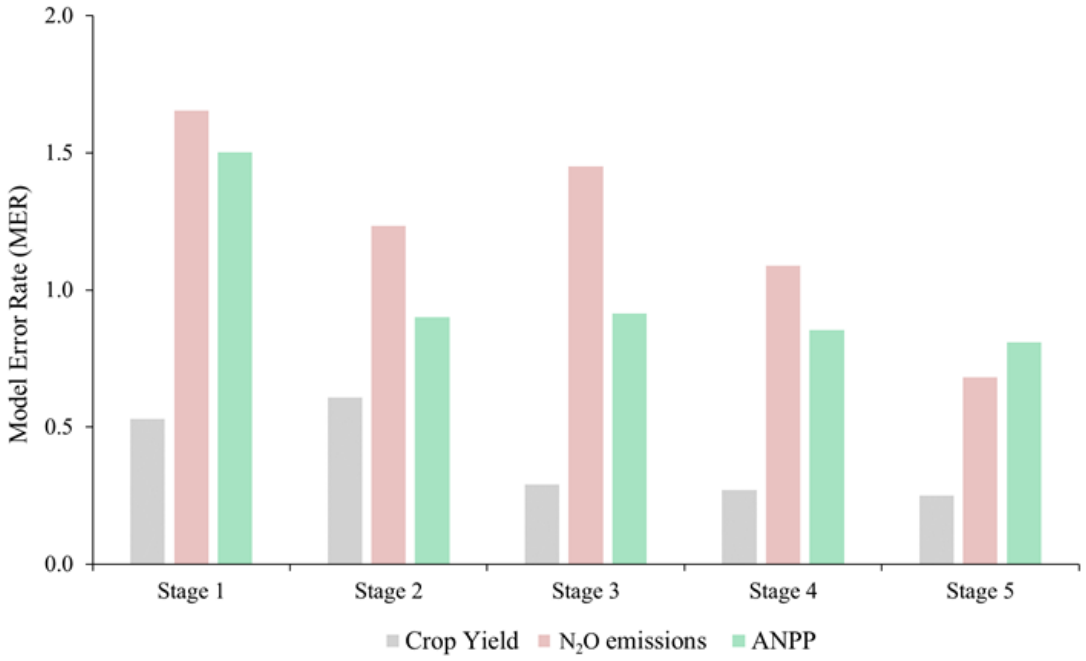


Table 1: Background information reported by age class (AC) of the modellers participating in the multistage intercomparison protocol and MCDM survey. N = number of modellers, F= proportion of modellers identified as female, PhD = proportion of modellers holding a PhD degree, FTC = proportion of modeller with a fix-term contract, MU = knowledge on number of models, MP = number of models published in peer-reviewed articles, and MT= type of model used.

AC	N	F	PhD	FTC	E	MU	MP	MT
25-34	4	0%	50%	75%	50%	from 1 to 4	from 1 to 4	Daycent, DNDC, Manure DNDC, Century, SPA/DALEC, EU-Rotate_N, FASSET, FarmAC
35-44	7	29%	100%	71%	86%	from 1 to 7	from 1 to 4	CERES-EGC, PaSim, FarmSim, EcoSys, Armosa, Daycent, DSSAT, EPIC, APEX, ModVege, Gemini, DairyMod, APSIM, GrassGro, AusFarm, GrazFeed, SGS, FarMax
45-54	7	43%	57%	43%	100%	from 1 to 7	from 1 to 7	AusFarm, DNDC, Daycent, Century, Tier II IPCC, RZWQM2, LEACHM, InfoCrop, DSSAT, STICS, Daycent, Century, SGS, DairyMod, RothC, DairyMod, GrassGro
55-64	1	100%	0%	0%	100%	3	3	Overseer, FarMax, APSIM

Table 2: Summary of the importance of the input variables and their categories, the consistency ratio of the modeller's judgments in the pairwise comparison matrix of each input category, and level of Kendal concordance between the modellers. The ranking scores (i.e importance) with letters m and e are significantly different at the $p < 0.05$ level between model types and modeller's experience groups, respectively. * indicates level of concordance significant within each variable categories at the $p < 0.05$ level.

Input Category (Importance)	Input variable	Ranking mean ± sd	Consistency Ratio mean ± sd	Kendal level mean
General site information (0.05 ± 0.04)	Crop type (crop rotation)	0.31 ± 0.16 m	0.06 ± 0.03	0.50*
	Mean regional yield	0.24 ± 0.11		
	Experimental length and season	0.23 ± 0.10 e		
	Location (country, latitude N)	0.12 ± 0.08		
	Terrain info	0.09 ± 0.06		
Climate during experiment (0.13 ± 0.08) e	Precipitation	0.32 ± 0.08	0.05 ± 0.03	0.81*
	Air temperature	0.26 ± 0.08		
	Solar radiation	0.21 ± 0.08		
	Other climate factors (Wind, [CO_2])	0.09 ± 0.05		
	Air humidity	0.07 ± 0.04 e		
	Atm. pressure	0.05 ± 0.02		
Long-term Climate (0.03 ± 0.02) m	Air temperature	0.26 ± 0.08	0.04 ± 0.04	0.60*
	Precipitation	0.32 ± 0.08		
	Solar radiation	0.21 ± 0.08		
	Air humidity	0.07 ± 0.04		
	Atm. pressure	0.05 ± 0.02		
	Other climate factors (Wind, [CO_2])	0.09 ± 0.05 m		
Experimental data from site (0.23 ± 0.10) e	Annual extracted yield	0.15 ± 0.12	0.08 ± 0.05	0.03
	Dynamic soil moisture	0.13 ± 0.07		
	Vegetation data (Phenology, LAI)	0.12 ± 0.08		
	N2O and/or CH_4	0.12 ± 0.08		
	GPP & NEP	0.10 ± 0.08 e		
	NEE & Reco	0.09 ± 0.06 e		
	Dynamic soil mineral N	0.08 ± 0.05 m		
	Dynamic SOC & SON stock	0.08 ± 0.05		
	Soil N losses	0.07 ± 0.04		
	Dynamic soil temperature	0.05 ± 0.03 e		
Management practices during experiment (0.26 ± 0.09)	Fertilization rates	0.23 ± 0.07	0.08 ± 0.04	0.12
	Irrigation	0.18 ± 0.06 e		
	Freq. harvest, grazing & cut in grass	0.13 ± 0.07		
	Fertilizer type	0.10 ± 0.04		
	Frequency of ploughing	0.09 ± 0.04		
	Crop residues	0.07 ± 0.04		
	Intercropping	0.08 ± 0.03		
	Fertilization mode	0.06 ± 0.04		
	Frequency of other activities	0.06 ± 0.04		
Long-term management practices (0.06 ± 0.04)	Fertilization rates	0.20 ± 0.07 m	0.06 ± 0.04	0.09
	Irrigation	0.17 ± 0.06		
	Frequency of harvest	0.11 ± 0.07		
	Intercropping	0.09 ± 0.05		
	Land use history	0.08 ± 0.06		
	Frequency of ploughing	0.09 ± 0.03		
	Crop residues	0.09 ± 0.03		
	Fertilizer type	0.07 ± 0.03		
	Frequency of other activities	0.06 ± 0.03		
	Fertilization mode	0.04 ± 0.03		
Soil information (0.23 ± 0.09) m	Soil texture	0.20 ± 0.10 m	0.09 ± 0.07	0.22*
	FC, WFPS, CEC	0.19 ± 0.09		
	Bulk density	0.15 ± 0.04		
	Initial SOC stock	0.13 ± 0.04 m, e		
	pH	0.09 ± 0.05 m		
	Initial SON stock	0.09 ± 0.04 m		
	Soil mineral N	0.07 ± 0.04		
	Other soil information	0.06 ± 0.03		
	Soil type	0.03 ± 0.02		

Footnote: The colour gradient indicates where the relative importance of each input variables falls within each variable category.

Table 3: Cumulative importance of the five stages of the model ensemble protocol. Within each stage, the ranking of the input variables shown in Table 2 was normalized over the importance score of their corresponding categories.

Input Category	Input Variable	Stage 1 (**0.67** ± 0.02)	Stage 2 (**0.11** ± 0.01)	Stage 3 (**0.06** ± 0.03)	Stage 4 (**0.05** ± 0.01)	Stage 5 (**0.11** ± 0.02)
Soil information	Soil type	0.01 ± 0.01				
	Soil texture	0.05 ± 0.04				
	Bulk density	0.04 ± 0.15				
	SOC stock	0.03 ± 0.13				
	SON stock	0.02 ± 0.09				
	pH	0.02 ± 0.09				
	Soil mineral N	0.02 ± 0.07				
	FC, WFPS, CEC	0.04 ± 0.19				
	Other soil information	0.01 ± 0.06				
Climate exp.	Air temperature	0.03 ± 0.03				
	Precipitation	0.04 ± 0.03				
	Solar radiation	0.03 ± 0.01				
	Air humidity	0.01 ± 0.01				
	Atm. pressure	0.01 ± <0.00				
	Other climate factors	0.01 ± <0.00				
Management practices during experiment	Crop residues	0.02 ± 0.01				
	Fertilization rates	0.06 ± 0.03				
	Fertilization mode	0.02 ± 0.01				
	Fertilizer type	0.03 ± 0.02				
	Irrigation	0.05 ± 0.02				
	Frequency of ploughing	0.02 ± 0.01				
	Frequency other activities	0.02 ± 0.01				
	Intercropping	0.02 ± 0.01				
	Freq. harvest, grazing & cut in grass	0.03 ± 0.01				
General site information	Crop type	0.02 ± 0.01				
	Location	0.01 ± 0.01				
	Terrain info	<0.00 ± <0.00				
	Experimental length	0.01 ± 0.02				
	Mean regional yield		0.01 ± 0.01			
Long-term Climate	Air temperature		0.01 ± <0.00			
	Precipitation		0.01 ± 0.01			
	Solar radiation		0.01 ± 0.01			
	Air humidity		<0.00 ± <0.00			
	Atm. pressure		<0.00 ± <0.00			
	Other climate factors		<0.00 ± <0.00			
Long-term management practices	Fertilization rates		0.01 ± 0.01			
	Fertilization mode		<0.00 ± <0.00			
	Fertilizer type		<0.00 ± <0.00			
	Irrigation		0.01 ± 0.01			
	Frequency of harvest		0.01 ± 0.01			
	Frequency of ploughing		0.01 ± <0.00			
	Frequency other activities		<0.01 ± <0.00			
	Crop residues		0.01 ± <0.00			
	Intercropping		0.01 ± <0.00			
	Land use history		0.01 ± 0.01			
Experimental data from site	Annual extracted yield			0.03 ± 0.03		
	Vegetation data (phenology, LAI)			0.03 ± 0.03		
	Soil temperature				0.01 ± 0.01	
	Soil moisture				0.03 ± 0.01	
	Soil mineral N				0.02 ± 0.01	
	SOC & SON					0.02 ± 0.01
	GPP & NEP					0.02 ± 0.02
	NEE & Reco					0.02 ± 0.02
	Soil N losses					0.02 ± 0.01
	N_2O and/or CH_4					0.03 ± 0.03

Footnote: SOC= soil organic carbon, SON= soil organic nitrogen, FC= field capacity, WFPS= water field pore space, CEC= cation exchange capacity, GPP= gross primary production, NEP= net ecosystem production, NEE= net ecosystem exchange, Reco= ecosystem respiration.