Multimodel ensembles of wheat growth: more models are better than one

Journal:	*Global Change Biology*
Manuscript ID:	Draft
Wiley - Manuscript type:	Primary Research Articles
Date Submitted by the Author:	n/a
Complete List of Authors:	Martre, Pierre; INRA, UMR1095 GDEC Wallach, Daniel; INRA, UMR1248 Agrosystèmes et Développement Territorial Asseng, Senthold; University of Florida, Agricultural & Biological Engineering Department Ewert, Frank; Crop Science, INRES Jones, Jim; University of Florida, Agricultural & Biological Engineering Department Rötter, Reimund; MTT Agrifood Research Finland, Plant Production Research Boote, Kenneth J; University of Florida, ; University of Florida, Agricultural & Biological Engineering Department Ruane, Alex; National Aeronautics and Space Administration, Goddard Institute for Space Studies Thorburn , Peter; CSIRO, Cammarano, Davide; University of Florida, Department of Agricultural and Biological Engineering Hatfield, jerry; USDA- ARS, NLAE; Rosenzweig, Cynthia; NASA GISS, Climate Impacts Group; National Aeronautics and Space Administration, Goddard Institute for Space Studies Aggarwal, P; International Water Management Institute, Consultative Group on International Agricultural Research, Research Program on Climate Change, Agriculture and Food Security Angulo, Carlos; Crop Science, INRES Basso, Bruno; Michigan State University, Department of Geological Sciences and Kellogg Biological Station Bertuzzi, Patrick; INRA, US1116 AgroClim Biernath, Christian; Helmholtz Zentrum München, Brisson, Nadine; INRA, UMR0211 Agronomie Challinor, Andrew; University of Leeds , Institute for Climate and Atmospheric Science, School of Earth and Environment Doltra, Jordi; Cantabrian Agricultural Research and Training Centre, Gayler, Sebastian; WESS competence centrer, Goldberg, Ritchie; National Aeronautics and Space Administration, Goddard Institute for Space Studies Grant, Robert; University of Alberta, Department of Renewable Resources Hooker, Josh; University of Reading, School of Agriculture Hunt, Anthony; University of Guelph, Department of Plant Agriculture

Ingwersen, Joachim; Universität Hohenheim, Institute of Soil Science and Land Evaluation
Izaurralde, Roberto; University of Maryland, Department of Geographical Sciences
Kersebaum, Christian; Leibniz-Center of Agricultural Landscape Research, Institute for Landscape Systems Analysis
Kumar, Naresh; Indian Agricultural Research Institute, Center for Environment Science and Climate Resilient Agriculture
Nendel, Claas; Leibniz-Center of Agricultural Landscape Research, Institute for Landscape Systems Analysis
O'Leary, Garry; Department of Primary Industries, Landscape & Water Sciences
Olesen, Jorgen; Aarhus University, Dept. of Agroecology and Environment
Osborne, Tom M; University of Reading, Meteorology
Palosuo, Taru; MTT Agrifood research Finland,
Priesack, Eckart; Helmholtz Zentrum München,
Ripoche, Dominique; INRA, US1116 AgroClim
Semenov, Mikhail; Rothamsted Research, Biomathematics and Bioinformatics
Shcherbak, Iurii; Michigan State University, Department of Geological Sciences and Kellogg Biological Station
Steduto, Pasquale; FAO,
Stöckle, Claudio
Stratonovitch, Pierre; Rothamsted Research, Biomathematics and Bioinformatics
Streck, Thilo; Universität Hohenheim, Institute of Soil Science and Land Evaluation<
Supit, Iwan; Wagenignen University, Earth System Sciences
Tao, F
Travasso, Maria; INTA, CIRN-Clima y Agua
Waha, Katharina; Potsdam Inst Climate Impact Res, PIK, Germany,
White, Jeffrey; USDA/ARS, ALARC,
Wolf, Joost; Wageningen University, Plant Production Systems

Keywords: Ecophysiological model, Ensemble modeling, model intercomparison, process-based model, uncertainty, wheat (Triticum aestivum L.)

Abstract: Crop models of crop growth are increasingly used to quantify the impact of global changes due to climate or crop management. Therefore, accuracy of simulation results is a major concern. Studies with ensembles of crop models can give valuable information about model accuracy and uncertainty, but such studies are difficult to organize and have only recently begun. We report on the largest ensemble study to date, of 27 wheat models tested in four contrasting locations for their accuracy in simulating multiple crop growth and yield variables. The relative error averaged over models was 24-38% for the different end-of-season variables including grain yield (GY) and grain protein concentration (GPC). There was little relation between error of a model for GY or GPC and error for in-season variables. Thus, most models did not arrive at accurate simulations of GY and GPC by accurately simulating preceding growth dynamics. Ensemble simulations, taking either the mean (e-mean) or median (e-median) of simulated values, gave better estimates than any individual model when all variables were considered. Compared to individual models, e-median ranked first in simulating measured GY and third in GPC. The error of e-mean and e-median declined with an increasing number of ensemble members, with little decrease beyond 10 models. We conclude that multimodel ensembles can be used to create new estimators with improved accuracy and consistency in simulating growth dynamics. We argue that these results are applicable to other crop species, and hypothesize that they apply more generally to ecological system models.

Running head: Multimodel ensembles of wheat growth

Multimodel ensembles of wheat growth: more models are better than one

PIERRE MARTRE[1,2], DANIEL WALLACH[3], SENTHOLD ASSENG[4], FRANK EWERT[5], JAMES W. JONES[4], REIMUND P. RÖTTER[6], KENNETH J. BOOTE[4], ALEX C. RUANE[7], PETER J. THORBURN[8], DAVIDE CAMMARANO[4], JERRY L. HATFIELD[9], CYNTHIA ROSENZWEIG[7], PRAMOD K. AGGARWAL[10], CARLOS ANGULO[5], BRUNO BASSO[11], PATRICK BERTUZZI[12], CHRISTIAN BIERNATH[13], NADINE BRISSON[14,15,*], ANDREW J. CHALLINOR[16,17], JORDI DOLTRA[18], SEBASTIAN GAYLER[19], RICHIE GOLDBERG[7], ROBERT F. GRANT[20], LEE HENG[21], JOSH HOOKER[22], LESLIE A HUNT[23], JOACHIM INGWERSEN[24], ROBERTO C IZAURRALDE[25], KURT CHRISTIAN KERSEBAUM[26], CHRISTOPH MÜLLER[27], SOORA NARESH KUMAR[28], CLAAS NENDEL[26], GARRY O'LEARY[29], JØRGEN E. OLESEN[30], TOM M. OSBORNE[31], TARU PALOSUO[6], ECKART PRIESACK[13], DOMINIQUE RIPOCHE[12], MIKHAIL A. SEMENOV[32], IURII SHCHERBAK[11], PASQUALE STEDUTO[33], CLAUDIO O. STÖCKLE[34], PIERRE STRATONOVITCH[32], THILO STRECK[24], IWAN SUPIT[35], FULU TAO[36], MARIA TRAVASSO[37], KATHARINA WAHA[27], JEFFREY W. WHITE[38] and JOOST WOLF[39]

[1]INRA, UMR1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), 5 chemin de Beaulieu, F-63 100 Clermont-Ferrand, France. [2]Blaise Pascal University, UMR1095 GDEC, F-63 170 Aubière, France. [3]INRA, UMR1248 Agrosystèmes et Développement Territorial, F-31 326 Castanet-Tolosan, France. [4]Agricultural & Biological Engineering Department, University of Florida, Gainesville, FL 32611. [5]Institute of Crop Science and Resource Conservation, Universität Bonn, D-53 115, Germany. [6]Plant Production Research, MTT Agrifood Research Finland, FI-50 100 Mikkeli, Finland. [7]National Aeronautics and Space Administration, Goddard Institute for Space Studies, New York, NY 10025. [8]Commonwealth Scientific and Industrial Research Organization, Ecosystem Sciences, Dutton Park QLD 4102, Australia. [9]National Laboratory for Agriculture and Environment, Ames, IA 50011. [10]Consultative Group on International Agricultural Research, Research Program on Climate Change, Agriculture and Food Security, International Water Management Institute, New Delhi 110012, India. [11]Department of Geological Sciences and Kellogg Biological Station, Michigan State University, East Lansing, MI. [12]INRA, US1116 AgroClim, F- 84 914 Avignon, France. [13]Institute of Soil Ecology, Helmholtz Zentrum München, German Research

Center for Environmental Health, Neuherberg, D-85 764, Germany. [14]INRA, UMR0211 Agronomie, F-78 750 Thiverval-Grignon, France. [15]AgroParisTech, UMR0211 Agronomie, F-78 750 Thiverval-Grignon, France. [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, A.A. 6713, Cali, Colombia. [18]Cantabrian Agricultural Research and Training Centre, 39600 Muriedas, Spain. [19]Water & Earth System Science Competence Cluster, c/o University of Tübingen, D-72 074 Tübingen, Germany. [20]Department of Renewable Resources, University of Alberta, Edmonton, AB, Canada T6G 2E3. [21]International Atomic Energy Agency, 1400 Vienna, Austria. [22]School of Agriculture, Policy and Development, University of Reading, RG6 6AR, United Kingdom. [23]Department of Plant Agriculture, University of Guelph, Guelph, Ontario, Canada, N1G 2W1. [24]Institute of Soil Science and Land Evaluation, Universität Hohenheim, D-70 599 Stuttgart, Germany. [25] Department of Geographical Sciences, University of Maryland, College Park, MD 20782. [26]Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research, D-15 374 Müncheberg, Germany. [27]Potsdam Institute for Climate Impact Research, D-14 473 Potsdam, Germany. [28]Centre for Environment Science and Climate Resilient Agriculture, Indian Agricultural Research Institute, New Delhi 110 012, India. [29]Landscape & Water Sciences, Department of Primary Industries, Horsham 3400, Australia. [30]Department of Agroecology, Aarhus University, 8830 Tjele, Denmark. [31]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, RG6 6BB, United Kingdom. [32]Computational and Systems Biology Department, Rothamsted Research, Harpenden, Herts, AL5 2JQ, United Kingdom. [33]Food and Agriculture Organization of the United Nations, Rome, Italy. [34]Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120. [35]Earth System Science-Climate Change, Wageningen University, 6700AA, The Netherlands. [36]Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science, Beijing 100101, China. [37]Institute for Climate and Water, INTA-CIRN, 1712 Castelar, Argentina. [38]Arid-Land Agricultural Research Center, USDA, Maricopa, AZ 85138. [39]Plant Production Systems, Wageningen University, 6700AA Wageningen, The Netherlands.

Correspondence: Pierre Martre, tel. +33 473 624 351, fax +33 473 624 457,
e-mail: pierre.martre@clermont.inra.fr.

[*]Dr Nadine Brisson passed away in 2011 while this work was being carried out.

Keywords: Ecophysiological model, Ensemble modeling, model intercomparison, process-based model, uncertainty, wheat (*Triticum aestivum* L.).

Type of paper: Primary Research

Abstract

Crop models of crop growth are increasingly used to quantify the impact of global changes due to climate or crop management. Therefore, accuracy of simulation results is a major concern. Studies with ensembles of crop models can give valuable information about model accuracy and uncertainty, but such studies are difficult to organize and have only recently begun. We report on the largest ensemble study to date, of 27 wheat models tested in four contrasting locations for their accuracy in simulating multiple crop growth and yield variables. The relative error averaged over models was 24-38% for the different end-of-season variables including grain yield (GY) and grain protein concentration (GPC). There was little relation between error of a model for GY or GPC and error for in-season variables. Thus, most models did not arrive at accurate simulations of GY and GPC by accurately simulating preceding growth dynamics. Ensemble simulations, taking either the mean (e-mean) or median (e-median) of simulated values, gave better estimates than any individual model when all variables were considered. Compared to individual models, e-median ranked first in simulating measured GY and third in GPC. The error of e-mean and e-median declined with an increasing number of ensemble members, with little decrease beyond 10 models. We conclude that multimodel ensembles can be used to create new estimators with improved accuracy and consistency in simulating growth dynamics. We argue that these results are applicable to other crop species, and hypothesize that they apply more generally to ecological system models.

Introduction

Global change with increased climatic variability are projected to strongly impact crop and food production, but the magnitude and trajectory of these impacts remain uncertain (Tubiello *et al.*, 2007). This uncertainty, together with the increasing demand for food of a growing world population (Bloom, 2011), has raised concerns about food security and the need to develop more sustainable agricultural practices (Godfray *et al.*, 2010). More confident understanding of global change impacts is needed to develop effective adaptation and mitigation strategies (Easterling *et al.*, 2007). Methodologies to quantify global change impacts on crop production include statistical models (Lobell *et al.*, 2011) and process-based crop simulation models (Porter & Semenov, 2005), which are increasingly used in basic and applied research and to support decision making at different scales (Angulo *et al.*, 2013, Challinor *et al.*, 2009, Ko *et al.*, 2010, Rosenzweig *et al.*, 2013b).

Different crop growth and development processes are affected by climatic variability via linear or non-linear relationships resulting in complex and unexpected responses (Trewavas, 2006). It has been argued that such responses can best be captured by process-based crop simulation models that quantitatively represent the interaction and feedback responses of crops to their environments (Bertin *et al.*, 2010, Porter & Semenov, 2005). Wheat is the most important staple crop in the world providing over 20% of the calories and proteins in human diet (FAOSTAT, 2012). It has therefore received much attention from the crop modeling community and over 40 wheat crop models are in use (White *et al.*, 2011). These differ in the processes included in the models and the mechanistic detail used to model individual processes like evapotranspiration or photosynthesis. Therefore, a thorough comparative evaluation of models is essential to understand the reliability of model simulations and to quantify and reduce the uncertainty of such simulations (Rötter *et al.*, 2011).

The Wheat Pilot study (Asseng *et al.*, 2013) of the Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig *et al.*, 2013b) compared twenty-seven wheat models, the largest ensemble of crop models created to date. The models vary greatly in their complexity and in the modeling approaches and equations used to represent the major physiological processes that determine crop growth and development and their responses to environmental factors, see Table S3 in supplemental in Asseng *et al.* (2013).

An initial study (Asseng *et al.*, 2013) analyzed the variability between crop models in simulating grain yield (GY) under climate change situations without specifically investigating multimodel ensemble estimators considering other end-of-season and in-season variables to better justify their possible application. The present analysis uses the resulting dataset to study how the multimodel ensemble average or median can reproduce in-season and end-of-season observations. In its simplest and most common form, a multimodel ensemble simulation is produced by averaging the simulations of member models weighted equally (Knutti, 2010). This method has been practiced in climate forecasting (Hagedorn *et al.*, 2005, Räisänen & Palmer, 2001) and in ecological modeling of species distribution (Grenouillet *et al.*, 2011), and it has been shown that multimodel ensembles can give better estimates than any individual model. Such improvement in skill of a multimodel ensemble may be also applicable to crop models. Preliminary evidence suggests that the average of ensembles of simulations is a good estimator of GY for several crops (Bassu *et al.*, 2014, Palosuo *et al.*, 2011, Rötter *et al.*, 2012) and possibly even better than the best individual model across different seasons and sites (Rötter *et al.*, 2012). However, a detailed quantitative analysis of the quality of simulators based on crop model ensembles, compared to individual models is lacking. By looking at outputs of multiple growth variables (both in-season and end-of-season), we would get a broader picture of how ensemble estimators perform and a better understanding of why they perform well compared to individual models. It is important therefore to consider not only GY but also other growth variables. If multimodel ensembles are truly more skillful than the best model in the ensemble, or even simply better than the average of the models, then using ensemble medians or means may be a powerful estimator to evaluating crop response to crop management and environmental factors.

Model evaluations can give quite different results depending on the use of the model that is studied. Here we investigate the situation where models are applied in environments for which they have not been specifically calibrated, which is typically the situation in global impact studies (Rosenzweig *et al.*, 2013a). The model results were compared to measured data from four contrasting growing environments. The modeling groups were provided with weather data, soil characteristics, soil initial conditions, management and flowering and harvest dates for each site. Although only four locations were tested in the AgMIP Wheat Pilot study, this limitation is partially compensated for by the diversity of the sites ranging from high to low yielding, from short to long season, and irrigated and not irrigated situations.

Two main approaches to evaluate the accuracy and uncertainty of the AgMIP wheat model ensemble were followed. First we evaluated the range of errors and the average error of the models for multiple growth variables, including both in-season and end-of-season variables. Secondly, we evaluated two ensemble-based models, the mean (e-mean) and the median (e-median) of the simulated values of the ensemble members. Finally, we studied how the error of e-mean and e-median changed with the size of the ensemble.

Materials and Methods

Experimental data

Quality-assessed experimental data from single crops at four contrasting locations representing diverse agro-ecological conditions were used. The locations were Wageningen, The Netherlands (NL; Groot & Verberne, 1991), Balcarce, Argentina (AR; Travasso *et al.*, 2005), New Delhi, India (IN; Naveen, 1986), and Wongan Hills, Australia (AU; Asseng *et al.*, 1998). Typical regional crop management was used at each site. In all experiments, the plots were kept weed-free, and plant protection methods were used as necessary to minimize damage from pests and diseases. Crop management and cultivar information, as given to each individual modeling group, are given in Table S1 in supplemental.

Daily values of solar radiation, maximum and minimum temperature and precipitation were recorded at weather stations at or near the experimental plots, except for IN solar radiation which was obtained from the NASA POWER dataset of modeled data (Stackhouse, 2006) that extends back to 1983. Daily values of 2-meter wind speed (m s^{-1}), dew point temperature (ºC), vapor pressure (hPa), and relative humidity (%) were estimated for each location from the NASA Modern Era Retrospective-Analysis for Research and Applications (Bosilovich *et al.*, 2011), except for NL wind speed and vapor pressure that were measured on site. Air CO_2 concentration was taken to be 360 ppm at all sites. A weather summary for each site is shown in Fig. S1 in supplemental.

For all sites, end-of-season (i.e. ripeness-maturity) values for GY (t DM ha^{-1}), total aboveground biomass ($AGBM_m$, t DM ha^{-1}), total aboveground nitrogen (N; AGN_m, kg N ha^{-1}), and grain N (GN_m, kg N ha^{-1}) were available. From these values, biomass harvest index ($HI = 100 \times GY/AGBM_m$, %), N harvest index ($NHI = 100 \times GN_m/AGN_m$, %), and grain protein concentration ($GPC = 0.57 \times GN_m/GY$, % of grain dry mass) were calculated. In-season measurements included leaf area index (LAI, m^2 m^{-2}; 15 measurements in total), total aboveground biomass (AGBM, t DM ha^{-1}; 28 measurements), total aboveground N (AGN, kg N ha^{-1}; 27 measurements) and soil water content to maximum rooting depth (mm, 28 measurements). Plant-available soil water to maximum rooting depth (PASW, mm) was calculated from the measured soil water content by layer (Θ_{V} , vol%), the estimated lower limit of water extraction (LL, vol%) , and the thickness of the soil layers (d, m):

$$\mathrm{PASW} = \sum_{i=1}^{k} d_i \times \left(\Theta_{\mathrm{V},i} - \mathrm{LL}_i\right) \quad (1)$$

where k is the number of sampled soil layers.

Based on the critical N dilution curve of wheat (Justes *et al.*, 1994), a N nutrition index (NNI, dimensionless) was calculated to quantify crop N status. Although this curve is empirical, it is based on solid theoretical grounds (Lemaire & Gastal, 1997). Climatic conditions can affect growth and N uptake differently, but the NNI reflects these effects in terms of crop N needs (Gonzalez-Dugo *et al.*, 2010, Lemaire *et al.*, 2008). For a given AGBM, NNI was calculated as the ratio between the actual and critical (N_{C} ; g N g^{-1} DM) AGN concentrations defined by the critical N dilution curve (Justes *et al.*, 1994):

$$N_{\mathrm{C}} = 5.35 \times \mathrm{AGBM}^{-0.442} \quad (2)$$

If the NNI value is close to 1 it indicates an optimal crop N status, a value lower than 1 indicates N deficiency and a value higher than 1 indicates N excess.

Models and setup of model intercomparison

The models considered here were the 27 wheat crop models (Table S2 in supplemental) used in the AgMIP Wheat Pilot study (Asseng *et al.*, 2013). All of these models have been described in publications and are currently in use. Not all models simulated all measured variables, either because the models did not simulate them or because they were not in the standard outputs. Of the 27 models, 23 models simulated PASW values, and 20 simulated

AGN and GN, and therefore NNI and GPC could be calculated for these 20 models. NHI could be calculated for 19 models.

All modeling groups were provided with daily weather data (i.e. precipitation, minimum and maximum air temperature, mean relative air humidity, dew point temperature, mean air vapor pressure, global radiation and mean wind speed), basic physical characteristics of soil, initial soil water and N content by layer and crop management information (Table S2 in supplemental). No indication of how to interpret or convert this information into parameter values was given to the modelers. Modelers were provided with observed anthesis and maturity dates for the cultivars grown at each site. Qualitative information on vernalization requirements and daylength responses were also provided.

In the simulations, phenology parameters were adjusted to reproduce the observed anthesis and maturity dates, but otherwise models were not specifically adjusted to the growth data, which were only revealed to the modelers at the end of the simulation phase of the project. Modelers were instructed to keep all parameters except for genotypic coefficients constant across all four sites.

For three of the four sites, the data used here were previously available in the literature, so some of these data may have been used in the past with some models as part of larger datasets. If so, this would concern only some of the data used here, only a few models and only part of the data used for testing and model calibration. We chose this over the alternative approach of only using unpublished data to avoid other potential problems (Kersebaum *et al.*, 2007, Palosuo *et al.*, 2011, Rötter *et al.*, 2012).

Except for the four Expert-N models which were run by the same group, all models were run by different groups without communication between the groups regarding the parameterization of the initial conditions or cultivar specific parameters. In most cases, the model developers ran their own model.

Model evaluation

Many different measures of the discrepancies between simulations and measurements have been proposed (Wallach, 2006). We concentrated on the root mean squared error (RMSE) and the root mean squared relative error (RMSRE), where each error is expressed as a percentage of the observed value. The RMSE has the advantage of expressing error in the same units as the variable. For comparing very different environments likely to give a broad range of crop responses, the relative error may be more meaningful than the absolute error as it gives more equal weight to each measurement. However, RMSRE needs to be interpreted with care because it is very sensitive to errors when measured values are small, as occurred for several early-season growth measurements.

RMSE was calculated as the square root of the mean squared error (MSE). MSE for model *m* and for a particular variable (MSE_m) was calculated as:

$$\mathrm{MSE}_m = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_{m,i}\right)^2 \quad (3)$$

where y_i is the value of the *i*th measurement of this variable, $\hat{y}_{m,i}$ is the corresponding value simulated by model *m*, and *N* is the total number of measurements of this variable (i.e. the sum over sites and over sampling dates per site for in-season variables).

RMSRE was calculated as:

$$\mathrm{RMSRE}_m = 100 \times \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_{m,i}}{y_i}\right)^2} \quad (4)$$

To assess whether a model that simulates well for one variable also performs well for other variables, Pearson's product-moment correlation between the RMSE or RMSRE value of each model was calculated across the variables. The adjusted two-sided *P*-values (*q*-values) resulting from the correction for multiple tests were calculated and reported here.

Multimodel ensemble estimators

We considered two models that are based on the ensemble of model simulations. The first ensemble estimator, e-mean, is the mean of the model simulations. The second ensemble estimator, e-median, is the median of the individual model simulations. For each of these ensemble models, e-mean and e-median, we calculated the same criteria as for the individual models, namely MSE, RMSE, and RMSRE.

In order to explore how e-mean MSE and e-median MSE varied with the number of models in the ensemble, we performed a bootstrap calculation for each value of *M'* (number of models in the ensemble) from 1 to 27. For each ensemble size *M'* we drew B = 25,600 bootstrap samples of *M'* models with replacement, so the same model might be represented more than once in the sample. A preliminary analysis showed that the results were essentially unchanged beyond 3,000 bootstrap samples. The final estimate of MSE for e-mean was then:

$$\mathrm{MSE}_{\text{e-mean}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N}\left(y_i - \hat{y}^{b}_{\text{e-mean},i}\right)^2 \tag{5}$$

where $\hat{y}^{b}_{e-mean,i}$ is the e-mean estimate in bootstrap sample b of the ith measurements of this variable, given by:

$$\hat{y}^{b}_{\text{e-mean},i} = \frac{1}{M'}\sum_{m=1}^{M'}\hat{y}^{b}_{m,i} \tag{6}$$

For e-median the estimate of MSE was calculated as:

$$\mathrm{MSE}_{\text{e-median}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N}\left(y_i - \hat{y}^{b}_{\text{e-median},i}\right)^2 \tag{7}$$

In the case of e-mean, we can calculate the theoretical expectation of MSE analytically as a function of *M'*. Consider a variable at a particular site. Let μ_i^* represent the true expectation of model simulations for that site (the mean over all possible models), and let $\hat{\mu}_{i,M'}$ represent an e-mean simulation which is based on a sample of models of size *M'*. The expectation of MSE (expectation over possible samples of *M'* models) for e-mean is then:

$$E\left(\mathrm{MSE}_{M'}\right)=E\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{\mu}_{i,M}\right)^2\right]=\frac{1}{N}\sum_{i=1}^{N}E\left[\left(y_i-\mu^*{}_i+\mu^*{}_i-\hat{\mu}_{i,M}\right)^2\right]$$
$$=\frac{1}{N}\sum_{i=1}^{N}\left[\left(y_i-\mu^*{}_i\right)^2+\frac{\mathrm{var}\left(\hat{y}_i\right)}{M}\right] \tag{7}$$

where $\mathrm{var}\left(\hat{y}_i\right)$ is the variance of the simulated values for the different models. The first term in the sum in (equation 8) is the squared bias of e-mean, when e-mean is based on a very large number of models. The second term is the variance of the model simulations divided by *M*. μ_i^* can be estimated as the average of the simulations over all the models in our study, and $\mathrm{var}\left(\hat{y}_i\right)$ can be estimated as the variance of those model simulations.

All calculations and graphs were made using the R statistical software R 3.0.1 (R Core Team, 2013). Pearson's product-moment correlation *P*-values were adjusted for false discovery rate using the 'LBE' package (Dalmasso *et al.*, 2005), and bootstrap sampling used the R function sample().

Results

Evaluation of a population of wheat crop models

In most cases, measured in-season LAI, PASW, AGBM, AGN, and NNI, and end-of-season GY and GPC values were within the range of model simulations (Fig. 1, 2). Even though measured GY ranged from 2.50 to 7.45 t DM ha^{-1} across the four sites, the ranges of simulated GY values were similar at the four sites with an average range between minimum and maximum simulations of 1.64 t DM ha^{-1} (Fig. 2a). The range between minimum and maximum simulations for GPC was also comparable at the four sites, averaging 7.1 percentage points (Fig. 2b).

On average over all models, the RMSRE was 29% (Fig. 3a and Table S3 in supplemental), and RMSE was 1.25 t DM ha^{-1} for GY (Fig. 3b and Table 1 and Table S4 in supplemental). The uncertainty in simulated GY was large, with RMSRE ranging from 8% to 73% among the 27 models, but 80% of the models had an RMSRE for GY comprised between 14% and 47% (Fig. 3a). For the other end-of-season variables RMSRE ranged from 7% to 60% for HI (averaging 24%), 22 to 61% for GN (averaging 38%), 15% to 52% for NHI (averaging 26%), and 8% to 122% for GPC (averaging 34%; Fig. 3a). For the in-season variables with multiple measurements per site, the RMSRE ranged from 48% to 1496% for LAI, 37% to 355% for PASW, 41% to 542% for AGBM, 49% to 472% for AGN, and 16% to 104% for NNI (Fig. 3a).

Of the three models with the smallest RMSE for GY, only the second-ranked model had RMSE values below the average of all models for all variables considered (Table 1). The other two models had an RMSE substantially higher than the average for at least one variable. The first- and second-ranked models simulated GY closely because of compensating errors. They underestimated LAI around anthesis and final AGBM which was compensated for by overestimating HI. For instance, the first-ranked model simulated that the canopy intercepted

83%, 74% and 51% of the incident radiation around anthesis in AR, IN and NL, respectively, while according to measured LAI values the percentage of radiation interception was close to 93% at the three sites (assuming an extinction coefficient of 0.55, an average value reported for wheat canopies (Sylvester-Bradley *et al.*, 2012)). This model compensated by having unrealistically high HI values that were 19% to 93% higher than measured HI. Theoretical maximum HI has been estimated at 62-64% for wheat (Foulkes *et al.*, 2011), while this model had simulated values up to 69% (in NL). The third-ranked model showed no significant compensation of errors. This model overestimated LAI around anthesis by 16% in AR and NL, but this translated into only a small effect on intercepted radiation, since the canopy intercepted more than 90% of incident radiation based on observed LAI.

Relation between the error for grain yield and that for underlying variables

There was little relation between the errors for different variables (Fig. 3a, b). There were some exceptions however. Notably, RMSE for AGBM was highly correlated with that for GY, and that for AGN was correlated with GN (Fig. 4). Similarly, RMSE for AGN was highly correlated with that for LAI, PASW, and NNI. Finally, RMSE for NNI was correlated with that for PASW, HI, and GN and to a lesser extent with that for NNI. RMSE for GPC was not significantly correlated with any other variable. Overall, the correlations between RMSRE for different variables were similar to that between RMSE for different variables (Fig. S2 in supplemental).

Multimodel ensemble estimators

Two multimodel ensemble estimators were tested. The first, the e-mean, uses the mean of the simulations of the ensemble members, a common practice in climate ensemble modeling (Knutti, 2010). The second, the e-median, uses the median of the simulations of the ensemble members. The e-median is expected to be less sensitive to outlier simulations than e-mean and therefore provide more robust estimates.

The e-median and e-mean values gave good agreement with measured values in almost all cases, despite the fact that the simulations of the individual models varied considerably (Fig. 1, 2). The e-median and e-mean models were much better than the average over models for all responses (Fig. 3). For most variables, e-mean and e-median had similar RMSE and RMSRE values, and their ranking among all models was close (Table 1 and Supplementary Table S3, S4). The largest difference in ranks was for RMSE for GPC, where e-median was ranked 3 and e-mean was ranked 7.

For most variables, e-mean and e-median were comparable to the best single model for that variable (Fig. 3a, b). When e-median was ranked with the other models based on RMSRE, it ranked fourth for GY and third for GPC (Table S3 in supplemental); and first for GY and third for GPC when ranked based on RMSE (Table S4 in supplemental). One way to quantify the overall skill of e-mean and e-median is to consider the sum of ranks over all the variables. The sum of ranks based on RMSE for the 10 variables analyzed in this study was 37 for e-median and 45 for e-mean, while the lowest sum of ranks for an individual model (among the 17 models that simulated all variables) was 53 (Table S3 in supplemental). If we only considered the four variables simulated by all 27 models (i.e. LAI, AGBM, GY, and HI), the sum of ranks for e-median and e-mean was 15 and 17, respectively, while the best sum of ranks for an individual model with these four variables was 28.

In order to analyze the relationship between the number of models in an ensemble and the RMSE of both e-mean and e-median, we used a bootstrap approach to create a large number of ensembles of different sizes. The RMSE of both e-mean and e-median in each bootstrap ensemble was calculated and averaged over bootstrap samples. The bootstrap results for e-mean were very close to the theoretical expectation of RMSE (Fig. 5). For all variables, the standard deviation of RMSE between bootstrap samples for e-mean decreased as the number of models in the ensemble increased. The average RMSE of e-median also decreased with the number of models, in a manner similar to, but not identical to, the average e-mean RMSE. The differences were most pronounced for GPC (Fig. 5j).

Discussion

Working with multimodel ensembles is well-established in climate modeling, but only recently has the necessary international coordination been developed to make this also possible for crop models (Rosenzweig *et al.*, 2013b). Here we examined the performance of an ensemble of 27 wheat models, created in the context of the AgMIP Wheat Pilot study (Asseng *et al.*, 2013). Multiple crop responses, including both end-of-season and in-season growth variables were considered. Among these, GY and GPC are the main determinants of wheat productivity and end-use value. The other variables helped indicate whether models are realistic and consistent in their description of the processes leading to GY and GPC. This provides more comprehensive information on crop system properties beyond GY and is essential for the analysis of adaptation and mitigation strategies to global changes (Challinor *et al.*, 2014).

In only a few cases there were significant correlations between a model's error for one variable and its error for other variables. Several individual models had relatively small errors for GY or GPC and large errors for in-season variables, including two of the three models with the lowest RMSE for GY. These models arrived at accurate simulations of GY or GPC without simulating crop growth accurately and thus got the right answer for, at least in part, the wrong reasons. That is, models can compensate for structural inconsistency. It has been argued that interactions among system components are largely empirical in most crop models (Ahuja & Ma, 2011) and that model error is minimized with different parameter values for different variables (Wallach, 2011), which would explain why a model might simulate one variable well and not others. However, it remains unclear whether such compensation will be effective in a wide range of environments. The lack of correlation between model errors for different variables illustrate the need for crop model ensemble assessment for multiple variables (Challinor *et al.*, 2014), as done in this study.

The behavior of the median and mean of the ensemble simulations was similar. Both estimators had much smaller errors and better skills than the average over models, for all variables. In comparing the sum of ranks of error for all variables, which provides an aggregated performance measure, the e-median was better than e-mean, but most importantly both were superior to even the best performing model in the ensemble. Different measures of performance might give slightly different results, but would not change the fact that e-median and e-mean compare well with even the best models.

E-mean and e-median had small errors in simulating not only end-of-season variables but also in-season variables. This suggests that multimodel ensembles could be useful not only for simulating GY and GPC, but also for relating those results to in-season growth processes. This is important if crop model ensembles are to be useful in exploring the consequences of global change and the benefits of adaptation or mitigation strategies.

A fundamental question is the origin of the advantage of ensemble predictors over individual models. Two possible explanations relate to compensation among errors in processes descriptions and to more coverage of the possible crop and soil phase spaces.

Certain models had large errors with compensations to achieve a reasonable yield simulation. In those cases, e-median can supply a better estimate when multiple responses are considered, since it gives reasonable results for all variables. In other cases, it is simply the fact that the errors in the different models tend to compensate each other well, that makes e-median the best estimator over multiple responses. The compensation of errors among models comes, at least in part, from the fact that models do not produce random outputs but are driven by environmental and management inputs and bio-physical processes and therefore they tend to converge to the measured crop response. It is an open question however as to whether the superiority of crop model ensemble estimators compared to individual models extends to conditions not tested in this study. In particular, will this still be the case if the models are used to predict the impact of climate change?

For climate models, the main reason for the superiority of multimodel ensemble estimators is that better coverage of the whole possible climate phase space leads to greater consistency (Hagedorn *et al.*, 2005). An analogous advantage holds as well for crop model ensembles, they have more associated knowledge and represent more processes than any individual model. Each of the individual models has been developed and calibrated based on a limited data set. The ensemble simulators are in a sense averaging over these data sets, which gives them the advantage of a much broader data base than any individual model and thus reduces the need for site- and varietal-specific model calibration.

The use of ensemble estimators to answer new questions in the future poses specific questions regarding the best procedure for creating an ensemble. Several of these questions have been debated in the climate science community (Knutti, 2010), but not always in a way that is directly applicable to crop models. One question is how performance varies with the number of models in the ensemble. Here we found that the change in ensemble error ($\mathrm{MSE}_{M'}$) with the number of model in the ensemble (M') follows the expectation of MSE. Thus when planning ensemble studies, one can estimate the potential reduction in $\mathrm{MSE}_{M'}$ and therefore, do a costs vs. benefits analysis for increasing M'. In the ensemble studied here, for all the variables, MSE for an ensemble of 10 models was close to the asymptotic limit for very large M'.

Other questions include how to choose the models in the ensemble, and whether one should weight the models in the ensemble differently, based on past performance and convergence for new situations (Tebaldi & Knutti, 2007). In this respect the crop modeling community might employ some of the ensemble weighting methods developed by the climate modeling community (Christensen *et al.*, 2010). There are also questions about the possible multiple uses of models. Would it be advantageous to have multiple simulations, based on a diversity of initial conditions (including 'spin-up' periods for models that depend on simulation of changes in soil organic matter) or multiple parameter sets from each model? In any case, the first step is to document the accuracy of multimodel ensemble estimators in specific situations, as done here.

In summary, by reducing simulation error and improving the consistency of simulation results for multiple variables, crop model ensembles could substantially increase the range of questions that could be addressed. A lack of correlation between end-of-season and in-season errors in the individual models indicates that further work is needed to improve the representation of the dynamics of growth and development processes leading to GY in crop models. This is crucial for their application under changed climatic or management conditions.

Most of the physical and physiological processes that are simulated in wheat models are the same as for other crops. In fact, several of the models in this study have a generic structure so that they can be applied to various crops, and for some of them the differences between crops are simply in the parameter values. It is thus reasonable to expect that the results obtained here for wheat are broadly applicable to other crop species. It would be worthwhile to study whether these results also apply more generally to biological and ecological system models.

Acknowledgements

P.M. is grateful to the INRA metaprogram "Adaptation of Agriculture and Forests to Climate Change" and Environment and Agronomy Division for supporting several stays at the University of Florida during this work.

References

Ahuja LR, Ma L (2011) A synthesis of current parameterization approaches and needs for further improvements. In: *Methods of introducing system models into agricultural research.* (eds Ahuja LR, Ma L) pp 427-440. Madison, WI, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Angulo C, Rötter R, Lock R, Enders A, Fronzek S, Ewert F (2013) Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe. *Agricultural and Forest Meteorology,* **170**, 32-46.

Asseng S, Ewert F, Rosenzweig C *et al.* (2013) Uncertainty in simulating wheat yields under climate change. *Nature Climate Change,* **3**, 827-832.

Asseng S, Keating BA, Fillery IRP *et al.* (1998) Performance of the APSIM-wheat model in Western Australia. *Field Crops Research,* **57**, 163-179.

Bassu S, Brisson N, Durand J-L *et al.* (2014) How do various maize crop models vary in their responses to climate change factors? *Global Change Biology*, in press.

Bertin N, Martre P, Genard M, Quilot B, Salon C (2010) Under what circumstances can process-based simulation models link genotype to phenotype for complex traits? Case-study of fruit and grain quality traits. *Journal of Experimental Botany,* **61**, 955-967.

Bloom DE (2011) 7 Billion and Counting. *Science,* **333**, 562-569.

Bosilovich MG, Robertson FR, Chen JY (2011) Global energy and water budgets in MERRA. *Journal of Climate,* **24**, 5721-5739.

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

Challinor AJ, Wheeler T, Hemming D, Upadhyaya HD (2009) Ensemble yield simulations: crop and climate uncertainties, sensitivity to temperature and genotypic adaptation to climate change. *Climate Research,* **38**, 117-127.

Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M (2010) Weight assignment in regional climate models. *Climate Research,* **44**, 179-194.

Dalmasso C, Broët P, Moreau T (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics,* **21**, 660-668.

Easterling WE, Aggarwal PK, Batima P *et al.* (2007) Food, fibre and forest products. In Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the intergovernmental Panel on Climate Change. (eds Parry ML, Canziani OF, Palutikof JP, Van De Linden P, Hanson CE) pp 273–313. Cambridge, UK, Cambridge University Press.

Faostat (2012) Food and Agricultural organization of the United Nations (FAO). FAO Statistical Databases. <apps.fao.org>.

Foulkes MJ, Slafer GA, Davies WJ *et al.* (2011) Raising yield potential of wheat. III. Optimizing partitioning to grain while maintaining lodging resistance. *Journal of Experimental Botany,* **62**, 469-486.

Godfray HCJ, Beddington JR, Crute IR *et al.* (2010) Food security: the challenge of feeding 9 billion people. *Science,* **327**, 812-818.

Gonzalez-Dugo V, Durand J-L, Gastal F (2010) Water deficit and nitrogen nutrition of crops. A review. *Agronomy for Sustainable Development,* **30**, 529-544.

Grenouillet G, Buisson L, Casajus N, Lek S (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography,* **34**, 9-17.

Groot JJR, Verberne ELJ (1991) Response of wheat to nitrogen fertilization, a data set to validate simulation models for nitrogen dynamics in crop and soil. In: *Nitrogen Turnover in the Soil-Crop System. Modelling of Biological Transformations, Transport of Nitrogen and Nitrogen Use Efficiency. Proceedings of a Workshop.* (eds Groot JJR, De Willigen P, Verberne ELJ) pp 349-383, Institute for Soil Fertility Research, Haren, The Netherlands.

Hagedorn T, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus,* **57A**, 219-233.

Justes E, Mary B, Meynard JM, Machet JM, Thelier-Huche L (1994) Determination of a critical nitrogen dilution curve for winter wheat crops. *Ann Bot,* **74**, 397-407.

Kersebaum K, Hecker J-M, Mirschel W, Wegehenkel M (2007) Modelling water and nutrient dynamics in soil–crop systems: a comparison of simulation models applied on common data sets. In: *Modelling water and nutrient dynamics in soil–crop systems.* (eds Kersebaum K, Hecker J-M, Mirschel W, Wegehenkel M) pp 1-17. Springer Netherlands.

Knutti R (2010) The end of model democracy? An editorial comment. *Climatic Change,* **102**, 395-404.

Ko J, Ahuja L, Kimball B *et al.* (2010) Simulation of free air CO2 enriched wheat growth and interactions with water, nitrogen, and temperature. *Agricultural and Forest Meteorology,* **150**, 1331-1346.

Lemaire G, Gastal F (1997) N uptake and distribution in plant canopies. In: *Diagnosis of the nitrogen status in crops.* (ed Lemaire G) pp 3-43. Berlin, Germany, Springer Verlag.

Lemaire G, Jeuffroy M-H, Gastal F (2008) Diagnosis tool for plant and crop N status in vegetative stage: Theory and practices for crop N management. *European Journal of Agronomy,* **28**, 614-624.

Lobell DB, Schlenker W, Costa-Roberts J (2011) Climate trends and global crop production since 1980. *Science,* **333**, 616-620.

Naveen (1986) Evaluation of soil water status, plant growth and canopy environment in relation to variable water supply to wheat. Unpublished Ph.D IARI, New Delhi.

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

Porter JR, Semenov MA (2005) Crop responses to climatic variation. *Philosophical Transactions of the Royal Society of London B Biological Sciences,* **360**, 2021-2035.

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Räisänen J, Palmer TN (2001) A probability and decision-model analysis of a multimodel ensemble of climate change simulations. *Journal of Climate,* **14**, 3212-3226.

Rosenzweig C, Elliott J, Deryng D *et al.* (2013a) Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences*.

Rosenzweig C, Jones JW, Hatfield JL *et al.* (2013b) The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. *Agricultural and Forest Meteorology,* **170**, 166-182.

Rötter RP, Carter TR, Olesen JE, Porter JR (2011) Crop-climate models need an overhaul. *Nature Climate Change,* **1**, 175-177.

Rötter RP, Palosuo T, Kersebaum KC *et al.* (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.

Stackhouse P (2006) Prediction of worldwide energy resources. http://power.larc.nasa.gov.

Sylvester-Bradley R, Riffkin P, O'leary G (2012) Designing resource-efficient ideotypes for new cropping conditions: Wheat (*Triticum aestivum* L.) in the High Rainfall Zone of southern Australia. *Field Crops Research,* **125**, 69-82.

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* **365**, 2053-2075.

Travasso MI, Magrin GO, Rodríguez R, Grondona MO (2005) Comparing CERES-wheat and SUCROS2 in the Argentinean Cereal Region. In: *MODSIM 2005 International Congress on Modelling and Simulation.* (eds Zerger A, Argent RM) pp 366-369. Modelling and Simulation Society of Australia and New Zealand.

Trewavas A (2006) A brief history of systems biology: "every object that biology studies is a system of systems." Francois Jacob (1974). *Plant Cell,* **18**, 2420-2430.

Tubiello FN, Soussana J-F, Howden SM (2007) Crop and pasture response to climate change. *Proceedings of the National Academy of Sciences,* **104**, 19686-19690.

Wallach D (2006) Evaluating crop models. In: *Working with Dynamic Crop Models.* (eds Wallach D, Makowski D, Jones JW) pp 11-53. Amsterdam, The Netherlands, Elsevier.

Wallach D (2011) Crop Model Calibration: A Statistical Perspective. *Agronomy Journal,* **103**, 1144-1151.

White JW, Hoogenboom G, Kimball BA, Wall GW (2011) Methodologies for simulating impacts of climate change on crop production. *Field Crops Research,* **124**, 357-368.

Supporting Information

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

Table S1. Details of the experimental sites and experiments provided to the modelers.

Table S2. Name, reference and source of the 27 wheat crop models used in this study.

Table S3. Root mean square relative error (RMSRE) for in-season and end-of-season variables.

Table S4. Root mean square error (RMSE) for in-season and end-of-season variables.

Figure S1. Weather data at the four studied sites.

Figure S2. Correlation matrix for Pearson's product-moment correlation (*r*) between the root mean squared relative error of simulated variables.

Figure Captions

Fig. 1. Measured and simulated values of five in-season wheat crop variables for four sites. (**a-d**) Leaf area index (LAI), (**e-h**) plant-available soil water (PASW), (**i-l**) total aboveground biomass (AGBM), (**m-p**) total aboveground nitrogen (AGN), and (**q-t**) nitrogen nutrition index (NNI) versus days after sowing in The Netherlands (NL), Argentina (AR), India (IN) and Australia (AU). Symbols are single measurements and solid lines are medians of the simulations (i.e. e-median). Dark grey areas indicate the 10th to 90th percentile range and light grey areas the 25th to 75th percentile range of the values generated by different wheat crop models. Twenty-seven models were used to simulate LAI and AGBM, 24 to simulate PASW, 20 to simulate AGN and NNI. In **e-h** the horizontal red lines indicate 50% soil water deficit.

Fig. 2. Measured and simulated values of two major end-of-season wheat crop variables for four sites. Measured (red crosses) and simulated (box plots) values for end-of-season (**a**) grain yield (GY) and (**b**) grain protein concentration (GPC) are shown for The Netherlands (NL), Argentina (AR), India (IN) and Australia (AU). Simulations are from 27 different wheat crop models for GY and 20 for GPC. Boxes show the 25th to 75th percentile range, horizontal lines in boxes show medians, and error bars outside boxes show the 10th to 90th percentile range.

Fig. 3. Wheat crop model errors for in-season and end-of-season variables. (**a**) Root mean squared relative error (RMSRE) and (**b**) root mean squared error (RMSE) for in-season leaf area index (LAI), plant-available soil water (PASW), total aboveground biomass (AGBM), total above ground nitrogen (AGN), nitrogen nutrition index (NNI), and for end-of-season grain yield (GY), biomass harvest index (HI), grain nitrogen yield (GN), nitrogen harvest index (NHI), and grain protein concentration (GPC). Twenty-seven models were used to simulate LAI, AGBM, GY, and HI, 20 to simulate AGN, GN, GPC and NNI, 24 to simulate

PASW, and 19 to simulate NHI. In **a** for GY the models are sorted from left to right in the order of increasing RMSE and this order of models was used to plot all other variables. The horizontal solid blue line shows RMSE or RMRSE averaged over all models and the horizontal red line shows RMSE or RMRSE for the median simulation of all models (e-median).

Fig. 4. Correlation matrix for Pearson's product-moment correlation (*r*) between the root mean squared error of simulated variables. In-season variables: leaf area index (LAI), plant-available soil water (PASW), total aboveground biomass (AGBM), total above ground nitrogen (AGN), nitrogen nutrition index (NNI). End-of-season variables: grain yield (GY), biomass harvest index (HI), grain nitrogen yield (GN), nitrogen harvest index (NHI), and grain protein concentration (GPC). Twenty-seven models were used to simulate LAI, AGBM, GY, and HI, 20 to simulate AGN, GN, GPC and NNI, 24 to simulate PASW, and 19 to simulate NHI. The numbers above the diagonal gap are *r* values and the numbers below are one-sided *q*-values (adjusted *P*-values for false discovery rate). The color (for *r* values only) and the shape of the ellipses indicate the strength (the narrower the ellipse the higher the *r* value) and the direction of the correlation, respectively.

Fig. 5. How the number of models in an ensemble affects error estimates. Average root mean squared error (RMSE) (± 1 s.d.) of e-mean and e-median for in-season (**a**) leaf area index (LAI), (**c**) plant-available soil water (PASW), (**e**) total above ground biomass (AGBM), (**g**) total above ground nitrogen (AGN) and (**i**) nitrogen nutrition index (NNI) and for end-of-season (**b**) grain yield (GY), (**d**) biomass harvest index (HI), (**f**) grain nitrogen yield (GN), (**h**) nitrogen harvest index (NHI), and (**j**) grain protein concentration (GPC) versus number of models in the ensemble. Values are calculated based on 10,000 bootstrap samples. The solid line is the analytical result for RMSE as a function of sample size (equation (8)). The blue dashed line shows the RMSE for e-mean and the red dashed line the RMSE for e-median of

the multimodel ensemble. The black dashed line is the RMSE for the individual model with lowest sum of ranks for RMSE. For visual clarity the RMSE for e-mean is plotted for even numbers of models, and the RMSE for e-median for odd numbers of models.

Table 1 RMSE for in-season and end-of-season variables. Ensemble averages and e-mean and e-median values are based on 27 different models for LAI, AGBM, GY, and HI, 24 for PASW, 20 for AGN, GN, GPC and NNI, and 19 for NHI. Values for the three best models for GY (based on RMSE) simulation are also given. Data for each individual model are given in Table S4 in supplemental. The numbers in parenthesis indicate the rank of the models (including e-mean and e-median) where 1 indicates the model with the lowest RMSE (i.e. best rank) for that variable. For each variable the model with the lowest RMSE is in bold type.

Estimator	**RMSE for in-season variables**					**RMSE for end-of-season variables**				
	LAI (m^{-2} m^{-2})	**PASW (mm)**	**AGBM (t DM ha^{-1})**	**AGN (kg N ha^{-1})**	**NNI (-)**	**GY (t DM ha^{-1})**	**HI (%)**	**GN (kg N ha^{-1})**	**NHI (%)**	**GPC (% of grain DM)**
Average over all models	1.90	47	2.07	39	0.35	1.25	8.5	38	18.7	3.93
Model ranked 1 for GY	2.31 (23)	60 (21)	2.26 (17)	89 (21)	0.92 (22)	0.42 (2)	20.0 (28)	100 (22)	23.6 (18)	6.91 (21)
Model ranked 2 for GY	1.24 (7)	36 (9)	1.71 (13)	24 (8)	0.26 (8)	0.56 (4)	7.2 (16)	27 (9)	9.1 (2)	2.75 (9)
Model ranked 3 for GY	1.75 (16)	63 (22)	**1.01 (3)**	22 (7)	**0.21 (4)**	0.63 (5)	3.8 (5)	29 (10)	11.7 (5)	2.13 (6)
e-median	**1.20 (6)**	**27 (3)**	1.20 (6)	15 (3)	0.25 (7)	**0.41 (1)**	2.8 (2)	**22 (5)**	**8.8 (1)**	**1.57 (3)**
e-mean	1.29 (8)	27 (5)	1.19 (5)	**13 (1)**	0.24 (6)	0.49 (3)	**2.2 (1)**	23 (6)	9.8 (3)	2.32 (7)

Supplementary Information

Multimodel ensembles of wheat growth: more models are better than one

PIERRE MARTRE, DANIEL WALLACH, SENTHOLD ASSENG, FRANK EWERT, JAMES W. JONES, REIMUND P. RÖTTER, KENNETH J. BOOTE, ALEX C. RUANE, PETER J. THORBURN, DAVIDE CAMMARANO, JERRY L. HATFIELD, CYNTHIA ROSENZWEIG, PRAMOD K. AGGARWAL, CARLOS ANGULO, BRUNO BASSO, PATRICK BERTUZZI, CHRISTIAN BIERNATH, NADINE BRISSON, ANDREW J. CHALLINOR, JORDI DOLTRA, SEBASTIAN GAYLER, RICHIE GOLDBERG, ROBERT F. GRANT, LEE HENG, JOSH HOOKER, LESLIE A HUNT, JOACHIM INGWERSEN, ROBERTO C IZAURRALDE, KURT CHRISTIAN KERSEBAUM, CHRISTOPH MÜLLER, SOORA NARESH KUMAR, CLAAS NENDEL, GARRY O'LEARY, JØRGEN E. OLESEN, TOM M. OSBORNE, TARU PALOSUO, ECKART PRIESACK, DOMINIQUE RIPOCHE, MIKHAIL A. SEMENOV, IURII SHCHERBAK, PASQUALE STEDUTO, CLAUDIO O. STÖCKLE, PIERRE STRATONOVITCH, THILO STRECK, IWAN SUPIT, FULU TAO, MARIA TRAVASSO, KATHARINA WAHA, JEFFREY W. WHITE& JOOST WOLF

Table S1. Details of the experimental sites and experiments provided to the modelers. Adapted from Asseng *et al.* (2013).

	Site			
	NL	AR	IN	AU
Site description				
Environment	High-yielding long-season	High/medium-yielding medium-season	Irrigated short-season	Low-yielding rain-fed short-season
Regional representation	Western and northern Europe	Argentina, northern China, western USA	India, Pakistan, southern China	Australia, southern Europe, northern Africa, South Africa, Middle East
Location name	Wageningen ('The Bouwing') The Netherlands	Balcarce Argentina	New Delhi India	Wongan Hills Australia
Coordinates	51° 58' N, 05° 37' E	37° 45' S, 58° 18' W	28° 22' N, 77° 7' E	30° 53' S, 116° 43' E
Soil characteristics				
Soil type[a]	Silty clay loam	Clay loam	Sandy loam	Loamy sand
Rooting depth (cm)	200	130	160	210
Apparent bulk density (m^3 m^{-3})	1.35	1.1	1.55	1.41
Top soil organic matter (%)	2.52	2.55	0.37	0.51
pH	6.0	6.3	8.3	5.7
Maximum plant available soil water (mm to maximum rooting depth)	354	222	109	125
Crop management				
Sowing density (seed m^{-2})	228	239	250	157
Cultivar				
Name	Arminda	Oassis	HD2009	Gamenya
Vernalization requirement	High	Little	None	Little
Daylength response	High	Moderate	None	Moderate
Ploughed crop residue	Potato (4 t ha^{-1})	Maize (7 t ha^{-1})	Maize (1.5 t ha^{-1})	Wheat/weeds (1.5 t ha^{-1})
Irrigation (mm)	0	0	383	0
N application (kg N ha^{-1})	120 (ZC30[b]) / 40 (ZC65)	120 (ZC00)	60 (ZC00) / 60 (ZC25)	50 (ZC10)
Initial top soil mineral N (kg N ha^{-1})	80	13	25	5
Sowing date	21 Oct. 1982	10 Aug. 1992	23 Nov. 1984	12 Jun. 1984
Anthesis date	20 Jun. 1983	23 Nov. 1992	18 Feb. 1985	1 Oct. 1984
Physiological maturity date	1 Aug. 1983	28 Dec. 1992	3 Apr. 1985	16 Nov. 1984

[a] Saturated soil water content, drainage upper limit and lower limit to water extraction were provided for 10 to 30-cm thick soil layers down to the maximum rooting depth.
[b] ZC, Zadoks stage(Zadoks *et al.*, 1974) at application is indicated in parenthesis (ZC00, sowing; ZC10, first leaf through coleoptile; ZC25, main shoot and five tillers; ZC30, pseudo stem erection; ZC65, anthesis half-way.

Table S2. Name, reference and source of the 27 wheat crop models used in this study. Modified from Asseng *et al.* (2013).

Model (version)	Reference to model description	Documentation/source (web link, e-mail address)
APSIM-Nwheats (V.1.55)	(Asseng *et al.*, 2004, Asseng *et al.*, 1998, Keating *et al.*, 2003)	http://www.apsim.info/Wiki/
APSIM (V.7.3)	(Keating *et al.*, 2003)	http://www.apsim.info/Wiki/
AquaCrop (V.3.1+)	(Steduto *et al.*, 2009)	http://www.fao.org/nr/water/aquacrop.html
CropSyst (V.3.04.08)	(Stöckle *et al.*, 2003)	http://www.bsyse.wsu.edu/CS_Suite/CropSyst/index.html
DSSAT-CERES (V.4.0.1.0)	(Hoogenboom & White, 2003, Jones *et al.*, 2003, Ritchie & Otter, 1985)	http://www.icasa.net/dssat/
DSSAT-CROPSIM (V.4.5.1.013)	(Hunt & Pararajasingham, 1995, Jones *et al.*, 2003)	http://www.icasa.net/dssat/
Ecosys	(Grant *et al.*, 2011)	http://www.rr.ualberta.ca/en/Research/EcosysModellingProject.aspx
EPIC wheat (V.1102)	(Izaurralde *et al.*, 2012, Kiniry *et al.*, 1995, Williams *et al.*, 1989)	http://epicapex.brc.tamus.edu/
Expert-N (V3.0.10) - CERES (V2.0)	(Biernath *et al.*, 2011, Priesack *et al.*, 2006, Stenger *et al.*, 1999)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
Expert-N (V3.0.10) – GECROS (V1.0)	(Biernath *et al.*, 2011, Priesack *et al.*, 2006, Stenger *et al.*, 1999, Yin & van Laar, 2005)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
Expert-N (V3.0.10) – SPASS (V2.0)	(Biernath *et al.*, 2011, Priesack *et al.*, 2006, Stenger *et al.*, 1999, Wang & Engel, 2000)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
Expert-N (V3.0.10) - SUCROS (V2)	(Biernath *et al.*, 2011, Goudriaan & Van Laar, 1994, Priesack *et al.*, 2006, Stenger *et al.*, 1999)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
FASSET (V.2.0)	(Berntsen *et al.*, 2003, Olesen *et al.*, 2002)	http://www.fasset.dk
GLAM-wheat (V.2)	(Challinor *et al.*, 2004, Li *et al.*, 2010)	http://www.see.leeds.ac.uk/see-research/icas/climate_change/glam/glam.html
HERMES (V.4.26)	(Kersebaum, 2007, Kersebaum, 2011)	http://www.zalf.de/en/forschung/institute/lsa/forschung/oekomod/hermes
InfoCrop (V.1)	(Aggarwal *et al.*, 2006)	Request from nareshkumar.soora@gmail.com
LINTUL-4 (V.1)	(Shibu *et al.*, 2010)	http://models.pps.wur.nl/models
LINTUL -FAST (V.1.0)	(Angulo *et al.*, 2013)	Request from frank.ewert@uni-bonn.de
LPJmL (V.3.2)	(Bondeau *et al.*, 2007)	http://www.pik-potsdam.de/research/projects/lpjweb
MCWLA-Wheat (V.2.0)	(Tao *et al.*, 2009)	Request from taofl@igsnrr.ac.cn
MONICA (V.1.0)	(Nendel *et al.*, 2011)	http://monica.agrosystem-models.com
O'Leary-model (V.7)	(O'Leary & Connor, 1996a, O'Leary & Connor, 1996b)	Request from author (gjoleary@yahoo.com)
SALUS (V.1.0)	(Basso *et al.*, 2010, Senthilkumar *et al.*, 2009)	http://www.salusmodel.net
Sirius (V.2010)	(Jamieson *et al.*, 2000, Jamieson *et al.*, 1998, Lawless *et al.*, 2005)	http://www.rothamsted.ac.uk/mas-models/sirius.html
SiriusQuality (V.2.0)	(Ferrise *et al.*, 2010, He *et al.*, 2012, Martre *et al.*, 2006)	http://www1.clermont.inra.fr/siriusquality/
STICS (V.1.1)	(Brisson *et al.*, 2003, Brisson *et al.*, 2009, Brisson *et al.*, 1998, Brisson *et al.*, 2002)	http://www7.avignon.inra.fr/agroclim_stics
WOFOST (V.7.1)	(Boogaard *et al.*, 1998, Van Diepen *et al.*, 1989)	http://www.wofost.wur.nl

Table S3. Root mean square relative error (RMSRE) for in-season and end-of-season variables.

Model*	RMSRE (%)¶										
	In-season					End-of-season					Sum of rank$
	LAI	PASW	AGBM	AGN	NNI	GY	HI	GN	NHI	GPC	
1	199	102	159	472	104	**8.1 (1)**	57.3 (28)	61.1 (22)	31.1 (15)	57.7 (19)	85/29
2	398	129	89	76	33	17.4 (9)	19.7 (16)	36.6 (12)	14.6 (3)	25.5 (12)	52/25
3	246	142	41	67	29	15.6 (6)	9.8 (4)	35.1 (11)	18.8 (6)	23.9 (10)	37/10
4	716	37	164	NA	NA	21.1 (12)	17.4 (15)	NA	NA	NA	–/27
5	319	177	129	NA	NA	13.3 (2)	24.3 (20)	NA	NA	NA	–/22
6	171	NA	47	NA	NA	19.3 (11)	20.3 (17)	NA	NA	NA	–/28
7	1496	50	132	60	28	23.5 (15)	15.3 (11)	23.2 (6)	18.5 (4)	58.3 (20)	56/26
8	172	95	114	123	38	13.7 (3)	11.3 (6)	29.5 (7)	19.4 (9)	36.9 (15)	40/9
9	140	**37**	67	63	**16**	14.4 (5)	13.3 (8)	22.2 (2)	22.8 (11)	10.7 (2)	28/13
10	821	68	542	384	35	16.4 (8)	14.3 (9)	44.1 (17)	28 (14)	26.8 (13)	61/17
11	692	59	52	**49**	56	27.8 (17)	23.5 (19)	39.3 (15)	48 (20)	28.8 (14)	85/36
12	133	45	103	145	48	18.2 (10)	24.5 (21)	NA	NA	NA	–/31
13	745	355	296	74	87	38.2 (22)	25.2 (22)	58 (21)	18.5 (5)	17.4 (5)	75/44
14	1150	150	53	72	32	42.5 (23)	16.6 (13)	31.6 (9)	19.2 (8)	121.8 (22)	75/36
15	58	40	84	75	34	22.8 (14)	7 (2)	37.9 (14)	40.3 (19)	23.2 (7)	56/16
16	219	NA	196	116	42	49.6 (28)	49.5 (26)	55.9 (19)	52.3 (21)	23.3 (8)	102/54
17	699	97	**41**	55	36	22.8 (13)	16.7 (14)	22.6 (4)	19.1 (7)	**8 (1)**	39/27
18	749	65	126	82	29	43.8 (25)	9.8 (4)	47.1 (18)	32.1 (16)	38.3 (17)	80/29
19	156	101	187	52	41	30.9 (20)	59.9 (29)	34.5 (10)	27.1 (13)	39.9 (18)	90/49
20	109	45	356	230	37	33.6 (21)	26.7 (23)	56.6 (20)	34.6 (18)	23.7 (9)	91/44
21	663	94	69	76	35	28.9 (18)	28.9 (24)	22.9 (5)	21.3 (10)	37.6 (16)	73/42
22	773	NA	193	192	49	29.9 (19)	11.8 (7)	30 (8)	23.8 (12)	15.3 (4)	50/26
23	294	40	199	NA	NA	45 (26)	44.6 (25)	NA	NA	NA	–/51
24	1085	79	77	73	61	27 (16)	22.3 (18)	37.6 (13)	33.6 (17)	24.8 (11)	75/34
25	**48**	59	91	NA	NA	43 (24)	15.7 (12)	40.9 (16)	NA	64 (21)	–/36
26	75	59	231	NA	NA	48.6 (27)	15.3 (10)	NA	NA	NA	–/37
27	1199	NA	306	NA	NA	72.6 (29)	53.8 (27)	NA	NA	NA	–/56
e-median	242	64	113	66	25	14 (4)	7.1 (3)	22.5 (3)	**13.7 (1)**	14.2 (3)	14/7
e- mean	442	70	133	79	24	15.6 (7)	**5.7 (1)**	**19.5 (1)**	14.3 (2)	20.8 (6)	17/8
Average over all models	501	92	154	127	44	29.2	24.3	38.3	27.5	35.3	–

Results are based on 27 different wheat crop models for LAI, AGBM, GY and HI, 20 for AGN, GN, GPC and NNI, 24 for PASW, and 19 for NHI.

* The models are sorted from top to bottom in the order of increasing RMSE for GY. For each variable the model with the lowest RMSRE is in bold type.

¶ NA, variables not available for a model. For end-of-season variables, the numbers in parentheses indicate the rank of the models (including e-mean and e-median) for each variable. Ranks were not calculated for in-season variables because several of the in-season measurements were very small causing large relative errors even the absolute errors were reasonable. Therefore RMSRE for in-season variables should be looked at with caution.

$ Sum of rank of RMSRE for end-of-season variables/sum of rank of RMSRE for the variables simulated by all 27 models (i.e., LAI, AGBM, GY, HI). For the reason mentioned above the sum of rank did not include in-season variables.

Table S4. Root mean square error (RMSE) for in-season and end-of-season variables.

Model[*]	RMSE[¶]										
	In-season					End-of-season					Sum of rank[$]
	LAI (m^2 m^{-2})	PASW (mm)	AGBM (t DM ha^{-1})	AGN (kg N ha^{-1})	NNI (-)	GY (t DM ha^{-1})	HI (%)	GN (kg N ha^{-1})	NHI (%)	GPC (% of grain DM)	
1	2.31 (23)	60 (21)	2.26 (17)	89 (21)	0.92 (22)	0.42 (2)	20.0 (28)	100 (22)	23.6 (18)	6.91 (21)	195/70
2	1.24 (7)	36 (9)	1.71 (13)	24 (8)	0.26 (8)	0.56 (4)	7.2 (16)	27 (9)	9.1 (2)	2.75 (9)	85/40
3	1.75 (16)	63 (22)	1.01 (3)	22 (7)	0.21 (4)	0.63 (5)	3.8 (5)	29 (10)	11.7 (5)	2.13 (6)	83/29
4	1.82 (19)	36 (8)	1.64 (12)	NA	NA	0.66 (6)	6.3 (13)	NA	NA	NA	–/50
5	1.13 (5)	46 (18)	2.30 (18)	NA	NA	0.69 (7)	9.9 (24)	NA	NA	NA	–/54
6	1.81 (18)	NA	1.41 (7)	NA	NA	0.74 (8)	7.6 (17)	NA	NA	NA	–/50
7	3.34 (28)	42 (16)	1.44 (9)	17 (4)	0.29 (11)	0.77 (9)	6.2 (12)	21 (3)	11.5 (4)	6.39 (20)	116/58
8	1.33 (10)	26 (2)	0.97 (2)	30 (10)	0.28 (9)	0.78 (10)	4.0 (6)	20 (2)	13.6 (9)	4.04 (16)	76/28
9	1.30 (9)	32 (7)	**0.87 (1)**	14 (2)	**0.16 (1)**	0.81 (11)	4.6 (9)	**20 (1)**	14.5 (10)	1.19 (2)	53/30
10	1.93 (21)	50 (20)	2.58 (23)	55 (19)	0.30 (12)	0.88 (12)	4.6 (8)	39 (15)	19.3 (14)	2.85 (10)	154/64
11	2.78 (26)	37 (14)	3.16 (28)	61 (20)	0.36 (16)	1.06 (13)	9.1 (22)	49 (18)	34.2 (21)	3.65 (15)	193/89
12	1.12 (4)	37 (12)	2.15 (15)	32 (13)	0.30 (13)	1.21 (14)	8.1 (18)	NA	NA	NA	–/51
13	4.50 (29)	77 (23)	1.90 (14)	92 (22)	0.79 (21)	1.24 (15)	8.5 (21)	31 (13)	13.5 (7)	2.01 (5)	170/79
14	1.90 (20)	37 (13)	2.60 (24)	21 (6)	0.20 (3)	1.25 (16)	6.9 (15)	26 (8)	12.1 (6)	13.2 (22)	133/75
15	1.12 (3)	30 (6)	1.62 (10)	30 (11)	0.20 (2)	1.26 (17)	2.9 (3)	60 (21)	29.2 (19)	3.42 (13)	105/33
16	**0.91 (1)**	NA	1.43 (8)	39 (15)	0.43 (19)	1.34 (18)	15.5 (26)	51 (19)	33.4 (20)	3.47 (14)	–/53
17	2.99 (27)	45 (17)	1.07 (4)	51 (18)	0.33 (15)	1.34 (19)	6.8 (14)	22 (4)	13.6 (8)	**1.04 (1)**	127/64
18	1.45 (11)	37 (11)	2.31 (19)	18 (5)	0.32 (14)	1.35 (20)	3.7 (4)	30 (12)	20.3 (15)	3.36 (12)	123/54
19	1.63 (14)	27 (4)	2.46 (21)	34 (14)	0.45 (20)	1.36 (21)	18.8 (27)	32 (14)	17.5 (12)	4.35 (17)	164/83
20	1.53 (13)	41 (15)	2.18 (16)	50 (17)	0.29 (10)	1.43 (22)	8.4 (20)	52 (20)	21.8 (16)	2.70 (8)	157/71
21	2.23 (22)	**25 (1)**	2.62 (25)	28 (9)	0.21 (5)	1.56 (23)	9.3 (23)	29 (11)	15.8 (11)	4.55 (18)	148/93
22	1.75 (17)	NA	2.73 (26)	32 (12)	0.36 (17)	1.59 (24)	4.1 (7)	43 (17)	18.0 (13)	1.64 (4)	–/74
23	1.67 (15)	47 (19)	2.47 (22)	NA	NA	1.61 (25)	14.3 (25)	NA	NA	NA	–/87
24	2.69 (25)	36 (10)	1.64 (11)	47 (16)	0.40 (18)	1.68 (26)	8.1 (19)	25 (7)	22.1 (17)	3.17 (11)	160/81
25	1.04 (2)	100 (24)	2.42 (20)	NA	NA	1.80 (27)	4.8 (11)	43 (16)	NA	5.73 (19)	–/60
26	1.52 (12)	112 (25)	3.76 (29)	NA	NA	2.17 (28)	4.8 (10)	NA	NA	NA	–/79
27	2.37 (24)	NA	3.07 (27)	NA	NA	3.63 (29)	20.3 (29)	NA	NA	NA	–/109
e-median	1.20 (6)	27 (3)	1.20 (6)	15 (3)	0.25 (7)	**0.41 (1)**	2.8 (2)	22 (5)	**8.8 (1)**	1.57 (3)	**37/15**
e- mean	1.29 (8)	27 (5)	1.19 (5)	**13 (1)**	0.24 (6)	0.49 (3)	**2.2 (1)**	23 (6)	9.8 (3)	2.32 (7)	45/17
Average over all models	1.90	47	2.07	39	0.35	1.25	8.5	38	18.7	3.93	–

Results are based on 27 different wheat crop models for LAI, AGBM, GY and HI, 20 for AGN, GN, GPC and NNI, 24 for PASW, and 19 for NHI.

[*] The models are sorted from top to bottom in the order of increasing RMSE for GY. For each variable the model with the lowest RMSE is in bold type.

[¶] NA, variables not available for a model. The numbers in parentheses indicate the rank of the models (including e-mean and e-median) for each variable.

[$] Sum of rank of RMSE for all variables/sum of rank of RMSE for the variables simulated by all 27 models (i.e., LAI, AGBM, GY, HI).

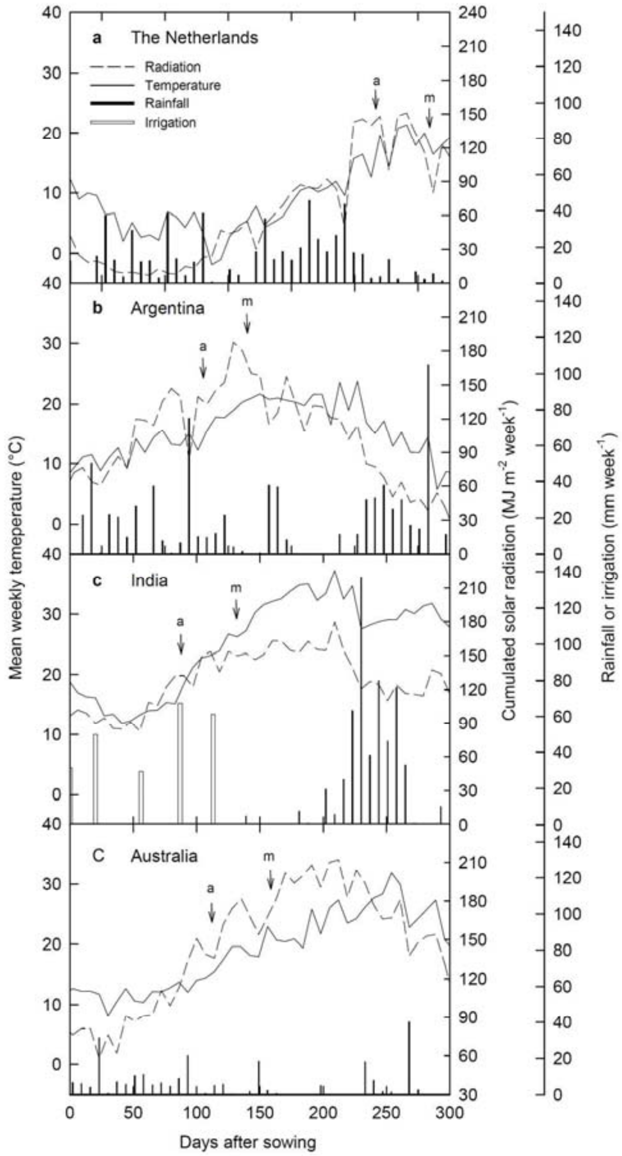


Figure S1. Weather data at the four studied sites. Mean weekly temperature (solid lines), cumulative weekly solar radiation (dashed lines), cumulative weekly rainfall (vertical solid bars) and irrigation (vertical open bars) in (**a**) Wageningen, The Netherlands, (**b**) Balcarce, Argentina, (**c**) New Delhi, India, and (**d**) Wongan Hills, Australia. Vertical arrows indicate (a) anthesis and (m) physiological maturity dates.

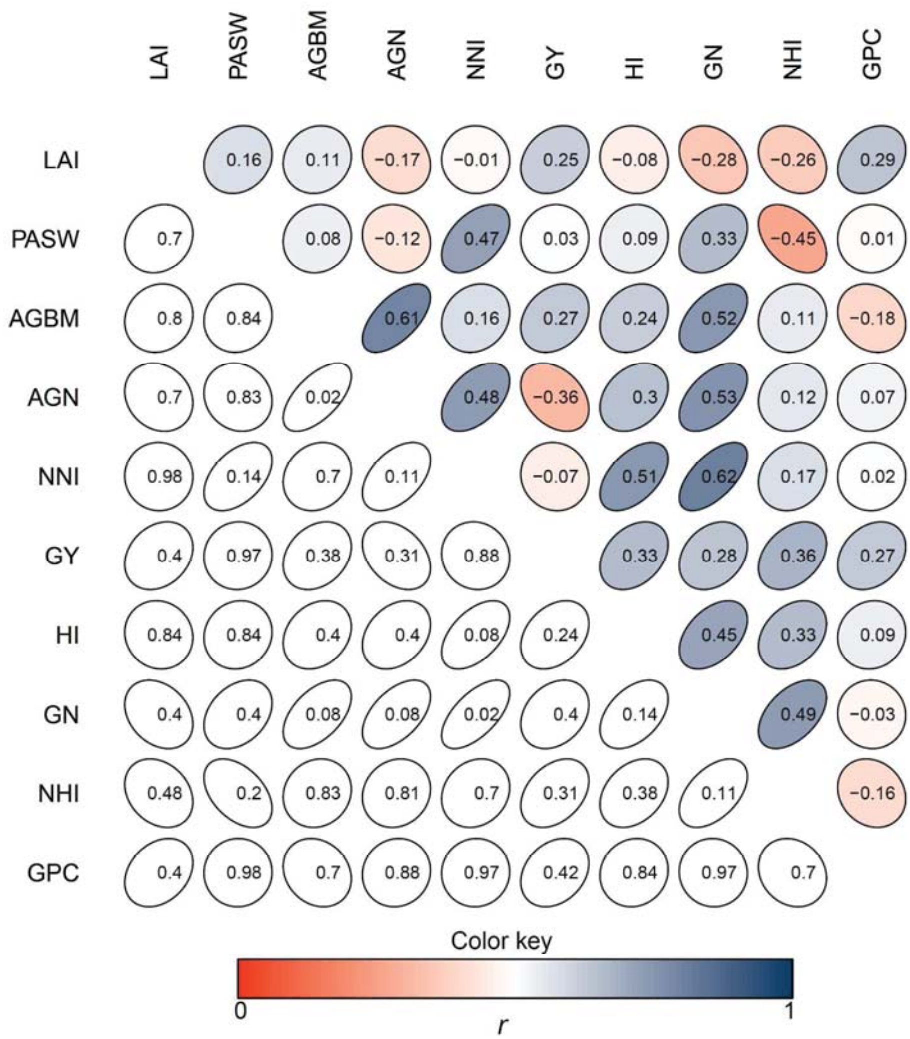


Figure S2. Correlation matrix for Pearson's product-moment correlation (*r*) between the root mean squared relative error of simulated variables. In-season variables: leaf area index (LAI), plant-available soil water (PASW), total aboveground biomass (AGBM), total above ground nitrogen (AGN), nitrogen nutrition index (NNI). End-of-season variables: grain yield (GY), biomass harvest index (HI), grain nitrogen yield (GN), nitrogen harvest index (NHI), and grain protein concentration (GPC). Twenty-seven models were used to simulate LAI, AGBM, GY, and HI, 20 to simulate AGN, GN, GPC and NNI, 24 to simulate PASW, and 19 to simulate NHI. The numbers above the diagonal gap are *r* values and the numbers below are one-sided *q*-values (adjusted *P*-values for false discovery rate). The color (for *r* values only) and the shape of the ellipses indicate the strength (the narrower the ellipse the higher the *r* value) and the direction of the correlation, respectively.

References

Aggarwal PK, Banerjee B, Daryaei MG *et al.* (2006) InfoCrop: A dynamic simulation model for the assessment of crop yields, losses due to pests, and environmental impact of agro-ecosystems in tropical environments. II. Performance of the model. *Agricultural Systems,* **89**, 47-67.

Angulo C, Rötter R, Lock R, Enders A, Fronzek S, Ewert F (2013) Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe. *Agricultural and Forest Meteorology,* **170**, 32-46.

Asseng S, Ewert F, Rosenzweig C *et al.* (2013) Uncertainty in simulating wheat yields under climate change. *Nature Climate Change,* **3**, 827-832.

Asseng S, Jamieson PD, Kimball B, Pinter P, Sayre K, Bowden JW, Howden SM (2004) Simulated wheat growth affected by rising temperature, increased water deficit and elevated atmospheric CO_2. *Field Crops Research,* **85**, 85-102.

Asseng S, Keating BA, Fillery IRP *et al.* (1998) Performance of the APSIM-wheat model in Western Australia. *Field Crops Research,* **57**, 163-179.

Basso B, Cammarano D, Troccoli A, Chen D, Ritchie JT (2010) Long-term wheat response to nitrogen in a rainfed Mediterranean environment: Field data and simulation analysis. *European Journal of Agronomy,* **33**, 132-138.

Berntsen J, Petersen BM, Jacobsen BH, Olesen JE, Hutchings NJ (2003) Evaluating nitrogen taxation scenarios using the dynamic whole farm simulation model FASSET. *Agricultural Systems,* **76**, 817-839.

Biernath C, Gayler S, Bittner S, Klein C, Högy P, Fangmeier A, Priesack E (2011) Evaluating the ability of four crop models to predict different environmental impacts on spring wheat grown in open-top chambers. *European Journal of Agronomy,* **35**, 71-82.

Bondeau A, Smith PC, Zaehle S *et al.* (2007) Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology,* **13**, 679-706.

Boogaard HL, Van Diepen CA, Rötter RP, Cabrera JCMA, Van Laar HH (eds) (1998) *User's guide for the WOFOST 7.1 crop growth simulation model and WOFOST control center 1.5.,* Wageningen, The Netherlands, Winand Staring Centre.

Brisson N, Gary C, Justes E *et al.* (2003) An overview of the crop model STICS. *Agronomy Journal,* **18**, 309-332.

Brisson N, Launay M, Mary B, Beaudoin N (2009) *Conceptual basis, formalisations and parameterization of the stics crop model* Paris, France, Quae.

Brisson N, Mary B, Ripoche D *et al.* (1998) STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. *Agronomie,* **18**, 311-346.

Brisson N, Ruget F, Gate P *et al.* (2002) STICS: a generic model for simulating crops and their water and nitrogen balances. II. Model validation for wheat and maize. *Agronomie,* **22**, 69-92.

Challinor AJ, Wheeler TR, Craufurd PQ, Slingo JM, Grimes DIF (2004) Design and optimisation of a large-area process-based model for annual crops. *Agricultural and Forest Meteorology,* **124**, 99-120.

Ferrise R, Triossi A, Stratonovitch P, Bindi M, Martre P (2010) Sowing date and nitrogen fertilisation effects on dry matter and nitrogen dynamics for durum wheat: An experimental and simulation study. *Field Crops Research,* **117**, 245-257.

Goudriaan J, Van Laar HH (1994) Modelling potential crop growth processes: Textbook with exercices. pp 238. Dordrecht, The Netherlands, Kluwer Academic Publishers.

Grant RF, Kimball BA, Conley MM, White JW, Wall GW, Ottman MJ (2011) Controlled warming effects on wheat growth and yield: field measurements and modeling. *Agronomy Journal,* **103**, 1742-1754.

He J, Le Gouis J, Stratonovitch P *et al.* (2012) Simulation of environmental and genotypic variations of final leaf number and anthesis date for wheat. *European Journal of Agronomy,* **42**, 22-33.

Hoogenboom G, White JW (2003) Improving physiological assumptions of simulation models by using gene-based approaches. *Agronomy Journal,* **95**, 82-89.

Hunt LA, Pararajasingham S (1995) CROPSIM-WEHAT: a model describing the growth and development of wheat. *Canadian Journal of Plant Science,* **75**, 619-632.

Izaurralde RC, Mcgill WB, J.R. W (2012) Development and application of the EPIC model for carbon cycle, greenhouse-gas mitigation, and biofuel studies. In: *Managing agricultural greenhouse gases: coordinated agricultural research through GRACEnet to address our changing climate.* (eds Franzluebbers A, Follett R, Liebig M) pp 409-429. Amsterdam, The Netherlands, Elsevier.

Jamieson PD, Berntsen J, Ewert F *et al.* (2000) Modelling CO_2 effects on wheat with varying nitrogen supplies. *Agriculture, Ecosystems and Environment,* **82**, 27-37.

Jamieson PD, Semenov MA, Brooking IR, Francis GS (1998) *Sirius*: a mechanistic model of wheat response to environmental variation. *European Journal of Agronomy,* **8**, 161-179.

Jones JW, Hoogenboom G, Porter CH *et al.* (2003) The DSSAT cropping system model. *European Journal of Agronomy,* **18**, 235-265.

Keating BA, Carberry PS, Hammer GL *et al.* (2003) An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy,* **18**, 267-288.

Kersebaum K (2007) Modelling nitrogen dynamics in soil–crop systems with HERMES. *Nutrient Cycling in Agroecosystems,* **77**, 39-52.

Kersebaum KC (2011) Special features of the HERMES model and additional procedures for parameterization, calibration, validation, and applications. In: *Methods of introducing system models into agricultural research.* (eds Ahuja LR, Ma L) pp 65-94. Madison, WI, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Kiniry JR, Williams JR, Major DJ *et al.* (1995) EPIC model parameters for cereal, oilseed, and forage crops in the northern Great Plains region. *Canadian Journal of Plant Science,* **75**, 679-688.

Lawless C, Semenov MA, Jamieson PD (2005) A wheat canopy model linking leaf area and phenology. *European Journal of Agronomy,* **22**, 19-32.

Li S, Wheeler T, Challinor A, Erda L, Xu Y, Hui J (2010) Simulating the impacts of global warming on wheat in China using a large area crop model. *Acta Meteorologica Sinica,* **24**, 123-125.

Martre P, Jamieson PD, Semenov MA, Zyskowski RF, Porter JR, Triboi E (2006) Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. *European Journal of Agronomy,* **25**, 138-154.

Nendel C, Berg M, Kersebaum KC *et al.* (2011) The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. *Ecological Modelling,* **222**, 1614-1625.

O'leary GJ, Connor DJ (1996a) A simulation model of the wheat crop in response to water and nitrogen supply : I. Model construction. *Agricultural Systems,* **52**, 1-29.

O'leary GJ, Connor DJ (1996b) A simulation model of the wheat crop in responses to water and nitrogen supply : II. Model validation. *Agricultural Systems,* **52**, 31-55.

Olesen JE, Petersen BM, Berntsen J, Hansen S, Jamieson PD, Thomsen AG (2002) Comparison of methods for simulating effects of nitrogen on green area index and dry matter growth in winter wheat. *Field Crops Research,* **74**, 131-149.

Priesack E, Gayler S, Hartmann HP (2006) The impact of crop growth sub-model choice on simulated water and nitrogen balances. *Nutrient Cycling in Agroecosystems,* **75**, 1-13.

Ritchie JT, Otter S (1985) Description of and performance of CERES-Wheat: A user-oriented wheat yield model. In: *ARS Wheat Yield Project.* (ed Willis WO) pp 159-175. Washington, DC, Department of Agriculture, Agricultural Research Service.

Senthilkumar S, Basso B, Kravchenko AN, Robertson GP (2009) Contemporary evidence of soil carbon loss in the US corn belt. *Soil Science Society of America Journal,* **73**, 2078-2086.

Shibu ME, Leffelaar PA, Van Keulen H, Aggarwal PK (2010) LINTUL3, a simulation model for nitrogen-limited situations: Application to rice. *European Journal of Agronomy,* **32**, 255-271.

Steduto P, Hsiao TC, Raes D, Fereres E (2009) AquaCrop—The FAO crop model to simulate yield response to water: I. Concepts and underlying principles. *Agronomy Journal,* **101**, 426-437.

Stenger R, Priesack E, Barkle GF, C. S (1999) Expert-N - A tool for simulating nitrogen and carbon dynamics in the soil-plant-atmosphere system. In: *Proceedings of the Technical Session No 20.* (eds Tomer M, Robinson M, Gielen G) pp 19-28, New Zealand Land Treatment Collective.

Stöckle CO, Donatelli M, Nelson R (2003) CropSyst, a cropping systems simulation model. *European Journal of Agronomy,* **18**, 289-307.

Tao F, Yokozawa M, Zhang Z (2009) Modelling the impacts of weather and climate variability on crop productivity over a large area: A new process-based model development, optimization, and uncertainties analysis. *Agricultural and Forest Meteorology,* **149**, 831-850.

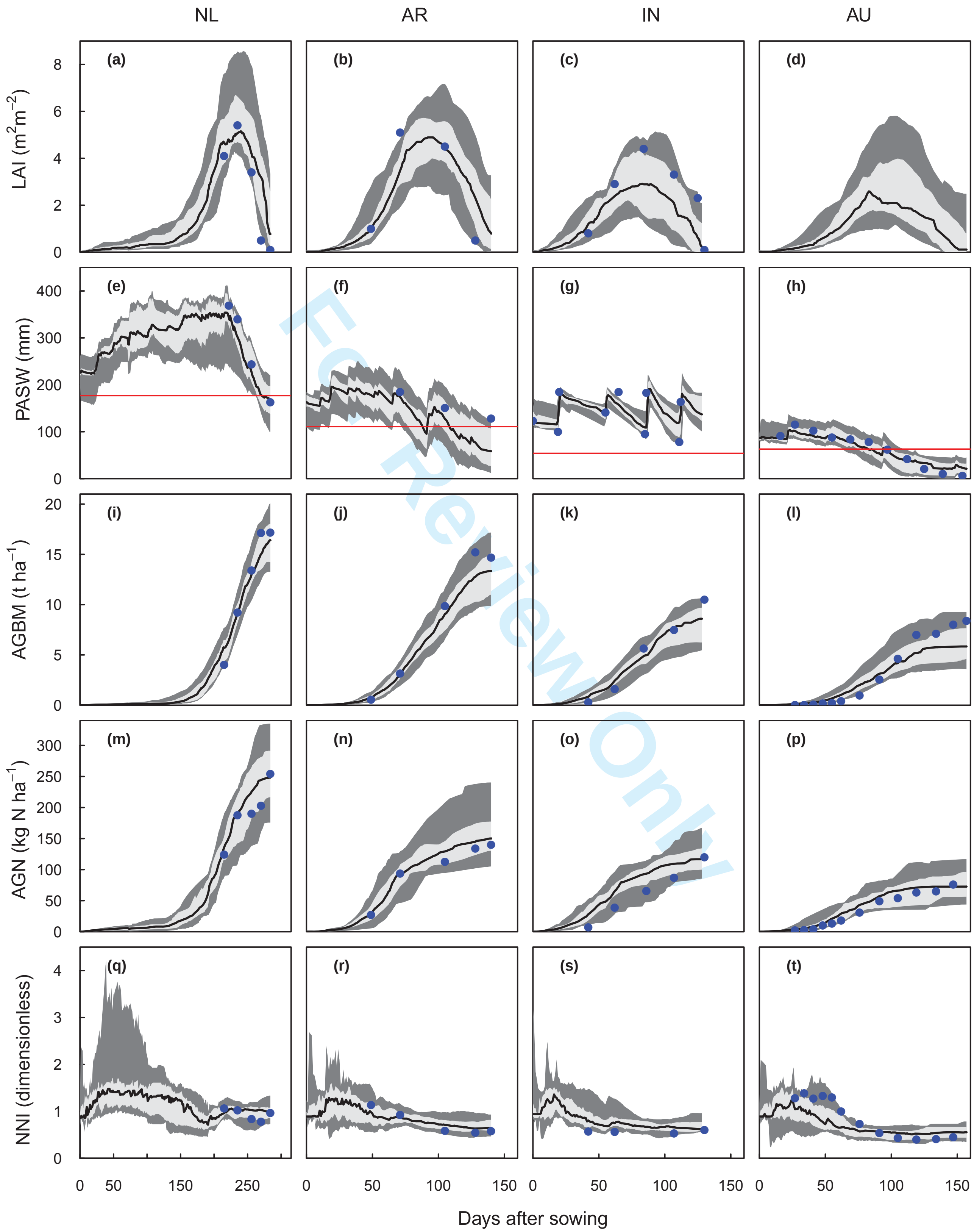
Van Diepen CA, Wolf J, Van Keulen H, Rappoldt C (1989) WOFOST: a simulation model of crop production. *Soil Use and Management,* **5**, 16-24.

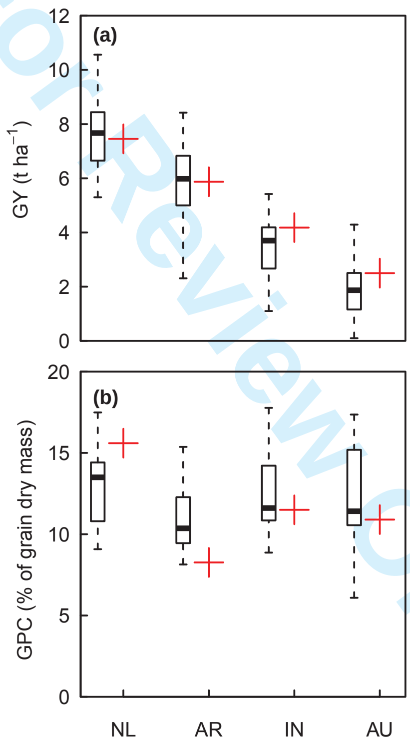
Wang E, Engel T (2000) SPASS: a generic process-oriented crop model with versatile windows interfaces. *Environmental Modelling and Software,* **15**, 179-188.

Williams JR, Jones CA, Kiniry JR, Spanel DA (1989) The EPIC crop growth model. *Transactions of the ASAE,* **32**, 497-511.

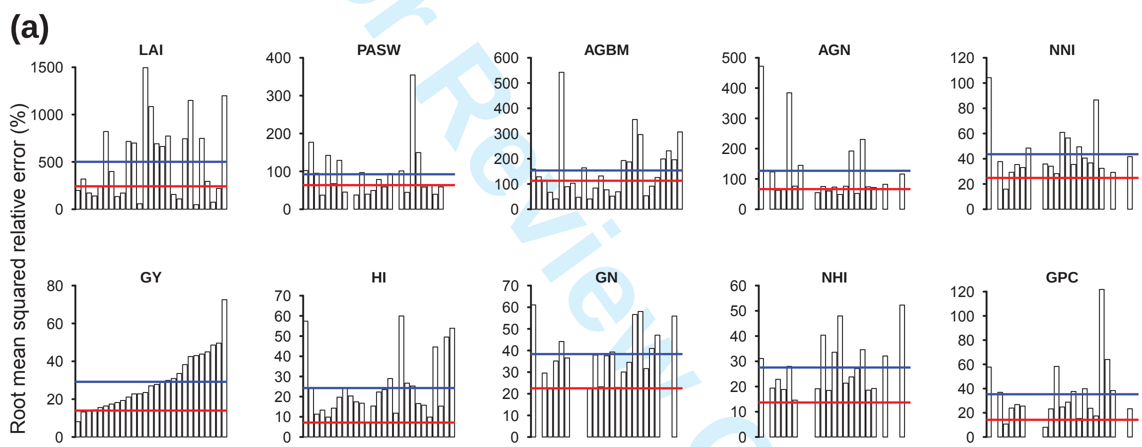
Yin X, Van Laar HH (2005) *Crop systems dynamics: an ecophysiological simulation model for genotype-by-environment interactions,* Wageningen, the Netherlands Wageningen Academic Publishers.

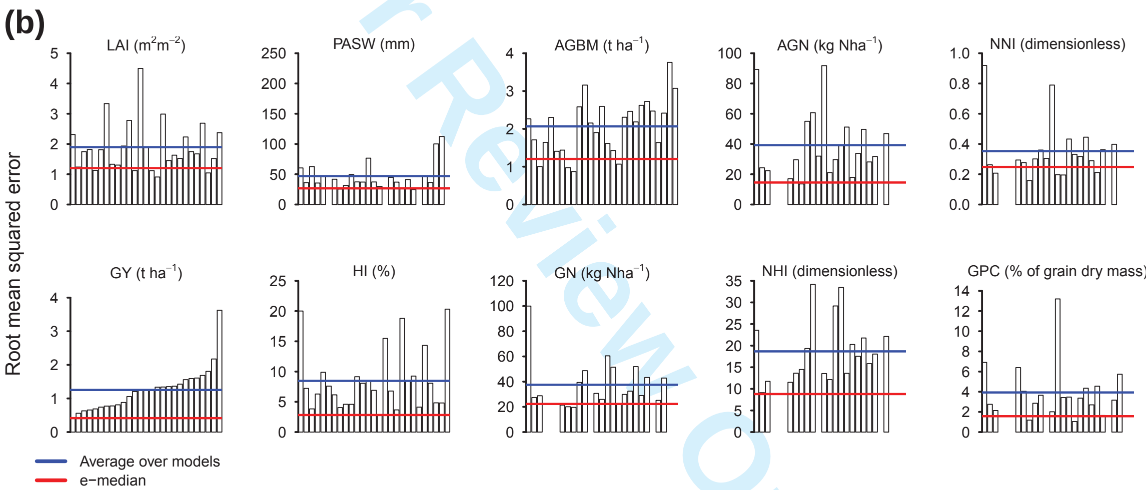
Zadoks JC, Chang TT, Konzak CF (1974) A decimal code for the growth stages of cereals. *Weed Research,* **14**, 415-421.

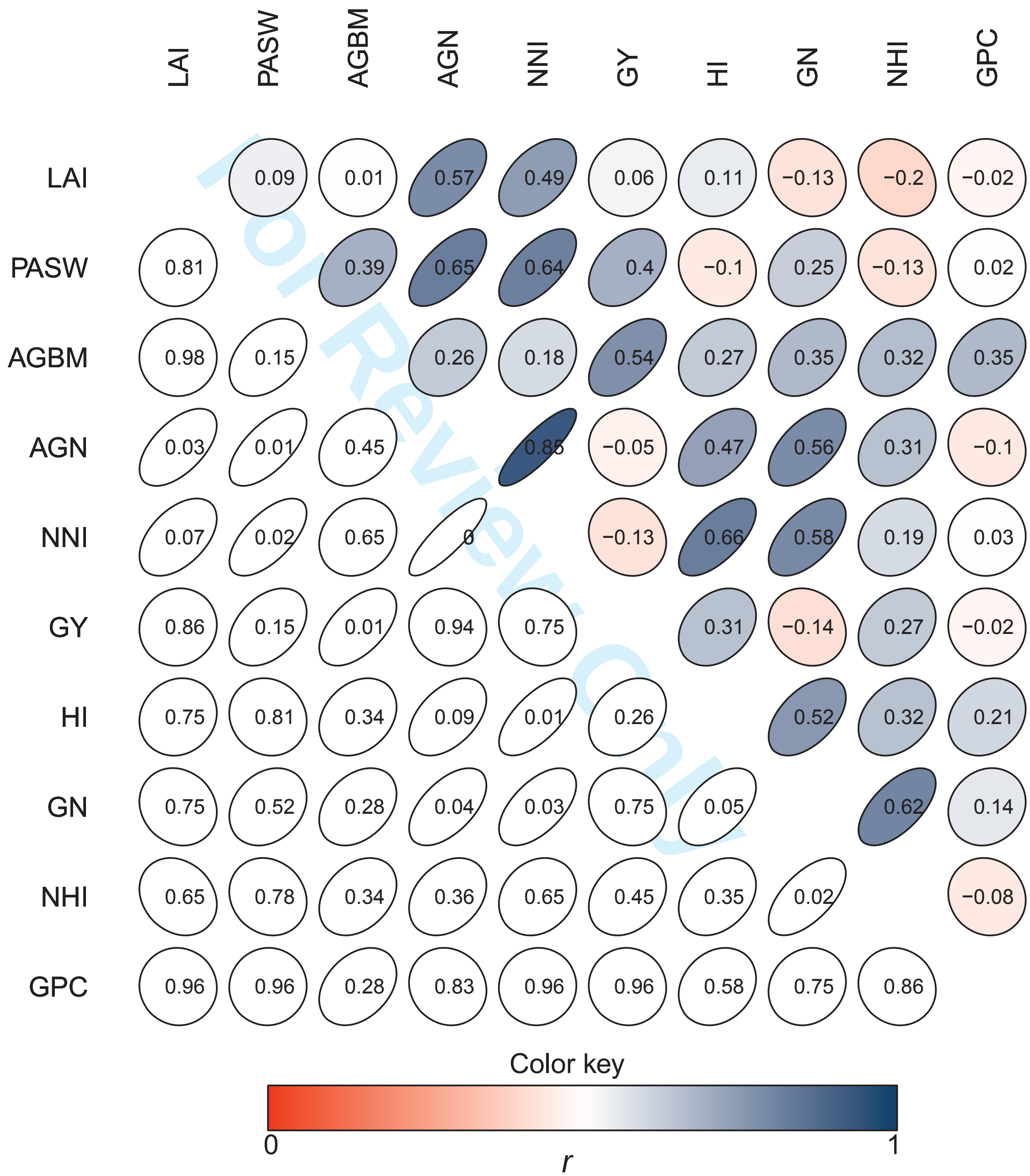

NL
AR
IN
AU
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)
LAI ($m^2 m^{-2}$)
PASW (mm)
AGBM (t ha^{-1})
AGN (kg N ha^{-1})
NNI (dimensionless)
Days after sowing

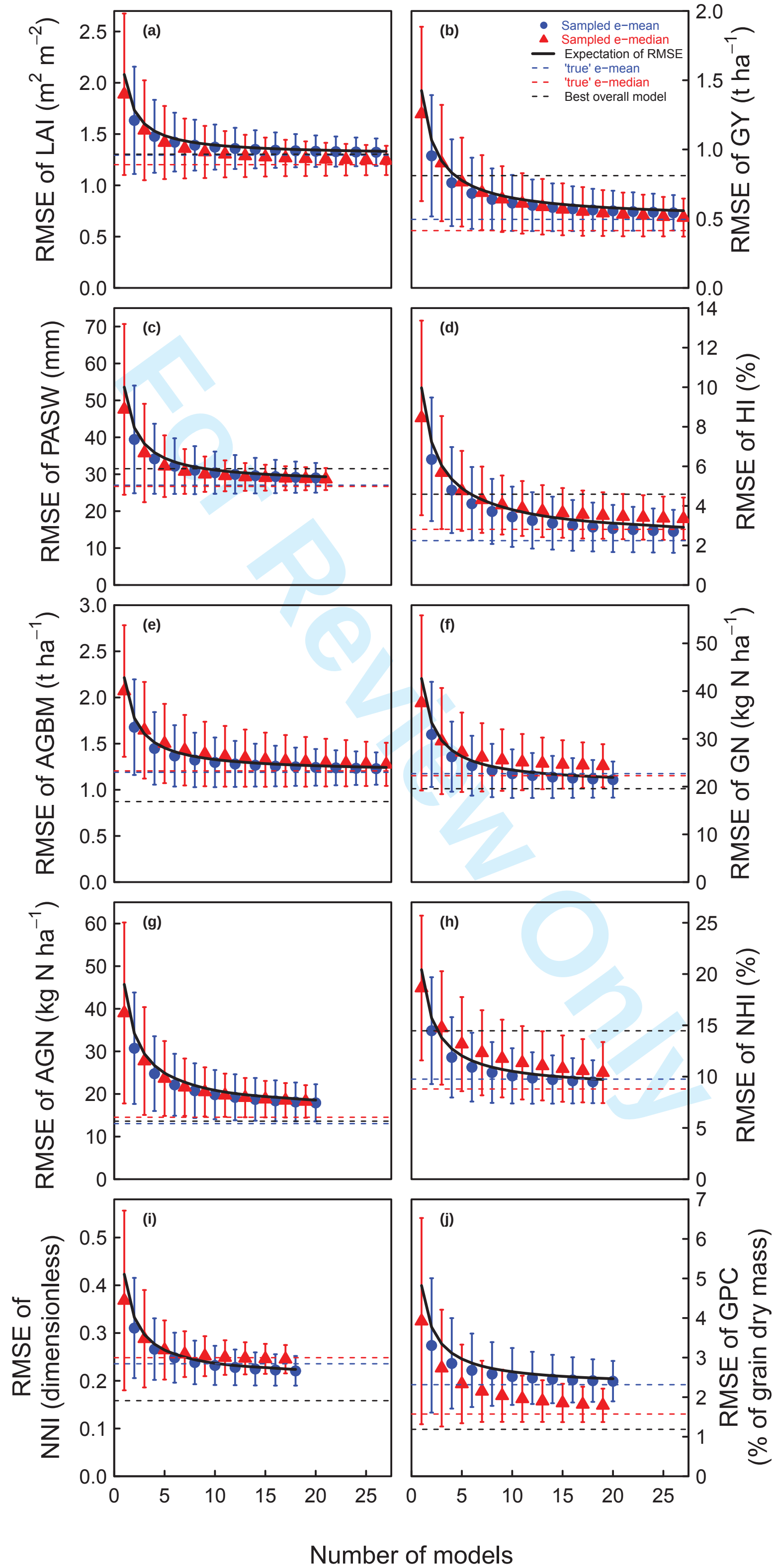


(a)
(b)
GY (t ha^{-1})
GPC (% of grain dry mass)
12
10
8
6
4
2
0
20
15
10
5
0
NL
AR
IN
AU


(a)
LAI
PASW
AGBM
AGN
NNI
GY
HI
GN
NHI
GPC
Root mean squared relative error (%)


(b)
LAI (m^{2}m^{-2})
PASW (mm)
AGBM (t ha^{-1})
AGN (kg Nha^{-1})
NNI (dimensionless)
GY (t ha^{-1})
HI (%)
GN (kg Nha^{-1})
NHI (dimensionless)
GPC (% of grain dry mass)
Root mean squared error
Average over models
e−median


LAI
PASW
AGBM
AGN
NNI
GY
HI
GN
NHI
GPC
Color key
0
1
r


Sampled e−mean
Sampled e−median
Expectation of RMSE
'true' e−mean
'true' e−median
Best overall model
RMSE of LAI (m2 m−2)
RMSE of GY (t ha−1)
RMSE of PASW (mm)
RMSE of HI (%)
RMSE of AGBM (t ha−1)
RMSE of GN (kg N ha−1)
RMSE of AGN (kg N ha−1)
RMSE of NHI (%)
RMSE of NNI (dimensionless)
RMSE of GPC (% of grain dry mass)
Number of models
For Review Only