White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/100059/

Version: Accepted Version

Article:
Maiorano, A, Martre, P, Asseng, S et al. (2017) Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. Field Crops Research, 202. pp. 5-20. ISSN: 0378-4290

https://doi.org/10.1016/j.fcr.2016.05.001

© 2016, Elsevier B.V. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International http://creativecommons.org/licenses/by-nc-nd/4.0/

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

*Manuscript
Click here to view linked References

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

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

[a] INRA, UMR759 Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux, 2 Place Viala, F-34 060 Montpellier, France

[b] Montpellier SupAgro, UMR759 Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux, F-34 060 Montpellier, France

[c] Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL-32611, USA

[d] Institute of Crop Science and Resource Conservation, University of Bonn, D-53 115 Bonn, Germany

[e] Potsdam Institute for Climate Impact Research, D-14 473 Potsdam, Germany

[f] Natural Resources Institute Finland (Luke), FI-01301 VantaaHelsinki, Finland

[g] NASA Goddard Institute for Space Studies, New York, NY-10025, USA

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

[i] INRA, UMR 1248 Agrosystèmes et développement territorial, F-31 326 Castanet-Tolosan, France

[j] CSIRO Agriculture, Black Mountain, ACT 2601, Australia

[k] CIMMYT Int. AP 6-641, D.F. Mexico 06600, Mexico

[l] Institute of Soil Ecology, Helmholtz Zentrum München,German Research Center for Environmental Health, D-85 764 Neuherberg, , Germany

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

[n] Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK

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

[p] Cantabrian Agricultural Research and Training Centre, 39600 Muriedas, Spain

[q] WESS-Water & Earth System Science Competence Cluster, c/o University of Tübingen, D-72 074 Tübingen, Germany

[r] Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research, D15 374 Müncheberg, Germany

[s] USDA, Agricultural Research Service, US Arid-Land Agricultural Research Center, Maricopa, AZ-85138, USA

[t] College of Agriculture, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China

[u] Grains Innovation Park, Department of Economic Development Jobs, Transport and Resources, Horsham 3400, Australia

[v] Department of Agroecology, Aarhus University, 8830 Tjele, Denmark

[w] The School of Plant Sciences, University of Arizona, Tucson, AZ-85721, USA

[x] Institute of Soil Science and Land Evaluation, University of Hohenheim, D-70 599 Stuttgart, Germany

[y] CSIRO Agriculture, St Lucia, Queensland, Australia

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

[aa] China Agricultural University, Beijing 100193, China

[†] Present address: James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK

[‡] Present address: CSIRO, St Lucia, Queensland, Australia

[§] Present address: Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK 74078-6028, USA

[*] Corresponding author, e-*mail address*: pierre.martre@supagro.inra.fr (P. Martre).

[**] Starting from P.D.A. the author list is in alphabetical order

Highlights

- 15 wheat crop models were improved for the simulation of the impact of heat stress
- Crop model improvements increased accuracy of simulations
- Improvements reduced multi-model ensemble yield impact uncertainty
- Required number of models for multi-model ensemble impact assessment was reduced

Abstract

To improve climate change impact estimates and to quantify their uncertainty, multi-model ensembles (MMEs) have been suggested. Model improvements can improve the accuracy of simulations and reduce the uncertainty of climate change impact assessments. Furthermore, they can reduce the number of models needed in a MME. Herein, 15 wheat growth models of a larger MME were improved through re-parameterization and/or incorporating or modifying heat stress effects on phenology, leaf growth and senescence, biomass growth, and grain number and size using detailed field experimental data from the USDA Hot Serial Cereal experiment (calibration data set). Simulation results from before and after model improvement were then evaluated with independent field experiments from a CIMMYT world-wide field trial network (evaluation data set). Model improvements decreased the variation (10^{th} to 90^{th} model ensemble percentile range) of grain yields simulated by the MME on average by 39% in the calibration data set and by 26% in the independent evaluation data set for crops grown in mean seasonal temperatures >24°C. MME mean squared error in simulating grain yield decreased by 37%. A reduction in MME uncertainty range by 27% increased MME prediction skills by 47%. Results suggest that the mean level of variation observed in field experiments and used as a benchmark can be reached with half the number of models in the MME. Improving crop models is therefore important to increase the certainty of model-based impact assessments and allow more practical, i.e. smaller MMEs to be used effectively.

Keywords:
Impact uncertainty,
High temperature,
Model improvement,
Multi-model ensemble,
Wheat crop model

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, m^2 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, 1994b). 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:

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

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:

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

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:

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

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), $\mathrm{var}_{\mathrm{M}}$ 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 ($\mathrm{bias}^2_{\mathrm{prediction}}$) and prediction variance ($\mathrm{var}_{\mathrm{prediction}}$). According to Wallach et al. (2016) the average squared error across treatments of MME-mean calculated using the known data set (hindcast) ($\mathrm{MSE}^{\mathrm{hindcast}}_{\mathrm{e\text{-}mean}}$) 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):

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

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

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

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):

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

where, $\mathrm{CV}(\hat{y}_{\text{e-median},M'})$ is the estimate of the coefficient of variation of e-median for the model ensemble of size *M'*, $\mathrm{sd}_B(\hat{y}^{M'}_{\text{e-median},i})$ and $\mathrm{mean}_B(\hat{y}^{M'}_{\text{e-median},i})$ 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:

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

where, $\mathrm{RMSRE}_{M'}$ is the RMSRE of e-median of the model ensemble of M' size, $\hat{y}^{b}_{\text{e-median},i}$ 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_{mean} > 22°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 ($T_{mean} \geq 29$°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).

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., 2015). 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 the modified or new temperature response functions and to LAI, HI, maximum potential leaf growth and transpiration, transpiration efficiency, and VPD calculation.

Table 1.
Continued.

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.
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, 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.

Table 1.
Continued.

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 expansnion, 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 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.

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 nonlinear) 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 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

Figure 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., 20015), 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.

Figure 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 10^{th} to 90^{th} 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 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, and above ground biomass (Fig 6, panel B). 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 increased by 27% (Fig. 6); and 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 B). 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). 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 to 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 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 heart 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 conditions, models had to improve how high temperature effects were captured by including or modifying some key biological processes involved in crop heat stress response. All the models improved their skills in simulating most of the tested variables. However, in several models HI simulation was not improved and in three models (APSIM-Wheat, GLAM-Wheat, Expert-N-SUCROSS) it was slightly worsened, showing that grain yield and above ground biomass did not improve proportionally to each other. As observed by Challinor et al. (2014) this might indicate some level of compensation error during the calibration phase despite the improvement of both yield and biomass. Furthermore, model improvement was focused on heat stress, and most of the improvement was observed for mean growing season temperature > 24°C which is also the range where most of the disagreement was observed before improvement.

Seven models included a sub-routine for simulating the acceleration of leaf senescence above a temperature threshold. Heat stress was reported to enhance leaf senescence with a consequent reduction in the total amount of intercepted light, reduction of the accumulation of assimilates, and shortening of the grain filling period(Chauhan et al., 2010; Wardlaw and Moncur, 1995; Wardlaw, 2002; Xu et al., 1995).

Most biological processes respond exponentially to temperature until an optimum and then they decline (Dell et al., 2011; Parent and Tardieu, 2012). The declining phase of a temperature function has become particularly important when considering climate change impacts (Schlenker and Roberts, 2009). Five

models modified their temperature sub-routines by including this declining phase with increasing temperatures, and 3 models that already included a declining phase used the HSC calibration data set to calibrate the implemented function or to change their shape (e.g. from trapezoidal to non-linear). Regarding cardinal temperatures used for describing the temperature response of phenological development and biomass growth (i.e. the minimum, optimum, and maximum temperatures), there was no clear accordance among models, with the exception of the optimum temperature for radiation use efficiency (~20°C) and the minimum temperature for both phenological development and biomass growth (~0°C) (Wang et al., unpublished). Some models calibrated the optimum and the maximum temperatures using the calibration data set and the best matching values obtained through calibration might have been influenced by the specificities of each model (Eitzinger et al., 2012).

Three models added a sub-routine for accounting for heat stress impact on grain number and/or size. Elevated temperatures before anthesis accelerate development of the spike and decrease grain number (Saini and Aspinall, 1982) and potential final grain size (Ferrise et al., 2010). Temperatures above 31°C around anthesis 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 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), 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 shoud 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 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 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.

Acknowledgements

AM has received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills fellowship under grant agreement n° 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.

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.

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. doi: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. doi:10.1016/j.agrformet.2012.11.017

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., Eyshi Rezaei, E., Ruane, A.C., Semenov, M.A., Shcherbak, I., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Thorburn, P.J., Waha, K., Wang, E., Wallach, D., Wolf, J., Zhao, Z., Zhu, Y., 2015. Rising temperatures reduce global wheat production. Nat. Clim. Chang. 5, 143–147. doi:10.1038/nclimate2470

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. Chang. 1–6. doi:10.1038/ncliamte1916

Asseng, S., Foster, I., Turner, N.C., 2011. The impact of temperature variability on wheat yields. Glob. Chang. Biol. 17, 997–1012. doi:10.1111/j.1365-2486.2010.02262.x

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. F. Crop. Res. 85, 85–102. doi:10.1016/S0378-4290(03)00154-0

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. F. Crop. Res. 57, 163–179.

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. doi: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. Chang. Biol. 20, 2301–20. doi:10.1111/gcb.12520

Bergez, J.E., Raynal, H., Launay, M., Beaudoin, N., Casellas, E., Caubel, J., Chabrier, P., Coucheney, 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. Model. Softw. 62, 370–384. doi: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. doi: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. doi: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 different environmental impacts on spring wheat grown in open-top chambers. Eur. J. Agron. 35, 71–82. doi: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. Chang. Biol. 13, 679–706. doi:10.1111/j.1365-2486.2006.01305.x

Burton, I., Lim, B., 2005. Achieving Adequate Adaptation in Agriculture. Clim. Change 70, 191–200. doi: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. Chang. 15, 413–431. doi:10.1007/s11027-010-9224-3

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. doi:10.1626/pps.5.248

Cao, W., Moss, D.N., 1997. Modelling phasic development in wheat: a conceptual integration of physiological components. J. Agric. Sci. 129, 163–172. doi:10.1017/S0021859697004668

Challinor, A., Martre, P., Asseng, S., Thornton, P., Ewert, F., 2014. Making the most of climate impacts ensembles. Nat. Clim. Chang. 4, 77–80. doi:10.1038/nclimate2117

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. doi:10.1016/j.agrformet.2004.01.002

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. doi: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. doi: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. Model. Softw. 39, 201–213. doi: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. doi:10.1073/pnas.1015178108

Denton, F., Wilbanks, T.J., Abeysinghe, A.C., Burton, I., Gao, Q., Lemos, M.C., Masui, T., O'Brien, K.L., Warner, K., 2014. Climate-resilient pathways: adaptation, mitigation, and sustainable development, in: Field, C.B., Barros, V.R., Dokken, D.J., Mach, K.J., Mastrandrea, M.D., Bilir, T.E., Chatterjee, M., Ebi, K.L., Estrada, Y.O., R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L.W. (Eds.), Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1101–1131.

Donatelli, M., Bregaglio, S., Confalonieri, R., De Mascellis, R., Acutis, M., 2014. A generic framework for evaluating hybrid models by reuse and composition – A case study on soil temperature simulation. Environ. Model. Softw. 62, 478–486. doi:10.1016/j.envsoft.2014.04.011

Donatelli, M., Duveiller, G., Fumagalli, D., Srivastava, A., Zucchini, A., Angileri, V., Fasbender, D., Loudjani, P., Kay, S., Juskevicius, V., Toth, T., Haastrup, P., M'barek, R., Espinosa, M., Ciaian, P., Niemeyer, S., 2012. Assessing agriculture vulnerabilities for the design of effective measures for adaptation to climate change - AVEMAC Project. Publ. Off. Eur. Union 176.

Eitzinger, J., Thaler, S., Schmid, E., Strauss, F., Ferrise, R., Mor, M., Bindi, M., Palosuo, T., Rötter, R., Kersebaum, K.C., Olesen, J.E., Patil, R.H., Şaylan, L., Çaldağ, B., Çaylak, O., 2012. Sensitivities of crop models to extreme weather conditions during flowering period demonstrated for maize and winter wheat in Austria. J. Agric. Sci. 151, 813–835. doi:10.1017/S0021859612000779

Fader, M., Rost, S., Müller, C., Bondeau, A., Gerten, D., 2010. Virtual water content of temperate cereals and maize: Present and potential future patterns. J. Hydrol. 384, 218–231. doi:10.1016/j.jhydrol.2009.12.011

Ferris, R., Ellis, R.H., Wheeler, T.R., Hadley, P., 1998. Effect of high temperature stress at anthesis on grain yield and biomass of field-grown crops of wheat. Ann. Bot. 82, 631–639. doi:10.1006/anbo.1998.0740

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. F. Crop. Res. 117, 245–257. doi:10.1016/j.fcr.2010.03.010

Gerten, D., Gerten, D., Schaphoff, S., Schaphoff, S., Haberlandt, U., Haberlandt, U., Lucht, W., Lucht, W., Sitch, S., Sitch, S., 2004. Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model. J. Hydrol. 286, 249–270. doi:10.1016/j.jhydrol.2003.09.029

Grant, R.F., Kimball, B.A., Conley, M.M., White, J.W., Wall, G.W., Ottman, M.J., 2011. Controlled warming effects on wheat growth and yield: Field measurements and modeling. Agron. J. 103, 1742–1754. doi:10.2134/agronj2011.0158

Hawker, J., Jenner, C., 1993. High temperature affects the activity of enzymes in the committed pathway of starch synthesis in developing wheat endosperm. Aust. J. Plant Physiol. 20, 197. doi:10.1071/PP9930197

He, J., Le Gouis, J., Stratonovitch, P., Allard, V., Gaju, O., Heumez, E., Orford, S., Griffiths, S., Snape, J.W., Foulkes, M.J., Semenov, M.A., Martre, P., 2012. Simulation of environmental and genotypic variations of final leaf number and anthesis date for wheat. Eur. J. Agron. 42, 22–33.

doi:10.1016/j.eja.2011.11.002

Holzworth, D.P., Snow, V., Janssen, S., Athanasiadis, I.N., Donatelli, M., Hoogenboom, G., White, J.W., Thorburn, P., 2015. Agricultural production systems modelling and software: Current status and future prospects. Environ. Model. Softw. doi:10.1016/j.envsoft.2014.12.013

Hu, J.C., Cao, W.X., Zhang, J.B., Jiang, D., Feng, J., 2004. Quantifying responses of winter wheat physiological processes to soil water stress for use in growth simulation modeling. Pedosphere 14, 509–518.

Jamieson, P.D., Semenov, M. a., 2000. Modelling nitrogen uptake and redistribution in wheat. F. Crop. Res. 68, 21–29. doi:10.1016/S0378-4290(00)00103-9

Jamieson, P.D., Semenov, M. a., Brooking, I.R., Francis, G.S., 1998. Sirius: a mechanistic model of wheat response to environmental variation. Eur. J. Agron. 8, 161–179. doi:10.1016/S1161-0301(98)00020-3

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288. doi:Pii S1161-0301(02)00108-9

Keeling, P., Banisadr, R., Barone, L., Wasserman, B., Singletary, G., 1994. Effect of Temperature on Enzymes in the Pathway of Starch Biosynthesis in Developing Wheat and Maize Grain. Aust. J. Plant Physiol. 21, 807. doi:10.1071/PP9940807

Kemp, D.R., Blacklow, W.M., 1982. The responsiveness to temperature of the extension rates of leaves of wheat growing in the field under different levels of nitrogen fertilizer. J. Exp. Bot. 33, 29–36. doi:10.1093/jxb/33.1.29

Kersebaum, K.C., 2007. Modelling nitrogen dynamics in soil-crop systems with HERMES. Nutr. Cycl. Agroecosystems 77, 39–52. doi:10.1007/s10705-006-9044-8

Kersebaum, K.C., Ahuja, L.R., Ma, L., 2011. Special features of the HERMES model and additional procedures for parameterization, calibration, validation, and applications, in: Ahuja, L.R., Ma, L. (Eds.), Methods of Introducing System Models into Agricultural Research. ASA_CSSA_SSSA, pp. 65–94. doi:10.2134/advagricsystmodel2.c2

Kimball, B.A., White, J.W., Ottman, M.J., Wall, G.W., Bernacchi, C.J., Morgan, J., Smith, D.P., 2015. Predicting canopy temperatures and infrared heater energy requirements for warming field plots. Agron. J. 107, 129. doi:10.2134/agronj14.0109

Lawless, C., Semenov, M.A., Jamieson, P.D., 2005. A wheat canopy model linking leaf area and phenology. Eur. J. Agron. 22, 19–32. doi:10.1016/j.eja.2003.11.004

Li, C.D., Cao, W.X., Zhang, Y.C., 2002. Comprehensive pattern of primordium initiation in shoot apex of wheat. Acta Bot. Sin. 44, 273–278.

Li, S.A., Wheeler, T., Challinor, A., Lin, E.D., Xu, Y.L., Ju, H., 2010. Simulating the impacts of global warming on wheat in China using a large area crop model. Acta Meteorol. Sin. 24, 123–135.

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T., Gaydon, D., Marcaida, M., Nakagawa, H., Oriol, P., Ruane, A.C., Ruget, F., Singh, B.-, Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z., Bouman, B., 2015. Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob. Chang. Biol. 21, 1328–41. doi:10.1111/gcb.12758

Lobell, D.B., Field, C.B., Cahill, K.N., Bonfils, C., 2006. Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties. Agric. For. Meteorol. 141, 208–218. doi:10.1016/j.agrformet.2006.10.006

Martre, P., Jamieson, P., Semenov, M.A., Zyskowski, R.F., Porter, J.R., Triboi, E., 2006. Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. Eur. J. Agron. 25, 138–154. doi:http://dx.doi.org/10.1016/j.eja.2006.04.007

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rötter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Glob. Chang. Biol. 21, 911–925. doi:10.1111/gcb.12768

Moradi, R., Koocheki, A., Nassiri Mahallati, M., 2013. Adaptation of maize to climate change impacts in Iran. Mitig. Adapt. Strateg. Glob. Chang. 19, 1223–1238. doi:10.1007/s11027-013-9470-2

Moriondo, M., Giannakopoulos, C., Bindi, M., 2010. Climate change impact assessment: the role of climate extremes in crop yield simulation. Clim. Change 1–23 ST – Climate change impact assessment: the ro. doi:10.1007/s10584-010-9871-0

Müller, C., 2011. Agriculture: Harvesting from uncertainties. Nat. Clim. Chang. 1, 253–254. doi:10.1038/nclimate1179

Müller, C., Eickhout, B., Zaehle, S., Bondeau, A., Cramer, W., Lucht, W., 2007. Effects of changes in CO2, climate, and land use on the carbon balance of the land biosphere during the 21st century. J. Geophys. Res. G Biogeosciences 112, G02032.

O'Leary, G.J., Connor, D.J., 1996a. A simulation model of the wheat crop in response to water and nitrogen supply: I. Model construction. Agric. Syst. 52, 1–29. doi:10.1016/0308-521X(96)00003-0

O'Leary, G.J., Connor, D.J., 1996b. A simulation model of the wheat crop in response to water and nitrogen supply: II. Model validation. Agric. Syst. 52, 31–55. doi:10.1016/0308-521X(96)00002-9

O'Leary, G.J., Connor, D.J., White, D.H., 1985. A simulation model of the development, growth and yield of the wheat crop. Agric. Syst. 17, 1–26. doi:10.1016/0308-521X(85)90019-8

Olesen, J.E., Petersen, B.M., Berntsen, J., Hansen, S., Jamieson, P.D., Thomsen, A.G., 2002. Comparison of methods for simulating effects of nitrogen on green area index and dry matter growth in winter wheat. F. Crop. Res. 74, 131–149. doi:10.1016/S0378-4290(01)00204-0

Ottman, M.J., Kimball, B.A., White, J.W., Wall, G.W., 2012. Wheat growth response to increased temperature from varied planting dates and supplemental infrared heating. Agron. J. doi:10.2134/agronj2011.0212

Palosuo, T., Kersebaum, K.C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J.E., Patil, R.H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F., Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B., Rötter, R., 2011. Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. Eur. J. Agron. 35, 103–114. doi:http://dx.doi.org/10.1016/j.eja.2011.05.001

Pan, J., Zhu, Y., Cao, W., 2007. Modeling plant carbon flow and grain starch accumulation in wheat. F. Crop. Res. 101, 276–284. doi:10.1016/j.fcr.2006.12.005

Pan, J., Zhu, Y., Jiang, D., Dai, T., Li, Y., Cao, W., 2006. Modeling plant nitrogen uptake and grain nitrogen accumulation in wheat. F. Crop. Res. 97, 322–336. doi:10.1016/j.fcr.2005.11.006

Parent, B., Tardieu, F., 2012. Temperature responses of developmental processes have not been affected by breeding in different ecological areas for 17 crop species. New Phytol. 194, 760–74. doi:10.1111/j.1469-8137.2012.04086.x

Pirttioja, N., Carter, T.R., Fronzek, S., Bindi, M., Hoffmann, H., Palosuo, T., Ruiz-Ramos, M., Tao, F., Trnka, M., Acutis, M., Asseng, S., Baranowski, P., Basso, B., Bodin, P., Buis, S., Cammarano, D., Deligios, P., Destain, M.-F., Dumont, B., Ewert, F., Ferrise, R., François, L., Gaiser, T., Hlavinka, P., Jacquemin, I., Kersebaum, K.C., Kollas, C., Krzyszczak, J., Lorite, I.J., Minet, J., Minguez, M.I., Montesino, M., Moriondo, M., Müller, C., Nendel, C., Öztürk, I., Perego, A., Rodríguez, A., Ruane, A.C., Ruget, F., Sanna, M., Semenov, M.A., Slawinski, C., Stratonovitch, P., Supit, I., Waha, K., Wang, E., Wu, L., Zhao, Z., Rötter, R.P., 2015. Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces. Clim. Res. in press. doi: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. Agroecosystems 75, 1–13. doi: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. doi:10.1038/sj.hdy.6800737

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.

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., 1994a. The International Heat Stress Genotype Experiment: results from 1990-1992. 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. doi:10.1071/PP9940717

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, 3268–73. doi:10.1073/pnas.1222463110

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. doi:10.1016/j.agrformet.2012.09.011

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. doi:10.1029/2007WR006331

Rötter, R.P., Carter, T.R., Olesen, J.E., Porter, J.R., 2011. Crop–climate models need an overhaul. Nat.

Clim. Chang. 1, 175–177. doi:10.1038/nclimate1152

RStudio Team, 2015. RStudio: Integrated development for R. Inc., Boston, MA. URL: http://www.rstudio.com/.

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. doi:10.1016/j.agrformet.2011.10.015

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. doi:10.1071/PP9840243

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. doi:10.1071/PP9830137

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.

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, 15594–8. doi: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. doi: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. doi: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. doi: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, 3599–609. doi:10.1093/jxb/erv070

Sundström, J.F., Albihn, 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. doi: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. doi: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, 2053–75. doi: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. Chang. Biol. 17, 2113–2133. doi: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. doi: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 iCROPM2016 Conference. Berlin, Germany, p. 437.

Wang, E., Engel, T., 2000. SPASS: A generic process-oriented crop model with versatile windows interfaces. Environ. Model. Softw. 15, 179–188. doi: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. doi:10.1016/S1161-0301(02)00100-4

Wardlaw, I., 2002. Interaction between drought and chronic high temperature during kernel filling in wheat in a controlled environment. Ann. Bot. 90, 469–476. doi:10.1093/aob/mcf219

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. doi:10.1071/PP9950391

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., 2015. Canopy temperature for simulation of heat stress in irrigated wheat in a semi-arid environment: a multi-model comparison. F. Crop Res. In press. doi: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. F. Crop. Res. 124, 357–368. doi: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. doi: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. doi: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. doi:10.1016/j.agee.2015.05.005

Supplemetary
Click here to view linked References

Supplementary information

Reducing impact uncertainty with model improvement

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 Camarrano, Andrew J. Challinor, Jordi Doltra, Benjamin Dumont, 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 P. Reynolds, Ehsan Eyshi Rezaei Pierre Stratonovitch, Thilo Streck, Peter J. Thorburn, Katharina Waha, Gary W. Wall, Jeffrey W. White, Zhigan Zhao, Yan Zhu

S1. Description of model improvements

S1.1 APSIM-E

In APSIM-E, the temperature response functions for phenological development and biomass growth (RUE) in the original APSIM-Wheat model were modified using a unique nonlinear temperature response function (Wang and Engel 2000). The function has three input parameters with a clear biological meaning, i.e., the minimum (T_{min}), optimum (T_{opt}), and maximum (T_{max}) temperature for the considered process:

$$f(T) = \frac{2(T - T_{\min})^{\alpha}(T_{\text{opt}} - T_{\min})^{\alpha} - (T - T_{\min})^{2\alpha}}{(T_{\text{opt}} - T_{\min})^{2\alpha}} \quad \text{(S1)}$$

with,

$$\alpha = \frac{\ln 2}{\ln\left(\frac{T_{\max} - T_{\min}}{T_{\text{opt}} - T_{\min}}\right)} \quad \text{(S2)}$$

In addition, the radiation use efficiency (RUE) was adjusted based on Meinke et al. (1997) (Table S1). The maximum specific leaf area was also adjusted.

Table S1
Estimated parameter values of the original and improved versions of APSIM-E.

Units	Parameter description	Original value	Improved value
°C	T_{min} for pre-anthesis phenological development	0	0
°C	T_{opt} for pre-anthesis development	25	27.5
°C	T_{max} for pre-anthesis development	35	40
°C	T_{min} for post-anthesis phenological development	0	0
°C	T_{opt} for post-anthesis development	25	27.5
°C	T_{max} for post-anthesis development	35	40
°C	T_{min} for biomass growth	0	0
°C	T_{opt} for biomass growth	22	20
°C	T_{max} for biomass growth	35	35
g MJ^{-1}	Radiation use efficiency	1.24	1.34
m^2 g^{-1}	Maximum specific leaf area	2.7, LAI < 5 m^2 m^{-2} 2.2, 5 m^2 m^{-2} ≤ LAI < 8 m^2 m^{-2} 2.2, LAI < 8 m^2 m^{-2}	3.2, LAI < 5 m^2 m^{-2} 3.0, 5 m^2 m^{-2} ≤ LAI < 8 m^2 m^{-2} 2.2 LAI < 8 m^2 m^{-2}

S1.2 APSIM-Wheat

The temperature response function for thermal time accumulation was modified from a triangular to a trapezoidal response curve and the heat stress effect on leaf senescence model was modified from the one proposed by Asseng et al (2011) to a linear response including a plateau for T_{max} > 43°C without discontinuity at the threshold temperature (34°C) (Stratonovitch and Semenov 2015).

APSIM-wheat (v7.5; http://www.apsim.info/) module was re-parameterized against the experimental data from the HSC calibration data set. Parameters were estimated with the Gauss-Marquardt-Levenberg algorithm using the parameter estimation software PEST (Doherty and Johnston 2003). The weighted sum of squared errors (WSSE) between observations and model predictions was minimized. Seven phenological data types from each of 28 experiments were used for calibration. These were two to three LAI observations, Sow-Ant, Sow-Mat, GY, AGBM, GNumber and GDM, summing up to a total of 226 data points in the objective function. To account for the different orders of magnitude of the different data types, data from each type were assigned with a different weight in the objective function. This was done to get a similar contribution of each data type to the objective function. Weighting factors were one for Sow-Mat and Sow-Ant, two for AGBM, two for GNumber and 10 for LAI, GY and GDM data. Parameters were estimated using a stepwise approach. First the phenological parameters and then the yield component parameters were estimated as suggested by Zhao et al. (2014). Parameters x_temp[3] and x_temp[4] were highly correlated, so both could not be estimated reliably. Therefore, x_temp[4] was fixed at 45 °C and the other parameters were estimated.

Table S2.
Estimated parameter values of the original and improved versions of APSIM-Wheat. Only the parameters that were re-calibrated or introduced with new sub-routines are shown

Parameter name	Units	Parameter description	Original value	Improved value
x_maxt_senescence[1]	°C	Threshold temperature for senescence heat response	-	33.40
x_maxt_senescence[2]	°C	Threshold temperature for the maximum heat response in senescence	-	42.53
y_heatsenescence_fac	-	Leaf senescence factor maximum value (i.e. value for the plateau)	-	0.157
x_temp[2]	°C	T_{opt1} for thermal time accumulation	26	28.50
x_temp[3]	°C	T_{opt2} for thermal time accumulation	26	34.48
x_temp[4]	°C	T_{max} for thermal time accumulation	34	45*
grains_per_gram_stem*	grains g^{-1}	Number of grains per stem dry mass at the beginning of grain filling	24	24.00
potential_grain_filling_rate*	g DM $grain^{-1}$ d^{-1}	Potential daily grain filling rate	0.0019	0.0029
max_grain_size†	g DM $grain^{-1}$	Maximum grain dry mass	0.041	0.042

*Parameters for cultivar Yecora-Rojo, †Set as a fixed value

S1.3 APSIM-Nwheat

The original version of Nwheat considers the effect of heat stress based on a concept that leaf senescence is accelerated three-folds when the daily maximum air temperature exceeds 34°C and six-folds at 40°C (Porter and Gawith 1999). However, this function provided a sudden jump in heat stress factor (SLFT) for a slight increase in temperature above 34°C, which was smoothed-out by changing the threshold temperature to 32°C.

A canopy temperature function was introduced to take into account canopy temperature effect on leaf senescence. Maximum daily canopy temperature (T_{canopy}) was observed to be about 6°C higher than maximum daily air temperature (T_{air}) when the crop is fully stressed and it is cooler than the air temperature on average by 6°C when the crop is non-stressed (Ayeneh et al 2002; Maes and Steppe 2012; Siebert et al 2014). Based on Idso et al. (1981) and Jackson et al. (1981) the difference in T_{canopy} and T_{air} is related to the ratio of actual (ET_a) to potential (ET_0) evapotranspiration and the vapor pressure deficit of the atmosphere (VPD). Thus, an empirical equation relating canopy temperature and air temperature has been included in Nwheat:

$$T_{canopy} = f\mathrm{VPD}\left[-12\left(\frac{\mathrm{ET_a}}{\mathrm{ET_0}} + 6\right)\right] + T_{max} \quad \text{(S3)}$$

with,

$$f\mathrm{VPD}=\begin{cases}0.5\mathrm{VPD}, & \mathrm{VPD} < 1\ \mathrm{kPa}\\ 0.125\mathrm{VPD}+0.375, & 1\ \mathrm{kPa} \le \mathrm{VPD} \le 5\ \mathrm{kPa}\\ 1, & \mathrm{VPD} > 5\ \mathrm{kPa}\end{cases} \quad \text{(S4)}$$

where, fVPD is a normalized factor of vapor pressure effect on Tcanopy .

S1.4 FASSET

A heat stress factor (F_h) accelerating wheat leaf senescence with high temperatures was implemented. The function (Fig S4) is based on the experimental data by Vignjevic et at. (2014) where 15 spring wheat cultivars were investigated and subjected to a post-anthesis (14 days after) high temperature period for five days. The derived function equation implemented in FASSET is:

$$F_h = 1 + b_{hs}\left(T_{max} - T_{hs}\right) \quad \text{(S5)}$$

In the previous version of the model, daily leaf senescence in FASSET was calculated as reported in Olesen et al. (2002). With the implementation of the heat stress function the algorithm is now:

$$\Delta L_g = \frac{\left(T-6\right)}{a_s\left(1-b_s\right)}\left(1-b_s\frac{E_{aT}}{E_{pT}}\right)L_{gx} + F_h \quad \text{(S6)}$$

where, L_{gx} is the maximum modeled green area index, a_s is the duration of senescence equivalent to the period from anthesis to yellow ripeness, and b_s is a factor that increases senescence under drought conditions. Parameters T_{hs} and b_{hs} were estimated by calibration for threshold temperatures up to 35°C (Table S3). Following the modification descripted above other parameters related to LAI, dry matter allocation and nitrogen content in storage organs were re-calibrated (Table S3).

Table S3.
Estimated parameter values of the original and improved versions of FASSET. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
MaxGLAI	$m^2\ m^{-2}$	Maximum crop green leaf area ndex	8	7
LAIDM	$m^2\ g^{-2}$	Maximum ratio between LAI and DM in vegetative top part	0.015	0.011
MaxAlloctoroot	-	Maximum fraction of dry matter production that is allocated to the root	0.3	0.6
MinN_store	-	Minimum content of nitrogen in storage organs	0.018	0.021
MaxN_store	-	Maximum content of nitrogen in storage organs	0.026	0.036
T_{hs}	°C	Threshold temperature for heat response in senescence.	-	30
b_{hs}	$°C^{-1}$	Coefficient increasing senescence due to heat stress	-	0.095

S1.5 GLAM-Wheat

Several temperature response functions were modified:

- The relationship between transpiration efficiency (TE) and temperature was modified from a bi-linear response function with no reduction towards the base temperature (T_{base} = 0°C) to trapezoidal response function.
- The temperature response function for phenological development was modified from a trapezoidal response function to a triangular response function.
- A trapezoidal temperature response functions (based on the mean daily temperature as input) for leaf growth was introduced.

The magnitude of canopy senescence for high temperature was modified using the approach described in Asseng et al. (2011) and the heat stress effect around anthesis on grain set and harvest index was removed as no substantial performance improvement was observed.

The definition of the phenological stage “anthesis” was modified: in the previous version it was reached at the beginning of flowering while in the new version it is reached at mid-flowering.

Various parameter values were modified in GLAM-Wheat. Some of them were introduced due to the modification of the temperature response functions, the others were re-calibrated to better match measurements in the HSC calibration data set (Table 4).

Table S4.
Estimated parameter values of the original and improved versions of GLAM-Wheat. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
TETR1	°C	T_{opt2} for TE	25.0	30.0
TETR2	°C	T_{opt} for TE	30.0	40.0
TETR3	°C	T_{opt} for TE	-	0.0
TETR4	°C	T_{opt1} for TE	-	17.0
TB	°C	T_{min} for phenological development	-	0.0
TO	°C	T_{opt} for phenological development	-	27.5
TM	°C	T_{max} for phenological development	-	45.0
TRLAIB	°C	T_{min} for leaf growth	-	0.0
TRLAIO1	°C	T_{op1} for leaf growth	-	17.0
TRLAIO2	°C	T_{op2} for leaf growth	-	24.0
TRLAIM	°C	T_{max} for leaf growth	-	40.0
CRTIT_LAI_T	m^2 m^{-2}	LAI above which potential transpiration = max value	5	1.2
DHDT	-	Increase in harvest index during grain filling period	0.0175	0.0125
DLDTMX	m^2 m^{-2} d^{-1}	Daily maximum LAI expansion	0.1	0.08
P_TRANS_MAX		Maximum value of potential transpiration	0.8	0.6
TE	-	Transpiration efficiency	5	6.5
TEN_MAX	-	Maximum value of normalized TE	6.8	8
VPD_CTE	-	Empirical parameter for vapour pressure deficit (VPD) calculation (Tanner and Sinclair 1983)	0.7	0.65
SENSTEP	-	Leaf senescence acceleration factor at the threshold temperature	-	2
SENSLOPE	-	Maximum leaf senescence acceleration factor	-	10
GCPLFL	°Cd	Thermal time from planting to flowering	1260	1150
GCFLPF	°Cd	Thermal time from flowering to start of grain filling	184	185
GCPFEN	°Cd	Thermal time duration of grain filling	441	635
TCRITMIN	°C	Temperature around flowering above which potential HI is reduced during flowering,	28.0	-
TLIMMIN	°C	Temperature around flowering above which seed set is null	36.0	-
DLDTMXA	m^2 m^{-2} d^{-1}	Daily decrease in LAI after peak LAI	0.02	DLDTMX

S1.6 HERMES

The previous version of HERMESS included a fixed percentage of grain (80% grain, 20% chaff) calculated on ear dry mass, which was replaced by a flexible function taken from Mirschel et al. (1986) which calculate the percentage depending on the duration from flowering to maturity.

Nitrogen curves for maximum and critical nitrogen concentration were fixed to a constant thermal time from emergence to maturity, now it is scaled to the varietal specific thermal time from emergence to maturity.

In the original version soil moisture and N simulations started just few days before sowing. In the improved version, initial soil moisture and mineral N conditions were determined starting soil moisture and N simulations at a fixed date at the beginning of the year allowing a longer equilibration according to the weather conditions.

103 In the original version of HERMES, the overhanging thermal time at the end of a growth phase was
104 lost. In the improved version it is transferred to the next phase, which required the recalibration of the
105 phenological parameters for both the calibration and the evaluation data sets (Table S5).

Table S5.
Estimated parameter values of the original and improved versions of HERMESS. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
TS1	C°d	Thermal time from sowing to emergence	164	164
Dlbase2	h	Base daylength for development between emergence and double ridge	5	6
TS3	C°d	Thermal time from double ridge to heading	500	498
TS5	C°d	Thermal time from flowering to maturity	440	480
Tbase5	°C	Base temperature from flowering to maturity	6	4

106 *S1.7 LPJmL*

107 Heat stress effect on leaf senescence was introduced. With daily mean air temperatures above 30°C
108 daily mean air temperature is multiplied with a factor (*as*) between 1 and 2:

109
$$as = \begin{cases} 1, & T_{\max} \le 30°\mathrm{C} \\ \dfrac{1}{40-30}\left(T_{\max} - 30\right) + 1, & 30°\mathrm{C} < T_{\max} < 40°\mathrm{C} \\ 2, & T_{\max} \ge 40°\mathrm{C} \end{cases} \quad \text{(S7)}$$

110 which accelerates growth and senescence when applied to the calculation of daily heat units for
111 calculating thermal accumulation (HU_{sum}):

112
$$hu_{\mathrm{sum}} = \begin{cases} (T_{\mathrm{mean}} as, & f\mathrm{HU} < f\mathrm{HU}_{\mathrm{sen}} \\ T_{\mathrm{mean}}, & f\mathrm{HU} \ge f\mathrm{HU}_{\mathrm{sen}} \end{cases} \quad \text{(S8)}$$

113 where $f\mathrm{HU}_{\mathrm{sen}}$ is the fraction of the heat units from sowing to maturity required for the starting of
114 senescence (Table S6).

Table S6.
Estimated parameter values of the original and improved versions of LPJmL. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
HU	°Cd	Heat units from sowing to maturity	2060	2120
psens	-	Sensitivity to the photoperiod effect	0.8	0.6
LAImax	m^2 m^{-2}	Maximum leaf area index	8	5
fHU_{sen}	-	Fraction of growing period at which LAI start decreasing	0.5	0.70
pb	h	Base photoperiod	-	10

115 *S1.8 Expert-N-SPASS and Expert-N-SUCROS*

116 In both ExpertN-Spass and ExpertN-Sucros models, the daily gross rate of canopy photosynthesis is
117 calculated based on temporal integration of the momentary photosynthesis rates over daytime (10 times
118 per day in case of SPSS and 6 in case of SUCROS) as a function of radiation and air temperature.
119 However, air temperatures were assumed to be constant over the day and corresponding to a weighted
120 mean temperature (daily maximum temperatures multiplied by 0.71, and daily minimum temperatures
121 multiplied by 0.29). The improved versions of the models include a routine for the calculation of hourly
122 air temperature based on a sinusoidal function:

123
$$T(t)=\begin{cases} T_{\min}(i)+\left(T_{\max}(i)-T_{\min}(i)\right)\cdot\sin\left(\dfrac{\pi\cdot\left(t-t_{SR}(i)\right)}{2\cdot\left(14-t_{SR}(i)\right)}\right), & t_{SR}(i)<t\le 14 \\ \dfrac{1}{2}\left[\begin{array}{l} T_{\max}(i)+T_{\min}(i+1)+ \\ \left(T_{\max}(i)-T_{\min}(i+1)\right)\cdot\cos\left(\dfrac{\pi\cdot\left(t-14\right)}{t_{SR}(i+1)-10}\right) \end{array}\right], & 14<t<t_{SR}(i+1) \end{cases} \quad \text{(S9)}$$

124 where, t_{SR} is the time of sunrise and i the day number of the year. In the improved model it is assumed
125 that T_{min} is at sunrise and T_{max} at 14:00.

126 The temperature response function of photosynthesis was modified from a triangular to a trapezoidal
127 response function allowing a wider range of temperatures that do not reduce the photosynthetic efficiency.
128 Table S7 shows the parameters that were adjusted in both models. No new parameters were introduced
129 during model improvements; in both models three parameters were re-calibrated.

Table S7.
Estimated parameter values of the original and improved versions of ExpertN-Spass and ExpertN-Sucros. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Cultivar	Original value	Improved value
ExpertN-Sucros					
LUE	g DM MJ^{-1}	Radiation use efficiency	Bacanora	0.69	0.70
			Nesser	0.6	0.68
SpcLW	g DM m^{-2}	Specific leaf dry mass	Bacanora	40.0	41.5
			Nesser	373	415
G1	grains g^{-1} DM	Grains per gram of stem dry mass at anthesis	Yecora Rojo	33	34.5
			Bacanora	24	23.5
			Nesser	28.1	28
ExpertN-Spass					
LUE	g DM MJ^{-1}	Radiation use efficiency	Bacanora	0.695	0.70
			Nesser	0.68	0.69
SpcLW	g m^{-2}	Specific leaf dry mass	Bacanora	425	385
			Nesser	41.9	39.4
G1	grains g^{-1} DM	Grains per gram stem at anthesis	Bacanora	28.8	30
			Nesser	28.5	36

S1.9 OLEARY

In the previous version of the model phenological development was driven by a linear relationship with temperature. In the improved version, the relationship with crop emergence and stem development rates were modified to a triangular response function equation, the relationship with booting and anthesis development rates were modified to a linear approach with cut-off at a maximum rate.

The following modifications were also applied:

- Added effects of elevation on psychrometric constant and radiation use efficiency
- The subroutine simulating N transfer to grain was modified from a generic implementation with a fixed duration (300°Cd) to a cultivar specific duration (parameter TTTDN, Table S8).
- A dry-sowing emergence routine was implemented to delay emergence under very dry conditions. A minimum threshold for soil water content to start emergence is applied (parameter THEM, Table S8).

Table S8 shows the parameters that were modified or introduced. Some of them were introduced due to the new subroutines (see above), the others re-calibrated

Table S8.
Estimated parameter values of the original and improved versions of OLEARY. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	cv Yecora rojo		cv Bacanora		cv Nesser	
			Original value	Improved value	Original value	Improved value	Original value	Improved value
THEM	g cm^{-3}	Minimum threshold for soil water content to start emergence	-	0.3	-	0.3	-	0.3
SLNOPT	g/m2	Optimum specific leaf nitrogen	3	3.6	3	2.1	3	2
OPT1	°C	Optimum temperature for sowing to emergence phase	20	33	20	33	20	33
OPT4	°C	Optimum temperature for sowing to anthesis phase	20	33	20	33	20	33
EMMDD	°Cd	Thermal time for emergence	180	110	180	110	180	110
STMDD	°Cd	Thermal time for stem extension	400	400	400	400	400	400
BOOTDL	°Cdh	Photothermal time for booting	3300	3300	3300	3500	3300	3500
ANTHDL	°Cdh	Photothermal time for anthesis	13800	13800	13800	18150	13800	16950
GRMAX	mg d^{-1}	Maximum grain growth rate	2.5	2.8	2.5	1.9	2.5	1.8
GXM	mg	Maximum grain size	55	55	55	50	55	50
PRES	g g^{-1}	Maximum proportion of biomass at anthesis to grain	0.2	0.15	0.2	0.15	0.2	0.15

S1.10 SALUS

No subroutine was modified or introduced in the improved version of SALUS. Model improvement in SALUS consisted in an extensive model re-calibration to obtain better performances, including harmonization of the cardinal temperatures of the temperature response of RUE, adjustment of the photoperiod - phyllochron relationship, and optimization of biomass allocation coefficients driving the source/sink ratio. Table S9 shows the parameters that were parameterized or calibrated.

Table S9.
Estimated parameter values of the original and improved versions of SALUS. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
Phyll	°Cd leaf^{-1}	Phyllochron	80	120
KrPGr	grain^{-1} d^{-1}	Daily rate of grain fill at T_{opt}	0.008	0.0019
KrNPt	grain^{-1} ear^{-1}	Maximum potential grain number per ear	800	24
Vcoef	-	Vernalization coefficient for winter cereals	20	0
PhLow	h	Photoperiod lower limit	8	6
EmgInt	leaf eq.	Intercept of the emergence leaf equivalents calculation	0.3	0.01
EmgSlp	leaf eq. cm^{-1}	Slope of the emergence leaf equivalents calculation	0.1	0.01
LEtg	leaf eq.	Leaf equivalents to germinate	0.5	0.8
LEJuv	leaf eq.	Leaf equivalents to end of juvenile stage	4	0
LEsec	leaf eq.	Leaf equivalent when first leaf starts senescing	3.5	2
LEear	leaf eq.	Leaf equivalents for ear growth	4.1	1.4
Legg	leaf eq.	Leaf equivalents for grain growth	5.5	4.5
PhotoC	-	Photoperiod vs. phyllochron relationship constant	0.007	0.012
RUE	g DM MJ^{-1}	Radiation use efficiency	2.9	2.5
SLWmax	g cm^{-2}	Maximum specific leaf dry mass	0.005	0.0065
Lncsf		factor for daily rate of leaf senescence	0.45	0.6
ToptP	°C	T_{opt} for photosynthesis	15	19
MxNVg	g N g^{-1}	Maximum concentration of N in vegetative parts	0.04	0.035
MxNKr	g N g^{-1}	Maximum concentration of N in grain	0.02	0.03
StemF-EG 1.0	g DM g^{-1}	Stem allocation factor at end (1.0) of the ear growth (EG) phase	1	0.5
GRF-EG 1.0	g DM g^{-1}	Grain allocation factor at end (1.0) of EG phase	0	0.5
StemF-GG 0.0	g DM g^{-1}	Stem allocation factor at begin (0.0) of the grain growth (GG) phase	0	0.5
GRF-GG 0.0	g DM g^{-1}	Grain allocation factor at begin (0.0) of GG phase	1	0.5
RTF-EG 0.0	g DM g^{-1}	Root fraction of tops sink at EG 1.0 phase	0.45	0.20
RTF-EG 0.5	g DM g^{-1}	Root fraction of tops sink at EG 0.5 phase	0.45	0.15
RTF-EG 1.0	g DM g^{-1}	Root fraction of tops sink at EG 1.0 phase	0.45	0.10
RTF-GG 0.0	g DM g^{-1}	Root fraction of tops sink at GG 0.0 phase	0.09	0.05
RTF-GG 0.5	g DM g^{-1}	Root fraction of tops sink at GG 0.5 phase	0.09	0.01
RTF-GG 1.0	g DM g^{-1}	Root fraction of tops sink at GG 1.0 phase	0.09	0.01
RES-EG 0.0	g DM g^{-1}	Reserve fraction of tops sink at EG 1.0 phase	0.45	0.10
RES-EG 0.5	g DM g^{-1}	Reserve fraction of tops sink at EG 0.5 phase	0.45	0.10
RES-EG 1.0	g DM g^{-1}	Reserve fraction of tops sink at EG 1.0 phase	0.45	0.05
RES-GG 0.0	g DM g^{-1}	Reserve fraction of tops sink at GG 0.0 phase	0.45	0.05
RES-GG 0.5	g DM g^{-1}	Reserve fraction of tops sink at GG 0.5 phase	0.45	0.01
RES-GG 1.0	g DM g^{-1}	Reserve fraction of tops sink at GG 1.0 phase	0.25	0.01

S1.11 SIMPLACE<LINTUL2-CC-HEAT>

The acceleration of leaf senescence model was introduced in the new version of the model as described by Asseng et al. (2011). The previous version of the model already included a routine for the simulation of heat stress on grain yield based on daily maximum temperature. In the improved version of the model the average diurnal temperature is used (Teixeira et al. 2013).

A function of post-anthesis biomass re-translocation to grains was introduced based on Jamieson et al. (1998) where 20% of accumulated biomass at anthesis is translocate to grains after anthesis. The rate of daily translocation is a function of total dry matter at anthesis, the fraction of dry matter available for re-translocation, and thermal time after anthesis.

Table S10 shows the parameters that were modified or introduced in the SIMPLACE<LINTUL2-CC-HEAT> model.

Table S10.
Estimated parameter values of the original and improved versions of SIMPLACE<LINTUL2-CC-HEAT>. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	New value
LUE	g MJ^{-1} m^{-2}	Radiation use efficiency	3	2.2
RGRL	-	Relative growth rate of LAI during exponential growth	0.009	0.03
LAII	m^2 m^{-2}	Initial LAI	0.017	0.022
HSTCritical	°C	Critical temperature threshold (heat stress component)	27	31

S1.12 Sirius2010

Improvements to Sirius2010 were described in Stratonovitch and Semenov (2015). In Sirius2010, the duration of leaf senescence is expressed in thermal time and linked to the rank of the leaf in the canopy, i.e. later emerged leaves have a longer period of senescence. Daily thermal time (ΔT) is calculated from 3-hourly canopy temperatures estimated as described in Jamieson et al. (1995) To account for shortening of the leaf mature and senescence phase caused by high temperature, the 3-hourlly temperatures T_i are multiplied by an accelerated leaf senescence factor R_i^{L} (dimensionless):

$$\Delta T = \sum_{i=1}^{8}\left(\frac{\max\left(0,\left(R_i^{\mathrm{L}} \times T_i - T_{base}\right)\right)}{8}\right) \quad \text{(S10)}$$

where, T_b is the base temperature (set at 0°C). R_i^{L} increases linearly form 1 when T_i exceeds T^{L}:

$$R_i^{\mathrm{L}} = 1 + \max\left(0, T_i - T^{\mathrm{L}}\right) S^{\mathrm{L}} \quad \text{(S11)}$$

where, S^{L} is the slope of the senescence acceleration per unit of canopy temperature above T^{L}. As in the original version, grain filling is stopped prematurely if the canopy has fully senesced.

The adverse effects of heat on grain number and size have been incorporated into Sirius2010 by modifying the calculation of the potential yield determinants: grain number and potential grain dry mass. In absence of heat stress, the sink capacity of the grains (Y_{pot}) is set to be the product of the potential number of grains by the potential dry mass of an individual grain (W_{pot} = 0.065 g $grain^{-1}$):

$$Y_{\text{pot}} = \text{DM}_{\text{ear}} N_{\text{pot}} W_{\text{pot}} \quad \text{(S12)}$$

where, DM_{ear} is the dry mass accumulated in ears prior to anthesis, and N_{pot} = 122.4 (grains g^{-1}) the maximum number of grain per unit of ear dry mass. To account for the effect of high temperature on meiosis and fertilization, the number of grain set per unit of ear dry mass is reduced when the daily maximum canopy temperature $T_{\text{max}}^{\text{C}}$ during a period from 10 days before to anthesis exceeds a threshold temperature T^{N} (Table S11). In this case, the rate of grain number per unit of ear dry mass decreases linearly from 1 when $T_{\text{max}}^{\text{C}}$ exceeds, T^{N}:

$$R_{\text{H}}^{\text{N}} = \max\left(0, \min\left(1, 1 - \left(T_{\text{max}}^{\text{C}} - T^{\text{N}}\right) S^{\text{N}}\right)\right) \quad \text{(S13)}$$

where, R_{H}^{N} is the rate of fertile grain number per unit of ear dry mass limited by heat stress and S^{N} is the slope of the grain number reduction per unit of $T_{\text{max}}^{\text{C}}$ above T^{N} (Table S11). The rate of grain number set per unit of ear dry mass is reduced if the minimum canopy temperature $T_{\text{min}}^{\text{C}}$ during a period from -3 to +3 days around anthesis decrease from a threshold temperature of 0°C to -1°C:

$$R_{\text{F}}^{\text{N}} = \max\left(0, \min\left(1, T_{\text{min}}^{\text{C}} + 1\right)\right) \quad \text{(S14)}$$

where, R_{F}^{N} is the rate of fertile grain number per unit of ear dry mass limited by frost. The actual number *N* of grain per unit of ear dry mass is the product of the potential number of grain by the heat and frost reduction rates:

$$N = N_{\text{pot}} R_{\text{H}}^{\text{N}} R_{\text{F}}^{\text{N}} \quad \text{(S15)}$$

where After the reduction of grain numbers at flowering, the potential dry mass of single grains is limited in the advent of heat stress during endosperm development. The potential dry mass of each grain is reduced if the maximum canopy temperature T_{max}^{S} occurring at the beginning of grain filling, i.e. a period of from 5 to 12 days after anthesis, exceeds a threshold temperature T^{W} (Table S11). The maximum dry mass of a grain is reduced linearly from W_{pot} when $T_{\text{max}}^{\text{S}}$ exceeds T^{W}:

$$W = W_{\text{pot}} \max\left(0, \min\left(1, 1 - \left(T_{\text{max}}^{\text{S}} - T^{\text{W}}\right) S^{\text{W}}\right)\right) \quad \text{(S16)}$$

where, W is the actual potential dry mass of a single grain limited by heat stress and S^{W} the slope of the potential dry mass reduction per unit of canopy temperature above T^{W}. Grain filling stops prematurely if the actual grain sink capacity $Y_{\lim} = \mathrm{DM}_{ear} \times N \times W$ has been filled.

Table S11.
Estimated parameter values of the original and improved versions of Sirius2010. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
T^{L}	°C	Temperature threshold for senescence acceleration	-	28.93
S^{L}	°C^{-1}	Slope of the senescence acceleration factor	-	0.108
T^{N}	°C	Temperature threshold for grain number reduction	-	27
S^{N}	°C^{-1}	Slope of grain number reduction	-	0.125
T^{W}	°C	Temperature threshold for maximum grain dry mass reduction	-	30
S^{W}	°C^{-1}	Slope of maximum grain dry mass reduction	-	0.004

S1.13 SiriusQuality

A nonlinear temperature response function (Yan and Hunt 1999) for phenological development and leaf expansion was introduced:

$$f(T) = \left(\left(\frac{T_{\max} - T}{T_{\max} - T_{opt}} \right) \left(\frac{T - T_{\min}}{T_{opt} - T_{\min}} \right)^{\frac{T_{opt} - T_{\min}}{T_{\max} - T_{opt}}} \right)^{\beta} \quad \text{(S17)}$$

where T_{min}, T_{opt}, and $T_{\max}$ are the cardinal temperatures, and β is a shape parameter.

Both phenological development and leaf expansion were parameterized with the same parameter values (Parent and Tardieu 2012) (Table S12). A heat stress response of the duration of the mature leaf phase and leaf senescence was introduced as described for Sirius2010.

Table S12.
Estimated parameter values of the original and improved versions of *SiriusQuality*. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
T^{L}	°C	Temperature threshold for senescence acceleration	-	35
S^{L}	°C^{-1}	Slope of the senescence acceleration	-	0.45
T_{minPL}	°C	T_{min} for phenological development and leaf expansion	-	0
T_{optPL}	°C	T_{opt} for phenological development and leaf expansion	-	32
T_{maxPL}	°C	T_{max} for phenological development and leaf expansion	-	55
$Shape_{PL}$	°C	Shape parameter of the nonlinear function	-	2.1
AreaPL	cm^2 lamina^{-1}	Maximum potential surface area of the penultimate leaf lamina	40	36
NLL	leaf	Number of leaves produced after floral initiation	6.5	6
P	°Cd	Phyllochron	120	115
SLDL	leaf h^{-1}	Daylength response of leaf production	0.62	0.47
VAI	[d°C]$^{-1}$	Response of vernalization rate to temperature between T_{min} and T_{opt} for vernalisation	0.00135	0.002
P_{decr}	-	Factor decreasing the phyllochron for leaf number less than 3	0.75	1
P_{incr}	-	Factor increasing the phyllochron for leaf number higher than 8	1.25	1

S1.14 WheatGrow

A subroutine for simulating phenological development under heat stress was introduced. In the improved version of the model, the daily thermal effect (TE) on phenological development is composed by (1) the daily thermal effect under normal temperature range (NTE – as in the previous version of the model); and (2) the high temperature effect for accelerating plants senescence (HTE – added in the improved version):

$$\mathrm{TE} = \mathrm{NTE} + \mathrm{HTE} \tag{S18}$$

with,

$$\mathrm{HTE}_i = \frac{\sum_{i=1}^{N} \mathrm{HDD}_i}{\mathrm{HTS} \times \mathrm{GDD}_\mathrm{R}} \tag{S19}$$

where, HDD is the accumulated thermal time above a threshold temperature (T_{hs}), i is the number of days after emergence, $\mathrm{GDD_R}$ is the thermal time required for the vegetative and reproductive growth stages to occur, set at 480°Cd and 520°Cd, respectively, and HTS is the high temperature sensitivity parameter. HTS is a genotypic parameter that indicate the heat tolerance of wheat cultivars. HDD is calculated as the accumulation of hourly temperature above T_h (Liu et al 2014):

$$\mathrm{HDD}_i = \frac{1}{24}\sum_{t=1}^{24} \max\left(0, T_{i,\mathrm{h}} - T_{\mathrm{hs}}\right) \quad \text{(S20)}$$

The hourly temperature is derived from the minimum and the maximum daily temperature using the cosine function described in Matthews and Hunt (1994).

Table S13 shows the parameters that were modified or introduced in the WheatGrow model following improvement. Some of them were introduced due to the new subroutines (see above), the others re-calibrated.

Table S13.
Estimated parameter values of the original and improved versions of WheatGrow. Only parameters that were re-parameterized, re-calibrated, or introduced with new subroutines are shown

Parameter name	Units	Parameter description	Original value	Improved value
T_h	°C	High temperature threshold (value for spring wheat)	-	34
IE*	-	Intrinsic earliness	0.91	0.86
HTS*	-	High temperature sensitivity	-	0.09
FDF*	-	Grain filling duration factor	0.95	0.85

* Parameters value for cultivar Yecora-Rojo

S2. Calculation of seasonal mean temperature

Seasonal mean air temperature was calculated from daily air temperature (Tt), which was derived from the sum of eight contributions of a cosine variation between maximum and minimum daily air temperatures (Weir et al 1984).

$$T_t = \frac{1}{8}\sum_{r=1}^{r=8} (T_h - T_b)$$

with

$$T_h(r) = T_{\min} + f_r\left(T_{\max} - T_{\min}\right)$$

and

$$f_r = \frac{1}{2}\left(1 + \cos\frac{90}{8}(2r-1)\right)$$

where r is an index for a particular 3-h period, T_b (°C) is the base temperature (0°C) and T_h (°C) is the calculated three hour temperature contribution to estimated daily mean temperature. Negative contributions of T_h were treated as zero.

References

Asseng S, Foster I, Turner NC (2011) The impact of temperature variability on wheat yields. Glob Chang Biol 17:997–1012. doi: 10.1111/j.1365-2486.2010.02262.x

Ayeneh A, van Ginkel M, Reynolds M., Ammar K (2002) Comparison of leaf, spike, peduncle and canopy temperature depression in wheat under heat stress. F Crop Res 79:173–184. doi: 10.1016/S0378-4290(02)00138-7

Doherty J, Johnston JM (2003) Methodologies for calibration and predictive analysis of a watershed model. J Am Water Resour Assoc 39:251–265. doi: 10.1111/j.1752-1688.2003.tb04381.x

Idso SB, Jackson RD, Pinter PJ, et al (1981) Normalizing the stress-degree-day parameter for environmental variability. Agric Meteorol 24:45–55. doi: 10.1016/0002-1571(81)90032-7

Jackson RD, Idso SB, Reginato RJ, Pinter PJ (1981) Canopy temperature as a crop water stress indicator. Water Resour Res 17:1133–1138. doi: 10.1029/WR017i004p01133

Jamieson PD, Francis GS, Wilson DR, Martin RJ (1995) Effects of water deficits on evapotranspiration from barley. Agric For Meteorol 76:41–58. doi: 10.1016/0168-1923(94)02214-5

Jamieson PD, Porter JR, Goudriaan J, et al (1998) A comparison of the models AFRCWHEAT2, CERES-Wheat, Sirius, SUCROS2 and SWHEAT with measurements from wheat grown under drought. F Crop Res 55:23–44. doi: 10.1016/S0378-4290(97)00060-9

Liu B, Liu L, Tian L, et al (2014) Post-heading heat stress and yield impact in winter wheat of China. Glob Chang Biol 20:372–81. doi: 10.1111/gcb.12442

Maes WH, Steppe K (2012) Estimating evapotranspiration and drought stress with ground-based thermal remote sensing in agriculture: a review. J Exp Bot 63:4671–712. doi: 10.1093/jxb/ers165

Matthews RB, Hunt LA (1994) GUMCAS: a model describing the growth of cassava (Manihot esculenta L. Crantz). F Crop Res 36:69–84. doi: 10.1016/0378-4290(94)90054-X

Meinke H, Hammer GL, van Keulen H, et al (1997) Perspectives for Agronomy - Adopting Ecological Principles and Managing Resource Use, Proceedings of the 4th Congress of the European Society for Agronomy. Dev Crop Sci. doi: 10.1016/S0378-519X(97)80012-8

Mirschel W, Wurbs A, Kretschmer H (1986) Zur Dynamik der Kornfüllung und Ermittlung der Kornbiomasse aus der Ährenbiomasse während der Kornfüllungsperiode für Winterweizen. Arch für Acker- und Pflanzenbau und Bodenkd 30:347–353.

Olesen JE, Petersen BM, Berntsen J, et al (2002) Comparison of methods for simulating effects of nitrogen on green area index and dry matter growth in winter wheat. F Crop Res 74:131–149. doi: 10.1016/S0378-4290(01)00204-0

Parent B, Tardieu F (2012) Temperature responses of developmental processes have not been affected by breeding in different ecological areas for 17 crop species. New Phytol 194:760–74. doi: 10.1111/j.1469-8137.2012.04086.x

283 Porter JR, Gawith M (1999) Temperatures and the growth and development of wheat: a review. Eur J
284 Agron 10:23–36.

285 Siebert S, Ewert F, Eyshi Rezaei E, et al (2014) Impact of heat stress on crop yield—on the importance of
286 considering canopy temperature. Environ Res Lett 9:044012. doi: 10.1088/1748-9326/9/4/044012

287 Stratonovitch P, Semenov MA (2015) Heat tolerance around flowering in wheat identified as a key trait
288 for increased yield potential in Europe under climate change. J Exp Bot 66:3599–609. doi:
289 10.1093/jxb/erv070

290 Tanner CB, Sinclair TR (1983) Efficient water use in crop production: research or re-search? Limitations
291 to Effic. Water Use Crop Prod. pp 1–27

292 Teixeira EI, Fischer G, van Velthuizen H, et al (2013) Global hot-spots of heat stress on agricultural crops
293 due to climate change. Agric For Meteorol 170:206–215. doi: 10.1016/j.agrformet.2011.09.002

294 Vignjevic M, Wang X, Olesen JE, Wollenweber B (2014) Traits in Spring Wheat Cultivars Associated
295 with Yield Loss Caused by a Heat Stress Episode after Anthesis. J Agron Crop Sci. doi:
296 10.1111/jac.12085

297 Wang E, Engel T (2000) SPASS: A generic process-oriented crop model with versatile windows
298 interfaces. Environ Model Softw 15:179–188. doi: 10.1016/S1364-8152(99)00033-X

299 Weir A, Bragg P, Porter JR, Rayner JH (1984) A winter wheat crop simulation model without water or
300 nutrient limitations. J Agric Sci 102:371–382.

301 Yan W, Hunt LA (1999) An Equation for Modelling the Temperature Response of Plants using only the
302 Cardinal Temperatures. Ann Bot 84:607–614. doi: 10.1006/anbo.1999.0955

303 Zhao G, Bryan B a., Song X (2014) Sensitivity and uncertainty analysis of the APSIM-wheat model:
304 Interactions between cultivar, environmental, and management parameters. Ecol Modell 279:1–11.
305 doi: 10.1016/j.ecolmodel.2014.02.003

306

FIGURE CAPTIONS

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

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 10^{th} to 90^{th} percentile range of the 15 original (unimproved) models, respectively. Solid red lines and light grey areas are e-median and the 10^{th} to 90^{th} percentile range of the 15 improved models, respectively. Areas are grey when improved and unimproved ranges overlap. Blue symbols are measured mean ± 1 s.d. for n = 3 independent replicates. (The figure is available in color in the online version of the article).

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.

Fig. 4. Log_2 difference of RMSRE for improved and unimproved models versus RMSRE of unimproved models 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. A positive difference of the log2RMSRE's indicate an improvement in model performance. The extent of model improvement in terms of RMSRE doubles for each unit of log2 RMSRE difference between the un-improved and the improved population of models.

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 10^{th} to 90^{th} percentile range of the 15 improved models, respectively. Symbols are

measured mean ± 1 s.d. for n = 3 independent replicates. Note that for LAI, there were no observations for the evaluation data set.

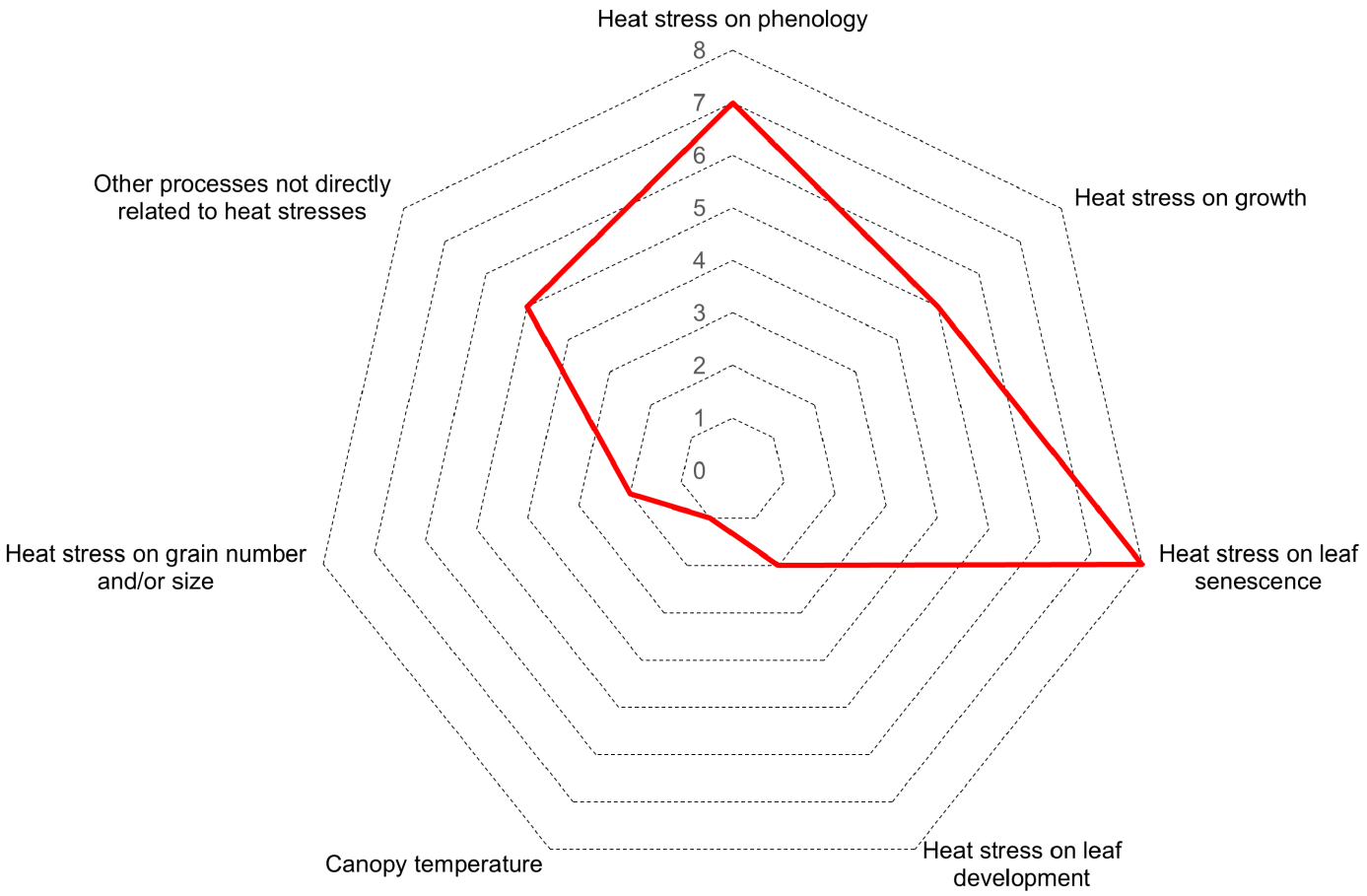
Fig. 6. Mean squared error (MSE) decomposition of grain yield simulated by the 15 unimproved and improved models for the calibration and evaluation (comparison with hindcast) data sets, and the prediction data set ("unknown" data set) (panel A). MSE decomposition for days from sowing to anthesis (panel 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 ± 1 s.e. for 15 (calibration) and 14 (evaluation and prediction) site/year/sowing dates combinations.

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 to 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 ±s.d. of the 20,000 sample e-median values, respectively.

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 ± 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.

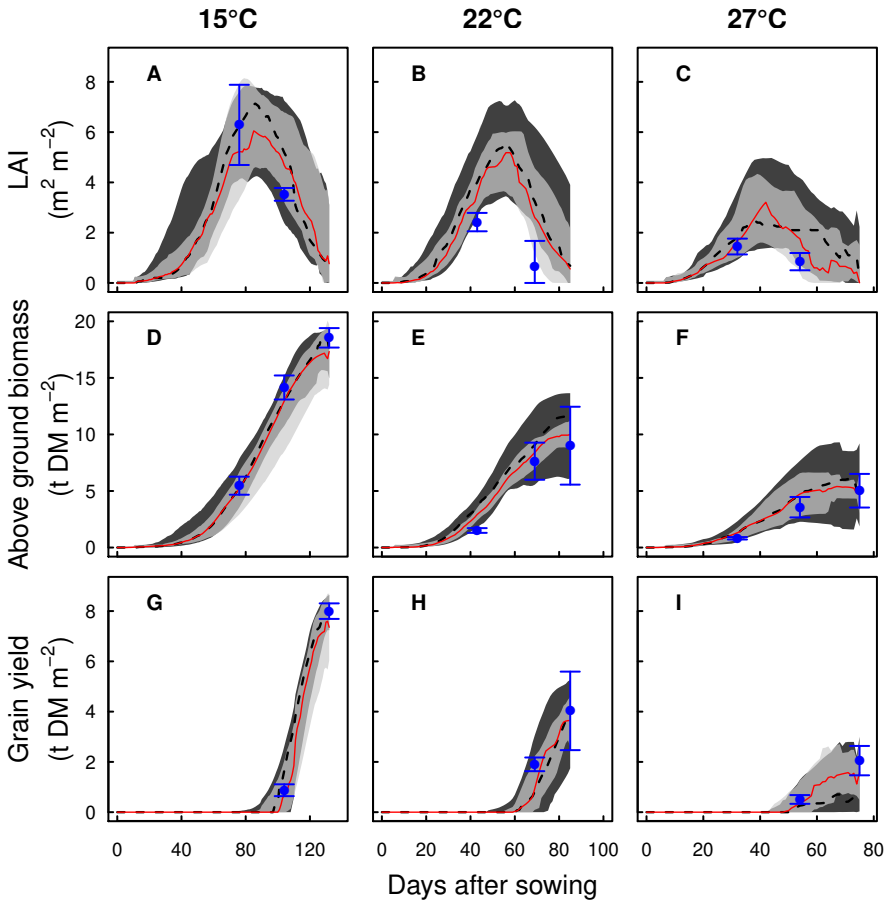
Figure

FIGURE 1



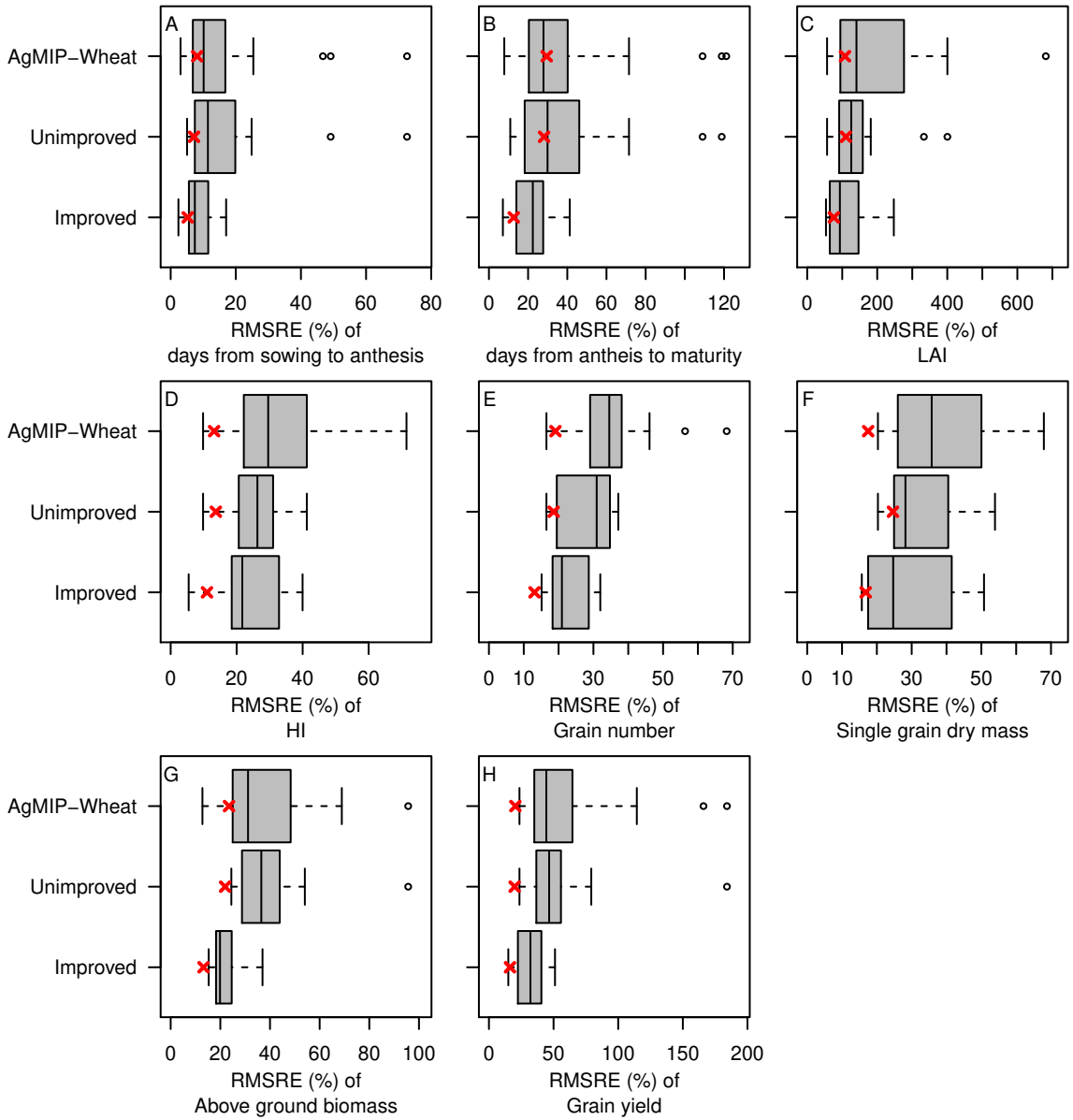
Figure

Figure 2



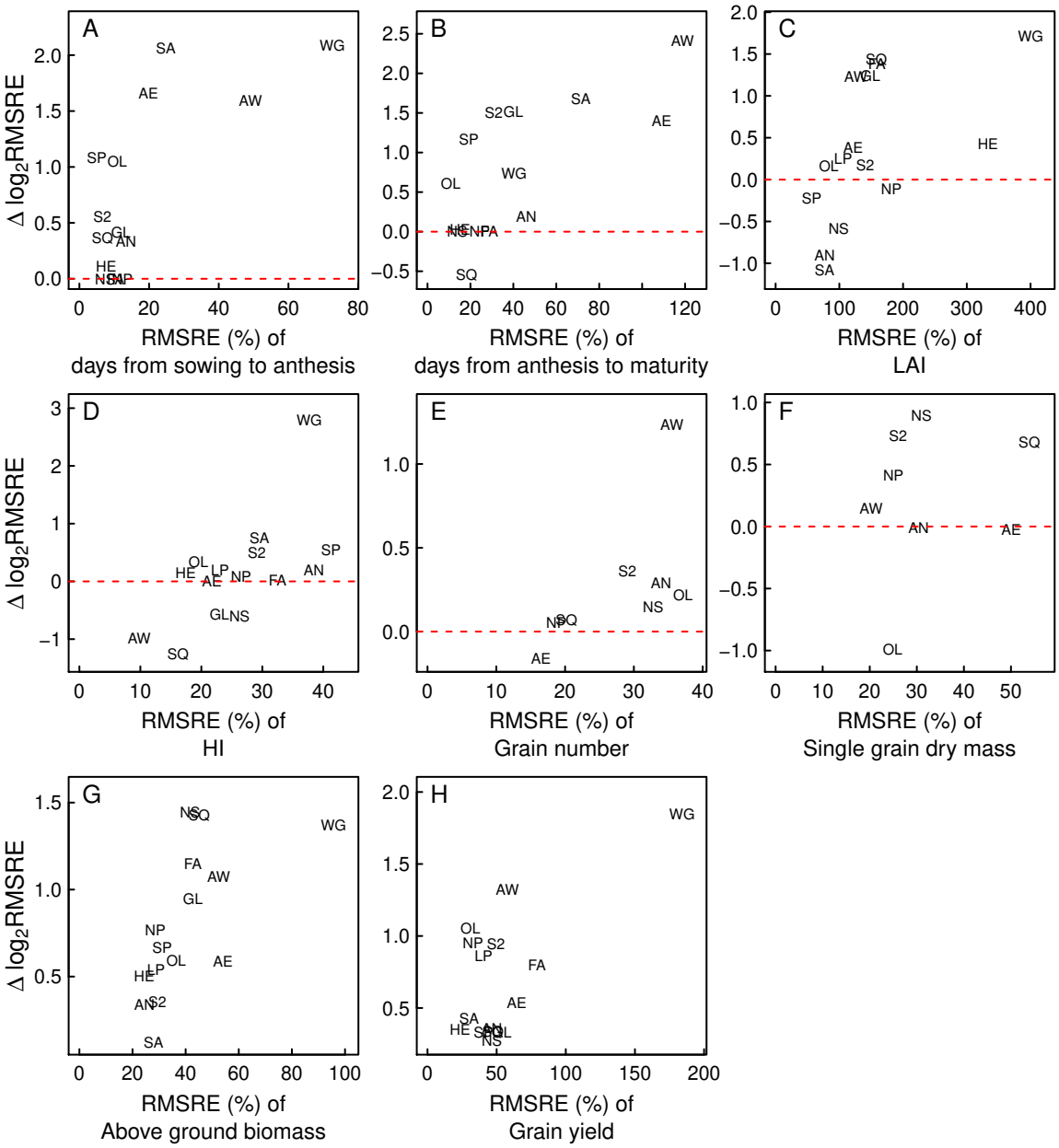
Figure

Figure 03



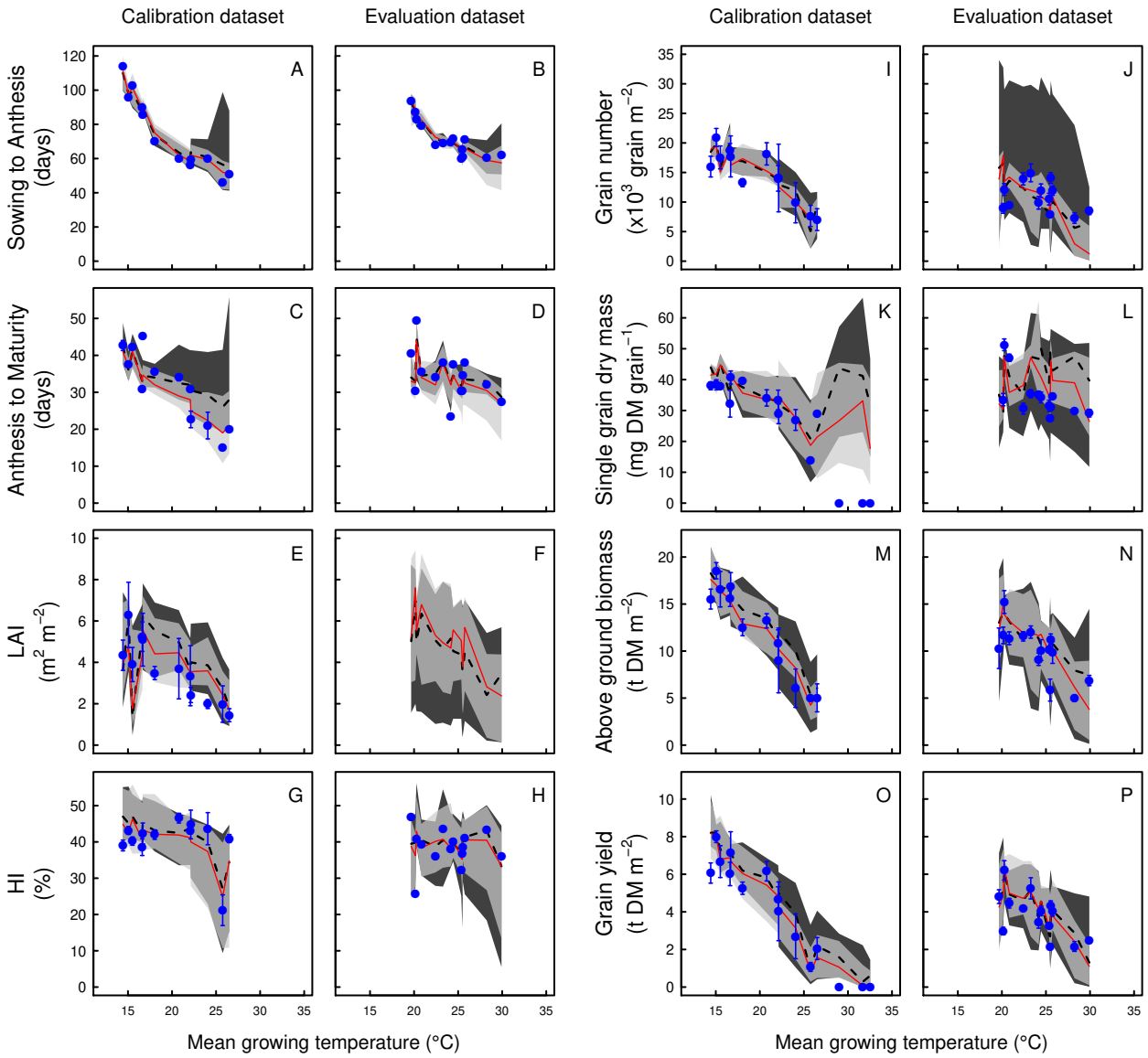
Figure

Figure 4



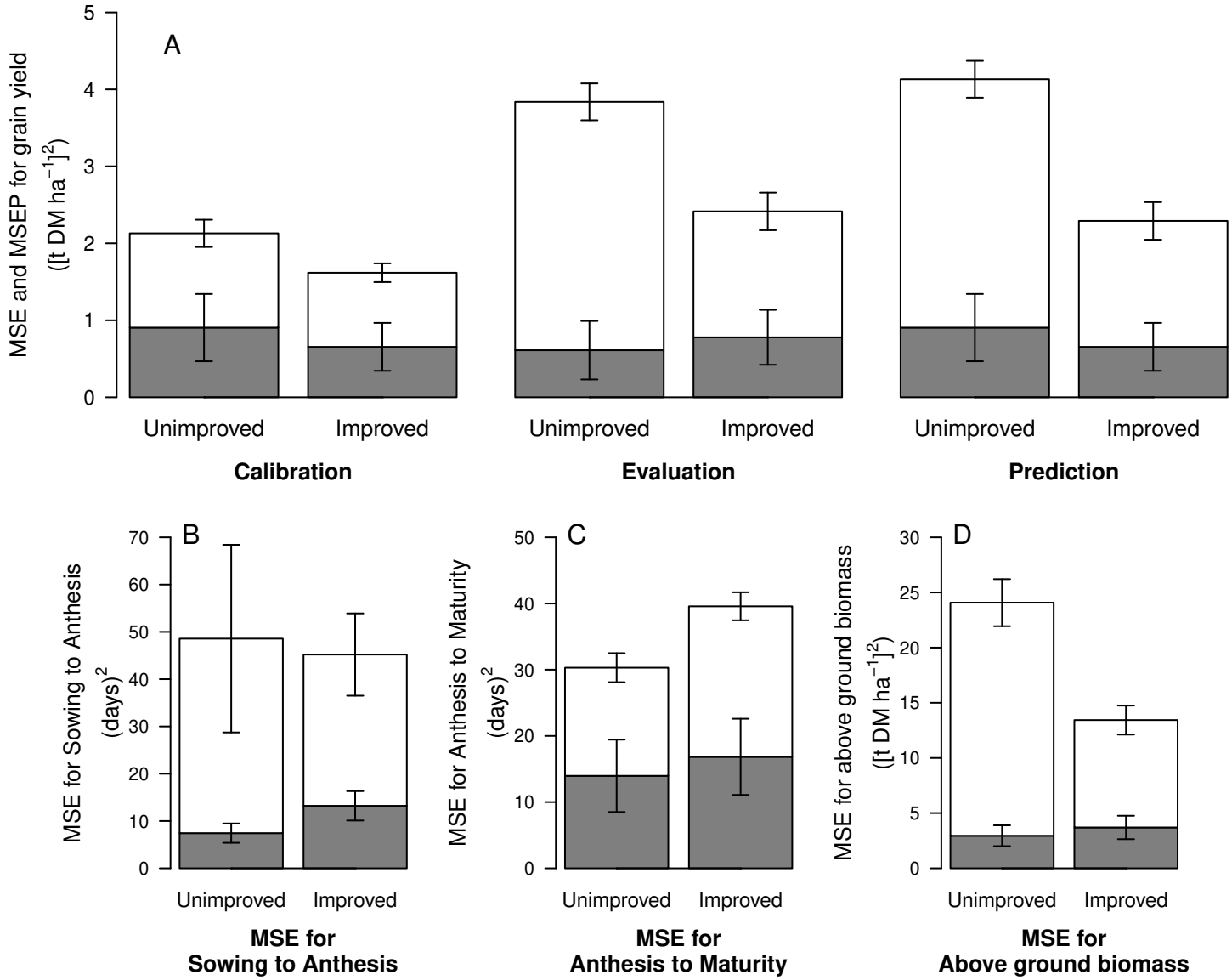
Figure

Figure 05



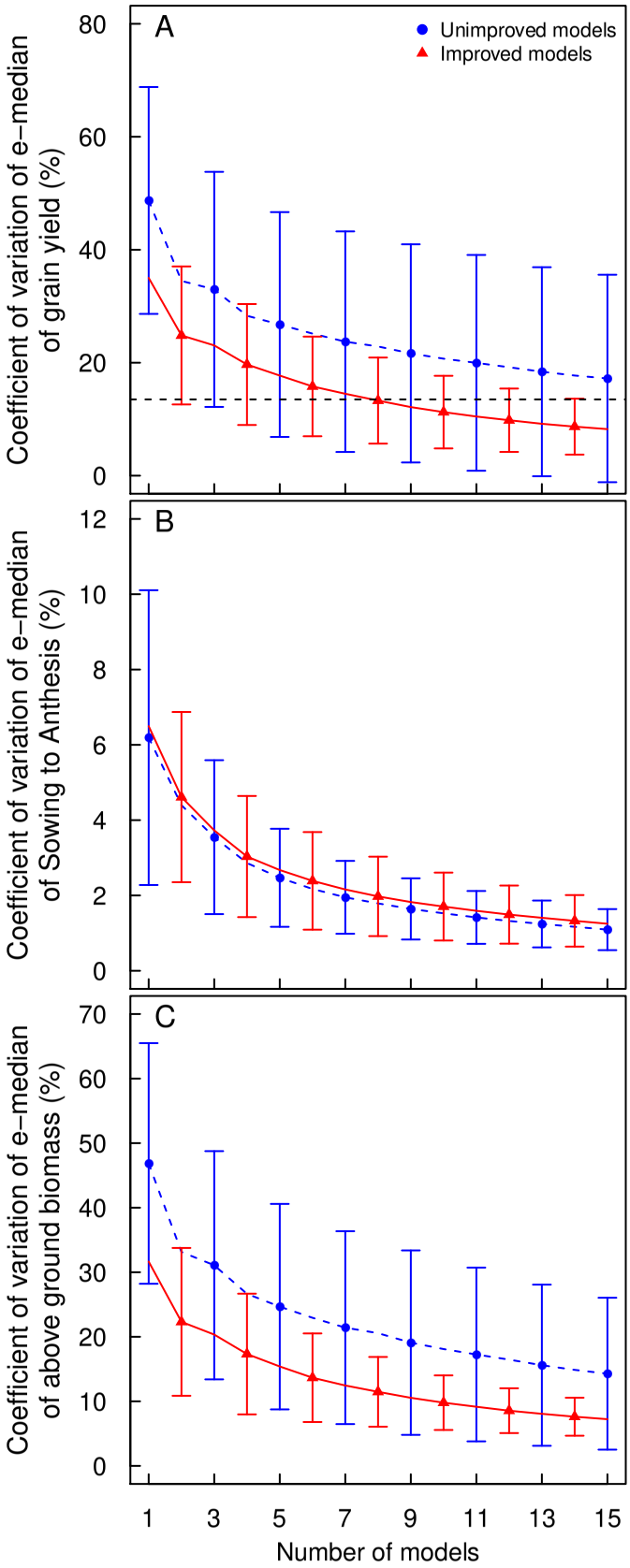
Figure

Figure 06



Figure

FIGURE 7



Figure

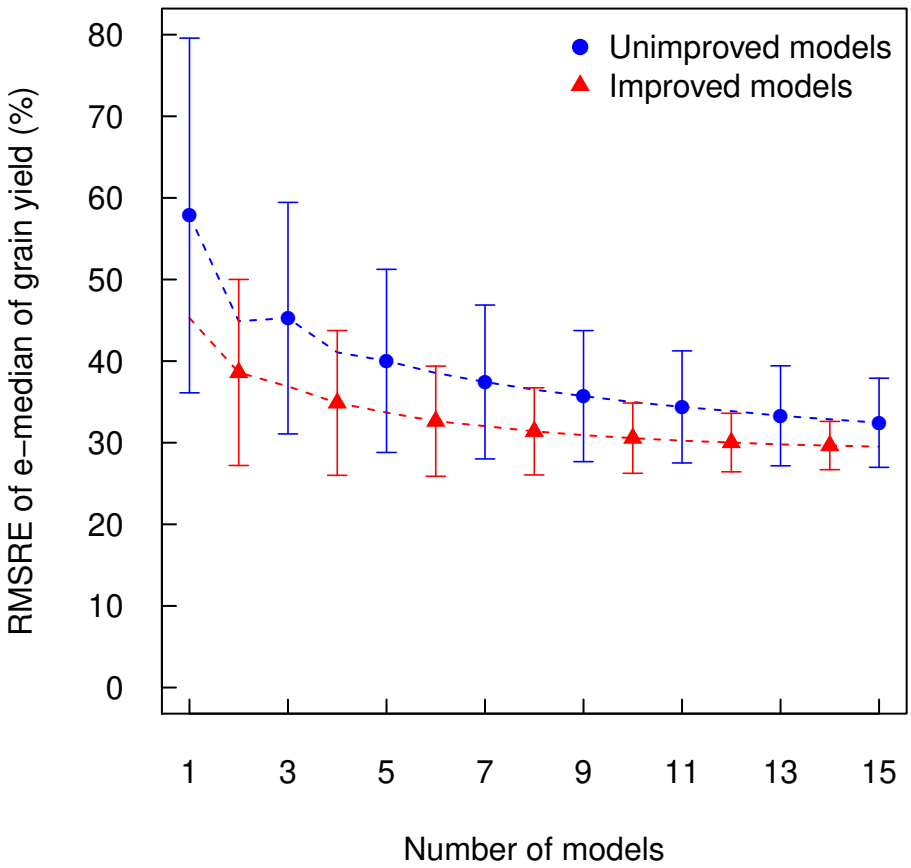



FIGURE 8