Article:

Ruane, AC, Hudson, NI, Asseng, S et al. (45 more authors) (2016) Multi-wheat-model ensemble responses to interannual climate variability. Environmental Modelling and Software, 81. pp. 86-101. ISSN 1364-8152

https://doi.org/10.1016/j.envsoft.2016.03.008

Multi-wheat-model ensemble responses to interannual climate variability

Alex C. Ruane[1], Nicholas I. Hudson[2], Senthold Asseng[3], Davide Camarrano[3,4], Frank Ewert[5], Pierre Martre[6,7], Kenneth J. Boote[3], Peter J. Thorburn[8], Pramod K. Aggarwal[9], Carlos Angulo[5], Bruno Basso[10], Patrick Bertuzzi[11], Christian Biernath[12], Nadine Brisson[13,14,*], Andrew J. Challinor[15,16], Jordi Doltra[17], Sebastian Gayler[18], Richard Goldberg[2], Robert F. Grant[19], Lee Heng[20], Josh Hooker[21], Leslie A. Hunt[22], Joachim Ingwersen[18], Roberto C Izaurralde[23], Kurt Christian Kersebaum[24], Soora Naresh Kumar[25], Christoph Müller[26], Claas Nendel[24], Garry O'Leary[27], Jørgen E. Olesen[28], Tom M. Osborne[29], Taru Palosuo[30], Eckart Priesack[10], Dominique Ripoche[11], Reimund P. Rötter[30], Mikhail A. Semenov[31], Iurii Shcherbak[32], Pasquale Steduto[33], Claudio O. Stöckle[34], Pierre Stratonovitch[31], Thilo Streck[18], Iwan Supit[35], Fulu Tao[30,36], Maria Travasso[37], Katharina Waha[26], Daniel Wallach[38], Jeffrey W. White[39] and Joost Wolf[40]

[1]National Aeronautics and Space Administration, Goddard Institute for Space Studies, New York, NY. [2]Columbia University Center for Climate Systems Research, New York, NY. [3]Agricultural & Biological Engineering Department, University of Florida, Gainesville, FL. [4]James Hutton Institute, Invergowrie, Dundee, Scotland, U.K.. [5]Institute of Crop Science and Resource Conservation, Universität Bonn, D-53 115, Germany. [6]National Institute for Agricultural Research (INRA), UMR1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), F-63 100 Clermont-Ferrand, France. [7]Blaise Pascal University, UMR1095 GDEC, F-63 170 Aubière, France, [8]Commonwealth Scientific and Industrial Research Organization, Agriculture Flagship, Dutton Park QLD 4102, Australia. [9]Consultative Group on International Agricultural Research, Research Program on Climate Change, Agriculture and Food Security, International Water Management Institute, New Delhi 110012, India. [10]Department of Geological Sciences and Kellogg Biological Station, Michigan State University, East Lansing, MI. [11]INRA, US1116 AgroClim, F- 84 914 Avignon, France. [12]Institute of Biochemical Plant Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, D-85 764, Germany. [13]INRA, UMR0211 Agronomie, F-78 750 Thiverval-Grignon, France. [14]AgroParisTech, UMR0211 Agronomie, F-78 750 Thiverval-Grignon, France. [15]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS29JT, UK. [16]CGIAR-ESSP Program on Climate Change, Agriculture and Food Security, International Centre for Tropical Agriculture, A.A. 6713, Cali, Colombia. [17]Cantabrian

Agricultural Research and Training Centre, 39600 Muriedas, Spain. [18]Institute of Soil Science and Land Evaluation, Universität Hohenheim, D-70 599 Stuttgart, Germany. [19]Department of Renewable Resources, University of Alberta, Edmonton, AB, Canada T6G 2E3. [20]International Atomic Energy Agency, 1400 Vienna, Austria. [21]School of Agriculture, Policy and Development, University of Reading, RG6 6AR, United Kingdom. [22]Department of Plant Agriculture, University of Guelph, Guelph, Ontario, Canada, N1G 2W1. [23]Department of Geographical Sciences, University of Maryland, College Park, MD 20782. [24]Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research, D-15 374 Müncheberg, Germany. [25]Centre for Environment Science and Climate Resilient Agriculture, Indian Agricultural Research Institute, New Delhi 110 012, India. [26]Potsdam Institute for Climate Impact Research, D-14 473 Potsdam, Germany. [27]Landscape & Water Sciences, Department of Primary Industries, Horsham 3400, Australia. [28]Department of Agroecology, Aarhus University, 8830 Tjele, Denmark. [29]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, RG6 6BB, United Kingdom. [30]Environmental Impacts Group, Natural Resources Institute Finland (Luke), FI-01370, Vantaa, Finland. [31]Computational and Systems Biology Department, Rothamsted Research, Harpenden, Herts, AL5 2JQ, United Kingdom. [32]Institute for Future Environments, Queensland University of Technology, Brisbane, QLD 4000, Australia, [33]Food and Agriculture Organization of the United Nations, Rome, Italy. [34]Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120. [35]Earth System Science-Climate Change and Adaptive Land-use and Water Management, Wageningen University, 6700AA, The Netherlands. [36]Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science, Beijing 100101, China. [37]Institute for Climate and Water, INTA-CIRN, 1712 Castelar, Argentina. [38]INRA, UMR1248 Agrosystèmes et Développement Territorial, F-31 326 Castanet-Tolosan, France. [39]Arid-Land Agricultural Research Center, USDA-ARS, Maricopa, AZ 85138. [40]Plant Production Systems, Wageningen University, 6700AA Wageningen, The Netherlands.

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

Re-Submission Draft – December 18th, 2015

Keywords: Crop modeling; uncertainty; multi-model ensemble; wheat; AgMIP; climate impacts; temperature; precipitation; interannual variability

Highlights:

- *Compares interannual climate response of 27 wheat models at four locations*
- *Calculates the diminishing return of constructing multi-model ensembles for assessment*
- *Identifies similarities and major differences of model responses*
- *Differentiates between interannual temperature sensitivity and climate change response*

Corresponding Author's Address
Alex C. Ruane
Climate Impacts Group, NASA Goddard Institute for Space Studies
2880 Broadway
New York, NY 10025, USA
Alexander.C.Ruane@nasa.gov
Phone: +1-212-678-5640; Fax: +1-212-678-5645

Abstract

We compare 27 wheat models' yield responses to interannual climate variability, analyzed at locations in Argentina, Australia, India, and The Netherlands as part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) Wheat Pilot. Each model simulated 1981-2010 grain yield, and we evaluate results against the interannual variability of growing season temperature, precipitation, and solar radiation. The amount of information used for calibration has only a minor effect on most models' climate response, and even small multi-model ensembles prove beneficial. Wheat model clusters reveal common characteristics of yield response to climate; however models rarely share the same cluster at all four sites indicating substantial independence. Only a weak relationship ($R^2 \leq 0.24$) was found between the models' sensitivities to interannual temperature variability and their response to long-term warming, suggesting that additional processes differentiate climate change impacts from observed climate variability analogs and motivating continuing analysis and model development efforts.

1. Introduction

Process-based crop simulation models have become increasingly prominent in the last several decades in climate impact research owing to their utility in understanding interactions among genotype, environment, and management to aid in planning key farm decisions including cultivar selection, sustainable farm management, and economic planning amidst a variable and changing climate (e.g., Ewert et al., 2015). In the coming decades climate change is projected to pose additional and considerable challenges for agriculture and food security around the world (Porter et al., 2014; Rosenzweig et al., 2014). Process-based crop simulation models have the potential to provide useful insight into vulnerability, impacts, and adaptation in the agricultural sector by simulating how cropping systems respond to changing climate, management, and variety choice. Such gains in insight require high-quality models and better understanding of model uncertainties for detailed agricultural assessment (Rötter et al., 2011). Although there have been a large number of studies utilizing crop models to assess climate impacts (Challinor et al., 2014a), a lack of consistency has made it very difficult to compare results across regions, crops, models, and climate scenarios (White et al., 2011a). The Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al., 2013; 2015) was launched in 2010 to establish a consistent climate-crop-economics modeling framework for agricultural impacts assessment with an emphasis on multi-model analysis, robust treatment of uncertainty, and model improvement.

A crop model's response to interannual climate variability provides a useful first indicator of model responses to variation in environmental conditions (Arnold and de Wit, 1976). A simulation model's ability to capture historical grain yield variability has shown it can serve as a sensible basis on which to demonstrate the utility of crop models among stakeholders and decision-makers (e.g., Dobermann et al., 2000). Considering the effort required in collecting data and calibrating a crop model for a particular application, previous studies have often relied upon only a single crop model and limited sets of observational data. This approach overlooks differences in plausible calibration methodologies as well as biases introduced in the selection of a single crop model and its parameterization sets; all of which may affect climate sensitivities (Pirttioja et al., 2015). The final decision-supporting information may therefore be biased depending on the amount of calibration data available and the crop model selected for simulations.

Here we present an agro-climatic analysis of 27 wheat models that participated in the AgMIP Wheat Model Intercomparison Pilot (described briefly in the next section and more completely in the text and supporting materials of Asseng et al., 2013; and Martre et al., 2015), with a focus on how interannual climate variability affects yield simulations and uncertainties across models. This is just one of several studies to emerge from the unprecedented Wheat Pilot multi-model intercomparison and it is intended to contribute to the overall effort by highlighting important areas for continuing analysis, model improvement, and data collection. As most climate impacts assessments cannot afford to run all 27 wheat models, for the first time we examine the consistency of agro-climatic responses across locations, models, and the extent of calibration information to determine whether a simpler, smaller multi-model assessment may be a suitable representation of the full AgMIP Wheat Pilot ensemble. The design of the AgMIP Wheat Pilot also enables a novel comparison of yield responses to interannual climate variability and to mean climate changes, testing the notion that the response to historical climate variability provides a reasonable analog for future climate conditions. The purpose of this analysis is to identify differences in model behaviors, data limitations, and areas for continuing research and model improvement.

2. Materials and Methods

2.1 The AgMIP Wheat Pilot

A total of 27 wheat modeling groups participated in the first phase of the AgMIP Wheat Model Intercomparison Pilot in order to investigate model performance across a variety of climates, management regimes, and climate change conditions (focusing on response sensitivity to temperature and carbon dioxide). This represented the largest multi-model intercomparison of crop models to date. Major climate change results for grain yields were presented by Asseng et al. (2013), while Martre et al. (2015) compared model performance across output variables against field observations. As those studies thoroughly documented the protocols and participating models of the Wheat Pilot's first phase, here we summarize the major elements with an emphasis on factors affecting interannual grain yield variability as simulated at four sites over the 1981-2010 historical period. Additional work from the Wheat Pilot's second phase have focused on response to increases in average temperature (Asseng et al., 2015), and the models are largely the same as those utilized in phase 1 and analyzed below.

2.1.1 Locations

The four locations simulated by participating wheat model groups are shown in **Table 1**, herein referred to as Argentina (AR), Australia (AU), India (IN), and the Netherlands (NL). Each location corresponded to a field trial ranked as either "gold" or "platinum" in AgMIP's field data standards (Boote et al., 2015), allowing for detailed model calibration and analysis with high-quality initial conditions, in-season measurements, phenology, and end-of-season records. Calibration in this study refers to the process of configuring a crop model for application at a given site, which typically entails the representation of soil properties, agricultural management, and coefficients representing the genetic properties of the cultivar planted; the core biophysical processes are properties of the model developed from extensive experimentation and are typically not adjusted to match field observations at these sites. These high-quality seasonal data unfortunately do not correspond to coincident long-term variety trials using the same management, cultivars, and soils that would be ideal to calibrate interannual variability (corresponding crop growth observations and long-term variety trials are quite rare, particularly in developing countries). Even where long-term variety trial data exist (and are publically available), considerable analysis is needed to attempt a direct comparison with multi-season crop model simulation given shifts in cultivars every 3-5 years (Piper et al., 1998; Dobermann et al., 2000; Mavromatis et al., 2001; Singh et al., 2014; Boote et al., 2015). As a result, analysis here follows many crop modeling studies in utilizing a single-year or short-period (~5 years or less) field dataset for calibration and then relying on soil properties, plant genetics, and established model biophysics to determine interannual variability rather than specifically calibrating internal parameters of response. Palosuo et al. (2011) examined the potential of a smaller multi-model ensemble to reproduce interannual yield variability of variety trial for wheat having only two sites with a longer yield series (14+ seasons) but limited data for calibration, finding errors in each model but much improved statistics for the multi-model ensemble mean. Rötter et al. (2012) came up with similar results for barley model simulations.

Daily climate data (maximum and minimum temperatures, solar radiation, precipitation, wind speed, vapor pressure, dew-point temperature, and relative humidity) were compiled from local observations with missing data filled using the NASA Modern Era Retrospective-analysis for Research and Applications (MERRA; Rienecker et al., 2011) and the NASA/GEWEX Solar Radiation Budget (Stackhouse et al., 2011; White et al., 2011b). The Indian site was irrigated according to the field trial applications. The irrigation (date and amount) of the experimental year (Table 1) was used as input to the models for simulating the 30-years historical period although this may not be sufficient for each year. The other sites were rain-fed. Calibration procedures varied from model to model (generally using the field data to detail crop management and soil properties and then configuring cultivar parameters to match growth stage periods). To isolate the climatic signal, the same configuration was used for the historical simulations, future simulations, and the temperature and CO_2 sensitivity tests at each site. The specific calibration approaches were discussed by Challinor et al. (2014b), who found no clear relationship between the number of parameters calibrated and the relative error of harvest index or grain yield. They further noted that this was consistent with compensating errors that can be a benefit of multi-model ensembles but found no evidence of over-tuning in the AgMIP Wheat Pilot.

Table 1: Locations simulated in AgMIP Wheat Pilot (for more details see Martre et al., 2015)

Parameter	**Location**			
	Argentina	**Australia**	**India**	**Netherlands**
Location	Balcarce	Wongan Hills	Delhi	Wageningen
Latitude	37.75°S	30.89°S	28.38°N	51.97°N
Longitude	58.30°W	116.72°E	77.12°E	5.63°E
Cultivar	Oassis	Gamenya	HD2009	Arminda
Irrigated	No	No	Yes (383 mm)	No
N fertilizer (kg N ha^{-1})	120	50	120	160
Planting date	10 August	12 June	23 November	21 October

Anthesis date	23 November	1 October	18 February	20 June
Harvest date	28 December	16 November	3 April	1 August
Year of experiment	1992	1984	1984-1985	1982-1983

Additional observations of yields in these regions potentially provide a target for accurate interannual variability that the models are challenged to match. We therefore examined 1981-2010 national level yield data from the UN Food and Agricultural Organization (http://faostat.fao.org/), overlapping district-level yields (in Australia; India: Ministry of Agriculture, Government of India; and the Netherlands: Central Bureau of Statistics, the Hague, STATLINE), and nearby variety trials (in Argentina: RET, www.inase.gov.ar; and the Netherlands: Central Bureau of Statistics, the Hague, STATLINE) as a point of comparison against simulated yields. It is not expected that these four modeling locations are precise representations of the surrounding region; each represents carefully-controlled field trials in one location within countries characterized by substantial differences in soils, climates, cultivars, and management practices.

2.1.2 Wheat Models

Table 2 lists the 27 wheat models that simulated each of the four sites. Details of the processes and parameter settings that distinguish each of these models are provided in the supplementary material (particularly Table S2) of Asseng et al. (2013). The AgMIP Wheat Pilot’s first phase agreed on a policy of model anonymity in the presentation of results, so for the purpose of this study the models will be referred to only by a number assigned at random. This allowed us to still determine the range of responses across these models’ native configurations and elucidate how the selection of a crop model contributes to uncertainty in interannual yield simulations and related decisions. The specific mechanisms for each model's response are being considered in ongoing analyses and future intercomparison design.

2.1.3 Types of simulation exercises

Wheat Pilot protocols were designed to investigate whether limitations in data (which hamper the calibration of crop models in many locations) substantially affect the accuracy of yield simulation and/or alter the simulated sensitivity to climate variability and climate changes. Participants were therefore instructed to perform simulations in two steps:

1) Low-information simulations: Weather data, planting, crop emergence, flowering, and physiological maturity dates, field management information, and soil characteristics and initial conditions were provided but no information was provided on end-of-season yields or in-season crop growth and soil water and nitrogen (N) dynamics. This subset of field experiment data was referred to as "blind test" simulations by Asseng et al. (2013), and represent the types of data that may be accessible for a large number of locations.
2) High-information simulations: In addition to the above data modelers were also provided with in-season growth dynamics from the same years' field trial, including, leaf area index (all sites but AU), total above ground biomass and N, root biomass (at IN only), cumulative evapotranspiration (at AU and IN only), plant available soil water and soil inorganic N contents within the season (at AU and NL only), and end-of-season grain yield and protein concentration, and grain density measurements. Plant components (green leaves, dead leaves, stem, and chaff) biomass and N contents were also available at NL. This full set of experimental data was referred to as "full calibration" simulations

by Asseng et al. (2013) and is equivalent to the more rare gold or platinum standards set by Kersebaum et al. (2015) and Boote et al. (2015).

Analysis by Asseng et al. (2013) revealed a considerable reduction of biases between field observations and yields using the high-information simulations, but noted that both the low- and high-information simulations showed a similar response to changes in mean temperature and CO_2 concentrations.

Table 2: Crop models included in AgMIP Wheat Pilot (in alphabetical order; for more information and details on the processes modeled in each model see supplementary materials of Asseng et al., 2013)

Model	Version	Model description and applications	Web address
APES-ACE[*]	V. 0.9.0.0	(Donatelli et al., 2010; Ewert et al., 2011a)	http://www.apesimulator.it/default.aspx
APSIM-Nwheats	V.1.55	(Asseng et al., 2004; Asseng et al., 1998; Keating et al., 2003)	http://www.apsim.info
APSIM-wheat	V.7.3	(Keating et al., 2003)	http://www.apsim.info/Wiki/
AquaCrop[*]	V.3.1+	(Steduto et al., 2009)	http://www.fao.org/nr/water/aquacrop.html
CropSyst	V.3.04.08	(Stockle et al., 2003)	http://www.bsyse.wsu.edu/CS_Suite/CropSyst/index.html
DSSAT-CERES-Wheat	V.4.0.1.0	(Hoogenboom and White 2003; Jones et al., 2003), (Ritchie et al., 1985)	http://www.icasa.net/dssat/
DSSAT-CROPSIM-Wheat		(Hunt and Pararajasingham 1995; Jones et al., 2003)	http://www.icasa.net/dssat/
Ecosys		(Grant et al., 2011)	https://portal.ales.ualberta.ca/ecosys/
EPIC wheat		(Kiniry et al., 1995; Williams et al., 1989)	http://epicapex.brc.tamus.edu/
Expert-N - CERES - wheat	ExpertN 3.0.10 Ceres 2.0	(Biernath et al., 2011; Priesack et al., 2006; Ritchie et al., 1987; Stenger et al., 1999)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
Expert-N - GECROS - wheat	ExpertN 3.0.10	(Biernath et al., 2011; Yin and van Laar 2005; Stenger et al., 1999)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
Expert-N - SPASS - wheat	ExpertN 3.0.10	(Biernath et al., 2011; Priesack et al., 2006; Stenger et al., 1999; Wang and Engel 2000)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
Expert-N - SUCROS – wheat	ExpertN 3.0.10 Sucros2	(Biernath et al., 2011; Goudriaan and Van Laar 1994; Priesack et al., 2006; Stenger et al., 1999)	http://www.helmholtz-muenchen.de/en/iboe/expertn/
FASSET	V.2.0	(Berntsen et al., 2003) (Olesen et al., 2002)	http://www.fasset.dk
GLAM-wheat[*]	V.2	(Challinor et al., 2004; Li et al.,	http://see-web-

		2010)	01.leeds.ac.uk/research/icas/climate_change/glam/download_glam.html
HERMES	V.4.26	(Kersebaum 1995; Kersebaum 2007; Kersebaum 2011; Kersebaum and Beblik 2001)	www.zalf.de/en/forschung/institute/lsa/forschung/oekomod/hermes
InfoCrop	V.1	(Aggarwal et al., 2006)	http://www.iari.res.in
LINTUL-4	v.1	(Shibu et al., 2010; Spitters and Schapendonk 1990)	http://models.pps.wur.nl/models
LPJmL*		(Bondeau et al., 2007; Fader et al., 2010; Waha et al., 2012)	http://www.pik-potsdam.de/research/projects/lpjweb
MCWLA-Wheat*	V2.0	(Tao et al., 2009a; Tao and Zhang 2010; Tao et al., 2009b; Tao and Zhang 2011)	---
MONICA	V.1.0	(Nendel et al., 2011)	http://monica.agrosystem-models.com
O'Leary-model	V.7	(Latta and O'Leary 2003; OLeary and Connor 1996a; b; Oleary et al., 1985)	Primary documentation for V7 (V3 (O'Leary and Connor 1996a; b), with incremental documentation thereafter.
SALUS	V.1.0	(Basso et al., 2010; Senthilkumar et al., 2009)	www.salusmodel.net
Sirius2010		(Jamieson and Semenov 2000; Jamieson et al., 1998; Lawless et al., 2005; Semenov and Shewry 2011)	http://www.rothamsted.ac.uk/mas-models/sirius.php
SiriusQuality	V.2.0	(Ferrise et al., 2010; He et al., 2011; He et al., 2010; Martre et al., 2006)	http://www1.clermont.inra.fr/siriusquality
STICS	V.1.1	(Brisson et al., 2003; Brisson et al., 1998)	http://www6.paca.inra.fr/stics_eng/
WOFOST*	V.7.1	(van Diepen et al., 1989; Supit and van Diepen, 1994; Boogard et al., 1998)	http://www.wofost.wur.nl

The 1981-2010 historical simulations that form the bulk of these analyses also served as the historical basis for climate change simulations conducted by each wheat-modeling group. The same model configurations were therefore forced by the same climate time series and baseline carbon dioxide concentrations but with historical temperatures adjusted by -3˚C, +3˚C, +6˚C, and +9˚C every day of the year. As initial soil conditions and crop management (including sowing date and nitrogen fertilizer application) were kept constant over the 30-year period, these simulations allow for a comparison between model responses to interannual climate variability and to mean climate changes. The re-initialization of soil conditions each year reduces the carry-over effects of multi-year droughts, which reduces overall interannual variability. This is common in agricultural modeling applications (particularly those that examine future climate change where the sequence of events is more difficult to project than mean conditions), but sequential simulations are an important developmental priority for more accurate representation of extreme events and soil degradation (Basso et al., 2016) and crop rotation effects (Kollas et al., 2015).

2.2 Performance of Ensemble

Martre et al. (2014) compared grain yield, protein content concentration, and in-season and end-of-season variables within the 27 wheat model simulations against observations at each of the four pilot locations. Although some models had the closest match to specific observations, across all observed variables the 27-model unweighted arithmetic ensemble mean performed best, in line with earlier findings based on smaller model ensembles even when used to reproduce interannual yield statistics (Palosuo et al., 2011; Rötter et al., 2012). Thus, while each wheat model has its own biases and accuracies, the errors across models tended to compensate and the resulting ensemble had additional value (see also Challinor et al., 2014b). The superior performance of the ensemble also reflected that wheat models have evolved with enough independence in approaches to achieve a random distribution of biases for most variables rather than leading to the emergence of common biases.

In light of the superior performance of the 27-member ensemble mean in reproducing field observations across the four sites (and the lack of long-term historical yield observations at each location), for the purposes of this study we utilize the full, 27-model unweighted arithmetic mean ensemble as the basis for comparison of each model's climate response.

2.3 Methods of analysis

2.3.1. Agro-climatic correlations

As each of the simulations held management constant throughout the 1981-2010 simulation period and soils were re-initialized each year (with the exception of LPJmL, which did not reinitialize soil water), interannual yield variability is a result of model responses to climate factors. Chief among these are precipitation, temperature, and solar radiation, which are likely to affect crop growth on a number of time scales. Here we focus on the effects of variability in mean values over the growing season, using Pearson's correlations against grain yield to determine key sensitivities within each crop model. Additional variance is likely explained by climate variables at sub-seasonal time scales (particularly when extreme conditions align with vulnerable phenological stages), which merits further examination in future studies. Correlation was chosen as a simple illustration of association between climate and crop model response, although aspects related to non-linearity and thresholds may not be captured. Future work may also consider associative metrics such as the probability of detection for extreme events as a way of isolating important properties of observations and models (Glotter et al., 2016).

As most studies will not have the luxury of running all 27 wheat models, we investigate the expected benefit of adding each additional member to a multi-model subset to converge on behaviors captured by the full 27-model ensemble. Without running the full analysis it is not possible to know whether the models that are available are among the best or worst for a given site's climate variability response, so we utilize an 80%-exceedance threshold as a practical risk in simulation design. Results therefore focus on the correlations that would be exceeded by 80% of the possible combinations for any number of combined models.

2.3.2 Agro-climatic clustering

We employed the k-means clustering technique to form clusters of wheat models that are characterized by similar correlations between yield and growing season temperature, precipitation, and solar radiation (with equal weighting for all). K-means is an iterative process by which models are regrouped until silhouette values (i.e., similarity between each model and the other members of its cluster) are maximized. For each location we examined the results with three, four, and five clusters and visually selected the number that best captured cohesive groupings in the climate-sensitivity space (this resulted in three clusters in both Argentina and India and four clusters in both Australia and the Netherlands). Fewer clusters than this grouped models with substantially different yield sensitivities to climate variability in the same cluster, while more clusters tended to unnecessarily divide similarly-responsive models. As each model belongs to a specific cluster at each location, we utilize the frequency that two models appear in the same clusters across the four sites as a metric of model similarity.

3. Results and discussion

3.1 Baseline interannual variability

Figure 1 presents the 1981-2010 yields for the four Wheat Pilot locations from 27 wheat models, the full model ensemble, and national and regional yields. These high-information simulation results indicate uncertainty across the model ensemble, although common differences in mean yield across the four locations are clear (as discussed by Asseng et al., 2013, and Martre et al., 2015). Simulations exceed national and regional yields in each location, as wheat models often do not include the effects of pests, diseases, poor crop management due to labor or equipment shortages, waterlogging, and other factors that are common on farms outside of experimental plots. Model results are therefore more representative of yield potential (Evans and Fischer, 1999) than the more complex conditions of a typical farmer's field. The other source of variation in the gray lines within Figure 1 comes from the less explored interannual variability of simulated yields, which is the focus of analyses below. Interannual variability is reduced in the model ensemble, as would be expected from averaging, although noteworthy variations suggest that there are common behaviors across the crop model responses. Simulated yields (which examine a single field) are characterized by greater interannual variance compared to the national and regional level observations, likely because heterogeneities in soils, climate, cultivars, and management reduces extreme year anomalies when aggregated to scales that may exceed those of a given extreme event (Ewert et al., 2011b). Only variety trials (in Argentina and the Netherlands) contain mean and variance of yields that are similar to the simulations, although differences in management and the varieties cultivated also reduce the utility of these records as a basis for truth in the comparison of models.

Discrepancies between various observational sources and the experimental field simulated by the wheat models are large enough to caution against an expectation that the models would reproduce national, regional, or trial-based observational records over the historical period. These discrepancies are often due to the set up of the simulations from the single field experiment not representing the diversity of soils, management and cultivars which affected the regional and national yield data (but are not documented). Also, yield variability is often driven by factors other than weather (Ray et al. 2015) and models that are driven by variations in weather only are bound to not reproduce observational records. As noted above, we therefore turn to the High-information ensemble average (dark line in Figure 1) as the standard for the individual crop models given its superior performance in producing the full range of field observations (Martre et al., 2015). The ensemble also reduces interannual variability through the averaging of multiple models' potentially uncorrelated anomalies.

3.2 Effect of calibration on climate sensitivity

The Wheat Pilot's protocol for Low-information and High-information experiments provides a useful examination of the ways in which model calibration has the potential to affect the resulting response to climate variability. **Figure 2** illustrates this sensitivity to calibration information via the correlation of each individual model's low-information results with the full ensemble of Low-information simulations (LL), the correlation of each model's Low-information result with the full ensemble of High-information simulations (LH), and the correlation of each model's High-information results with the full ensemble of High-information simulations (HH).

Correlations do not change dramatically between the Low- and High-information simulations for the vast majority of wheat models at each of the four locations. The exceptions feature both substantial improvements (e.g., Model #25 in Argentina) and declines (e.g., Model #10 in Australia) in correlations as additional information is provided. In these cases calibration to cultivars, soil conditions, or other internal parameters may have improved the experimental year's results but also affected climate sensitivity via shifts in the resilience to heat, water, and/or frost stresses. Effects of calibration strategy on simulations of climate change impact were also examined by Challinor et al. (2014b) and for simulations of crops across Europe (Angulo et al., 2013). The relative lack of different sensitivities between the Low- and High-information simulations could also be explained by the fact that each was simulated by the same model experts for a given model, and that additional data provided for the High-information

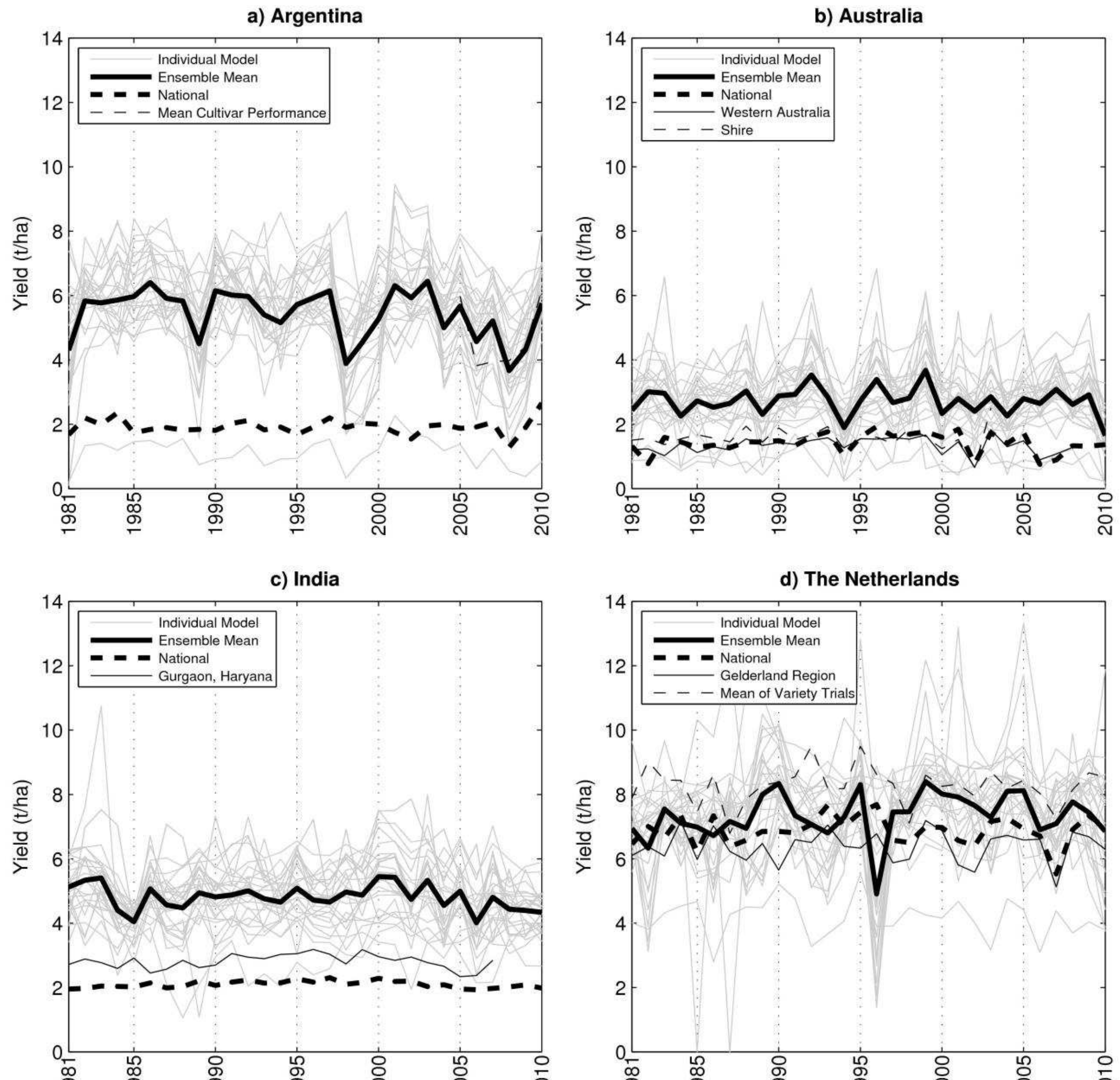

Figure 1: Historical period grain yields for a) Argentina, b) Australia, c) India, and d) The Netherlands, including the individual crop models at single simulation locations (gray lines), mutli-model ensemble mean (black solid line), and observations from national, regional, and local field trial data. Linear trends were removed from observational data at all but the Argentinian site (which had no significant trend). Modeled yields are the result of the high-information calibration simulations.

simulations were mostly limited to details on the crop itself. Additional information about the soil environment, in particular, would have potentially altered the sensitivity to interannual rainfall anomalies.

A comparison between the LL and HH correlations indicates that most models have the same relationship with the full ensemble regardless of the level of calibration information. Where LH and HH correlations are similar for a given model there is little benefit from additional calibration in terms of interannual climate response, as the Low-information results perform just as well as the High-information results against the High-information ensemble standard. HH correlations are at least higher than LL correlations in the majority of cases, suggesting that additional calibration information does tighten the spread of models around the ensemble mean and thus improve the performance of several models. This benefit is blurred by the likelihood that the fully-calibrated set of models would be expected to have closer agreement among members; however, it is important to note that calibration data at each site were only provided for a single year, making it impossible to directly calibrate the interannual variability examined here. This is a typical limitation for crop model simulations, as there are few long-term field trials that would allow full calibration of interannual variability. Also calibration in many cases focuses on minimizing error between modelled and observed results for the calibration dataset, which may have little influence on model responses to variation in environmental conditions that may be controlled by model structure and parameters other than those in focus for the calibration. The remainder of this study will focus on the High-information simulation sets, as these are likely to be of highest fidelity. Agro-climatological mechanisms at the root of these correlations are explored in Section 3.4 below.

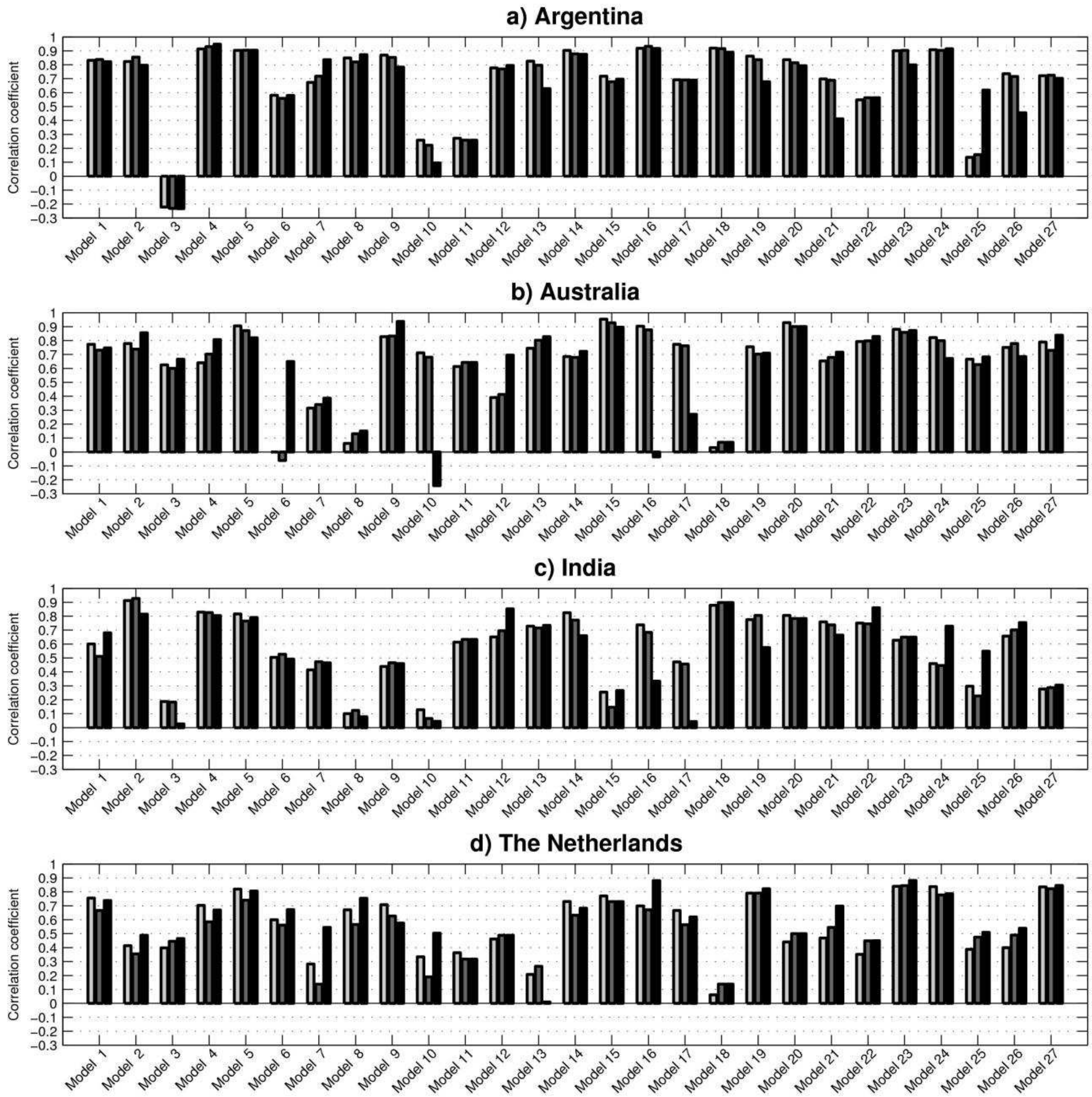

Figure 2: Single model run correlations against ensemble mean during 1981-2010 for (a) Argentina; (b) Australia; (c) India; and (d) The Netherlands. The correlation between the Low-information model runs and the Low-information ensemble mean (LL) is displayed in light gray, the correlation between the Low-Information model runs and the High-information ensemble mean (LH) is displayed in dark gray, and the correlation between the High-information model runs and the High-information ensemble mean (HH) is displayed in black.

3.3 Benefit of multi-model ensemble

The 27-model community approach of the AgMIP Wheat Pilot is not possible in the vast majority of crop model applications. Instead, what is needed is prior information that aids in the construction of a practical subset of models with a high likelihood of representing the larger ensemble. Beginning on the left-hand side of **Figure 3** (representing the use of a randomly selected single model), the plotted value represents the Pearson's correlation (against the full High-information ensemble) that would be exceeded by 80% of the individual models. This value is highest for Argentina (where 80% of the models exceed r = 0.50) and lowest for India (r = 0.28). Introducing a second model results in (27*26)/2=351 possible combinations, but 80% of them have a correlation of at least r = 0.71 in Argentina and r = 0.53 in India. Across the four sites, the benefit of adding a second model to a climate variability analysis is therefore an increase of +0.23 in its likely correlation with the full ensemble, with gains highest in Australia (+0.33) and lowest in the Netherlands (+0.13). Adding a third model also substantially increases the 80%-likely correlation, although the average increase is reduced (+0.11). The additions of a fourth and fifth model (increasing correlations by an average of 0.06 and 0.04, respectively) to the subset are also beneficial and lead to very high correlations, but the increases begin to be small in comparison to the effort likely required to calibrate an additional model (and collaborate with an additional modeling group) for the effort.

Efforts to include a second and third model therefore provide substantial benefit to climate variability simulations; however, investment in including additional models has a diminishing return. These results suggest a benefit at smaller subsets to account for interannual climate variability than the 5- to 10-member subsets that AgMIP crop model pilots identified as beneficial by comparing multi-model convergence against the 13.5% error that is common in field observations for wheat (Asseng et al., 2013) and maize (Bassu et al., 2014) or the 15% observational error for rice (Li et al., 2014). The analyses were also conducted using a 70% and 90% threshold, with consistent patterns of benefit but the higher thresholds further emphasizing the risks of the worst model being randomly selected.

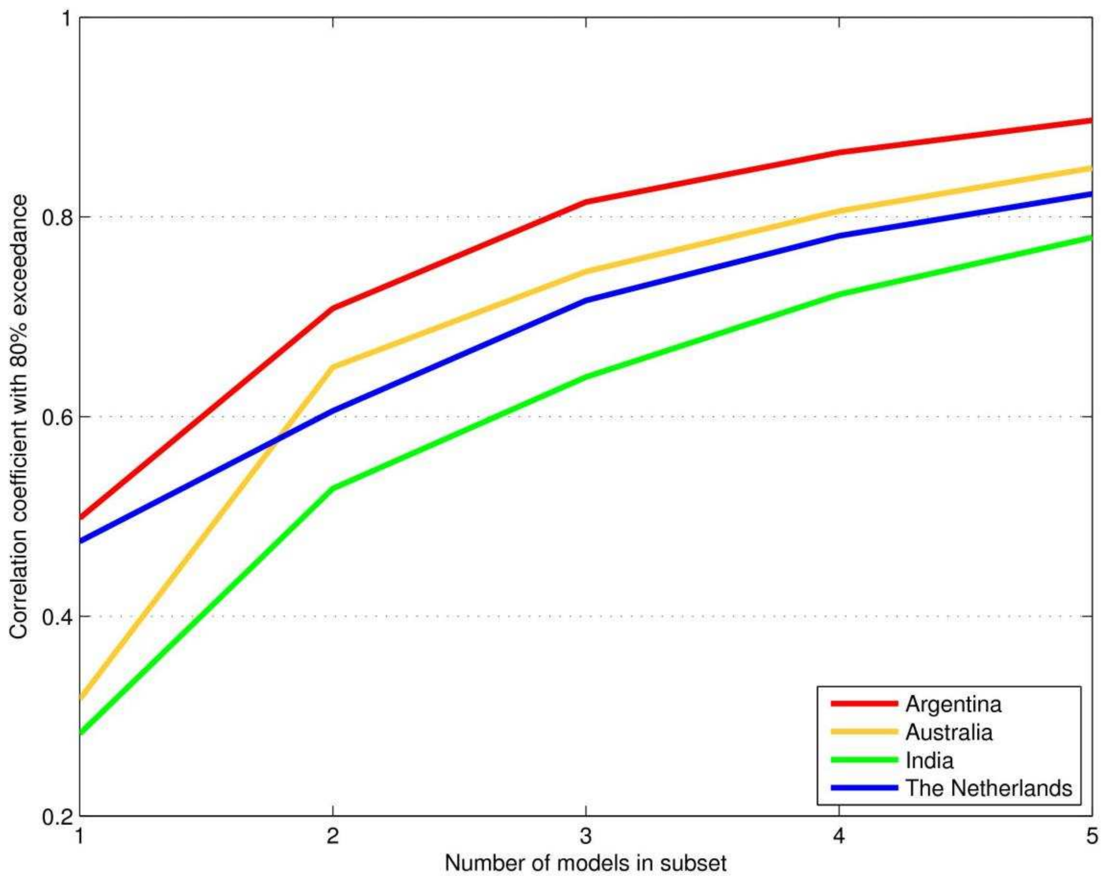

Figure 3: Improvement in correlations with each additional model within a multi-model subset of the full ensemble. For each number of models included in the subset N, the value shown represents Pearson's correlation coefficient between the subset's mean yield and the full ensemble's mean yield and that would be exceeded 80% of the time given a random selection of N models from the full set of 27 wheat models. Simulations were performed at single locations in each country (see Table 1) after calibration with High information, and all possible combinations of N models were tested.

3.4. Agro-climatic Sensitivity

Correlations of the 1981-2010 modeled grain yields and observed grain yields with mean growing season solar radiation, temperature, and precipitation are shown in **Figure 4** across the four locations. In Argentina simulated grain yields are positively correlated with wet seasons in all but one model, with more than 75% of the models demonstrating significant correlations. A strong sensitivity to rainfall anomalies is also seen in the cultivar trials; however, national grain yields are not significantly correlated with the precipitation at Balcarce, Argentina, as the wheat area covers a much larger region. The simulations and cultivar trials agree that lower temperatures significantly favor grain yields, with even the national grain yields following suit as warm and cooler seasons tend to spread more widely than the precipitation anomalies. At all sites, for both temperature and precipitation, the magnitude of the ensemble average's correlation is substantially higher than that of the median model; indicating that precipitation and temperature sensitivities are a unifying factor describing grain yield across the model members. Solar radiation variability is not significantly correlated for the bulk of models.

The Australian location is characterized by an even stronger sensitivity to rainfall. This site is also significantly sensitive to solar radiation anomalies, with negative correlations suggesting interdependence as cloudier seasons correspond with wetter conditions. National and regional yields are less responsive to precipitation anomalies and are governed more by temperature, as temperature anomalies may be widespread while droughts in the east are often offset by wetter conditions in the west.

Simulated yields at the Indian site are significantly correlated with precipitation despite irrigation applications totaling 383 mm over the growing season using fixed application dates (as applied in the field experiment). While an irrigation amount of 383 mm was sufficient for the 1984-1985 field trial, in other years the amount and timing of these applications may not have been adequate to prevent water stresses from influencing crop growth and final yields. It is also possible that precipitation anomalies are correlated with particular temperature and solar radiation regimes

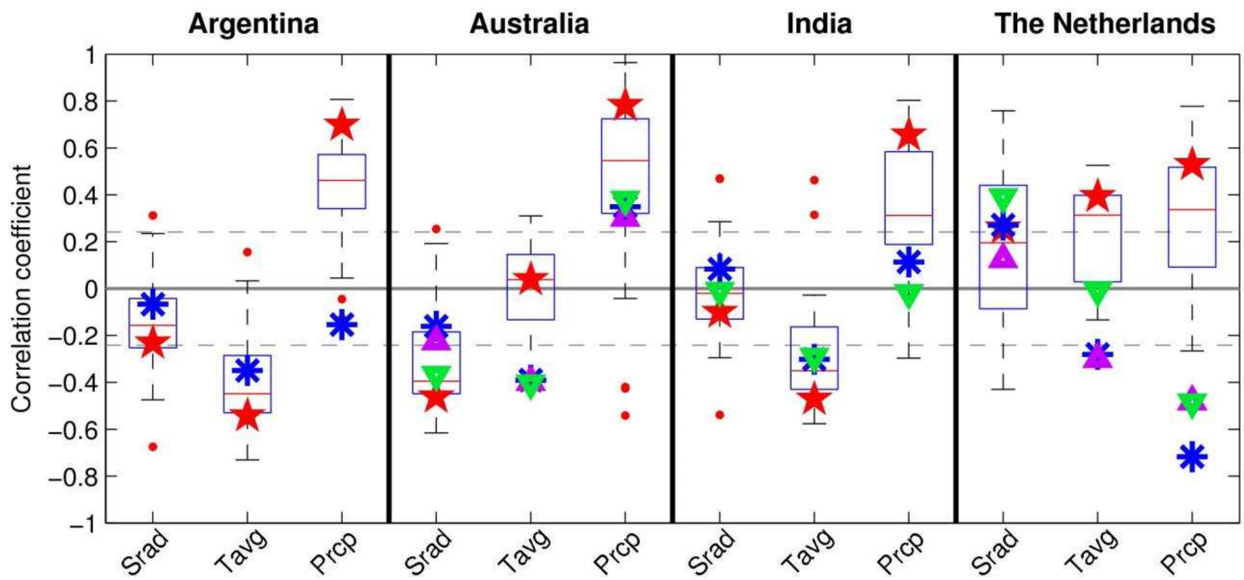

Figure 4: Box-and-whiskers plots of Pearson's correlation coefficients between the 27 wheat models' 1981-2010 simulated grain yields at single locations in each country and corresponding growing season mean solar radiation (Srad), average temperature (Tavg) and precipitation (Prcp). The median of the model simulations is marked by the red line, the box contains the middle two quartiles (from 25% to 75%), and the whiskers extend to the most extreme data points of the simulations that are not considered outliers (displayed as red dots). The correlation of the ensemble performance (red star), national observations (blue asterisk), regional observations (magenta triangles; where available), and the mean of other field trial results or local observations (green triangles) over the years data were available are also presented (as in Figure 1). Dashed lines indicate thresholds for correlations that are significant at the 90th percentile (t-test).

that are favorable for irrigated wheat growth. Cool seasons here are favorable for wheat production, and solar radiation correlations are not significant. National level correlations with the Delhi weather series are understandably weaker for all variables, as heterogeneous climate across India's wheat-growing regions reduces the prominence of anomalies and results in insignificant correlations in all but average temperature.

Wheat at the Netherlands site follows a different agro-climatic pattern from that at the other three sites. Warm seasons are positively correlated with yields in the bulk of models, suggesting a growing degree day limitation. Simulations and observations also suggest a radiation limitation at this high latitude, with sunnier seasons (and the associated temperature and rainfall patterns) favoring higher yields. The field site is notably different from the regional and national level observations in that the aggregated observations are either not correlated with temperature or suggest that yields favor cooler temperatures. The models also indicate stronger yields in wet years, while observations indicate better production during drier seasons. This likely comes from the fact that local and regional management of shallow groundwater tables in this region helps control against water stress but this management is not considered in the models at the test site. Contrary to the models' perception of drought, elevated regional yields are recorded in dry seasons as higher solar radiation and groundwater provisions increase yield potential (Asseng et al., 2000).

3.5. Clusters of agro-climatic response

Figure 5 shows each of the 27 wheat models as plotted on a three-dimensional space of temperature, precipitation, and solar radiation correlations with that model's grain yield. Models falling in the same agro-climatic cluster are represented with a common symbol and color. The full ensemble average and cluster averages do not fall as an average of the individual model members' correlations as the ensemble averaging reduces individual models' yearly anomalies to produce a unique time series. The results illustrate that the model spread is not randomly distributed in the agro-climatic sensitivity space, but rather distinct families of responses are evident. Several clusters also correspond much more closely with the full ensemble average responses.

Figure 6 shows the spread of model correlations within each cluster as well as the cluster ensemble correlations against the full 27-model ensemble's interannual yield variability. One or two clusters at each location demonstrate substantially better coherence to the ensemble average than the others. Even within a given cluster there are substantial differences in correlation between individual models and the ensemble average; particularly among clusters that are furthest from the ensemble average sensitivities (e.g., the "x" cluster in Argentina or the diamond cluster in Australia). The ensemble average for each cluster is also a marked improvement on the median model within that cluster, although occasionally there is one model that outperforms even the cluster mean.

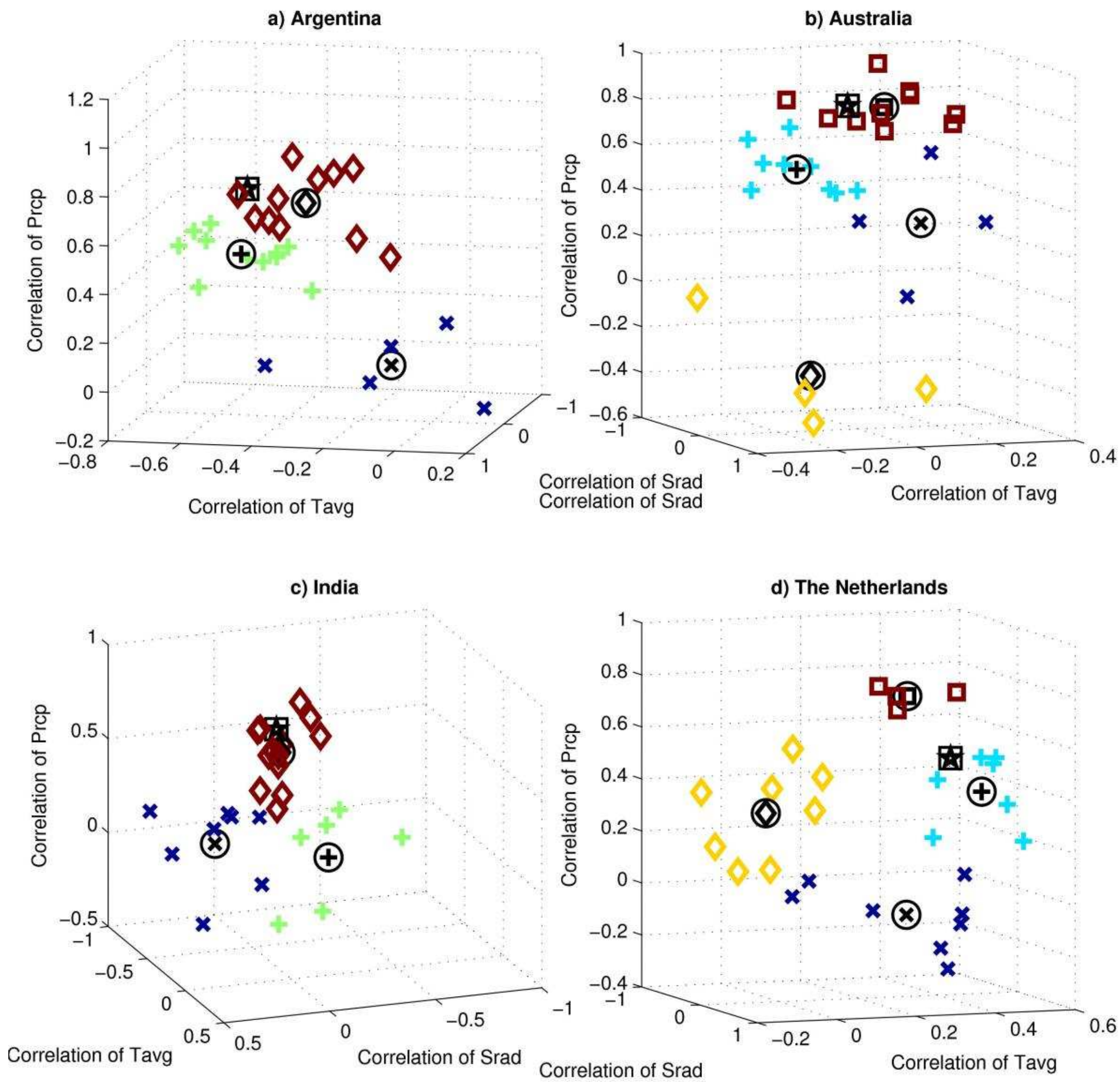

Figure 5: Clusters of the 27 wheat model simulations (cluster membership denoted by shape of symbols), the ensemble average, and observational data according to their grain yield correlation coefficients versus mean growing season solar radiation (Srad), average temperature (Tavg), and precipitation (Prcp) from 1981-2010 for single locations in (a) Argentina; (b) Australia; (c) India; and (d) The Netherlands. The correlation coefficients of the ensemble yield performance (boxed star) and the centroids of the clusters (corresponding symbols with circles) are also presented. Note that the perspective is rotated and axes limits adjusted in each panel in order to best visualize the differences in the model clusters.

Despite the fact that many of these wheat models have common heritage in pioneering crop modeling groups and approaches developed in the last 30 years, only two pairs of models (#1/#5 and #20/#22 from Figure 2; making <0.3% of possible combinations and thus potentially just a coincidence) fall in the same agro-climatic cluster at all four Pilot locations. 7% of model pairs fall in the same cluster at three of the four sites, while 24% of model pairs are never in the same cluster. The remaining 69% of model pairs share one or two clusters, which would be expected for independent models. No individual model stands out as being particularly divergent from the others, as each model has at least three other models that never appear in the same cluster, and at least four models that fall in the same cluster for two or more sites. Only one model falls into the highest-correlating cluster at all four locations, and likewise only a single model always falls into the lowest-correlating cluster. In total 15 different models are included in the lowest-correlating cluster for at least one site, and 21 different models are part of the highest-correlating cluster at least once. This independence likely contributes to the strength of the full ensemble, as more independent models are less likely to share common response biases. Model similarities and differences from site to site also cautions against assuming that performance of a given model at a limited number of sites is indicative of its likely performance at a new site. The high sensitivity of the models' response to climate variability demonstrates high sensitivity to location, representing different growing environments. Results suggest that there is little basis on which to categorize groups of models based upon expected commonalities in climate variability response, as these responses show high sensitivity to location rather than models imposing the same response to all sites.

We created subsets of models with the rule that only one model could be drawn from each cluster to test the hypothesis that diverse model combinations would more efficiently capture responses of the full ensemble than would a random combination of wheat models. However, performance of these subsets was not significantly different from the random subsets tested in Section 3.3 above. Selecting more diverse models via cluster analysis is therefore not an

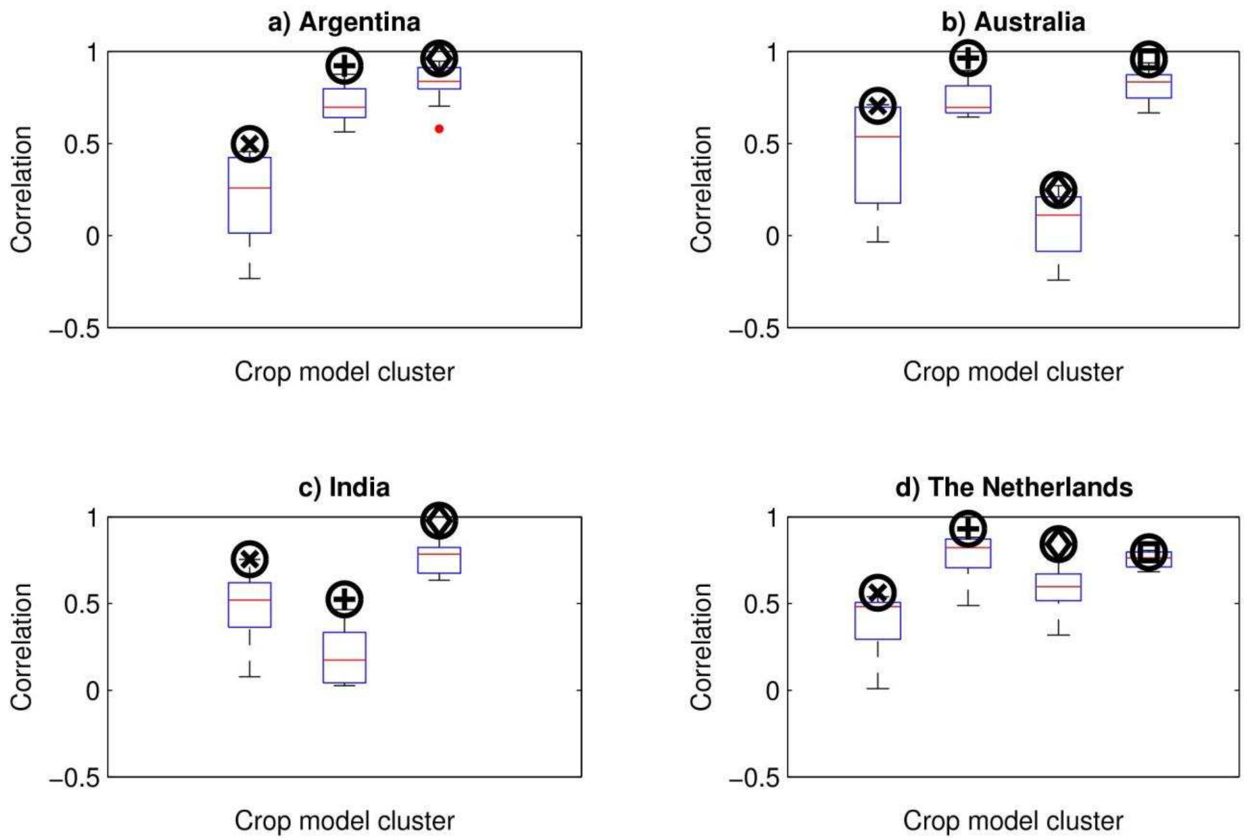

Figure 6: Correlations between simulated grain yield by the wheat models against the 27-member ensemble average series of interannual grain yields for single locations in (a) Argentina; (b) Australia; (c) India; and (d) the Netherlands. The correlations of the cluster ensembles are shown in the dark black symbol above the box-and-whiskers distribution of individual models within that cluster (corresponding to the symbols from Figure 5).

effective strategy for creating multi-model subsets for new studies, although the construction of subsets based upon model structure and parameter sets (rather than response characteristics) merits further study. Additional work may also explore agro-climatic responses in perturbed physics ensembles as an alternative to multi-model ensembles (PPEs and MMEs, respectively; Wallach et al., 2015).

3.6. Relationship between interannual and climatological temperature sensitivities

While the above analyses focused on the ways in which simulated grain yields are sensitive to interannual variability in temperature, rainfall, and solar radiation, the temperature sensitivity tests (-3˚C, +3˚C, +6˚C, and +9˚C) isolate the effect of mean changes in temperature. Popular impressions of climate change impacts are often based upon temporal proxies, or the assumption that an x-degree warmer mean climate at a given location would have grain yields similar to the yields observed in that location in past years when an x-degree anomaly occurred. Empirical models based upon historical regressions are often premised on such an assumption, although developed to a greater extent (e.g., Lobell and Burke, 2010). This is indeed a logical hypothesis as one would expect that a crop's response to mean warming would mimic its response to interannual temperature anomalies. Models that are most responsive to interannual temperature variability would therefore be expected to also be the most sensitive to mean temperature changes.

For example, consider two models: Model A (which simulates higher yields in warm years and thus whose response is positively correlated with interannual temperatures) and Model B (which simulates lower yields in warm years and thus whose response is negatively correlated with interannual temperatures). A temporal proxy assumption would anticipate that Model A would have more positive simulated yield changes (as a percentage of the historical simulations' yields) than Model B if both were exposed to warmer mean conditions. Likewise, if both models were simulated under cooler mean conditions Model A would have more negative yield changes than Model B. These comparisons between climate variability sensitivities and climate change responses are informative not only for the relationship of a single given model, but the pattern of the full ensemble provides a basis on which to evaluate model consistency and simple statistical modeling approaches.

The 27 wheat models' interannual temperature sensitivity and mean temperature change responses are compared for each of the temperature sensitivity tests and each of the four locations in **Figure 7**, with each dot representing a single wheat model. A model's position on the x-axis represents the correlation of its interannual yields against growing season temperature anomalies in the 1980-2010 period, and its position on the y-axis represents the percentage change in mean yield (over the 30 growing seasons) for each of the temperature sensitivity tests in comparison to the 1980-2010 mean yield (with CO_2 held at historical concentrations of 360 ppm). A linear fit is also drawn for each color-coded sensitivity test (quadratic fits were not substantially better).

As expected, the slopes of the linear fits indicate that models with greater interannual temperature sensitivity are more sensitive to mean temperature changes. The +3˚C, +6˚C, and +9˚C sensitivity tests' linear fits have a positive slope at all sites. This indicates that the mean warming tended to lead to relatively higher simulated grain yields in models with more positive correlations between interannual temperature and grain yield compared to models with more negative correlations (which had lower simulated grain yield in the sensitivity tests). Also as expected, the -3˚C sensitivity test's linear fit has a negative slope, as decreases in mean temperature lead to larger grain yield losses when models' interannual temperature anomalies are more positively correlated with yields compared to models.

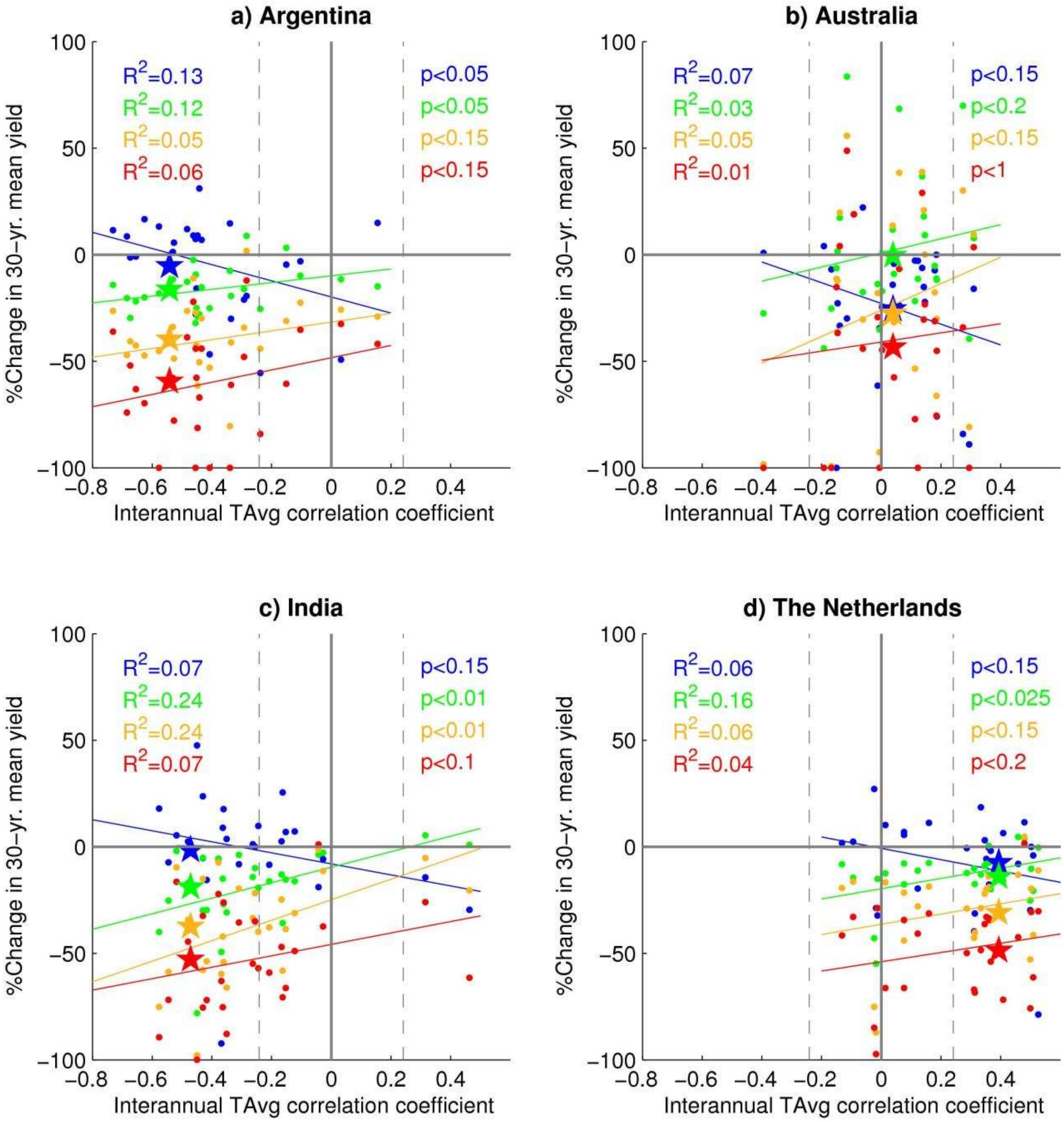

Figure 7: Comparison between each model's Pearson's correlation coefficient of interannual temperature and grain yield with its response to mean temperature change sensitivity tests (-3˚C, blue; +3˚C, green; +6˚C, orange; +9˚C, red; each compared against 1981-2010 historical period at single location in each country). Ensemble averages for each sensitivity test are represented by a star, and colored lines represent the least-squares linear fit for each sensitivity test with R^2 correlation and t-test significance documented for each fit. Vertical dashed lines indicate t-test significance at the 90th-percentile level for interannual correlations between average temperature and simulated

grain yield. The R^2 correlation and significance level for the fitted slope of each least-squares-fitted line is also presented in text of the corresponding color in each panel. p-levels presented for the slope were the lowest possible among 0.005, 0.01, 0.025, 0.05, 0.1, 0.15, 0.2, 0.25, and 1.

While the slopes of these lines support the use of temporal proxies for climate impact analyses, other aspects of the analysis cast serious doubt on the utility of the temporal proxy approach (even when CO_2 is held constant). Firstly, there is a dramatic spread among the 27 wheat models around the fitted line, with the sign of many models' mean temperature change responses opposite from what would be predicted by the interannual temperature response. As shown in Figure 7, R^2 correlations are quite low (between 0 and 0.24), with lowest values in the +9°C sensitivity test. Correlations are particularly low in Australia ($R^2 \leq 0.07$) where interannual temperature sensitivity was weak in most models, and are highest in the +3°C and +6°C sensitivity tests for India ($R^2 = 0.24$) where irrigation likely enabled a stronger temperature signal. t-test evaluations of the least-squares fit reveal many instances where the slopes are not statistically significant at the $p=0.05$ level, particularly in Australia and for the higher temperature change sensitivity tests (where only India is significant at the $p<0.1$ level). Together, these low correlations and the weak significance of fitted slopes suggest that the temporal proxy cannot be reliably applied, especially for conditions that are substantially warmer than the calibration period.

Secondly, a temporal proxy would predict that models with no sensitivity to interannual temperature variability would have no response to climate change (as represented by the temperature sensitivity tests), and therefore all linear fits should intersect at the origin of the axes. This is not the case as nearly all temperature sensitivity test lines fall below the origin with increasing distance as temperatures rise, suggesting that additional factors impart a mean grain yield reduction above what would be expected from examining the impacts of historical temperature variability. Several potential explanations for these differences merit further study.

A first candidate factor is that this simple temporal proxy based solely on temperature lends itself to biases as a result of interdependence of climate variables (Sheehy et al., 2006). For example, temperature anomalies may correlate with yield losses only because they coincide with dry seasons, which would suggest that a rainfall-based empirical model would be more appropriate. Interdependence of climate variables would somewhat explain the deviations of the wheat models around the least-squares fitted lines in Figure 7 as the interannual correlation would not be solely a temperature sensitivity. This factor cannot explain the extent of these deviations, however, nor is this explanation sufficient to explain the offset at the origin.

A second factor is the non-linearity in grain yield responses as mean climate change pushes systems beyond critical thresholds and tipping points, some of which may not have been present in the historical conditions. Within each temperature sensitivity test there are 30 years of climate variability including warm seasons with extreme events that are amplified by an increasing mean temperature and which may have a disproportionate impact on the mean yield shift. In combination, the mean warming and interannual extremes can produce conditions never experienced during the 1981-2010 period. In many cases this leads to a non-linear impact on grain yields beyond a simple extrapolation of interannual proxies (Porter and Semenov, 1995). For example, Lobell et al. (2012) found an acceleration of leaf senescence in Indian wheat during extreme heat events beyond what would have been expected from average temperatures alone. Interactions with other variables can also compound yield losses. Chief among these are increases in water stress during critical growth stages, as warmer temperatures lead to increased vapor pressure deficit and higher potential evapotranspiration (although accumulated water requirements may be partially counter-balanced by a shorter growing season). Non-linear effects could be identified if particular years in the sensitivity tests experienced much larger losses than the average year (compared to the historical climate). Thresholds and plant stresses at critical growth stages can also lead to complete loss of grain yields, as is clear in the number of models reporting 100% grain yield loss under the highest temperature conditions (Figure 7).

A third factor relates to different responses of grain yield to temperature variability and change during different parts of the crop growing season or during different parts of the year. This is probably particularly relevant for crops with a long growing period such as winter wheat in the Netherlands. An example of this is winter wheat in Denmark, where Kristensen et al. (2011) found a positive response of yield to increased temperature at low temperatures during winter, but a highly negative response during summer. Also Liu et al. (2013) found differential effects of warming on winter wheat yield in the North China Plain depending on whether the warming mainly affected winter or summer conditions. The effects of warming for crops that have long growing seasons with large seasonal differences may therefore be obscured by positive effects of warming in some parts of the growing season and negative ones in other parts of the growth period.

A final candidate factor for the differences between interannual temperature variability and mean warming is the extent of within-season climate variability. In the historical record extremely warm seasons tend to be only marginally warm on the average day but feature a substantial heat wave (or several), which has a fundamentally different effect on plant function from that of a season where a slight warming is relentless (even if the average temperature is the same). With prolonged warming maturation is accelerated and yields may be reduced as a result of lower net radiation interception. There is also an increased chance that warm temperatures will negatively affect key phenological stages and/or interact with precipitation or solar radiation to create evaporative demand that the plants cannot meet. These alterations to phenological development and/or heat and water stresses can have cascading effects on plant growth throughout the season with net yield reductions on average compared to the historical temperature variability. The models respond to high temperatures according to a large variety of parameterizations (Alderman et al., 2013), with responses to extreme heat an area in particular need of development (Lobell et al., 2012).

4. Conclusions and next steps

Analysis of the 27 models participating in the AgMIP Wheat Model Intercomparison Pilot reveals substantial differences in the ways that models respond to interannual variations in rainfall, temperature, and solar radiation at four diverse locations. These differences provide useful context to differences in the abilities of the same models to reproduce detailed field observations (Martre et al., 2015) and climate change responses (Asseng et al., 2013, 2015). The large differences apparent in interannual climate sensitivity suggest that multiple years of consistent field trials are desirable to enable proper initialization of field conditions, and field experiments during extreme conditions would benefit the calibration of crop models for both mean yields and interannual variability. Such long-term agricultural research datasets are rare, unfortunately, so in typical applications such as those done here it is likely that any biases in calibration are amplified when a single-year's calibration is used for multiple seasons. It is therefore useful to take advantage of the tendency of multi-model ensemble statistics to reduce overall errors beyond the calibration period.

The AgMIP Wheat Pilot offers a far larger multi-model sample than would be expected in the applications for which each of the participating models was designed; however several of the interannual response results help guide the formation of practical subsets and application protocols. Although calibration information has been shown to reduce errors in mean yields and details in crop growth (Asseng et al., 2013), the results presented here suggest that interannual yield variability for most models is not strongly affected by the availability of more detailed field observations (e.g., evapotranspiration, biomass, leaf-area index, plant available soil moisture) for calibration. This is encouraging as high-information field trials are much less common. Adding a second (and third) wheat model dramatically increases the likelihood that the simulated results will reproduce the interannual behavior of the full 27-model ensemble, with a diminishing benefit to efforts that utilize additional models beyond that. This information is directly relevant to the design of new studies looking to take advantage of multi-model ensemble statistics despite resource constraints, including AgMIP efforts to form crop modeling tools that may link with global agricultural monitoring and outlooks on a sub-seasonal to seasonal scale (Singh et al., 2012; Vitart et al., 2012). Use of an ensemble also highlights the sensitivity of simulated yields to interannual climate variability as common features rise above the ensemble's diminished noise more easily than the individual models' larger noise.

The wheat models demonstrate several common patterns of climate variability response at each tested location. In some cases there is a fundamental disagreement between models about whether grain yield responds positively or negatively to a given anomaly, although interdependence of climate variables (e.g., wet and cool years vs. hot and dry years) muddles the picture. Even when two models respond in a very similar manner at one location, differences in calibration method and quality, parameters, model structure, and environmental conditions can lead to strong deviations in model response at other sites. These results therefore suggest that there are still strong differences in wheat models' climate sensitivities, and that further work is needed to create models that are truly applicable across a wide range of current and future conditions. The analysis presented here focuses on mean growing season climate anomalies at four locations; however consideration of intra-seasonal variability and extremes (e.g., heat waves, dry spells, frosts, floods, waterlogging, monsoon dynamics) require further study. Comparing multi-model simulation experiments against long-term field trials (e.g., Dobermann et al., 2000) would also be desirable in order to provide true observations upon which to evaluate simulated outputs (rather than assuming the value of the ensemble average as done here).

The effects of interannual temperature variability and mean climate warming were shown to be only weakly related among the 27 wheat models, indicating that a temporal proxy for climate change is likely oversimplified. State-of-the-art empirical models use far more than interannual temperature for climate impacts projection, however these findings underscore the importance of considering complex interactions between variables and non-linear responses that may not be present in the historical period datasets to which models are fit. Further work is needed to elucidate additional physiological factors that differentiate the effects of a warm season from those of a warmer climate (Porter and Semenov et al., 2005).

Follow-on phases of the AgMIP Wheat Pilot are focusing on more sites and experiments designed to better distinguish between heat waves and warmer mean climate conditions. The analyses presented here would also be of interest for other completed AgMIP Crop Model Pilots (e.g., for maize, Bassu et al., 2014; rice, Li et al., 2014; and sugarcane, Singels et al., 2013) as well as pilots planned for millet and sorghum, potato, canola, and grasslands. AgMIP's Coordinated Climate-Crop Modeling Project (C3MP; Ruane et al., 2014; McDermid et al., 2015) and Global Gridded Crop Model Intercomparison (GGCMI; Rosenzweig et al., 2014; Elliott et al., 2015), as well as the impact response surface studies conducted in FACCE MACSUR (Pirttioja et al., 2015) provide additional fora in which to compare climate sensitivities across multiple locations and crop models, assuming that observational yield data also are available for those points or aggregated grid cells. This study's yield response analyses are currently being applied to GGCMI's historical period intercomparison, helping to determine the causes for differences in interannual yield variation for more than a dozen models with global coverage of multiple crops (Elliott et al., 2015). Wheat model development would benefit from a future intercomparison centered upon a region where long-term variety trials overlap with similar detailed field experiments so that calibration and the response to interannual climate variability may be more comprehensively evaluated. Of particular interest would be the way in which interannual yield observations affect calibration and the resulting climate variability and climate change sensitivities.

Results from this study underscore the need for model intercomparison results to avoid anonymity in order to enable careful analysis of structural and parameter differences that cause differences in yield response. Current and future phases of the AgMIP Wheat intercomparisons no longer hold the models anonymous, and evaluation of the mechanisms driving different climate responses is a crucial line of continuing inquiry (as was performed for the AgMIP Rice Pilot; Li et al., 2015). Through these activities the efforts of the AgMIP Wheat Pilot will better accomplish integrated assessments of climate impact on the agricultural sector.

Acknowledgements

We acknowledge the efforts of AgMIP Leaders Cynthia Rosenzweig, Jim Jones, John Antle, and Jerry Hatfield in their efforts to initiate the AgMIP Wheat Pilot and encourage explorations such as these. Support for many European participants in the AgMIP Wheat Pilot was provided by the Modelling European Agriculture with Climate Change for Food Security (MACSUR) knowledge hub within the Joint Research Programming Initiative on Agriculture, Food Security and Climate Change (FACCE-JPI). Alex Ruane's contributions were supported by the NASA Earth Sciences Research Program (#281945.02.03.03.96) and the NASA Modeling, Analysis, and Prediction Program (#509496.02.08.04.24). Participation of Claudio Stöckle and CropSyst simulations were supported by the project Regional Approaches to Climate Change for Pacific Northwest Agriculture (REACCH-PNA), funded through award #2011-68002-30191 from the US National Institute for Food and Agriculture.

References

Aggarwal, P., B. Banerjee, M. Daryaei, A. Bhatia, A. Bala, S. Rani, S. Chander, H. Pathak, and N. Kalra, 2006: InfoCrop: A dynamic simulation model for the assessment of crop yields, losses due to pests, and environmental impact of agro-ecosystems in tropical environments. II. Performance of the model. Agricultural Systems 89: 47-67. doi: 10.1016/j.agsy.2005.08.003|10.1016/j.agsy.2005.08.003.

Alderman, P.D., E. Quilligan, S. Asseng, F. Ewert, and M.P. Reynolds, 2013: Proceedings of the Workshop: Modeling Wheat Response to High Temperature. CIMMYT, El Batan, Texcoco, Mexico. 19-21 June, 2013. 138pp.

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

Arnold, G.W., de Wit, C.T., 1976. Critical evaluation of system analysis in ecosystem management. Simulation Monographs, Wageningen, The Netherlands.

Asseng, S., H. van Keulen, and W. Stol, 2000: Performance and application of the APSIM Nwheat model in the Netherlands. European Journal of Agronomy 12: 37-54.

Asseng, S., P.D. Jamieson, B. Kimball, P. Pinter, K. Sayre, J.W. Bowden, and S.M. Howden, 2004: Simulated wheat growth affected by rising temperature, increased water deficit and elevated atmospheric CO2. Field Crops Research 85: 85-102.

Asseng, S., B.A. Keating, I.R.P. Fillery, P.J. Gregory, J.W. Bowden, N.C. Turner, J.A. Palta, and D.G. Abrecht, 1998: Performance of the APSIM-wheat model in Western Australia. Field Crops Research 57: 163-179.

Asseng, S., H. van Keulen, and W. Stol, 2000: Performance and application of the APSIM Nwheat model in the Netherlands. European Journal of Agronomy 12: 37-54.

Asseng, S., F. Ewert, C. Rosenzweig, J.W. Jones, J.L. Hatfield, A.C. Ruane, K.J. Boote, P. Thorburn, R.P. Rötter, D. Cammarano, N. Brisson, B. Basso, P. Martre, P.K. Aggarwal, C. Angulo, P. Bertuzzi, C. Biernath, A.J. Challinor, J. Doltra, S. Gayler, R. Goldberg, R. Grant, L. Heng, J. Hooker, L.A. Hunt, J. Ingwersen, R.C. Izaurralde, K.C. Kersebaum, C. Müller, S. Naresh Kumar, C. Nendel, G. O'Leary, J.E. Olesen, T. M. Osborne, T. Palosuo, E. Priesack, D. Ripoche, M.A. Semenov, I. Shcherbak, P. Steduto, C. Stöckle, P. Stratonovitch, T. Streck, I. Supit, F. Tao, M. Travasso, K. Waha, D. Wallach, J.W. White, J.R. Williams and J. Wolf, 2013: Uncertainty in simulating wheat yields under climate change. Nature Clim. Change **3**, 827-832. doi: 10.1038/NCLIMATE1916

Asseng, S., F. Ewert, P. Martre, R.P. Rötter, D.B. Lobell, D. Cammarano, B.A. Kimball, M.J. Ottman, G.W. Wall, J.W. White, M.P. Reynolds, P.D. Alderman, P.V. V. Prasad, P.K. Aggarwal, J. Anothai, B. Basso, C. Biernath, A.J. Challinor, G. De Sanctis, J. Doltra, E. Fereres, M. Garcia-Vila, S. Gayler, G. Hoogenboom, L.A. Hunt, R.C. Izaurralde, M. Jabloun, C.D. Jones, K.C. Kersebaum, A-K. Koehler, C. Müller, S. Naresh Kumar, C. Nendel, G. O'Leary, J.E. Olesen, T. Palosuo, E. Priesack, E. Eyshi Rezaei, A.C. Ruane, M.A. Semenov, I. Shcherbak, C. Stockle, P. Stratonovitch, T. Streck, I. Supit, F. Tao, P.J. Thorburn, K. Waha, E. Wang, D. Wallach, J. Wolf, Z. Zhao, and Y. Zhu, 2015: Rising temperatures reduce global wheat production. Nature Clim. Change, 5(2), 143-147, doi:10.1038/nclimate2470.

Basso B, Cammarano D, Troccoli A, Chen D, Ritchie J (2010) Long-term wheat response to nitrogen in a rainfed Mediterranean environment: Field data and simulation analysis. European Journal of Agronomy 33: 132-138. doi: DOI 10.1016/j.eja.2010.04.004.

Bassu S., N. Brisson, J-L. Durand, K.J. Boote, J. Lizaso, J.W. Jones, C. Rosenzweig, A.C. Ruane, M. Adam, C. Baron, B. Basso, C. Biernath, H. Boogaard, S. Conijn, M. Corbeels, D. Deryng, G. De Sanctis, S. Gayler, P. Grassini, J. Hatfield, S. Hoek, C. Izaurralde, R. Jongschaap, A. Kemanian, K.C. Kersebaum, D. Makowski, C. Müller, C. Nendel, E. Priesack, M.V. Pravia, F. Sau, I. Shcherbak, F. Tao, E. Teixeira, D. Timlin, and K. Waha, 2014: Do various maize crop models give the same responses to climate change factors? Global Change Biology **20**, 2301–2320, doi: 10.1111/gcb.12520

Berntsen, J., B. Petersen, B. Jacobsen, J. Olesen, and N. Hutchings, 2003: Evaluating nitrogen taxation scenarios using the dynamic whole farm simulation model FASSET. Agricultural Systems 76: 817-839. doi: 10.1016/S0308-521X(02)00111-7|10.1016/S0308-521X(02)00111-7.

Biernath, C., S. Gayler, S. Bittner, C. Klein, P. Hogy, A. Fangmeier, and E. Priesack, 2011: Evaluating the ability of four crop models to predict different environmental impacts on spring wheat grown in open-top chambers. European Journal of Agronomy 35: 71-82. doi: DOI 10.1016/j.eja.2011.04.001.

Bondeau, A., P.C. Smith, S. Zaehle, S. Schaphoff, W. Lucht, W. Cramer, and D. Gerten, 2007: Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biology 13: 679-706.

Boogaard, H.L., C.A. Van Diepen, R.P. Rötter, J.C.M.A. Cabrera, and H.H. Van Laar (eds), 1998: User's guide for the WOFOST 7.1 crop growth simulation model and WOFOST control center 1.5. Winand Staring Centre, Wageningen, The Netherlands, 144pp.

Boote, K.J., C.H. Porter, J.W. Jones, P.J. Thorburn, K.C. Kersebaum, G. Hoogenboom, J.W. White, and J.L. Hatfield, 2015: Sentinel Site Data for Model Improvement – Definition and Characterization. In: Hatfield, J.L., Fleisher, D. (Eds.), Improving Modeling Tools to Assess Climate Change Effects on Crop Response, Advances in Agricultural Systems Modeling. ASA, CSSA, and SSSA, Madison, WI, USA. doi:10.2134/advagricsystmodel7.2014.0019

Brisson, N., C. Gary, E. Justes, R. Roche, B. Mary, D. Ripoche, D. Zimmer, J. Sierra, P. Bertuzzi, P. Burger, F. Bussiere, Y. Cabidoche, P. Cellier, P. Debaeke, J. Gaudillere, C. Henault, F. Maraux, B. Seguin, and H. Sinoquet, 2003: An overview of the crop model STICS. European Journal of Agronomy 18: 309-332. doi: PII S1161-0301(02)00110-7.

Brisson, N., B. Mary, D. Ripoche, M. Jeuffroy, F. Ruget, B. Nicoullaud, P. Gate, F. Devienne-Barret, R. Antonioletti, C. Durr, G. Richard, N. Beaudoin, S. Recous, X. Tayot, D. Plenet, P. Cellier, J. Machet, J. Meynard, and R. Delecolle, 1998: STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. Agronomie 18: 311-346.

Challinor, A., T. Wheeler, P. Craufurd, J. Slingo, and D. Grimes, 2004: Design and optimisation of a large-area process-based model for annual crops. Agricultural and Forest Meteorology 124: 99-120. doi: DOI 10.1016/j.agrformet.2004.01.002.

Challinor, A.J., and Coauthors, 2014a: A meta-analysis of crop yield under climate change and adaptation. Nature Climate Change 4(4), 287-291. doi:10.1038/NCLIMATE2153

Challinor, A.J., P. Martre, S. Asseng, P. Thornton, and F. Ewert, 2014b: Making the most of climate impact ensembles. Nature Climate Change 4(2), 77-80.

Dobermann, A., D. Dawe, R.P. Rötter, and K.G. Cassman, 2000: Reversal of rice yield decline in a long-term continuous cropping experiment. Agronomy Journal, 92, 633-643.

Donatelli, M., and Co-authors, 2010: A Component-Based Framework for Simulating Agricultural Production and Externalities. Environmental and Agricultural Modelling: Integrated Approaches For Policy Impact Assessment: 63-108. doi: 10.1007/978-90-481-3619-3_4|10.1007/978-90-481-3619-3_4.

Elliott, J., C. Müller, D. Deryng, J. Chryssanthacopoulos, K.J. Boote, M. Büchner, I. Foster, M. Glotter, J. Heinke, T. Iizumi, R.C. Izaurralde, N.D. Mueller, D.K. Ray, C. Rosenzweig, A.C. Ruane, and J. Sheffield, 2015: The Global Gridded Crop Model Intercomparison: Data and modeling protocols for Phase 1 (v1.0). Geosci. Model Dev., 8, 261-277, doi:10.5194/gmd-8-261-2015.

Evans, L.T., and R.A. Fischer, 1999: Yield potential: Its definition, measurement, and significance. Crop Science, 39, 1544-1551.

Ewert, F., C. Angulo, C. Rumbaur, R. Lock, A. Enders, M. Andenäuer, T. Heckelei, M. van Ittersum, J. Wolf, R. Rötter, 2011a: Assessing the adaptive capacity of agriculture in the Netherlands to the impacts of climate change under different market and policy scenarios. AgriAdapt Project Reports no 2 & 3 Dutch Research Program Climate Change and Spatial Planning: 49.

Ewert F., M.K. van Ittersum, T. Heckelei, O. Therond, I. Bezlepkina, and E. Andersen, 2011b: Scale changes and model linking methods for integrated assessment of agri-environmental systems. Agriculture, Ecosystems and Environment **142**, 6-17.

Ewert, F., R.P. Rötter, M. Bindi, H. Webber, M. Trnka, K-C. Kersebaum, J.E. Olesen, M.K. van Ittersum, S. Janssen, M. Rivington, M. Semenov, D. Wallach, J.R. Porter, D. Stewart, J. Verhagen, T. Gaiser, T. Palosuo, F. Tao, C. Nendel, P-P. Roggero, L. Bartosova, and S. Asseng, 2015: Crop modelling for integrated assessment of risk to food production from climate change. Environmental Modelling & Software, 72, 287-303. doi:10.1016/j.envsoft.2014.12.003

Fader, M., S. Rost, C. Muller, A. Bondeau, and D. Gerten, 2010: Virtual water content of temperate cereals and maize: Present and potential future patterns. Journal of Hydrology 384: 218-231. doi: 10.1016/j.jhydro1.2009.12.011.

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

Goudriaan, J., H.H. Van Laar (eds), 1994: Modelling Potential Crop Growth Processes. Textbook With Exercises. Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 238.

Grant, R., B. Kimball, M. Conley, J. White, G. Wall, and M. Ottman, 2011: Controlled Warming Effects on Wheat Growth and Yield: Field Measurements and Modeling. Agronomy Journal 103: 1742-1754. doi: 10.2134/agronj2011.0158.

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

He, J., P. Stratonovitch, V. Allard, M.A. Semenov, and P. Martre, 2010: Global Sensitivity Analysis of the Process-Based Wheat Simulation Model SiriusQuality1 Identifies Key Genotypic Parameters and Unravels Parameters Interactions. Procedia - Social and Behavioral Sciences 2: 7676-7677.

Hoogenboom, G., and J. White, 2003: Improving physiological assumptions of simulation models by using gene-based approaches. Agronomy Journal 95: 82-89.

Hunt, L.A., and S. Pararajasingham, 1995: CROPSIM-wheat - a model describing the growth and development of wheat. Canadian Journal of Plant Science 75: 619-632.

Jamieson, P., and M. Semenov, 2000: Modelling nitrogen uptake and redistribution in wheat. Field Crops Research 68: 21-29. doi: 10.1016/S0378-4290(00)00103-9.

Jamieson, P., M. Semenov, I. Brooking, and G. Francis, 1998: Sirius: a mechanistic model of wheat response to environmental variation. European Journal of Agronomy 8: 161-179. doi: 10.1016/S1161-0301(98)00020-3.

Jones, J., G. Hoogenboom, C. Porter, K. Boote, W. Batchelor, L. Hunt, P. Wilkens, U. Singh, A. Gijsman, and J. Ritchie, 2003: The DSSAT cropping system model. European Journal of Agronomy 18: 235-265. doi: PII S1161-0301(02)00107-7.

Keating, B.A., and Co-Authors, 2003: An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy 18: 267-288.

Kersebaum, K., 1995: Application of a simple management model to simulate water and nitrogen dynamics. Ecological Modelling 81: 145-156.

Kersebaum, K., 2007: Modelling nitrogen dynamics in soil-crop systems with HERMES. Nutrient Cycling in Agroecosystems 77: 39-52. doi: DOI 10.1007/s10705-006-9044-8.

Kersebaum, K.C., 2011: Special features of the HERMES model and additional procedures for parameterization, calibration, validation, and applications. Ahuja, LR and Ma, L (eds) Methods of introducing system models into agricultural research Advances in Agricultural Systems Modeling Series 2, Madison (ASA-CSSA-SSSA): 65-94.

Kersebaum, K.C., and A.J. Beblik, 2001: Performance of nitrogen dynamics model applied to evaluate agricultural management practices. Lewis Publishers.

Kersebaum, K.C., K.J. Boote, J.S. Jorgenson, C. Nendel, M. Bindi, C. Frühauf, T. Gaiser, G. Hoogenboom, C. Kollas, J.E. Olesen, R.P. Rötter, F. Ruget, P. Thorburn, M. Trnka, M. Wegehenkel, 2015: Analysis and classification of data sets for calibration and validation of agro-ecosystem models. Environmental Modelling & Software, 72, 402-417. DOI: 10.1016/j.envsoft.2015.05.009.

Kiniry, J., D. Major, R. Izaurralde, J. Williams, P. Gassman, M. Morrison, R. Bergentine, and R. Zentner, 1995: EPIC model parameters for cereal, oilseed, and forage crops in the northern great-plains region. Canadian Journal of Plant Science 75: 679-688.

Kollas, C., K.-C. Kersebaum, C. Nendel, K. Manevski, C. Müller, T. Palosuo, C.M. Armas-Herrera, N. Beaudoin, M. Bindi, M. Charfeddine, T. Conradt, J. Constantin, J. Eitzinger, F. Ewert, R. Ferrise, T. Gaiser, I. Garcia de Cortazar-Atauri, L. Giglio, P. Hlavinka, H. Hoffmann, M.P. Hoffmann, M. Launay, R. Manderscheid, B. Mary, W. Mirschel, M. Moriondo, J.E. Olesen, I. Öztürk, A. Pacholski, D. Ripoche-Wachter, P.P. Roggero, S. Roncossek, R.P. Rötter, F. Ruget, B. Sharif, M. Trnka, D. Ventrella, K. Waha, M. Wegehenkel, H-J. Weigel, and L. Wu, 2015: Crop rotation modelling – a European model intercomparison. Eur. J. Agron. 70, 98-111. DOI: 10.1016/j.eja.2015.06.007

Kristensen, K., K. Schelde, and J.E. Olesen, 2011: Winter wheat yield response to climate variability in Denmark. Journal of Agricultural Science 149: 33-47.

Latta, J., and G. O'Leary, 2003: Long-term comparison of rotation and fallow tillage systems of wheat in Australia. Field Crops Research 83: 173-190. doi: 10.1016/S0378-4290(03)00073-X|10.1016/S0378-4290(03)00073-X.

Lawless, C., M. Semenov, and P. Jamieson, 2005: A wheat canopy model linking leaf area and phenology. European Journal of Agronomy 22: 19-32. doi: 10.1016/j.eja.2003.11.004.

Li, S., T. Wheeler, A. Challinor, E. Lin, Y. Xu, and H. Ju, 2010: Simulating the Impacts of Global Warming on Wheat in China Using a Large Area Crop Model. Acta Meteorologica Sinica 24: 123-135.

Li, T., T. Hasegawa, X. Yin, Y. Zhu, K. Boote, M. Adam, S. Bregaglio, S. Buis, R. Confalonieri, T. Fumoto, D. Gaydon, M. Marcaida, III, H. Nakagawa, P. Oriol, A.C. Ruane, F. Ruget, B. Singh, U. Singh, L. Tang, F. Tao, P. Wilkens, H. Yoshida, Z. Zhang, and B. Bouman, 2015: Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob. Change Biol., 21, no. 3, 1328-1341, doi:10.1111/gcb.12758.

Liu, L., C. Hu, J.E. Olesen, Z. Ju, P. Yang, Y. Zhang, 2013. Warming and nitrogen fertilization effects on winter wheat yields in northern China varied between four years. Field Crops Research 151: 56-64.

Lobell, D.B., and M.B. Burke, 2010: On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology 150, 1443–1452.

Lobell, D.B., A. Sibley, and J.I. Ortiz-Monasterio, 2012: Extreme heat effects on wheat senescence in India. Nature Climate Change 2, 186-189. doi:10.1038/NCLIMATE1356

Martre, P., P.D. Jamieson, M.A. Semenov, R.F. Zyskowski, J.R. Porter, and E. Triboi, 2006: Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. Eur J Agron 25: 138-154.

Martre, P., and Coauthors, 2014: Error of multimodel ensembles of wheat growth: more models are better than one. Global Change Biology, 21, 911–925, doi: 10.1111/gcb.12768.

McDermid, S.P., A.C. Ruane , C. Rosenzweig, N.I. Hudson, M.D. Morales, P. Agalawatte, S. Ahmad, L.R. Ahuja, I. Amien, S.S. Anapalli, J. Anothai, S. Asseng, J. Biggs, F. Bert, P. Bertuzzi, V. Bhatia, M. Bindi, I. Broad, D. Cammarano, R. Carretero, A.A. Chattha, U. Chung, S. Debats, P. Deligios, G. De Sanctis, T. Dhliwayo, B. Dumont, L. Estes, F. Ewert, R. Ferrise, T. Gaiser, G. Garcia, S. Gbegbelegbe, V. Geethalakshmi, E. Gerardeaux, R. Goldberg, B. Grant, E. Guevara, J. Hickman, H. Hoffmann, H. Huang, J. Hussain, F.B. Justino, A.S. Karunaratne, A.-K. Koehler, P.K. Kouakou, S.N. Kumar, A. Lakshmanan, M. Lieffering, X. Lin, Q. Luo, G. Magrin, M. Mancini, F.R. Marin, A.D. Marta, Y. Masutomi, T. Mavromatis, G. McLean, S. Meira, M. Mohanty, M. Moriondo, W. Nasim, N. Negm, F. Orlando, S. Orlandini, I. Ozturk, H.M.S. Pinto, G. Podesta, Z. Qi, J. Ramarohetra, H.H. Rahman, H. Raynal, G. Rodriguez, R. Rötter, V. Sharda, L. Shuo, W. Smith, V. Snow, A. Soltani, K. Srinivas, B. Sultan, D.K. Swain, F. Tao, K. Tesfaye, M.I. Travasso, G. Trombi, A. Topaj, E. Vanuytrecht, F.E. Viscarra, S.A. Wajid, E. Wang, H. Wang, J. Wang, E. Wijekoon, L. Byun-Woo, Y. Xiaoguang, B.H. Young, J.I. Yun, Z. Zhao, and L. Zubair, 2015: The AgMIP Coordinated Climate-Crop Modeling Project (C3MP): Methods and protocols. In Handbook of Climate Change and Agroecosystems: The Agricultural Model Intercomparison and Improvement Project (AgMIP). C. Rosenzweig, and D. Hillel, Eds., ICP Series on Climate Change Impacts, Adaptation, and Mitigation Vol. 3. Imperial College Press, 191-220, doi:10.1142/9781783265640_0008.

Nendel, C., M. Berg, K. Kersebaum, W. Mirschel, X. Specka, M. Wegehenkel, K. Wenkel, and R. Wieland, 2011: The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. Ecological Modelling 222: 1614-1625. doi: DOI 10.1016/j.ecolmodel.2011.02.018.

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

O'Leary, G., and D. Connor, 1996b: A simulation model of the wheat crop in response to water and nitrogen supply .2. Model validation. Agricultural Systems 52: 31-55. doi: 10.1016/0308-521X(96)00002-9.

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

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

Palosuo, T., and co-authors, 2011: Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. Europ. J. Agronomy 35, 103– 114. doi:10.1016/j.eja.2011.05.001

Pirttioja, N., T.R. Carter, S. Fronzek, M. Bindi, H. Hoffmann, T. Palosuo, M. Ruiz-Ramos, F. Tao, M. Trnka, M. Acutis, S. Asseng, P. Baranowski, B. Basso, P. Bodin, S. Buis, D. Cammarano, P. Deligios, M.F. Destain, B. Dumont, F. Ewert, R. Ferrise, L. François, T. Gaiser, P. Hlavinka, I. Jacquemin, K.C. Kersebaum, C. Kollas, J. Krzyszczak, I.J. Lorite, J. Minet, MI. Minguez, M. Montesino, M. Moriondo, C. Müller, C. Nendel, I. Öztürk, A. Perego, A. Rodríguez A.C., Ruane, F. Ruget, M. Sanna, M.A. Semenov, C. Slawinski, P. Stratonovitch, I. Supit, K. Waha, E. Wang, L. Wu, Z. Zhao, and R.P. Rötter, 2015: A crop model ensemble analysis of temperature and precipitation effects on wheat yield across a European transect using impact response surfaces. Climate Research 65: 87–105. doi:10.3354/cr01322

Porter, J.R., and M.A. Semenov. 1995: Non-linearity in climate change impact assessments. Journal of Biogeography, 22(4/5): 597-600.

Porter, J.R., and M.A. Semenov. 2005: Crop responses to climatic variation. Philosophical Transactions of the Royal Society of London B Biological Sciences 360: 2021-2035.

Porter, J.R., L. Xie, and Coauthors, 2014: Food Security and Food Production Systems. Chapter 7 of the Working Group 2 contribution to the Intergovernmental Panel on Climate Change Fifth Assessment Report. 82pp.

Priesack, E., S. Gayler, and H. Hartmann, 2006: The impact of crop growth sub-model choice on simulated water and nitrogen balances. Nutrient Cycling in Agroecosystems 75: 1-13. doi: DOI 10.1007/s107500-006-9006-1.

Ray, D.K., J.S. Gerber, G.K. MacDonald, and P.C. West, 2015: Climate variation explains a third of global crop yield variability, Nat Commun, 6, doi: 10.1038/ncomms6989.

Rienecker, M.M., M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich, S.D. Schubert, L. Takacs, G.K. Kim, S. Bloom, J.Y. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R.D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C.R. Redder, R. Reichle, F.R. Robertson, A.G. Ruddick, M. Sienkiewicz, and J. Woollen, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. J. Climate, 24 (14), 3624–3648, doi: 10.1175/JCLI-D-11-00015.1.

Ritchie, J.T., D.C. Godwin, and S. Otter-Nacke, 1985: CERES-wheat: A user-oriented wheat yield model. Preliminary documentation.

Ritchie, S., H. Nguyen, and A. Holaday, 1987: Genetic diversity in photosynthesis and water-use efficiency of wheat and wheat relatives. Journal of Cellular Biochemistry: 43-43.

Rosenzweig, C., J.W. Jones, J.L. Hatfield, J.M. Antle, A.C. Ruane, and C.Z. Mutter, 2015: The Agricultural Model Intercomparison and Improvement Project: Phase I activities by a global community of science. In Handbook of Climate Change and Agroecosystems: The Agricultural Model Intercomparison and Improvement Project (AgMIP). C. Rosenzweig, and D. Hillel, Eds., ICP Series on Climate Change Impacts, Adaptation, and Mitigation Vol. 3. Imperial College Press, 3-24, doi:10.1142/9781783265640_0001.

Rosenzweig, C., J.W. Jones, J.L. Hatfield, A.C. Ruane, K.J. Boote, P. Thorburn, J.M. Antle, G.C. Nelson, C. Porter, S. Janssen, S. Asseng, B. Basso, F. Ewert, D. Wallach, G. Baigorria, J.M Winter. 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agric. Forest Meteorol. 170, 166-182. http://dx.doi.org/10.1016/j.agrformet.2012.09.011

Rosenzweig, C., J. Elliott, D. Deryng, A.C. Ruane, C. Müller, et al., 2014: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proc. Natl. Acad. Sci., 111, 3268-3273, doi:10.1073/pnas.1222463110.

Rötter, R.P., T.R. Carter, J.E. Olesen, and J.R. Porter, 2011: Crop-climate models need an overhaul. Nature Climate Change, 1(4): 175-177.

Rötter, R.P., T. Palosuo, K.-C. Kersebaum, C. Angulo, M. Bindi, F. Ewert, R. Ferrise, P. Hlavinka, M. Moriondo, J.E. Olesen, J. Takáč, and M. Trnka: 2012. Simulation of spring barley yield in different climatic zones of Northern and Central Europe: A comparison of nine crop growth models. Field Crops Research, 133, 23-36.

Ruane, A.C., S. McDermid, C. Rosenzweig, G.A. Baigorria, J.W. Jones, C.C. Romero, and L.D. Cecil, 2014: Carbon–Temperature–Water change analysis for peanut production under climate change: a prototype for the AgMIP Coordinated Climate-Crop Modeling Project (C3MP). Global Change Biology, **20**, 394–407, doi: 10.1111/gcb.12412

Semenov, M.A., and P.R. Shewry, 2011: Modelling predicts that heat stress, not drought, will increase vulnerability of wheat in Europe. Nature Scientific Reports 1:66, doi: 10.1038/srep00066

Senthilkumar, K., P. Bindraban, W. de Boer, N. de Ridder, T. Thiyagarajan, and K. Giller, 2009: Characterising rice-based farming systems to identify opportunities for adopting water efficient cultivation methods in Tamil Nadu, India. Agricultural Water Management 96: 1851-1860. doi: 10.1016/j.agwat.2009.08.007.

Shibu, M., P. Leffelaar, H. van Keulen, and P. Aggarwal, 2010: LINTUL3, a simulation model for nitrogen-limited situations: Application to rice. European Journal of Agronomy 32: 255-271. doi: DOI 10.1016/j.eja.2010.01.003.

Singels, A., M. Jones, F. Marin, A.C. Ruane, and P. Thorburn, 2013: Predicting climate change impacts on sugarcane production at sites in Australia, Brazil and South Africa using the Canegro model. Sugar Tech, doi: 10.1007/s12355-013-0274-1.

Singh, P.J., C. Justice, J. Soares, O. Leo, P. Kosuth, I. Jarvis, D. Williams, W. Bingfang, J. Latham, and I. Becker-Reshef, 2012: GEO-GLAM: A GEOSS-G20 initiative on Global Agricultural Monitoring. In 39th COSPAR Scientific Assembly. Held 14-22 July 2012, in Mysore, India. Abstract D2.2-9-12, Vol. 39, p. 1451. Retrieved from http://adsabs.harvard.edu/

Spitters, C.J.T., and A.H.C.M. Schapendonk, 1990: Evaluation of breeding strategies for drought tolerance in potato by means of crop growth simulation. Plant and Soil 123: 193-203.

Stackhouse, Jr., P.W., S.K. Gupta, S.J. Cox, J.C. Mikovitz, T. Zhang, and L.M. Hinkelman, 2011: The NASA/GEWEX Surface Radiation Budget Release 3.0: 24.5-Year Dataset. GEWEX News, 21, No. 1, February, 10-12.

Steduto P., T. Hsiao, D. Raes, and E. Fereres, 2009: AquaCrop-The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles. Agronomy Journal 101: 426-437. doi: DOI 10.2134/agronj2008.0139s.

Stenger, R., E. Priesack, G. Barkle, and C. Sperr, 1999: Expert-N A tool for simulating nitrogen and carbon dynamics in the soil-plant-atmoshpere system. Land Treatment collective proceedings Technical Session, New Zealand.

Stockle, C., M. Donatelli, and R> Nelson, 2003: CropSyst, a cropping systems simulation model. European Journal of Agronomy 18: 289-307. doi: PII S1161-0301(02)00109-0.

Supit, I., A.A. Hooijer, and C.A. van Diepen, 1994: System Description of the WOFOST 6.0 Crop Simulation Model Implemented in CGMS, Volume 1: Theory and Algorithms.

EUR 15956 EN, Joint Research Center, Commission of the European Communities, Luxembourg.

Tao F., and Z. Zhang, 2010: Adaptation of maize production to climate change in North China Plain: Quantify the relative contributions of adaptation options. European Journal of Agronomy 33: 103-116. doi: DOI 10.1016/j.eja.2010.04.002.

Tao F, and Z. Zhang, 2011: Climate change, wheat productivity and water use in the North China Plain: A new super-ensemble-based probabilistic projection. Agricultural and Forest Meteorology ISSN 0168-1923,.

Tao, F., M. Yokozawa, and Z. Zhang, 2009a: Modelling the impacts of weather and climate variability on crop productivity over a large area: A new process-based model development, optimization, and uncertainties analysis. Agricultural and Forest Meteorology 149: 831-850. doi: 10.1016/j.agrformet.2008.11.004.

Tao F, Z. Zhang, J. Liu, and M. Yokozawa, 2009b: Modelling the impacts of weather and climate variability on crop productivity over a large area: A new super-ensemble-based probabilistic projection. Agricultural and Forest Meteorology 149: 1266-1278. doi: 10.1016/j.agrformet.2009.02.015.

van Diepen, C.A., J. Wolf, and H. van Keulen, 1989: WOFOST: a simulation model of crop production. Soil Use and Management, 5: 16-24

Vitart, F., A.W. Robertson, and D.L.T. Anderson, 2012: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. WMO Bulletin, 61(2), 23-28

Waha, K., L.G.J. van Bussel, , C. Müller, and A. Bondeau, 2012: Climate-driven simulation of global crop sowing dates. Global Ecology and Biogeography, 21: 247–259. doi:10.1111/j.1466-8238.2011.00678.x

Wallach, D., L.O. Mearns, M. Rivington, J.M. Antle, and A.C Ruane, 2015: Uncertainty in agricultural impact assessment. In Handbook of Climate Change and Agroecosystems: The Agricultural Model Intercomparison and Improvement Project (AgMIP). C. Rosenzweig, and D. Hillel, Eds., ICP Series on Climate Change Impacts, Adaptation, and Mitigation Vol. 3. Imperial College Press, 223-260, doi:10.1142/9781783265640_0009.

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

White, J.W., G. Hoogenboom, B.A. Kimball, and G.W. Wall, 2011a: Methodologies for simulating impacts of climate change on crop production. Field Crops Research 124:357-368.

White, J.W., G. Hoogenboom, P.W. Wilkens, P.W Stackhouse, J.M. Hoel, 2011b: Evaluation of Satellite-Based, Modeled-Derived Daily Solar Radiation Data for the Continental United States. Agron. J., 103(4): 1242–1251. doi: 10.2134/agronj2011.0038.

Williams J., C. Jones, J. Kiniry, D. Spanel, 1989: The EPIC crop growth-model. Transactions of the ASAE 32: 497-511.

Yin, X., and H.H van Laar, 2005: Crop Systems Dynamics: An Ecophysiological Simulation Model for Genotype-by-environment Interactions. Wageningen Academic Pub., 155pp.