

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23

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<sup>1,\*</sup>, P. Martre<sup>2</sup>, B. Liu<sup>3,4</sup>, S. Asseng<sup>4</sup>, F. Ewert<sup>5,6</sup>, P.J. Thorburn<sup>7</sup>, M. van Ittersum<sup>8</sup>, P.K. Aggarwal<sup>9,†</sup>, M. Ahmed<sup>10,11</sup>, B. Basso<sup>12,13</sup>, C. Biernath<sup>14</sup>, D. Cammarano<sup>15</sup>, A.J. Challinor<sup>16,17</sup>, G. De Sanctis<sup>18,‡</sup>, B. Dumont<sup>19</sup>, E. Eyshi Rezaei<sup>5,20</sup>, E. Fereres<sup>21</sup>, G.J. Fitzgerald<sup>22,23</sup>, Y. Gao<sup>4</sup>, M. Garcia-Vila<sup>21</sup>, S. Gayler<sup>24</sup>, C. Girousse<sup>25</sup>, G. Hoogenboom<sup>4,26</sup>, H. Horan<sup>7</sup>, R.C. Izaurralde<sup>27,28</sup>, C.D. Jones<sup>28</sup>, B.T. Kassie<sup>4</sup>, K.C. Kersebaum<sup>29</sup>, C. Klein<sup>30</sup>, A.K.

This is the author manuscript accepted for publication and has 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](https://doi.org/10.1111/gcb.14411)

24 Koehler<sup>16</sup>, A. Maiorano<sup>2,31</sup>, S. Minoli<sup>32</sup>, C. Müller<sup>32</sup>, S. Naresh Kumar<sup>33</sup>, C. Nendel<sup>29</sup>, G.J.  
25 O’Leary<sup>34</sup>, T. Palosuo<sup>35</sup>, E. Priesack<sup>30</sup>, D. Ripoche<sup>36</sup>, R.P. Rötter<sup>37,38</sup>, M.A. Semenov<sup>39</sup>, C.  
26 Stöckle<sup>10</sup>, P. Stratonovitch<sup>39</sup>, T. Streck<sup>24</sup>, I. Supit<sup>40</sup>, F. Tao<sup>41,35</sup>, J. Wolf<sup>42</sup>, and Z. Zhang<sup>43</sup>  
27

## 28 **Affiliations**

29 <sup>1</sup>UMR AGIR, INRA 31326 Castanet-Tolosan, France.

30 <sup>2</sup>UMR LEPSE, INRA, Montpellier SupAgro, 34 060, Montpellier, France.

31 <sup>3</sup>National Engineering and Technology Center for Information Agriculture, Key Laboratory for  
32 Crop System Analysis and Decision Making, Ministry of Agriculture, Jiangsu Key Laboratory  
33 for Information Agriculture, Jiangsu Collaborative Innovation Center for Modern Crop  
34 Production, Nanjing Agricultural University, Nanjing, Jiangsu 210095, P. R. China.

35 <sup>4</sup>Agricultural & Biological Engineering Department, University of Florida, Gainesville, FL  
36 32611, USA.

37 <sup>5</sup>Institute of Crop Science and Resource Conservation INRES, University of Bonn, 53115,  
38 Germany.

39 <sup>6</sup>Leibniz Centre for Agricultural Landscape Research, 15374 Müncheberg, Germany.

40 <sup>7</sup>CSIRO Agriculture and Food, Brisbane, St Lucia Queensland 4067, Australia.

41 <sup>8</sup>Plant Production Systems Group, Wageningen University, 6700 AK Wageningen, The  
42 Netherlands.

43 <sup>9</sup>CGIAR Research Program on Climate Change, Agriculture and Food Security, BISA-  
44 CIMMYT, New Delhi-110012, India.

45 <sup>10</sup>Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120.

46 <sup>11</sup>Department of Agronomy, Pir Mehr Ali Shah Arid Agriculture University Rawalpindi-46300,  
47 Pakistan.

48 <sup>12</sup>Department of Earth and Environmental Sciences, Michigan State University, East Lansing,  
49 Michigan 48823, USA.

50 <sup>13</sup>W.K. Kellogg Biological Station, Michigan State University East Lansing, Michigan 48823,  
51 USA.

- 52 <sup>14</sup>Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research  
53 Center for Environmental Health, Neuherberg, 85764, Germany.
- 54 <sup>15</sup>James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK.
- 55 <sup>16</sup>Institute for Climate and Atmospheric Science, School of Earth and Environment, University of  
56 Leeds, Leeds LS29JT, UK.
- 57 <sup>17</sup>CGIAR-ESSP Program on Climate Change, Agriculture and Food Security, International  
58 Centre for Tropical Agriculture (CIAT), A.A. 6713, Cali, Colombia.
- 59 <sup>18</sup>European Food Safety Authority, GMO Unit, Via Carlo Magno 1A, Parma, IT-43126, Italy.
- 60 <sup>19</sup>Department Terra & AgroBioChem, Gembloux Agro-Bio Tech, University of Liege,  
61 Gembloux 5030, Belgium.
- 62 <sup>20</sup>Center for Development Research (ZEF), Walter-Flex-Straße 3, 53113 Bonn, Germany.
- 63 <sup>21</sup>IAS-CSIC and University of Cordoba, Apartado 3048, 14080 Cordoba, Spain.
- 64 <sup>22</sup>Agriculture Victoria Research, Department of Economic Development, Jobs, Transport and  
65 Resources, Ballarat, Victoria 3350 Australia.
- 66 <sup>23</sup>Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, 4 Water Street,  
67 Creswick, VIC 3363, Australia.
- 68 <sup>24</sup>Institute of Soil Science and Land Evaluation, University of Hohenheim, 70599 Stuttgart,  
69 Germany.
- 70 <sup>25</sup>UMR GDEC, INRA, Université Clermont Auvergne, 63000, Clermont-Ferrand, France.
- 71 <sup>26</sup>Institute for Sustainable Food Systems, University of Florida, Gainesville, FL 32611, USA.
- 72 <sup>27</sup>Department of Geographical Sciences, Univ. of Maryland, College Park, MD 20742, USA.
- 73 <sup>28</sup>Texas A&M AgriLife Research and Extension Center, Texas A&M Univ., Temple, TX 76502,  
74 USA.
- 75 <sup>29</sup>Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research,  
76 15374 Müncheberg, Germany.
- 77 <sup>30</sup>Institute of Biochemical Plant Pathology, Helmholtz Zentrum München-German Research  
78 Center for Environmental Health, Neuherberg, 85764, Germany.

79 <sup>31</sup>Present address:- European Food Safety Authority – EFSA, via Carlo Magno 1/A, 43126 Parma  
80 - Italy

81 <sup>32</sup>Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany.

82 <sup>33</sup>Centre for Environment Science and Climate Resilient Agriculture, Indian Agricultural  
83 Research Institute, IARI PUSA, New Delhi 110 012, India.

84 <sup>34</sup>Grains Innovation Park, Agriculture Victoria Research, Department of Economic  
85 Development, Jobs, Transport and Resources, Horsham 3400, Australia.

86 <sup>35</sup>Natural Resources Institute Finland (Luke), 00790 Helsinki, Finland.

87 <sup>36</sup>US AgroClim, INRA, 84 914 Avignon, France.

88 <sup>37</sup>University of Göttingen, Tropical Plant Production and Agricultural Systems Modelling  
89 (TROPAGS), Grisebachstraße 6, 37077 Göttingen.

90 <sup>38</sup>University of Göttingen, Centre of Biodiversity and Sustainable Land Use (CBL), Buesgenweg  
91 1, 37077 Göttingen, Germany.

92 <sup>39</sup>Computational and Systems Biology Department, Rothamsted Research, Harpenden, Herts,  
93 AL5 2JQ, UK.

94 <sup>40</sup>Water & Food and Water Systems & Global Change Group, Wageningen University, 6700AA  
95 Wageningen, The Netherlands.

96 <sup>41</sup>Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of  
97 Science, Beijing 100101, China.

98 <sup>42</sup>Plant Production Systems, Wageningen University, 6700AA Wageningen, The Netherlands.

99 <sup>43</sup>State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of  
100 Geographical Science, Beijing Normal University, Beijing, China.

101

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

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

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

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

110       Paper type: Primary research

111

## 112 **Abstract**

113       A recent innovation in assessment of climate change impact on agricultural production  
114 has been to use crop multi model ensembles (MMEs). These studies usually find large variability  
115 between individual models but that the ensemble mean (e-mean) and median (e-median) often  
116 seem to predict quite well. However few studies have specifically been concerned with the  
117 predictive quality of those ensemble predictors. We ask what is the predictive quality of e-mean  
118 and e-median, and how does that depend on the ensemble characteristics. Our empirical results  
119 are based on five MME studies applied to wheat, using different data sets but the same 25 crop  
120 models . We show that the ensemble predictors have quite high skill and are better than most and  
121 sometimes all individual models for most groups of environments and most response variables.  
122 Mean squared error of e-mean decreases monotonically with the size of the ensemble if models  
123 are added at random, but has a minimum at usually 2-6 models if best-fit models are added first.  
124 Our theoretical results describe the ensemble using four parameters; average bias, model effect  
125 variance, environment effect variance and interaction variance. We show analytically that mean  
126 squared error of prediction (MSEP) of e-mean will always be smaller than MSEP averaged over  
127 models, and will be less than MSEP of the best model if squared bias is less than the interaction  
128 variance. If models are added to the ensemble at random, MSEP of e-mean will decrease as the  
129 inverse of ensemble size, with a minimum equal to squared bias plus interaction variance. This  
130 minimum value is not necessarily small, and so it is important to evaluate the predictive quality  
131 of e-mean for each target population of environments. These results provide new information on  
132 the advantages of ensemble predictors, but also show their limitations.

## 133 Introduction

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

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

153 Crop MME studies have often noted that the ensemble mean (e-mean) and ensemble  
154 median (e-median) of simulated values give good agreement with observations (Bassu et al.,  
155 2014; Palosuo et al., 2011; Rötter et al., 2012). This suggests that in practice, it might be better to  
156 create a MME and then use the predictions of e-mean or e-median rather than use the predictions  
157 of an individual model. Several recent impact assessment studies have based conclusions on  
158 ensemble predictors (Asseng et al., 2014; Liu et al., 2016).

159 Only a few studies have examined the properties of crop MME predictors in more detail,  
160 in each case for one set of environmental conditions. One study, based on prediction of multiple  
161 response variables in four environments, found that e-mean and e-median were both better than  
162 the best model, for a composite criterion including all outputs and environments (Pierre Martre et

163 al., 2015). Yin et al. (2017) found that e-mean predicted grain N better than a randomly chosen  
164 model. Of particular practical interest is the behavior of e-mean and e-median as a function of  
165 the number of models in the ensemble. This has been studied by treating the ensemble as the full  
166 population of models, and drawing sub samples from that population. The conclusions have been  
167 that prediction error decreases systematically as the number of models increases. Li et al. (2015)  
168 suggested that eight models would be sufficient to obtain errors of e-mean below 10% of  
169 observed yield. All of these studies have been empirical, based on a single MME study. The  
170 general behavior of crop ensemble predictors has not been addressed. Studies in other fields,  
171 including group intelligence (Surowiecki, 2005), hydrologic modeling (Duan, Ajami, Gao, &  
172 Sorooshian, 2007), air quality modeling (Solazzo & Galmarini, 2015) and climate modeling  
173 (Tebaldi & Knutti, 2007) have also found that averaging over multiple opinions or solutions can  
174 give good predictions, often better than any individual model. The basis for using MME  
175 predictors has received particular attention in the field of climate modeling (Hagedorn et al.,  
176 2005; Weigel et al., 2008). However, the context there is quite different than for crop models; for  
177 example in climate modeling each MME member is often itself an ensemble based on a single  
178 model with different initial conditions (DelSole, Nattala, & Tippet, 2014) whereas in crop  
179 modeling, each model normally provides a single simulation, a major interest in climate  
180 modeling is in probabilistic predictions rather than the deterministic predictions of crop models  
181 (DelSole et al., 2013; Wang et al., 2009) and in climate modeling spatial patterns of prediction  
182 play an important role (DelSole et al., 2013).

183 One can easily imagine situations where e-mean and e-median for crop models do not  
184 predict well. For example, if all models have large positive bias, then e-mean and e-median will  
185 also have large positive bias, and e-median will be worse than half the models. Thus, one cannot  
186 automatically assume that one will obtain reliable predictions by using MME predictors. The  
187 question we ask then is what is the predictive quality of e-mean and e-median, and how does that  
188 depend on the ensemble characteristics? We break this down into specific sub-questions. First,  
189 how does the predictive quality of MME predictors compare to predictive quality of a model  
190 chosen at random from the models in the ensemble, or to that of the best individual model in the  
191 ensemble, and how does that depend on the ensemble characteristics? The answer to this  
192 question affects the choice between using an individual model and a MME predictor. Second,

193 what is the level of error of the MME predictors? This is a major determinant of the potential  
194 usefulness of these predictors. Finally, how does the level of error of the MME predictors depend  
195 on the number of models in the ensemble? This affects the very practical decision as to the  
196 number of models to include in a MME.

## 197 **Materials and Methods**

### 198 **Data**

199 The data sets simulated in the five wheat MME studies considered here are described in  
200 Table 1. Details are available in the cited references. Each data set concerns a different range of  
201 environmental conditions, where an environment is to be understood as a combination of  
202 physical environment and management. We consider each data set as representative of some  
203 infinite range of environments, the target population. The target population corresponding to the  
204 AgMIP wheat pilot data set is worldwide wheat environments. The data set is a sample from that  
205 population, and the prediction problem is prediction for a randomly chosen individual  
206 environment from that population. In the case of the HSC data set, the target population of  
207 environments is considered to be all possible weather sequences for wheat in Maricopa,  
208 Arizona, generated by different years and planting dates. The data set can be considered a sample  
209 from that distribution of environments, where the heat treatments are meant to increase  
210 artificially the diversity of the sampled conditions. In the case of the HSGE data set, the target  
211 population of environments is taken to be worldwide hot environments for wheat, including all  
212 possible weather sequences and all locations. The target population for the C3-GEM data set is  
213 taken to be all possible weather sequences at the location of the study, with or without heat  
214 shocks during grain filling. Finally, the target population corresponding to the AGFACE data set  
215 is considered to be wheat crops under different weather sequences at the location of the study,  
216 with or without irrigation and with either current or enhanced CO<sub>2</sub> levels. We consider here four  
217 output variables that were measured in most or all of these studies: grain yield (yield), grain  
218 protein concentration (protein), final aboveground biomass (biomass) and maximum leaf area  
219 index during the course of growth (maximum LAI).



## 220 **Models and calibration**

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

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

239

## 240 **Evaluation metrics**

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

$$243 \quad MSE = 1 / N \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

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

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

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

252  
253 where  $MSE_{model}$  is MSE for the model in question and  $MSE_{\bar{y}}$  is MSE when all predictions use  
254 the average of observed values for that data set ( $\bar{y}$ ). Since  $\bar{y}$  is a constant, it explains none of  
255 the variability in the data set. A perfect model has  $EF=1$ . A model that does worse than  $\bar{y}$  has  
256  $EF < 0$  and can be considered to have no skill in explaining variability between environments.

257 The above criteria refer to the data in the data set. As a criterion of prediction accuracy  
258 for the target population we use mean squared error of prediction (MSEP), defined as the  
259 expectation of squared error over the target population. It is well known that if the same data are  
260 used for calibration and for evaluation, MSE tends to under-estimate MSEP. To examine how  
261 important this is, we calculated MSE for yield, using either all environments or leaving out all  
262 those environments which provided yield for calibration. The resulting MSE values for e-mean  
263 and e-median, and their ranks among all models, were very similar (Supplementary Table S2).  
264 We therefore use MSE based on all environments of a data set as an estimate of MSEP for the  
265 corresponding target population.

266

## 267 **Statistical description of multi-model ensemble**

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

$$269 \quad e_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (1)$$

270 where  $e_{ij}$  is error (observed value for environment  $j$  minus value simulated by model  $i$ ),  $\mu$  is  
271 the overall bias (error averaged over models and environments),  $\alpha_i$  is a random model effect  
272 with mean 0 and variance  $\sigma_\alpha^2$ ,  $\beta_j$  is a random environment effect with mean 0 and variance  $\sigma_\beta^2$   
273 and  $\gamma_{ij}$  is the random interaction term, with mean 0 and variance  $\sigma_\gamma^2$  (Scheffé, 1959). Thus the

274 random effects model characterizes a MME and target population using four parameters:  $\mu$ ,  $\sigma_\alpha^2$ ,  
275  $\sigma_\beta^2$  and  $\sigma_\gamma^2$ .

276 If there is bias, this implies that predictions, averaged over models and environments, are  
277 too small or too large. For example, if models tended to underestimate potential yield for the  
278 cultivars of the HSGE data set, this could lead on the average to systematic under-prediction of  
279 yield and therefore to a positive bias. The bias term contributes equally to all individual models  
280 and therefore also to e-mean, for all environments of the target population. The model effect  
281 indicates to what extent a specific model over- or under- predicts, on the average over  
282 environments. The larger  $\sigma_\alpha^2$ , the larger the variability between errors of different models. The  
283 environment effect indicates to what extent there is over- or under-prediction for individual  
284 environments, averaged over models. For example, if all models tended to over-predict  
285 specifically for the highest temperatures of the HSC target population, this would lead to an  
286 environment effect. The larger  $\sigma_\beta^2$ , the larger the variability between errors for different  
287 environments. Finally, the interaction effect measures the effect of interaction between a  
288 specific model and a specific environment on model error.

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

297

## 298 Results

### 299 Empirical results

300 Figure 1 shows RMSE relative to e-median ( $RMSE_{\text{model}} - RMSE_{\text{e-median}}$ ) for yield for each  
301 model and each data set. Models with negative values have smaller RMSE than e-median. It is

302 seen that e-median is better than all individual models (all individual models have positive values  
303 of RMSE relative to e-median) except for the HSGE and AGFACE studies, where there are  
304 respectively four and two individual models out of 25 that are better than e-median. E-mean is  
305 slightly worse than e-median (slightly positive RMSE relative to e-median) except for the HSGE  
306 data set. Its worst ranking for yield is seventh (among the 25 individual models, e-mean and e-  
307 median) . For protein, biomass and maximum LAI, the rankings of e-median and e-mean are  
308 more variable. At worst e-median is ranked sixth and e-mean tenth. E-median is better than e-  
309 mean in 13 out of the 17 combinations of data set and output variable (Supplementary Figures  
310 S1-S3). Figure 2 shows as an example the fit of e-mean, e-median and the individual models to  
311 the HSC yield data.

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

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

330 Figure 3 shows empirical results for the effect of number of models on MSE of e-mean,  
331 for predicting yield. These results are averages over multiple choices of models, and correspond

332 to choosing the models to add to the ensemble at random. There is an almost monotonic decrease  
333 in MSE as more models are added to the ensemble. Similar behavior is exhibited for the other  
334 output variables (Supplementary Figure S5).

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

### 341 Theoretical results

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

353 MSE of e-mean based on a MME of size n is

$$354 \quad MSE_{e\text{-mean}}(n) = E \left\{ \left[ \mu + (1/n) \sum_{i=1}^n \alpha_i + \beta_j + (1/n) \sum_{i=1}^n \gamma_{ij} \right]^2 \right\} \quad (2)$$

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

$$356 \quad MSE_{e\text{-mean}}(n) = \mu^2 + \sigma_\alpha^2 / n + \sigma_\beta^2 + \sigma_\gamma^2 / n \quad (3)$$

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

358 
$$MSEP_{e-mean} = \mu^2 + \sigma_\beta^2 \quad (4)$$

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

360 
$$\overline{MSEP} = E\left\{\left[\mu + \alpha_i + \beta_j + \gamma_{ij}\right]^2\right\} = \mu^2 + \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 \quad (5)$$

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

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

367 
$$E\left[(\mu + \alpha_i + \beta_j + \gamma_{ij} | \alpha_i = a)^2\right] = (\mu + a)^2 + \sigma_\beta^2 + \sigma_\gamma^2 \quad (6)$$

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

370 
$$P\left[(\mu + a)^2 + \sigma_\beta^2 + \sigma_\gamma^2 \leq \mu^2 + \sigma_\beta^2\right] = P\left[a' \leq (\mu^2 - \sigma_\gamma^2) / \sigma_\alpha^2\right] \quad (7)$$

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

380 Equation (4) shows that  $MSEP_{emean}$  is not necessarily small, even in the limit of a very  
 381 large MME. It will only be small if both  $\mu^2$  and  $\sigma_\beta^2$  are small. In the limit of large MME, the

382 model effect and the interaction effect cancel out between models and thus don't contribute to  
 383  $MSEP_{e-mean}$ . Empirically, it is found that  $\mu^2$  is always relatively small, but this is not the case for  
 384  $\sigma_\beta^2$ . As a result there are several cases where e-mean has no skill.

385 Consider now the effect of the size of the MME. Eq. (3) shows that  $MSEP_{e-mean}(n)$   
 386 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  
 387 there are infinitely many models. This assumes that models in the ensemble are chosen at  
 388 random from the distribution of models. Figure 3 and Supplementary Figure S5 show how  
 389  $MSEP_{e-mean}(n)$  decreases with the size of the MME, based on the estimated variance components  
 390 and eq. 3. The results generalize the empirical results to prediction for the target population.

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

$$394 \quad MSEP_{e-mean}(P1) = \mu_{(P1)}^2 + \sigma_{\beta(P1)}^2 + (1/n)(\sigma_{\alpha(P1)}^2 + \sigma_{\gamma(P1)}^2) \quad (8)$$

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

$$396 \quad MSEP_{e-mean}(P2) = \mu_{(P2)}^2 + \sigma_{\beta(P2)}^2 + (1/(n+1))(\sigma_{\alpha(P2)}^2 + \sigma_{\gamma(P2)}^2) \quad (9)$$

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

401

## 402 Discussion

403 There have been several publications that have documented the good performance of e-  
 404 mean and e-median for crop models, including for the same data sets considered here (Asseng et  
 405 al., 2014; Martre et al., 2015) and also for other crops than wheat (Bassu et al., 2014; Fleisher et  
 406 al., 2017; Li et al., 2015; Rötter et al., 2012). However, here for the first time we analyze the

407 results using MMEs for five different data sets, each representing a different range of  
408 environmental variability, in a common framework.

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

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

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

431 The empirical results show that MSE of e-median and e-mean are always smaller than the  
432 average MSE of the individual models in the MME. This has also been observed with respect to  
433 climate models (Wang et al., 2009). The theoretical results show that this will always be true for  
434 MSE of e-mean compared to MSE averaged over models, for any size of the MME. The  
435 advantage of e-mean will increase as the ensemble size increases. Thus theory and empirical  
436 results agree that it is better (less prediction error) to use e-mean than a model chosen at random



437 from the population of models, on average over the chosen model. The statistical basis for the  
438 superiority of e-mean is that the model and interaction effects cancel out between models. One  
439 possible modeling explanation could be that different models have different errors in the  
440 parameters, and averaging over models averages out the parameter errors. A similar mechanism  
441 has been suggested for climate models (Wang et al., 2009).

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

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

464 The value of  $MSEP_{e-mean}$  is not necessarily small; it is equal to the sum of squared bias and  
465 the variance of the environment effect. Since  $MSEP_{e-mean}$  can be large, the skill of e-mean can be  
466 poor. It is thus essential to verify, for each application of crop models, that e-mean is indeed

467 sufficiently skillful for the application intended. Model improvement, to the extent that it  
468 reduces bias and/or leads to models which track the effects of environment more closely (i.e.  
469 reduces the variance of the environment effect) will reduce  $MSEP_{e-mean}$ . Thus model  
470 improvement is not only important in its own right, but can also be a path to improved prediction  
471 by e-mean, as shown in (Maiorano et al., 2016) where improving wheat models by calibration  
472 and/or taking better account of heat stress improved prediction accuracy of e-median. Simply  
473 making models more similar, in the absence of improvement, reduces the variance of the model  
474 effect, but this does not reduce  $MSEP_{e-mean}$ . It is easy to show that according to the mixed model,  
475 the covariance between errors of two different models for a given environment is equal to  $\sigma_\beta^2$ ,  
476 the variance of the environment effect. Thus, everything else being equal, the smaller the  
477 covariance (the less the model outputs are related), the smaller  $MSEP_{e-mean}$  will be. The fact that  
478 bias is small for all the data sets here might be partially a consequence of calibration. The  
479 calibration data allow modelers to verify that their simulated values are close to reality for at  
480 least some environments.

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

493 If, instead of choosing models at random, one is capable of identifying the best models  
494 and builds the MME by successively adding models with larger prediction error, then the

495 empirical results show that  $MSE_{e-mean}(n)$  has a minimum at some small number of models,  
496 almost always greater than 1. That is, even if the best model is assumed to be known, it is almost  
497 always found to be advantageous to create at least a small MME by including less well-  
498 performing models. The theoretical results show that this is due to cancellation of errors  
499 between models which reduces the model effect and interaction contributions to  $MSEP_{e-mean}(n)$ .  
500 In this case it is not advantageous to make the MME as large as possible. Adding increasingly  
501 poorly performing models eventually increases  $MSE_{e-mean}(n)$ . To take advantage of this  
502 behavior, one would need to identify the best models (to be included in the MME) and/or the  
503 worst models (to be excluded). However, the empirical results show that identifying the best  
504 models can be very difficult, since all models had a wide range of rankings for fit to the  
505 observations. Thus actually creating an MME which contains only the best models or at least  
506 avoids the worst models is a challenge. We examined here the rather simple strategy of adding  
507 models in inverse order of MSE. For climate models, it has been suggested that the optimal  
508 choice of models should take into account both the skill of the individual models (high skill  
509 better) and their degree of dependency (less dependency better) (Yoo & Kang, 2005).

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

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

525

## 526 **Acknowledgements**

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

529

530

## 531 **References**

- 532 Asseng, S., Ewert, F., Martre, P., Rosenzweig, C., Jones, J., Hatfield, J., ... Wolf, J. (2016).  
533 Benchmark data set for wheat growth models: field experiments and AgMIP multi-model  
534 simulations. *Open Data Journal for Agricultural Research*, 1(1).  
535 <http://doi.org/10.18174/odjar.v1i1.14746>
- 536 Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., ... Zhu, Y.  
537 (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, 5(2),  
538 143–147. <http://doi.org/10.1038/nclimate2470>
- 539 Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., ... Wolf, J.  
540 (2013). Uncertainty in simulating wheat yields under climate change. *Nature Climate*  
541 *Change*, 3(9), 827–832. <http://doi.org/10.1038/nclimate1916>
- 542 Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J. W., ... Waha, K. (2014).  
543 How do various maize crop models vary in their responses to climate change factors?  
544 *Global Change Biology*, 20(7), 2301–20. <http://doi.org/10.1111/gcb.12520>
- 545 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models  
546 Using lme4. *Journal of Statistical Software*, 67(1), 1–48.  
547 <http://doi.org/10.18637/jss.v067.i01>
- 548 Chenu, K., Porter, J. R., Martre, P., Basso, B., Chapman, S. C., Ewert, F., ... Asseng, S. (2017).  
549 Contribution of Crop Models to Adaptation in Wheat. *Trends in Plant Science*.  
550 <http://doi.org/10.1016/j.tplants.2017.02.003>
- 551 DelSole, T., Jia, L., Tippet, M. K., DelSole, T., Jia, L., & Tippet, M. K. (2013). Scale-Selective  
552 Ridge Regression for Multimodel Forecasting. *Journal of Climate*, 26(20), 7957–7965.

553 <http://doi.org/10.1175/JCLI-D-13-00030.1>

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

557 Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic  
558 prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–  
559 1386. <http://doi.org/10.1016/j.advwatres.2006.11.014>

560 Ewert, F., Rötter, R. P., Bindi, M., Webber, H., Trnka, M., Kersebaum, K. C., ... Asseng, S.  
561 (2015). Crop modelling for integrated assessment of risk to food production from climate  
562 change. *Environmental Modelling & Software*, *72*, 287–303.  
563 <http://doi.org/10.1016/j.envsoft.2014.12.003>

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

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

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

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

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

584 Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., ... Bouman, B. (2015).  
585 Uncertainties in predicting rice yield by current crop models under a wide range of climatic  
586 conditions. *Global Change Biology*, 21(3), 1328–41. <http://doi.org/10.1111/gcb.12758>  
587 Liu, B., Asseng, S., Müller, C., Ewert, F., Elliott, J., Lobell, D. B., ... Zhu, Y. (2016). Similar  
588 estimates of temperature impacts on global wheat yield by three independent methods.  
589 *Nature Climate Change*, 6(12). <http://doi.org/10.1038/nclimate3115>  
590 Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., ... Zhu, Y. (2016).  
591 Crop model improvement reduces the uncertainty of the response to temperature of multi-  
592 model ensembles. *Field Crops Research*. <http://doi.org/10.1016/j.fcr.2016.05.001>  
593 Majoul-Haddad, T., Bancel, E., Martre, P., Triboi, E., & Branlard, G. (2013). Effect of short heat  
594 shocks applied during grain development on wheat (*Triticum aestivum* L.) grain proteome.  
595 *Journal of Cereal Science*, 57(3), 486–495. <http://doi.org/10.1016/j.jcs.2013.02.003>  
596 Martre, P., Reynolds, M. P., Asseng, S., Awer, F., Alderman, D. P., Cammarano, D. C., ... Al.,  
597 E. (2017). The International Heat Stress Genotype Experiment for modeling wheat response  
598 to heat: field experiments and AgMIP-Wheat multi-model simulations. *Open Data Journal*  
599 *for Agricultural Research*, in press.  
600 Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., ... Wolf, J. (2015).  
601 Multimodel ensembles of wheat growth: many models are better than one. *Global Change*  
602 *Biology*, 21(2), 911–25. <http://doi.org/10.1111/gcb.12768>  
603 O'Leary, G. J., Christy, B., Nuttall, J., Huth, N., Cammarano, D., Stöckle, C., ... Asseng, S.  
604 (2015). Response of wheat growth, grain yield and water use to elevated CO<sub>2</sub> under a Free-  
605 Air CO<sub>2</sub> Enrichment (FACE) experiment and modelling in a semi-arid environment.  
606 *Global Change Biology*, 21(7), 2670–2686. <http://doi.org/10.1111/gcb.12830>  
607 Palosuo, T., Kersebaum, K. C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J. E., ... Rötter,  
608 R. (2011). Simulation of winter wheat yield and its variability in different climates of  
609 Europe: A comparison of eight crop growth models. *European Journal of Agronomy*, 35(3),  
610 103–114. <http://doi.org/10.1016/j.eja.2011.05.001>  
611 Porter, J. R., Xie, L., Challinor, A. J., Cochrane, K., Howden, S. M., Iqbal, M. M., ... Travasso,  
612 M. I. (2014). Food security and food production systems. In C. B. Field, V. R. Barros, D. J.  
613 Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, ... L. L. White (Eds.), *Climate Change*  
614 *2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects*.

615 *Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental*  
616 *Panel on Climate Change* (pp. 485–533). Cambridge, United Kingdom and New York, NY,  
617 USA: Cambridge University Press.

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

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

623 Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., ... Jones, J. W.  
624 (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded  
625 crop model intercomparison. *Proceedings of the National Academy of Sciences of the*  
626 *United States of America*, 111(9), 3268–73. <http://doi.org/10.1073/pnas.1222463110>

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

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

633 Rötter, R. P., Palosuo, T., Kersebaum, K. C., Angulo, C., Bindi, M., Ewert, F., ... Trnka, M.  
634 (2012). Simulation of spring barley yield in different climatic zones of Northern and Central  
635 Europe: A comparison of nine crop models. *Field Crops Research*, 133, 23–36.  
636 <http://doi.org/10.1016/j.fcr.2012.03.016>

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

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

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

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

645 Wallach, D., Mearns, L. O., Ruane, A. C., Rötter, R. P., & Asseng, S. (2016). Lessons from

646 climate modeling on the design and use of ensembles for crop modeling. *Climatic Change*,  
647 139(3–4), 551–564. <http://doi.org/10.1007/s10584-016-1803-1>

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

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

655 Yin, X., Kersebaum, K. C., Kollas, C., Baby, S., Beaudoin, N., Manevski, K., ... Olesen, J. E.  
656 (2017). Multi-model uncertainty analysis in predicting grain N for crop rotations in Europe.  
657 *European Journal of Agronomy*, 84, 152–165. <http://doi.org/10.1016/j.eja.2016.12.009>

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

661

Author Manuscript



	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 <sup>-1</sup>	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 <sup>-1</sup>	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 T <sub>max</sub> >31°C ranged from 28 to 74.  2 spring cultivars (Bacanora 88 and Nesser)  Yields 1.9-8.0 t ha <sup>-1</sup>	Detailed crop measurements for four environments at one location (Obregon, 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 <sup>-1</sup>	Detailed crop measurements for the 3 control environments.	Majoul-Haddad, Bancel, Martre, Triboi, & Branlard (2013)
AGFACE (18)	Elevated free air CO <sub>2</sub> concentration experiment, over three years, early or late sowing, CO <sub>2</sub> concentrations of 385 or 550 ppm, rain-fed or irrigated.  1 spring cultivar (Yitpi)  Yields 1.1-4.6 t ha <sup>-1</sup>	Detailed crop measurements for one environment (385 ppm CO <sub>2</sub> , early sowing, irrigated).  Parameters used previously in 6 models.	O'Leary et al. (2015)

663

664

**Table 1.**

665

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

666

**is given in parentheses.**

667 Figure legends

668 Figure 1.

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

676 Figure 1

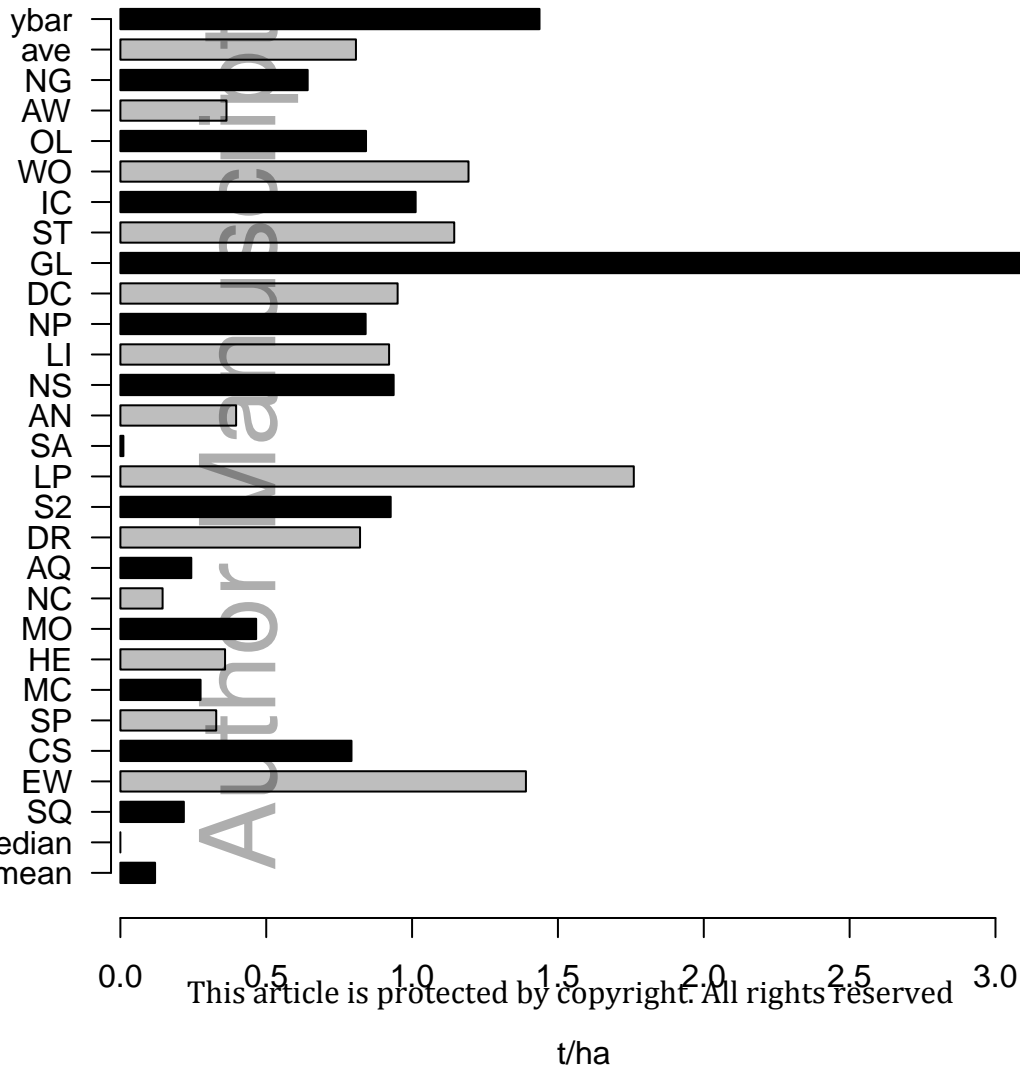
677 Fit of models to HSC yield data. Each environment number corresponds to a different  
678 sowing date, either without (“C”) or with (“H”) supplementary heating. Solid diamonds are  
679 observed yields. Circles and triangles show respectively e-mean and e-median. Values simulated  
680 by the 25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

681 Figure 3.

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

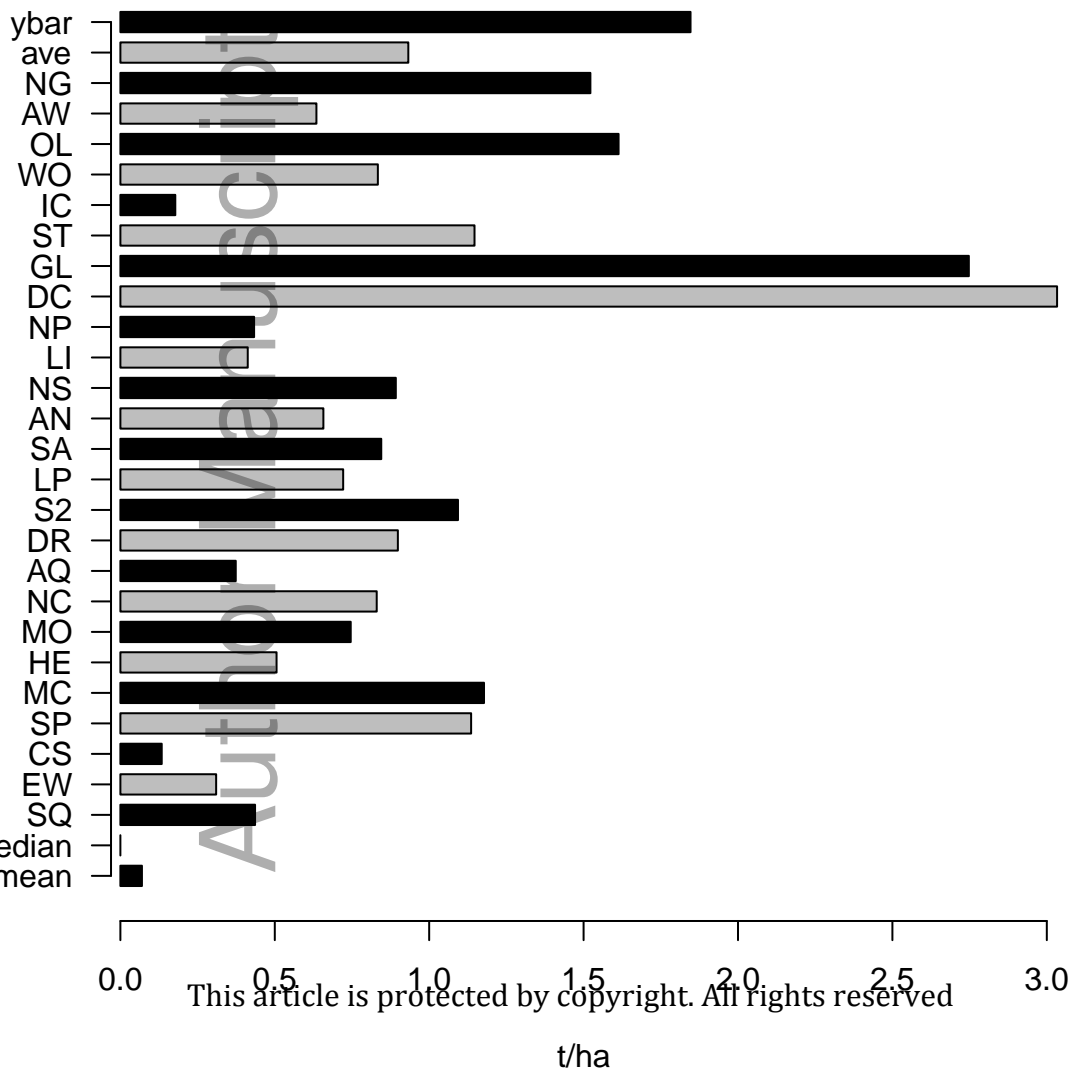
a

## Yield AgMIP–Wheat pilot



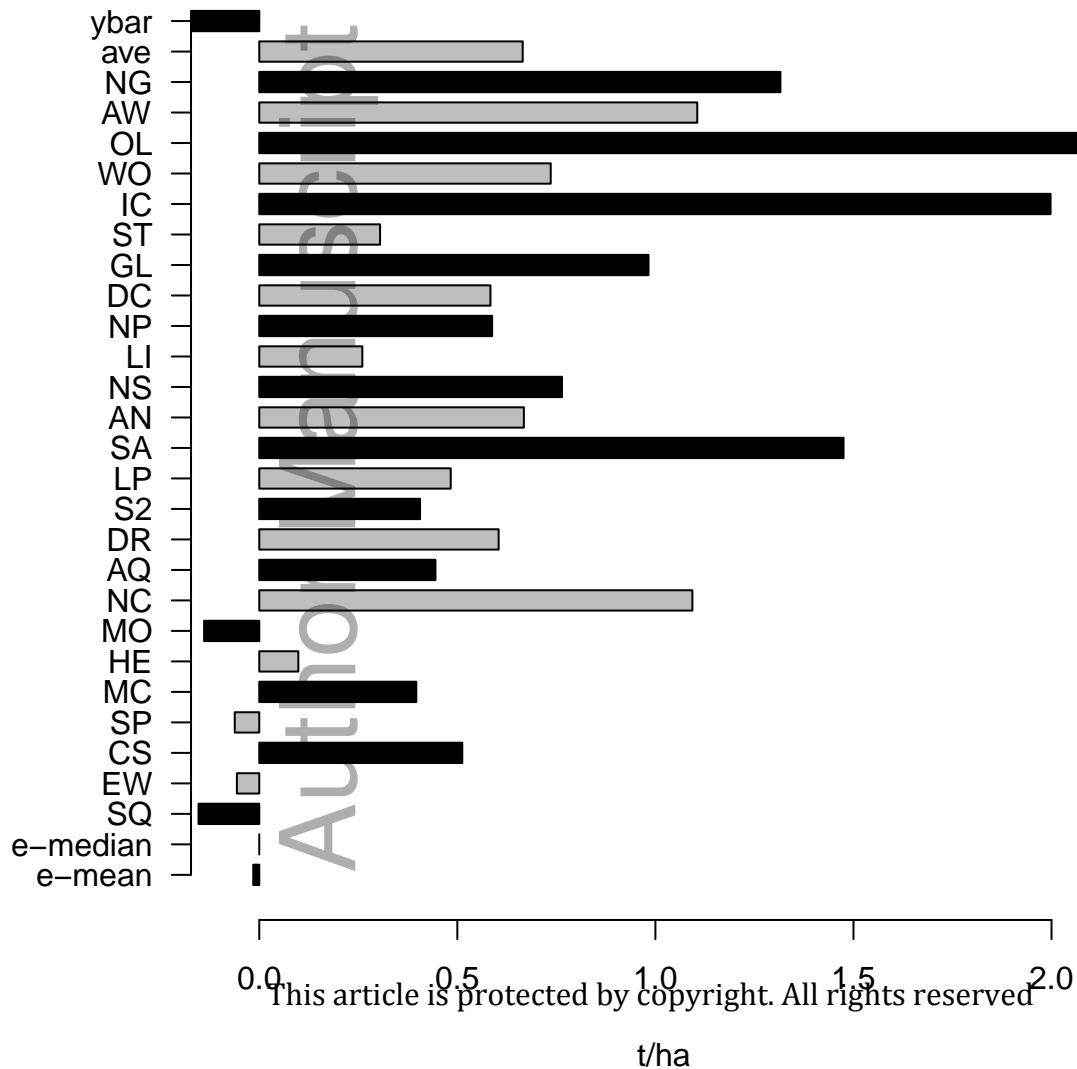
## Yield HSC

b



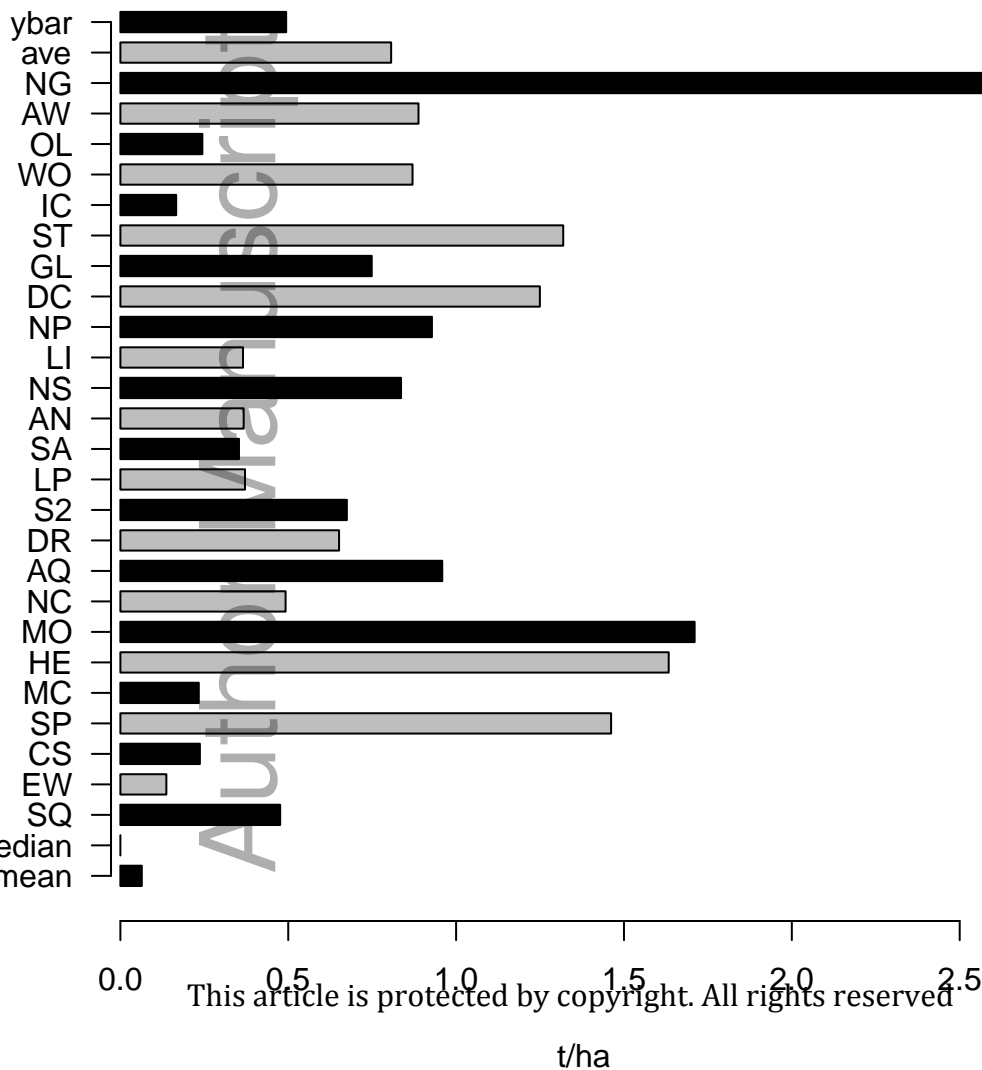
C

## Yield HSGE



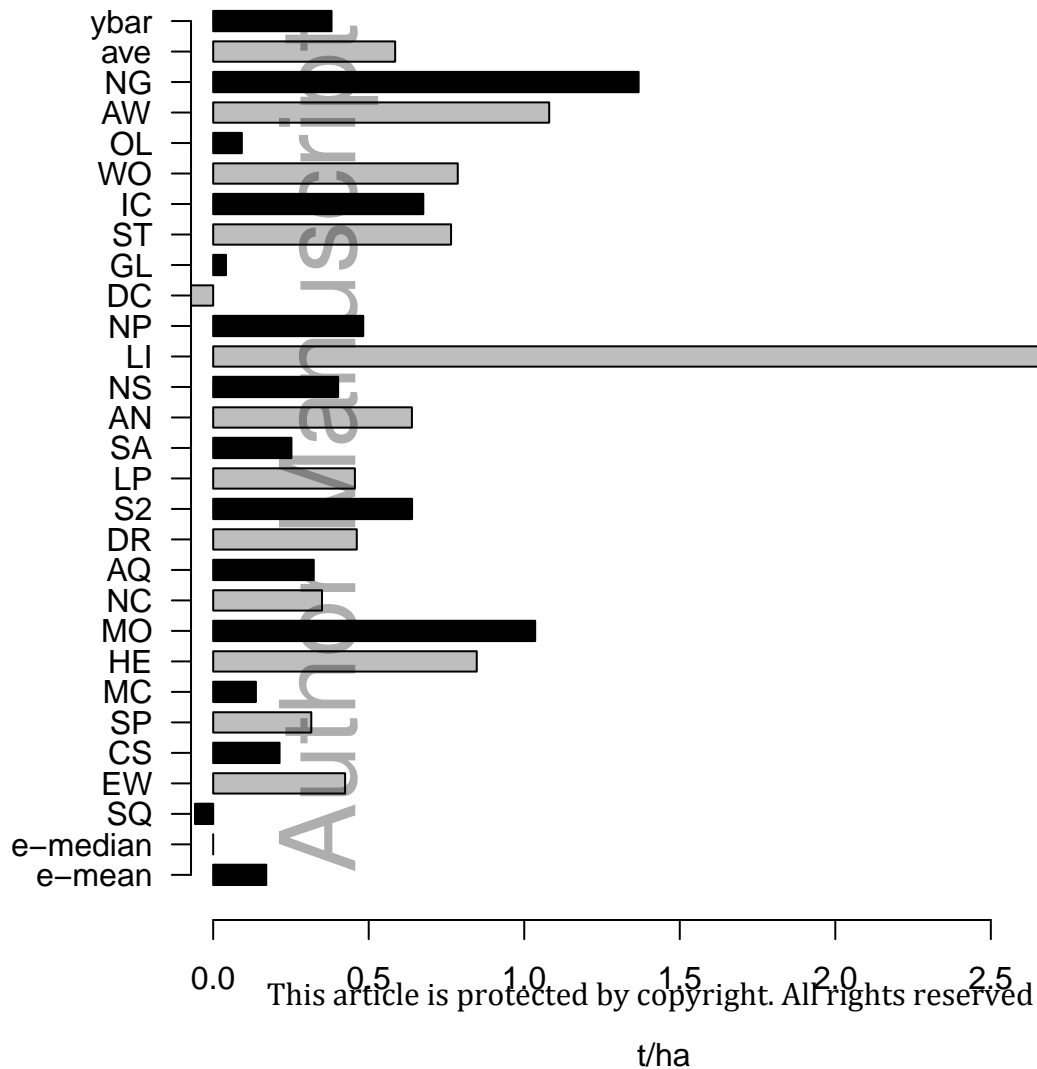
## Yield C3-GEM

d



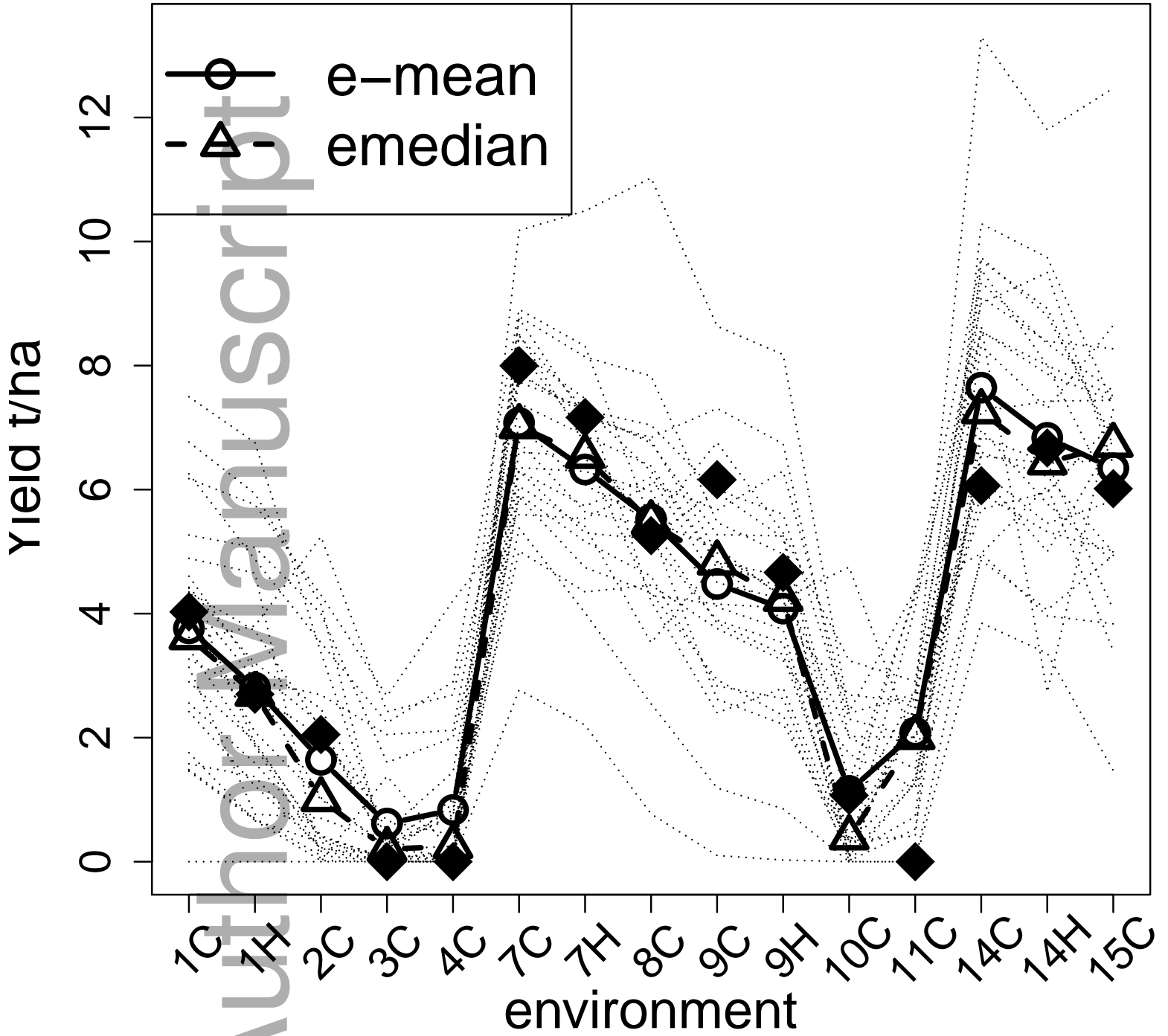
e

## Yield AGFACE





# HSC Yield

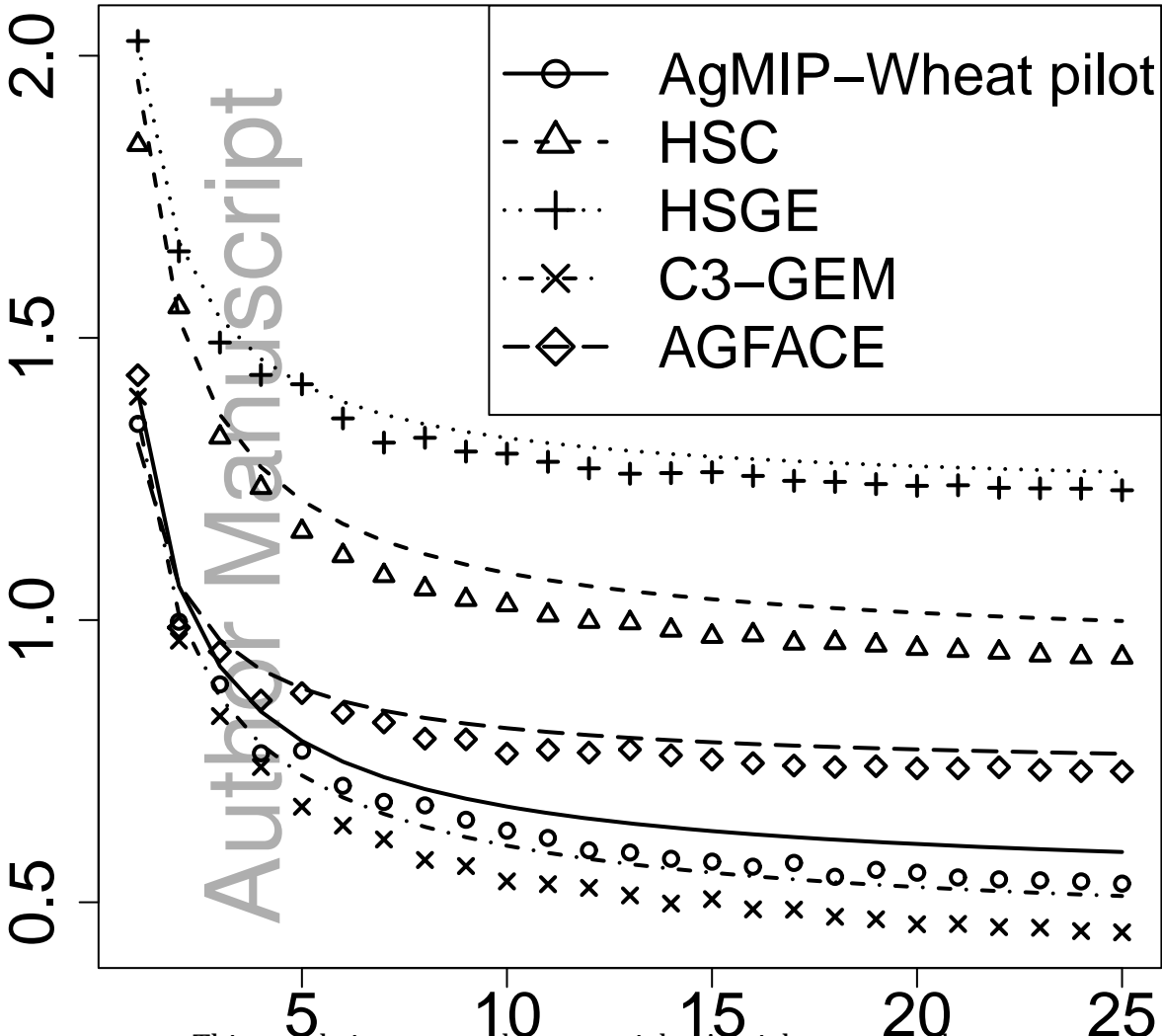


gcb\_14411\_f2.eps

# Yield

a

RMSE e-mean t/ha



This article is protected by copyright. All rights reserved

number of models

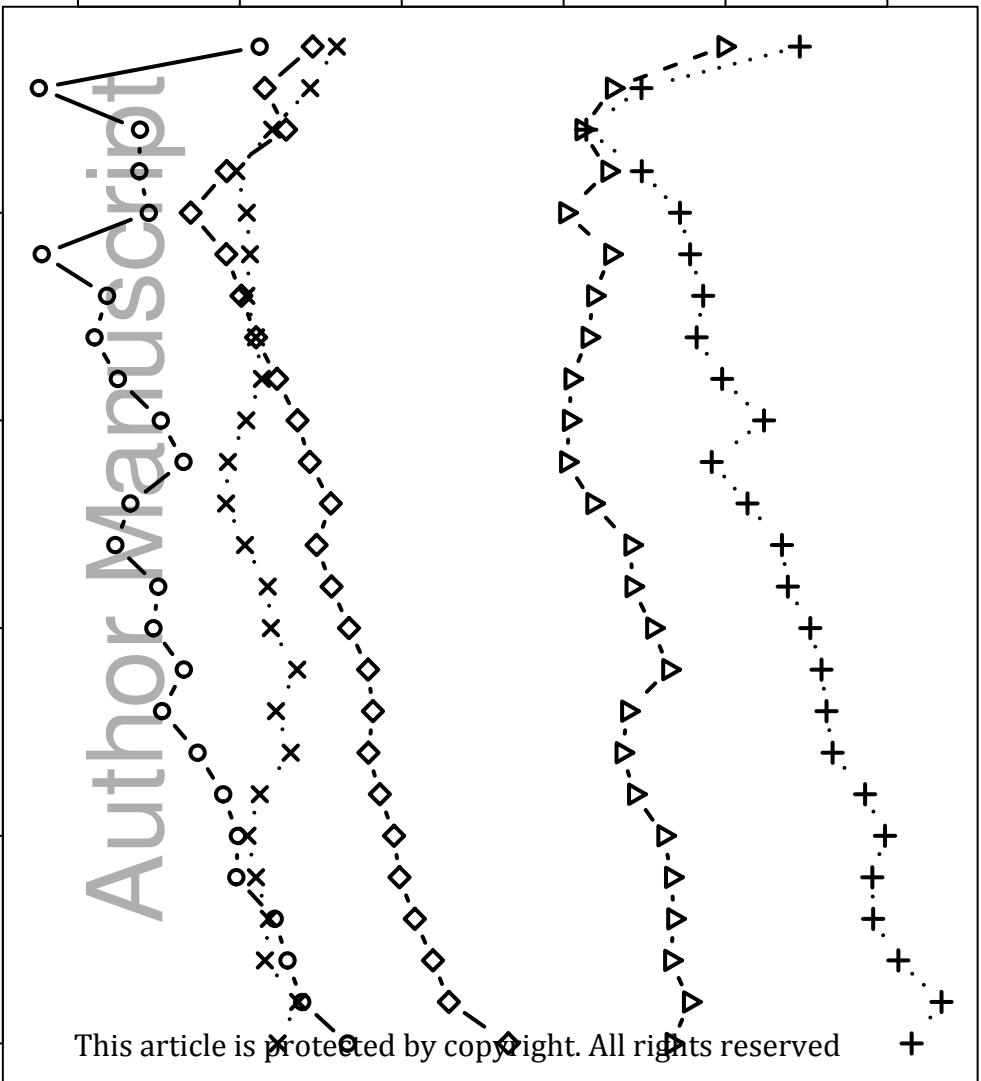
RMSE e-mean t/ha

0.2 0.4 0.6 0.8 1.0 1.2

b

number of models

Yield



This article is protected by copyright. All rights reserved



Minerva Access is the Institutional Repository of The University of Melbourne

**Author/s:**

Wallach, D;Martre, P;Liu, B;Asseng, S;Ewert, F;Thorburn, PJ;van Ittersum, M;Aggarwal, PK;Ahmed, M;Basso, B;Biernath, C;Cammarano, D;Challinor, AJ;De Sanctis, G;Dumont, B;Rezaei, EE;Feres, E;Fitzgerald, GJ;Gao, Y;Garcia-Vila, M;Gayler, S;Girousse, C;Hoogenboom, G;Horan, H;Izauralde, RC;Jones, CD;Kassie, BT;Kersebaum, KC;Klein, C;Koehler, A-K;Maiorano, A;Minoli, S;Mueller, C;Kumar, SN;Nendel, C;O'Leary, GJ;Palosuo, T;Priesack, E;Ripoche, D;Roetter, RP;Semenov, MA;Stockle, C;Stratonovitch, P;Streck, T;Supit, I;Tao, F;Wolf, J;Zhang, Z

**Title:**

Multimodel ensembles improve predictions of crop-environment-management interactions

**Date:**

2018-11-01

**Citation:**

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P. J., van Ittersum, M., Aggarwal, P. K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A. J., De Sanctis, G., Dumont, B., Rezaei, E. E., Feres, E., Fitzgerald, G. J., Gao, Y., ... Zhang, Z. (2018). Multimodel ensembles improve predictions of crop-environment-management interactions. *GLOBAL CHANGE BIOLOGY*, 24 (11), pp.5072-5083. <https://doi.org/10.1111/gcb.14411>.

**Persistent Link:**

<http://hdl.handle.net/11343/284698>