The GGCMI Phase II experiment: simulating and emulating global crop yield responses to changes in CO_2, temperature, water, and nitrogen levels

James Franke[1,2], Joshua Elliott[2,3], Christoph Müller[4], Alexander Ruane[5], Abigail Snyder[6], Jonas Jägermeyr[3,2,4,5], Juraj Balkovic[7,8], Philippe Ciais[9,10], Marie Dury[11], Pete Falloon[12], Christian Folberth[7], Louis François[11], Tobias Hank[13], Munir Hoffmann[14,23], R. Cesar Izaurralde[15,16], Ingrid Jacquemin[11], Curtis Jones[15], Nikolay Khabarov[7], Marian Koch[14], Michelle Li[2,17], Wenfeng Liu[18,9], Stefan Olin[19], Meridel Phillips[5,20], Thomas A. M. Pugh[21,22], Ashwan Reddy[15], Xuhui Wang[9,10], Karina Williams[12], Florian Zabel[13], and Elisabeth Moyer[1,2]

[1]Department of the Geophysical Sciences, University of Chicago, Chicago, IL, USA
[2]Center for Robust Decision-making on Climate and Energy Policy (RDCEP), University of Chicago, Chicago, IL, USA
[3]Department of Computer Science, University of Chicago, Chicago, IL, USA
[4]Potsdam Institute for Climate Impact Research, Leibniz Association (Member), Potsdam, Germany
[5]NASA Goddard Institute for Space Studies, New York, NY, United States
[6]Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA
[7]Ecosystem Services and Management Program, International Institute for Applied Systems Analysis, Laxenburg, Austria
[8]Department of Soil Science, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovak Republic
[9]Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, 91191 Gif-sur-Yvette, France
[10]Sino-French Institute of Earth System Sciences, College of Urban and Env. Sciences, Peking University, Beijing, China
[11]Unité de Modélisation du Climat et des Cycles Biogéochimiques, UR SPHERES, Institut d'Astrophysique et de Géophysique, University of Liège, Belgium
[12]Met Office Hadley Centre, Exeter, United Kingdom
[13]Department of Geography, Ludwig-Maximilians-Universität, Munich, Germany
[14]Georg-August-University Göttingen, Tropical Plant Production and Agricultural Systems Modelling, Göttingen, Germany
[15]Department of Geographical Sciences, University of Maryland, College Park, MD, USA
[16]Texas Agrilife Research and Extension, Texas A&M University, Temple, TX, USA
[17]Department of Statistics, University of Chicago, Chicago, IL, USA
[18]EAWAG, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[19]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[20]Earth Institute Center for Climate Systems Research, Columbia University, New York, NY, USA
[21]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK.
[22]Birmingham Institute of Forest Research, University of Birmingham, Birmingham, UK.
[23]Leibniz Centre for Agricultural Landscape Research (ZALF), D-15374 Müncheberg, Germany

Correspondence: James Franke (jfranke@uchicago.edu)

Abstract. Concerns about food security under climate change motivate efforts to better understand future changes in crop yields. Process-based crop models, which represent plant physiological processes, are necessary tools for this purpose since they allow representing future conditions not sampled in the historical record and new locations where cultivation may shift. However, models remain uncertain and differ in many critical details. The Global Gridded Crop Model Intercomparison (GGCMI) Phase II experiment, an activity of the Agricultural Model Intercomparison and Improvement Project (AgMIP),

is designed to allow systematic evaluation and "emulation" of model responses to multiple interacting factors, including carbon dioxide, temperature, water availability, and nitrogen application (CTWN). In this paper we describe the GGCMI Phase II experimental protocol (simulations run with systematic uniform perturbations of historical climate) and its simulations (from nine crop models and five crops); identify responses that are robust across models and those that remain uncertain; and present an emulator or statistical representation of model responses. Modeled yields show robust decreases to warmer mean climatologies in almost all regions, with a nonlinear dependence that means that yield changes in warmer baseline locations are more sensitive to temperature increases. Inter-model uncertainty is qualitatively similar across all the four input dimensions, but is highest in high-latitude regions where crops may be grow in the future. For emulating these responses, the GGCMI Phase II systematic parameter sweep protocol allows disentangling the climate-driven mean response from year-over-year variations; we show that the two responses have very different relationships to standard climate metrics such as mean growing season temperature. The climatological mean yield response can be readily represented with a simple polynomial in almost all locations where crops are currently grown, permitting a tool that captures model responses in a lightweight, computationally tractable form. Crop model emulation should therefore facilitate both model comparison and integrated assessment of climate impacts.

1 Introduction

Understanding crop yield response to a changing climate is critically important, especially as the global food production system will face pressure from increased demand over the next century. Climate-related reductions in supply could therefore have severe socioeconomic consequences. Multiple studies using different crop or climate models concur in predicting sharp yield reductions on currently cultivated cropland under business-as-usual climate scenarios, although their yield projections show considerable spread (e.g. Rosenzweig et al., 2014; Schauberger et al., 2017; Porter et al. (IPCC), 2014, and references therein). Modeling crop responses continues to be challenging, as crop growth is a function of complex interactions between climate inputs and management practices. Intercomparison projects targeting model responses to important drivers are critical to improve future projections.

Computational models have been used to project crop yields since the 1950's, beginning with statistical models that attempt to capture the relationship between input factors and resultant yields (e.g. Heady, 1957; Heady and Dillon, 1961). These statistical models were typically developed on a small scale for locations with extensive histories of yield data. The emergence of electronic computers allowed development of numerical models that simulate the process of photosynthesis and the biology and phenology of individual crops (first proposed by de Wit (1957) and Duncan et al. (1967) and attempted by Duncan (1972); for a history of crop model development see Rosenzweig et al. (2014)). A half-century of improvement in both models and computing resources means that researchers can now run crop simulations for many years at higher spatial resolution on the global scale.

Both types of models continue to be used, and comparative studies have concluded that when done carefully, both approaches can provide similar yield estimates (e.g. Lobell and Burke, 2010; Moore et al., 2017; Roberts et al., 2017; Zhao et al., 2017). Models tend to agree broadly in major response patterns, including a reasonable representation of the spatial pattern in historical yields of major crops (e.g. Elliott et al., 2015; Müller et al., 2017) and projections of shifts in yield under future climate scenarios.

Process-based models do continue to struggle with some important details, including reproducing historical year-to-year variability (e.g. Müller et al., 2017), reproducing historical yields when driven by reanalysis weather (e.g. Glotter et al., 2014), and low sensitivity to extreme events (e.g. Glotter et al., 2015; Jägermeyr and Frieler, 2018; Schewe et al., 2019). These issues are driven in part by the diversity of new cultivars and genetic variants, which outstrips the ability of academic modeling groups to capture them (e.g. Jones et al., 2017). Models also do not simulate many additional factors affecting production, including but not limited to: pests, diseases, and weeds. For these reasons, individual studies must generally re-calibrate models to ensure that short-term predictions reflect current cultivars and management levels, and long-term projections retain considerable uncertainty (Wolf and Oijen, 2002; Jagtap and Jones, 2002; Iizumi et al., 2010; Angulo et al., 2013; Asseng et al., 2013, 2015). Inter-model discrepancies can also be high in areas not yet cultivated (e.g. Challinor et al., 2014; White et al., 2011). Finally, process-based models present additional difficulties for high-resolution global studies because of their complexity and computational requirements. For global economic impacts assessments, it is often impossible to integrate a set of process-based crop models directly into an integrated assessment model to estimate the potential cost of climate change to the agricultural sector.

Nevertheless, process-based models are necessary for understanding the future yield impacts of climate change. Cultivation may shift to new areas, where no yield data are currently available and therefore statistical models cannot be applied. Yield data are also often limited in the developing world, where future climate impacts may be the most critical. Finally, only process-based models can capture the growth response to novel conditions and practices that are not represented in historical data (e.g. Pugh et al., 2016; Roberts et al., 2017). These novel changes can include the direct fertilization effect of elevated CO_2, and changes in management practices that may mitigate climate-induced damages.

Interest has been rising in statistical emulation, which allows combining advantageous features of both statistical and process-based models. The approach involves constructing a statistical representation or "surrogate model" of complicated numerical simulations by using simulation output as the training data for a statistical model (e.g. O'Hagan, 2006; Conti et al., 2009). Emulation is particularly useful in cases where simulations are complex and output data volumes are large, and has been used in a variety of fields, including hydrology (e.g. Razavi et al., 2012), engineering (e.g. Storlie et al., 2009), environmental sciences (e.g. Ratto et al., 2012), and climate (e.g. Castruccio et al., 2014; Holden et al., 2014). For agricultural impacts studies, emulation of process-based models allows capturing key relationships between input variables in a lightweight, flexible form that is compatible with economic studies.

In the past decade, multiple studies have developed emulators of process-based crop simulations. Early studies proposing or describing potential crop yield emulators include Howden and Crimp (2005); Räisänen and Ruokolainen (2006); Lobell and Burke (2010), and Ferrise et al. (2011), who used a machine learning approach to predict Mediterranean wheat yields. Studies developing single-model emulators include Holzkämper et al. (2012) for the CropSyst model, Ruane et al. (2013) for

the CERES wheat model, and Oyebamiji et al. (2015) for the LPJmL model (for multiple crops, using multiple scenarios as a training set). More recently, emulators have begun to be used in the context of multi-model intercomparisons, with Blanc and Sultan (2015); Blanc (2017); Ostberg et al. (2018) and Mistry et al. (2017) using them to analyze the five crop models of the Inter-Sectoral Impacts Model Intercomparison Project (ISIMIP) (Warszawski et al., 2014), which simulated yields for maize, soy, wheat, and rice. Choices differ: Blanc and Sultan (2015) and Blanc (2017) base their emulation on historical simulations and three climate scenarios for one Representative Concentration Pathway (RPC8.5), which represents a high level of global warming; and use local weather variables and yields in their regression across soil regions; Ostberg et al. (2018) use global mean temperature change (and CO_2) as regressors then pattern-scale to emulate local yields; while Mistry et al. (2017) compare emulated and observed historical yields, using local weather data and a historical crop simulation. These efforts do share important common features: all emulate annual crop yields across the entire scenario or scenarios, and when future scenarios are considered, they are non-stationary, i.e. their input climate parameters evolve over the course of the simulations.

An alternative approach is to construct a training set of multiple stationary scenarios in which parameters are systematically varied. Such a "parameter sweep" offers several advantages for emulation over scenarios in which climate evolves over time. First, it allows separating the effects of different variables that impact yields but that are highly correlated in realistic future scenarios (e.g. CO_2 and temperature). Second, it allows making a distinction between year-over-year yield variations and climatological changes, which may involve different responses to the particular climate regressors used (e.g. Ruane et al., 2016). For example, if year-over-year yield variations are driven predominantly by variations in the distribution of temperatures throughout the growing period, and long-term climate changes are driven predominantly by shifts in means, then regressing on the mean growing period temperature will produce different yield responses at annual vs. climatological timescales. Disadvantages of this approach include neglecting changes in seasonality and some implausible combinations of input settings (e.g. colder temperature and high CO_2).

Systematic parameter sweeps have begun to be used in crop model evaluation and emulation, with early efforts in 2014 and 2015 (Ruane et al., 2014; Makowski et al., 2015; Pirttioja et al., 2015), and several recent studies in 2018 (Fronzek et al., 2018; Snyder et al., 2018; Ruiz-Ramos et al., 2018). All three 2018 studies sample multiple perturbations to temperature and precipitation (with Snyder et al. (2018) and Ruiz-Ramos et al. (2018) adding CO_2 as well), in 132, 99 and 220 different combinations, respectively, and take advantage of the structured training set to construct emulators ("response surfaces") of climatological mean yields, omitting year-over-year variations. All studies focus on a limited number of sites; Fronzek et al. (2018) and Ruiz-Ramos et al. (2018) simulate only wheat (over many models) and Snyder et al. (2018) analyzes four crops (maize, wheat, rice, soy) for agricultural impacts experiments with the GCAM (Calvin et al., 2019) model.

In this paper we describe a new comprehensive dataset designed to expand the parameter sweep approach still further. The Global Gridded Crop Model Intercomparison (GGCMI) Phase II experiment involves running a suite of process-based crop models across historical conditions perturbed by a set of discrete steps in different input parameters, including an applied nitrogen dimension. The experimental protocol involves 756 different parameter combinations for each model and crop, with simulations providing near-global coverage at a half degree spatial resolution. The experiment was conducted as part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013, 2014), an international

Table 1. GGCMI Phase II input levels. Temperature and precipitation values indicate the perturbations from the historical climatology. W-percentage does not apply to the irrigated (W_{inf}) simulations, which are all simulated at the maximum beneficial levels of water. One model provided simulations at the T + 5 level. See Figure S3 in the supplement for number of simulations associated with each combination of input levels.

Input variable	**Tested range**	**Unit**
CO_2 (C)	**360**, 510, 660, 810	ppm
Temperature (T)	-1, **0**, 1, 2, 3, 4, 6	°C
Precipitation (W)	-50, -30, -20, -10, **0**, 10, 20, 30, (and W_{inf})	%
Applied nitrogen (N)	10, 60, **200**	kg ha^{-1}

effort conducted under a framework similar to the Climate Model Intercomparison Project (CMIP) (Taylor et al., 2012; Eyring et al., 2016). The GGCMI protocol builds on the AgMIP Coordinated Climate-Crop Modeling Project (C3MP) (Ruane et al., 2014; McDermid et al., 2015) and contributes to the AgMIP Coordinated Global and Regional Assessments (CGRA) (Ruane et al., 2018; Rosenzweig et al., 2018). GGCMI Phase II is designed to allow addressing goals such as understanding where highest-yield regions may shift under climate change; exploring future adaptive management strategies; understanding how interacting input drivers affect crop yield; quantifying uncertainties across models and major drivers; and testing strategies for producing lightweight emulators of process-based models. In this paper, we describe the GGCMI Phase II experiments, present initial results, and introduce a spatially explicit emulator for climatological time scales that allows for representing crop model responses in economic assessment models and other applications.

2 Simulation – Methods

GGCMI Phase II is the continuation of a multi-model comparison exercise begun in 2014. The initial Phase I compared harmonized yields of 21 models for 19 crops over a 30-year historical (1981-2010) scenario with a primary goal of model evaluation (Elliott et al., 2015; Müller et al., 2017). Phase II compares simulations of 12 models for 5 crops (maize, rice, soybean, spring wheat, and winter wheat) over the same historical time series (1981-2010) used in Phase I, but with individual climate or management inputs adjusted from their historical values. The reduced set of crops includes the three major global cereals and the major legume and accounts for over 50% of human calories (in 2016, nearly 3.5 billion tons or 32% of total global crop production by weight (Food and Agriculture Organization of the United Nations, 2018).

The guiding scientific rationale of GGCMI Phase II is to provide a comprehensive, systematic evaluation of the response of process-based crop models to CO_2, temperature, water, and applied nitrogen (collectively referred to as "CTWN"). The dataset is designed to allow researchers to:

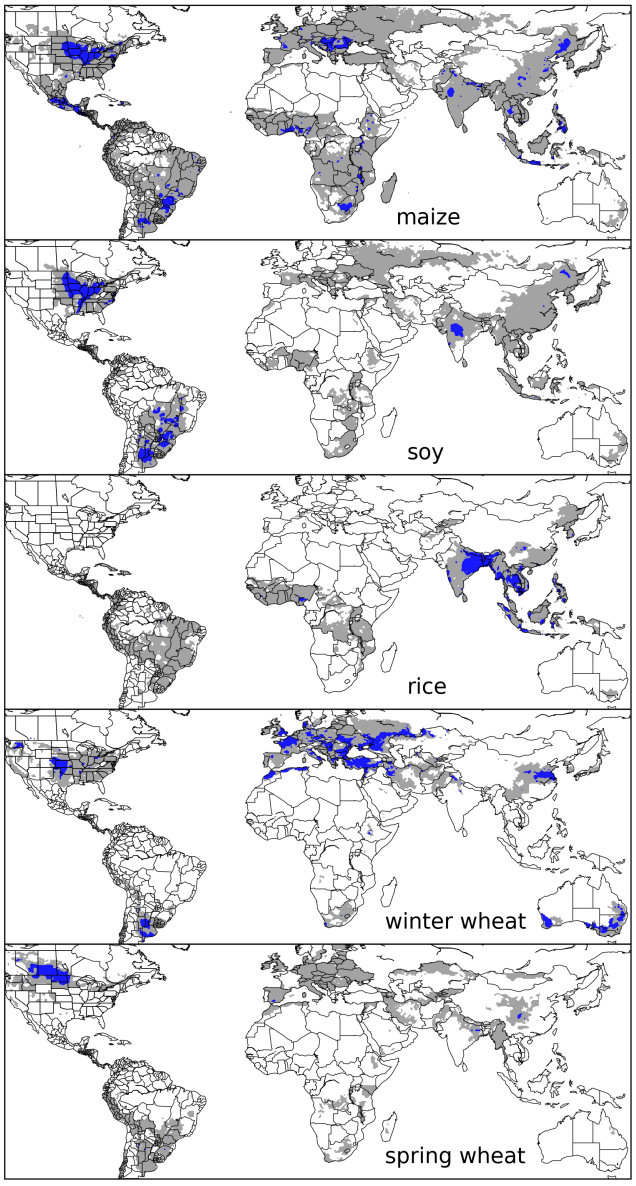

Figure 1. Presently cultivated area for rainfed crops. Blue indicates grid cells with more that 20,000 hectares (∼10% of the equatorial grid cell). Gray contour shows area with more that 10 hectares cultivated. Cultivated areas for maize, rice, and soy are taken from the MIRCA2000 ("monthly irrigated and rainfed crop areas around the year 2000") dataset (Portmann et al., 2010). Areas for winter and spring wheat areas are adapted from MIRCA2000 data and sorted by growing period. For analogous figure of irrigated crops, see Figure S1.

- Enhance understanding of how models work by characterizing their sensitivity to input climate and nitrogen drivers.
- Study the interactions between climate variables and nitrogen inputs in driving modeled yield impacts.

Table 2. Models included in GGCMI Phase II and the number of C, T, W, and N simulations that each performs, with 756 as the maximum. "N-Dim." indicates whether the simulations include varying nitrogen levels. Two models provide only one nitrogen level. All models provide the same set of simulations across all modeled crops, but some omit individual crops. (For example, APSIM does not simulate winter wheat.)

Model (Key Citations)	Maize	Soy	Rice	Winter wheat	Spring wheat	N dim.	Simulations per crop
APSIM-UGOE, Keating et al. (2003); Holzworth et al. (2014)	X	X	X	–	X	X	44
CARAIB, Dury et al. (2011); Pirttioja et al. (2015)	X	X	X	X	X	–	252
EPIC-IIASA, Balkovič et al. (2014)	X	X	X	X	X	X	39
EPIC-TAMU, Izaurralde et al. (2006)	X	X	X	X	X	X	765
JULES, Osborne et al. (2015); Williams and Falloon (2015); Williams et al. (2017)	X	X	X	–	X	–	252
GEPIC, Liu et al. (2007); Folberth et al. (2012)	X	X	X	X	X	X	430
LPJ-GUESS, Lindeskog et al. (2013); Olin et al. (2015)	X	–	–	X	X	X	756
LPJmL, von Bloh et al. (2018)	X	X	X	X	X	X	756
ORCHIDEE-crop, Wu et al. (2016)	X	–	X	–	X	X	33
pDSSAT, Elliott et al. (2014); Jones et al. (2003)	X	X	X	X	X	X	672
PEPIC, Liu et al. (2016a, b)	X	X	X	X	X	X	149
PROMET, Hank et al. (2015); Mauser et al. (2015)	X	X	X	X	X	X†	261
Totals	12	10	11	9	12	10	5240

– Explore differences in crop response to warming across the Earth's climate regions.

– Provide a dataset that allows statistical emulation of crop model responses for downstream modelers.

The experimental protocol consists of 9 levels for precipitation perturbations, 7 for temperature, 4 for CO_2, and 3 for applied nitrogen, for a total of 672 simulations for rainfed agriculture and an additional 84 for irrigated (Table 1). For irrigated simulations, limitations from actual water supply are not considered. Temperature perturbations are applied as absolute offsets from the daily mean, minimum, and maximum temperature time series for each grid cell. Precipitation perturbations are applied

as fractional changes at the grid cell level, and CO_2 and nitrogen levels are specified as discrete values applied uniformly over all grid cells. Limits for the climate variable perturbations are selected to represent reasonable ranges for potential climate changes in the medium term. In most cases, historical daily climate inputs are taken from the 0.5 degree NASA AgMERRA daily gridded re-analysis product specifically designed for agricultural modeling, with satellite-corrected precipitation (Ruane et al., 2015), but two models (JULES and PROMET) require sub-daily input data and use alternative sources. Note that CO_2 changes are applied independently of changes in climate variables, so that higher CO_2 is not associated with higher temperatures. The resulting GGCMI Phase II dataset captures a distribution of crop responses over the potential space of future climate conditions.

The 12 models included in GGCMI Phase II are all process-based crop models that are widely used in impacts assessments (Table 2). Although some models share a common base (e.g. the LPJ family or the EPIC family of models), they have subsequently developed independently. Differences in model structure mean that several key factors are not standardized across the experiment, including "non-nitrogen" nutrients, carry-over effects across growing years including residue management and soil moisture, and the extent of simulated area for different crops. Growing seasons are standardized across models (with assumptions based on Sacks et al. (2010) and Portmann et al. (2008, 2010)), but vary by crop and by location on the globe. For example, maize is sown in March in Spain, in July in Indonesia, and in December in Namibia. All stresses are disabled other than factors related to nutrients, temperature, and water (e.g. alkalinity and salinity). No additional nitrogen inputs, such as atmospheric deposition, are considered, but some model treatments of soil organic matter allows additional nitrogen release through mineralization. See Elliott et al. (2015) for further details on model setup for intercomparison in the GGCMI protocol. Not all modeling teams provide the full simulation protocol, for instance, CARIAB and JULES do not simulate the nitrogen dimension and some crops are not parameterized in each model (see Table 2 for details). Note that the three models that provide less than 50 simulations are excluded from the emulator analysis (APSIM-UGOE, EPIC-IIASA, and ORCHIDEE-crop).

Each model is run at 0.5 degree spatial resolution and covers all currently cultivated areas and much of the uncultivated land area. (See Figure 1 for the present-day cultivated area of rainfed crops, and Figure S1 in the Supplemental Material for irrigated crops.) Coverage extends considerably outside currently cultivated areas because cultivation will likely shift under climate change. However, areas are not simulated in some cases if they are assumed to remain non-arable even under an extreme climate change; these regions include Greenland, far-northern Canada, Siberia, Antarctica, the Gobi and Sahara Deserts, and central Australia. All models produce as output crop yields (tons ha^{-1} $year^{-1}$) for each 0.5 degree grid cell. Because both yields and yield changes vary substantially across models and across grid cells, we primarily analyze relative change from a baseline. We take as the baseline the scenario with historical climatology (i.e. T and P changes of 0), C of 360 ppm, and applied N at 200 kg $ha^{-}1$. The GGCMI Phase II simulations are designed for evaluating changes in yield but not absolute yields, since they omit detailed calibrations. To provide some evaluation of the skill of the process-based models used, we repeat the evaluation exercises of Müller et al. (2017) for GGCMI Phase I. See Appendix A for details on simulation model evaluation.

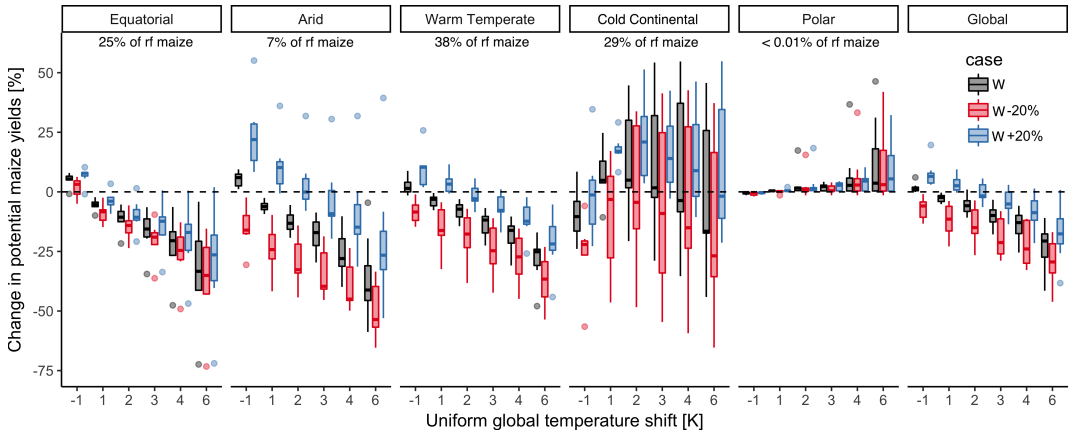

Figure 2. Illustration of the distribution of regional yield changes across the multi-model ensemble, split by Köppen-Geiger climate regions (Rubel and Kottek, 2010). We show responses of a single crop (rainfed maize) to applied uniform temperature perturbations, for three discrete precipitation perturbation levels (-20%, 0%, and +20%), with CO_2 and nitrogen held constant at baseline values (360 pmm and 200 kg ha^{-1} yr^{-1}). Y-axis is fractional change in the regional average climatological potential yield relative to the baseline. Box-and-whiskers plots show distribution across models, with median marked; edges are first and third quartiles, i.e. box height is the interquartile range (IQR). If all models like within 1.5·IQR then whiskers extend to maximum and minimum of simulations, else the outlier is shown separately. Outliers in the tropics (strong negative impact of temperature increases) are the pDSSAT model; outliers in the high-rainfall case (strong positive impact of precipitation increases) are the JULES model. Figure shows all modeled land area; see Figure S4 in the supplemental material for currently-cultivated land and Figure S5 for other crops. Panel text gives the percentage of rainfed maize presently cultivated in each climate zone (data from Portmann et al., 2010). Note that Rubel and Kottek (2010) use the name 'Snow' rather than 'Cold continental'. Outside high-latitude regions ('Cold continental' and 'Polar'), models generally agree, with projected declines under increasing temperatures larger than inter-model variance. The right panel (Global) shows yield responses to a globally uniform temperature shift; note that these results are not directly comparable to simulations of more realistic climate scenarios.

3 Simulation – Results

Crop models in the GGCMI Phase II ensemble show broadly consistent responses to climate and management perturbations in most regions, with a strong negative impact of increased temperature in all but the coldest regions. We illustrate this result for rainfed maize in Figure 2, which shows yields across all grid cells for the primary Köppen-Geiger climate regions (Rubel and Kottek, 2010). In warming scenarios with precipitation held constant, all models show decreases in maize yield in the 'warm temperate', 'equatorial', and 'arid' regions that account for nearly three-quarters of global maize production. These impacts are robust for even moderate climate perturbations. In the 'warm temperate' zone, even a 1 degree temperature rise with other variables held fixed leads to a median yield reduction that exceeds the variance across models. A 6 degree temperature rise results in median loss of ∼25% of yields with a signal to noise ratio of nearly three to one. A notable exception is the 'cold continental' region, where models disagree strongly, extending even to the sign of impacts. Other crops show similar responses

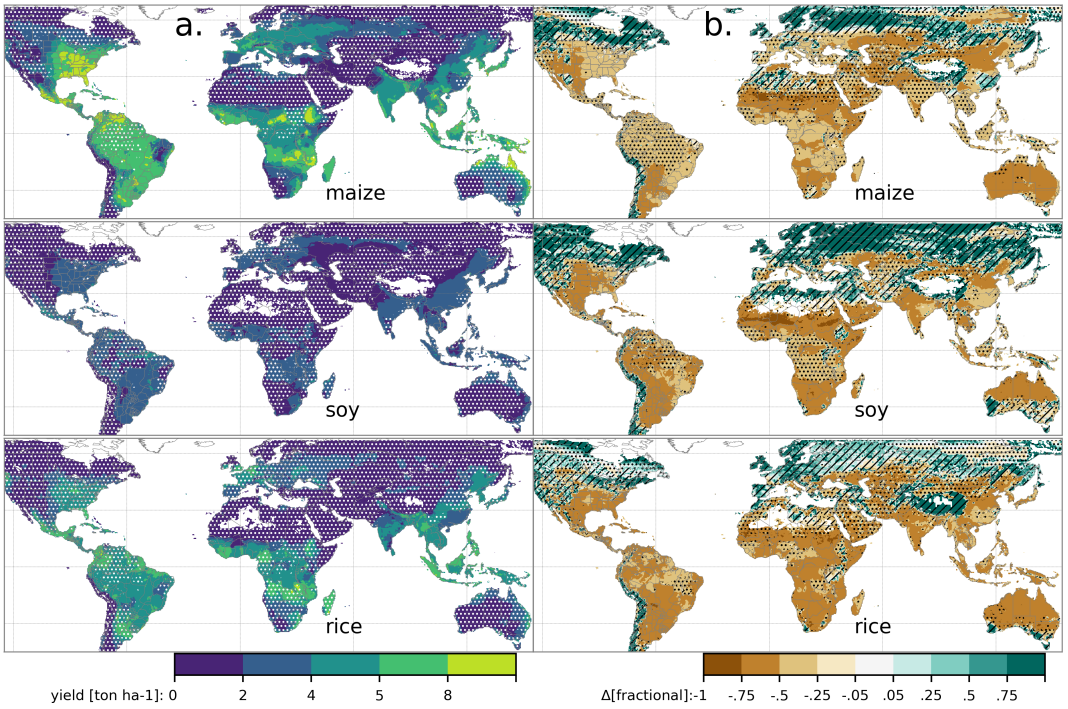

Figure 3. Illustration of the spatial pattern of potential yields and potential yield changes in the GGCMI Phase II ensemble, for three major crops. Left column (a) shows multi-model mean climatological yields for the baseline scenario for (top–bottom) rainfed maize, soy, and rice. Wheat shows a qualitatively similar response, see Figure S16 in the supplemental material. White stippling indicates areas where these crops are not currently cultivated. Absence of cultivation aligns well with the lowest yield contour (0-2 ton ha^{-1}). Right column (b) shows the multi-model mean fractional yield change in the extreme T + 4 °C scenario (with other inputs at baseline values). Areas without hatching or stippling are those where confidence in projections is high: the multi-model mean fractional change exceeds two standard deviations of the ensemble. ($\Delta > 2\sigma$). Hatching indicates areas of low confidence ($\Delta < 1\sigma$), and stippling areas of medium confidence ($1\sigma < \Delta < 2\sigma$). Crop model results in cold areas, where yield impacts are on average positive, also have the highest uncertainty.

to warming, with robust yield losses in warmer locations and high inter-model variance in the 'cold continental' regions (Figure S5).

The effects of rainfall changes on maize yields shown in Figure 2 are also as expected and are consistent across models. Increased rainfall mitigates the negative effect of higher temperatures by counteracting the increased evapo-transpiration to some degree, most strongly in arid regions. Decreased rainfall amplifies yield losses and also increases inter-model variance; i.e. models agree that the response to decreased water availability is negative in sign but disagree on its magnitude. We show only rainfed maize here; see Figure S6 for comparison between rainfed and irrigated case. As expected, irrigated crops are more resilient to temperature increases in all regions, especially so where water is limiting. See Figures S7-15 in the supplement for other crops.

Mapping the distribution of baseline yields and yield changes shows the geographic dependencies that underlie these results. Crop cultivation areas and yield changes with respect to the T+4 scenario show distinct geographic pattern (Figure 3). Absolute yield potentials show strong spatial variation, with much of the Earth's surface area unsuitable for any of these crops. In general, models agree most on yield response in regions where yield potentials are currently high and therefore where crops are currently grown. Models show robust decreases in yields at low latitudes, and highly uncertain median increases at most high latitudes, possibly due to how crop failures are considered across different models. For wheat crops see Figure S16 wheat projections are more uncertain, possibly because simulation calibration is especially important (e.g. Asseng et al., 2013).

4 Emulation – Methods

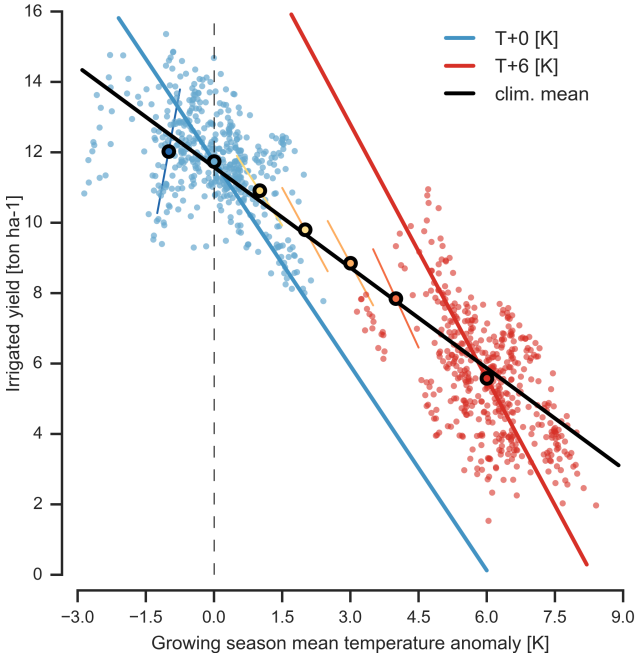

Figure 4. Example showing distinction between crop yield responses to year-to-year and climatological mean temperature shifts. Figure shows irrigated maize for a representative high-yield region (nine adjacent grid cells in northern Iowa) from the pDSSAT model, for the baseline 1981-2010 historical climate (blue) and for the scenario of maximum temperature change (+6 K, red). Other variables are held at baseline values, and the choice of irrigated yields means that precipitation is not a factor. Open black circles mark climatological mean yield values for all six temperature scenarios (T-1, +0, +1, +2, +3, +4, +6). Colored lines show total least squares linear regressions of year-over-year variations in each scenario. Black line shows the fit through the climatological mean values. Responses to year-over-year temperature variations (colored lines) are 100–200% larger than those to long-term climate perturbations, rising under warmer conditions. Linear fits are shown for illustration purposes and are not used in the emulation models.

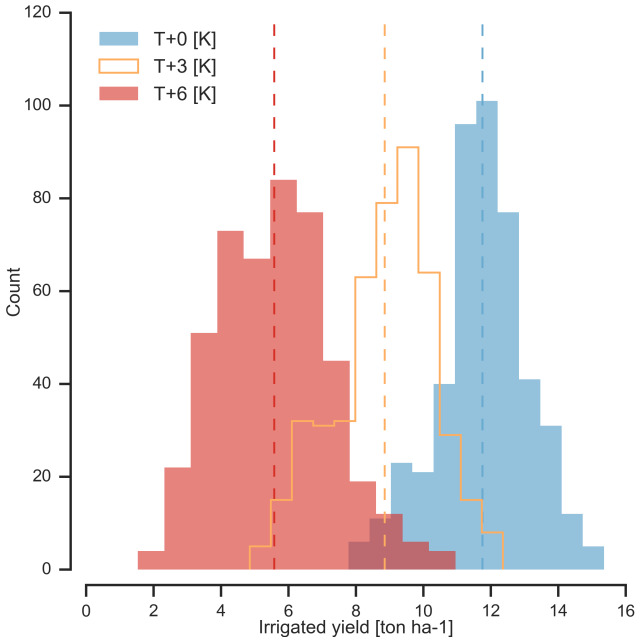

Figure 5. Example showing climatological mean yields and distribution of yearly yields for three 30-year scenarios. Figure shows irrigated maize for nine adjacent high-yield grid cells of Figure 4 from the pDSSAT model, for the baseline 1981-2010 historical climate (blue) and for scenarios with temperature shifted by T+3 (orange) and T+6 K (red), with other variables held at baseline values. The stronger year-over-year temperature response with higher temperatures seen in Figure 4 is manifested here as larger variance in annual yields even though the variance in climate drivers is identical. In this work we emulate not the year-over-year distributions but the climatological mean response (dashed vertical lines).

As part of our demonstration of the properties of the GGCMI Phase II dataset, we construct an emulator of 30-year climatological mean yields. This approach is made possible by the structured set of simulations involving systematic perturbations. In the GGCMI Phase II dataset, the year-over-year responses are generally quantitatively distinct from (and larger than) climatological mean responses. In the example Figure 4, responses to year-over-year temperature variations are 100% larger than those to long-term perturbations in the baseline case, and larger still under warmer conditions, rising to nearly 200% more in the T+6 case. The stronger year-over-year response under warmer conditions also manifests as a wider distribution of yields (Figure 5). As discussed previously, year-over-year and climatological responses can differ for many reasons including memory in the crop model, lurking covariants, and differing associated distributions of daily growing-season daily weather (e.g. Ruane et al., 2016). Note that the GGCMI Phase II datasets do not capture one climatological factor, potential future distributional shifts, because all simulations are run with fixed offsets from the historical climatology. Prior work has suggested that mean changes are the dominant drivers of climatological crop yield shifts in non-arid regions (e.g. Glotter et al., 2014).

Emulation involves fitting individual regression models for each crop, simulation model, and 0.5 degree geographic pixel from the GGCMI Phase II dataset; the regressors are the applied constant perturbations in CO_2, temperature, water, and nitrogen

(C,T, W, N). We regress 30-year climatological mean yields against a third-order polynomial in C, T, W, and N with interaction terms. We aggregate the entire 30-year run in each case to improve signal to noise ratio. The higher-order terms are necessary to capture any nonlinear responses, which are well-documented in observations for temperature and water perturbations (e.g. Schlenker and Roberts (2009) for T and He et al. (2016) for W). We include interaction terms (both linear and higher-order) because past studies have shown them to be significant effects. For example, Lobell and Field (2007) and Tebaldi and Lobell (2008) showed that in real-world yields, the joint distribution in T and W is needed to explain observed yield variance. (C and N are fixed in these data.) Other observation-based studies have shown the importance of the interaction between water and nitrogen (e.g. Aulakh and Malhi, 2005), and between nitrogen and CO_2 (Osaki et al., 1992; Nakamura et al., 1997). To avoid over-fitting or unstable parameter estimation, we apply a feature selection procedure (described below) that reduces the potential 34-term polynomial (for the rainfed case) to 23 terms.

We do not focus on comparing different functional forms in this study, and instead choose a relatively simple parametrization that allows for some interpretation of coefficients. Some prior studies have used other statistical specifications, e.g. 39 terms in Blanc and Sultan (2015) and Blanc (2017), who borrow information across space by fitting grid points simultaneously across soil region in a panel regression. The simple functional form used here allows emulation at the grid cell level. The emulation therefore indirectly includes any yield response to geographically distributed factors such as soil type, insolation, and the baseline climate. We hold the statistical specification constant across all crops and models to facilitate parameter by parameter simulation model comparison.

4.1 Feature selection procedure

Although the GGCMI Phase II sampled variable space is large, it is still sufficiently limited that use of the full polynomial expression described above can be problematic. We therefore reduce the number of terms through a feature selection cross-validation process in which terms in the polynomial are tested for importance. In this procedure higher-order and interaction terms are added successively to the regression model; we then follow the reduction of the aggregate mean squared error with increasing terms and eliminate those terms that do not contribute significant reductions. See section 1 in the supplemental documents for more details. We select terms by applying the feature selection process to three example models that provided the complete set of 672 rainfed simulations (pDSSAT, EPIC-TAMU, and LPJmL); the resulting choice of terms is then applied for all emulators and all crops.

Feature importance is remarkably consistent across all three models and across all crops (see Figure S17 in the supplemental material). The feature selection process results in a final polynomial in 23 terms, with 11 terms eliminated. We omit the N^3 term, which cannot be fitted because we sample only three nitrogen levels. We eliminate many of the C terms: the cubic, the CT, CTN, and CWN interaction terms, and all higher order interaction terms in C. Finally, we eliminate two 2nd-order interaction terms in T and one in W. Implication of this choice include that nitrogen interactions are complex and important, and that water interaction effects are more nonlinear than those in temperature. The resulting statistical model (Equation 1) is used for all grid cells, models, and rainfed crops. (The regressions for irrigated crops do not contain the W terms and the models that do not sample the nitrogen levels omit the N terms).

$$
\begin{aligned}
Y \; &= K_1 \\
&+ K_2 C + K_3 T + K_4 W + K_5 N \\
&+ K_6 C^2 + K_7 T^2 + K_8 W^2 + K_9 N^2 \\
&+ K_{10} CW + K_{11} CN + K_{12} TW + K_{13} TN + K_{14} WN \\
&+ K_{15} T^3 + K_{16} W^3 + K_{17} TWN \\
&+ K_{18} T^2 W + K_{19} W^2 T + K_{20} W^2 N \\
&+ K_{21} N^2 C + K_{22} N^2 T + K_{23} N^2 W
\end{aligned}
\tag{1}
$$

To fit the parameters K, we use a Bayesian Ridge probabilistic estimator (MacKay, 1991), which reduces volatility in parameter estimates when the sampling is sparse, by weighting parameter estimates towards zero. The Bayesian Ridge method is necessary to maintain a consistent functional form across all models and locations. We use the implementation of the Bayesian Ridge estimator from the scikit-learn package in Python (Pedregosa et al., 2011). In the GGCMI Phase II experiment, the most problematic fits are those for models that provided a limited number of cases or for low-yield geographic regions where some modeling groups did not run all scenarios. We do not attempt to emulate models that provided less than 50 simulations. The lowest number of simulations emulated across the full parameter space is then 130 (for the PEPIC model). The yield output for a single GGCMI Phase II model that simulates all scenarios and all five crops is ∼12.5 GB; the emulator is ∼100 MB, a reduction by over two orders of magnitude.

5 Emulation – Results

Emulation provides not only a computational tool but a means of understanding and interpreting crop yield response across the parameter space. Emulation is only possible when crop yield responses are sufficiently smooth and continuous to allow fitting with a relatively simple functional form, but this condition largely holds in the GGCMI Phase II simulations. Responses are quite diverse across locations, crops, and models, but in most cases local responses are regular enough to permit emulation. We show illustrations of emulation fidelity in this section; for more detailed discussion see Appendix B.

Crop yield responses are geographically diverse, even in high-yield and high-cultivation areas. Geographic diversity is high within a single crop and model (Figure 6 rainfed maize in pDSSAT); this heterogeneity supports the choice of emulating at the grid cell level. Each panel in Figure 6 shows simulated yield output from scenarios varying only along a single dimension (CO_2, temperature, precipitation, or nitrogen addition), with other inputs held fixed at baseline levels, compared to the full 4D emulation across the parameter space. Yields evolve smoothly across the space sampled, and the polynomial fit captures the climatological response to perturbations. Crop yield responses generally follow similar functional forms across models, though with a large spread in magnitude partly due to the lack of calibration. Iter-model diversity for a single crop and location is also high (Figure 7, rainfed maize in northern Iowa, also shown in Figure 6). Differences in response shape can lead to differences

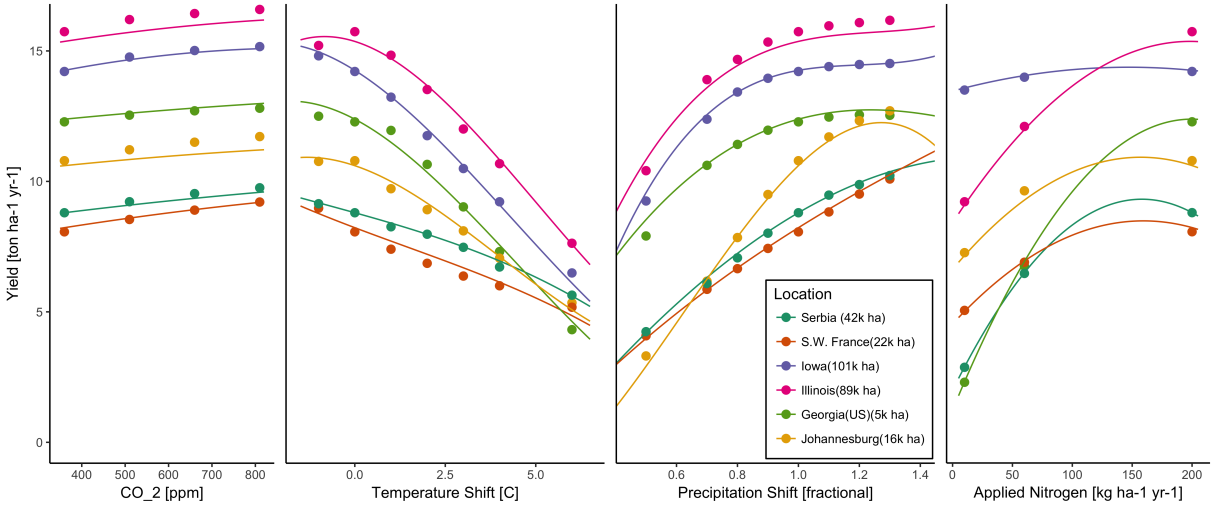

Figure 6. Illustration of spatial variations in yield response and emulation ability. We show rainfed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares cultivated in each selected grid cell. Each panel shows variation along a single variable, with others held at baseline values. Dots show climatological mean yields and lines the results of the full 4D emulator of Equation 1. In general the climatological response surface is sufficiently smooth that it can be represented within the sampled variable space by the simple polynomial used in this work. Extrapolation can however produce misleading results. Nitrogen fits in some cases may not be realistic at intermediate values given limited sampling. For more detailed emulator assessment, see Appendix B.

in the fidelity of emulation, though comparison here is complicated by the different simulation experiment sampling regimes across models. Note that models are most similar in their responses to temperature perturbations.

While the nitrogen dimension is important, it is also the most problematic to emulate in this work because of its limited sampling. The GGCMI Phase II protocol specified only three nitrogen levels (10, 60 and 200 kg N y^{-1} ha^{-1}), so a third-order fit would be over-determined but a second-order fit can result in potentially unphysical results. Steep and nonlinear declines in yield with lower nitrogen levels mean that some regressions imply a peak in yield between the 100 and 200 kg N y^{-1} ha^{-1} levels. While it is possible that over-application of nitrogen at the wrong time in the growing period could lead to reduced yields, these features are potentially an artifact of under sampling. In addition, the polynomial fit cannot capture the well-documented saturation effect of nitrogen application (e.g. Ingestad, 1977) as accurately as would be possible with a non-parametric model.

The emulation fidelity demonstrated here is sufficient to allow using emulated response surfaces to compare model responses and derive insight about impacts projections. Because the emulator or "surrogate model" transforms the discrete simulation sample space into a continuous response surface at any geographic scale, it can be used for a variety of applications, including construction of continuous damage functions. As an example, we show a damage function constructed from the 4D emulation, aggregated to global yield, with simulated values shown for comparison (Figure 8, which shows maize on currently cultivated land; see Figures S18-S21 for other crops and dimensions). The emulated values closely match simulations even at this aggre-

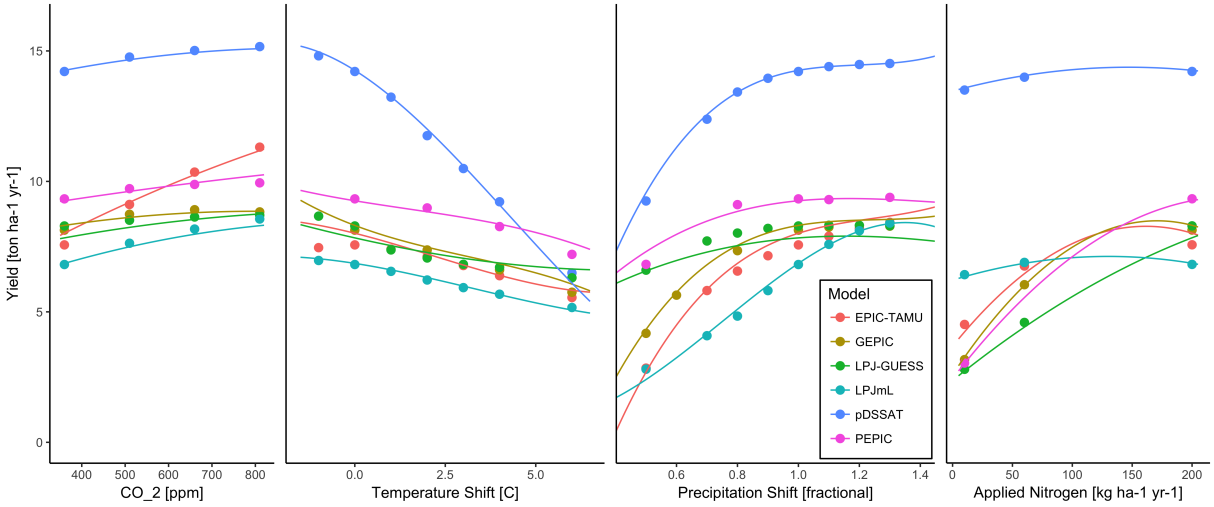

Figure 7. Illustration of across-model variations in yield response. Figures shows simulations and emulations from six models for rainfed maize in the same Iowa grid cell shown in Figure 6, with the same plot conventions. Models that do not simulate the nitrogen dimension are omitted for clarity. Note that models are uncalibrated, increasing spread in absolute yields. While most model responses can readily emulated with a simple polynomial, some response surfaces diverge slightly from the polynomial approach (e.g. LPJ-GUESS here) and lead to emulation error, though error generally remains small relative to inter-model uncertainty. For more detailed emulator assessment, see Appendix B. As in Figure 6, extrapolation out of the sample space is potentially problematic.

gation level. Note that these functions are presented only as examples and do not represent true global projections, because they are developed from simulation data with a uniform temperature shift while increases in global mean temperature should manifest non-uniformly in space and distributions (Sippel et al., 2015). The global coverage of the GGCMI Phase II simulations allows impacts modelers to apply arbitrary geographically-varying climate projections, as well as arbitrary aggregation masks, to develop damage functions for any climate scenario and any geopolitical or geographic level.

6 Discussion and Conclusions

The GGCMI Phase II experiment provides a database targeted to allow detailed study of crop yields from process-based models under climate change. The systematic input parameter variations are designed to facilitate not only comparing the sensitivities of process-based crop yield models to changing climate and management inputs but also evaluating the complex interactions between driving factors (CO_2, temperature, precipitation, and applied nitrogen). Its global nature also allows identifying geographic shifts in high yield potential locations. We expect that the simulations will yield multiple insights in future studies, and show here a selection of preliminary results to illustrate their potential uses.

First, the GGCMI Phase II simulations allow identifying major areas of uncertainty. Inter-model uncertainty is qualitatively similar across all four inputs tested at the globally aggregate level with some notable exceptions. For example, uncertainty in

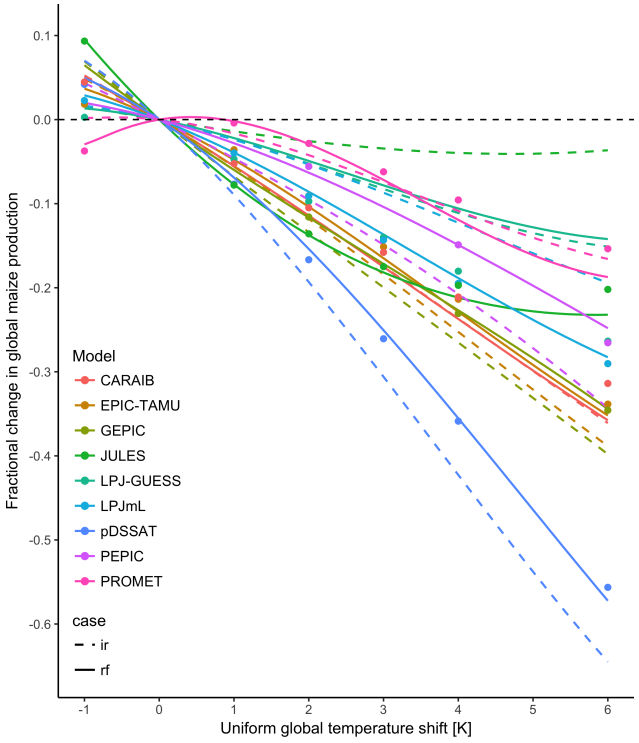

Figure 8. Global emulated damages for maize on currently cultivated lands for the GGCMI Phase II models emulated, for uniform temperature shifts with other inputs held at baseline. (The damage function is created from aggregating up emulated values at the grid cell level, not from a regression of global mean yields.) Lines are emulations for rainfed (solid) and irrigated (dashed) crops; for comparison, dots are the simulated values for the rainfed case. For most models, irrigated crops show a sharper reduction than rainfed because of the locations of cultivated areas: irrigated crops tend to be grown in warmer areas where impacts are more severe for a given temperature shift (the exceptions are PROMET, JULES, and LPJmL). For other crops and scenarios see Figures S18-21 in the supplemental material.

the response to CO_2 and water dominates for wheat, and soy is insensitive to nitrogen (see Figure S22). Across geographic regions, projections are most uncertain in the high latitudes where yields may increase, and most robust in low latitudes where yield impacts are largest. Inter-model uncertainty in the high latitudes is like due in part to the differences in the way crop failures are reported across models.

Second, the GGCMI Phase II simulations allow an understanding of the way that climate-driven changes and locations of cultivated land combine to produce yield impacts. One counterintuitive result immediately apparent is that irrigated maize shows steeper yield reductions under warming than does rainfed maize when considered only over currently cultivated land (compare Figure 8 to Figure 2 and Figures S6 and S7). The effect results from geographic differences in cultivation. In any given location, irrigation (or additional rainfall) increases crop resiliency to temperature increases, but irrigated maize is grown in warmer locations where the impacts of warming are more severe (Figures S6 and S7). The same behavior holds for rice and winter wheat, but not for soy or spring wheat (Figures S8-S15). Irrigated wheat and maize are also more sensitive to nitrogen

fertilization levels than the analogous non-irritated crops, presumably because those rainfed crops are limited by water as well as nitrogen availability (Figure S21). Soy as an efficient atmospheric nitrogen-fixer is relatively insensitive to nitrogen, and rice is not generally grown in water-limited conditions.

Third, we show that the GGCMI Phase II climatological crop yield response can be emulated with a relatively simple reduced-form statistical model. The systematic parameter sampling in the GGCMI Phase II procedure provides information on the influence of multiple interacting factors in a way that realistic climate model simulations cannot. Emulating the climatological response isolates long-term impacts from any confounding factors that complicate year-over-year changes. The use of a relatively simple functional form offers the possibility of physical interpretation of parameter values and application to model intercomparison.

While the GGCMI Phase II database should offer the foundation for multiple future studies, several cautions need to be noted. Emulating these responses allows estimating agricultural impacts under arbitrary climate scenarios, but extrapolating should be avoided. Additionally, because the simulation protocol was designed to focus on change in yield under climate perturbations and not on replicating real-world yields, the models are not formally calibrated so cannot be used for impacts projections unless used in conjunction with historical data (or data products). Finally, the GGCMI Phase II simulations apply uniform perturbations to historical climate inputs and they do not sample potential changes in climate variability. Although such changes in climate are uncertain and remain poorly characterized (e.g. Alexander et al., 2006; Kodra and Ganguly, 2014), follow-up experiments may wish to consider them. Several recent studies have described procedures for generating simulations that combine historical data with model projections of changes not only in temperature and precipitation means but in their marginal distributions or temporal dependence (e.g. Leeds et al. (2015); Poppick et al. (2016); Chang et al. (2016) and Haugen et al. (2018)).

The GGCMI Phase II output dataset invites a broad range of potential future avenues of analysis. A major target area of research is studying the simulation models themselves including: a detailed examination of interaction terms between the major input drivers, a robust quantification of the sensitivity of different models to the input drivers, and comparisons with field-level experimental data. The parameter space tested in GGCMI Phase II may allow investigations into yield variability and response to extremes under changing management and CO_2 levels and allow the study of geographic shifts in optimal growing regions for different crops. Emulation studies that are possible include a more systematic evaluation of different statistical model specifications and formal calculation of uncertainties in derived parameters. The development of multi-model ensembles such as GGCMI Phase II provides a way to begin to better understand crop responses to a range of potential climate inputs, improve process based models, and explore the potential benefits of adaptive responses included shifting growing period, cultivar types and cultivar geographic extent.

Code and data availability. The resulting polynomial emulator parameter matrices for all crop model emulators are available on location, as are the raw simulation yield outputs. Code to run the emulator is available upon request to the author.

Appendix A

A1 Simulations – Assessment

The Müller et al. (2017) procedure evaluates response to year-to-year temperature and precipitation variations in a control run driven by historical climate and compares it to detrended historical yields from the FAO (Food and Agriculture Organization of the United Nations, 2018) by calculating the Pearson product moment correlation coefficient. The procedure is sensitive to the detrending method and the area mask used to aggregate yields. Here we use a 5-year running mean removal and the MICRA area mask for aggregation. Sometimes the time series are shifted by one year to account for errors in FAO or model year reporting. The procedure offers no means of assessing CO_2 fertilization, since CO_2 has been relatively constant over the historical data collection period. Nitrogen introduces another source of uncertainty into the analysis, since the GGCMI Phase II runs impose fixed, uniform nitrogen application levels that are not realistic for individual countries. We evaluate up to three control runs for each model, since some modeling groups provide historical runs for three different nitrogen levels.

Results are similar to those of GGCMI Phase I, with reasonable fidelity at capturing year-over-year variation, with differences by region and crop stronger than difference between models. (That is, Figure A1 shows more similarity in horizontal than vertical bars.) No single model is dominant, with each model providing near best-in-class performance in at least one location-crop combination. For example, maize in the United States is consistently well-simulated while maize in Indonesia is problematic (mean Pearson correlation coefficients of 0.68 and 0.18, respectively). In some cases, especially in the developing world, low correlation coefficients may indicate not only model failure but also problems in FAO yield data. In general, correlation coefficients in GGCMI Phase II are slightly below those of Phase I, likely because of unrealistic nitrogen levels, lack of country level calibration in some models, and restriction to only the MICRA aggregation mask in this study. (Compare Figure A1 to Müller et al. (2017) Figures 1–4 and 6.) Note that in this methodology, simulations of crops with low year-to-year variability such as irrigated rice and wheat will tend to score more poorly than those with higher variability.

Some models do show particular strength for particular crops. For example, the EPIC family of models, and especially the EPIC-TAMU model, perform particularly well for soy across all regions. In other cases a model has particular strength in only certain crop and region combinations. For example, the strongest correlation coefficient in Figure A1 is that for the pDSSAT model for maize in the U.S. (the example crop-model-location used in many example figures in this paper), but pDSSAT slightly under performs for maize in other regions. These model assessment results are similar to those for GGCMI Phase I in Müller et al. (2017).

A2 Emulation – Assessment

No general criteria exist for defining an acceptable crop model emulator, so we present two different metrics. First, for a multi-model comparison exercise like GGCMI Phase II, one reasonable criterion is what we term the "normalized error", which compares the fidelity of an emulator for a given model and scenario to the inter-model uncertainty. We define the normalized

error e for each scenario as the difference between the fractional yield change from the emulator and that in the original simulation, divided by the standard deviation of the multi-model spread (Equations A1 and A2):

$$F_{scn.} = \frac{Y_{scn.} - Y_{baseline}}{Y_{baseline}} \tag{A1}$$

$$e_{scn.} = \frac{F_{em,\,scn.} - F_{sim,\,scn.}}{\sigma_{sim,\,scn.}} \tag{A2}$$

Here $F_{scn.}$ is the fractional change in a model's mean emulated or simulated yield from a defined baseline, in a certain setting or scenario (scn.) in C, T, W, and N space; $Y_{scn.}$ and $Y_{baseline}$ are the absolute emulated or simulated mean yields. The normalized error e is the difference between the emulated fractional change in yield and that actually simulated, normalized by σ_{sim}, the standard deviation in simulated fractional yields change $F_{sim,\,scn.}$ across all models. The emulator is fitted across all available simulation outputs for each grid cell, model, and crop, and then the error is calculated across the each of the simulation scenarios provided by all nine models (Figure S3).

This metric implies that emulation is generally satisfactory, with several distinct exceptions. Almost all model-crop combination emulators have normalized errors less than one over nearly all currently cultivated hectares (Figure A3), but some individual model-crop combinations are difficult to emulate (e.g. PROMET for rice and soy, JULES for soy and winter wheat, Figures S23-S24). Problems with emulating PROMET for rice and soy may have to do with the parametrization of the phenology for those crops which lengthens the growing season in some cases. Normalized errors for soy are somewhat higher across all models not because emulator fidelity is worse but because models agree more closely on yield changes for soy than for other crops (see Figure S18), lowering the denominator. Emulator performance often degrades in geographic locations where crops are not currently cultivated. For example, emulator performance may be satisfactory over cultivated areas for all crops, but uncultivated regions may show some problematic areas (Figure A3 shows a CARAIB model case, see also Figure S25).

This first assessment procedure is relatively forgiving for several reasons. First, each emulation is evaluated against the simulation actually used to train the emulator. Had we used a spline interpolation the error would necessarily be zero. Second, the performance metric scales emulator fidelity not by the magnitude of yield changes but by the inter-model spread in those changes. The normalized error e for a model depends not only on the fidelity of its emulator in reproducing a given simulation but on the particular suite of models considered in the intercomparison exercise. Where models differ more widely, the standard for emulators becomes less stringent. This effect is readily seen when comparing assessments of emulator performance in simulations at baseline CO_2 (Figure A2) with those at higher CO_2 levels (Figure S26) because models disagree on the magnitude of CO_2 fertilization. The rationale for this choice of assessment metric is to relate the fidelity of the emulation to an estimate of true uncertainty, which we take as the multi-model spread. We therefore do not provide a formal parameter uncertainty analysis, but note that the GGCMI Phase II dataset is well-suited to statistical exploration of emulation approaches and quantification of emulator fidelity. More rigorous emulator assessments that could be preformed in future work include: testing other statistical specifications including non-parametric models and calculating standard error on emulator parameters.

Table A1. Mean absolute error of emulator representation of a simulation as a percentage of baseline yield for the cross-validation process. A 3-fold stratified k-fold cross validation scheme is utilized where the model is trained on two-thirds of the data and validated on the held-out remaining third (repeated three times). The split does not represent a uniform number of samples in each location or in each model because simulation sampling extent in variable spaces is heterogeneous. The mean absolute error is then divided by the baseline yield for the control case in each grid cell. The calculation only includes grid cells with at least 1 % of surface area cultivated with a specific crop (approximately 1000 grid cells in each case). The table displays area weighted mean ('WM') shows the mean error weighted by hectares grown in each grid cell (Portmann et al., 2010) and 'MD' shows the unweighted median across grid cell values. * Indicates cases where the OLS linear model is unstable.

	Maize		**Soy**		**Rice**		**S. Wheat**		**W. Wheat**	
Model	WM (%)	MD (%)	WM (%)	MD (%)	WM (%)	MD (%)	WM (%)	MD (%)	WM (%)	MD (%)
CARAIB	0.00	1.71	0.02	2.39	0.03	2.95	0.02	4.40	0.01	2.36
EPIC-TAMU	0.00	4.30	0.01	6.24	0.00	3.35	0.01*	6.82*	0.01	3.51
JULES	0.11	6.13	0.01	10.2	0.01	6.97	0.04	15.1	NA	NA
GEPIC	0.00	5.78	0.00	3.75	0.01	5.64	0.01	6.76	0.01	7.01
LPJ-GUESS	0.00	1.78	NA	NA	NA	NA	0.05	6.22	0.02	3.35
LPJmL	0.00	9.44	0.00	3.25	0.01	8.37	0.01	9.83	0.01	4.98
pDSSAT	0.00	2.93	0.05	3.02	0.01	3.97	0.01	2.97	0.01	4.67
PROMET	0.01	4.19	0.00	6.03	0.01	9.85	0.01	7.04	0.01	3.68
PEPIC	0.00*	3.71*	0.00*	2.80*	0.00*	2.89*	0.00*	4.83*	0.02*	6.70*

We also provide a more stringent test of emulator performance; a three-fold cross validation. Here the training data is split and the model is trained on two thirds of the data and tested on the held out portion (the process is then repeated three times to cover all data in the training set). We normalize the error in each grid cell by dividing by the yield in that grid cell in the baseline (T+0, W+0, C=360, N=200) case and show aggregations by grid cell and weighted by area cultivated per grid cell. Errors are generally low as a percentage of yield –even for this strict protocol– and when weighted by area, essentially zero in most cases (Table A1). Note that the cross validation process often does not include edge simulations in the training set that are then predicted in the test phase. This extrapolation during cross validation is not realistic based on the actual use of the emulator.

Author contributions. J.E., C.M, A.R., J.F., and E.M. designed the research. C.M., J.J., J.B., P.C., M.D., P.F., C.F., L.F., M.H., C.I., I.J., C.J., N.K., M.K., W.L., S.O., M.P., T.P., A.R., X.W., K.W., and F.Z. performed the simulations. J.F., J.J., A.S., M.L., and E.M. performed the analysis and J.F. and E.M. prepared the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. We thank Michael Stein and Kevin Schwarzwald, who provided helpful suggestions that contributed to this work. This research was performed as part of the Center for Robust Decision-Making on Climate and Energy Policy (RDCEP) at the University of Chicago, and was supported through a variety of sources. RDCEP is funded by NSF grant #SES-1463644 through the Decision Making Under Uncertainty program. J.F. was supported by the NSF NRT program, grant #DGE-1735359. C.M. was supported by the MACMIT project (01LN1317A) funded through the German Federal Ministry of Education and Research (BMBF). C.F. was supported by the European Research Council Synergy grant #ERC-2013-SynG-610028 Imbalance-P. P.F. and K.W. were supported by the Newton Fund through the Met Office Climate Science for Service Partnership Brazil (CSSP Brazil). A.S. was supported by the Office of Science of the U.S. Department of Energy as part of the Multi-sector Dynamics Research Program Area. S.O. acknowledges support from the Swedish strong research areas BECC and MERGE together with support from LUCCI (Lund University Centre for studies of Carbon Cycle and Climate Interactions). R.C.I. acknowledges support from the Texas Agrilife Research and 634 Extension, Texas AM University. This is paper number 35 of the Birmingham Institute of Forest Research. Computing resources were provided by the University of Chicago Research Computing Center (RCC).

References

Alexander, L., Zhang, X., Peterson, T., Caesar, J., BA, G., Tank, A., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Rupa Kumar, K., Revadekar, J., Griffiths, G., Vincent, L., B. Stephenson, D., Burn, J., Aguilar, E., Brunet, M., and L. Vazquez-Aguirre, J.: Global Observed Changes in Daily Climate Extremes of Temperature and Precipitation, Journal of Geophysical Research, 111, https://doi.org/10.1029/2005JD006290, 2006.

Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S., and Ewert, F.: Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe, Agric. For. Meteorol., 170, 32 – 46, https://doi.org/10.1016/j.agrformet.2012.11.017, 2013.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J., Hatfield, J., Ruane, A., J. Boote, K., Thorburn, P., Rötter, R. P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A., Doltra, J., and Wolf, J.: Uncertainty in simulating wheat yields under climate change, Nature Climate Change, 3, 827–832, https://doi.org/10.1038/nclimate1916, 2013.

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., B. Lobell, D., Cammarano, D., A. Kimball, B., Ottman, M., W. Wall, G., White, J., Reynolds, M., D. Alderman, P., Prasad, P. V. V., Aggarwal, P., Anothai, J., Basso, B., Biernath, C., Challinor, A., De Sanctis, G., and Zhu, Y.: Rising temperatures reduce global wheat production, Nature Climate Change, 5, 143–147, https://doi.org/10.1038/nclimate2470, 2015.

Aulakh, M. S. and Malhi, S. S.: Interactions of Nitrogen with Other Nutrients and Water: Effect on Crop Yield and Quality, Nutrient Use Efficiency, Carbon Sequestration, and Environmental Pollution, Advances in Agronomy, 86, 341 – 409, https://doi.org/10.1016/S0065-2113(05)86007-9, 2005.

Balkovič, J., van der Velde, M., Skalský, R., Xiong, W., Folberth, C., Khabarov, N., Smirnov, A., Mueller, N. D., and Obersteiner, M.: Global wheat production potentials and management flexibility under the representative concentration pathways, Global and Planetary Change, 122, 107 – 121, https://doi.org/10.1016/j.gloplacha.2014.08.010, 2014.

Blanc, E.: Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models, Agricultural and Forest Meteorology, 236, 145 – 161, https://doi.org/10.1016/j.agrformet.2016.12.022, 2017.

Blanc, E. and Sultan, B.: Emulating maize yields from global gridded crop models using statistical estimates, Agricultural and Forest Meteorology, 214-215, 134 – 147, https://doi.org/10.1016/j.agrformet.2015.08.256, 2015.

Calvin, K., Patel, P., Clarke, L., Asrar, G., Bond-Lamberty, B., Cui, R. Y., Di Vittorio, A., Dorheim, K., Edmonds, J., Hartin, C., Hejazi, M., Horowitz, R., Iyer, G., Kyle, P., Kim, S., Link, R., McJeon, H., Smith, S. J., Snyder, A., Waldhoff, S., and Wise, M.: GCAM v5.1: representing the linkages between energy, water, land, climate, and economic systems, Geoscientific Model Development, 12, 677–698, https://doi.org/10.5194/gmd-12-677-2019, 2019.

Castruccio, S., McInerney, D. J., Stein, M. L., Liu Crouch, F., Jacob, R. L., and Moyer, E. J.: Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs, Journal of Climate, 27, 1829–1844, https://doi.org/10.1175/JCLI-D-13-00099.1, 2014.

Challinor, A., Watson, J., Lobell, D., Howden, S., Smith, D., and Chhetri, N.: A meta-analysis of crop yield under climate change and adaptation, Nature Climate Change, 4, 287 – 291, https://doi.org/10.1038/nclimate2153, 2014.

Chang, W., Stein, M., Wang, J., Kotamarthi, V., and Moyer, E.: Changes in Spatio-temporal Precipitation Patterns in Changing Climate Conditions, Journal of Climate, 29, https://doi.org/10.1175/JCLI-D-15-0844.1, 2016.

Conti, S., Gosling, J. P., Oakley, J. E., and O'Hagan, A.: Gaussian process emulation of dynamic computer codes, Biometrika, 96, 663–676, https://doi.org/10.1093/biomet/asp028, 2009.

de Wit, C.: Transpiration and crop yields, Verslagen van Landbouwkundige Onderzoekingen : 64.6, 1957.

Duncan, W.: SIMCOT: a simulation of cotton growth and yield, in: Proceedings of a Workshop for Modeling Tree Growth, Duke University, Durham, North Carolina, edited by Murphy, C., pp. 115–118, Durham, North Carolina, 1972.

Duncan, W., Loomis, R., Williams, W., and Hanau, R.: A model for simulating photosynthesis in plant communities, Hilgardia, pp. 181–205, https://doi.org/10.3733/hilg.v38n04p181, 1967.

Dury, M., Hambuckers, A., Warnant, P., Henrot, A., Favre, E., Ouberdous, M., and François, L.: Responses of European forest ecosystems to 21st century climate: assessing changes in interannual variability and fire intensity, iForest - Biogeosciences and Forestry, pp. 82–99, https://doi.org/10.3832/ifor0572-004, 2011.

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., and Foster, I.: The parallel system for integrating impact models and sectors (pSIMS), Environmental Modelling and Software, 62, 509–516, https://doi.org/10.1016/j.envsoft.2014.04.008, 2014.

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., Foster, I., Glotter, M., Heinke, J., Iizumi, T., Izaurralde, R. C., Mueller, N. D., Ray, D. K., Rosenzweig, C., Ruane, A. C., and Sheffield, J.: The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0), Geoscientific Model Development, 8, 261–277, https://doi.org/10.5194/gmd-2016-207, 2015.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmdd-8-10539-2015, 2016.

Ferrise, R., Moriondo, M., and Bindi, M.: Probabilistic assessments of climate change impacts on durum wheat in the Mediterranean region, Natural Hazards and Earth System Sciences, 11, 1293–1302, https://doi.org/10.5194/nhess-11-1293-2011, 2011.

Folberth, C., Gaiser, T., Abbaspour, K. C., Schulin, R., and Yang, H.: Regionalization of a large-scale crop growth model for sub-Saharan Africa: Model setup, evaluation, and estimation of maize yields, Agriculture, Ecosystems & Environment, 151, 21 – 33, https://doi.org/10.1016/j.agee.2012.01.026, 2012.

Food and Agriculture Organization of the United Nations: FAOSTAT Database, http://www.fao.org/faostat/en/home, 2018.

Fronzek, S., Pirttioja, N., Carter, T. R., Bindi, M., Hoffmann, H., Palosuo, T., Ruiz-Ramos, M., Tao, F., Trnka, M., Acutis, M., Asseng, S., Baranowski, P., Basso, B., Bodin, P., Buis, S., Cammarano, D., Deligios, P., Destain, M.-F., Dumont, B., Ewert, F., Ferrise, R., François, L., Gaiser, T., Hlavinka, P., Jacquemin, I., Kersebaum, K. C., Kollas, C., Krzyszczak, J., Lorite, I. J., Minet, J., Minguez, M. I., Montesino, M., Moriondo, M., Müller, C., Nendel, C., Öztürk, I., Perego, A., Rodríguez, A., Ruane, A. C., Ruget, F., Sanna, M., Semenov, M. A., Slawinski, C., Stratonovitch, P., Supit, I., Waha, K., Wang, E., Wu, L., Zhao, Z., and Rötter, R. P.: Classifying multi-model wheat yield impact response surfaces showing sensitivity to temperature and precipitation change, Agricultural Systems, 159, 209–224, https://doi.org/10.1016/j.agsy.2017.08.004, 2018.

Glotter, M., Elliott, J., McInerney, D., Best, N., Foster, I., and Moyer, E. J.: Evaluating the utility of dynamical downscaling in agricultural impacts projections, Proceedings of the National Academy of Sciences, 111, 8776–8781, https://doi.org/10.1073/pnas.1314787111, 2014.

Glotter, M., Moyer, E., Ruane, A., and Elliott, J.: Evaluating the Sensitivity of Agricultural Model Performance to Different Climate Inputs, Journal of Applied Meteorology and Climatology, 55, 151113145618 001, https://doi.org/10.1175/JAMC-D-15-0120.1, 2015.

Hank, T., Bach, H., and Mauser, W.: Using a Remote Sensing-Supported Hydro-Agroecological Model for Field-Scale Simulation of Heterogeneous Crop Growth and Yield: Application for Wheat in Central Europe, Remote Sensing, 7, 3934–3965, https://doi.org/10.3390/rs70403934, 2015.

Haugen, M., Stein, M., Moyer, E., and Sriver, R.: Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression, Journal of Climate, 31, 8573–8588, https://doi.org/10.1175/JCLI-D-17-0782.1, 2018.

He, W., Yang, J., Zhou, W., Drury, C., Yang, X., D. Reynolds, W., Wang, H., He, P., and Li, Z.-T.: Sensitivity analysis of crop yields, soil water contents and nitrogen leaching to precipitation, management practices and soil hydraulic properties in semi-arid and humid regions of Canada using the DSSAT model, Nutrient Cycling in Agroecosystems, 106, 201–215, https://doi.org/10.1007/s10705-016-9800-3, 2016.

Heady, E. O.: An Econometric Investigation of the Technology of Agricultural Production Functions, Econometrica, 25, 249–268, 1957.

Heady, E. O. and Dillon, J. L.: Agricultural production functions, Iowa State University Press, 1961.

Holden, P., Edwards, N., PH, G., Fraedrich, K., Lunkeit, F., E, K., Labriet, M., Kanudia, A., and F, B.: PLASIM-ENTSem v1.0: A spatiotemporal emulator of future climate change for impacts assessment, Geoscientific Model Development, 7, 433–451, https://doi.org/10.5194/gmd-7-433-2014, 2014.

Holzkämper, A., Calanca, P., and Fuhrer, J.: Statistical crop models: Predicting the effects of temperature and precipitation changes, Climate Research, 51, 11–21, https://doi.org/10.3354/cr01057, 2012.

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., Chenu, K., van Oosterom, E. J., Snow, V., Murphy, C., Moore, A. D., Brown, H., Whish, J. P., Verrall, S., Fainges, J., Bell, L. W., Peake, A. S., Poulton, P. L., Hochman, Z., Thorburn, P. J., Gaydon, D. S., Dalgliesh, N. P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F. Y., Wang, E., Hammer, G. L., Robertson, M. J., Dimes, J. P., Whitbread, A. M., Hunt, J., van Rees, H., McClelland, T., Carberry, P. S., Hargreaves, J. N., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., and Keating, B. A.: APSIM – Evolution towards a new generation of agricultural systems simulation, Environmental Modelling and Software, 62, 327 – 350, https://doi.org/doi.org/10.1016/j.envsoft.2014.07.009, 2014.

Howden, S. and Crimp, S.: Assessing dangerous climate change impacts on Australia's wheat industry, Modelling and Simulation Society of Australia and New Zealand, pp. 505–511, https://doi.org/-, 2005.

Iizumi, T., Nishimori, M., and Yokozawa, M.: Diagnostics of Climate Model Biases in Summer Temperature and Warm-Season Insolation for the Simulation of Regional Paddy Rice Yield in Japan, Journal of Applied Meteorology and Climatology, 49, 574–591, https://doi.org/10.1175/2009JAMC2225.1, 2010.

Ingestad, T.: Nitrogen and Plant Growth; Maximum Efficiency of Nitrogen Fertilizers, Ambio, 6, 146–151, 1977.

Izaurralde, R., Williams, J., Mcgill, W., Rosenberg, N., and Quiroga Jakas, M.: Simulating soil C dynamics with EPIC: Model description and testing against long-term data, Ecological Modelling, 192, 362–384, https://doi.org/10.1016/j.ecolmodel.2005.07.010, 2006.

Jägermeyr, J. and Frieler, K.: Spatial variations in crop growing seasons pivotal to reproduce global fluctuations in maize and wheat yields, Science Advances, 4, 4517, https://doi.org/10.1126/sciadv.aat4517, 2018.

Jagtap, S. S. and Jones, J. W.: Adaptation and evaluation of the CROPGRO-soybean model to predict regional yield and production, Agriculture, Ecosystems & Environment, 93, 73 – 85, https://doi.org/10.1016/S0167-8809(01)00358-9, 2002.

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., and Ritchie, J.: The DSSAT cropping system model, European Journal of Agronomy, 18, 235 – 265, https://doi.org/10.1016/S1161-0301(02)00107-7, 2003.

Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H. C. J., Herrero, M., Howitt, R. E., Janssen, S., Keating, B. A., Munoz-Carpena, R., Porter, C. H., Rosenzweig, C., and Wheeler, T. R.: Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science, Agricultural Systems, 155, 269 – 288, https://doi.org/doi.org/10.1016/j.agsy.2016.09.021, 2017.

Keating, B., Carberry, P., Hammer, G., Probert, M., Robertson, M., Holzworth, D., Huth, N., Hargreaves, J., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J., Silburn, M., Wang, E., Brown, S., Bristow, K., Asseng, S., Chapman, S., McCown, R.,

Freebairn, D., and Smith, C.: An overview of APSIM, a model designed for farming systems simulation, European Journal of Agronomy, 18, 267 – 288, https://doi.org/10.1016/S1161-0301(02)00108-9, 2003.

Kodra, E. and Ganguly, A.: Asymmetry of projected increases in extreme temperature distributions, Scientific reports, 4, 5884, https://doi.org/10.1038/srep05884, 2014.

Leeds, W. B., Moyer, E. J., and Stein, M. L.: Simulation of future climate under changing temporal covariance structures, Advances in Statistical Climatology, Meteorology and Oceanography, 1, 1–14, https://doi.org/10.5194/ascmo-1-1-2015, 2015.

Lindeskog, M., Arneth, A., Bondeau, A., Waha, K., Seaquist, J., Olin, S., and Smith, B.: Implications of accounting for land use in simulations of ecosystem carbon cycling in Africa, Earth System Dynamics, 4, 385–407, https://doi.org/10.5194/esd-4-385-2013, 2013.

Liu, J., Williams, J. R., Zehnder, A. J., and Yang, H.: GEPIC - modelling wheat yield and crop water productivity with high resolution on a global scale, Agricultural Systems, 94, 478 – 493, https://doi.org/10.1016/j.agsy.2006.11.019, 2007.

Liu, W., Yang, H., Folberth, C., Wang, X., Luo, Q., and Schulin, R.: Global investigation of impacts of PET methods on simulating crop-water relations for maize, Agricultural and Forest Meteorology, 221, 164 – 175, https://doi.org/10.1016/j.agrformet.2016.02.017, 2016a.

Liu, W., Yang, H., Liu, J., Azevedo, L. B., Wang, X., Xu, Z., Abbaspour, K. C., and Schulin, R.: Global assessment of nitrogen losses and trade-offs with yields from major crop cultivations, Science of The Total Environment, 572, 526 – 537, https://doi.org/10.1016/j.scitotenv.2016.08.093, 2016b.

Lobell, D. B. and Burke, M. B.: On the use of statistical models to predict crop yield responses to climate change, Agricultural and Forest Meteorology, 150, 1443 – 1452, https://doi.org/10.1016/j.agrformet.2010.07.008, 2010.

Lobell, D. B. and Field, C. B.: Global scale climate-crop yield relationships and the impacts of recent warming, Environmental Research Letters, 2, 014 002, https://doi.org/10.1088/1748-9326/2/1/014002, 2007.

MacKay, D.: Bayesian Interpolation, Neural Computation, 4, 415–447, https://doi.org/10.1162/neco.1992.4.3.415, 1991.

Makowski, D., Asseng, S., Ewert, F., Bassu, S., Durand, J., Martre, P., Adam, M., Aggarwal, P., Angulo, C., Baron, C., Basso, B., Bertuzzi, P., Biernath, C., Boogaard, H., Boote, K., Brisson, N., Cammarano, D., Challinor, A., Conijn, J., and Wolf, J.: Statistical Analysis of Large Simulated Yield Datasets for Studying Climate Effects, p. 1100, https://doi.org/10.13140/RG.2.1.5173.8328, 2015.

Mauser, W., Klepper, G., Zabel, F., Delzeit, R., Hank, T., Putzenlechner, B., and Calzadilla, A.: Global biomass production potentials exceed expected future demand without the need for cropland expansion, Nature Communications, 6, https://doi.org/10.1038/ncomms9946, 2015.

McDermid, S., Dileepkumar, G., Murthy, K., Nedumaran, S., Singh, P., Srinivasa, C., Gangwar, B., Subash, N., Ahmad, A., Zubair, L., and Nissanka, S.: Integrated assessments of the impacts of climate change on agriculture: An overview of AgMIP regional research in South Asia, Chapter in: Handbook of Climate Change and Agroecosystems, pp. 201–218, 2015.

Mistry, M. N., Wing, I. S., and De Cian, E.: Simulated vs. empirical weather responsiveness of crop yields: US evidence and implications for the agricultural impacts of climate change, Environmental Research Letters, 12, https://doi.org/10.1088/1748-9326/aa788c, 2017.

Moore, F. C., Baldos, U., Hertel, T., and Diaz, D.: New science of climate change impacts on agriculture implies higher social cost of carbon, Nature Communications, 8, https://doi.org/10.1038/s41467-017-01792-x, 2017.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S., Iizumi, T., Izaurralde, R. C., Jones, C., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T. A. M., Ray, D. K., Reddy, A., Rosenzweig, C., Ruane, A. C., Sakurai, G., Schmid, E., Skalsky, R., Song, C. X., Wang, X., de Wit, A., and Yang, H.: Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications, Geoscientific Model Development, 10, 1403–1422, https://doi.org/10.5194/gmd-10-1403-2017, 2017.

Nakamura, T., Osaki, M., Koike, T., Hanba, Y. T., Wada, E., and Tadano, T.: Effect of CO2 enrichment on carbon and nitrogen interaction in wheat and soybean, Soil Science and Plant Nutrition, 43, 789–798, https://doi.org/10.1080/00380768.1997.10414645, 1997.

O'Hagan, A.: Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering & System Safety, 91, 1290 – 1300, https://doi.org/10.1016/j.ress.2005.11.025, 2006.

Olin, S., Schurgers, G., Lindeskog, M., Wårlind, D., Smith, B., Bodin, P., Holmér, J., and Arneth, A.: Modelling the response of yields and tissue C:N to changes in atmospheric CO_2 and N management in the main wheat regions of western Europe, Biogeosciences, 12, 2489–2515, https://doi.org/10.5194/bg-12-2489-2015, 2015.

Osaki, M., Shinano, T., and Tadano, T.: Carbon-nitrogen interaction in field crop production, Soil Science and Plant Nutrition, 38, 553–564, https://doi.org/10.1007/BF00025019, 1992.

Osborne, T., Gornall, J., Hooker, J., Williams, K., Wiltshire, A., Betts, R., and Wheeler, T.: JULES-crop: a parametrisation of crops in the Joint UK Land Environment Simulator, Geoscientific Model Development, 8, 1139–1155, https://doi.org/10.5194/gmd-8-1139-2015, 2015.

Ostberg, S., Schewe, J., Childers, K., and Frieler, K.: Changes in crop yields and their variability at different levels of global warming, Earth System Dynamics, 9, 479–496, https://doi.org/10.5194/esd-9-479-2018, 2018.

Oyebamiji, O. K., Edwards, N. R., Holden, P. B., Garthwaite, P. H., Schaphoff, S., and Gerten, D.: Emulating global climate change impacts on crop yields, Statistical Modelling, 15, 499–525, https://doi.org/10.1177/1471082X14568248, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

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

Poppick, A., McInerney, D. J., Moyer, E. J., and Stein, M. L.: Temperatures in transient climates: Improved methods for simulations with evolving temporal covariances, Ann. Appl. Stat., 10, 477–505, https://doi.org/10.1214/16-AOAS903, 2016.

Porter et al. (IPCC): Food security and food production systems. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change., in: IPCC Fifth Assessment Report, edited by et al., C. F., pp. 485–533, Cambridge University Press, Cambridge, UK, 2014.

Portmann, F., Siebert, S., Bauer, C., and Doell, P.: Global dataset of monthly growing areas of 26 irrigated crops, https://doi.org/-, 2008.

Portmann, F., Siebert, S., and Doell, P.: MIRCA2000 - Global Monthly Irrigated and Rainfed Crop Areas around the Year 2000: A New High-Resolution Data Set for Agricultural and Hydrological Modeling, Global Biogeochemical Cycles, 24, GB1011, https://doi.org/10.1029/2008GB003435, 2010.

Pugh, T., Müller, C., Elliott, J., Deryng, D., Folberth, C., Olin, S., Schmid, E., and Arneth, A.: Climate analogues suggest limited potential for intensification of production on current croplands under climate change, Nature Communications, 7, 12 608, https://doi.org/10.1038/ncomms12608, 2016.

Räisänen, J. and Ruokolainen, L.: Probabilistic forecasts of near-term climate change based on a resampling ensemble technique, Tellus A: Dynamic Meteorology and Oceanography, 58, 461–472, https://doi.org/10.1111/j.1600-0870.2006.00189.x, 2006.

Ratto, M., Castelletti, A., and Pagano, A.: Emulation techniques for the reduction and sensitivity analysis of complex environmental models, Environmental Modelling & Software, 34, 1 – 4, https://doi.org/10.1016/j.envsoft.2011.11.003, 2012.

Razavi, S., Tolson, B. A., and Burn, D. H.: Review of surrogate modeling in water resources, Water Resources Research, 48, https://doi.org/10.1029/2011WR011527, 2012.

Roberts, M., Braun, N., R Sinclair, T., B Lobell, D., and Schlenker, W.: Comparing and combining process-based crop models and statistical models with some implications for climate change, Environmental Research Letters, 12, https://doi.org/10.1088/1748-9326/aa7f33., 2017.

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

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. A. M., Schmid, E., Stehfest, E., Yang, H., and Jones, J. W.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proceedings of the National Academy of Sciences, 111, 3268–3273, https://doi.org/10.1073/pnas.1222463110, 2014.

Rosenzweig, C., Ruane, A. C., Antle, J., Elliott, J., Ashfaq, M., Chatta, A. A., Ewert, F., Folberth, C., Hathie, I., Havlik, P., Hoogenboom, G., Lotze-Campen, H., MacCarthy, D. S., Mason-D'Croz, D., Contreras, E. M., Müller, C., Perez-Dominguez, I., Phillips, M., Porter, C., Raymundo, R. M., Sands, R. D., Schleussner, C.-F., Valdivia, R. O., Valin, H., and Wiebe, K.: Coordinating AgMIP data and models across global and regional scales for 1.5°C and 2.0°C assessments, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 376, https://doi.org/10.1098/rsta.2016.0455, 2018.

Ruane, A., I. Hudson, N., Asseng, S., Camarrano, D., Ewert, F., Martre, P., J. Boote, K., Thorburn, P., Aggarwal, P., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A., Doltra, J., Gayler, S., Goldberg, R., Grant, R., and Wolf, J.: Multi-wheat-model ensemble responses to interannual climate variability, Environmental Modelling and Software, 81, 86–101, https://doi.org/10.1016/j.envsoft.2016.03.008, 2016.

Ruane, A. C., Cecil, L. D., Horton, R. M., Gordon, R., McCollum, R., Brown, D., Killough, B., Goldberg, R., Greeley, A. P., and Rosenzweig, C.: Climate change impact uncertainties for maize in Panama: Farm information, climate projections, and yield sensitivities, Agricultural and Forest Meteorology, 170, 132 – 145, https://doi.org/10.1016/j.agrformet.2011.10.015, 2013.

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

Ruane, A. C., Goldberg, R., and Chryssanthacopoulos, J.: Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation, Agric. Forest Meteorol., 200, 233–248, https://doi.org/10.1016/j.agrformet.2014.09.016, 2015.

Ruane, A. C., Antle, J., Elliott, J., Folberth, C., Hoogenboom, G., Mason-D'Croz, D., Müller, C., Porter, C., Phillips, M. M., Raymundo, R. M., Sands, R., Valdivia, R. O., White, J. W., Wiebe, K., and Rosenzweig, C.: Biophysical and economic implications for agriculture of +1.5° and +2.0°C global warming using AgMIP Coordinated Global and Regional Assessments, Climate Research, 76, 17–39, https://doi.org/10.3354/cr01520, 2018.

Rubel, F. and Kottek, M.: Observed and projected climate shifts 1901-2100 depicted by world maps of the Köppen-Geiger climate classification, Meteorologische Zeitschrift, 19, 135–141, https://doi.org/10.1127/0941-2948/2010/0430, 2010.

Ruiz-Ramos, M., Ferrise, R., Rodríguez, A., Lorite, I., Bindi, M., Carter, T., Fronzek, S., Palosuo, T., Pirttioja, N., Baranowski, P., Buis, S., Cammarano, D., Chen, Y., Dumont, B., Ewert, F., Gaiser, T., Hlavinka, P., Hoffmann, H., Höhn, J., Jurecka, F., Kersebaum, K., Krzyszczak, J., Lana, M., Mechiche-Alami, A., Minet, J., Montesino, M., Nendel, C., Porter, J., Ruget, F., Semenov, M., Steinmetz, Z., Stratonovitch, P., Supit, I., Tao, F., Trnka, M., de Wit, A., and Rötter, R.: Adaptation response surfaces for managing wheat under perturbed climate and CO2 in a Mediterranean environment, Agricultural Systems, 159, 260 – 274, https://doi.org/10.1016/j.agsy.2017.01.009, 2018.

Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N.: Crop planting dates: an analysis of global patterns, Global Ecology and Biogeography, 19, 607–620, https://doi.org/10.1029/2009GB003765, 2010.

Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., A. M. Pugh, T., Rolinski, S., Schaphoff, S., Schmid, E., Wang, X., Schlenker, W., and Frieler, K.: Consistent negative response of US crops to high temperatures in observations and crop models, Nature Communications, 8, 13 931, https://doi.org/10.1038/ncomms13931, 2017.

Schewe, J., Gosling, S. N., Reyer, C., Zhao, F., Ciais, P., Elliott, J., Francois, L., Huber, V., Lotze, H. K., Seneviratne, S. I., van Vliet, M. T. H., Vautard, R., Wada, Y., Breuer, L., Büchner, M., Carozza, D. A., Chang, J., Coll, M., Deryng, D., de Wit, A., Eddy, T. D., Folberth, C., Frieler, K., Friend, A. D., Gerten, D., Gudmundsson, L., Hanasaki, N., Ito, A., Khabarov, N., Kim, H., Lawrence, P., Morfopoulos, C., Müller, C., Müller Schmied, H., Orth, R., Ostberg, S., Pokhrel, Y., Pugh, T. A. M., Sakurai, G., Satoh, Y., Schmid, E., Stacke, T., Steenbeek, J., Steinkamp, J., Tang, Q., Tian, H., Tittensor, D. P., Volkholz, J., Wang, X., and Warszawski, L.: State-of-the-art global models underestimate impacts from climate extremes, Nature Communications, 10, 1005–, https://doi.org/10.1038/s41467-019-08745-6, 2019.

Schlenker, W. and Roberts, M. J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change, Proceedings of the National Academy of Sciences, 106, 15 594–15 598, https://doi.org/10.1073/pnas.0906865106, 2009.

Sippel, S., Zscheischler, J., Heimann, M., Otto, F. E. L., Peters, J., and Mahecha, M. D.: Quantifying changes in climate variability and extremes: Pitfalls and their overcoming, Geophysical Research Letters, 42, 9990–9998, https://doi.org/10.1002/2015GL066307, 2015.

Snyder, A., Calvin, K. V., Phillips, M., and Ruane, A. C.: A crop yield change emulator for use in GCAM and similar models:Persephone v1.0, Accepted for publication in Geoscientific Model Development, pp. 1–42, https://doi.org/10.5194/gmd-2018-195, in open review, 2018.

Storlie, C. B., Swiler, L. P., Helton, J. C., and Sallaberry, C. J.: Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models, Reliability Engineering & System Safety, 94, 1735 – 1763, https://doi.org/10.1016/j.ress.2009.05.007, 2009.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Tebaldi, C. and Lobell, D. B.: Towards probabilistic projections of climate change impacts on global crop yields, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL033423, 2008.

von Bloh, W., Schaphoff, S., Müller, C., Rolinski, S., Waha, K., and Zaehle, S.: Implementing the Nitrogen cycle into the dynamic global vegetation, hydrology and crop growth model LPJmL (version 5.0), Geoscientific Model Development, 11, 2789–2812, https://doi.org/10.5194/gmd-11-2789-2018, 2018.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework, Proceedings of the National Academy of Sciences, 111, 3228–3232, https://doi.org/10.1073/pnas.1312330110, 2014.

White, J. W., Hoogenboom, G., Kimball, B. A., and Wall, G. W.: Methodologies for simulating impacts of climate change on crop production, Field Crops Research, 124, 357 – 368, https://doi.org/10.1016/j.fcr.2011.07.001, 2011.

Williams, K., Gornall, J., Harper, A., Wiltshire, A., Hemming, D., Quaife, T., Arkebauer, T., and Scoby, D.: Evaluation of JULES-crop performance against site observations of irrigated maize from Mead, Nebraska, Geoscientific Model Development, 10, 1291–1320, https://doi.org/10.5194/gmd-10-1291-2017, 2017.

Williams, K. E. and Falloon, P. D.: Sources of interannual yield variability in JULES-crop and implications for forcing with seasonal weather forecasts, Geoscientific Model Development, 8, 3987–3997, https://doi.org/10.5194/gmd-8-3987-2015, 2015.

Wolf, J. and Oijen, M.: Modelling the dependence of European potato yields on changes in climate and CO2, Agricultural and Forest Meteorology, 112, 217 – 231, https://doi.org/10.1016/S0168-1923(02)00061-8, 2002.

Wu, X., Vuichard, N., Ciais, P., Viovy, N., de Noblet-Ducoudré, N., Wang, X., Magliulo, V., Wattenbach, M., Vitale, L., Di Tommasi, P., Moors, E. J., Jans, W., Elbers, J., Ceschia, E., Tallec, T., Bernhofer, C., Grünwald, T., Moureaux, C., Manise, T., Ligne, A., Cellier, P., Loubet, B., Larmanou, E., and Ripoche, D.: ORCHIDEE-CROP (v0), a new process-based agro-land surface model: model description and evaluation over Europe, Geoscientific Model Development, 9, 857–873, https://doi.org/10.5194/gmd-9-857-2016, 2016.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J. L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., and Asseng, S.: Temperature increase reduces global yields of major crops in four independent estimates, Proc. Natl. Acad. Sci., 114, 9326–9331, https://doi.org/10.1073/pnas.1701762114, 2017.

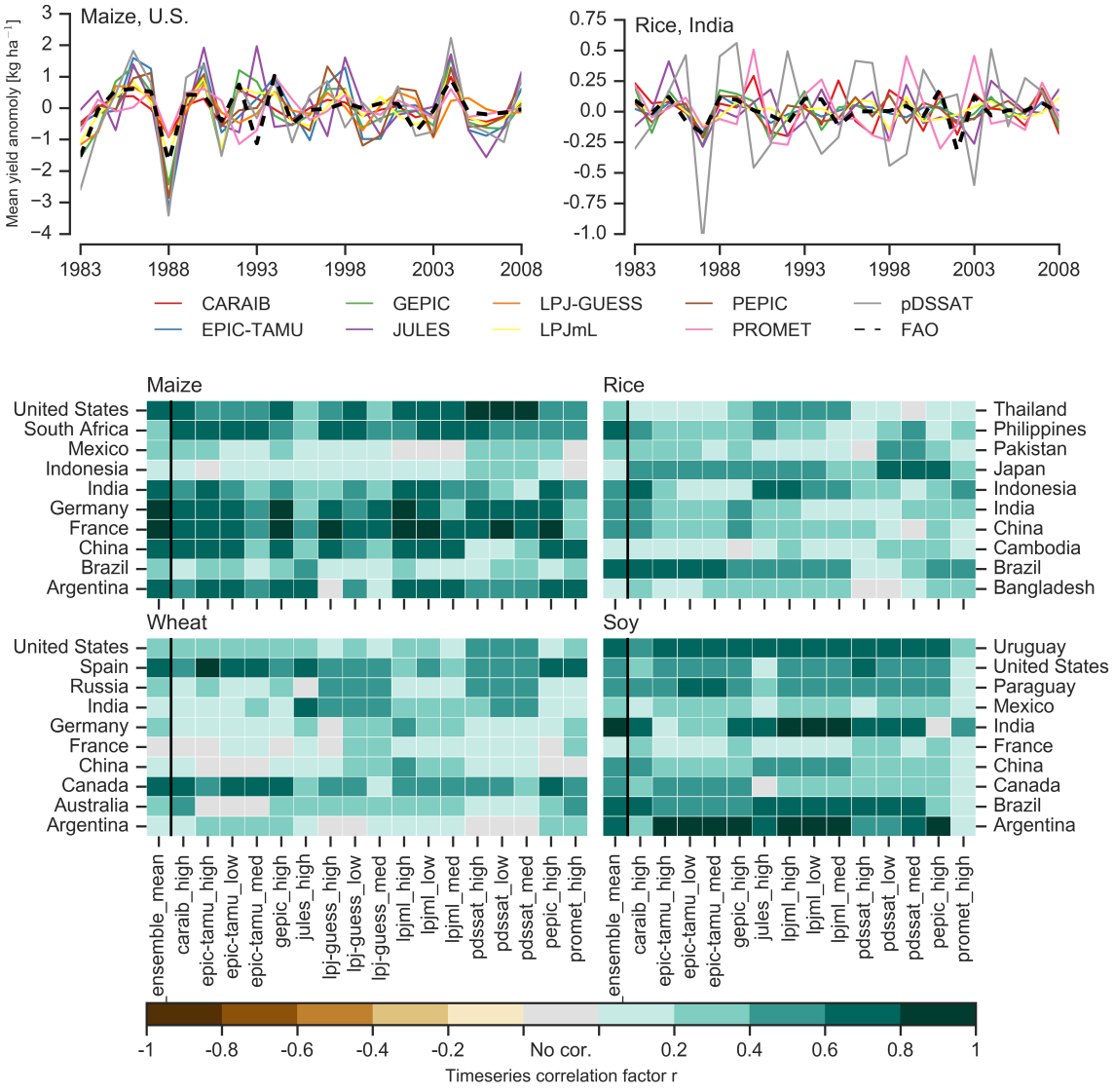

Figure A1. Time series of correlation coefficients between simulated crop yield and FAO data (Food and Agriculture Organization of the United Nations, 2018) at the country level. The top panels indicate two example cases: US maize (a good case), and rice in India (mixed case), both for the high nitrogen application case. The heatmaps illustrate the Pearson r correlation coefficient between the detrended simulation mean yield at the country level compared to the detrended FAO yield data for the top producing countries for each crop with continuous FAO data over the 1981-2010 period. Models that provided different nitrogen application levels are shown with low, med, and high label (models that did not simulate different nitrogen levels are analogous to a high nitrogen application level). The ensemble mean yield is also correlated with the FAO data (not the mean of the correlations). Wheat contains both spring wheat and winter wheat simulations where supplied, else one or the other (see Table 2). The Pearson r correlation coefficients are similar to those of GGCMI Phase I, with reasonable fidelity at capturing year-over-year variation, with differences by region and crop stronger than difference between models as indicated by more horizontal bars than vertical bars of the same color.

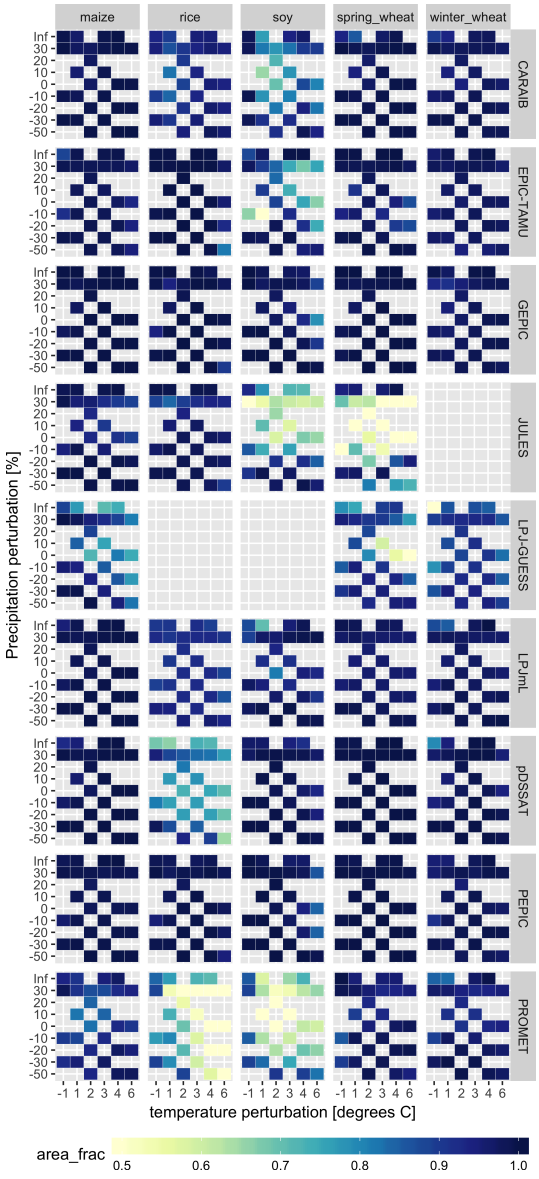

Figure A2. Assessment of emulator performance over currently cultivated areas based on normalized error (Equations A2, A1). We show performance of all 9 models emulated, over all crops and all sampled T and P inputs, but with CO_2 and nitrogen held fixed at baseline values. Large columns are crops and large rows models; squares within are T,P scenario pairs. Colors denote the fraction of currently cultivated hectares ('area frac') for each crop with normalized area e less that 1 indicating the the error between the emulation and simulation less that one standard deviation of the ensemble simulation spread. Of the 756 scenarios with these CO_2 and N values, we consider only those for which all 9 models submitted data (Figure S3). JULES did not simulate spring wheat and LPJ-GUESS did not simulate rice and soy. Emulator performance is generally satisfactory, with some exceptions. Emulator failures (significant areas of poor performance) occur for individual crop-model combinations, with performance generally degrading for hotter and wetter scenarios.

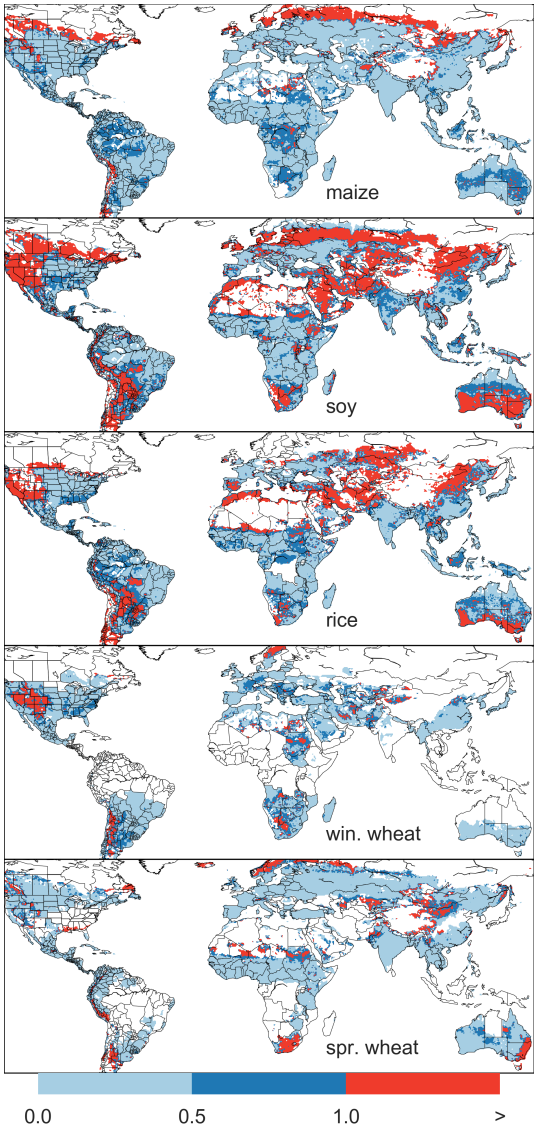

Figure A3. Illustration of our test of emulator performance, applied to the CARAIB model for the T+4 scenario for rainfed crops. Contour colors indicate the normalized emulator error e, where $e > 1$ means that emulator error exceeds the multi-model standard deviation. White areas are those where crops are not simulated by this model. Models differ in their areas omitted, meaning the number of samples used to calculate the multi-model standard deviation is not spatially consistent in all locations. Emulator performance is generally good relative to model spread in areas where crops are currently cultivated (compare to Figure 1) and in temperate zones in general; emulation issues occur primarily in marginal areas with low yield potentials. For CARAIB, emulation of soy is more problematic, as was also shown in Figure A2.