The GGCMI Phase II experiment: global crop yield responses to changes in carbon dioxide, temperature, water, and nitrogen levels

James Franke[a,b,*], Joshua Elliott[b,c], Christoph Müller[d], Alexander Ruane[e], Abigail Snyder[f], Jonas Jägermeyr[c,b,d,e], Juraj Balkovic[g,h], Philippe Ciais[i,j], Marie Dury[k], Pete Falloon[l], Christian Folberth[g], Louis François[k], Tobias Hank[m], Munir Hoffmann[n], Cesar Izaurralde[o,p], Ingrid Jacquemin[k], Curtis Jones[o], Nikolay Khabarov[g], Marian Koch[n], Michelle Li[b,1], Wenfeng Liu[r,i], Stefan Olin[s], Meridel Phillips[e,t], Thomas Pugh[u,v], Ashwan Reddy[o], Xuhui Wang[i,j], Karina Williams[l], Florian Zabel[m], Elisabeth Moyer[a,b]

[a] *Department of the Geophysical Sciences, University of Chicago, Chicago, IL, USA*
[b] *Center for Robust Decision-making on Climate and Energy Policy (RDCEP), University of Chicago, Chicago, IL, USA*
[c] *Department of Computer Science, University of Chicago, Chicago, IL, USA*
[d] *Potsdam Institute for Climate Impact Research, Leibniz Association (Member), Potsdam, Germany*
[e] *NASA Goddard Institute for Space Studies, New York, NY, United States*
[f] *Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA*
[g] *Ecosystem Services and Management Program, International Institute for Applied Systems Analysis, Laxenburg, Austria*
[h] *Department of Soil Science, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovak Republic*
[i] *Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, 91191 Gif-sur-Yvette, France*
[j] *Sino-French Institute of Earth System Sciences, College of Urban and Environmental Sciences, Peking University, Beijing, China*
[k] *Unité de Modélisation du Climat et des Cycles Biogéochimiques, UR SPHERES, Institut d'Astrophysique et de Géophysique, University of Liège, Belgium*
[l] *Met Office Hadley Centre, Exeter, United Kingdom*
[m] *Department of Geography, Ludwig-Maximilians-Universität, Munich, Germany*
[n] *Georg-August-University Göttingen, Tropical Plant Production and Agricultural Systems Modelling, Göttingen, Germany*
[o] *Department of Geographical Sciences, University of Maryland, College Park, MD, USA*
[p] *Texas Agrilife Research and Extension, Texas A&M University, Temple, TX, USA*
[q] *Department of Statistics, University of Chicago, Chicago, IL, USA*
[r] *EAWAG, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland*
[s] *Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden*
[t] *Earth Institute Center for Climate Systems Research, Columbia University, New York, NY, USA*
[u] *Karlsruhe Institute of Technology, IMK-IFU, 82467 Garmisch-Partenkirchen, Germany.*
[v] *School of Geography, Earth and Environmental Science, University of Birmingham, Birmingham, UK.*

Abstract

Concerns about food security under climate change have motivated efforts to better understand the future changes in yields by using detailed process-based models in agronomic sciences. Process-based models differ on many details affecting yields and considerable uncertainty remains in future yield projections. Phase II of the Global Gridded Crop Model Intercomparison (GGCMI), an activity of the Agricultural Model Intercomparison and Improvement Project (AgMIP), consists of a large simulation set with perturbations in atmospheric CO_2 concentrations, temperature, precipitation, and applied nitrogen inputs and constitutes a data-rich basis of projected yield changes across twelve models and five crops (maize, soy, rice, spring wheat, and winter wheat) using global gridded simulations. In this paper we present the simulation output database from Phase II of the GGCMI effort, a targeted experiment aimed at understanding the sensitivity to and interaction between multiple climate variables (as well as management) on yields, and illustrate some initial summary results from the model intercomparison project. We also present the construction of a simple "emulator or statistical representation of the simulated 30-year mean climatological output in each location for each crop and model. The emulator captures the response of the process-based models in a lightweight, computationally tractable form that facilitates model comparison as well as potential applications in subsequent modeling efforts such as integrated assessment.

Keywords: climate change, food security, model emulation, AgMIP, crop model

1. Introduction

Projecting crop yield response to a changing climate is of great importance, 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 with different crop or climate models predict sharp reduction in yields on currently cultivated cropland under business-as-usual climate scenarios, although their yield projections show considerable spread (e.g. Porter et al. (IPCC), 2014, Rosenzweig et al., 2014, Schauberger et al., 2017, and references therein). Model differences are unsurprising because crop responses in models can be complex, with crop growth a function of complex interactions between climate inputs and management practices.

Computational Models have been used to project crop yields since the 1950's, beginning with statistical models (Heady, 1957, Heady & Dillon, 1961) that attempt to capture the relationship between input factors and resultant yields. These statistical models were typically developed on a small scale for locations with extensive histories of yield data. The emergence of 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), Duncan et al. (1967) and attempted by Duncan (1972)). For a history of crop model development see the appendix of Rosenzweig et al. (2014). A half-century of improvement in both models and computing resources means that researchers can now run crop simulation models for many years at high 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 & 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 decreases in yield under future climate scenarios.

Process 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). 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 do not simulate many additional factors affecting production, including pests/diseases/weeds. For these reasons, individual studies must generally re-calibrate models to ensure that short-term predictions reflect current cultivar mixes, and long-term projections retain considerable uncertainty (Wolf & Oijen, 2002, Jagtap & 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 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 global future yield impacts of climate change for many reasons. First, cultivation may shift to new areas, where no yield data are currently available and therefore statistical models cannot apply. Yield data are also often limited in the de-

[*]Corresponding author at: 4734 S Ellis, Chicago, IL 60637, United States. email: jfranke@uchicago.edu

veloping world, where future climate impacts may be the most critical. Second, only process-based models can capture the growth response to elevated CO_2, novel conditions that are not represented in historical data (e.g. Pugh et al., 2016, Roberts et al., 2017). Similarly process-based models can represent novel changes in management practices (e.g. fertilizer input) that may ameliorate climate-induced damages.

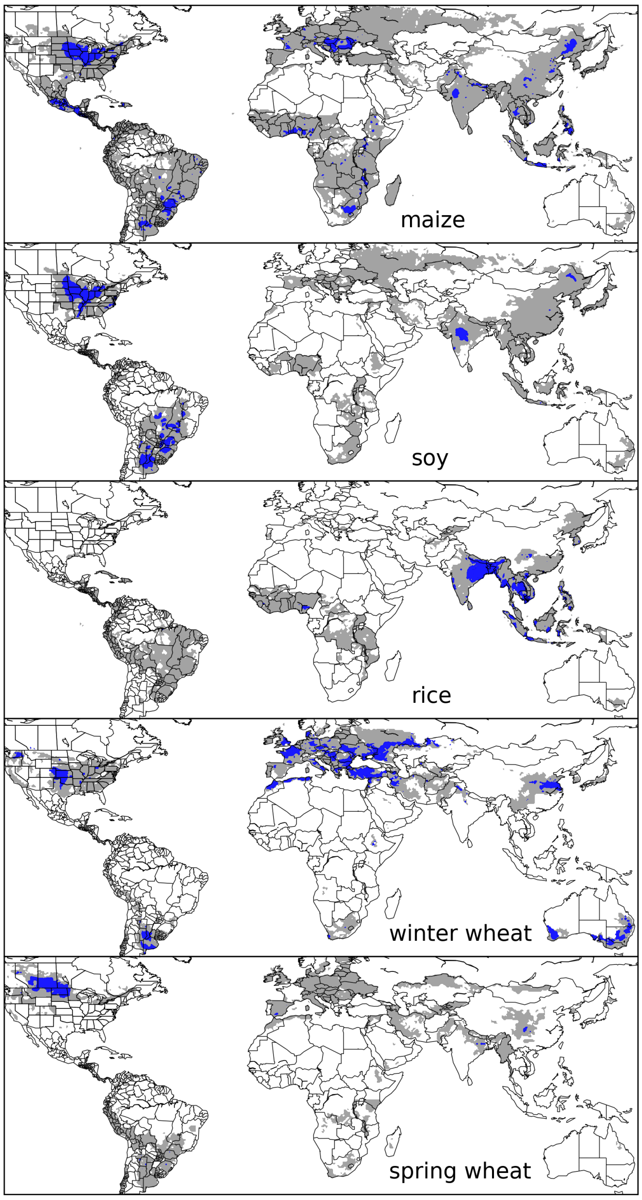

Figure 1: Presently cultivated area for rain-fed crops in the real world. Blue contour area indicates grid cells with more that 20,00 hectares (approx. 10% of the land in a grid cell near the equator for example) of crop cultivated. Gray contour shows area with more that 10 hectares cultivated. Cultivated areas for maize, rice, and soy are taken from the MIRCA2000 ("monthly irrigated and rain-fed 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 season. For analogous figure of irrigated crops, see Figure S1.

Statistical emulation of crop simulations has been used to combine advantageous features of both statistical and process-based models. The statistical representation of complicated numerical simulation (e.g. O'Hagan, 2006, Conti et al., 2009), in which simulation output acts as the training data for a statistical model, has been of increasing interest with the growth of simulation complexity and volume of output. Such emulators or "surrogate models" have 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 exploring crop yields in regions outside ranges of current cultivation and with input variables outside historical precedents, in a lightweight, flexible form that is compatible with economic studies.

In the past decade, many studies have developed emulators of crop yields from process-based models. Early studies proposing or describing potential emulators include Howden & Crimp (2005), Räisänen & Ruokolainen (2006) and Lobell & Burke (2010). In an early application, Ferrise et al. (2011) used a Artificial Neural Net trained on simulation outputs to predict wheat yields in the Mediterranean. Studies developing single-model emulators include Holzkämper et al. (2012) for the CropSyst model, Ruane et al. (2013) for the CERES wheat model, Oyebamiji et al. (2015) for the LPJmL model (for multiple crops, using multiple scenarios as a training set). In recent years, emulators have begun to be used in the context of multi-model intercomparisons, with Blanc & 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) (for maize, soy, wheat, and rice). Approaches differ: Blanc & Sultan

(2015) and Blanc (2017) used local weather variables (and CO_2 values) and yields but emulate across soil types using historical simulations and a future climate scenario (RCP8.5 over multiple climate models); Ostberg et al. (2018) used global mean temperature change (and CO_2) as regressors but pattern-scale to emulate local yields using multiple climate scenarios; Mistry et al. (2017) used local weather and yields and a historical simulation and compare with data.

ine 115 limitations of all existing studies, reason for parameter sweep - key paragraph. don't be negative as on line 131, don't end on negative. blame datasets, not authors of papers

A systematic parameter sweep offers advantages over analyses on small number of realistic scenario in which climate varies over time.

It allows highlighting the distinction between year-over-year and climatological changes, which can be different.

- removes correlation of the key variables which makes them difficult to disentangle in realistic scenarios

- provides fully stationary simulations

- large suite of simulations allows testing ability of emulator to reproduce yields by using only some in training set and testing ability to reproduce those exclude.

(maybe paragraph break here?)

Say that trend is increasing to use these - cite Frontzek and Snyder, respectively and Makowski and Pirttioja as "earlier efforts incude..".. Both Frontzek and Snyder do tempreature and water and Snyder adds Co2, 50 and 100? of simulations per model. Both take advantage of size of database to construct climatological mean emulators. however, both papers investigate just a few individual locations, Frontzek many models only wheat, Snyder multiple crops but only one model. In this paper we describe a new comprehensive dataset designed to expand this approach still further. GGCMI Phase II experiments provide global coverage, add nitrogen dimension (over 700 simulations), full suite of models. [all the acronym stuff]

might delete or move stuff from line 136, just say what GGCMI Phase II is

last para might be good as is, other than that this might be a dangerous ending. better say that it's tractable to emulation and resulting emulator can provide interpretable parameters and insight

present a simple climatological emulator as a potential tool for impacts assessments.

Line 276 - move caveats to later in this section- make separate caveat para to split it up. consider shortening.

Line 300 delete which are of most interest to impact modelers

what I would do is say that you make a climatological mean emulator, go immediately to saying "we test the necessity for this approach by using the GGCMI Phase II dataset to evaluate whether year-over-year responses are quantitatively distinct from climatological mean responses" -¿ these can be different for many reasons –¿ introduce Figure 3, say yes they are different, so we isolate the climatological signal of reponse to long-term perturbations by emulating on the mean yield for each scenario in the parameter sweep.

separate paragraph - note that we don't capture distributional

Input variable	**Abbr.**	**Tested range**	**Unit**
CO_2	C	360, 510, 660, 810	ppm
Temperature	T	-1, 0, 1, 2, 3, 4, 5*, 6	°C
Precipitation	W	-50, -30, -20, -10, 0, 10, 20, 30, (and W_{inf})	%
Applied nitrogen	N	10, 60, 200	kg ha^{-1}

Table 1: Phase II input variable test levels. Temperature and precipitation values indicate the perturbations from the historical, climatology. * Only simulated by one model. W-percentage does not apply to the irrigated (W_{inf}) simulations.

shift and if you wanted to do this use methods from A, B, or C. Leeds and Poppick are for temporal dependence and Matz is for marginal distribution in temperature. Can add Chang for precpitation.

Unclear where you cop to not having done a formal uncertainty assessment - in Methods? In conclusion as a suggestion for future work?

Conclusions - line 750-751 should be removed because you already did that.

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 parameters 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, summarize output and present initial results, demonstrate that it is tractable to emulation, and present a simple climatological emulator as a potential tool for impacts assessments.

2. Materials and Methods

2.1. GGCMI Phase II: experiment design

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 historical (1980-2010) 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 hundreds of scenarios in which individual climate or management inputs are 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 major goals of GGCMI Phase II are to:

- 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.
- 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.
- Illustrate differences in potential adaptation via growing season changes.

The guiding scientific rationale of GGCMI Phase II is to provide a comprehensive, systematic evaluation of the response of process-based crop models to different values for carbon dioxide, temperature, water, and applied nitrogen (collectively known as "CTWN"). Phase II of the GGCMI project consists of a series of simulations, each with one or more of the CTWN dimensions perturbed over the 31-year historical time series (1980-2010) used in Phase I. 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). Two models require sub-daily input data and use alternative sources. See Elliott et al. (2015) for additional details.

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 rain-fed agriculture and an additional 84 for irrigated (Table 1). For irrigated simulations, soil water is held at either field capacity or, for those models that include water-log damage, at maximum

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	Yes	37
CARAIB, Dury et al. (2011), Pirttioja et al. (2015)	X	X	X	X	X	No	224
EPIC-IIASA, Balkovi et al. (2014)	X	X	X	X	X	Yes	35
EPIC-TAMU, Izaurralde et al. (2006)	X	X	X	X	X	Yes	672
JULES*, Osborne et al. (2015), Williams & Falloon (2015), Williams et al. (2017)	X	X	X	–	X	No	224
GEPIC, Liu et al. (2007), Folberth et al. (2012)	X	X	X	X	X	Yes	384
LPJ-GUESS, Lindeskog et al. (2013), Olin et al. (2015)	X	–	–	X	X	Yes	672
LPJmL, von Bloh et al. (2018)	X	X	X	X	X	Yes	672
ORCHIDEE-crop, Valade et al. (2014)	X	–	X	–	X	Yes	33
pDSSAT, Elliott et al. (2014), Jones et al. (2003)	X	X	X	X	X	Yes	672
PEPIC, Liu et al. (2016a,b)	X	X	X	X	X	Yes	130
PROMET*†, Mauser & Bach (2015), Hank et al. (2015), Mauser et al. (2009)	X	X	X	X	X	Yes†	239
Totals	12	10	11	9	12	–	3993 (maize)

Table 2: Models included in the GGCMI Phase II and the number of C, T, W, and N simulations performed for rain-fed crops ("Sims per Crop"), with 672 as the maximum. "N-Dim." indicates if simulations include varying nitrogen levels; two models omit this dimension. All models provided the same set of simulations across all modeled crops, but some omitted individual crops in some cases. (For example, APSIM did not simulate winter wheat.) Irrigated simulations are provided at the level of the other covariates for each model (for an additional 84 simulations for the fully-sampled models). Geographic extent of simulation varies to some extent within a certain model for different scenarios (672 rain-fed simulations does not necessarily equal 672 climatological yields in all areas). This geographic variance only applies for areas far outside the area of currently cultivated crops. Two models (marked with *) use non-AgMERRA climate inputs. For further details on models, see Elliott et al. (2015). †PROMET provided simulations at only two nitrogen levels so is not emulated across the nitrogen dimension.

beneficial level. Temperature perturbations are applied as absolute offsets from the daily mean, minimum, and maximum temperature time series for each grid cell used as inputs. Precipitation perturbations are applied as fractional changes at the grid cell level, and carbon dioxide and nitrogen levels are specified as discrete values applied uniformly over all grid cells. Note that CO_2 changes are applied independently of changes in climate variables, so that higher CO_2 is not associated with higher temperatures. An additional, identical set of scenarios (at the same C, T, W, and N levels) simulate adaptive agronomy under climate change by varying the growing season for crop production. (These adaptation simulations are not shown or analyzed here.) The resulting GGCMI data set captures a distribution of crop responses over the potential space of future climate conditions.

The 12 models included in GGCMI Phase II are all mechanistic process-based crop models that are widely used in impacts assessments (Table 2). Although some of the models shares a common base (e.g. LPJmL and LPJ-GUESS and the EPIC models), they have developed independently from this shared base, for more details on the genealogy of the models see Figure S1 in Rosenzweig et al. (2014). Differences in model structure does mean that several key factors are not standardized across the experiment, including secondary soil nutrients, carry over effects across growing years including residue management and soil moisture, and extent of simulated area for different crops. Growing seasons are identical across models, but vary by crop and by location on the globe. All stresses except factors related to nitrogen, temperature, and water (e.g. Alkalinity, salinity) are disabled. No additional nitrogen inputs, such as atmospheric deposition, are considered, but some models have individual assumptions on soil organic matter that may release additional nitrogen through mineralization. See Rosenzweig et al. (2014), Elliott et al. (2015) and Müller et al. (2017) for further details on models and underlying assumptions.

Each model is run at 0.5 degree spatial resolution and covers all currently cultivated areas and much of the uncultivated land area. Coverage extends considerably outside currently cultivated areas because cultivation will likely shift under climate change. See Figure 1 for the present-day cultivated area of rain-fed crops, and Figure S1 in the supplemental material for irrigated crops. Some areas such as Greenland, far-northern Canada, Siberia, Antarctica, the Gobi and Sahara deserts, and central Australia are not simulated as they are assumed to remain non-arable even under an extreme climate change. Growing seasons are standardized across models with data adapted from several sources (Sacks et al., 2010, Portmann et al., 2008, 2010).

The participating modeling groups provide simulations at any of four initially specified levels of participation, so the number of simulations varies by model, with some sampling only a part of the experiment variable space. Most modeling groups simulate all five crops in the protocol, but some omitted one or more. Table 2 provides details of coverage for each model. Note that the three models that provide less than 50 simulations are excluded from the emulator analysis.

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. We show absolute yields in some cases to illustrate geographic differences in yields for a single model.

2.2. *Simulation model validation approach*

To verify the skill of the process-based models used, we repeat the validation exercises presented in Müller et al. (2017) for GGCMI Phase I. Note however that the GGCMI Phase II simulations are designed for evaluating changes in yield but not absolute yields, and so omit the calibrations used in predicting modeling to account for cultivar, pest loss, and management differences. The Phase II simulations also do not reproduce realistic nitrogen application levels for individual countries, since nitrogen is one of the parameters systematically varied. The Müller et al. (2017) validation 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 correlation coefficient. The procedure offers no means of assessing CO2 fertilization, since CO2 has been relatively constant over the historical data collection period. Nitrogen data are limited for many countries, and as mentioned the GGCMI Phase II runs impose fixed and uniform nitrogen application, introducing some uncertainty into the analysis. We evaluate one or more control runs for each model, since some modeling groups provide historical runs for three different nitrogen levels.

2.3. *Climatological-mean yield emulator design*

To demonstrate the properties of the GGCMI Phase II dataset, we construct an emulator of 30-year climatological mean yields, which are of most interest to impact modelers. This approach differs from previous studies of crop model emulation, which have typically emulated at the annual level. Annual emulation is required when the input training set consists of non-stationary projections of evolving yields (such as an RCP run). Recent studies (e.g. Fronzek et al., 2018, Snyder et al., 2018) that used a training set of stationary simulations with fixed variations in parameters allow emulating the climatological mean response instead. The two can differ for multiple reasons, including any year-to-year memory in the crop model, or if the distribution of growing-season daily temperatures associated with interannual variability is different from that associated with long-term CO2-driven changes. The confounding

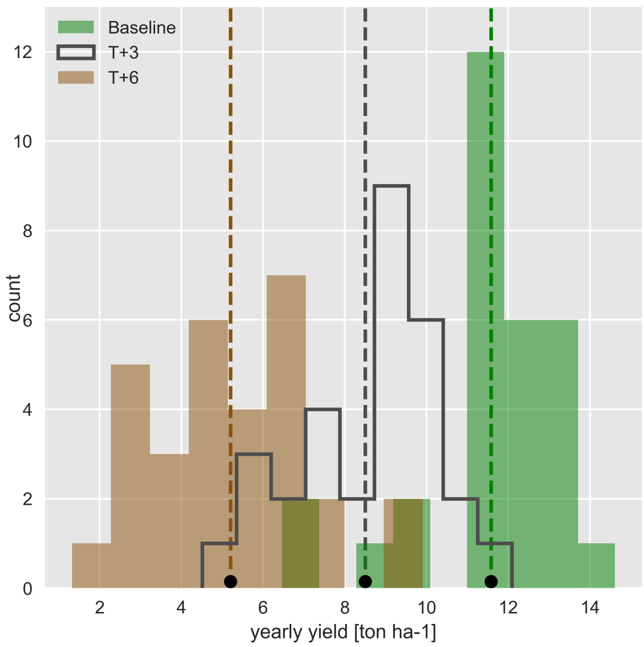

Figure 2: Example showing both climatological mean yields and distribution of yearly yields for three 30-year scenarios. Figure shows rain-fed maize for a grid cell in northern Iowa (a representative high-yield region) from the pDSSAT model, for the baseline climatology (1981-2010) and for scenarios with temperature shifted by (T) +3 and (T) +6 °C, with other variables held at baseline values. Dashed vertical lines and black dots indicate the climatological mean yield.

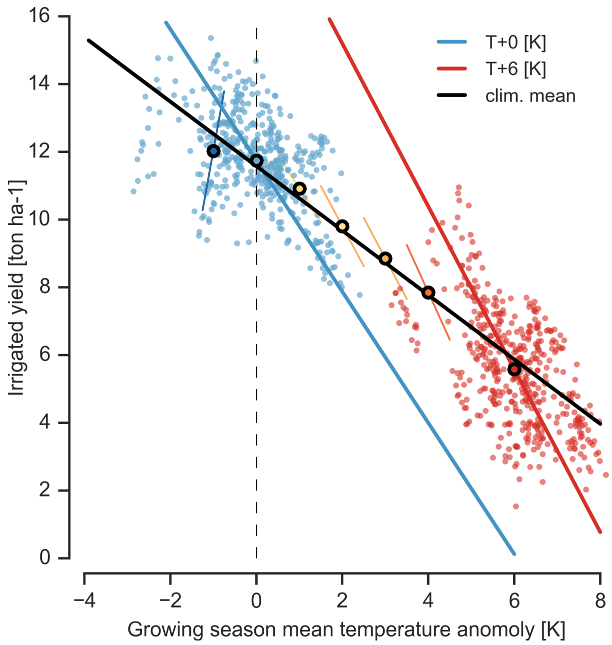

Figure 3: Example showing temperature relationship developed from year-to-year values vs. climatological mean values. Figure shows irrigated maize for nine adjacent grid cell in northern Iowa (a representative high-yield region) from the pDSSAT model, for the baseline climatology (1981-2010) and for scenarios with temperature shifted (T) +6 °C, with other variables held at baseline values. Irrigated yields are shown to control for precipitation effects. Blue and red lines indicate total least squares linear regression across each temperature scenario. Black ringed points indicate the climatological mean yield values for each climatological temperature scenario in the study (T-1, +0, +1, +2, +3, +4, +6 [K]). Short colored lines indicate slope of best fit (TLS) for year-to-year relationship at each climatalogical mean value. Bold black line indicates the fit (OLS) through the climatalogical mean values.

of year-to-year response with climatological response might be considered a class of Simpson's paradox. Crop yields in process based models do not respond to the mean growing season temperature, they respond the the full distribution in temperature over the growing season (or, specifically the exact temperature time series). Much of the variance is left unexplained if one tries to fit a statistical model between yields and some aggregate temperature variable (mean growing season temperature, monthly temperature etc.). Application of relationships obtained from such statistical models to mean changes in climate may provide problematic. The year-over-year yield response to individual factors in GGCMI Phase II do in fact often exceeds the climatological response (Figure 3). Note that the GGCMI Phase II datasets will not capture distributional shifts, because all simulations are run with fixed offsets from the historical climatology. (For methods to generate adjust historical climate data inclusive of distributional changes, see Haugen et al. (2018) and Poppick et al. (2016)). Emulation approaches are an area of active ongoing study and one of the goals of the GGCMI Phase II dataset is to facilitate these efforts.

Emulation involves fitting individual regression models for each crop, simulation model, and 0.5 degree geographic pixel from the GGCMI Phase II data set. The regressors are the applied constant perturbations in temperature, water, nitrogen and CO_2, we aggregate the simulation outputs in the time dimension, and regress on the 30-year mean yields. (See Figure 2 for illustration). The regression therefore omits information about yield responses to year-to-year climate perturbations, which are more complex. Emulating inter-annual yield variations would likely require considering statistical details of the historical climate time series, including changes in marginal distribution and temporal dependencies. (Future work should explore this). The climatological emulation indirectly includes any yield response to geographically distributed factors such as soil type, insolation, and the baseline climate itself, because we construct separate emulators for each grid cell. The emulator parameter matrices are portable and the yield computations are cheap even at the half-degree grid cell resolution, so we do not aggregate in space at this time.

We regress climatological mean yields against a third-order polynomial in C, T, W, and N with interaction terms. 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 & 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 & Field (2007) and Tebaldi & 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 & Malhi, 2005), and between nitrogen and carbon dioxide (Osaki et al., 1992, Nakamura et al., 1997). 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 more complex functional forms and larger numbers of parameters, e.g. 39 in Blanc & Sultan (2015) and Blanc (2017), who borrow information across space by fitting grid points simultaneously across a large region in a panel regression. We choose a simpler emulation at grid-cell level to avoid the requirement of assuming responses are uniform across space and to maximize interpretability.

The limited GGCMI variable sample space means that use of the full polynomial expression described above, which has

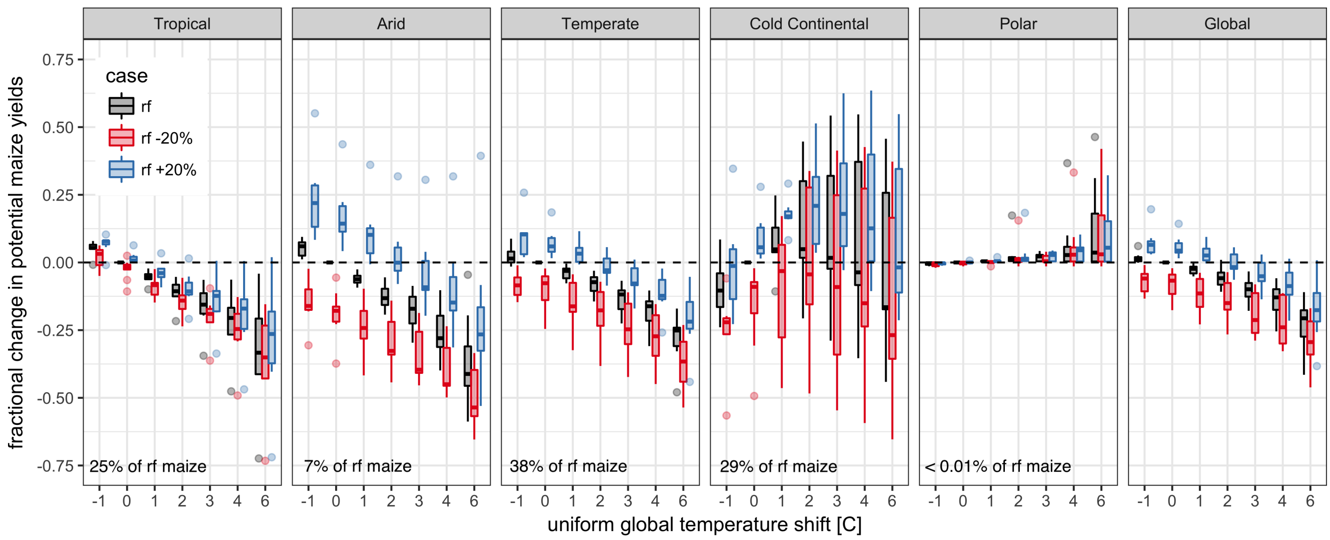

Figure 4: Illustration of the distribution of regional yield changes across the multi-model ensemble, split by Köppen-Geiger climate regions (Rubel & Kottek (2010)). We show responses of a single crop (rain-fed maize) to applied uniform temperature perturbations, for three discrete precipitation perturbation levels (rain-fed (rf) -20%, rain-fed (0), and rain-fed +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. The figure shows all modeled land area; see Figure S6 in the supplemental material for only currently-cultivated land. Panel text gives the percentage of rain-fed maize presently cultivated in each climate zone (data from Portmann et al., 2010). Box-and-whiskers plots show distribution across models, with median marked. Box edges are first and third quartiles, i.e. box height is the interquartile range (IQR). Whiskers extend to maximum and minimum of simulations but are limited at 1.5·IQR; otherwise the outlier is shown. Models generally agree in most climate regions (other than cold continental), with projected changes larger than inter-model variance. 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. Inter-model variance increases in the case where precipitation is reduced, suggesting uncertainty in model response to water limitation. The right panel with global changes shows yield responses to an globally uniform temperature shift; note that these results are not directly comparable to simulations of more realistic climate scenarios with the same global mean change.

34 terms for the rain-fed case (12 for irrigated), can be problematic, and can lead to over-fitting and unstable parameter estimations. 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 model; we then follow the reduction of the the aggregate mean squared error with increasing terms and eliminate those terms that do not contribute significant reductions. See supplemental documents for more details. We select terms by applying the feature selection process to the three models that provided the complete set of 672 rain-fed simulations (pDSSAT, EPIC-TAMU, and LPJmL); the resulting choice of terms is then applied for all emulators.

Feature importance is remarkably consistent across all three models and across all crops (see Figure S4 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 crops:

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

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 as the linear least squares fails to provide a stable result in many cases. 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. Because 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 130 (for the PEPIC model). We use the implementation of the Bayesian Ridge estimator from the scikit-learn package in Python (Pedregosa et al., 2011).

The resulting parameter matrices for all crop model emulators are available on request, as are the raw simulation data and a Python application to emulate yields. The yield output for a single GGCMI 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.

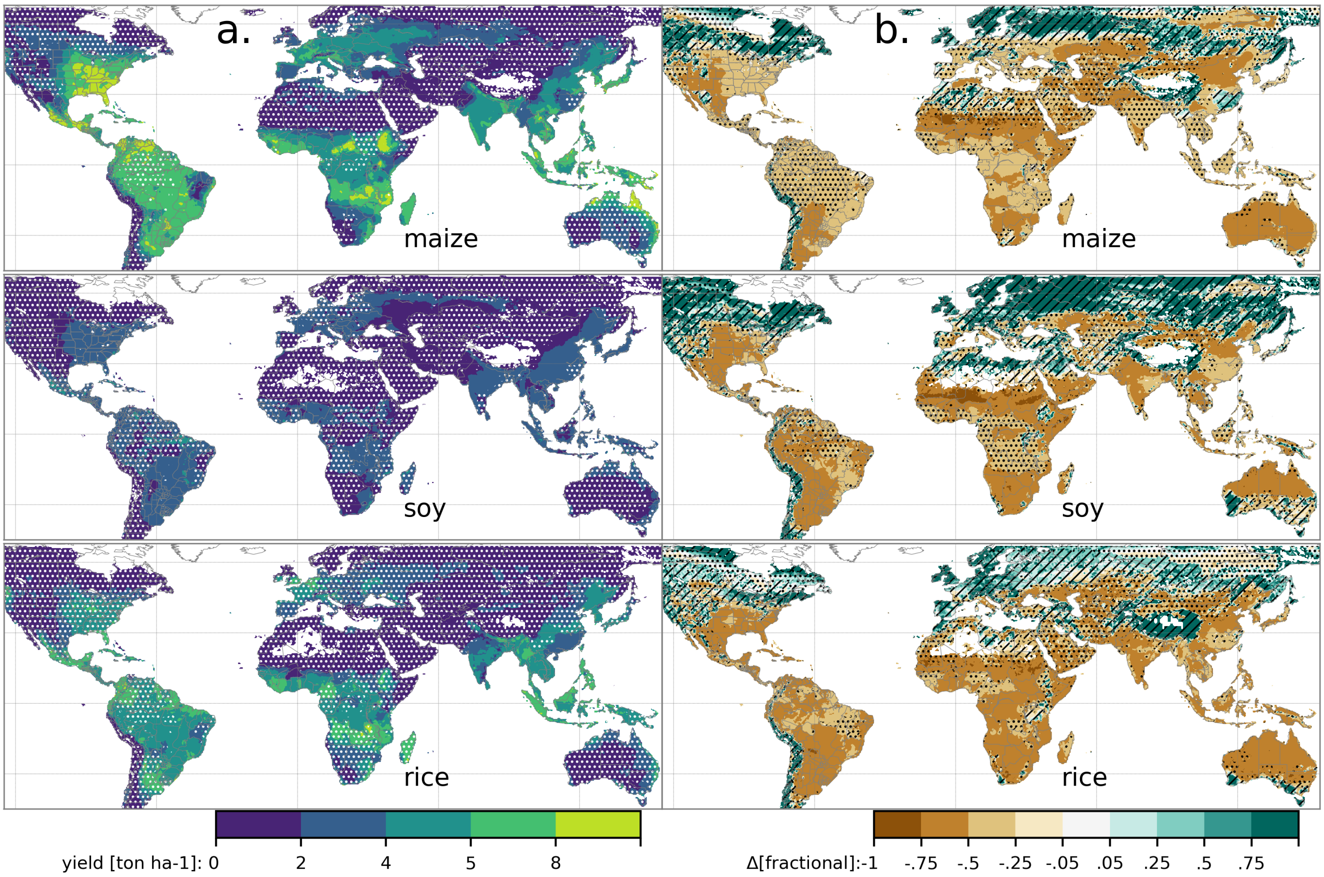

Figure 5: Illustration of the spatial pattern of potential yields and potential yield changes in the GGCMI Phase II ensemble, for three major crops. Left column (a) shows multi-model mean climatological yields for the baseline scenario for (top–bottom) rain-fed maize, soy, and rice. (For wheat see Figure S11 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.

2.4. Emulator evaluation

Because no general criteria exist for defining an acceptable model emulator, we develop a metric of emulator performance specific to GGCMI. For a multi-model comparison exercise like GGCMI, a 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 2 and 3):

$$F_{scn.} = \frac{Y_{scn.} - Y_{baseline}}{Y_{baseline}} \tag{2}$$

$$e_{scn.} = \frac{F_{em,\, scn.} - F_{sim,\, scn.}}{\sigma_{sim,\, scn.}} \tag{3}$$

Here $F_{scn.}$ is the fractional change in a model's mean emulated or simulated yield from a defined baseline, in some 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 $F_{sim,\, scn.}$ across all models. The emulator is fit across all available simulation outputs, and then the error is calculated across the simulation scenarios provided by all nine models (Figure 9 and Figures

S12 and Figures S13 in supplemental documents). Note that 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. The rationale for this choice is to relate the fidelity of the emulation to an estimate of true uncertainty, which we take as the multi-model spread.

3. Results

3.1. Simulation results

Crop models in the GGCMI ensemble show a 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 rain-fed maize in Figure 4, which shows yields for the primary Köppen-Geiger climate regions (Rubel & Kottek, 2010). In warming scenarios, models show decreases in maize yield in the temperate, tropical, and arid regions that account for nearly three-quarters of global maize production. These impacts are robust for even moderate climate perturbations. In the temperate zone, even a 1 degree temperature rise with other variables held fixed leads to a median yield reduction that outweighs the variance across models. A 6 degree temperature rise results in median loss of ~25% of yields with a signal to noise of nearly three. A notable exception is the cold continental region, where models disagree strongly, extending even to the sign of impacts. Model simulations of other crops produce similar responses to warming, with robust yield losses in warmer locations and high inter-model variance in the cold continental regions (Figures S7).

The effects of rainfall changes on maize yields are also as expected and are consistent across models. Increased rainfall mitigates the negative effect of higher temperatures, most strongly in arid regions. Decreased rainfall amplifies yield losses and also increases inter-model variance more strongly, suggesting that models have difficulty representing crop response to water stress. We show only rain-fed maize here; see Figure S5 for the irrigated case. As expected, irrigated crops are more resilient to temperature increases in all regions, especially so where water is limiting.

Mapping the distribution of baseline yields and yield changes shows the geographic dependencies that underlie these results. Figure 5 shows baseline and changes in the T+4 scenario for rain-fed maize, soy, and rice in the multi-model ensemble mean, with locations of model agreement marked.Absolute yield potentials are have strong spatial variation, with much of the Earth's surface area unsuitable for any given crop. 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. For wheat crops see Figure S11; wheat projections are both more uncertain and show fewer areas of increased yield in the inter-model mean.

3.2. Simulation model validation results

Figure 6 shows the Pearson time series correlation between the simulation model yield and FOA yield data. Figure 6 can be compared to Figures 1,2,3,4 and 6 in Müller et al. (2017). The results are mixed, with many regions for rice and wheat being difficult to model. No single model is dominant, with each model providing near best-in-class performance in at least one location-crop combination. The presence of very few vertical dark green color bars clearly illustrates the power of a multi-model intercomparison project like the one presented here. The ensemble mean does not beat the best model in each case, but shows positive correlation in over 75% of the cases presented here. The EPIC-TAMU model performs best for soy, CARIAB, EPIC-TAMU, and PEPIC perform best for maize, PROMET

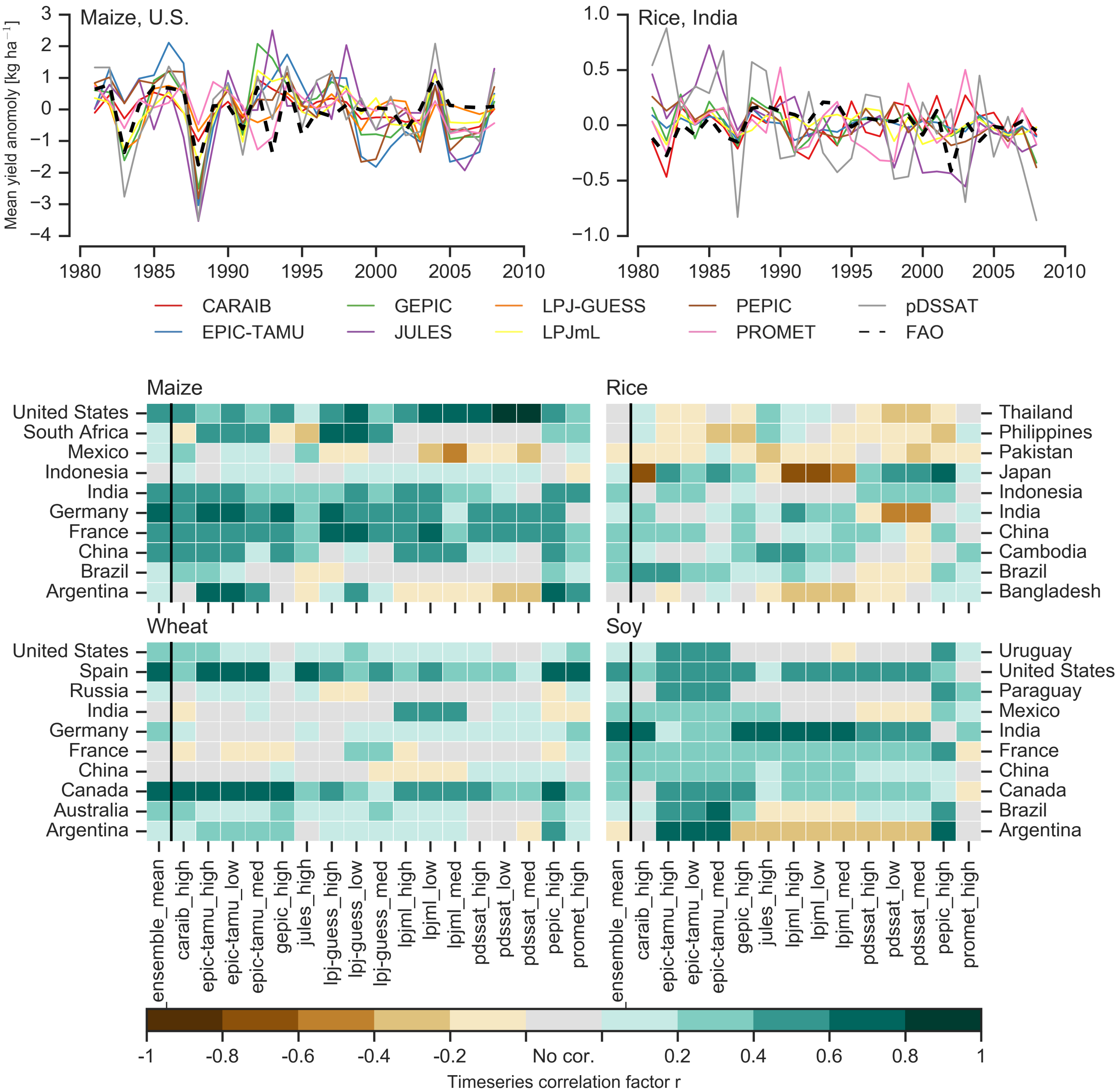

Figure 6: Time series correlation coefficients between simulated crop yield and FAO data (Food and Agriculture Organization of the United Nations, 2018) at the country level. The top panels indicate two example cases: US maize (a good case), and rice in India (mixed case), both for the high nitrogen 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 1980-2010 period. Models that provided different nitrogen application levels are shown with low, med, and high label (models that did not simulated 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.

performs best for wheat, and the EPIC family of models perform best for rice. Reductions in skill over the performance illustrated in Müller et al. (2017) can be attributed to the nitrogen levels or lack of calibration in some models.

*** or harmonization *** Christoph

Soy is qualitatively the easiest crop to represent (except in Argentina), which is likely due in part to the invariance of the response to nitrogen application (soy fixes atmospheric nitrogen very efficiently). Comparison to the FAO data is therefore easier than the other crops because the nitrogen application levels do not matter. US maize has the best performance across models, with nearly every model representing the historical variability to a reasonable extent. Especially good example years for US maize are 1983, 1988, and 2004 (top left panel of Figure 6), where every model gets the direction of the anomaly compared to surrounding years correct. 1983 and 1988 are famously bad years for US maize along with 2012 (not shown). US maize is possibly both the most uniformly industrialized (in terms of management practices) crop and the one with the best data collection in the historical period of all the cases presented here.

The FAO data is at least one level of abstraction from ground truth in many cases, especially in developing countries. The failure of models to represent the year-to-year variability in rice in some countries in southeast Asia is likely partly due to model failure and partly due to lack of data. It is possible to speculate that the difference in performance between Pakistan (no successful models) and India (many successful models) for rice may reside at least in part in the FAO data and not the models themselves. The same might apply to Bangladesh and India for rice. Partitioning of these contributions is impossible at this stage. Additionally, there is less year-to-year variability in rice yields (partially due to the fraction of irrigated cultivation). Since the Pearson r metric is scale invariant, it will tend to score the rice models more poorly than maize and soy. An example of very poor performance can be seen with the pDSSAT model for rice in India (top right panel of Figure 6).

3.3. *Emulator performance*

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, however, when crop yield responses are sufficiently smooth and continuous to allow fitting with a relatively simple functional form. In the GGCMI simulations, this condition largely but not always holds. Responses are quite diverse across locations, crops, and models, but in most cases local responses are regular enough to permit emulation. Figure 7 illustrates the geographic diversity of responses even in high-yield areas for a single crop and model (rain-fed maize in pDSSAT for various high-cultivation areas). This heterogeneity validates the choice of emulating at the grid cell level.

Each panel in Figure 7 shows model 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; in all cases yields evolve smoothly across the space sampled. For reference we show the results of the full emulation fitted across the parameter space. The polynomial fit readily captures the climatological response to perturbations.

Crop yield responses generally follow similar functional forms across models, though with a spread in magnitude. Figure 8 illustrates the inter-model diversity of yield responses to the same perturbations, even for a single crop and location (rain-fed maize in northern Iowa, the same location shown in the Figure 7). The differences make it important to construct emulators separately for each individual model, and the fidelity of emulation can also differ across models. This figure illustrates a common phenomenon, that models differ more in response to perturbations in CO_2 and nitrogen perturbations than to those in temperature or precipitation. (Compare also Figures 4 and S18.) For this location and crop, CO_2 fertilization effects can range from ~5–50%, and nitrogen responses from nearly flat to a 60% drop in the lowest-application simulation.

While the nitrogen dimension is important and uncertain, it is also the most problematic to emulate in this work because of its limited sampling. The GGCMI 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 de-

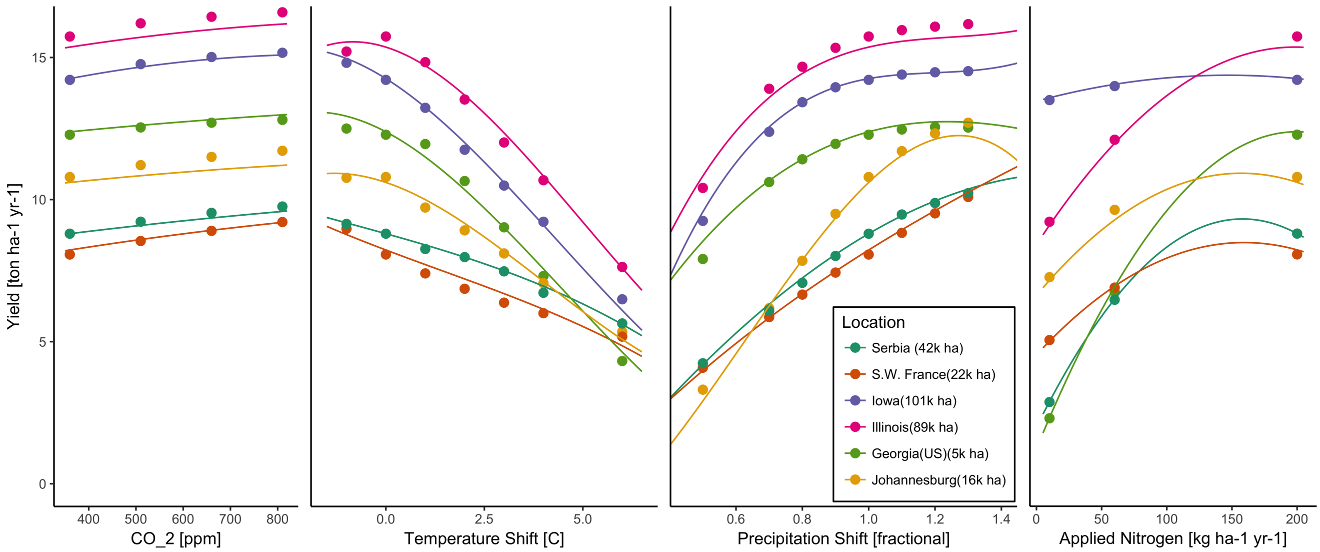

Figure 7: Example illustrating spatial variations in yield response and emulation ability. We show rain-fed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares cultivated in each selected grid cell. Each panel shows variation along a single variable, with others held at baseline values. Dots show climatological mean yield, solid lines a simple polynomial fit across this 1D variation, and dashed lines the results of the full emulator of Equation 1. The climatological response is sufficiently smooth to be be represented by a simple polynomial. Because precipitation changes are imposed as multiplicative factors rather than additive offsets, locations with higher baseline precipitation have a larger absolute spread in inputs sampled than do drier locations (not visible here).

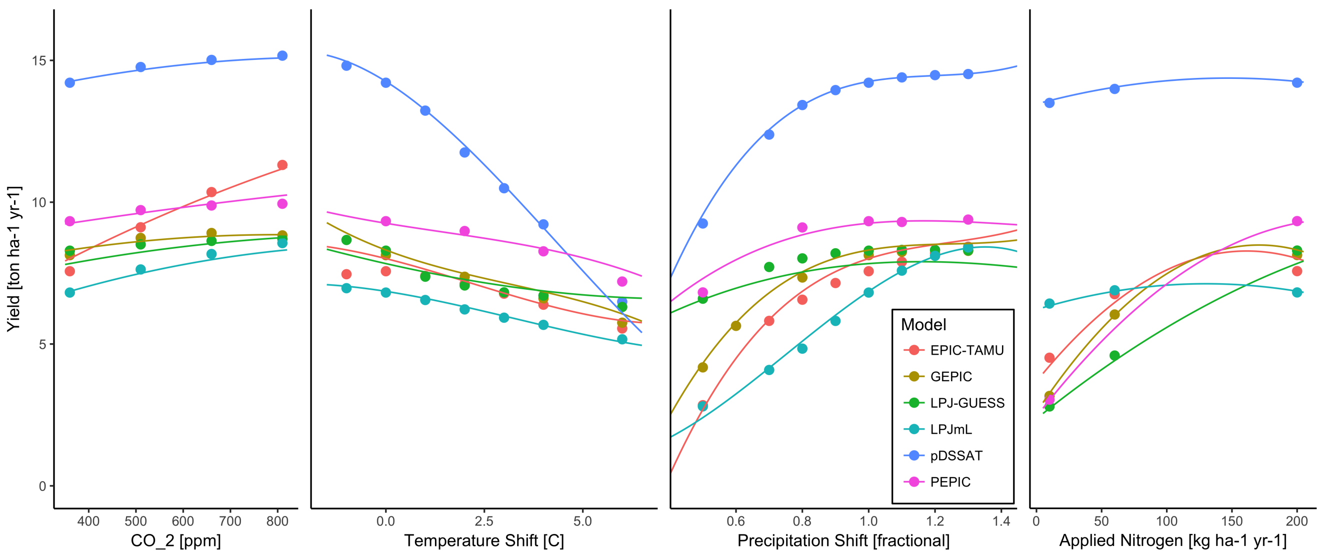

Figure 8: Example illustrating across-model variations in yield response. We show simulations and emulations from six models for rain-fed maize in the same Iowa location shown in Figure 7, with the same plot conventions. Models that did not simulate the nitrogen dimension are omitted for clarity. The inter-model standard deviation is larger for the perturbations in CO_2 and nitrogen than for those in temperature or precipitation illustrating that the sensitivity to weather is better constrained than to management at elevated CO_2. While most model responses can readily emulated with a simple polynomial, some response surfaces are more complicated and lead to emulation error.

clines in yield with lower nitrogen levels means that some regressions imply a peak in yield between the 100 and 200 kg N y^{-1} ha^{-1} levels. While there may be some reason to believe over-application of nitrogen at the wrong time in the growing season could lead to reduced yields, these features are almost certainly 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.

To assess the ability of the polynomial emulation to capture

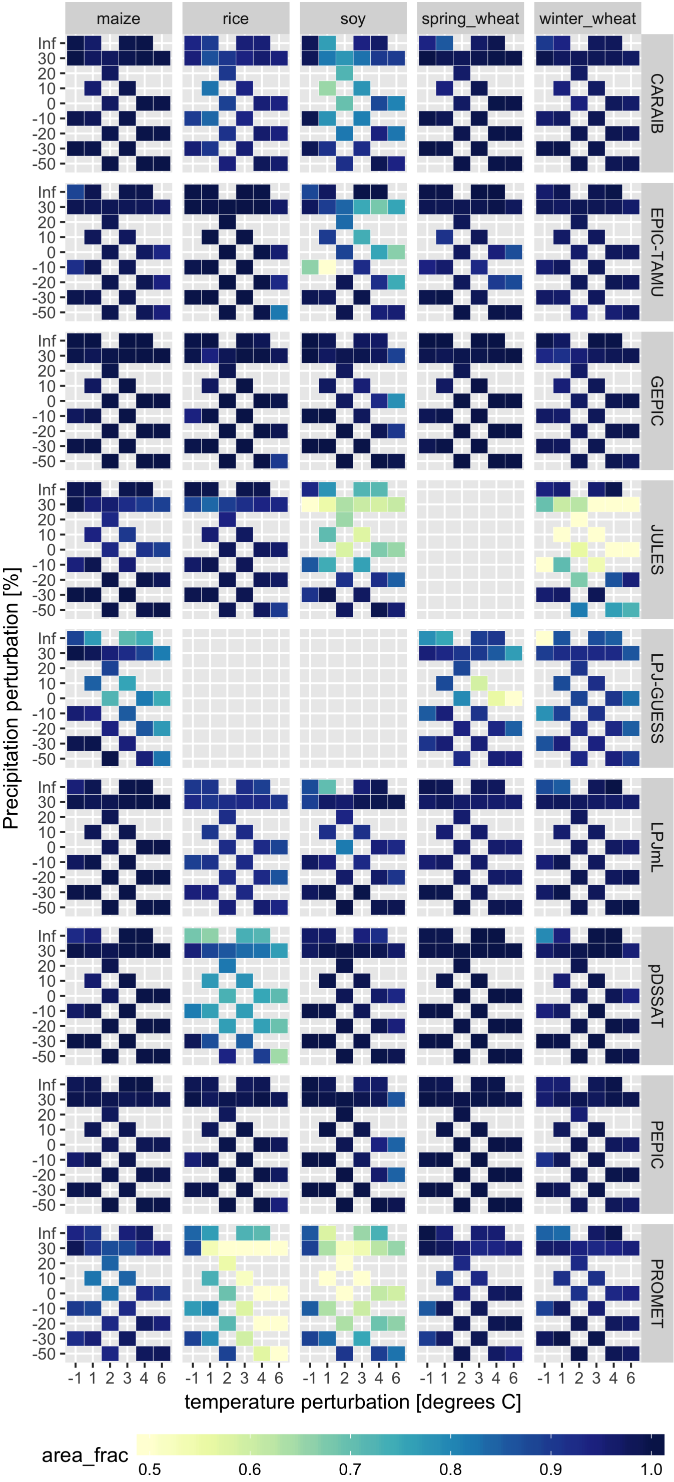

Figure 9: Assessment of emulator performance over currently cultivated areas based on normalized error (Equations 3, 2). We show performance of all 9 models emulated, over all crops and all sampled T and P inputs, but with CO_2 and nitrogen held fixed at baseline values. Large columns are crops and large rows models; squares within are T,P scenario pairs. Colors denote the fraction of currently cultivated hectares with $e < 1$. Of the 756 scenarios with these CO_2 and N values, we consider only those for which all 9 models submitted data. 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.

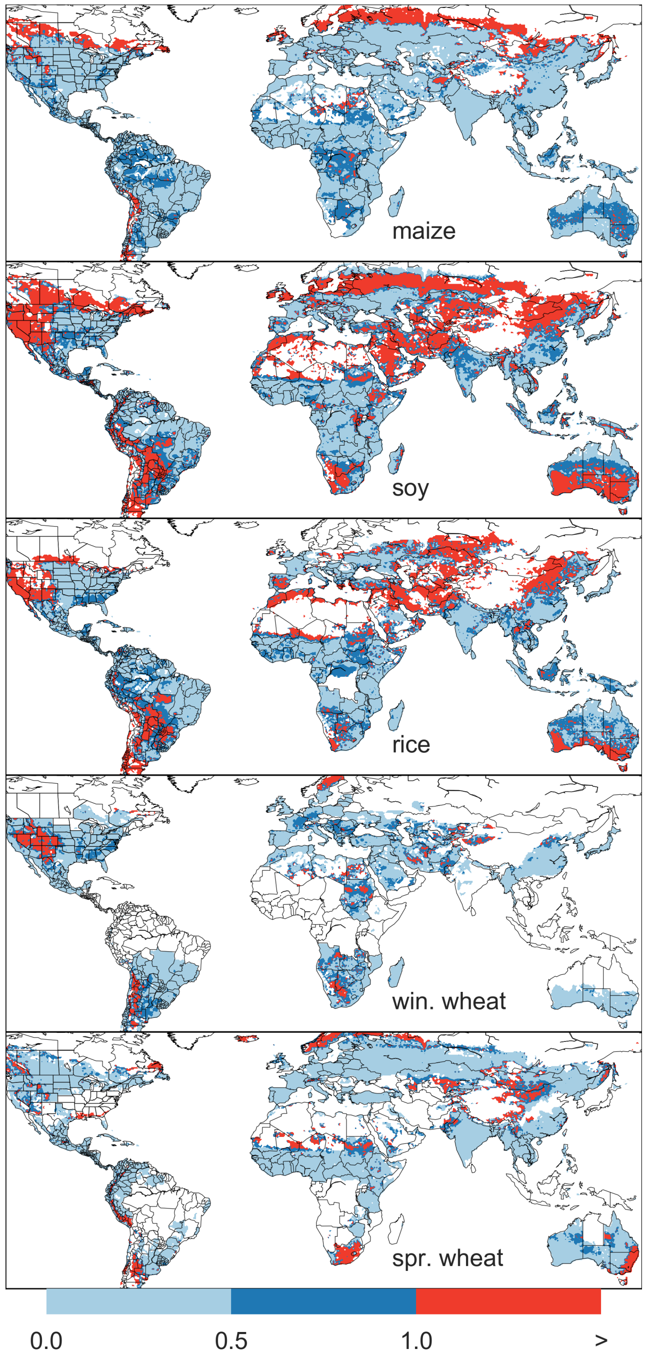

Figure 10: Illustration of our test of emulator performance, applied to the CARAIB model for the T+4 scenario for rain-fed crops. Contour colors indicate the normalized emulator error e, where $e > 1$ means that emulator error exceeds the multi-model standard deviation. White areas are those where crops 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 9.

the behavior of complex process-based models, we evaluate the normalized emulator error. That is, for each grid cell, model, and scenario we evaluate the difference between the model yield and its emulation, normalized by the inter-model standard deviation in yield projections. 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 9), but some individual model-crop combinations are problematic (e.g. PROMET for rice and soy, JULES for soy and winter wheat, Figures S14–S15). 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 S16, lowering the denominator. Emulator performance often degrades in geographic locations where crops are not currently cultivated. Figure 10 shows a CARAIB case as an example, where emulator performance is satisfactory over cultivated areas for all crops other than soy, but uncultivated regions show some problematic areas.

It should be noted that this assessment metric is relatively forgiving. 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. Where models differ more widely, the standard for emulators becomes less stringent. Because models disagree on the magnitude of CO_2 fertilization, this effect is readily seen when comparing assessments of emulator performance in simulations at baseline CO_2 (Figure 9) with those at higher CO_2 levels (Figure S13). Widening the inter-model spread leads to an apparent increase in emulator skill.

3.4. *Emulator applications*

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. Emulators provide a easy way to compare a ensembles of climate or impacts projections. They also provide a means for generating continuous damage functions. As an example, we show a damage function constructed from 4D emulations for aggregated yield at the global scale, for maize on currently cultivated land, with simulated values shown for comparison. (Figure 11; see Figures S16- S19 in the supplemental material for other crops and dimensions.) The emulated values closely match simulations even at this aggregation 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. The global coverage of the GGCMI simulations allows impacts modelers to apply arbitrary geographically-varying climate projections, as well as arbitrary aggregation mask, to develop damage functions for any climate scenario and any geopolitical or geographic level.

4. Conclusions and discussion

The GGCMI Phase II experiment provides a database targeted to allow detailed study of crop yields from process-based models under climate change. The experiment is 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

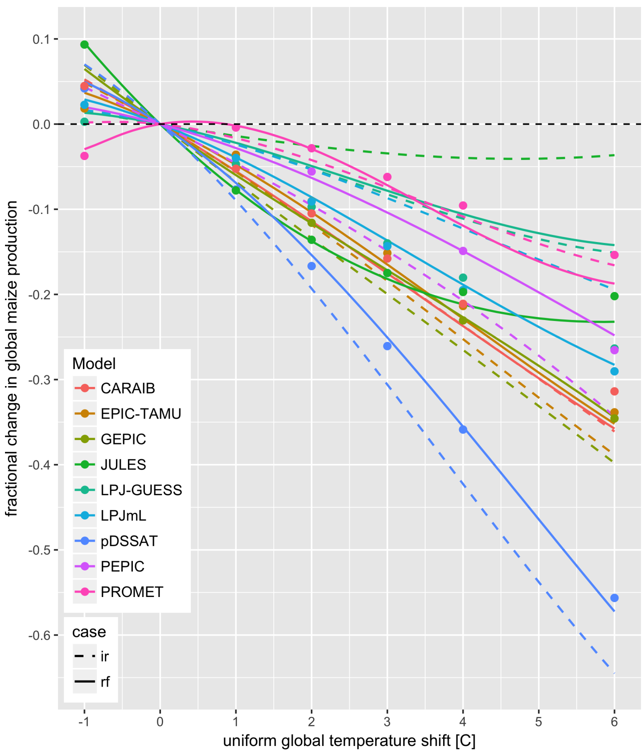

Figure 11: Global emulated damages for maize on currently cultivated lands for the GGCMI models emulated, for uniform temperature shifts with other inputs held at baseline. (The damage function is created from aggregating up emulated values at the grid cell level, not from a regression of global mean yields.) Lines are emulations for rain-fed (solid) and irrigated (dashed) crops; for comparison, dots are the simulated values for the rain-fed case. For most models, irrigated crops show a sharper reduction than do rain-fed 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 S16- S19 in the supplemental material.

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. Across the major crops, inter-model uncertainty is greatest for wheat and least for soy. Across factors impacting yields, inter-model uncertainty is largest for CO_2 fertilization and nitrogen response effects. 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.

Second, the GGCMI Phase II simulations allow understanding the way that climate-driven changes and locations of cultivated land combine to produce yield impacts. One counterintuitive result immediate apparent is that irrigated maize shows steeper yield reductions under warming than does rain-fed maize when considered only over currently cultivated land. The effect results from geographic differences in cultivation. In any given location, irrigation increases crop resiliency to temperature increase, but irrigated maize is grown in warmer locations where the impacts of warming are more severe (Figures S5–S6). The same behavior holds for rice and winter wheat, but not for soy or spring wheat (Figures S8-S10). Irrigated wheat and maize are also more sensitive to nitrogen fertilization levels than are analogous non-irritated crops, presumably because those rain-fed crops are limited by water as well as nitrogen availability (Figure S19). (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 even the relatively limited GGCMI Phase II sampling space allows emulation of the climatological response of crop models 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 single projections cannot, and emulating the resulting response surface then produces a tool that can aid in both physical interpretation of the process-based models and in assessment of agricultural impacts under arbitrary climate scenarios. Emulating the climatological response isolates long-term impacts from any confounding factors that complicate year-over-year changes, and the use of simple functional forms offer the possibility of physical interpretation of parameter values. Care should be taken in applying relationships developed at the yearly level to shifts in the mean climatology. We anticipate that systematic parameter sampling will become the norm in future model intercomparison exercise.

While the GGCMI Phase II database should offer the foundation for multiple future studies, several cautions need to be noted. 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). Because the GGCMI simulations apply uniform perturbations to historical climate inputs, they do not sample changes in higher order moments, and cannot address the additional crop yield impacts of potential changes in climate variability. Although distributional changes in model projections are fairly uncertain at present, follow-on experiments may wish to consider them. Several recent studies have described procedures for generating simulations that combine historical data with model projections of not only mean changes in temperature and precipitation but changes in their marginal distributions (e.g. Chang et al., 2016) or temporal dependence (e.g. Leeds et al., 2015).

The GGCMI phase II output dataset invites a broad range of potential future avenues of analysis. A major target area involves studying the models themselves with a detailed examination of interaction terms between the major input drivers, a more 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 will allow detailed investigations into yield variability and response to extremes under changing management and CO_2 levels. As mentioned previously, the database allows study of geographic shifts in optimal growing regions for different crops and studying the viability of switching crop types in some areas. The output dataset also contains other runs and variables not analyzed or shown here. Runs include several which allowed adaptation to climate changes by altering growing seasons, and additional variables include above ground biomass, LAI, and root biomass (as many as 25 output variables for some models). Emulation studies that are possible include study of year-over-year vs climatological emulation, and more systematic evaluation of different statistical model specifications and formal calculation of uncertainties in derived parameters.

The future of food security is one of the larger challenges facing humanity at present. 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 season, cultivar types and cultivar geographic extent.

5. Acknowledgments

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. Computing resources were provided by the University of Chicago Research Computing Center (RCC). S.O. acknowledges support from the Swedish strong research areas BECC and MERGE together with support from LUCCI (Lund University Centre for studies of Car-

bon Cycle and Climate Interactions).

6. References

Angulo, C., Rtter, R., Lock, R., Enders, A., Fronzek, S., & Ewert, F. (2013). Implication of crop model calibration strategies for assessing regional impacts of climate change in europe. *Agric. For. Meteorol.*, *170*, 32 – 46.

Asseng, S., Ewert, F., Martre, P., Rtter, 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., & Zhu, Y. (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, *5*, 143–147. doi:10.1038/nclimate2470.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J., Hatfield, J., Ruane, A., J. Boote, K., Thorburn, P., Rtter, R. P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A., Doltra, J., & Wolf, J. (2013). Uncertainty in simulating wheat yields under climate change. *Nature Climate Change*, *3*, 827832. doi:10.1038/nclimate1916.

Aulakh, M. S., & Malhi, S. S. (2005). 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.

Balkovi, J., van der Velde, M., Skalsk, R., Xiong, W., Folberth, C., Khabarov, N., Smirnov, A., Mueller, N. D., & Obersteiner, M. (2014). Global wheat production potentials and management flexibility under the representative concentration pathways. *Global and Planetary Change*, *122*, 107 – 121.

Blanc, E. (2017). Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models. *Agricultural and Forest Meteorology*, *236*, 145 – 161.

Blanc, E., & Sultan, B. (2015). Emulating maize yields from global gridded crop models using statistical estimates. *Agricultural and Forest Meteorology*, *214-215*, 134 – 147.

von Bloh, W., Schaphoff, S., Müller, C., Rolinski, S., Waha, K., & Zaehle, S. (2018). Implementing the Nitrogen cycle into the dynamic global vegetation, hydrology and crop growth model LPJmL (version 5.0). *Geoscientific Model Development*, *11*, 2789–2812.

Castruccio, S., McInerney, D. J., Stein, M. L., Liu Crouch, F., Jacob, R. L., & Moyer, E. J. (2014). Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs. *Journal of Climate*, *27*, 1829–1844.

Challinor, A., Watson, J., Lobell, D., Howden, S., Smith, D., & Chhetri, N. (2014). A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change*, *4*, 287 – 291.

Chang, W., Stein, M., Wang, J., Kotamarthi, V., & Moyer, E. (2016). Changes in spatio-temporal precipitation patterns in changing climate conditions. *Journal of Climate*, *29*. doi:10.1175/JCLI-D-15-0844.1.

Conti, S., Gosling, J. P., Oakley, J. E., & O'Hagan, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika*, *96*, 663–676.

Duncan, W. (1972). SIMCOT: a simulation of cotton growth and yield. In C. Murphy (Ed.), *Proceedings of a Workshop for Modeling Tree Growth, Duke University, Durham, North Carolina* (pp. 115–118). Durham, North Carolina.

Duncan, W., Loomis, R., Williams, W., & Hanau, R. (1967). A model for simulating photosynthesis in plant communities. *Hilgardia*, (pp. 181–205).

Dury, M., Hambuckers, A., Warnant, P., Henrot, A., Favre, E., Ouberdous, M., & François, L. (2011). Responses of European forest ecosystems to 21st century climate: assessing changes in interannual variability and fire intensity. *iForest - Biogeosciences and Forestry*, (pp. 82–99).

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., & Foster, I. (2014). The parallel system for integrating impact models and sectors (pSIMS). *Environmental Modelling and Software*, *62*, 509–516.

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., & Sheffield, J. (2015). The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0). *Geoscientific Model Development*, *8*, 261–277.

Ferrise, R., Moriondo, M., & Bindi, M. (2011). Probabilistic assessments of climate change impacts on durum wheat in the mediterranean region. *Natural Hazards and Earth System Sciences*, *11*, 1293–1302.

Folberth, C., Gaiser, T., Abbaspour, K. C., Schulin, R., & Yang, H. (2012). 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.

Food and Agriculture Organization of the United Nations (2018). FAOSTAT database. URL: `http://www.fao.org/faostat/en/home`.

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., & Rötter, R. P. (2018). Classifying multi-model wheat yield impact response surfaces showing sensitivity to temperature and precipitation change. *Agricultural Systems*, *159*, 209–224.

Glotter, M., Elliott, J., McInerney, D., Best, N., Foster, I., & Moyer, E. J. (2014). Evaluating the utility of dynamical downscaling in agricultural impacts projections. *Proceedings of the National Academy of Sciences*, *111*, 8776–8781.

Glotter, M., Moyer, E., Ruane, A., & Elliott, J. (2015). Evaluating the Sensitivity of Agricultural Model Performance to Different Climate Inputs. *Journal of Applied Meteorology and Climatology*, *55*, 151113145618001.

Hank, T., Bach, H., & Mauser, W. (2015). 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.

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

He, W., Yang, J., Zhou, W., Drury, C., Yang, X., D. Reynolds, W., Wang, H., He, P., & Li, Z.-T. (2016). 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.

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

Heady, E. O., & Dillon, J. L. (1961). *Agricultural production functions*. Iowa State University Press.

Holden, P., Edwards, N., PH, G., Fraedrich, K., Lunkeit, F., E, K., Labriet, M., Kanudia, A., & F, B. (2014). Plasim-entsem v1.0: A spatiotemporal emulator of future climate change for impacts assessment. *Geoscientific Model Development*, *7*, 433–451. doi:10.5194/gmd-7-433-2014.

Holzkämper, A., Calanca, P., & Fuhrer, J. (2012). Statistical crop models: Predicting the effects of temperature and precipitation changes. *Climate Research*, *51*, 11–21.

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., & Keating, B. A. (2014). APSIM Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling and Software*, *62*, 327 – 350.

Howden, S., & Crimp, S. (2005). Assessing dangerous climate change impacts on australia's wheat industry. *Modelling and Simulation Society of Australia and New Zealand*, (pp. 505–511).

Iizumi, T., Nishimori, M., & Yokozawa, M. (2010). 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.

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

Izaurralde, R., Williams, J., Mcgill, W., Rosenberg, N., & Quiroga Jakas, M. (2006). Simulating soil C dynamics with EPIC: Model description and testing against long-term data. *Ecological Modelling*, *192*, 362–384.

Jagtap, S. S., & Jones, J. W. (2002). Adaptation and evaluation of the CROPGRO-soybean model to predict regional yield and production. *Agriculture, Ecosystems & Environment*, *93*, 73 – 85.

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., & Ritchie, J. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, *18*, 235 – 265.

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., & Wheeler, T. R. (2017). Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agricultural Systems*, *155*, 269 – 288.

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., & Smith, C. (2003). An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, *18*, 267 – 288.

Leeds, W. B., Moyer, E. J., & Stein, M. L. (2015). Simulation of future climate under changing temporal covariance structures. *Advances in Statistical Climatology, Meteorology and Oceanography*, *1*, 1–14. URL: `https://www.adv-stat-clim-meteorol-oceanogr.net/1/1/2015/`. doi:10.5194/ascmo-1-1-2015.

Lindeskog, M., Arneth, A., Bondeau, A., Waha, K., Seaquist, J., Olin, S., & Smith, B. (2013). Implications of accounting for land use in simulations of ecosystem carbon cycling in Africa. *Earth System Dynamics*, *4*, 385–407.

Liu, J., Williams, J. R., Zehnder, A. J., & Yang, H. (2007). GEPIC - modelling wheat yield and crop water productivity with high resolution on a global scale. *Agricultural Systems*, *94*, 478 – 493.

Liu, W., Yang, H., Folberth, C., Wang, X., Luo, Q., & Schulin, R. (2016a). Global investigation of impacts of PET methods on simulating crop-water relations for maize. *Agricultural and Forest Meteorology*, *221*, 164 – 175.

Liu, W., Yang, H., Liu, J., Azevedo, L. B., Wang, X., Xu, Z., Abbaspour, K. C., & Schulin, R. (2016b). Global assessment of nitrogen losses and trade-offs with yields from major crop cultivations. *Science of The Total Environment*, *572*, 526 – 537.

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

Lobell, D. B., & Field, C. B. (2007). Global scale climate-crop yield relationships and the impacts of recent warming. *Environmental Research Letters*, *2*, 014002.

MacKay, D. (1991). Bayesian Interpolation. *Neural Computation*, *4*, 415–447.

Mauser, W., & Bach, H. (2015). PROMET - Large scale distributed hydrological modelling to study the impact of climate change on the water flows of mountain watersheds. *Journal of Hydrology*, *376*, 362 – 377.

Mauser, W., Klepper, G., Zabel, F., Delzeit, R., Hank, T., Putzenlechner, B., & Calzadilla, A. (2009). Global biomass production potentials exceed expected future demand without the need for cropland expansion. *Nature Communications*, *6*.

Mistry, M. N., Wing, I. S., & De Cian, E. (2017). Simulated vs. empirical weather responsiveness of crop yields: US evidence and implications for the agricultural impacts of climate change. *Environmental Research Letters*, *12*.

Moore, F. C., Baldos, U., Hertel, T., & Diaz, D. (2017). New science of climate change impacts on agriculture implies higher social cost of carbon. *Nature Communications*, *8*.

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., & Yang, H. (2017). Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. *Geoscientific Model Development*, *10*, 1403–1422.

Nakamura, T., Osaki, M., Koike, T., Hanba, Y. T., Wada, E., & Tadano, T. (1997). Effect of CO2 enrichment on carbon and nitrogen interaction in wheat and soybean. *Soil Science and Plant Nutrition*, *43*, 789–798.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, *91*, 1290 – 1300.

Olin, S., Schurgers, G., Lindeskog, M., Wårlind, D., Smith, B., Bodin, P., Holmér, J., & Arneth, A. (2015). 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. doi:10.5194/bg-12-2489-2015.

Osaki, M., Shinano, T., & Tadano, T. (1992). Carbon-nitrogen interaction in field crop production. *Soil Science and Plant Nutrition*, *38*, 553–564.

Osborne, T., Gornall, J., Hooker, J., Williams, K., Wiltshire, A., Betts, R., & Wheeler, T. (2015). JULES-crop: a parametrisation of crops in the Joint UK Land Environment Simulator. *Geoscientific Model Development*, *8*, 1139–1155.

Ostberg, S., Schewe, J., Childers, K., & Frieler, K. (2018). Changes in crop yields and their variability at different levels of global warming. *Earth System Dynamics*, *9*, 479–496.

Oyebamiji, O. K., Edwards, N. R., Holden, P. B., Garthwaite, P. H., Schaphoff, S., & Gerten, D. (2015). Emulating global climate change impacts on crop yields. *Statistical Modelling*, *15*, 499–525.

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., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

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., & Rötter, R. (2015). 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.

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

Porter et al. (IPCC) (2014). 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 C. F. et al. (Ed.), *IPCC Fifth Assessment Report* (pp. 485–533). Cambridge, UK: Cambridge University Press.

Portmann, F., Siebert, S., Bauer, C., & Doell, P. (2008). Global dataset of monthly growing areas of 26 irrigated crops.

Portmann, F., Siebert, S., & Doell, P. (2010). 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.

Pugh, T., Müller, C., Elliott, J., Deryng, D., Folberth, C., Olin, S., Schmid, E., & Arneth, A. (2016). Climate analogues suggest limited potential for intensification of production on current croplands under climate change. *Nature Communications*, *7*, 12608.

Räisänen, J., & Ruokolainen, L. (2006). Probabilistic forecasts of near-term climate change based on a resampling ensemble technique. *Tellus A: Dynamic Meteorology and Oceanography*, *58*, 461–472.

Ratto, M., Castelletti, A., & Pagano, A. (2012). Emulation techniques for the reduction and sensitivity analysis of complex environmental models. *Environmental Modelling & Software*, *34*, 1 – 4.

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*.

Roberts, M., Braun, N., R Sinclair, T., B Lobell, D., & Schlenker, W. (2017). Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters*, *12*.

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

Ruane, A. C., Cecil, L. D., Horton, R. M., Gordon, R., McCollum, R., Brown, D., Killough, B., Goldberg, R., Greeley, A. P., & Rosenzweig, C. (2013). Climate change impact uncertainties for maize in panama: Farm information, climate projections, and yield sensitivities. *Agricultural and Forest Meteorology*, *170*, 132 – 145.

Ruane, A. C., Goldberg, R., & Chryssanthacopoulos, J. (2015). Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation. *Agric. Forest Meteorol.*, *200*, 233–248.

Rubel, F., & Kottek, M. (2010). Observed and projected climate shifts 1901-2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteorologische Zeitschrift*, *19*, 135–141.

Sacks, W. J., Deryng, D., Foley, J. A., & Ramankutty, N. (2010). Crop planting dates: an analysis of global patterns. *Global Ecology and Biogeography*, *19*, 607–620.

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., & Frieler, K. (2017). Consistent negative response of US crops to high temperatures in observations and crop models. *Nature Communications*, *8*, 13931.

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, *106*, 15594–15598.

Snyder, A., Calvin, K. V., Phillips, M., & Ruane, A. C. (2018). A crop yield change emulator for use in gcam and similar models: Persephone v1.0. *Geoscientific Model Development Discussions*, *2018*, 1–42.

Storlie, C. B., Swiler, L. P., Helton, J. C., & Sallaberry, C. J. (2009). Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering & System Safety*, *94*, 1735 – 1763.

Tebaldi, C., & Lobell, D. B. (2008). Towards probabilistic projections of climate change impacts on global crop yields. *Geophysical Research Letters*, *35*.

Valade, A., Ciais, P., Vuichard, N., Viovy, N., Caubel, A., Huth, N., Marin, F., & Martin, J. F. (2014). Modeling sugarcane yield with a process-based model from site to continental scale: Uncertainties arising from model structure and parameter values. *Geoscientific Model Development*, *7*, 1225–1245.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences*, *111*, 3228–3232.

White, J. W., Hoogenboom, G., Kimball, B. A., & Wall, G. W. (2011). Methodologies for simulating impacts of climate change on crop production. *Field Crops Research*, *124*, 357 – 368.

Williams, K., Gornall, J., Harper, A., Wiltshire, A., Hemming, D., Quaife, T., Arkebauer, T., & Scoby, D. (2017). Evaluation of JULES-crop performance against site observations of irrigated maize from Mead, Nebraska. *Geoscientific Model Development*, *10*, 1291–1320.

Williams, K. E., & Falloon, P. D. (2015). Sources of interannual yield variability in JULES-crop and implications for forcing with seasonal weather forecasts. *Geoscientific Model Development*, *8*, 3987–3997.

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

Wolf, J., & Oijen, M. (2002). Modelling the dependence of european potato yields on changes in climate and co2. *Agricultural and Forest Meteorology*, *112*, 217 – 231.

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., Mller, C., Peng, S., Peuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., & Asseng, S. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci.*, *114*, 9326–9331.