

Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles

Andrea Maiorano, Pierre Martre, Senthold Asseng, Frank Ewert, Christoph Müller, Reimund P. Rötter, Alex C. Ruane, Mikhail A. Semenov, Daniel Wallach, Enli Wang, Phillip D. Alderman, Belay T. Kassie, Christian Biernath, Bruno Basso, Davide Cammarano, Andrew J. Challinor, Jordi Doltra, Benjamin Dumont, Ehsan Eyshi Rezaei, Sebastian Gayler, Kurt Christian Kersebaum, Bruce A. Kimball, Ann-Kristin Koehler, Bing Liu, Garry J. O’Leary, Jørgen E. Olesen, Michael J. Ottman, Eckart Priesack, Matthew Reynolds, Pierre Stratonovitch, Thilo Streck, Peter J. Thorburn, Katharina Waha, Gerard W. Wall, Jeffrey W. White, Zhigan Zhao, Yan Zhu

Angaben zur Veröffentlichung / Publication details:

Maiorano, Andrea, Pierre Martre, Senthold Asseng, Frank Ewert, Christoph Müller, Reimund P. Rötter, Alex C. Ruane, et al. 2017. “Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles.” *Field Crops Research* 202: 5-20.
<https://doi.org/10.1016/j.fcr.2016.05.001>.

Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles

Andrea Maiorano^{a,*}, Pierre Martre^{a,*}, Senthold Asseng^b, Frank Ewert^{c,1}, Christoph Müller^d, Reimund P. Rötter^{e,2}, Alex C. Ruane^f, Mikhail A. Semenov^g, Daniel Wallach^h, Enli Wangⁱ, Phillip D. Alderman^{j,5,6}, Belay T. Kassie^b, Christian Biernath^k, Bruno Basso^l, Davide Cammarano^{b,3}, Andrew J. Challinor^{m,n}, Jordi Doltra^o, Benjamin Dumont^l, Ehsan Eyshi Rezaei^{c,y}, Sebastian Gayler^p, Kurt Christian Kersebaum^q, Bruce A. Kimball^r, Ann-Kristin Koehler^m, Bing Liu^s, Garry J. O'Leary^t, Jørgen E. Olesen^u, Michael J. Ottman^v, Eckart Priesack^k, Matthew Reynolds^j, Pierre Stratonovitch^g, Thilo Streck^w, Peter J. Thorburn^x, Katharina Waha^{d,4}, Gerard W. Wall^r, Jeffrey W. White^r, Zhigan Zhao^{i,z}, Yan Zhu^s

^a UMR LEPSE, INRA, Montpellier SupAgro, 2 Place Viala, 34 060 Montpellier, France

^b Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL-32611, USA

^c Institute of Crop Science and Resource Conservation, University of Bonn, D-53 115 Bonn, Germany

^d Potsdam Institute for Climate Impact Research, D-14 473 Potsdam, Germany

^e Natural Resources Institute Finland (Luke), FI-01301 Vantaa, Finland

^f NASA Goddard Institute for Space Studies, New York, NY-10025, USA

^g Computational and Systems Biology Department, Rothamsted Research, Harpenden, Herts AL5 2JQ, UK

^h INRA, UMR 1248 Agrosystèmes et développement territorial, F-31 326, Castanet-Tolosan, France

ⁱ CSIRO Agriculture, Black Mountain, ACT 2601, Australia

^j CIMMYT Int. AP 6-641, D.F. Mexico 06600, Mexico

^k Institute of Biochemical Plant Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health, D-85764 Neuherberg, Germany

^l Department of Geological Sciences and W.K. Kellogg Biological Station, Michigan State University, East Lansing, MI-48 823, USA

^m Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK

ⁿ CGIAR-ESSP Program on Climate Change, Agriculture and Food Security, International Centre for Tropical Agriculture, A.A. 6713, Cali, Colombia

^o Cantabrian Agricultural Research and Training Centre, 39600 Muriedas, Spain

^p Institute of Soil Science and Land Evaluation, University of Hohenheim, D-70 599 Stuttgart, Germany

^q Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research, D15 374 Müncheberg, Germany

^r USDA, Agricultural Research Service, US Arid-Land Agricultural Research Center, Maricopa, AZ 85138, USA

^s College of Agriculture, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China

^t Grains Innovation Park, Department of Economic Development Jobs, Transport and Resources, Horsham 3400, Australia

^u Department of Agroecology, Aarhus University, 8830 Tjele, Denmark

^v The School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

^w Institute of Soil Science and Land Evaluation, University of Hohenheim, D-70 599 Stuttgart, Germany

^x CSIRO Agriculture, 306 Carmody Road, St Lucia Queensland 4067, Australia

^y Center for Development Research (ZEF), Walter-Flex-Straße 3, 53113 Bonn, Germany

^z China Agricultural University, Beijing 100193, China

* Corresponding authors at: UMR LEPSE, INRA, Montpellier SupAgro, 2 Place Viala, 34 060 Montpellier, France.

E-mail addresses: maiorano.andrea@gmail.com (A. Maiorano), pierre.martre@supagro.inra.fr (P. Martre).

¹ Present address: Leibniz Centre for Agricultural 36 Landscape Research (ZALF), D15374 Müncheberg, Germany.

² Present address: Department of Crop Sciences, Division Crop Production Systems in the Tropics, Georg-August-Universität Göttingen, D-37077 Göttingen, Germany

³ Present address: James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA, UK.

⁴ Present address: CSIRO Agriculture, 306 Carmody Rd, 4067 St Lucia, Australia.

⁵ Present address: Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK, 74078-6028, USA.

⁶ Starting from P.D.A. the author list is in alphabetical order.

1. Introduction

Wheat is the most widely grown crop in the world and provides more than 20% of the daily protein and food calories for the world population (Shiferaw et al., 2013). With a predicted world population of 9 billion in 2050, the demand for food including wheat is expected to increase by then (Alexandratos and Bruinsma, 2012). Climate trends are significantly affecting agricultural production systems, including wheat, in several regions of the world, thereby posing risks to global food supply and security (Sundström et al., 2014). Therefore, quantifying the potential impact of climate variability on crops has become a priority in order to develop effective adaptation and mitigation strategies (Burton and Lim, 2005; Denton et al., 2014).

Process-based crop simulation models are useful tools to assess the impact of climate as they consider the interaction between climate variables and crop management and their effects on crop productivity. Their use in climate impact studies and for analyzing and developing adaptation and mitigation strategies has increased during the recent years (Byjesh et al., 2010; Donatelli et al., 2012; Moradi et al., 2013; Rosenzweig et al., 2014). Nevertheless, most of the current crop models lack explicit definitions of relevant physiological thresholds and crop responses to extreme weather events, particularly for temperatures exceeding these thresholds (Rötter et al., 2011). These omissions might be one of the reason for the considerable differences in estimates of grain yield observed among models especially for high temperatures, and between models and field observations (Palosuo et al., 2011). In addition, since a clear methodology is lacking, most climate change impact assessments for agriculture have not addressed crop model uncertainties (Müller, 2011), which have become a major concern recently in climate impact assessments (Lobell et al., 2006; Ruane et al., 2013; Zhang et al., 2015).

White et al. (2011) reported that over 40 wheat crop models are in use worldwide. They differ in the processes they include, or in the modelling approaches used to simulate physiological processes. A recent work carried out by the Wheat Team of the Agricultural Model Inter-comparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013) compared 27 wheat crop models and showed that a greater portion of the uncertainty in climate change impact projections was due to variations among crop models than to variations among climate models, and that uncertainties in simulated yield increased dramatically under high temperature conditions (Asseng et al., 2013). Following the example of the climate modelling community, to increase reliability of impact estimates and to give better estimates of uncertainty, use of crop multi-model ensembles (MME) has been suggested (Asseng et al., 2015; Bassu et al., 2014; Li et al., 2015; Pirttioja et al., 2015). Model improvements have been suggested for improving the accuracy of

simulations and reducing the uncertainty of climate impact assessments (Asseng et al., 2013; Challinor et al., 2014; Rötter et al., 2011). Martre et al. (2015) argued that one of the consequences of model improvements will be the reduction of the number of models required for an acceptable level of simulation uncertainty. Furthermore, the improvement of the models in an ensemble using good quality field-based experimental data could substantially widen the range of research questions to be addressed and increase the confidence in simulation results of applications under changed climatic or management conditions (Martre et al., 2015).

Herein, we investigated the effects of model improvements in 15 wheat crop models with regards to heat stress and its impact on model performances, uncertainty, and the number of crop models required in multi-model ensembles used for impact studies.

2. Materials and methods

2.1. Experimental data

Detailed quality-assessed data from the USDA 'Hot Serial Cereal' (HSC) experiment (Grant et al., 2011; Kimball et al., 2015; Ottman et al., 2012; Wall et al., 2011) and from the 'International Heat Stress Genotype Experiment' (IHSGE) coordinated by CIMMYT (Reynolds et al., 1994b) were used. Both experiments were well watered and fertilized to avoid drought and nutritional stress to assure that temperature would be the main environmental variable. Daily global solar radiation, maximum and minimum air temperature, average wind speed, dew point temperature and precipitation were recorded at weather stations near the experimental plots. The mean daily average air temperature for the growing season (sowing to physiological maturity) was calculated from minimum and maximum daily air temperatures as described in Asseng et al. (2015) and reported in Supplementary Information S2. In both experiments phenological development measurements included: emergence date (Zadock scale 10), anthesis date (Zadock scale 65), and maturity date (Zadock scale 89). From these measurements the number of days from sowing to anthesis (days), from anthesis to maturity (days), and from sowing to maturity (days) were calculated. In both experiments, the plots were kept weed-free, and plant protection methods were used as necessary to minimize damage from pest and diseases. The two data sets are further described in Asseng et al. (2015). Following is a brief description with focus on the measurement data that were available for this study.

The HSC experiment was conducted at Maricopa, AZ, USA (33.07°N, 111.97°W, 361 m a.s.l.): The spring wheat cultivar 'Yecora Rojo' was sown about every six weeks for two years, and infrared heaters were deployed on six of the sowing dates in a T-FACE (temperature free-air controlled enhancement) system which warmed the canopies of the heated plots on average by 1.3 °C and 2.7 °C

during the day and the night, respectively (targets were 1.5°C and 3.0°C; modes were 1.4°C and 3.0°C; Kimball et al., 2015). Yecora Rojo is of short stature, requires little to no vernalization, is not or little photoperiod sensitive, and matures early (Qualset et al., 1985). In-season measurements included leaf area index (LAI, m² m⁻²), total above ground dry biomass, dry matter weight of grain per square meter and nitrogen content measured at milk stage and maturity. End-of-season (i.e. ripeness-maturity) measurements were total above ground dry biomass (t DM ha⁻¹), grain yield (t DM ha⁻¹), single grain dry mass (mg DM grain⁻¹), and grain number (grain m⁻²). Biomass harvest index was calculated as HI = 100 × (grain yield)/(above ground biomass) (%).

Data from the IHSGE experiments used in this study includes two spring wheat cultivars (Bacanora 88 and Nesser) grown during the 1990–1991 and 1991–1992 winter cropping cycles at hot, irrigated, and low latitude sites in Mexico (Ciudad Obregon, 27.34°N, 109.92°W, 38 m a.s.l.; and Tlatizapan, 19.69°N, 99.13°W, 940 m a.s.l.), Egypt (Aswan, 24.1°N, 32.9°E, 200 m a.s.l.), India (Dharwar, 15.49°N, 74.98°E, 940 m a.s.l.), Sudan (Wad Medani, 14.40°N, 33.49°E, 411 m a.s.l.), Bangladesh (Dinajpur, 25.65°N, 88.68°E, 29 m a.s.l.), and Brazil (Londrina, 23.34°S, 51.16°W, 540 m a.s.l.) (Reynolds, 1993; Reynolds et al., 1994a,b). Experiments in Mexico included normal (December) and late (March) sowing dates. Bacanora 88 has moderate vernalization requirement and low photoperiod sensitivity and Nesser has low to no vernalization requirement and photoperiod sensitivity. The seven sites (out of the original 12 locations) were chosen to represent a range of temperature as detailed in Asseng et al. (2015). Bacanora 88 and Nesser were chosen (out of the original 16 cultivars) for their low photoperiod sensitivity and low vernalization requirements. Variables measured in the experiment included plant number per square meter, anthesis and final above ground biomass, final grain yield and yield components (number of ear per square meter, number of grain per ear, and single grain dry mass). These experimental data were not publicly available and could therefore be used in a blind model evaluation.

2.2. Model inter-comparison and improvement protocols

Of the 30 models that participated in the original study using the HSC data (Asseng et al., 2015), 15 models accepted to participate in this new study. There was no explicit criterion of inclusion, so this would be an “ensemble of opportunity” as defined in the climate model community (Tebaldi and Knutti, 2007). All of the models have been described in publications and are currently in use. For the evaluation data set measurements, above ground biomass and grain yield were simulated by all the models. 7 out of 15 models did not simulated single grain dry mass and grain number but used a harvest index approach.

For both experiments, all modeling groups were provided with daily weather data, crop management, soil, and cultivar information. Qualitative information on vernalization requirements and day length response for each cultivar were also provided.

The HSC experiment (calibration data set) was used to improve the models. All available measurements from the HSC experiment were provided to modelers to improve and refine the parameterization and processes of their model. The objective was to improve wheat models for the simulation of the impact of high temperature and heat stresses on crop development and growth. Modelling groups were allowed to decide how to improve and implement heat stress impact in their models.

The IHSGE experiment (evaluation data set) was used as independent evaluation data set to test single models and model ensemble performances before and after improvement. All measurements of the evaluation data set were withheld from modelers (blind test) with the exception of phenology for all treatments and

grain yield for one of the treatments (one year at Ciudad Obregon, Mexico) which was used to calibrate genotypic coefficients.

The experimental data used in this study were not previously used to develop or calibrate any of the 15 models used in this study. Except for the two Expert-N models which were executed by the same group, all models were simulated 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 executed their own models.

2.3. Evaluation of model improvement effects on single models and on multi-model ensemble accuracy

We evaluated the effect of model improvement on two different performance characteristics, *accuracy* and *uncertainty*, and on three model entities: (i) single models (accuracy only); (ii) multi-model ensemble (MME, the ensemble of 15 models in this experiment exercise); and (iii) MME median (e-median).

Accuracy was measured using the mean squared error (MSE), the root mean squared error (RMSE), and the root mean squared relative error (RMSRE).

For measuring single model error in reproducing the calibration and the evaluation data set we concentrated on the root mean squared relative error (RMSRE). This error indicator has the advantage of giving more equal weight to each measurement, and it's meaningful when comparing very different environments likely to give a broad range of responses (Martre et al., 2015). RMSRE was calculated as:

$$\text{RMSRE}_m = 100 \times \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{Y_i - \hat{Y}_{m,i}}{Y_i} \right)^2} \quad (1)$$

where RMSRE_m is the RMSRE of model m , i is the site/year/sowing dates combinations (treatment), N is the total number of treatments, Y_i is the observed variable for treatment i , $\hat{Y}_{m,i}$ is the variable simulated by model m for treatment i . Since this indicator is very sensitive to errors when measured values are small, RMSE was used as additional supporting information for a better understanding of RMSRE when needed. RMSE was calculated as:

$$\text{RMSE}_m = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_{m,i})^2} \quad (2)$$

where, RMSE_m is the RMSE of model m .

The accuracy of the population of 15 models before and after improvement was analyzed using the mean squared error (MSE) and its two components squared bias and variance, averaged across treatments:

$$\begin{aligned} \text{MSE}_{\text{MME}} &= \frac{1}{N} \frac{1}{M} \sum_{i=1}^N \sum_{m=1}^M (Y_i - \hat{Y}_{m,i})^2 \\ &= \frac{1}{N} \sum_{i=1}^N \text{var}_M(\hat{Y}_{m,i}) + \frac{1}{N} \sum_{i=1}^N (\text{bias}_M(\hat{Y}_{m,i}, Y_i))^2 \end{aligned} \quad (3)$$

where, MSE_{pm} is the MSE of the population of models in the ensemble, N is the number of treatments, M is the total number of models in the ensemble (i.e. 15), var_m and bias_M are the variance and the bias for the model population, respectively. From Eq. (3) it is evident that while bias is based on both observations and simulations, variance only takes into account simulated values.

2.4. Evaluation of model improvement effects on MME prediction uncertainty

To assess the MME prediction uncertainty we considered both the variability in MME and the comparison with hindcast (i.e. retrospective forecasts using known inputs and known field measurements) (Wallach et al., 2015) using the two available measurement data sets. In order to evaluate the prediction uncertainty of the MME before and after improvement we used the HSC calibration-data set to simulate model hindcast in respect to observed data, and the IHSGE experiment as the “unknown” data set used to simulate model prediction to unknown data and to evaluate the predictive skills of the models in the ensemble. As a measure of uncertainty we used the mean squared error of prediction (MSEP) and its decomposition in prediction squared bias ($\text{bias}_{\text{prediction}}^2$) and prediction variance ($\text{var}_{\text{prediction}}$). According to Wallach et al. (2016) the average squared error across treatments of MME-mean calculated using the known data set (hindcast) ($\text{MSE}_{\text{e-mean}}^{\text{hindcast}}$) can be used as a reference estimate of the model population squared bias when calculating prediction estimates. This corresponds to the average squared bias of hindcasts as calculated in Eq. (4):

$$\begin{aligned} \text{bias}_{\text{prediction}}^2 &= \text{MSE}_{\text{e-mean}}^{\text{hindcast}} \\ &= \frac{1}{N_{\text{hindcast}}} \sum_{i=1}^{N_{\text{hindcast}}} \left(Y_i^{\text{hindcast}} - \frac{1}{M} \sum_{m=1}^M \hat{Y}_{m,i}^{\text{hindcast}} \right)^2 = \text{bias}_{\text{hindcast}}^2 \end{aligned} \quad (4)$$

where, N_{hindcast} is the number of treatments in the known data set, Y_i^{hindcast} is the observed variable for treatment i of the known data set, $\hat{Y}_{m,i}^{\text{hindcast}}$ is the hindcast of the simulated variable for treatment i by the model m . The prediction variance $\text{var}_{\text{prediction}}$ is the variance of the values simulated by the population of models for the unknown data set averaged across treatments:

$$\text{var}_{\text{prediction}} = \frac{1}{N_{\text{prediction}}} \sum_{i=1}^{N_{\text{prediction}}} \text{var}_{M}(\hat{Y}_i^{\text{prediction}}) \quad (5)$$

where, $N_{\text{prediction}}$ is the number of treatments in the unknown data set, $Y_i^{\text{prediction}}$ is the simulated variable for the treatment i of the unknown data set. Therefore an estimate of MSEP can be composed as:

$$\text{MSEP} = \text{bias}_{\text{prediction}}^2 + \text{var}_{\text{prediction}} \quad (6)$$

2.5. Evaluation of model improvement effects on MME-median

Following Asseng et al. (2015) and Martre et al. (2015), we used the median of the model simulations (e-median) as the estimator of the ensemble model simulations. In order to evaluate the overall e-median accuracy we calculated the same criteria as for the individual models, namely RMSRE (Eq (1)).

To explore how the e-median and its error (RMSRE) varied with the number of models and with the random selection of models in the ensemble, we performed a bootstrap calculation (i.e. random sampling with replacement) for each value of M' (number of models in the ensemble) from 1 to 15. For each ensemble of size M' we drew 20×10^3 bootstrap samples (substantially higher than the 3200 samples found by Martre et al. (2015) as a sufficient number of samples for 27 models) of M' models with replacement, so the same model might be represented more than once in a sample. The variation of e-median across the bootstrap samples due

to random model selection was estimated with the coefficient of variation (CV):

$$\text{CV}(\hat{Y}_{\text{e-median},M'}) = \frac{1}{N} \sum_{i=1}^N \left(100 \times \frac{\text{sd}_B(\hat{Y}_{\text{e-median},i}^{M'})}{\text{mean}_B(\hat{Y}_{\text{e-median},i}^{M'})} \right) \quad (7)$$

where, $\text{CV}(\hat{Y}_{\text{e-median},M'})$ is the estimate of the coefficient of variation of e-median for the model ensemble of size M' , $\text{sd}_B(\hat{Y}_{\text{e-median},i}^{M'})$ and mean_B are the standard deviation and the mean of B (number of bootstrap samples) e-medians of model ensembles of size M' for the i th treatment. A benchmark CV of 13.5%, previously established through a meta-analysis of field trials (Taylor et al., 1999) was used to evaluate the minimum number of models required within a MME.

The final estimate of RMSRE for e-median was calculated as:

$$\text{RMSRE}_{M'} = \frac{1}{B} \sum_{b=1}^B 100 \times \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{y_i - \hat{y}_{\text{e-median},i}^b}{y_i} \right)^2} \quad (8)$$

where, $\text{RMSRE}_{M'}$ is the RMSRE of e-median of the model ensemble of M' size, $\hat{y}_{\text{e-median},i}^b$ is the e-median estimate in bootstrap sample b of the i th treatment.

All calculations and graphs were made using the R statistical software R 3.1.3 (R Core Team, 2013) and the development environment RStudio (RStudio Team, 2015). Bootstrap sampling used the R function `sample`.

3. Results

3.1. Individual model improvements

The major draw backs in simulating the HSC experiment were related to the impact of the higher temperature range ($T_{\text{mean}} > 22^\circ\text{C}$) on yield, biomass and phenology (Asseng et al., 2015). Furthermore it was shown that the few models that already included heat stress routines affecting canopy senescence were the only ones able to reproduce the impact of very high mean seasonal temperatures ($T_{\text{mean}} \geq 29^\circ\text{C}$) on grain yield and above ground biomass. Therefore, the process that received most attention was leaf senescence, followed by heat stress effects on processes related to biomass growth and/or phenological development, grain number and/or size, leaf development (Table 1, Fig. 1). Based on experimental evidences (e.g. Parent and Tardieu, 2012; Porter and Gawith, 1999), in several models linear temperature responses were replaced by non-linear (APSIM-E and *SiriusQuality*) or trapezoidal (APSIM-Wheat, GLAM-Wheat, Expert-N-SPASS, Expert-N-SUCROSS) response functions. The cardinal temperatures for these processes were fixed using values reported in the literature or calibrated using the HSC experimental data set. One model (APSIM-Nwheat) added a canopy temperature sub-routine. In addition to the inclusion/modification of heat stress impacts on physiological processes, five models improved processes not directly related to heat stress using the HSC data set or other published data sets (Table 1). One model (GLAM-wheat) removed the sub-routine for heat stress effect on grain set and potential harvest index as they observed no substantial improvement and decided not to increase the complexity of their model (Table 1 and Supplementary Methods).

In the case of heat stress impacts on leaf senescence, a similar approach, based on Asseng et al. (2011), was adopted in all models (Table 1 and Supplementary Methods). A factor for accelerating leaf senescence is calculated as a linear function of air or canopy temperature (daily maximum, average or tri-hourly according to the different model implementations) above a threshold temperature value. Some models included a plateau to the senescence factor.

Table 1
Outline of individual model improvement. More details are given in the Supplementary Data.

Model code	Model name	Reference	Description of model improvements	
			Introduction and/or modification of process representation	Calibration
AE	APSIM-E	(Chen et al., 2010; Keating et al., 2003; Wang et al., 2002; Zhao et al., 2015)	Introduction of a nonlinear temperature response function for phenological development and biomass growth.	Calibration of 14 parameters related to the modified temperature response functions and to radiation use efficiency and maximum specific leaf area.
AW	APSIM-Wheat	(Keating et al., 2003)	Modification of the temperature response function for thermal time accumulation from a triangular to a trapezoidal function. Modification of heat stress effect on leaf senescence to remove discontinuity around the threshold temperature.	Calibration of nine parameters related to the modified temperature response function for thermal time accumulation, canopy senescence, grain number, and grain filling rate.
AN	APSIM-Nwheat	(Asseng et al., 2004, 1998; Keating et al., 2003)	Introduction of an empirical model of canopy temperature as a function of evapotranspiration and daily mean air VPD (described in Webber et al., 2017). Modification of heat stress effect on leaf senescence to remove discontinuity around the threshold temperature.	Calibration of seven parameters related to the new canopy temperature model and the modified leaf senescence heat stress response.
FA	FASSET	(Berntsen et al., 2003; Olesen et al., 2002)	Introduction of a heat stress effect on leaf senescence.	Calibration of seven parameters related to the new leaf senescence response and to LAI, DM allocation to roots, N concentration in storage organs.
GL	GLAM-Wheat	(Challinor et al., 2004; Li et al., 2010)	Introduction of a trapezoidal temperature response function for leaf growth. Modification of the temperature response function for photosynthesis and transpiration efficiency from a bi-linear function with no reduction towards the base temperature to a trapezoidal function. Modification of the temperature response of phenological development from a trapezoidal to a triangular function. Modification of the magnitude of the response of canopy senescence to high temperature. Removed heat stress effect around anthesis on grain set and potential harvest index. Modification of the definition of anthesis (from beginning of flowering to mid-flowering).	Calibration of 26 parameters related to the modified or new temperature response functions and to LAI, HI, maximum potential leaf growth and transpiration, transpiration efficiency, and VPD calculation.
HE	HERMESS	(Kersebaum, 2007; Kersebaum et al., 2011)	Correction of an error in the calculation of thermal time accumulation. Constant grain-to-chaff dry mass ratio at maturity replaced by a function based on the duration of the flowering-to-maturity period. N dilution curves for maximum and critical N concentration were fixed to a constant thermal time from emergence to maturity, now it is scaled to the varietal thermal time from emergence to maturity. Simulation of soil moisture and mineral N starts at the beginning of the year for equilibration based on given weather conditions.	Calibration of thermal time for phenological development and of five parameters related to the correction of thermal time accumulation.
LP	LPJmL	(Beringer et al., 2011; Bondeau et al., 2007; Fader et al., 2010; Gerten et al., 2004; Müller et al., 2007; Rost et al., 2008)	Introduction of a heat stress effect on leaf senescence.	Calibration of five parameters related to phenological development, the sensitivity to photoperiod and LAI.

Table 1 (Continued)

Model code	Model name	Reference	Description of model improvements	
			Introduction and/or modification of process representation	Calibration
NP	Expert-N-SPASS	(Biernath et al., 2011; Priesack et al., 2006; Wang and Engel, 2000)	Introduction of a function to calculate hourly temperature. Modification of the temperature response functions for photosynthesis from a triangular to a trapezoidal function.	Calibration of three parameters related to radiation use efficiency, specific leaf dry mass and grain number.
NS	Expert-N-SUCROSS	(Biernath et al., 2011; Priesack et al., 2006)	Introduction of a function to calculate hourly temperature. Modification of the temperature response functions for photosynthesis from a triangular to a trapezoidal function.	Calibration of three parameters related to radiation use efficiency, specific leaf dry mass and grain number.
OL	OLEARY	(O'Leary and Connor, 1996a; O'Leary and Connor, 1996b; O'Leary et al., 1985)	Modification of the temperature response functions for phenological development and stem development from a linear to a triangular or bi-linear with a maximum function. Introduction of a dry-sowing emergence subroutine. Introduction of an effect of elevation on the psychometric constant and radiation use efficiency.	Modification of the routine simulating transfer of N to grains from generic to cultivar specific.
SA	SALUS	(Basso et al., 2010; Senthilkumar et al., 2009)		Calibration of 35 parameters related to phyllochron, vernalization requirement, sensitivity to photoperiod, LAI, cardinal temperatures of the temperature response function for radiation use efficiency, leaf expansion, root growth, grain filling, grain number, grain N concentration and DM partitioning.
SP	SIMPLACE <LINTUL2-CC-HEAT>	(Angulo et al., 2013)	Introduction of a heat stress effect on leaf senescence. Reduction of yield due to heat stress calculated using T_{mean} instead of T_{max} . Introduction of a sub-routine for post-anthesis biomass re-translocation to grains.	Calibration of four parameters related to radiation use efficiency, LAI, and critical heat stress response.
S2	Sirius2010	(Jamieson and Semenov, 2000; Jamieson et al., 1998; Lawless et al., 2005; Stratonovitch and Semenov, 2015)	Introduction of a heat stress effect on leaf maturation and senescence. Introduction of a heat stress and frost effects on grain number Introduction of a heat stress effect on potential grain dry mass.	Calibration of six parameters related to the new heat stress and frost responses.
SQ	SiriusQuality	(Ferrise et al., 2010; He et al., 2012; Martre et al., 2006)	Introduction of a heat stress effect on leaf maturation and senescence. Modification of the temperature response functions for phenological development and leaf expansion from a linear to a non-linear function.	Calibration of 13 parameters related to heat stress effect on leaf maturation and senescence, the non-linear temperature response function for development and leaf expansion, daylength sensitivity, and vernalization requirement.
WG	WheatGrow	(Cao and Moss, 1997; Cao et al., 2002; Hu et al., 2004; Li et al., 2002; Pan et al., 2007, 2006)	Introduction of a heat stress effect on phenological development. Introduction of function to calculate hourly temperature.	Calibration of four parameters related to the heat stress effect on phenological development and grain filling duration.

In the case of improvements related to heat stress impact on phenological and/or growth processes, the impact of heat stress was modeled by introducing a temperature response function which included a decreasing phase (triangular, trapezoidal, or non-

linear) at high temperatures and which substituted for a linear response function with or without a plateau. Only in one model (OLEARY) a linear response for phenological development was substituted for a linear with a plateau for some phenological stages. In

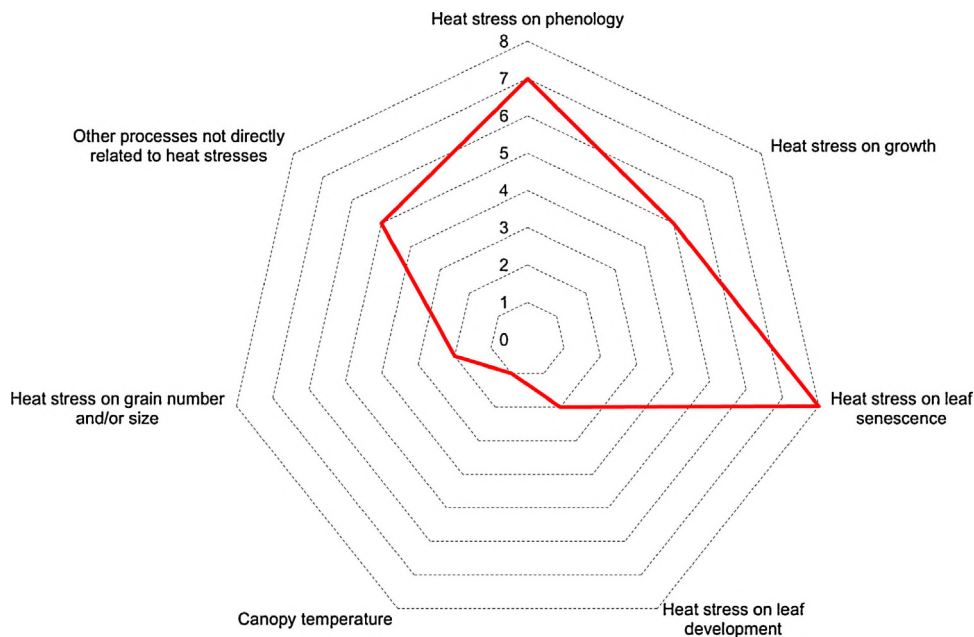


Fig. 1. Number of models that included or modified (if already included) key processes related to heat stress during the model improvement exercise.

the APSIM-wheat model the temperature effect on the phenological development was previously modeled using a function with a single optimum temperature (triangular function) that was now changed to a function with a range of optimum temperatures (trapezoidal function). The crop models that did not introduce such a type of response for phenological development and biomass growth already included this type of response for both processes (APSIM-NWheat, SIMPLACE), or already had a function with a decreasing phase above an optimum temperature for biomass growth and kept a linear temperature response function for phenological development (HERMESS, LPJmL, Sirius2010), or kept a linear approach for both processes (FASSET).

3.2. Effects of model improvement on single models accuracy

Fig. 2 illustrates the effects of model improvement on the simulations of three treatments of the HSC calibration data set whose mean growing seasonal temperatures were different. In most cases, measured in-season and end-of-season LAI, above ground biomass, and grain yield were in the range of model simulations for both the un-improved and the improved models. Nevertheless, the improved models showed a lower level of variation (measured through the 10th to 90th percentile range of the 15 model simulations). For grain yield and above ground biomass, the improved MME was more precise at high temperatures than the unimproved MME (mean growing season temperature of 22 °C and 27 °C in Fig. 2). Most unimproved and improved models underestimated the impact of high temperature on LAI, but this was true to a lower extent for the improved compared to the un-improved models. Considering the e-median of the model ensemble, the simulations of the improved MME appeared similar to the un-improved population at 15 °C but more accurate at 22 °C and 27 °C for LAI and above ground biomass, and for grain yield at 27 °C.

In order to explore if the population of 15 models used in this study had skills similar to that of the 30 models that had previously been used to simulated the calibration data set (Asseng et al., 2015), we compared the RMSRE distribution of these two populations of models for the calibration data set (Fig. 3). The RMSRE distribution for almost all the variables was similar for the 30 models and the 15 unimproved models included in this study. Therefore, we could

reasonably exclude any “model sampling” effects on the results of this work. Comparison of RMSRE distribution of the 15 unimproved and improved models for the calibration data set showed a reduction in the median values for RMSRE of most of the variables: 53% for days from sowing to maturity, 36% for above ground biomass, 31% for grain yield, 18% for HI, 32% for grain number, 12.4% for single grain dry mass. However, RMSRE range for HI, grain number, and single grain dry mass remained almost unchanged.

Fig. 4 shows the effect of model improvement on the accuracy (as measured by RMSRE) of each model for grain yield and for the key variables leading to final yield for the calibration data set. In general, models were improved for almost all measured variables. As expected, models that had large errors for a specific variable were the ones that improved the most for that variable. All models had lower RMSRE for simulating above ground biomass and grain yield after model improvement. The only variables for which more than one model worsened after model improvements were LAI and HI. Five models (APSIM-Nwheat, Expert-N–SPASS, Expert-N–SUCROSS, SALUS, and SIMPLACE<LINTUL2-CC-HEAT>) increased the error for LAI after improvements (Fig. 4).

Two of these models were among the ones that included or modified a sub-routine for heat stress impact on leaf senescence (APSIM-Nwheat and SIMPLACE<LINTUL2-CC-HEAT>). Four models had higher RMSRE of HI after improvement (APSIM-Wheat, GLAM-Wheat, Expert-N – SUCROSS, and SiriusQuality), although they had lower RMSRE for both above ground biomass and grain yield after model improvement. For both the calibration and evaluation data sets, model improvement decreased the variation (measured through the 10th to 90th model ensemble percentile range) of most simulated variables at high mean seasonal temperatures (Fig. 5). For the calibration dataset the reduction of the variability between models and their convergence is an expected result as all the teams used the same dataset to improve and recalibrate their model. For grain yield, an increase in precision was observed for temperature >24 °C for both the calibration and the evaluation data set: grain yield variation decreased by 4% and 21% considering the whole temperature range of the calibration and the evaluation data sets, respectively, and by 39% and 26% considering only mean seasonal temperatures >24 °C. For the evaluation data set, consistent reduction of the range of variation among models was also observed

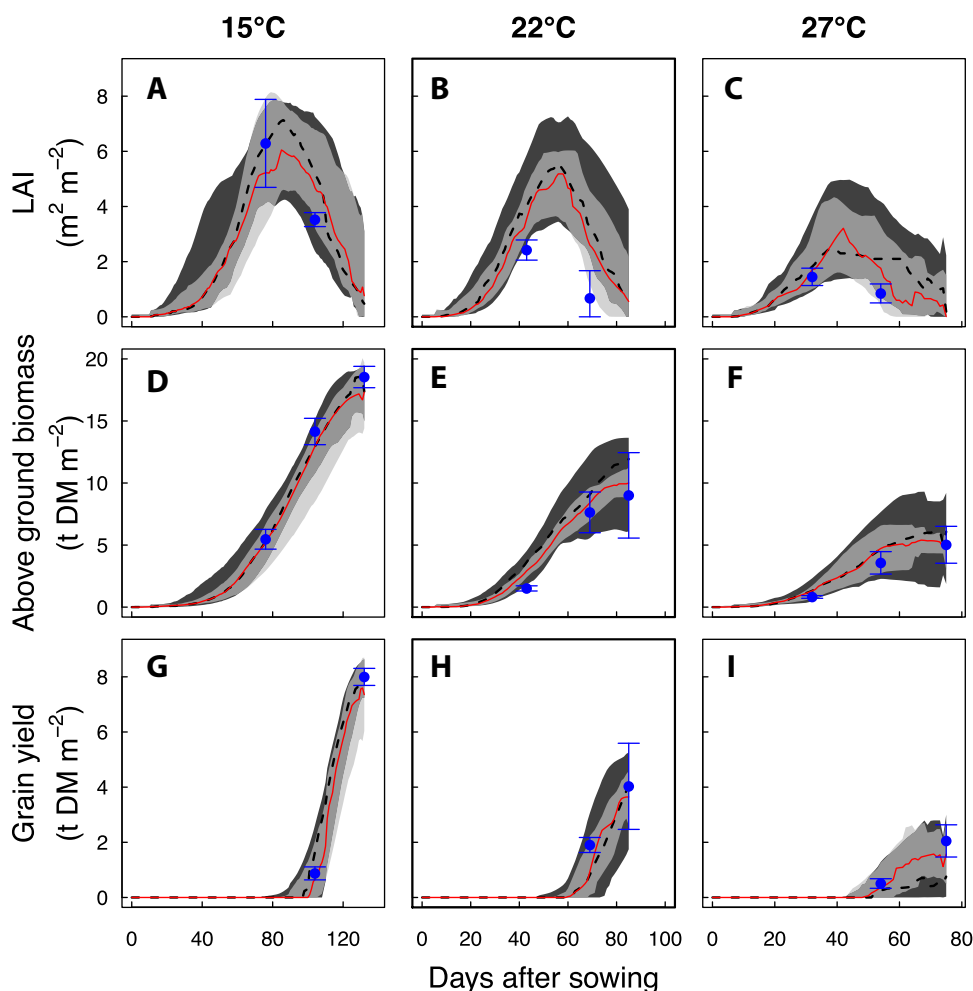


Fig. 2. Simulated and measured wheat growth dynamics for the calibration data set. (A–C) leaf area index (LAI), (D–F) total above ground biomass, and (G–I) grain yield versus days after sowing for mean growing season temperatures 15 °C (A, D, and G), 22 °C (B, E, and H) and 27 °C (C, F, and I). Black dotted lines and dark grey areas are e-median (MME median) and the 10th to 90th percentile range of the 15 original (unimproved) models, respectively. Solid red lines and light grey areas are e-median and the 10th to 90th percentile range of the 15 improved models, respectively. Areas are grey when improved and unimproved ranges overlap. Blue symbols are measured mean \pm 1 s.d. for $n=3$ independent replicates (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for HI (20%), grain number (71%), and single grain dry mass (44%) (Fig. 5).

3.3. Effects of individual model improvement on MME accuracy and prediction skills

For both the calibration and evaluation data sets, model improvements decreased MSE of models for grain yield (Fig. 6, panel A), phenology (Fig. 6, panels B and C), and above ground biomass (Fig. 6, panel D). This reduction was mainly due to a reduction in MME variance. Considering the calibration data set (Fig. 6, panel A), MSE of grain yield decreased on average by 29%, equally due to decrease in squared bias (–33%) and variance (–27%). Considering the evaluation data set, MSE of grain yield was reduced by 37%, due to a 49% reduction in variance, while the squared bias did not increase significantly (Fig. 6, panel A). MSE of above ground biomass was reduced by 44% due to a 54% reduction in variance, while the squared bias did not change significantly (Fig. 6, panel D). Analysis of the prediction skills of the model population showed that the level of prediction error (MSEP) when simulating the “unknown” data set was reduced after improvement by 47% (Fig. 6, panel A). As the MSEP is the sum of the squared bias for the calibration data set and the variance for the evaluation data set

(Eq. (6)), changes in bias and variance of MSEP followed the same reduction patterns.

3.4. Effect of individual model improvement on MME e-median skill

The RMSRE of e-median was reduced by 38% for grain yield and by 46% for above ground biomass, in the calibration data set, and by 2% for grain yield and 11% for above ground biomass in the evaluation data set (Fig. 3). The relationship between the number of models in an ensemble and the CV and RMSRE of e-median estimation of grain yield and above ground biomass was analyzed through a bootstrap approach to create a large number of random ensembles of 1–15 models. Independently of the number of models in the ensembles, for the evaluation data set the CV of e-median was about 41% lower for improved models compared with unimproved models (Fig. 7, panel A and B).

Therefore, model improvement decreased variation of e-median in a range between 15% for $M' = 1$ and 7% for $M' = 15$ for above ground biomass and between 14% at $M' = 1$, and 9% for $M' = 15$ for grain yield. As a consequence, while with the unimproved models the benchmark CV% of 13.5% (Taylor et al., 1999) was not achieved for grain yield even with the maximum model ensemble size, with the improved models this threshold was reached with eight models in

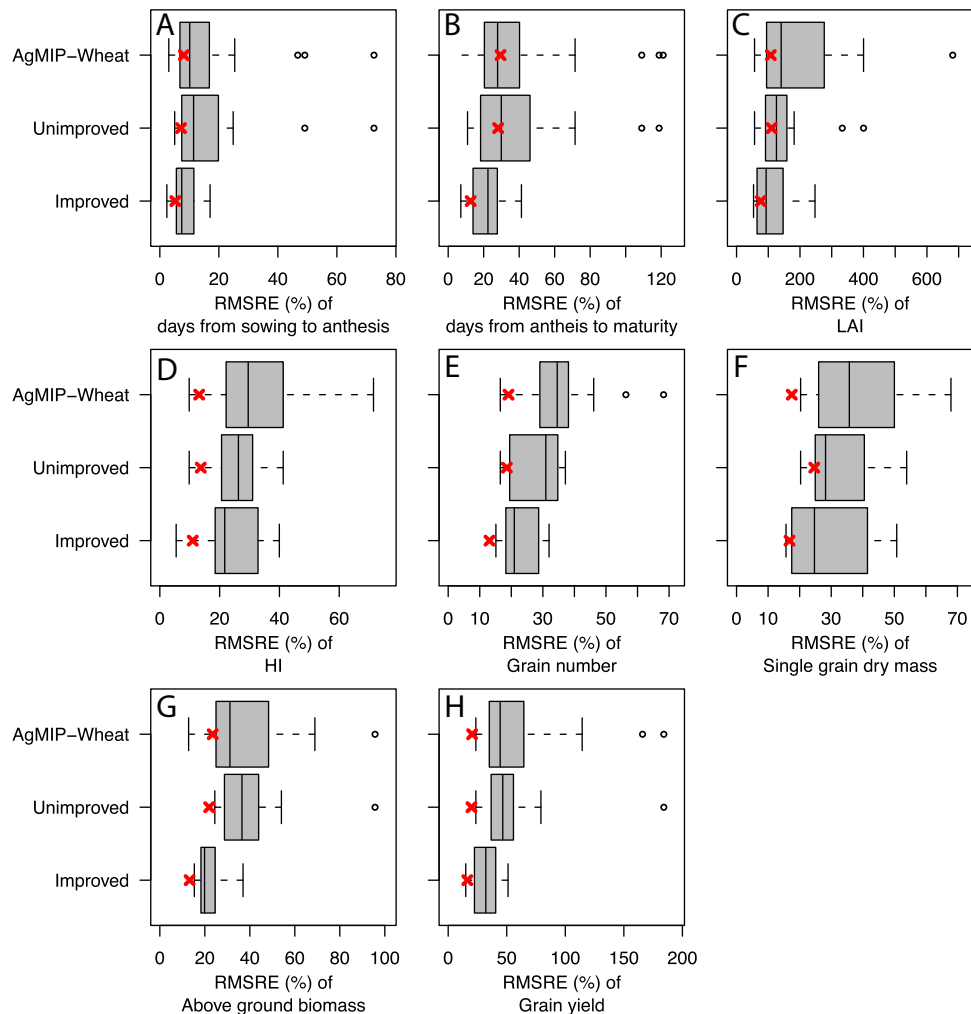


Fig. 3. Effect of model improvement on root mean squared relative error (RMSRE) distribution for days from sowing to anthesis (A), days from anthesis to maturity (B), leaf area index (LAI) (C), harvest index (HI) (D), grain number (E), single grain dry mass (F), final total above ground biomass (G), final grain yield (H), for the calibration data set. RMSRE was calculated for the 30 models included in a previous study (AgMIP-Wheat) (Asseng et al., 2015) and the 15 unimproved and improved models included in the model improvement study. The left and the right side of the box are the first and third RMSRE quartiles. The line inside the box is the RMSRE second quartile or median of individual model errors. The ends of the whiskers indicate the RMSRE 10th and 90th percentile respectively. The empty points are the outliers. The red crosses indicate the e-median RMSRE (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the ensemble. Model improvements reduced e-median RMSRE of grain yield in a range between 12% at $M' = 1$, and 2% at $M' = 15$ for grain yield for the evaluation data set (Fig. 8).

4. Discussion

For the first time, using two unique experimental field data sets with a large range of temperature, we improved the predictive skills of a MME of 15 wheat models. As a result we increased MME accuracy while reducing model ensemble uncertainty. As a consequence, the number of required models for MME impact assessments on yield to achieve observed levels of field experimental variation was halved. This is a significant step forward for crop modelling and future climate impact studies as until now very few models have explicitly considered heat stress impacts on wheat development and growth (Asseng et al., 2011; Moriondo et al., 2010).

4.1. Model improvements

Model improvements increased the accuracy of single models in reproducing heat stress impact on wheat crops. As a consequence,

the accuracy of the models and of the e-median in simulating the impact of high temperatures and heat stress increased and the variance among models in the population was reduced.

As we focused on the effects of model improvements on a MME of 15 models and on the possible consequences for future MME impact assessments studies, we did not analyze each model improvement in detail. In this exercise, the concept of “model improvement” was implemented as an improvement of the applicability of models across diverse environments and climates including climate extremes. Each crop model aimed to improve how high temperature effects were captured by incorporating and/or improving a range of different processes using a high-quality data set. The process descriptions in the models were mostly updated using new information from the literature, e.g. a new approach to heat stress, or they accounted of a harmful effect of high temperatures for the first time. Each team was left free to decide how to implement heat stress in their model. This choice was made considering the diversity of implementation of key physiological processes, and/or the diversity in the level of empiricism/mechanism in their approaches (see supplementary information in Asseng et al., 2015, 2013). In most cases, being primarily developed to simulate “standard” climate condi-

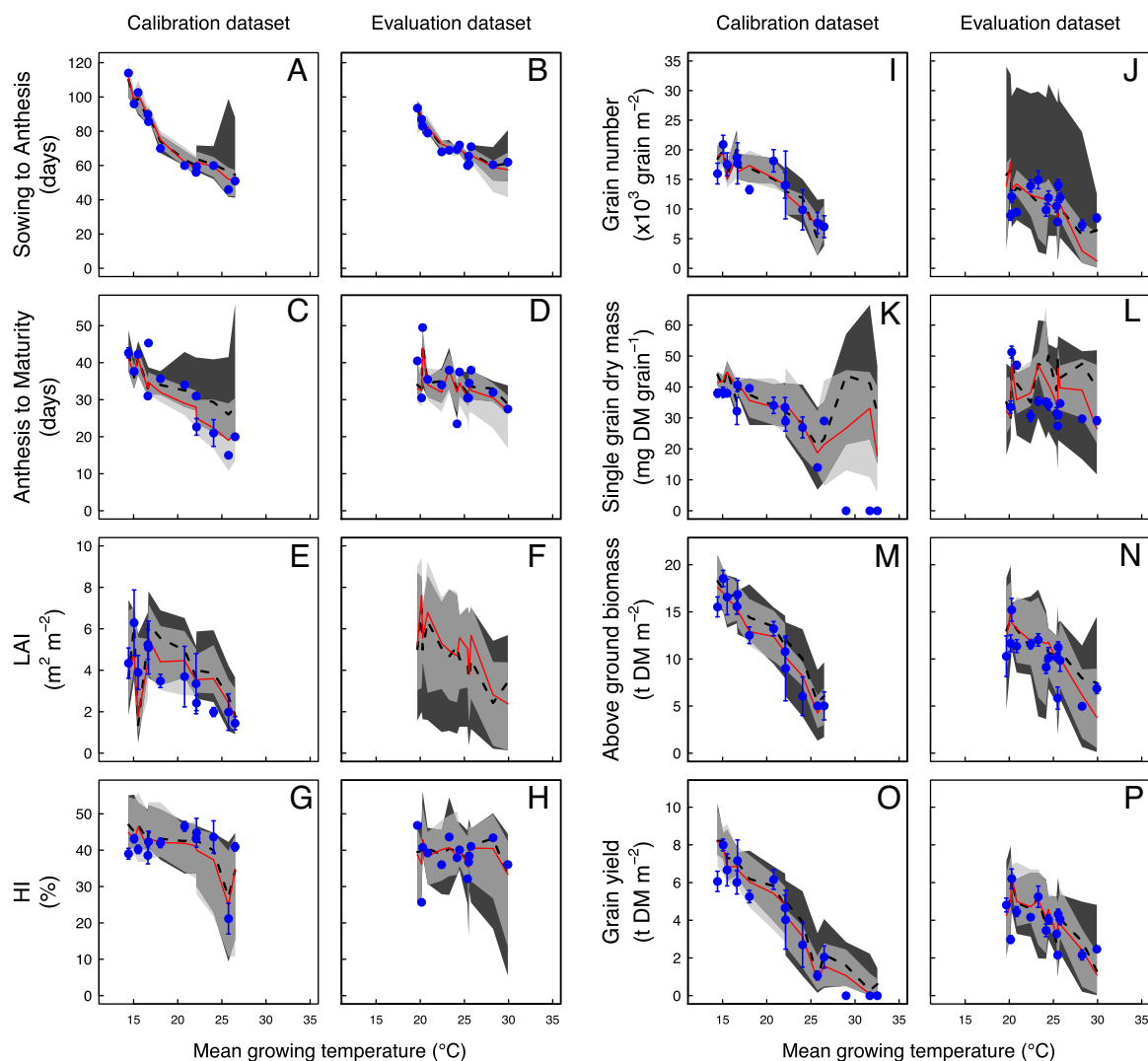


Fig. 5. Simulated and measured days from sowing to anthesis (A and B), days from anthesis to maturity (C and D), leaf area index (LAI) (E and F), harvest index (HI) (G and H), grain number (I and J), single grain dry mass (K and L), final total above ground biomass (M and N), final grain yield (O and P), versus mean growing season temperature for the calibration (A, C, E, G, I, K, M, O) and evaluation (B, D, F, H, J, L, N, P) data sets. Black dotted lines and dark grey areas are e-median (ensemble median) and the 10th to 90th percentile range of the 15 original (unimproved) models, respectively. Solid red lines and light grey areas are e-median and the 10th to 90th percentile range of the 15 improved models, respectively. Symbols are measured mean \pm 1 s.d. for $n=3$ independent replicates. Note that for LAI, there were no observations for the evaluation data set (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

were reported to reduce ear fertility and grain set and consequently grain number (Alghabari et al., 2014; Ferris et al., 1998), and temperatures above 35 °C at the beginning of grain filling were reported to reduce potential final grain size (Hawker and Jenner, 1993; Keeling et al., 1994; Saini et al., 1984, 1983).

Two models considered heat stress impact on leaf development and expansion growth, which was reported to slow down under heat stress (Kemp and Blacklow, 1982). Some models improved the performances by including or modifying canopy temperature routines.

However, modelling of such temperature responses are currently limited by the availability of experimental data sets where these responses can be quantified. Further modeling and experimental work are also needed to reach agreement among models regarding the cardinal temperature of key physiological processes determining wheat development and growth. Furthermore, improved model versions should be further tested through sensitivity analysis in order to better understand the impact of new and revised processes and additional parameters in model structures on simulated variables.

4.2. Model improvement effects on the accuracy and predictive skills of MME

After improvement, the variation range of the MME was reduced at high temperatures in the evaluation data set. The reduction of the variation between the models at high temperatures does not eliminate the value of using MME as model structures remain still different and uncertainty will continue to be part of impact assessment. Grain yield predictive skill (quantified in this study by MSE) of the MME was doubled, and after improvement it was comparable to that of hindcasts, suggesting that the improved model predictions related to the impact of heat stress can be considered reliable and consistent in relation to the observed error.

MME accuracy for grain yield and above ground biomass was also doubled after improvement. The unimproved and the improved MME had similar squared bias, indicating that the main source of variation in the considered MME was due to differences between models. These results suggest that the current level of bias might be an intrinsic property of current simulations or of the considered MME or also possibly linked to other uncertainty factors

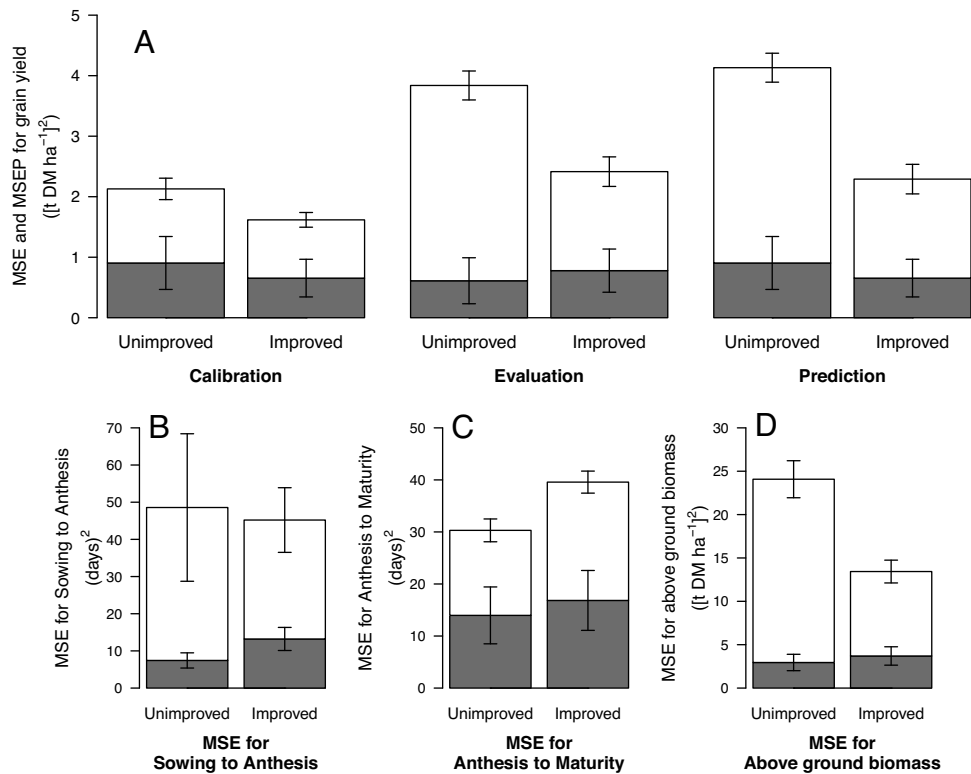


Fig. 6. Mean squared error (MSE) decomposition of grain yield simulated by the 15 unimproved and improved models for the calibration and evaluation (comparison with hindcast) data sets, and the prediction data set (“unknown” data set) (panel A). MSE decomposition for days from sowing to anthesis (panel B), anthesis to maturity (panel C) and final total above ground biomass (panel D) simulated by the 15 unimproved and improved models for the evaluation data set. In panel A, the prediction data set is the same as the evaluation data set but is used as an “unknown” data set to be predicted. MSE was decomposed into squared bias (grey) and variance (white). Data are mean \pm 1 s.e. for 15 (calibration) and 14 (evaluation and prediction) site/year/sowing dates combinations.

that are still not considered explicitly. Due to the similarity of the improved and unimproved MME squared biases, the results related to the analysis of the predictive skills of the MME were similar to the evaluation results. The agreement between the evaluation and the prediction results is an important result and is related to the usefulness of crop models in exploring the consequences on climate change. A fundamental question in crop model impact assessments is the quality assessment of estimates of uncertainty (Wallach et al., 2015). For the first time, the quality of a MME was measured, and it showed that at the current state of crop model development, especially after improvement, prediction uncertainties and hindcast errors are at the same level. Therefore, given a certain level of squared bias measured with hindcast and applied to predictions, we can assume that predictions with these models are reliable. Since in this work the level of prediction uncertainty was measured using the squared bias for a data set that was also used for calibration, we suggest that for future prediction uncertainty assessments done with this MME, the squared bias of the improved models calculated for the evaluation data sets is used as the reference prediction squared bias.

4.3. Model improvement effects on e-median uncertainty

Two fundamental questions in MME uncertainty are what is the uncertainty of the MME predictor and how does the quality of the uncertainty estimates vary with the number of models (Wallach et al., 2015).

As expected, the CV and the RMSRE of e-median decreased with the number of models. On average the unimproved version of MME was not able to reach the benchmark of $CV \leq 13.5\%$ for grain yield (Taylor et al., 1999): even with a random model population of 15

models the average CV was 17%. On the contrary, the improved MME reached $CV \leq 13.5\%$ with 8 models in the ensemble and at this model ensemble size the RMSRE of e-median was reduced by 16%. MME can be a powerful tool for climate impact assessments as they take advantage of the presence of different models in the ensemble (Martre et al., 2015), but they are costly to execute. Execution of MME imply public availability of crop models and/or the interest of modeling groups in participating in coordinated simulation exercises, their availability of funding and/or computational resources to do the requested simulations (Tebaldi and Knutti, 2007). Crop models are developed using different software languages and/or implementations which makes their use by third parties difficult. A model framework that is able to host multiple crop models most probably will overcome these limitations in the future (Bergez et al., 2014; David et al., 2013; Donatelli et al., 2014; Holzworth et al., 2015 Holzworth et al., 2015), but the number of crop models included in these platforms is still limited and, even when available, executing several crop models requires at least some knowledge about the specifics of each model in order to correctly interpret results. Therefore the reduction of the required number of models in an ensemble is a fundamental result key conclusion of this study that makes multi-model impact assessments more realistic practical and less costly to be executed.

Until now the constitution of crop MMEs has been based on the “ensemble of opportunity” approach without an a priori specification that defines the characteristics of a model that should or should not be part of an ensemble (Solazzo and Galmarini, 2015). In most cases, the only requirement for participation has been that there must be a published description of the model. However, one could envisage a more pro-active choice of models. For example, Solazzo and Galmarini (2015) proposed screening models to be

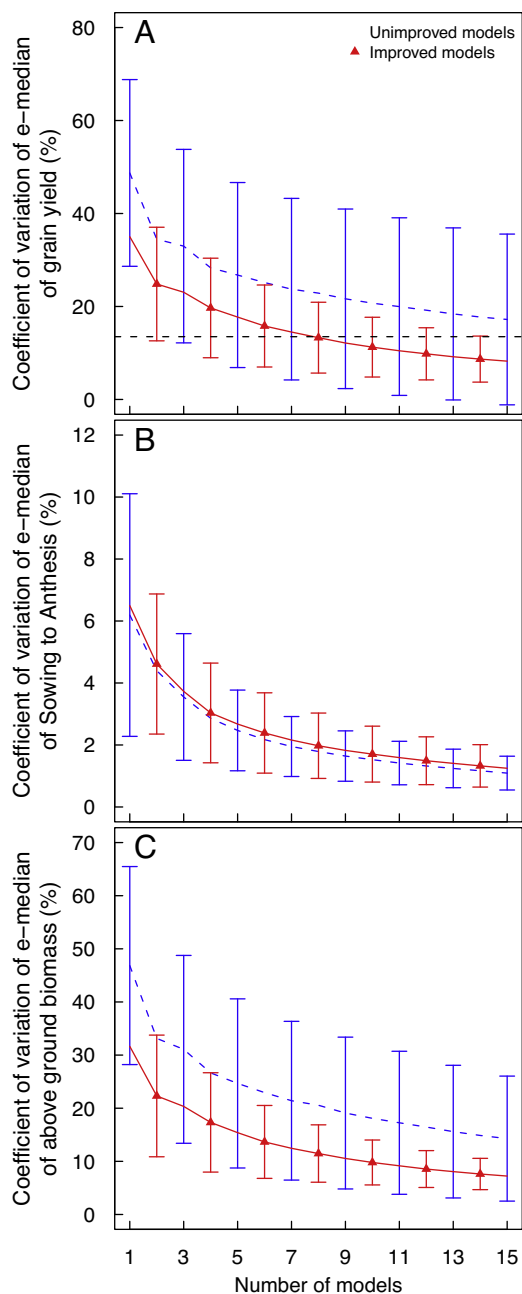


Fig. 7. Coefficient of variation of multi-model ensemble e-median for final grain yield (panel A), days from sowing to maturity (panel B) and final total above ground biomass (Panel C), versus number of models in an ensemble. Values were calculated based on 20,000 bootstrap samples of 1–15 original (unimproved) (blue circles) and improved (red triangles) models for the independent evaluation data set. The horizontal black dashed line in panel A indicates the mean coefficient of variation of GY calculated from a meta-analysis of agronomic field trials (Taylor et al., 1999). For readability, results for unimproved and improved models are shown for odd and even number of models, respectively. Symbols and error bars indicate mean and \pm s.d. of the 20,000 sample e-median values, respectively (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

included in a MME in order to reduce redundancy. They propose doing this in three steps: i) determination to what extent the variability present in the observations is reproduced by the MME, ii) determination of the minimum number of models necessary to represent the observed variability iii) identification of the models to be included in a reduced MME to be used for subsequent analysis. An alternative approach to excluding some models would

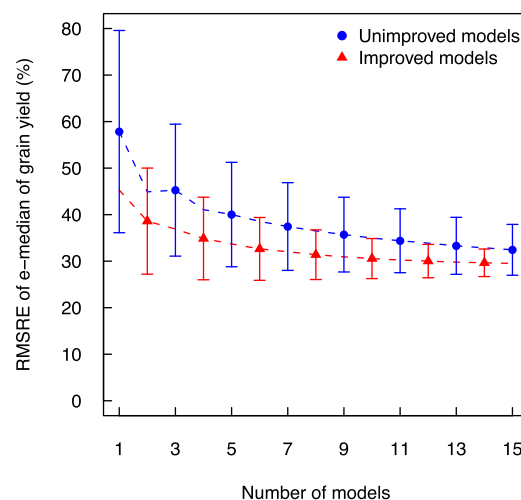


Fig. 8. Root mean squared relative error (RMSRE) of multi-model ensemble e-median for final grain yield (GY) versus number of models in the ensemble for original, unimproved models (blue circles) and improved models (red triangles) for the evaluation field data set. Values are mean \pm 1 s.d. for 20,000 bootstrap samples. For readability, results for unimproved and improved models are shown for odd and even number of models, respectively (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

be to differentially weight the different models in a MME in order to obtain a weighted average prediction. In the climate modeling community weighting methods based on model performance have been reported to improve performance of a MME predictor (Tebaldi and Knutti, 2007). However, weighting based on fit of hindcasts is difficult, because it requires a choice of which output variables to consider and how to combine them in an overall criterion. Another open question is related to the quantification of the global uncertainty in impact assessments. Here we focused our attention on the uncertainty related to model simulations and MME assuming a fixed (non-varietal) parameter set for each model. Furthermore we did not include uncertainty related to weather, soil, and management inputs. In the case of climate change impact assessments the uncertainty related to weather inputs may have a higher importance.

5. Conclusions

Following the example of the climate science community, the crop model community has recently proposed the use of MME as a valid approach to analyze impact assessment uncertainties for current and future climate conditions. However, differently from climate models, the performance of crop models can be evaluated against controlled field experiments from environments that already experience higher than normal growing season temperatures creating conditions that might become common in the future. Using a unique set of experiments for testing the impact of heat stress on wheat crops, we demonstrated that crop model improvements can increase the accuracy of simulations, increase predictive skills of MME's, reduce MME uncertainty, and reduce the number of models needed for reliable impact assessments.

Author contributions

AM, PM, SA, and FE, Conceived and designed research. AM, PM, and DW analyzed the simulation results. AM and PM wrote the manuscript. SA, FW, DW, EW, CM, RPM, ACR, and MAS revised the manuscript. AM, PM, SA, FE, CM, RPR, MAS, EW, BTK, CB, BB, DC, AJC, JD, BD, EER, SG, KCK, AKK, BL, GJO, JEO, EP, PS, TS, PJT, KW, ZZ, and YZ performed simulations, improved individual models and

discussed the results. PDA, BAK, MJO, MPR, AR, GWW, and JWW provided experimental data.

Acknowledgements

AM has received the support of the EU in the framework of the Marie-Curie FP7COFUND People Programme, through the award of an AgreeSkills fellowship under grant agreement no. PCOFUND-GA-2010-267196. PM and DW acknowledge support from the FACCE JPI MACSUR project (031A103B) through the metaprogram Adaptation of Agriculture and Forests to Climate Change (AAFCC) of the French National Institute for Agricultural Research (INRA). SA and DC received financial support from the International Food Policy Research Institute (IFPRI) and the International Maize and Wheat Improvement Center (CIMMYT). FE received support from the FACCE MACSUR project (031A103B) funded through the German Federal Ministry of Education and Research (2812ERA115) and EER was funded through the German Federal Ministry of Economic Cooperation and Development (Project: PARI). EW was funded by the by CSIRO and the Chinese Academy of Sciences through the project 'Advancing crop yield while reducing the use of water and nitrogen'. CM received financial support from the KULUNDA project (01LL0905L) and the MACMIT project (01LN1317A) funded through the German Federal Ministry of Education and Research (BMBF). RPR received financial support from FACCE MACSUR project funded through the Finnish Ministry of Agriculture and Forestry. MPR and PDA received funding from the CGIAR Research Program on Climate Change, Agriculture, and Food Security (CCAFS). CB was funded through the Helmholtz project 'REKLIM-Regional Climate Change: Causes and Effects' Topic 9: 'Climate Change and Air Quality'. KCK and CN were funded by the FACCE MACSUR project through the German Federal Office for Agriculture and Food (BLE). GO'L was funded through the Australian Grains Research and Development Corporation and the Department of Environment and Primary Industries Victoria, Australia. JEO were funded through the FACCE MACSUR project by the Danish Strategic Research Innovation Foundation. ZZ received scholarship from the China Scholarship Council through the CSIRO and Chinese Ministry of Education PhD Research Program. Rothamsted Research is supported via the 20:20 Wheat Programme by the UK Biotechnology and Biological Sciences Research Council.

References

- Alexandratos, N., Bruinsma, J., 2012. *World Agriculture Towards 2030/2050: The 2012 Revision*. FAO, Rome.
- Alghabari, F., Lukac, M., Jones, H.E., Gooding, M.J., 2014. Effect of Rht alleles on the tolerance of wheat grain set to high temperature and drought stress during booting and anthesis. *J. Agron. Crop Sci.* 200, 36–45, <http://dx.doi.org/10.1111/jac.12038>.
- 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. *Agric. For. Meteorol.* 170, 32–46, <http://dx.doi.org/10.1016/j.agrformet.2012.11.017>.
- Asseng, S., Keating, B.A., Fillery, I.R.P., Gregory, P.J., Bowden, J.W., Turner, N.C., Palta, J.A., Abrecht, D.G., 1998. *Performance of the APSIM-wheat model in Western Australia*. *Field Crops Res.* 57, 163–179.
- Asseng, S., Jamieson, P., Kimball, B., Pinter, P., Sayre, K., Bowden, J., Howden, S., 2004. Simulated wheat growth affected by rising temperature, increased water deficit and elevated atmospheric CO₂. *Field Crops Res.* 85, 85–102, [http://dx.doi.org/10.1016/S0378-4290\(03\)00154-0](http://dx.doi.org/10.1016/S0378-4290(03)00154-0).
- Asseng, S., Foster, I., Turner, N.C., 2011. The impact of temperature variability on wheat yields. *Glob. Change Biol.* 17, 997–1012, <http://dx.doi.org/10.1111/j.1365-2486.2010.02262.x>.
- Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thornburn, P.J., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, S., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J.W., Williams, J.W., Wolf, J., 2013. Uncertainty in simulating wheat yields under climate change. *Nat. Clim. Change* 1–6, <http://dx.doi.org/10.1038/nclimate1916>.
- Asseng, S., Ewert, F., Martre, P., Rötter, R.P., Lobell, D.B., Cammarano, D., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., Reynolds, M.P., Alderman, P.D., Prasad, P.V.V., Aggarwal, P.K., Anothai, J., Basso, B., Biernath, C., Challinor, A.J., De Sanctis, G., Doltra, J., Fereres, E., Garcia-Vila, M., Gayler, S., Hoogenboom, G., Hunt, L.A., Izaurralde, R.C., Jabloun, M., Jones, C.D., Kersebaum, K.C., Koehler, A.-K., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Palosuo, T., Priesack, E., Eysih Rezaei, E., Ruane, A.C., Semenov, M.A., Shcherbak, I., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Thornburn, P.J., Waha, K., Wang, E., Wallach, D., Wolf, J., Zhao, Z., Zhu, Y., 2015. Rising temperatures reduce global wheat production. *Nat. Clim. Change* 5, 143–147, <http://dx.doi.org/10.1038/nclimate2470>.
- Basso, B., Cammarano, D., Troccoli, A., Chen, D., Ritchie, J.T., 2010. Long-term wheat response to nitrogen in a rainfed Mediterranean environment: field data and simulation analysis. *Eur. J. Agron.* 33, 132–138, <http://dx.doi.org/10.1016/j.eja.2010.04.004>.
- Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.-H., Kumar, N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? *Glob. Change Biol.* 20, 2301–2320, <http://dx.doi.org/10.1111/gcb.12520>.
- Bergez, J.E., Raynal, H., Launay, M., Beaudoin, N., Casellas, E., Caubel, J., Chabrier, P., Coucheny, E., Dury, J., Garcia de Cortazar-Atauri, I., Justes, E., Mary, B., Ripoche, D., Ruget, F., 2014. Evolution of the STICS crop model to tackle new environmental issues: new formalisms and integration in the modelling and simulation platform RECORD. *Environ. Modell. Softw.* 62, 370–384, <http://dx.doi.org/10.1016/j.envsoft.2014.07.010>.
- Beringer, T., Lucht, W., Schaphoff, S., 2011. Bioenergy production potential of global biomass plantations under environmental and agricultural constraints. *GCB Bioenergy* 3, 299–312, <http://dx.doi.org/10.1111/j.1757-1707.2010.01088.x>.
- Berntsen, J., Petersen, B.M., Jacobsen, B.H., Olesen, J.E., Hutchings, N.J., 2003. Evaluating nitrogen taxation scenarios using the dynamic whole farm simulation model FASSET. *Agric. Syst.* 76, 817–839, [http://dx.doi.org/10.1016/S0308-521X\(02\)00111-7](http://dx.doi.org/10.1016/S0308-521X(02)00111-7).
- Biernath, C., Gayler, S., Bittner, S., Klein, C., Högy, P., Fangmeier, A., Priesack, E., 2011. Evaluating the ability of four crop models to predict differential environmental impacts on spring wheat grown in open-top chambers. *Eur. J. Agron.* 35, 71–82, <http://dx.doi.org/10.1016/j.eja.2011.04.001>.
- Bondeau, A., Smith, P.C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-campen, H., Müller, C., Reichstein, M., Smith, B., 2007. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Glob. Change Biol.* 13, 679–706, <http://dx.doi.org/10.1111/j.1365-2486.2006.01305.x>.
- Burton, I., Lim, B., 2005. Achieving adequate adaptation in agriculture. *Clim. Change* 70, 191–200, <http://dx.doi.org/10.1007/s10584-005-5942-z>.
- Byjesh, K., Kumar, S.N., Aggarwal, P.K., 2010. Simulating impacts, potential adaptation and vulnerability of maize to climate change in India. *Mitig. Adapt. Strateg. Glob. Change* 15, 413–431, <http://dx.doi.org/10.1007/s11027-010-9224-3>.
- Cao, W., Moss, D.N., 1997. Modelling phasic development in wheat: a conceptual integration of physiological components. *J. Agric. Sci.* 129, 163–172, <http://dx.doi.org/10.1017/S0021859697004668>.
- Cao, W., Liu, T., Luo, W., Wang, S., Pan, J., Guo, W., 2002. Simulating organ growth in wheat based on the organ-weight fraction concept. *Plant Prod. Sci.* <http://dx.doi.org/10.1626/pp.5.248>.
- Challinor, A.J., Wheeler, T.R., Craufurd, P.Q., Slingo, J.M., Grimes, D.I.F., 2004. Design and optimisation of a large-area process-based model for annual crops. *Agric. For. Meteorol.* 124, 99–120, <http://dx.doi.org/10.1016/j.agrformet.2004.01.002>.
- Challinor, A., Martre, P., Asseng, S., Thornton, P., Ewert, F., 2014. Making the most of climate impacts ensembles. *Nat. Clim. Change* 4, 77–80, <http://dx.doi.org/10.1038/nclimate2117>.
- Chauhan, S., Srivalli, S., Nautiyal, A.R., Khanna-Chopra, R., 2010. Wheat cultivars differing in heat tolerance show a differential response to monocarpic senescence under high-temperature stress and the involvement of serine proteases. *Photosynthetica* 47, 536–547, <http://dx.doi.org/10.1007/s11099-009-0079-3>.
- Chen, C., Wang, E., Yu, Q., 2010. Modeling wheat and maize productivity as affected by climate variation and irrigation supply in North China Plain. *Agron. J.* 102, 1037, <http://dx.doi.org/10.2134/agronj2009.0505>.
- David, O., Ascough, J.C., Lloyd, W., Green, T.R., Rojas, K.W., Leavesley, G.H., Ahuja, L.R., 2013. A software engineering perspective on environmental modeling framework design: the object modeling system. *Environ. Modell. Softw.* 39, 201–213, <http://dx.doi.org/10.1016/j.envsoft.2012.03.006>.
- Dell, A.I., Pawar, S., Savage, V.M., 2011. Systematic variation in the temperature dependence of physiological and ecological traits. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10591–10596, <http://dx.doi.org/10.1073/pnas.1015178108>.

- Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces. *Clim. Res.* 65, 87–105, <http://dx.doi.org/10.3354/cr01322>.
- Porter, J.R., Gawith, M., 1999. *Temperatures and the growth and development of wheat: a review*. *Eur. J. Agron.* 10, 23–36.
- Priesack, E., Gayler, S., Hartmann, H.P., 2006. The impact of crop growth sub-model choice on simulated water and nitrogen balances. *Nutr. Cycl. Agroecosyst.* 75, 1–13, <http://dx.doi.org/10.1007/s10705-006-9006-1>.
- Qualset, C.O., Vogt, H.E., Borlaug, N.E., 1985. *Registration of yecora rojo wheat*. *Crop Sci.* 25, 1130.
- R Core Team, 2013. R: A Language and Environment for Statistical Computing, 3.0.1 ed. R Foundation for Statistical Computing, Vienna, Austria, <http://dx.doi.org/10.1038/sj.hdy.6800737>.
- Rötter, R.P., Carter, T.R., Olesen, J.E., Porter, J.R., 2011. Crop–climate models need an overhaul. *Nat. Clim. Change* 1, 175–177, <http://dx.doi.org/10.1038/nclimate1152>.
- RStudio Team, 2015. RStudio: Integrated development for R. Inc., Boston, MA. URL: <http://www.rstudio.com/>.
- Reynolds, M., Ageeb, O.A.A., Cesar-Albrecht, J., Costa-Rodrigues, G., Ghanem, E., Hanchinal, R.R., Mann, C., Okuyama, L., Olugbemi, L.B., Ortiz-Ferrara, G., Rajaram, S., Razzaque, M.A., Tandon, J.P., Fischer, R.A., 1994. The International Heat Stress Genotype Experiment: results from 1990 to 1992. Wheat Special Report No. 32. DF, Mexico.
- Reynolds, M., Balota, M., Delgado, M., Amani, I., Fischer, R., 1994b. Physiological and morphological traits associated with spring wheat yield under hot, irrigated conditions. *Aust. J. Plant Physiol.* 21, 717, <http://dx.doi.org/10.1071/PP9940717>.
- Reynolds, M., 1993. Summary of data from the 1st and 2nd International Heat Stress Genotype Experiments. Wheat heat-stressed Environ. Irrig. dry areas rice-wheat farming Syst. Proc. Int. Conf. held Wad Medani, Sudan, 1–4 February, 1993 Dinajpur, Bangladesh 13–15 Febr. 1993.
- Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. *Agric. For. Meteorol.* 170, 166–182, <http://dx.doi.org/10.1016/j.agrformet.2012.09.011>.
- Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid, E., Stehfest, E., Yang, H., Jones, J.W., 2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proc. Natl. Acad. Sci. U. S. A.* 111, 3268–3273, <http://dx.doi.org/10.1073/pnas.1222463110>.
- Rost, S., Gerten, D., Bondeau, A., Lucht, W., Rohwer, J., Schaphoff, S., 2008. Agricultural green and blue water consumption and its influence on the global water system. *Water Resour. Res.* 44, <http://dx.doi.org/10.1029/2007wr006331>.
- Ruane, A.C., Cecil, L.D., Horton, R.M., Gordón, R., McCollum, R., Brown, D., Killough, B., Goldberg, R., Greeley, A.P., Rosenzweig, C., 2013. Climate change impact uncertainties for maize in Panama: farm information, climate projections, and yield sensitivities. *Agric. For. Meteorol.* 170, 132–145, <http://dx.doi.org/10.1016/j.agrformet.2011.10.015>.
- Saini, H.S., Aspinall, D., 1982. *Abnormal sporogenesis in wheat (Triticum aestivum L.) induced by short periods of high temperature*. *Ann. Bot.* 49, 835–846.
- Saini, H., Sedgley, M., Aspinall, D., 1983. Effect of heat stress during floral development on pollen tube growth and ovary anatomy in Wwheat (*Triticum aestivum* L.). *Aust. J. Plant Physiol.* 10, 137, <http://dx.doi.org/10.1071/PP9830137>.
- Saini, H., Sedgley, M., Aspinall, D., 1984. Development anatomy in wheat of male sterility induced by heat stress, water deficit or abscisic acid. *Aust. J. Plant Physiol.* 11, 243, <http://dx.doi.org/10.1071/PP9840243>.
- Schlenker, W., Roberts, M.J., 2009. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proc. Natl. Acad. Sci. U. S. A.* 106, 15594–15598, <http://dx.doi.org/10.1073/pnas.0906865106>.
- Senthilkumar, S., Basso, B., Kravchenko, A.N., Robertson, G.P., 2009. Contemporary evidence of soil carbon loss in the U.S. corn belt. *Soil Sci. Soc. Am. J.* 73, 2078, <http://dx.doi.org/10.2136/sssaj2009.0044>.
- Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., Muricho, G., 2013. Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Secur.* 5, 291–317, <http://dx.doi.org/10.1007/s12571-013-0263-y>.
- Solazzo, E., Galmarini, S., 2015. A science-based use of ensembles of opportunities for assessment and scenario studies. *Atmos. Chem. Phys.* 15, 2535–2544, <http://dx.doi.org/10.5194/acp-15-2535-2015>.
- Stratonovitch, P., Semenov, M.A., 2015. Heat tolerance around flowering in wheat identified as a key trait for increased yield potential in Europe under climate change. *J. Exp. Bot.* 66, 3599–3609, <http://dx.doi.org/10.1093/jxb/erv070>.
- Sundström, J.F., Albiñ, A., Boqvist, S., Ljungvall, K., Marstorp, H., Martiin, C., Nyberg, K., Vågsholm, I., Yuen, J., Magnusson, U., 2014. Future threats to agricultural food production posed by environmental degradation, climate change, and animal and plant diseases—a risk analysis in three economic and climate settings. *Food Secur.* 6, 201–215, <http://dx.doi.org/10.1007/s12571-014-0331-y>.
- Taylor, S.L., Payton, M.E., Raun, W.R., 1999. Relationship between mean yield, coefficient of variation, mean square error, and plot size in wheat field experiments 1. *Commun. Soil Sci. Plant Anal.* 30, 1439–1447, <http://dx.doi.org/10.1080/00103629909370298>.
- Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. A Math. Phys. Eng. Sci.* 365, 2053–2075, <http://dx.doi.org/10.1098/rsta.2007.2076>.
- Wall, G.W., Kimball, B.A., White, J.W., Ottman, M.J., 2011. Gas exchange and water relations of spring wheat under full-season infrared warming. *Glob. Change Biol.* 17, 2113–2133, <http://dx.doi.org/10.1111/j.1365-2486.2011.02399.x>.
- Wallach, D., Mearns, L.O., Rivington, M., Antle, J.M., Ruane, A.C., 2015. Uncertainty in agricultural impact assessment. In: Rosenzweig, C., Hillel, D. (Eds.), *Handbook of Climate Change and Agroecosystems: The Agricultural Model Intercomparison and Improvement Project (AgMIP)*. Imperial College Press, London, United Kingdom, pp. 223–260, <http://dx.doi.org/10.1142/9781783265640.0009>.
- Wallach, D., Thorburn, P.J., Asseng, S., Challinor, A.J., Ewert, F., Jones, J.W., Rötter, R.P., Ruane, A.C., 2016. A framework for evaluating uncertainty in crop model predictions, in: Ewert, F., Boote, K.J., Rötter, R.P., Thorburn, P.J., Nendel, C. (Eds.), *Crop Modelling for Agriculture and Food Security under Global Change—Abstracts of the iCROP2016 Conference*. Berlin, Germany, p. 437.
- Wang, E., Engel, T., 2000. SPASS: a generic process-oriented crop model with versatile windows interfaces. *Environ. Modell. Softw.* 15, 179–188, [http://dx.doi.org/10.1016/S1364-8152\(99\)00033-X](http://dx.doi.org/10.1016/S1364-8152(99)00033-X).
- Wang, E., Robertson, M.J., Hammer, G.L., Carberry, P.S., Holzworth, D., Meinke, H., Chapman, S.C., Hargreaves, J.N.G., Huth, N.I., McLean, G., 2002. Development of a generic crop model template in the cropping system model APSIM. *Eur. J. Agron.* 18, 121–140, [http://dx.doi.org/10.1016/S1161-0301\(02\)00100-4](http://dx.doi.org/10.1016/S1161-0301(02)00100-4).
- Wardlaw, I., Moncur, L., 1995. The response of wheat to high temperature following anthesis. I. The rate and duration of kernel filling. *Aust. J. Plant Physiol.* 22, 391, <http://dx.doi.org/10.1071/PP950391>.
- Wardlaw, I., 2002. Interaction between drought and chronic high temperature during kernel filling in wheat in a controlled environment. *Ann. Bot.* 90, 469–476, <http://dx.doi.org/10.1093/aob/mcf219>.
- Webber, H., Martre, P., Asseng, S., Kimball, B., White, J., Ottman, M., Wall, G.W., De Sanctis, G., Doltra, J., Grant, R., Kassie, B., Maiorano, A., Olesen, J.E., Ripoche, D., Eyshi Rezaei, E., Semenov, M.A., Stratonovitch, P., Ewert, F., 2017. Canopy temperature for simulation of heat stress in irrigated wheat in a semi-arid environment: a multi-model comparison. *Field Crops Res.* 202, 21–35, <http://dx.doi.org/10.1016/j.fcr.2015.10.009>.
- White, J.W., Hoogenboom, G., Kimball, B.A., Wall, G.W., 2011. Methodologies for simulating impacts of climate change on crop production. *Field Crops Res.* 124, 357–368, <http://dx.doi.org/10.1016/j.fcr.2011.07.001>.
- Xu, Q., Paulsen, A.Q., Guikema, J.A., Paulsen, G.M., 1995. Functional and ultrastructural injury to photosynthesis in wheat by high temperature during maturation. *Environ. Exp. Bot.* 35, 43–54, [http://dx.doi.org/10.1016/0098-8472\(94\)00030-9](http://dx.doi.org/10.1016/0098-8472(94)00030-9).
- Zhang, Y., Zhao, Y., Chen, S., Guo, J., Wang, E., 2015. Prediction of maize yield response to climate change with climate and crop model uncertainties. *J. Appl. Meteorol. Climatol.* 54, 785–794, <http://dx.doi.org/10.1175/JAMC-D-14-0147.1>.
- Zhao, Z., Qin, X., Wang, E., Carberry, P., Zhang, Y., Zhou, S., Zhang, X., Hu, C., Wang, Z., 2015. Modelling to increase the eco-efficiency of a wheat–maize double cropping system. *Agric. Ecosyst. Environ.* 210, 36–46, <http://dx.doi.org/10.1016/j.agee.2015.05.005>.