DR. DANIEL WALLACH (Orcid ID : 0000-0003-3500-8179)

DR. SENTHOLD ASSENG (Orcid ID : 0000-0002-7583-3811)

DR. MUKHTAR AHMED (Orcid ID : 0000-0002-7223-5541)

DR. DAVIDE CAMMARANO (Orcid ID : 0000-0003-0918-550X)

DR. CURTIS DINNEEN JONES (Orcid ID : 0000-0002-4008-5964)

DR. F TAO (Orcid ID : 0000-0001-8342-077X)

Article type : Primary Research Articles

Multi-model ensembles improve predictions of crop-environment-management interactions

Running head

Multi-model ensembles improve predictions

Authors

D. Wallach[1,*], P. Martre[2], B. Liu[3,4], S. Asseng[4], F. Ewert[5,6], P.J. Thorburn[7], M. van Ittersum[8], P.K. Aggarwal[9,†], M. Ahmed[10,11], B. Basso[12,13], C. Biernath[14], D. Cammarano[15], A.J. Challinor[16,17], G. De Sanctis[18,‡], B. Dumont[19], E. Eyshi Rezaei[5,20], E. Fereres[21], G.J. Fitzgerald[22,23], Y. Gao[4], M. Garcia-Vila[21], S. Gayler[24], C. Girousse[25], G. Hoogenboom[4,26], H. Horan[7], R.C. Izaurralde[27,28], C.D. Jones[28], B.T. Kassie[4], K.C. Kersebaum[29], C. Klein[30], A.K. Koehler[16], A. Maiorano[2,31], S. Minoli[32], C. Müller[32], S. Naresh Kumar[33], C. Nendel[29], G.J. O'Leary[34], T. Palosuo[35], E. Priesack[30], D. Ripoche[36], R.P. Rötter[37,38], M.A. Semenov[39], C. Stöckle[10], P. Stratonovitch[39], T. Streck[24], I. Supit[40], F. Tao[41,35], J. Wolf[42], and Z. Zhang[43]

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.14411

Affiliations

[1]UMR AGIR, INRA 31326 Castanet-Tolosan, France.

[2]UMR LEPSE, INRA, Montpellier SupAgro, 34 060, Montpellier, France.

[3]National Engineering and Technology Center for Information Agriculture, Key Laboratory for Crop System Analysis and Decision Making, Ministry of Agriculture, Jiangsu Key Laboratory for Information Agriculture, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing, Jiangsu 210095, P. R. China.

[4]Agricultural & Biological Engineering Department, University of Florida, Gainesville, FL 32611, USA.

[5]Institute of Crop Science and Resource Conservation INRES, University of Bonn, 53115, Germany.

[6]Leibniz Centre for Agricultural Landscape Research, 15374 Müncheberg, Germany.

[7]CSIRO Agriculture and Food, Brisbane, St Lucia Queensland 4067, Australia.

[8]Plant Production Systems Group, Wageningen University, 6700 AK Wageningen, The Netherlands.

[9]CGIAR Research Program on Climate Change, Agriculture and Food Security, BISA-CIMMYT, New Delhi-110012, India.

[10]Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120.

[11]Department of Agronomy, Pir Mehr Ali Shah Arid Agriculture University Rawalpindi-46300, Pakistan.

[12]Department of Earth and Environmental Sciences, Michigan State University, East Lansing, Michigan 48823, USA.

[13]W.K. Kellogg Biological Station, Michigan State University East Lansing, Michigan 48823, USA.

[14]Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research Center

for Environmental Health, Neuherberg, 85764, Germany.

[15]James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK.

[16]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS29JT, UK.

[17]CGIAR-ESSP Program on Climate Change, Agriculture and Food Security, International Centre for Tropical Agriculture (CIAT), A.A. 6713, Cali, Colombia.

[18]European Food Safety Authority, GMO Unit, Via Carlo Magno 1A, Parma, IT-43126, Italy.

[19]Department Terra & AgroBioChem, Gembloux Agro-Bio Tech, University of Liege, Gembloux 5030, Belgium.

[20]Center for Development Research (ZEF), Walter-Flex-Straße 3, 53113 Bonn, Germany.

[21]IAS-CSIC and University of Cordoba, Apartado 3048, 14080 Cordoba, Spain.

[22]Agriculture Victoria Research, Department of Economic Development, Jobs, Transport and Resources, Ballarat, Victoria 3350 Australia.

[23]Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, 4 Water Street, Creswick, VIC 3363, Australia.

[24]Institute of Soil Science and Land Evaluation, University of Hohenheim, 70599 Stuttgart, Germany.

[25]UMR GDEC, INRA, Université Clermont Auvergne, 63000, Clermont-Ferrand, France.

[26]Institute for Sustainable Food Systems, University of Florida, Gainesville, FL 32611, USA.

[27]Department of Geographical Sciences, Univ. of Maryland, College Park, MD 20742, USA.

[28]Texas A&M AgriLife Research and Extension Center, Texas A&M Univ., Temple, TX 76502, USA.

[29]Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research, 15374 Müncheberg, Germany.

[30]Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, 85764, Germany.

[31]Present address: European Food Safety Authority – EFSA, via Carlo Magno 1/A, 43126 Parma - Italy

[32]Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany.

[33]Centre for Environment Science and Climate Resilient Agriculture, Indian Agricultural Research Institute, IARI PUSA, New Delhi 110 012, India.

[34]Grains Innovation Park, Agriculture Victoria Research, Department of Economic Development, Jobs, Transport and Resources, Horsham 3400, Australia.

[35]Natural Resources Institute Finland (Luke), 00790 Helsinki, Finland.

[36]US AgroClim, INRA, 84 914 Avignon, France.

[37] University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Grisebachstraße 6, 37077 Göttingen.

[38]University of Göttingen, Centre of Biodiversity and Sustainable Land Use (CBL), Buesgenweg 1, 37077 Göttingen, Germany.

[39]Computational and Systems Biology Department, Rothamsted Research, Harpenden, Herts, AL5 2JQ, UK.

[40]Water & Food and Water Systems & Global Change Group, Wageningen University, 6700AA Wageningen, The Netherlands.

[41]Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science, Beijing 100101, China.

[42]Plant Production Systems, Wageningen University, 6700AA Wageningen, The Netherlands.

[43]State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China.

*Corresponding author: Daniel Wallach (INRA Occitanie-Toulouse, UMR AGIR, Chemin de Borde Rouge, CS52627, 31326 Castanet-Tolosan Cedex, Tél: +33 5 61 28 50 35, daniel.wallach@inra.fr)

†Authors after P.K.A. contributed equally to this work and are listed in alphabetical order.

‡The views expressed in this paper are the views of the author and do not necessarily represent the views of the organization or institution to which he is currently affiliated.

Keywords: multi-model ensemble, climate change impact, prediction, crop models, ensemble mean, ensemble median

Abstract

A recent innovation in assessment of climate change impact on agricultural production has been to use crop multi model ensembles (MMEs). These studies usually find large variability between individual models but that the ensemble mean (e-mean) and median (e-median) often seem to predict quite well. However few studies have specifically been concerned with the predictive quality of those ensemble predictors. We ask what is the predictive quality of e-mean and e-median, and

how does that depend on the ensemble characteristics. Our empirical results are based on five MME studies applied to wheat, using different data sets but the same 25 crop models . We show that the ensemble predictors have quite high skill and are better than most and sometimes all individual models for most groups of environments and most response variables. Mean squared error of e-mean decreases monotonically with the size of the ensemble if models are added at random, but has a minimum at usually 2-6 models if best-fit models are added first. Our theoretical results describe the ensemble using four parameters; average bias, model effect variance, environment effect variance and interaction variance. We show analytically that mean squared error of prediction (MSEP) of e-mean will always be smaller than MSEP averaged over models, and will be less than MSEP of the best model if squared bias is less than the interaction variance. If models are added to the ensemble at random, MSEP of e-mean will decrease as the inverse of ensemble size, with a minimum equal to squared bias plus interaction variance. This minimum value is not necessarily small, and so it is important to evaluate the predictive quality of e-mean for each target population of environments. These results provide new information on the advantages of ensemble predictors, but also show their limitations.

Introduction

Climate change is expected to have an important impact on crop production and its geographic variability, with most results to date showing a negative influence of climate change on crop yields (IPCC, 2014). Crop simulation models are important tools for impact assessment, that allow one to generalize to environmental conditions and management options beyond those observed experimentally (Ewert et al., 2015; Porter et al., 2014). This makes possible for example a detailed spatial analysis of the impact of climate change (Rosenzweig et al., 2014) (Rosenzweig et al., 2014) and evaluation of adaptation strategies for climate change (Chenu et al., 2017).

A recent innovation in the use of crop models for impact assessment is the use of crop multi-model ensembles (MMEs), largely as a result of recent international cooperative programs (Ewert et al., 2015; Rosenzweig et al., 2013), although the first studies go back to 2011 (Palosuo et al., 2011). In these studies, different modeling groups running different models are given the same input information and requested to provide simulated values for the same output variables. An initial objective of these studies was to evaluate the uncertainty in crop model predictions. These studies found that there is large variability in predictions between models, implying large uncertainty in predictions when a single model is used (Asseng et al., 2013; Bassu et al., 2014; Hasegawa et al., 2017; Rötter, Carter, Olesen, & Porter, 2011). We use here the term "prediction" in the sense of calculating an output based on known inputs, rather than forecasting the future.

Crop MME studies have often noted that the ensemble mean (e-mean) and ensemble median (e-median) of simulated values give good agreement with observations (Bassu et al., 2014; Palosuo et al., 2011; Rötter et al., 2012). This suggests that in practice, it might be better to create a MME and then use the predictions of e-mean or e-median rather than use the predictions of an individual model. Several recent impact assessment studies have based conclusions on ensemble predictors (Asseng et al., 2014; Liu et al., 2016).

Only a few studies have examined the properties of crop MME predictors in more detail, in each case for one set of environmental conditions. One study, based on prediction of multiple response variables in four environments, found that e-mean and e-median were both better than the best model, for a composite criterion including all outputs and environments (Pierre Martre et al., 2015). Yin et al. (2017) found that e-mean predicted grain N better than a randomly chosen model. Of particular practical interest is the behavior of e-mean and e-median as a function of the number of

models in the ensemble. This has been studied by treating the ensemble as the full population of models, and drawing sub samples from that population. The conclusions have been that prediction error decreases systematically as the number of models increases. Li et al. (2015) suggested that eight models would be sufficient to obtain errors of e-mean below 10% of observed yield. All of these studies have been empirical, based on a single MME study. The general behavior of crop ensemble predictors has not been addressed. Studies in other fields, including group intelligence (Surowiecki, 2005), hydrologic modeling (Duan, Ajami, Gao, & Sorooshian, 2007), air quality modeling (Solazzo & Galmarini, 2015) and climate modeling (Tebaldi & Knutti, 2007) have also found that averaging over multiple opinions or solutions can give good predictions, often better than any individual model. The basis for using MME predictors has received particular attention in the field of climate modeling (Hagedorn et al., 2005; Weigel et al., 2008). However, the context there is quite different than for crop models; for example in climate modeling each MME member is often itself an ensemble based on a single model with different initial conditions (DelSole, Nattala, & Tippett, 2014) whereas in crop modeling, each model normally provides a single simulation, a major interest in climate modeling is in probabilistic predictions rather than the deterministic predictions of crop models (DelSole et al., 2013; Wang et al., 2009) and in climate modeling spatial patterns of prediction play an important role (DelSole et al., 2013).

One can easily imagine situations where e-mean and e-median for crop models do not predict well. For example, if all models have large positive bias, then e-mean and e-median will also have large positive bias, and e-median will be worse than half the models. Thus, one cannot automatically assume that one will obtain reliable predictions by using MME predictors. The question we ask then is what is the predictive quality of e-mean and e-median, and how does that depend on the ensemble characteristics? We break this down into specific sub-questions. First, how does the predictive quality of MME predictors compare to predictive quality of a model chosen at random

from the models in the ensemble, or to that of the best individual model in the ensemble, and how does that depend on the ensemble characteristics? The answer to this question affects the choice between using an individual model and a MME predictor. Second, what is the level of error of the MME predictors? This is a major determinant of the potential usefulness of these predictors. Finally, how does the level of error of the MME predictors depend on the number of models in the ensemble? This affects the very practical decision as to the number of models to include in a MME.

Materials and Methods

Data

The data sets simulated in the five wheat MME studies considered here are described in Table 1. Details are available in the cited references. Each data set concerns a different range of environmental conditions, where an environment is to be understood as a combination of physical environment and management. We consider each data set as representative of some infinite range of environments, the target population. The target population corresponding to the AgMIP wheat pilot data set is worldwide wheat environments. The data set is a sample from that population, and the prediction problem is prediction for a randomly chosen individual environment from that population. In the case of the HSC data set, the target population of environments is considered to be all possible weather sequences for wheat in Maricopa, Arizona, generated by different years and planting dates. The data set can be considered a sample from that distribution of environments, where the heat treatments are meant to increase artificially the diversity of the sampled conditions. In the case of the HSGE data set, the target population of environments is taken to be worldwide hot environments for wheat, including all possible weather sequences and all locations. The target population for the C3-GEM data set is taken to be all possible weather sequences at the location of the study, with or without heat shocks during grain filling. Finally, the target population

corresponding to the AGFACE data set is considered to be wheat crops under different weather sequences at the location of the study, with or without irrigation and with either current or enhanced CO_2 levels. We consider here four output variables that were measured in most or all of these studies: grain yield (yield), grain protein concentration (protein), final aboveground biomass (biomass) and maximum leaf area index during the course of growth (maximum LAI).

Models and calibration

We consider only the 25 crop models that provided simulation results for all of the data sets for at least yield and biomass (Supplementary Table S1). All of these models have been described in detail in separate publications (see references in Table S1). All are dynamic system models; they describe crop development, crop growth and soil processes of a homogeneous field over a single growing season, using differential or difference equations, often with a time step of one day. The explanatory variables include daily weather over the growing season, management (sowing date and cultivar, irrigation and fertilization, etc.) and soil characteristics and initial conditions. While there are certainly similarities between some of the crop models, it seems reasonable to consider them as independent since each has undergone at least some development independently of other models. Each model produces a single prediction of a specific output (e.g. yield) for each environment. In addition to the individual models in the MME we consider the two most common MME predictors, namely e-mean and e-median.

In all of these studies, some of the data were provided to the modeling groups for calibration (Table 1). The calibration data consisted of detailed crop data, including yield, from one environment for the HSC and AGFACE data sets, from the three control environments for the C3-GEM data set and from four environments for the HSGE data set, plus some peripheral information related to, but not the same as, the variables to be simulated (crop phenology information, parameter values of some models that had previously seen the data).

Evaluation metrics

Our basic criterion of simulation accuracy is mean squared error (MSE), i.e. squared error averaged over environments of a data set:

$$MSE = 1/N \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2$$

where y_i is the observed value for the i[th] environment of the data set, $\hat{y}_i$ is the corresponding simulated value, and N is the number of environments in the data set. *MSE* is calculated separately for each output variable and each model. Often it is more convenient to look at root mean squared error; $RMSE = \sqrt{MSE}$.

MSE is an important measure of model error, but skill measures are better at conveying the usefulness of model simulations, since they compare model errors to errors of some alternative, simple predictor. The skill measure commonly used for crop models is modelling efficiency (EF), defined as

$$EF = 1 - MSE_{model} / MSE_{\bar{y}}$$

where MSE_{model} is MSE for the model in question and $MSE_{\bar{y}}$ is MSE when all predictions use the average of observed values for that data set ($\bar{y}$). Since $\bar{y}$ is a constant, it explains none of the variability in the data set. A perfect model has EF=1. A model that does worse than $\bar{y}$ has $EF < 0$ and can be considered to have no skill in explaining variability between environments.

The above criteria refer to the data in the data set. As a criterion of prediction accuracy for the target population we use mean squared error of prediction (MSEP), defined as the expectation of squared error over the target population. It is well known that if the same data are used for

calibration and for evaluation, MSE tends to under-estimate MSEP. To examine how important this is, we calculated MSE for yield, using either all environments or leaving out all those environments which provided yield for calibration. The resulting MSE values for e-mean and e-median, and their ranks among all models, were very similar (Supplementary Table S2). We therefore use MSE based on all environments of a data set as an estimate of MSEP for the corresponding target population.

Statistical description of multi-model ensemble

We propose a random effects statistical model for describing model errors:

$$e_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (1)$$

where e_{ij} is error (observed value for environment *j* minus value simulated by model *i*), μ is the overall bias (error averaged over models and environments), α_i is a random model effect with mean 0 and variance σ_α^2, β_j is a random environment effect with mean 0 and variance σ_β^2 and γ_{ij} is the random interaction term, with mean 0 and variance σ_γ^2 (Scheffé, 1959). Thus the random effects model characterizes a MME and target population using four parameters: μ, σ_α^2, σ_β^2 and σ_γ^2.

If there is bias, this implies that predictions, averaged over models and environments, are too small or too large. For example, if models tended to underestimate potential yield for the cultivars of the HSGE data set, this could lead on the average to systematic under-prediction of yield and therefore to a positive bias. The bias term contributes equally to all individual models and therefore also to e-mean, for all environments of the target population. The model effect indicates to what extent a specific model over- or under- predicts, on the average over environments. The larger σ_α^2, the larger the variability between errors of different models. The environment effect indicates to what extent there is over- or under-prediction for individual environments, averaged over models. For example,

if all models tended to over-predict specifically for the highest temperatures of the HSC target population, this would lead to an environment effect. The larger σ_{β}^{2}, the larger the variability between errors for different environments. Finally, the interaction effect measures the effect of interaction between a specific model and a specific environment on model error.

If it is assumed that models are drawn at random from some underlying distribution of models, and that environments are drawn at random from the target population of environments, then all the random effects are mutually uncorrelated (Scheffé, 1959). If there is random measurement error it affects the observations of each environment and thus is included in the environment effect. The bias and variance components were estimated for each data set using the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2012) with the REML option. The variance components for yield, calculated with or without the environments that provided yield data for calibration, were quite similar (SupplementaryTable S5).

Results

Empirical results

Figure 1 shows RMSE relative to e-median ($RMSE_{model}$-$RMSE_{e\text{-}median}$) for yield for each model and each data set. Models with negative values have smaller RMSE than e-median. It is seen that e-median is better than all individual models (all individual models have positive values of RMSE relative to e-median) except for the HSGE and AGFACE studies, where there are respectively four and two individual models out of 25 that are better than e-median. E-mean is slightly worse than e-median (slightly positive RMSE relative to e-median) except for the HSGE data set. Its worst ranking for yield is seventh (among the 25 individual models, e-mean and e-median) . For protein, biomass

and maximum LAI, the rankings of e-median and e-mean are more variable. At worst e-median is ranked sixth and e-mean tenth. E-median is better than e-mean in 13 out of the 17 combinations of data set and output variable (Supplementary Figures S1-S3). Figure 2 shows as an example the fit of e-mean, e-median and the individual models to the HSC yield data.

The ranking of e-mean improves more or less systematically as one considers more environments, up to the actual number of environments for each data set (Supplementary Figure S4). A final step in this progression of averaging over more situations is to average over data sets. When RMSE values are averaged across data sets, e-mean is ranked 2, 6, 2 and 3 for the output variables yield, protein, biomass and maximum LAI, respectively (Supplementary Table S3). The corresponding ranks for e-median are 1, 1, 1 and 2. Among the individual models, the average rankings are more variable. The model SQ is systematically quite well ranked (3, 3, 3 and 8 for yield, protein, biomass and maximum LAI respectively) but the best individual model for protein has rankings of 13, 2, 18 and 23 for the four output variables and the best individual model for maximum LAI has rankings 12, 11, 21 and 1. In all cases, both e-mean and e-median are better than the average over individual models (bar labeled “ave” in Figure 1 and Supplementary Figures S1-S3).

Figure 1 shows that RMSE using the average of observed values (bar labeled “ybar”) is appreciably larger than RMSE for e-mean or e-median for yield for four of the studies, implying that the ensemble predictors have substantial skill values for those studies. However, no model, including e-mean and e-median, has skill for the HSGE data set (i.e. “ybar” has the smallest RMSE value). Over all combinations of study and output variable, e-mean and e-median have no skill in a little over one third of the situations (Supplementary Table S4).

Figure 3 shows empirical results for the effect of number of models on MSE of e-mean, for predicting yield. These results are averages over multiple choices of models, and correspond to choosing the models to add to the ensemble at random. There is an almost monotonic decrease in MSE as more models are added to the ensemble. Similar behavior is exhibited for the other output variables (Supplementary Figure S5).

Rather than building the MME by adding models chosen at random, suppose that one starts from the model with smallest RMSE and then adds models in the order of increasing RMSE. The general result of doing so is an initial decrease in RMSE and then a trend of increasing RMSE as the number of models in the ensemble increases. In 12 out of 17 combinations of data set and output, minimum RMSE is reached with 2-6 models in the ensemble (Figure 3 and Supplementary Figure S5).

Theoretical results

In the following we focus only on e-mean, which is more amenable to theoretical treatment than e-median. The analysis is based on eq. (1), which separates model error into a bias component and model, environment and model x environment interaction effects. The estimated values of μ, σ_α^2, σ_β^2 and σ_γ^2 for each data set and output variable are shown in Supplementary Tables S5-S8. The results are that squared bias μ^2 is usually much smaller than any of the variance components. That is, model error averaged over models and environments for each data set is small. The contributions of the other variance components are quite variable. Depending on the data set and the variable that is predicted, the major variability can arise from the variability in errors between models (e.g. maximum LAI prediction for the C3-GEM data set), the variability in errors between environments (e.g. biomass prediction for the AGFACE data set) or from the interaction (e.g. prediction of protein for the HSC data set).

MSEP of e-mean based on a MME of size n is

$$MSEP_{e-mean}(n) = E\left\{\left[\mu + (1/n)\sum_{i=1}^{n}\alpha_i + \beta_j + (1/n)\sum_{i=1}^{n}\gamma_{ij}\right]^2\right\} \quad (2)$$

Using the properties of the random effects model, this leads directly to

$$MSEP_{e-mean}(n) = \mu^2 + \sigma_\alpha^2/n + \sigma_\beta^2 + \sigma_\gamma^2/n \quad (3)$$

Letting n tend toward infinity, it is seen that in the limit of a very large MME

$$MSEP_{e-mean} = \mu^2 + \sigma_\beta^2 \quad (4)$$

On the other hand, the expectation of MSEP over individual models ($\overline{MSEP}$) is

$$\overline{MSEP} = E\left\{\left[\mu + \alpha_i + \beta_j + \gamma_{ij}\right]^2\right\} = \mu^2 + \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 \quad (5)$$

Thus $\overline{MSEP}$ is always as large as or larger than $MSEP_{e-mean}$. This is a generalization of the empirical results in Figure 1 and Supplementary Figures S1-S3, which show that e-mean has smaller RMSE than the average over models (the bar labeled “ave”) in all the cases considered.

Assuming the a_i values have a normal distribution, we can also obtain results for the probability that e-mean is better than any individual model. A model with random effect $\alpha_i = a$ has an MSEP value of

$$E\left[(\mu + \alpha_i + \beta_j + \gamma_{ij} \mid \alpha_i = a)^2\right] = (\mu + a)^2 + \sigma_\beta^2 + \sigma_\gamma^2 \quad (6)$$

If the a_i have a normal distribution, then in the limit of a very large MME, the probability that an individual model will have MSEP less than or equal to $MSEP_{emean}$ is

$$P\left[(\mu+a)^2+\sigma_\beta^2+\sigma_\gamma^2 \le \mu^2+\sigma_\beta^2\right]=P\left[a' \le \left(\mu^2-\sigma_\gamma^2\right)/\sigma_\alpha^2\right] \quad (7)$$

where $(a')^2$ is distributed as a noncentral chi squared variable with 1 degree of freedom and non-centrality parameter μ^2/σ_α^2 (Supplementary Figure S6). If $\sigma_\gamma^2 \ge \mu^2$ (interaction variance greater than squared bias), then in the limit of a very large MME this probability is 0. The result just depends on the relative values of squared bias and interaction variance, and not on how good the individual models are. The inequality is satisfied for every data set and output variable here, implying that in the limit of many models and averaged over environments, e-mean should be better than every model in the ensemble. This is an extension of the empirical results, which concern a finite number of models and environments. Those results show that there are relatively few models that are better than e-mean.

Equation (4) shows that $MSEP_{emean}$ is not necessarily small, even in the limit of a very large MME. It will only be small if both μ^2 and σ_β^2 are small. In the limit of large MME, the model effect and the interaction effect cancel out between models and thus don't contribute to $MSEP_{e-mean}$. Empirically, it is found that μ^2 is always relatively small, but this is not the case for σ_β^2. As a result there are several cases where e-mean has no skill.

Consider now the effect of the size of the MME. Eq. (3) shows that $MSEP_{e-mean}(n)$ decreases as 1/*n*, going from $\mu^2+\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2$ when there is a single model to $\mu^2+\sigma_\beta^2$ when there are infinitely many models. This assumes that models in the ensemble are chosen at random from the distribution of models. Figure 3 and Supplementary Figure S5 show how $MSEP_{e-mean}(n)$ decreases with the size of the MME, based on the estimated variance components and eq. 3. The results generalize the empirical results to prediction for the target population.

Eq. (3) also helps understand the empirical behavior of MSE of e-mean when the ensemble is built from successively worse models. Suppose that one starts from a sample of size *n* from some population P1 of models, for which MSEP of e-mean is

$$MSEP_{e-mean}(P1) = \mu^2_{(P1)} + \sigma^2_{\beta(P1)} + (1/n)\left(\sigma^2_{\alpha(P1)} + \sigma^2_{\gamma(P1)}\right) \quad (8)$$

To obtain an MME of size *n+1*, one must enlarge the sampled population to P2, with say

$$MSEP_{e-mean}(P2) = \mu^2_{(P2)} + \sigma^2_{\beta(P2)} + (1/(n+1))\left(\sigma^2_{\alpha(P2)} + \sigma^2_{\gamma(P2)}\right) \quad (9)$$

Since models are added in order of increasing MSEP, $\mu^2 + \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\gamma$ is larger for P2 than for P1. However, the contribution of the term $\sigma^2_\alpha + \sigma^2_\gamma$ is divided by n for P1 and by n+1 for P2, which can offset the increase in $\mu^2 + \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\gamma$, especially for small n. The empirical result is a minimum in MSE of e-mean for some value of n almost always larger than 1.

Discussion

There have been several publications that have documented the good performance of e-mean and e-median for crop models, including for the same data sets considered here (Asseng et al., 2014; Martre et al., 2015) and also for other crops than wheat (Bassu et al., 2014; Fleisher et al., 2017; Li et al., 2015; Rötter et al., 2012). However, here for the first time we analyze the results using MMEs for five different data sets, each representing a different range of environmental variability, in a common framework.

Empirical evidence is essential, but necessarily limited. It is important to complement the empirical evidence with theoretical results. The theoretical framework that we propose helps explain and generalize the empirical results. The framework assumes that there is some essentially infinite underlying distribution of crop models, from which the models in the MME are sampled at random. This assumption could be questioned, on the basis that there are in fact a limited number of existing crop models. However, it has been found that even crop models derived from the same underlying model but differing in parameterization can give quite different results (Folberth et al., 2016), implying that the number of effectively different crop models is in fact essentially infinite.

The theoretical results are based on variance components, which are simple to calculate. It may be worthwhile doing so systematically for MME studies, because the random effects model then provides a diagnostic tool for relating results to the characteristics of the MME and also a tool for extrapolating to the target population of environments and to different numbers of models.

The theoretical results all concern the simple mean of the values simulated by the individual models. It might be possible to improve the performance of e-mean by weighting different models depending on agreement with observations, using for example Bayesian model averaging (Raftery, Balabdaoui, Gneiting, & Polakowski, 2003). This is however difficult for crop models, because each environment involves growing a crop for a full season and as a consequence there are in general relatively few data available for estimating the weighting coefficients. Simple averaging is also often used for climate model ensembles (for example Wang et al., 2009).

The empirical results show that MSE of e-median and e-mean are always smaller than the average MSE of the individual models in the MME. This has also been observed with respect to climate models (Wang et al., 2009). The theoretical results show that this will always be true for MSEP of e-mean compared to MSEP averaged over models, for any size of the MME. The advantage of e-mean will increase as the ensemble size increases. Thus theory and empirical results agree that it is better (less prediction error) to use e-mean than a model chosen at random from the population of models, on average over the chosen model. The statistical basis for the superiority of e-mean is that the model and interaction effects cancel out between models. One possible modeling explanation could be that different models have different errors in the parameters, and averaging over models averages out the parameter errors. A similar mechanism has been suggested for climate models (Wang et al., 2009).

The empirical results show that e-median often has smaller MSE values than even the best individual model, and if not, it has an MSE value quite close to that of the best model. E-mean is not as highly ranked, but also is always close to the best MSE value. The theoretical results show that in the limit of a very large MME, MSEP of e-mean will be smaller than MSEP of the best model when squared bias is smaller than the variance of the interaction effect. The bias refers to error averaged over models, and thus bias contributes to MSEP of e-mean. An individual model however may have a model effect that is the negative of the bias, which is simply to say that the best individual model may have very small or zero error averaged over environments. Thus the existence of bias tends to make e-mean a worse predictor than the best model. A large interaction variance implies that model error is sometimes small, sometimes large for different environments. The average over models of the interaction term however tends to zero for large MMEs, for each environment. Thus the existence of interaction tends to make e-mean a better predictor than any model. Overall then, the relative values of squared bias and interaction variance determine whether there will be individual models better than e-mean.

Based on the estimated variance components, squared bias is smaller than the variance of the interaction effect for all the data sets and outputs considered here. Together, the empirical and theoretical results suggest that in a wide variety of cases, e-mean or e-median will be a better choice as predictor than any individual model, with e-median seeming to be empirically somewhat better than e-mean. The fact that the ensemble predictors out-perform most or all models not only for yield but also for protein, biomass and maximum LAI, suggests that they are useful not only for predicting final yield but also for prediction of the growth trajectory and quality of the crop.

The value of $MSEP_{e-mean}$ is not necessarily small; it is equal to the sum of squared bias and the variance of the environment effect. Since $MSEP_{e-mean}$ can be large, the skill of e-mean can be poor. It is thus essential to verify, for each application of crop models, that e-mean is indeed sufficiently skillful for the application intended. Model improvement, to the extent that it reduces bias and/or leads to models which track the effects of environment more closely (i.e. reduces the variance of the environment effect) will reduce $MSEP_{e-mean}$. Thus model improvement is not only important in its own right, but can also be a path to improved prediction by e-mean, as shown in (Maiorano et al., 2016) where improving wheat models by calibration and/or taking better account of heat stress improved prediction accuracy of e-median. Simply making models more similar, in the absence of improvement, reduces the variance of the model effect, but this does not reduce $MSEP_{e-mean}$. It is easy to show that according to the mixed model, the covariance between errors of two different models for a given environment is equal to σ_{β}^2, the variance of the environment effect. Thus, everything else being equal, the smaller the covariance (the less the model outputs are related), the smaller $MSEP_{e-mean}$ will be. The fact that bias is small for all the data sets here might be partially a consequence of calibration. The calibration data allow modelers to verify that their simulated values are close to reality for at least some environments.

The effect of number of models in a MME is of practical importance, and has received attention in several studies. For example, Li et al. (Li et al., 2015) suggested that eight models would be sufficient to obtain errors of e-mean below 10% of observed yield. The results here shed further light on this question. Our results indicate that the behavior of MSE_{e-mean} as a function of ensemble size depends on how the MME is created. If models are added at random, then $MSEP_{e-mean}(n)$ depends on *n*, the number of models, through the term $(\sigma_\alpha^2 + \sigma_\gamma^2)/n$, which decreases monotonically with n. In this case, a larger ensemble size always leads in expectation to a smaller value of $MSE_{e-mean}(n)$. Even going from 1 to 2 models is of interest, since it reduces that term by half. With five models, one obtains 80% of the potential improvement from adding more models. Note that the theoretical reduction in MSE_{e-mean} with n is in expectation, not for each sample of models. Wang et al. (2009) similarly found that improvement of a MME of climate models was very slight beyond 5-6 models.

If, instead of choosing models at random, one is capable of identifying the best models and builds the MME by successively adding models with larger prediction error, then the empirical results show that $MSE_{e-mean}(n)$ has a minimum at some small number of models, almost always greater than 1. That is, even if the best model is assumed to be known, it is almost always found to be advantageous to create at least a small MME by including less well-performing models. The theoretical results show that this is due to cancellation of errors between models which reduces the model effect and interaction contributions to $MSEP_{e-mean}(n)$. In this case it is not advantageous to make the MME as large as possible. Adding increasingly poorly performing models eventually increases $MSE_{e-mean}(n)$. To take advantage of this behavior, one would need to identify the best models (to be included in the MME) and/or the worst models (to be excluded). However, the empirical results show that identifying the best models can be very difficult, since all models had a wide range of rankings for fit to the observations. Thus actually creating an MME which contains only the best models or at least

avoids the worst models is a challenge. We examined here the rather simple strategy of adding models in inverse order of MSE. For climate models, it has been suggested that the optimal choice of models should take into account both the skill of the individual models (high skill better) and their degree of dependency (less dependency better) (Yoo & Kang, 2005).

The practical conclusion of this study is that predicting with e-mean or e-median of a fairly small MME of around five models which have been shown to be well-suited to the predictions of interest, will often be a good strategy. If the models are chosen in a way that is equivalent to choosing models at random, then this ensemble size captures, in expectation, most of the cancellation of errors that arises from having multiple models. If this includes only the best models, then this size is consistent with the number of models that empirically gives smallest error for e-mean.

While the emphasis here has been on ensemble predictors, it should be noted that there are other objectives of ensemble studies (Wallach, Mearns, Ruane, Rötter, & Asseng, 2016). A major objective is to obtain information on model uncertainty, based on the spread between models. Another important objective is to foster collaboration between modeling groups. Those objectives could lead to different considerations concerning ensemble size. Also, it is important to emphasize that using ensemble predictors is not a substitute for model improvement. Both model improvement and use of ensemble predictors, either singly or in combination, could contribute to extending the usefulness of crop models.

Acknowledgements

The authors acknowledge the Agricultural Model Intercomparison and Improvement Project (AgMIP) which led to the collaboration underlying this study.

References

Asseng, S., Ewert, F., Martre, P., Rosenzweig, C., Jones, J., Hatfield, J., … Wolf, J. (2016). Benchmark data set for wheat growth models: field experiments and AgMIP multi-model simulations. *Open Data Journal for Agricultural Research*, *1*(1). http://doi.org/10.18174/odjar.v1i1.14746

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., … Zhu, Y. (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, *5*(2), 143–147. http://doi.org/10.1038/nclimate2470

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., … Wolf, J. (2013). Uncertainty in simulating wheat yields under climate change. *Nature Climate Change*, *3*(9), 827–832. http://doi.org/10.1038/nclimate1916

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J. W., … Waha, K. (2014). How do various maize crop models vary in their responses to climate change factors? *Global Change Biology*, *20*(7), 2301–20. http://doi.org/10.1111/gcb.12520

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Chenu, K., Porter, J. R., Martre, P., Basso, B., Chapman, S. C., Ewert, F., … Asseng, S. (2017). Contribution of Crop Models to Adaptation in Wheat. *Trends in Plant Science*. http://doi.org/10.1016/j.tplants.2017.02.003

DelSole, T., Jia, L., Tippett, M. K., DelSole, T., Jia, L., & Tippett, M. K. (2013). Scale-Selective Ridge Regression for Multimodel Forecasting. *Journal of Climate*, *26*(20), 7957–7965. http://doi.org/10.1175/JCLI-D-13-00030.1

DelSole, T., Nattala, J., & Tippett, M. K. (2014). Skill improvement from increased ensemble size and model diversity. *Geophysical Research Letters*, *41*(20), 7331–7342. http://doi.org/10.1002/2014GL060133

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction

using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. http://doi.org/10.1016/j.advwatres.2006.11.014

Ewert, F., Rötter, R. P., Bindi, M., Webber, H., Trnka, M., Kersebaum, K. C., … Asseng, S. (2015). Crop modelling for integrated assessment of risk to food production from climate change. *Environmental Modelling & Software*, *72*, 287–303. http://doi.org/10.1016/j.envsoft.2014.12.003

Fleisher, D. H., Condori, B., Quiroz, R., Alva, A., Asseng, S., Barreda, C., … Woli, P. (2017). A potato model intercomparison across varying climates and productivity levels. *Global Change Biology*, *23*(3), 1258–1281. http://doi.org/10.1111/gcb.13411

Folberth, C., Elliott, J., Müller, C., Balkovic, J., Chryssanthacopoulos, J., Izaurralde, R. C., … Wang, X. (2016). Uncertainties in global crop model frameworks: effects of cultivar distribution, crop management and soil handling on crop yield estimates. *Biogeosciences Discussions*, 1–30. http://doi.org/10.5194/bg-2016-527

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi- model ensembles in seasonal forecasting–I. Basic concept. *Tellus A*, 219–233. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0870.2005.00103.x/full

Hasegawa, T., Li, T., Yin, X., Zhu, Y., Boote, K., Baker, J., … Zhu, J. (2017). Causes of variation among rice models in yield response to CO<inf>2</inf>examined with Free-Air CO<inf>2</inf>Enrichment and growth chamber experiments. *Scientific Reports*, *7*(1). http://doi.org/10.1038/s41598-017-13582-y

IPCC. (2014). Summary for policy makers. In C. B. Field, V. R. Barros, D. J. Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, … L. L. White (Eds.), *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1–32). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., … Bouman, B. (2015). Uncertainties in

predicting rice yield by current crop models under a wide range of climatic conditions. *Global Change Biology*, *21*(3), 1328–41. http://doi.org/10.1111/gcb.12758

Liu, B., Asseng, S., Müller, C., Ewert, F., Elliott, J., Lobell, D. B., … Zhu, Y. (2016). Similar estimates of temperature impacts on global wheat yield by three independent methods. *Nature Climate Change*, *6*(12). http://doi.org/10.1038/nclimate3115

Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., … Zhu, Y. (2016). Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research*. http://doi.org/10.1016/j.fcr.2016.05.001

Majoul-Haddad, T., Bancel, E., Martre, P., Triboi, E., & Branlard, G. (2013). Effect of short heat shocks applied during grain development on wheat (Triticum aestivum L.) grain proteome. *Journal of Cereal Science*, *57*(3), 486–495. http://doi.org/10.1016/j.jcs.2013.02.003

Martre, P., Reynolds, M. P., Asseng, S., Awer, F., Alderman, D. P., Cammarano, D. C., … Al., E. (2017). The International Heat Stress Genotype Experiment for modeling wheat response to heat: field experiments and AgMIP-Wheat multi-model simulations. *Open Data Journal for Agricultural Research*, *in press*.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., … Wolf, J. (2015). Multimodel ensembles of wheat growth: many models are better than one. *Global Change Biology*, *21*(2), 911–25. http://doi.org/10.1111/gcb.12768

O'Leary, G. J., Christy, B., Nuttall, J., Huth, N., Cammarano, D., Stöckle, C., … Asseng, S. (2015). Response of wheat growth, grain yield and water use to elevated CO_2 under a Free-Air CO_2 Enrichment (FACE) experiment and modelling in a semi-arid environment. *Global Change Biology*, *21*(7), 2670–2686. http://doi.org/10.1111/gcb.12830

Palosuo, T., Kersebaum, K. C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J. E., … Rötter, R. (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*(3), 103–114. http://doi.org/10.1016/j.eja.2011.05.001

Porter, J. R., Xie, L., Challinor, A. J., Cochrane, K., Howden, S. M., Iqbal, M. M., … Travasso, M. I. (2014). Food security and food production systems. In C. B. Field, V. R. Barros, D. J. Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, … L. L. White (Eds.), *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 485–533). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austrai: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Raftery, A. E., Balabdaoui, F., Gneiting, T., & Polakowski, M. (2003). *Using Bayesian Model Averaging to Calibrate Forecast Ensembles*. Retrieved from http://www.stat.washington.edu/www/research/reports/2003/tr440.pdf

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., … Jones, J. W. (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 of the United States of America*, *111*(9), 3268–73. http://doi.org/10.1073/pnas.1222463110

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., … Winter, J. M. (2013). The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. *Agricultural and Forest Meteorology*, *170*, 166–182. http://doi.org/10.1016/j.agrformet.2012.09.011

Rötter, R. P., Carter, T. R., Olesen, J. E., & Porter, J. R. (2011). Crop–climate models need an overhaul. *Nature Climate Change*, *1*(4), 175–177. http://doi.org/10.1038/nclimate1152

Rötter, R. P., Palosuo, T., Kersebaum, K. C., Angulo, C., Bindi, M., Ewert, F., … Trnka, M. (2012). Simulation of spring barley yield in different climatic zones of Northern and Central Europe: A comparison of nine crop models. *Field Crops Research*, *133*, 23–36. http://doi.org/10.1016/j.fcr.2012.03.016

Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley & Sons.

Solazzo, E., & Galmarini, S. (2015). A science-based use of ensembles of opportunities for assessment and scenario studies. *Atmospheric Chemistry and Physics*, *15*(5), 2535–2544. http://doi.org/10.5194/acp-15-2535-2015

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *365*(1857), 2053–75. http://doi.org/10.1098/rsta.2007.2076

Wallach, D., Mearns, L. O., Ruane, A. C., Rötter, R. P., & Asseng, S. (2016). Lessons from climate modeling on the design and use of ensembles for crop modeling. *Climatic Change*, *139*(3–4), 551–564. http://doi.org/10.1007/s10584-016-1803-1

Wang, B., Lee, J.-Y., Kang, I.-S., Shukla, J., Park, C.-K., Kumar, A., … Yamagata, T. (2009). Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dynamics*, *33*(1), 93–117. http://doi.org/10.1007/s00382-008-0460-0

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260. http://doi.org/10.1002/qj.210

Yin, X., Kersebaum, K. C., Kollas, C., Baby, S., Beaudoin, N., Manevski, K., … Olesen, J. E. (2017). Multi-model uncertainty analysis in predicting grain N for crop rotations in Europe. *European Journal of Agronomy*, *84*, 152–165. http://doi.org/10.1016/j.eja.2016.12.009

Yoo, J. H., & Kang, I.-S. (2005). Theoretical examination of a multi-model composite for seasonal prediction. *Geophysical Research Letters*, *32*(18), n/a-n/a. http://doi.org/10.1029/2005GL023513

Table 1

	Environments	Data furnished for Calibration	References
AgMIP-Wheat Pilot (4)	Four global sites, corresponding to four different mega-environments. 3 spring cultivars (Gamenya, HD 2009, and Oasis), 1 winter cultivar (Arminda) Yields 2.5-7.5 t ha^{-1}	Anthesis and maturity date, all environments	Asseng et al. (2016); Martre et al. (2015)
HSC (15)	Maricopa, Arizona. Gradient of mean growing season temperature from 15.0°C to 33.4°C created by varying sowing date and artificial heating. 1 spring cultivar (Yecora Rojo) Yields 0-8 t ha^{-1}	Detailed crop measurements for one environment (average temperature of 15.4°C). Phenology parameters used previously in one model.	Asseng et al. (2014)
HSGE (34)	6 high temperature global sites, two years, one or two planting dates. Number of days with Tmax>31°C ranged from 28 to 74. 2 spring cultivars (Bacanora 88 and Nesser) Yields 1.9-8.0 t ha^{-1}	Detailed crop measurements for four environments at one location (Obregan, Mexico). Anthesis and maturity dates for all other environments.	Asseng et al. (2014); Martre et al. (2017)
C3-GEM (10)	Control and heat shock environments in outdoor controlled environment chambers. Heat shock of Tmax=38°C for 4 hours for 2 or 4 days during the lag or linear grain filling period or both. 1 winter cultivar (Récital) Yields 5.6-8.4 t ha^{-1}	Detailed crop measurements for the 3 control environments.	Majoul-Haddad, Bancel, Martre, Triboi, & Branlard (2013)

AGFACE (18)	Elevated free air CO_2 concentration experiment, over three years, early or late sowing, CO_2 concentrations of 385 or 550 ppm, rain-fed or irrigated. 1 spring cultivar (Yitpi) Yields 1.1-4.6 t ha^{-1}	Detailed crop measurements for one environment (385 ppm $C0_2$, early sowing, irrigated). Parameters used previously in 6 models. .	O'Leary et al. (2015)

Table 1.

Data sets. The five wheat data sets that provided the empirical evidence. *The number of environments in the data set is given in parentheses.

Figure legends

Figure 1.

RMSE relative to RMSE of e-median ($RMSE_{model}$-$RMSE_{e\text{-}median}$) for each data set. A negative value means that the model has smaller RMSE than e-median. The two letter codes represent different crop models, see Table S1 for model identification information. “ybar” refers to the predictor that uses the same predicted value, equal to the average of observed values for the data set, for all environments. Models with relative RMSE values larger than "ybar" have no skill. Relative RMSE for “ave” is obtained by averaging MSE over all individual models, taking the square root and subtracting $RMSE_{e\text{-}median}$.

Figure 2

Fit of models to HSC yield data. Each environment number corresponds to a different sowing date, either without (“C”) or with (“H”) supplementary heating. Solid diamonds are observed yields. Circles and triangles show respectively e-mean and e-median. Values simulated by the 25 individual models are connected by thin dotted lines.

Figure 3.

Effect of ensemble size on root mean squared error (RMSE) of e-mean for yield. Left panel. Effect of ensemble size on RMSE of e-mean for yield when models are chosen at random. Each point is the RMSE of e-mean averaged over 100 samples of n (n=1,...,25) models drawn at random, without replacement, from the models of the original MME. The lines are based on equation 3, using the variance components estimated for each data set. Right panel. Effect of ensemble size on RMSE of e-mean for yield when models are added from best (smallest RMSE) to worst.