The GGCMI phase II emulators: global gridded crop model responses to changes in CO_2, temperature, water, and nitrogen (version 1.0)

James Franke[1,2], Christoph Müller[3], Joshua Elliott[2,4], Alex C. Ruane[5], Jonas Jägermeyr[3,2,4,5], Abigail Snyder[6], Marie Dury[7], Pete Falloon[8], Christian Folberth[9], Louis François[7], Tobias Hank[10], R. Cesar Izaurralde[11,12], Ingrid Jacquemin[7], Curtis Jones[11], Michelle Li[2,13], Wenfeng Liu[14,15], Stefan Olin[16], Meridel Phillips[5,17], Thomas A. M. Pugh[18,19], Ashwan Reddy[11], Karina Williams[8], Ziwei Wang[1,2], Florian Zabel[10], 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]Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany
[4]Department of Computer Science, University of Chicago, Chicago, IL, USA
[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]Unité de Modélisation du Climat et des Cycles Biogéochimiques, UR SPHERES, Institut d'Astrophysique et de Géophysique, University of Liège, Belgium
[8]Met Office Hadley Centre, Exeter, United Kingdom
[9]Ecosystem Services and Management Program, International Institute for Applied Systems Analysis, Laxenburg, Austria
[10]Department of Geography, Ludwig-Maximilians-Universität, Munich, Germany
[11]Department of Geographical Sciences, University of Maryland, College Park, MD, USA
[12]Texas Agrilife Research and Extension, Texas A&M University, Temple, TX, USA
[13]Department of Statistics, University of Chicago, Chicago, IL, USA
[14]EAWAG, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[15]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
[16]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[17]Earth Institute Center for Climate Systems Research, Columbia University, New York, NY, USA
[18]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK.
[19]Birmingham Institute of Forest Research, University of Birmingham, Birmingham, UK.

Correspondence: James Franke (jfranke@uchicago.edu)

Abstract. Statistical emulation of process-based crop models provides the opportunity to combine advantageous features of statistical and process-based crop models. The Global Gridded Model Intercomparison Project (GGCMI) Phase II consists of a set of simulations run on a suit of process based models with an explicit goal of producing a structured training dataset for crop model emulator development across four dimensions: atmospheric carbon dioxide (CO_2) concentrations, temperature, water supply, and nitrogen inputs (CTWN). These training datasets are available for two contrasting assumptions on growing season adaption (A): fixed growing seasons (through adaptation in cultivar choice) and growing seasons shortening in warmer climates (assuming no adaptation in cultivar choice). In this study we present the construction of a set of crop model emulators of mean-climatological yield for nine process-based crop models and five crops. After presenting the rationale and technical

implementation of the emulator construction, we evaluate the emulator performance and discuss the general applicability of these emulators as well as individual cases, where the emulators show unexpected behavior. 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. The crop model emulators presented here should therefore facilitate both model comparison and integrated assessment of climate impacts.

1 Introduction

Improving our understanding of the impacts of future climate change on crop yields is critical for global food security in the twenty-first century. Projections of future yields under climate change are generally made with one of two approaches: either process-based models, which simulate the process of photosynthesis and the biology and phenology of individual crops, or statistical models, which use historical weather and yield data to capture relationships between observed crop yields and major drivers. Process-based crop models provide some advantages, including capturing the direct effects of CO_2 fertilization and allowing projections in areas where crops are not currently grown. However, they are computationally expensive, and can be difficult or impossible to directly integrate into integrated climate change impacts assessments. Statistical crop models can only capture crop responses under the range of current conditions, but have several advantages: they implicitly include management and behavioral practices that are difficult to model explicitly, and they are typically simple analytical expressions that are easily implemented by downstream impact modelers. Both types of models are routinely 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; Liu et al., 2016a).

Statistical emulation allows combining some of the advantageous features of both statistical and process-based models. The approach involves constructing a "surrogate model" of numerical simulations by using their output as training data for a statistical representation (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. The resultant statistical model can produce yield projections under arbitrary emissions scenarios and is an important diagnostic tool for model comparison and model evaluation.

Interest is rising in applying statistical emulation to crop models, and multiple studies have developed crop model emulators in the past decade. 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). 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. More recently, emulators have begun to be used in the context of multi-model intercomparison, with multiple authors (Blanc and Sultan, 2015; Blanc, 2017; Ostberg et al., 2018; Mistry et al., 2017) using them to

analyze the five crop models of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). ISIMIP offers a relatively large training set – control, historical, and several Representative Concentration Pathway (RCP) scenarios using output from up to five climate models (Warszawski et al., 2014; Frieler et al., 2017) – and choices of emulation strategy differ. Blanc and Sultan (2015) and Blanc (2017) use historical and RPC8.5 scenarios, combine multiple climate model projections for RCP8.5, and regress across soil regions. Ostberg et al. (2018) use global mean temperature change (and CO_2) as regressors, and then pattern-scales to emulate local yields. Mistry et al. (2017) compare emulated and observed historical yields, using local weather data and a historical crop simulation. The constraints of the ISIMIP experiment mean that all these efforts do share important common features. All emulate annual crop yields along an entire scenario or scenarios, and all future climate scenarios are non-stationary, with important covariates (temperature and precipitation for example) evolving simultaneously.

An alternative approach to emulation involves construction of a "parameter sweep" training set, a collection of multiple stationary scenarios that systematically cover a range of input parameter values. A parameter sweep offers several important advantages for emulation over an experiment in which climate evolves over time. First, it allows separating the effects of different variables that affect yields but that are highly correlated in realistic future scenarios like those used in ISIMIP (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.

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 and 2019 (Fronzek et al., 2018; Snyder et al., 2019; Ruiz-Ramos et al., 2018). These three studies sample multiple perturbations to temperature and precipitation, and two of the three add CO_2 as well, for a total of 132, 99 and 220 different combinations, respectively. All take advantage of the structured training set to construct emulators ("response surfaces") of climatological mean yields, omitting year-over-year variations. All studies have some limitations, however, for assessing global agricultural impacts. None of the 2018 papers offer responses in every grid cell globally, most instead focus on a limited number of sites. Two involve many crop models but only one crop (wheat) (Fronzek et al., 2018; Ruiz-Ramos et al., 2018), while Snyder et al. (2019) analyzes yield responses for four crops from a variety of voluntarily submitted site-specific crop model results, but extrapolates from a network of site yield to latitude zone responses due to data limitations. Snyder et al. analyzes yield responses for four crops from a variety of voluntarily submitted site-specific crop model results, but extrapolates from a network of site yield to latitude zone responses due to data limitations.

In this paper we describe a set of globally-gridded crop model emulators developed from the new parameter-sweep dataset of the Global Gridded Crop Model Intercomparison (GGCMI) Phase II effort. GGCMI Phase II, a part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013, 2014), provides the first near-global-coverage systematic parameter sweep of multi-model crop simulations consisting of up to 756 combinations in CO_2, temperature, water supply, and applied nitrogen (CTWN). The experiment is specifically designed for construction of crop model emulators, and

to allow diagnosing the impacts on crop yields of both individual factors and their joint effects. In the following, we describe the training dataset (Section 2), the statistical model used for emulation (Section 3), measures of emulator fidelity (Section 4), and examples of preliminary results (Section 5).

2 Training dataset

2.1 The GGCMI Phase II dataset

Table 1. Crop models included in GGCMI Phase II emulators and the number of CTWN-A (Carbon, Temperature, Water, Nitrogen, Adaptation) simulations performed for each model. The maximum number is 756 for A0 (no adaptation) experiments, and 648 for A1 (maintaining growing length) experiments, since T0 is not simulated under A1. "N-Dim." indicates whether the models are able to represent varying nitrogen levels. Each model provides the same set of CTWN simulations across all its modeled crops, but some models omit individual crops. (For example, CARAIB does not simulate spring wheat.) Table adapted from Franke et al. (2019b). For clarity, three simulation models included in Phase II have been removed, those that provided a training set too small to be used in emulation.

Model (Key Citations)	**Maize**	**Soybean**	**Rice**	**Winter wheat**	**Spring wheat**	**N dim.**	**Sims per crop (A0 / A1)**
CARAIB, Dury et al. (2011); Pirttioja et al. (2015)	X	X	X	X	X	–	**252 / 216**
EPIC-TAMU, Izaurralde et al. (2006)	X	X	X	X	X	X	**756 / 648**
JULES, Osborne et al. (2015); Williams and Falloon (2015); Williams et al. (2017)	X	X	X	–	X	–	**252 / 0**
GEPIC, Liu et al. (2007); Folberth et al. (2012)	X	X	X	X	X	X	430 / 181
LPJ-GUESS, Lindeskog et al. (2013); Olin et al. (2015)	X	–	–	X	X	X	**756 / 648**
LPJmL, von Bloh et al. (2018)	X	X	X	X	X	X	**756 / 648**
pDSSAT, Elliott et al. (2014); Jones et al. (2003)	X	X	X	X	X	X	**756 / 648**
PEPIC, Liu et al. (2016b, c)	X	X	X	X	X	X	149 / 121
PROMET, Hank et al. (2015); Mauser et al. (2015); Zabel et al. (2019)	X	X	X	X	X	X	261 / 232

The GGCMI Phase II simulations are described in detail in Franke et al. (2019a), but we summarize briefly here. The experiment involves nine different globally gridded crop models, each simulating multiple crops (maize, rice, soybean, and

Table 2. GGCMI Phase II input levels for the parameter sweep. Values for temperature and water supply are perturbations from the historical climatology. For water supply, perturbations are fractional changes to historical precipitation, except in the irrigated (W_∞) simulations, which are all performed with the maximum beneficial levels of water. Bold font indicates the 'baseline' historical level. One model (XX) also provided simulations at the T+5 level. The full protocol samples across all parameter combinations for a total of 756 cases. Table repeated from Franke et al. (2019b).

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_∞)	%
Applied nitrogen (N)	10, 60, **200**	kg ha^{-1}
Adaptation (A)	**A0: none**, A1: new cultivar to maintain original growing season length	-

spring and winter wheat) across a systematic parameter sweep of as many as 756 combinations, each driven by a historical climate timeseries with systematic perturbations to CO_2, temperature, water supply, and nitrogen application (CTWN). Table 1 shows the participating models and the number of simulation scenarios that each provides, and Table 2 shows the specified 4 levels of atmospheric CO_2, 7 of temperature, 9 of water supply, and 3 of applied nitrogen. See Table 2 for all values associated with each dimension; we sample across all parameter combinations.

These simulations are repeated for two adaptation scenarios: "A0" simulations assume no adaptation in cultivar choice, so that growing seasons shorten in warmer climates, while "A1" simulations assume that adaptation in cultivar choice maintains fixed growing seasons. The complete protocol for each modeling group involves up to 43,524 years of global simulated output for each crop. Because the computational demand is high, modeling groups were allowed to submit at various specified levels of participation, with the lowest recommended level of participation consisting of 20% of the maximum possible simulations; the mean participation level is 65%. Three models (APSIM-UGOE, EPIC-IIASA, and ORCHIDEE-crop) that contributed data to the CTWN-A experiment (Franke et al., 2019a) below this recommended threshold (providing samples under 5% of the full protocol) could not be robustly emulated with the method described here and so are excluded here.

Each individual crop model simulation is run for 31 years over historic weather for the period of 1980-2010, with added uniform perturbations to any of the CTWN variables. Historical weather is taken for most models from the AgMERRA (Ruane et al., 2015) historical daily climate data product, but the PROMET model uses the ERA-Interim reanalysis (Dee et al., 2011) and the JULES model uses a bias-corrected version of ERA-Interim, WFDEI (WATCH-Forcing-Data-ERA-Interim) (Weedon et al., 2014) as these groups have specific sub-daily input data requirements. Temperature perturbations are applied as additive

mean shifts, water supply as fractional multipliers to precipitation (except W_∞), and CO_2 and nitrogen application as fixed values. Models provide global output at 0.5 degree latitude and longitude resolution for each simulation year.

2.2 Climatological vs. year-to-year response

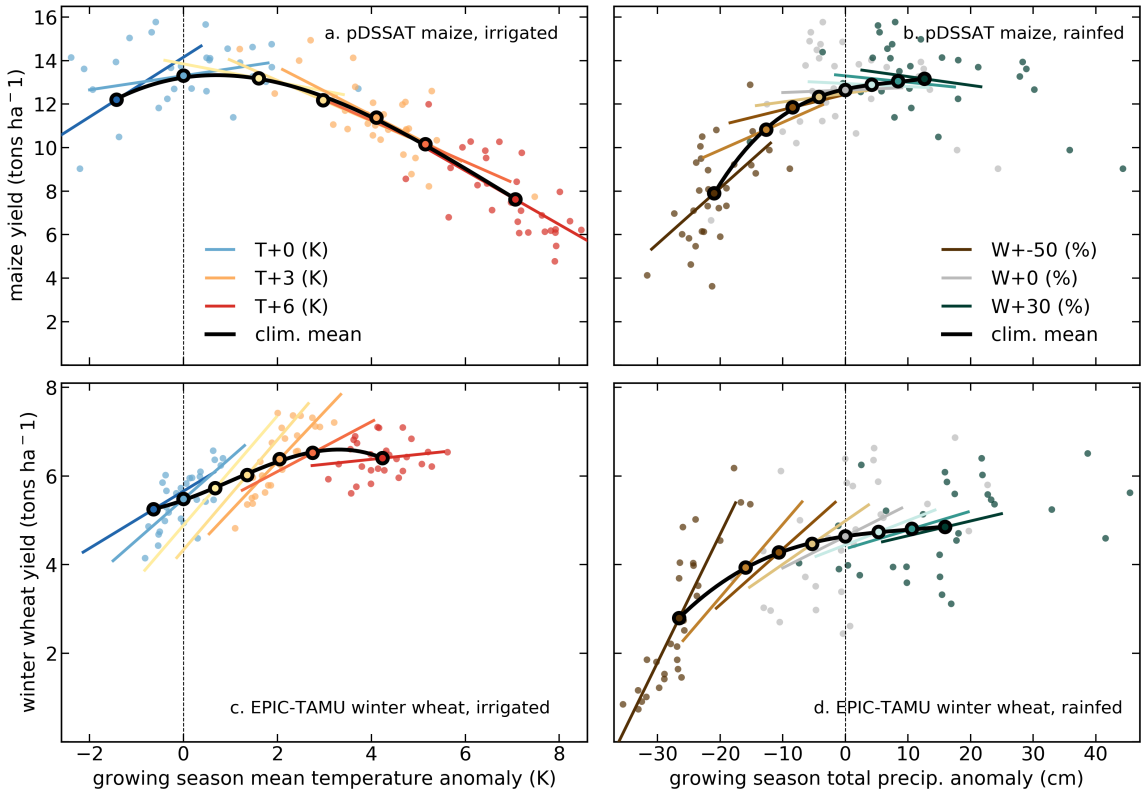

Figure 1. Example showing distinction between crop yield responses to year-to-year and climatological mean shifts in climate variables. For maize from the pDSSAT model in a representative high-yield region (grid cell in northern Iowa, top row) and for winter wheat in France in the EPIC-TAMU model (bottom row). **a & c:** irrigated crops, all temperature cases with other variables held at baseline values, and **b & d:** rainfed crops, all precipitation cases. Open black circles mark climatological mean yields and bold black lines show a 3rd order polynomial fit through them. Colored lines show linear regressions (by orthogonal distance regression) through the 30 annual yields of each parameter case. Colored circles show annual yields for selected cases Responses to year-over-year fluctuations can be very different from those to longer-term climate shifts – slope of colored and black lines differ – with differences generally stronger for wheat (bottom) than maize (top). Note that for rain-fed crops, slope differences in this representation could result from correlated precipitation and temperature fluctuations in the baseline timeseries, but P-T correlations do not contribute to the differences shown here. Such correlation would complicate emulations based on year-over-year yields but would not necessarily bias them.

We emulate the climatological mean response, because that is the response of interest in assessments of climate change impacts. The year-over-year response can be significantly different from the forced climatological one, so we do not use

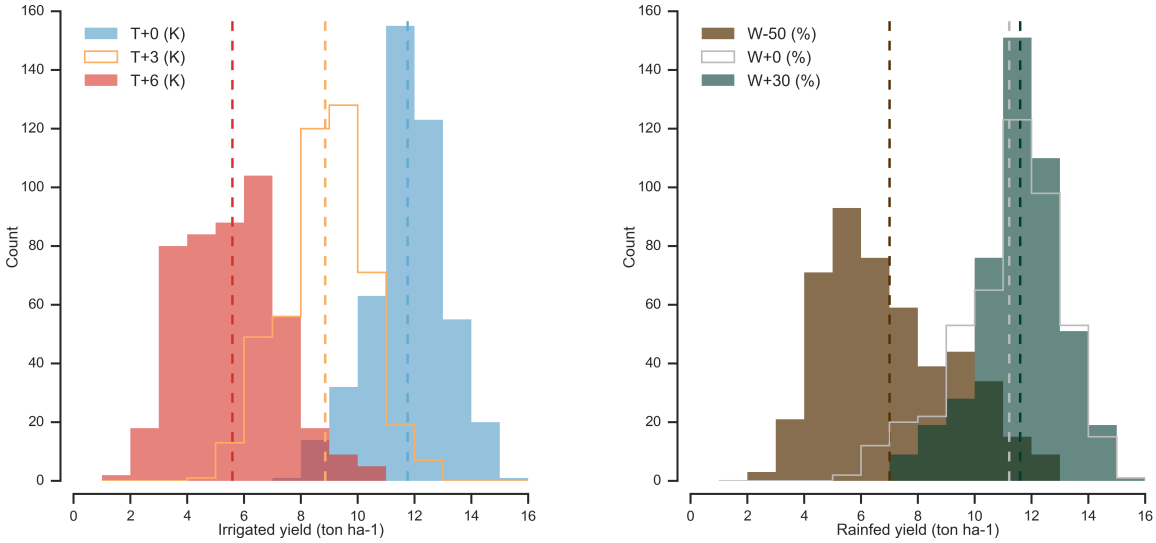

Figure 2. Example showing results of increased crop yield sensitivity to year-over-year climate variations under climate stress. Yield distributions are from examples of Figure 1, top row, of maize in Iowa, (**left**) for irrigated maize in scenarios of altered temperature and (**right**) for rainfed maize in scenarios of altered precipitation. Because yield sensitivities rise under strong warming or drying, distributions of year-over-year crop yields widen in T+6 and P-50% scenarios relative to present-day simulations, even though all input climate timeseries have identical variance for temperature. Note: precipitation changes have different variance since the perturbations are fractional.

information from year-to-year variability but instead emulate the aggregated mean yield in each 30-year simulation. Emulation then becomes relatively straightforward, since changes in time-averaged yields are also considerably smoother than those in year-to-year yield response.

In the GGCMI Phase II simulation output dataset, year-over-year responses to weather can be quantitatively distinct from responses to climatological shifts. This behavior is illustrated in Figure 1, which shows irrigated and rainfed maize and wheat in representative locations; open circles and black lines show the climatological mean response, and solid circles and colored lines the responses for the 30 individual years in individual scenarios. Year-over-year and climatological responses do not generally match, with the discrepancy especially strong in wheat and rice. When discrepancies are large, year-over-year responses are generally stronger than climatological ones, but exact responses differ by crop and region and even by model within GGCMI Phase II.

While differences between year-over-year and climatological temperature responses can arise for many reasons, including memory in the crop model or lurking covariates, the most likely explanation here is that the regressors used, mean growing-season temperature or precipitation, do not fully describe the conditions that affect crop yields. First, the mean growing-season value is only a proxy for the distribution of daily climatic conditions that crops are sensitive to, and present-day interannual variability can be very different from future forced changes. That is, variations in growing-season *means* from year to year at present are associated with changes in growing-season *distributions* that are unrelated to any changes in future warmer

climates (e.g. Ruane et al., 2016): a warm year at present may be quite different from a warm year in the future. Changes in temperature distributions have been shown to strongly affect crop yields (e.g. Hansen and Jones, 2000; Gadgil et al., 2002), though precipitation effects should be smaller since crops respond not to rainfall but to soil moisture, which integrates over weeks or even months (e.g. Potter et al., 2005; Glotter et al., 2014; Challinor et al., 2004). Second, any nonlinearity in crop responses will itself lead to a distinction between climatological and year-over-year fits, even if distributional differences are irrelevant. Given the interannual variations in the climate timeseries, the mean annual yield response to a perturbation is not the same as the response of the climatological mean yield. The effect of nonlinearity may be particularly relevant for precipitation, since model crop yields drop steeply and nonlinearly with increasing dryness. (Crop yields should drop under excess precipitation as well, but process-based models do not capture losses in saturated conditions well (e.g. Glotter et al., 2015; Li et al., 2019).)

In the GGCMI Phase II experiment, the imposed perturbations involve no changes in underlying distributions. The choice is reasonable, since climate model projections of changes in temperature distributions are small relative to existing year-to-year variations, and often inconsistent between models.

See supplemental Figure SXX for example variability change under climate change for some example climate models.

Note that even though distributions of climate variables are unchanged in the GGCMI Phase II simulations, the spread in annual yields becomes wider in highly impacted climate states because of the nonlinearity of yield responses (Figure 2). Higher sensitivity in conditions of extreme climate stress produces greater year-to-year yield variance for all crops except rice, which is typically irrigated and experiences no water stress. Similar results have been reported in other studies: for example, in a study of statistical models projecting with a variety of climate models, Urban et al. (2012) shows 20% increase per degree K in variance in U.S. maize yields under climate change.

3 Emulation

Emulation involves fitting individual regression models from GGCMI Phase II output for each crop and model and 0.5 degree geographic pixel; the regressors are the applied perturbations in CO_2, temperature, water, and nitrogen (CTWN). We discuss here largely emulations of climatological mean crop yield with no growing season adaptation (A0 scenarios), but note that any output of the crop models can potentially be emulated. We provide separate emulations of not only irrigated and rainfed yields but also applied irrigation water (pirrww in mm yr^{-1}, see (Franke et al., 2019a)) in both the A0 and A1 growing season, meaning that each model and crop combination results in six regressions. (See Supplementary Material SXX for more information on additional cases not shown.)

3.1 Statistical model

For the statistical model of crop yields as a function of CTWN, we choose a relatively simple parametric model with a 3rd-order polynomial basis function. If the climatological mean response is relatively smooth, then a simpler form provides a reasonable fit that allows for some interpretation of resultant parameter weights. A relativity simple parametric form also allows fast model

emulation at the grid cell level as opposed to the global or large regional level. By emulating at the grid cell level, we indirectly include any yield response to geographically distributed factors such as soil type, insolation, and the baseline climate, and preserve the spatial resolution of the parent models. To facilitate potential parameter-by-parameter comparison across crop models, we hold the functional form constant in space, across all crops, and models. That is, the same statistical model is used for all grid cells, models, and rainfed crops. Note however that regressions for irrigated crops do not contain W terms and models that do not sample the nitrogen levels omit the N terms.

Both higher-order and interaction terms are expected to be important for representing crop yields. Higher order terms are needed because crop yield responses to weather are well-documented to be nonlinear: e.g. Schlenker and Roberts (2009) for T perturbations and He et al. (2016) for W (precipitation). Interaction terms are needed since the yield response is expected to depend on interactions between the major inputs. For example, Lobell and Field (2007) and Tebaldi and Lobell (2008) showed that in real-world yields (with C and N fixed), the joint distribution in T and W is needed to explain observed yield variance. Other observation-based studies have shown the importance of the interaction between W and N (e.g. Aulakh and Malhi, 2005), and between N and C (Osaki et al., 1992; Nakamura et al., 1997).

A full third order polynomial with interaction terms for the four regressors (CTWN) has 35 total terms (Equation 1), too many for robust fitting even with the large GGCMI Phase II dataset. We therefore reduce the number of free parameters through a feature selection process (discussed in detail below), eliminating 12 terms that do not play a significant role in predicting crop yields; these are shown in gray in Equation 1. The resulting 23-parameter model (Equation 1) can be well-fitted to crop model response in nearly all regions, with the only exceptions being extremely low-yield regions where crops are not currently grown.

$$
\begin{aligned}
Y \; &= K_1 \\
&+ K_2 C + K_3 T + K_4 W + K_5 N + K_6 C^2 \\
&+ K_7 T^2 + K_8 W^2 + K_9 N^2 + K_{10} CW \\
&+ K_{11} CN + K_{12} TW + K_{13} TN + K_{14} WN \\
&+ K_a CT + K_{15} T^3 + K_{16} W^3 + K_b C^3 + K_c N^3 \\
&+ K_{17} TWN + K_{18} T^2 W + K_{19} W^2 + K_{20} W^2 N \\
&+ K_d CWN + K_e CTN + K_f CTW + K_{21} N^2 C \\
&+ K_{22} N^2 T + K_{23} N^2 W + K_g T^2 N + K_h T^2 C \\
&+ K_i W^2 C + K_j C^2 W + K_k C^2 T + K_l C^2 N
\end{aligned}
\tag{1}
$$

We do not focus in this study on comparing different functional forms or non-parametric models. Some prior studies have used other statistical specifications in crop model emulation: for example, Blanc and Sultan (2015) and Blanc (2017) use a 39 term fractional polynomial. Such a high-dimensional model is difficult to fit, especially for a training set of realistic simulations in which input parameters are highly correlated, and Blanc and Sultan (2015) and Blanc (2017) "borrow information across

space" by fitting grid points simultaneously across soil region in a panel regression. Our simpler functional form can be fit independently at each grid cell while still providing a satisfactory emulation of all GGCMI crop models and crops. (See Section 4 for evaluation of emulator fidelity.)

3.2 Feature selection

To reduce the number of terms in our statistical model, we apply 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 one by one, and we calculate an aggregate mean absolute error with each increasing terms and eliminate those terms that do not contribute significant reductions in error (top row of Figure 3). Some terms that did not reduce the aggregate error are included if a higher order version of that term provided a decrease in mean squared error: for example, the T^3 term cannot be included without also taking the T^2 and T terms. We select terms by applying the feature selection process to three example models: two that provided the complete set of 672 rainfed simulations (pDSSAT, EPIC-TAMU, and one that provided the smallest training set (120 input combinations, PEPIC). Feature importance is not uniform due to spatial heterogeneity across models and crops, so we weight the loss function by current cultivation area (Portmann et al., 2010) during this step. The resulting choice of terms is then applied for all emulators and all crops. Since the goal of the emulator is interpolation within the sample space and not extrapolation, we err on the side of including terms that are useful in at least some cases, because the added predictive ability outweighs the costs to distribution of the residuals or over-fitting.

Feature importance is remarkably consistent across models (Figure 3). Even though the models exhibit different absolute levels of error, all three models agree remarkably well on feature importance, that is on which terms reduce error and which provide no predictive benefit. (Agreement means that line slopes match in Figure 3.) The feature selection process allows us to eliminate 11 terms, leaving a final polynomial in 23 terms. We necessarily omit the N^3 term, which cannot be fitted because we sample only three nitrogen levels, but retain other higher-order N terms. The eliminated terms include many of those in C: the cubic; the CT, CTN, CTW, and CWN interaction terms; and all higher order interaction terms in C. Finally, we eliminate one 2nd-order interaction term in W and two in T. Implications of this choice include that nitrogen interactions are complex and important, and that water interaction effects are more nonlinear than those in temperature.

3.3 Model fitting

To fit the parameters K, we use a Bayesian Ridge regularization method (MacKay, 1991) rather than standard ordinary least squares (OLS). The Bayesian Ridge method reduces volatility in parameter estimates when the sampling is sparse, by weighting parameter estimates towards zero, allowing and use of a consistent functional form across all models and locations. The choice slightly reduces mean absolute error for some of the high-order interaction terms in the model (Figure 3, top row) but drastically reduces standard parameter error in the model by stabilizing the estimates (Figure 3, third row). The estimation method scores relatively lower on adjusted R^2 for the simplest parameter specifications, but reaches parity with the OLS at the number of terms included in this study. We use adjusted R^2 as a metric because additional terms are penalized (Equation 2, where n is

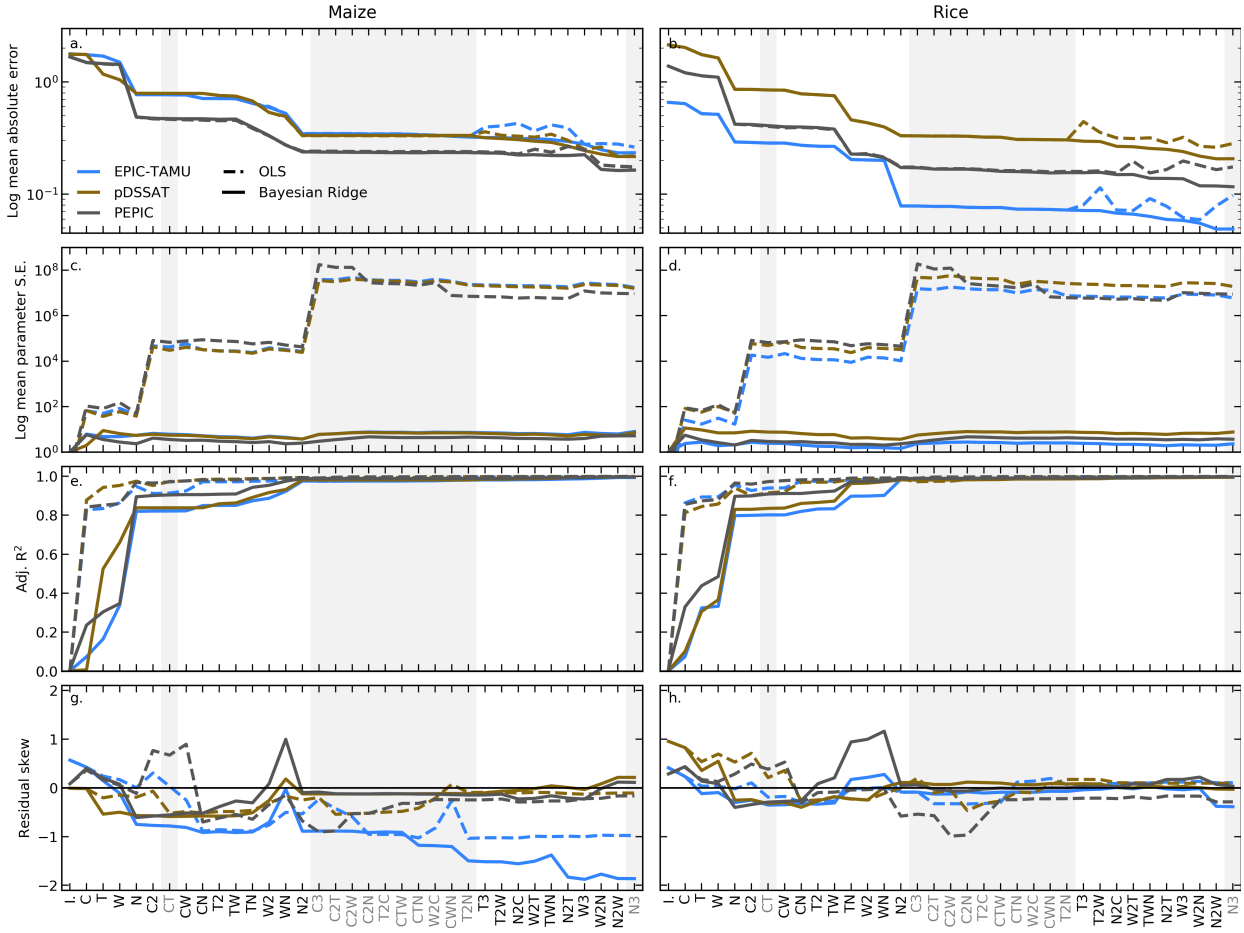

Figure 3. Illustration of results from the polynomial feature selection process for three different crop models (colors), for all grid cells with more than 1000 ha cultivated for maize (**left**) and rice (**right**). Solid lines are Bayesian Ridge regression results and dashed lines those for standard OLS. Rows show four metrics of fit quality and x axes the terms successively tested in the statistical model, sequentially added to the model in order from left to right. Terms that do not reduce the aggregate error are marked in gray are not included in the final model. **a & b:** log mean absolute error between emulated yield and simulated values calculated with a three fold cross validation process, where the emulator is trained on two thirds of the data and predicts the remaining third. **c & d:** adjusted R^2 score for the fit at each model specification. **e & f:** log mean standard parameter error. The Bayesian Ridge method strongly reduces parameter error and results in more stable estimates. **g & h:** distribution of the residuals.

the number of samples and k is the number of features):

$$R^2_{adj} = 1 - \frac{(n-1)\cdot(1-R^2)}{n-k} \tag{2}$$

We use the implementation of the Bayesian Ridge estimator from the scikit-learn package in Python (Pedregosa et al., 2011).

We also Normally distributed residuals indicates that the errors from the. regression model are random. The distribution of the residuals depends on the number of features included in the regression, the method for estimating the parameters, and the target distribution in the training set. Including additional higher order terms in the model tends to reduce the skew in the residuals in most cases. The residuals are only normally distributed (Shapiro–Wilk test (Shapiro and Wilk, 1965) pvalue > 0.05) for the PEPIC model for any specification tested. The EPIC-TAMU and pDSSAT crop module emulator residuals are never normally distributed by this metric for any feature specification proposed here.

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. The lowest number of simulations emulated across the full parameter space is 120 (for the PEPIC model), since we do not attempt to emulate models that provided less than 50 simulations. Should this statement move to emulator evaluation? What is the metric of a problematic fit?

4 Emulator evaluation

In this section we show illustrations of the GCCMI models yield responses to climate perturbations, and evaluate metrics of emulator performance. Model emulation with the parametric method used here requires that crop yield responses be sufficiently smooth and continuous to allow fitting with a relatively simple functional form; we demonstrate that this condition largely holds in the GGCMI Phase II simulations. Emulation errors – discrepancies between emulation and simulation – are generally small, especially when compared to the differences across crop models or climate model inputs. Emulation errors become problematic only in certain cases: in models that provide limited sampling of the GGCMI parameter space, and in some geographic locations where crops are currently not grown.

4.1 Yield response

Crop yields show strong spatial differentiation across geographic regions, and emulators are able to readily reproduce these.Figure 4 illustrates the spatial yield pattern under current climate for one crop and model (maize in LPJmL). Absolute emulation errors are low – nearly all (99.8%) of grid cells have errors below 0.5 tons ha^{-1} – but emulation errors as a percentage of baseline yield can be large in areas with low potential yield and no current cultivation in the real world (e.g. the Sahara, Patagonia). These regions are not currently viable for agriculture and may never become viable even under extreme climate change. Emulation spatial skill differs across models and crops, with maize being the qualitatively easiest to emulate across all models and XX showing larger discrepancies outside currently cultivated areas. See Supplemental Figures SXX-SXX for more crop and model examples.

Yield responses to the four main drivers considered here (C, T, W, and N) are also quite diverse across locations, crops, and models, but in nearly all cases the local climatological mean responses are smooth enough to permit emulation with the functional form used here. Figure 5 illustrates geographic diversity of responses within a single crop and model, for rainfed maize in pDSSAT. While CO_2 responses (in t ha^{-1}/ppm) are quite similar, precipitation response is stronger in more arid locations and nitrogen responses appear strongly soil-dependent. This heterogeneity supports the choice of emulating at the

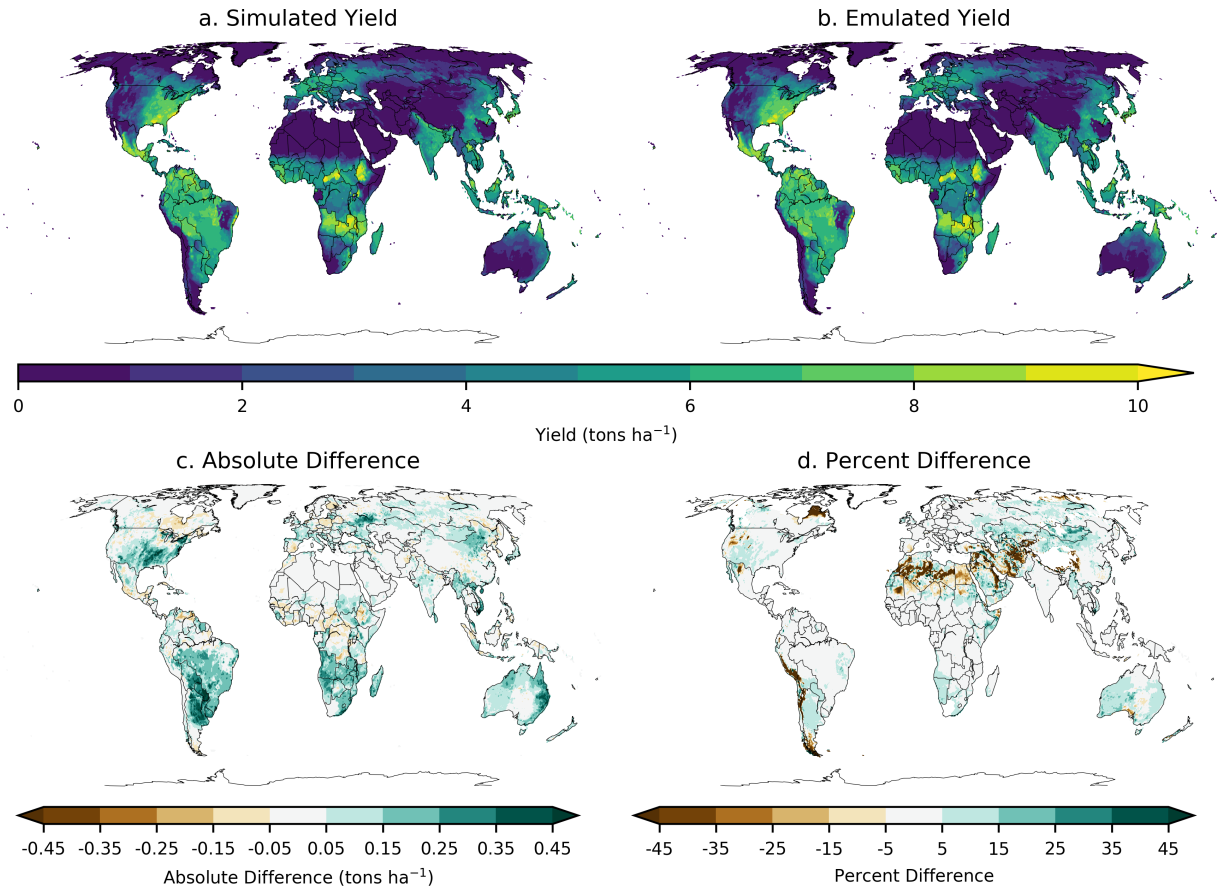

Figure 4. Illustration of spatial pattern in baseline yield successfully captured by the emulator. Simulated (**a.**) and emulated (**b.**) yield under historical (1981-2010) conditions for rainfed maize from the LPJmL model. Absolute yield differences (**c.**) are less that 0.5 ton ha^{-1} in almost all (99.8%) grid cells across the globe. Percent difference (from simulated baseline, **d.**) is below 5% in most (75%) grid cells currently cultivated in the real world. Approximately 7% of grid cells have errors over 20% different from baseline, but only 3% of grid cells with current cultivation (Portmann et al., 2010) have errors over 20%. Notable exceptions include areas with very low baseline yield in the simulations including, for example, the Sahara, the Andes, and northern Quebec. Percent error weighted by cultivation area globally is essentially zero (see also Table 3). Performance varies by crop and model. See Supplementary Material for more examples.

grid cell level. In regions with current cultivation, yields evolve smoothly across the space sampled, and the polynomial fit captures the climatological-mean response to perturbations well. Emulators do perform poorly in a few regions that involve discontinuous or irregular yield responses. This condition is illustrated here with maize from the PROMET model in northern Canada, which is considered too cold for maize at present (0 ton ha^{-1} yield), but which shows an abrupt rise to moderate yields once temperature rise 4 degrees. Under these conditions, the 3rd order polynomial cannot fit the response, and errors are high. See Section 4.2 for additional discussion. Why not show PROMET for all in Figure 5? Looks weird to show two different models.

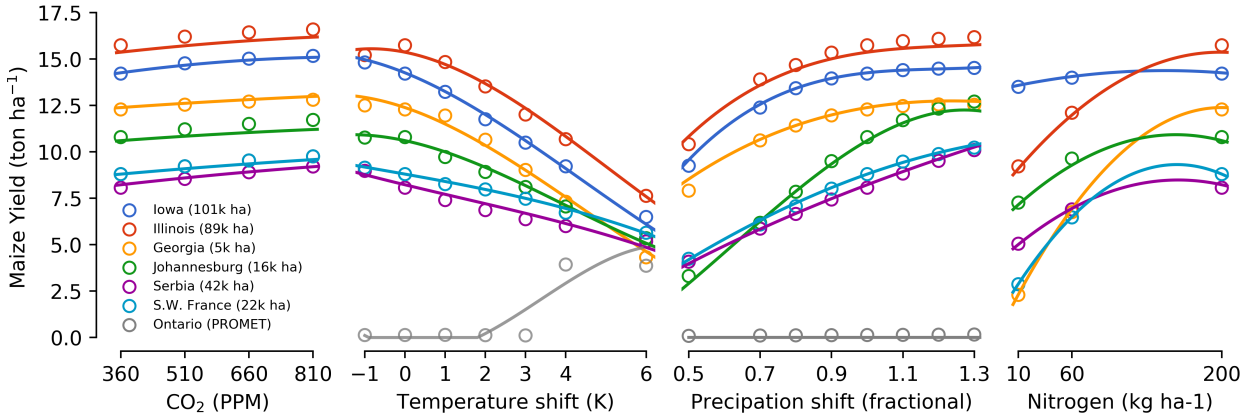

Figure 5. Illustration of spatial variations in yield response, which are successfully captured by the emulator. Panels show simulations and emulations of rainfed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares 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. In some cases extrapolation would produce misleading results, and the emulator fails in conditions where yield response changes abruptly. Failure is illustrated here by the rainfed maize response in north-central Ontario for the PROMET model, which shows present-day yields of zero rising abruptly if temperature warms by 4 degrees.

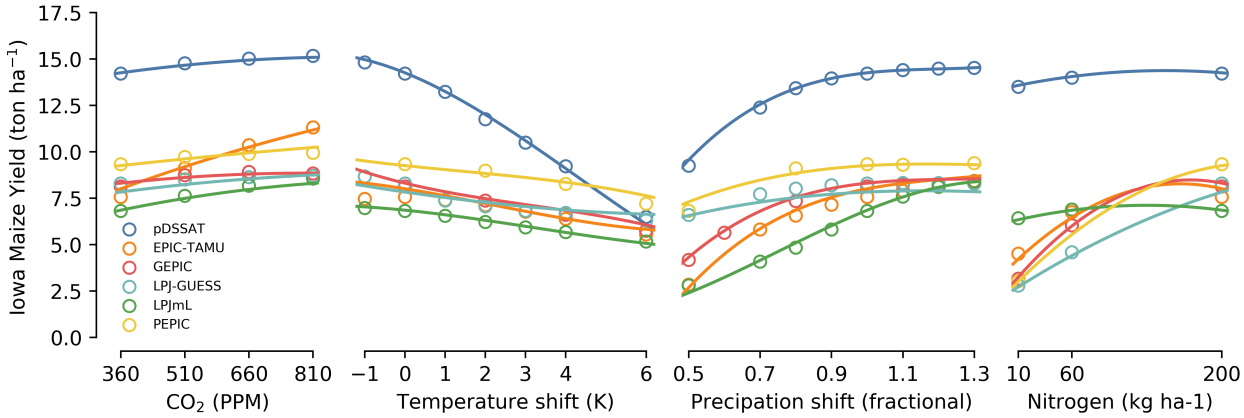

Figure 6. Illustration of variations in yield response across models, which are successfully captured by the emulator. Panels show simulations and emulations from six representative GGCMI models for rainfed maize in the same Iowa grid cell shown in Figure 5, with the same plot conventions. Three models (PROMET, JULES, and CARAIBN) that do not simulate the nitrogen dimension are omitted for clarity. Models are uncalibrated, producing spread in absolute yields. While most model responses can readily emulated with a simple polynomial, some response surfaces diverge slightly from the polynomial form, producing emulation error (e.g. LPJ-GUESS here, for all parameter dimensions). Resulting error generally remains small relative to differences across models.

Crop yield responses in all models generally follow similar functional forms at any given location, though with a spread in magnitude (Figure 6, which shows rainfed maize in northern Iowa in a selection of GGCMI models). Absolute yield differences between models can be substantial because models are uncalibrated. In general, models are most similar in their responses to temperature perturbations, and least similar to changes in CO_2. That is, CO_2 fertilization effects *within* a single model are consistent across locations, but CO_2 effects differ strongly *across* models. Differences in response shape can lead to some differences in the fidelity of emulation for different crop-model combinations; we evaluate these in more detail in the following section.

Note that while the nitrogen dimension is important, it is also the most troublesome to emulate in this work because of its limited sampling compared to other dimensions. 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 (Figure 6, right). While reduced yields under high nitrogen levels are physically possible and could reflect over-application at particular times in the growing period, they are implausible at the magnitude shown here and likely an artifact of the fit. The Bayesian Ridge estimator mitigates the 'peak-decline effect' in the nitrogen dimension relative to ordinary least squares, but does not entirely remove it. The polynomial fit also 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.

4.2 Emulator performance metrics

Our emulators collectively consist of nearly 3 million individual regressions, so developing concise performance metrics poses a challenge. No general agreed-upon criteria exist for defining an acceptable crop model emulator, so we present two different metrics, one relatively loose and one more stringent. These are described in detail below. Both metrics assess the ability of the emulator to reproduce model output from the GGCMI Phase II experiment. Finally, we also demonstrate the emulator's ability to simulate a realistic projection of evolving climate.

1. Normalized error. We take as our first metric what we term the "normalized error", which compares the fidelity of an emulator to the inter-model spread. For a multi-model comparison exercise like GGCMI Phase II, a reasonable though loose emulator criterion is that its errors be small relative to intermodel differences. The normalized error e is defined separately for each C,T,W,N scenario s as the difference between emulated and simulated fractional yield changes, normalized by the standard deviation in simulated changes across all models:

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

where F is the fractional change in yields Y between scenario s and baseline b:

$$F_{\,s} = \frac{Y_s - Y_b}{Y_b} \tag{4}$$

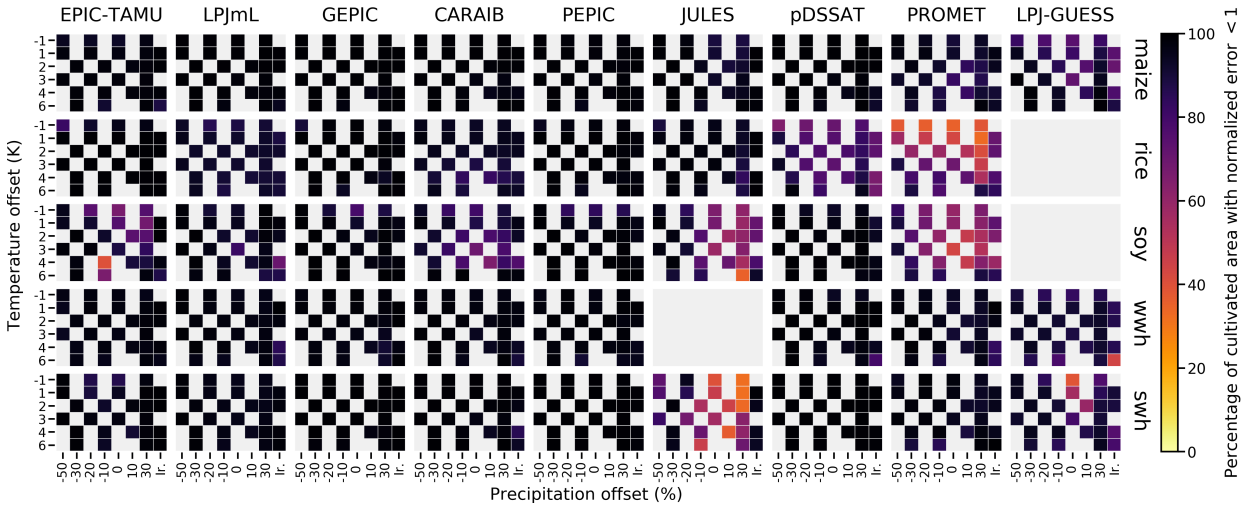

Figure 7. Assessment of emulator performance over currently cultivated areas based on normalized error (Equations 4). We show performance of all 9 models emulated, over all crops and all sampled T and W inputs ("ir." indicates the irrigated W_∞ setting), but with CO_2 and nitrogen held fixed at baseline values. Large columns are crops and large rows models; squares within are T, W 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 63 scenarios at a single CO_2 and N value, we consider only those for which all 9 models submitted data (Figure SX) so the model ensemble standard deviation can be calculated uniformly in each case. JULES did not simulate winter wheat and LPJ-GUESS did not simulate rice and soybean. 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 colder and wetter scenarios.

Evaluation of this metric implies that GGCMI Phase II emulators are generally satisfactory. We calculate the mean error for each grid cell, model, and crop in each C,T,W,N scenario by comparing emulated and simulated yields, and show the averages over currently cultivated area in Figure 8. A normalized error $e < 1$ means that any deviation of the emulation from the simulation is less than 1 standard deviation of the inter-model spread. For maize, for example, 4 (CARAIB, GEPIC, LPJmL, and pDSSAT) of the 9 models produce $e < 1$ over 99% of currently cultivated area and an additional 3 models produce $e < 1$ over 97% of currently cultivated area. (LPJ-GUESS emulation is lowest in fidelity for maize, as previously shown in Figure 6, and produces 92%.) A few individual model-crop combinations are particularly problematic, including PROMET for soybeans and rice (only 67% of cultivated area) and JULES for soybeans and spring wheat (74% of cultivated area). (See Figures SXX for additional examples). In crop-model cases where performance is low, it tends to degrade most in the coldest and wettest scenarios, suggesting that decreasing temperature or increasing precipitation is hard to fit with the polynomial because yield increases potentially are saturating in these conditions. Emulator performance will also be poor in cases where models show steep yield losses or complete crop failures, because a discontinuity (crop failure) cannot be captured with a 3rd order polynomial.

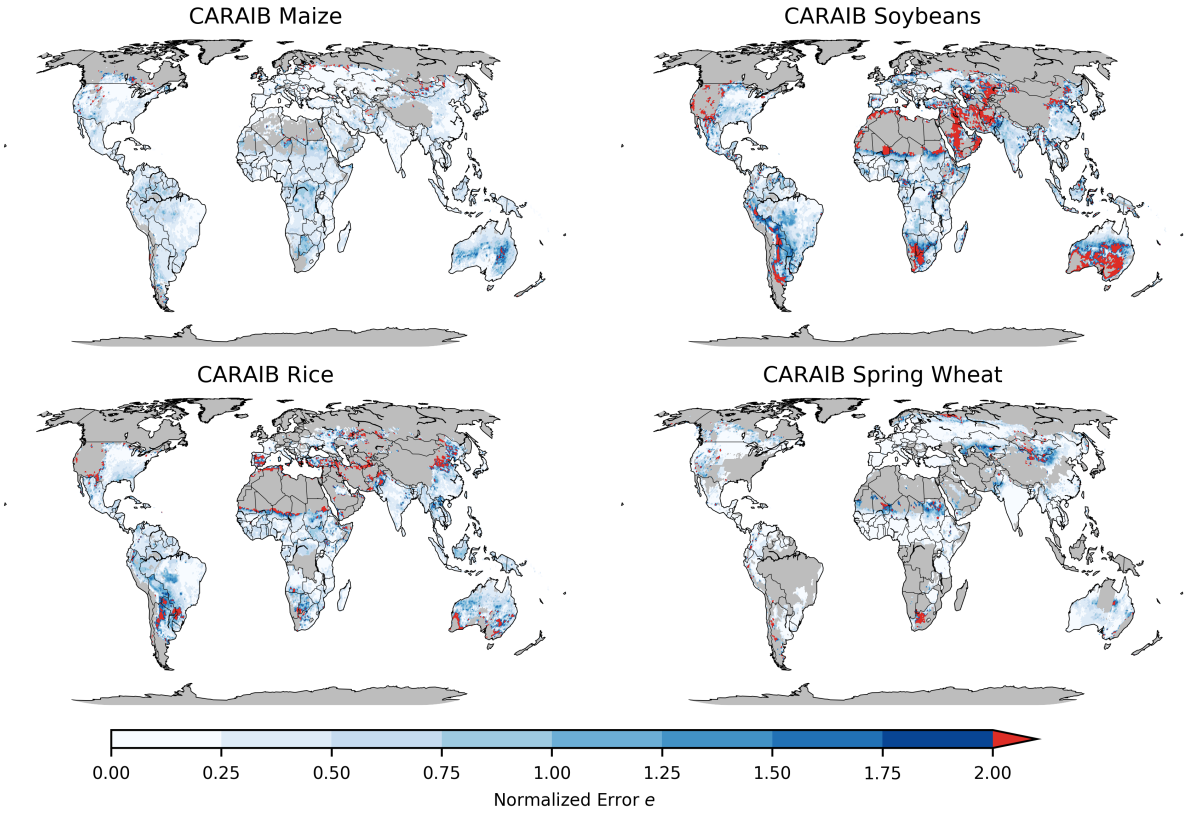

Figure 8. Illustration of our test of emulator performance, applied to the CARAIB model for the T+4 scenario for rainfed crops. Contour colors indicate the normalized emulator error e, where $e > 1$ means that emulator error exceeds the multi-model standard deviation. White areas are those where crops 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 soybean is more problematic, as was also shown in Figure 7.

While Figure 8 shows only currently cultivated land, performance can be worse in locations where crops are not currently cultivated, or on marginal lands where potential yields are low. Figure 8 shows normalized error for CARAIB in the T+4 scenario over all simulated area with a non-zero baseline yield and with at least 6 models providing simulations. Emulator performance can be poor ($e > 2$) in arid or mountainous zones: the edges of the Sahara, the Near East, S. Africa and Southern Australia. See Figures SXX-SXX for analogous figures for all models. Over all crop model combinations, the emulator produces $e < 1$) on average for 95% of currently cultivated area.

Note that our normalized error assessment is relatively forgiving for several reasons. First, it is an in-sample validation, with the emulation evaluated against the simulations actually used to train the emulator. Had we used a spline interpolation, the error would necessarily be zero. (Our second metric involves evaluating on scenarios not included in the training set.) Second, the

metric scales emulator fidelity not by the magnitude of yield changes in the evaluated model but by the spread in yield changes across models. The normalized error e for a given model then depends on the particular suite of other models considered in the intercomparison exercise. The rationale for the choice is to relate the fidelity of the emulation to the true uncertainty, which we take as the multi-model spread. The metric then has the property that where models differ more widely, the standard for emulators becomes less stringent, and vice versa. In GGCMI Phase II the effect is manifested in the higher normalized errors for soybeans across all models, which results not because soybean yields are difficult to emulate but because models agree more closely on yield changes for soybeans than for the other crops. (See also Section 5).

Table 3. Mean squared error of emulator representation of a simulation as a percentage of baseline simulated yield for the cross-validation process for rainfed and irrigated crops. A 3-fold stratified k-fold cross validation scheme is utilized where the model is trained on 66% of the data and validated on the held-out 33% (repeated four times). The split does not represent a uniform number of samples in each location or in each model because simulation sampling extent in variable space is heterogeneous. The table shows the median grid cell error (as a percentage of baseline yield) over all currently cultivated grid cell (Portmann et al., 2010). Values in parenthesis indicate the values across irrigated crops. * Indicates cases where the OLS linear model fails. Note that PEPIC has the lowest number of samples (n=130), so is difficult to fit with the OLS.

Model	Maize	Soybean	Rice	S. Wheat	W. Wheat
CARAIB	0.45 (0.35)	1.22 (0.50)	1.67 (0.32)	0.48 (0.26)	1.21 (0.47)
EPIC-TAMU	2.08 (0.61)	4.59 (0.20)	0.74 (0.27)	2.81* (0.20)	1.34 (0.14)
JULES	**7.46** (0.03)	**27.0** (10.21)	**11.8** (0.23)	**19.0** (5.23)	NA
GEPIC	3.13 (0.18)	1.35 (0.20)	3.12 (0.36)	2.28 (0.13)	3.55 (0.20)
LPJ-GUESS	0.66 (0.85)	NA	NA	1.13 (2.83)	0.67 (0.98)
LPJmL	2.04 (0.14)	1.52 (0.24)	1.09 (0.22)	0.51 (0.20)	0.65 (0.22)
pDSSAT	2.82 (0.52)	2.30 (0.25)	3.85 (2.30)	0.51 (0.20)	3.27 (0.97)
PROMET	2.45 (0.30)	2.96 (1.91)	**11.8** (1.87)	4.98 (0.13)	4.47 (0.21)
PEPIC	2.04* (0.79)	0.82* (3.27)	1.21* (1.82)	1.25* (0.54)	4.33* (0.51)

2. Out-of-sample validation. We also provide a second, more stringent test of emulator performance via a four-fold cross validation (also termed an out-of-sample validation). In this test the GGCMI Phase II dataset is split randomly into two parts, with 66% of the data used to train the model and the held-out 33% used to test the fidelity of the resulting emulator. We calculate the mean square error (MSE) between emulated (predicted) and actual simulated values across the test set, repeat the process three times so that test sets include all data randomly divided, and average the results of the three splits. As a last step, we normalize the MSE in each grid cell by dividing by its simulated yield in the baseline case (T+0, W+0, C=360, N=200). Geographic locations with very low baseline yields are especially problematic in this metric because minor disparities result in high errors. For this reason we mask all areas with less than 0.5 ton ha^{-1} baseline yields. We do not compare to yield *changes*; for more detailed potential evaluation metrics see (e.g Castruccio et al., 2014). Note that one aspect of this test can lead to increased errors and is not representative of the intended use of the emulator. The simple randomized sampling protocol for

dividing training and test sets can mean that the training set omits edge simulations (those at the highest or lowest value of CTWN space). The test prediction then involves extrapolating out of the training set range (e.g. predicting a T+6 case when the training set extends only to T+4), an improper use of an emulator. In this sense the metric is overly conservative, i.e. should lead to improved values than under a more realistic sampling strategy.

The resulting error metric is generally low as a percentage of yield. Table 3 shows MSE error by this metric over currently cultivated land for each model-crop combination for rainfed and irrigated crops. Means are weighted by cultivated area and include all simulations in CTWN space. Values are below 5% for all but six problematic model-crop combinations, highlighted in bold. These include the cases flagged previously (JULES for soybeans and spring wheat and PROMET for rice) but now also JULES for maize and rice. Error for irrigated emulators are considerably lower and below 1% in all but 7 crop-model combinations and below 5% for all cases except JULES soybean and winter wheat.

Errors are s; see supplemental Figures SXX-SXX for maps. Concluding statement.

3. Climate RCP validation. Finally, we test the ability of the emulator to reproduce a crop model response driven by an evolving climate scenario. The HADGEM2-ES climate model (Jones et al., 2011) shows comparably standard variability changes for growing season temperature compared to other CMIP-5 models. (See supplemental Figure SXX.) The emulator constructed from the GGCMI perturbed mean training set can reliably reproduce a realistic RCP climate scenario (Figure 9). This suggests that the temperature distribution in a location within the growing season is relatively insignificant when compared to the mean change. we show absolute global production to illustrate the impact of differences in baseline climate between AgMERRA and HADGEM2-ES... The growing season used for calculating the temperature and precipitation changes in the climate model run in especially important here because growing seasons vary dynamically in models based on temperature. In the case shown here, using the growing seasons from the baseline GGCMI Phase II simulations provides the best result because the emulator implicitly includes the model response to changing growing season. If the growing season from the RCP crop model is utilized (an unrealistic use-case because these simulations do not exist for every possible emulator application) you are likely to over- or under-estimate the growing season temperature or precipitation changes depending on the seasonality of weather in the growing season.

5 Emulator results and products

Because the emulator or "surrogate model" transforms the discrete simulation samples into a continuous response surface at any geographic scale, it can be used for a variety of applications, including construction of continuous damage functions in a flexible format. As an example, we present global damage functions constructed from the 4D emulation, for all four dimensions tested in this study (Figure 10) with the ensemble median and ensemble spread shown in bold line and ribbon. This is helpful in the crop model intercomparison project context for diagnosing model differences. In general, across model spread is qualitatively similar across different crops and different dimensions with some notable exceptions. Model spread is highest for spring wheat in general and the CO_2 response for the wheats and soybean. On the other side, muted responses include soybean, an efficient atmospheric nitrogen-fixer, is relatively insensitive to nitrogen, rice is not generally grown in

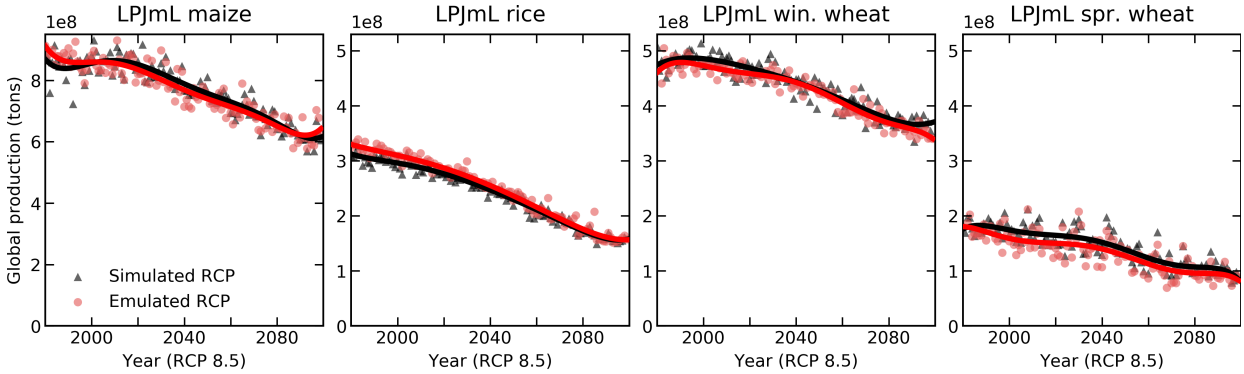

Figure 9. Simulated global production (black) and emulated global production (red) for four crops from the LPJmL model. Points show yearly global production values and lines show spline fit. Both simulation and emulation are driven with temperature and precipitation outputs from the HADGEM2-ES climate model (Jones et al., 2011) for RCP 8.5 with Nitrogen and CO_2 held fixed. The emulator trained on uniform climatological offset simulation outputs is able to capture the general response of the transient climate run in most cases. Notable exceptions include low simulated production maize for some years in the 1980s, overestimated historical rice, and a persistent low bias for emulated spring wheat.

water-limited conditions so it shows the lowest response to changes in precipitation, and maize has a muted response to CO_2 as a C4 plant.

Note that these functions are presented here in Figure 10 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 (e.g 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 bigger than 0.5 degrees in latitude and longitude.

Nitrogen response is relatively consistent across crops, with most models showing saturation at values less that 200 kg ha^{-1} Large increases in precipitation may cause yield losses that are not well-captured by current process-based crop models (Glotter et al., 2015; Li et al., 2019).

The emulator can also be used for investigating the contributions of the different major climate drivers to production outcomes as it can project many different climate scenarios or models quickly. The emulated crop model yield responses to a high-end climate change scenario (Representative Concentration Pathway (RCP) 8.5) are shown in Figure 11 for 5 climate models from the CMIP-5 archive (Taylor et al., 2012) at the decadal scale. See supplemental Figures SXX for a winter wheat example.

The differences between the emulation and the simulations for a uniform T shift are small compared to the differences across different crop models. PROMET, the quantitatively most difficult model to emulate for maize is shown in Figure 11 to illustrate that emulation error at the global production scale is still small compared to the spread across crop models or the spread across

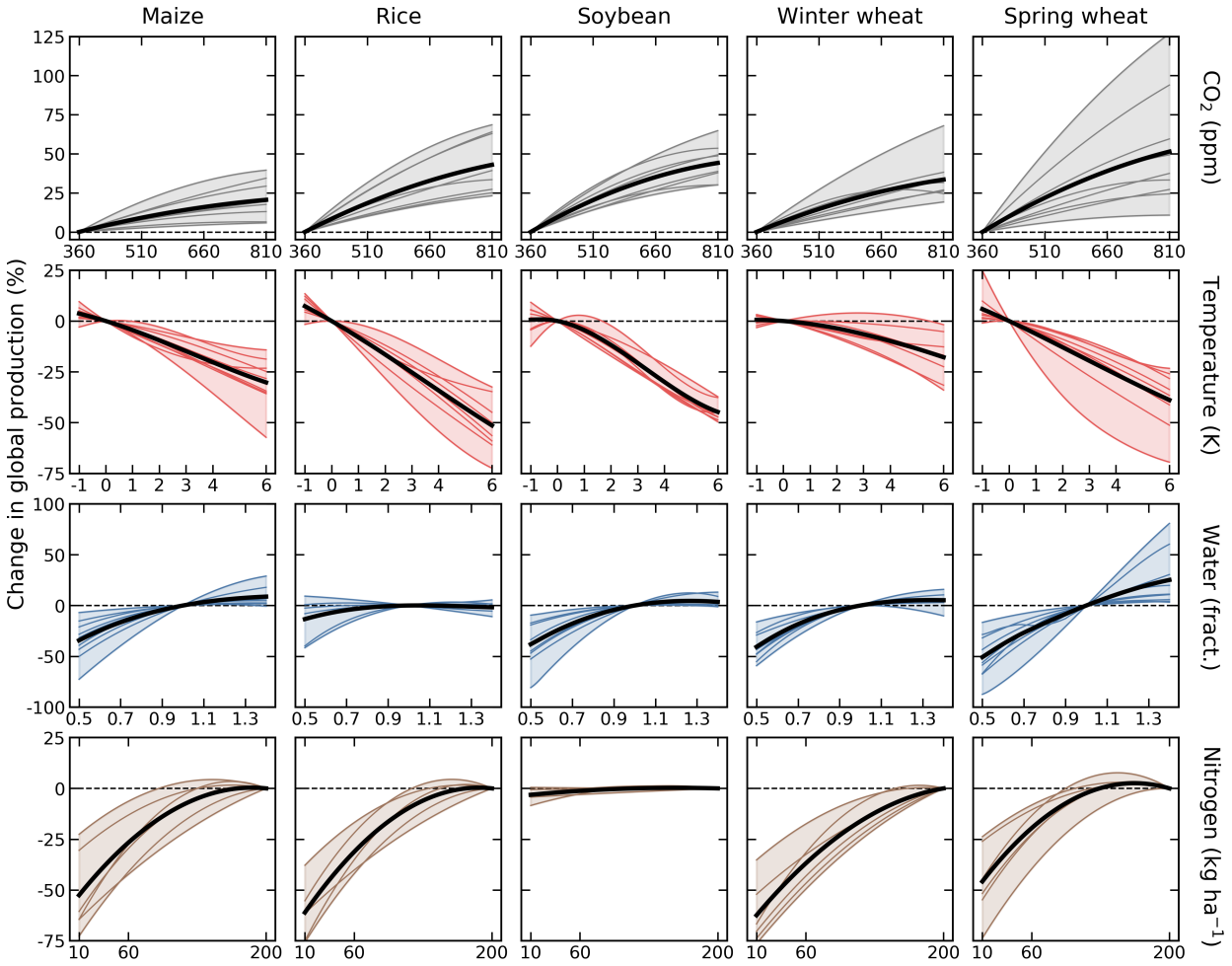

Figure 10. Emulated global damage functions for the five crops included in GGCMI Phase II, from the multi-model mean, for the four dimensions varied: CO_2, temperature, water, and nitrogen (collectively "CTWN"). Black line shows the multi-model ensemble mean and the shaded area and colored lines show the individual model projections. All other covariates held constant at baseline values (T+0K, W+0%, C = 360ppm, and N = 200kg ha-1). Damages are reported as percent change in global production relative to the baseline (1980-2010) case over currently cultivated land (Portmann et al., 2010).

climate projections when all factors are included. The uniform temperature shift over growing area is not very different from realistic scenarios. That is, projected temperature change distribution in space do not matter that much over currently cultivated area, but are likely important to production in high latitude regions which will warm much faster than lower latitudes.

Precipitation changes introduce some noise because different climate models have different P response for a given temperature change. Including the direct effects of elevated CO_2 introduces the largest intermodel uncertainty with different crop models showing very different responses. This difference in model response to CO_2 is much larger than differences resulting from different climate model sensitivity to CO_2 ($T(CO_2)$). Emulation based on GGCMI climatological mean shifts is useful

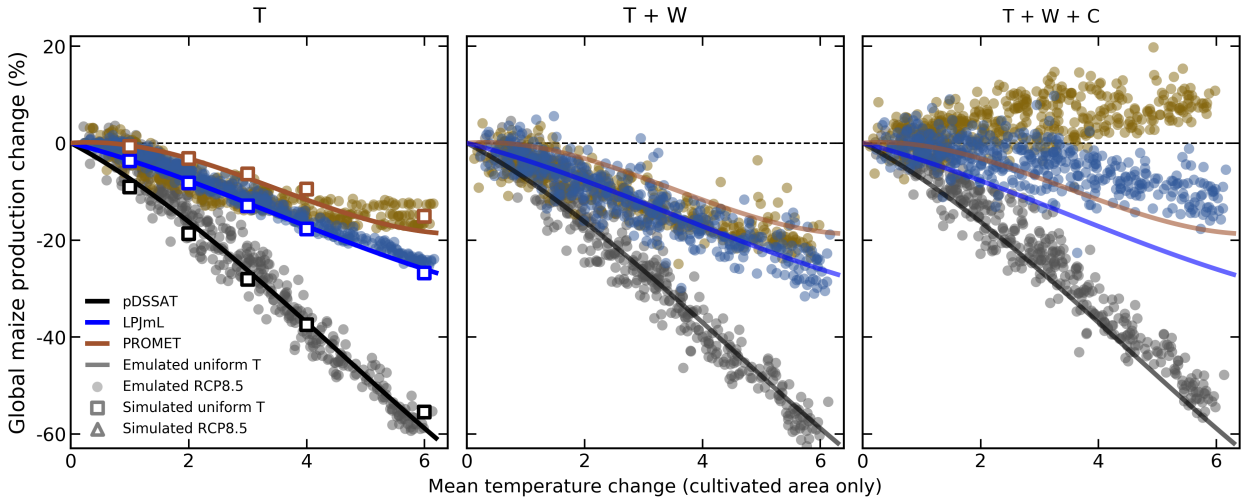

Figure 11. Illustration of the effects of factors affecting yields in more realistic climate scenarios. Figure shows emulated yield changes for maize on currently cultivated land under RCP8.5 (relative to 1980-2010 mean) for three representative crop models, with changes to T only (**a**), to T and W (**b**), and to T, W, and CO_2 (**c**). Circles are emulated yearly global production changes for 2010-2100 in scenarios from 5 CMIP-5 climate models, i.e. 50 total, with x-axis the mean T shift over all grid cells where maize are grown (unweighted by within-cell cultivated area). Bold lines are the emulated values over uniform T shifts. Open squares in panel **a** are GGCMI Phase II simulated values for each T level (with CWN at baseline). Emulations capture simulated behavior well (compare squares to lines), with the exception of PROMET at extreme temperature change. (See also Figure 5.) Mean yields are very similar whether T changes occur as a uniform temperature shift or in a more realistic spatial pattern (compare lines to circles). **b:** adding in projected precipitation changes depresses yields slightly for PROMET and increases spread between projections for a given temperature change for the other models. **c:** adding in CO_2 changes produces very different responses across models. CO_2 fertilization is small in pDSSAT, moderate in LPJmL, and very large in PROMET. Emulation uncertainty is small compared to the differences across climate and crop models.

for realistic scenarios and emulator errors are small compared to climate model differences and tiny compared to crop model differences.

6 Discussion and conclusions

We show that the systematic parameter sampling in the GGCMI Phase II experiments allow emulating climatological crop yield responses with a relatively simple reduced-form statistical model. The sampling provides information on the influence of multiple interacting factors in a way that realistic climate model simulations cannot, and allows isolating long-term impacts from confounding factors that lead to different year-over-year responses. The use of a relatively simple functional form in turn offers the possibility of physical interpretation of parameter values that can assist in model intercomparison and evaluation. The yield output for a single GGCMI Phase II model that simulates all scenarios and all five crops is $\sim$12.5 GB; the emulator is $\sim$20 MB, a reduction of nearly three orders of magnitude.

Several cautions should be noted when using the emulator. While the emulator allows estimating agricultural impacts under arbitrary climate scenarios, extrapolation outside the sample space should be avoided. Emulators by design reduce the complexity of process-based simulations and can thus deviate from the raw simulation output. This limitation is especially prominent in models with limited sampling or in geographic regions outside the current cultivated area. 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 they should not be used for impacts projections of absolute yields except in conjunction with historical yield information. Finally, because the GGCMI Phase II simulations apply uniform perturbations to historical climate inputs, they do not sample potential changes in climate variability. but this only matters so much... Although such changes 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 in the marginal distributions or temporal dependence of temperature and precipitation (e.g. Leeds et al. (2015); Poppick et al. (2016); Chang et al. (2016) and Haugen et al. (2018)).

The GGCMI Phase II dataset invites a broad range of potential future avenues of analysis, especially because emulation allows statistical distillation of the large dataset (40 billion simulated yields) into a tractable form. Potential studies might include a detailed examination of interaction terms between the major input drivers, robust quantification of model sensitivities to input drivers, exploration of yield responses to extremes, and evaluation of geographic shifts in optimal growing regions. The dataset also enables studies of emulation itself, including a more systematic evaluation of different statistical and machine learning model specifications. In general, the development of multi-model ensembles involving systematic parameters sweeps has large promise for better understanding potential future crop responses and for improving process-based crop models.

Code and data availability. The polynomial emulator parameter matrices for all crop model emulators are available at doi.org/XXXXX

Author contributions. J.E., C.M, A.R., J.F., and E.M. designed the research. C.M., J.J., P.F., C.F., L.F., R.C.I., I.J., C.J., W.L., S.O., M.P., T.P., A.Re., 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., C.M., 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). K.W. was supported by the IMPREX research project supported by the European Commission under the Horizon 2020 Framework programme, grant #641811. 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 Extension, Texas A & M 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). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. (DGE-1746045). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

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.

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.

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.

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., Wheeler, T., Craufurd, P., Slingo, J., and Grimes, D.: Design and optimisation of a large-area process-based model for annual crops, Agricultural and Forest Meteorology, 124, 99 – 120, https://doi.org/https://doi.org/10.1016/j.agrformet.2004.01.002, http://www.sciencedirect.com/science/article/pii/S0168192304000085, 2004.

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.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society, 137, 553–597, 2011.

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.

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.

Franke, J., Müller, C., Elliott, J., Ruane, A., Snyder, A., Jägermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P., Folberth, C. François, L., Hank, T., Hoffmann, M., Izaurralde, R., Jacquemin, I., Jones, C. Khabarov, N., Koch, M., Li, M. Liu, W., Olin, S., Phillips, M., Pugh, T.,

Reddy, A., Wang, X., Williams, K., Zabel, F., and Moyer, E.: The GGCMI phase II experiment: global gridded crop model simulations under uniform changes in CO_2, temperature, water, and nitrogen levels (protocol version 1.0)., Geoscientific Model Development, in open review, 2019a.

Franke, J., Müller, C., Elliott, J., Ruane, A. C., Jagermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P., Folberth, C., Francois, L., Hank, T., Hoffmann, M., Izaurralde, R. C., Jacquemin, I., Jones, C., Khabarov, N., Koch, M., Li, M., Liu, W., Olin, S., Phillips, M., Pugh, T. A. M., Reddy, A., Wang, X., Williams, K., Zabel, F., and Moyer, E.: The GGCMI Phase II experiment: global gridded crop model simulations under uniform changes in CO_2, temperature, water, and nitrogen levels (protocol version 1.0), Geoscientific Model Development Discussions, 2019, 1–30, https://doi.org/10.5194/gmd-2019-237, https://www.geosci-model-dev-discuss.net/gmd-2019-237/, 2019b.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., van Vliet, M., Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and Yamagata, Y.: Assessing the impacts of 1.5°C global warming — Simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), Geosci. Model Dev., 10, 4321–4345, https://doi.org/10.5194/gmd-10-4321-2017, 2017.

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.

Gadgil, S., Rao, P. S., and Rao, K. N.: Use of climate information for farm-level decision making: rainfed groundnut in southern India, Agricultural Systems, 74, 431 – 457, https://doi.org/10.1016/S0308-521X(02)00049-5, 2002.

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.

Hansen, J. and Jones, J.: Scaling-up crop models for climate variability applications, Agricultural Systems, 65, 43 – 72, https://doi.org/10.1016/S0308-521X(00)00025-1, 2000.

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.

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.

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.

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.

Jones, C. D., Hughes, J. K., Bellouin, N., Hardiman, S. C., Jones, G. S., Knight, J., Liddicoat, S., O'Connor, F. M., Andres, R. J., Bell, C., Boo, K.-O., Bozzo, A., Butchart, N., Cadule, P., Corbin, K. D., Doutriaux-Boucher, M., Friedlingstein, P., Gornall, J., Gray, L., Halloran, P. R., Hurtt, G., Ingram, W. J., Lamarque, J.-F., Law, R. M., Meinshausen, M., Osprey, S., Palin, E. J., Parsons Chini, L., Raddatz, T., Sanderson, M. G., Sellar, A. A., Schurer, A., Valdes, P., Wood, N., Woodward, S., Yoshioka, M., and Zerroukat, M.: The HadGEM2-ES implementation of CMIP5 centennial simulations, Geoscientific Model Development, 4, 543–570, https://doi.org/10.5194/gmd-4-543-2011, 2011.

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.

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.

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., and Peng, B.: Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States, Global Change Biology, 25, 2325–2337, https://doi.org/10.1111/gcb.14628, 2019.

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, B., Asseng, S., Müller, C., Ewert, F., Elliott, J., Lobell, D. B., Martre, P., Ruane, A. C., Wallach, D., Jones, J. W., et al.: Similar estimates of temperature impacts on global wheat yield by three independent methods, Nature Climate Change, 6, 1130, 2016a.

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, 2016b.

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, 2016c.

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, World Scientific Publishing Co, 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.

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.

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.

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.

Potter, N. J., Zhang, L., Milly, P. C. D., McMahon, T. A., and Jakeman, A. J.: Effects of rainfall seasonality and soil moisture capacity on mean annual water balance for Australian catchments, Water Resources Research, 41, https://doi.org/10.1029/2004WR003697, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004WR003697, 2005.

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.

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.

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.

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.

Shapiro, S. and Wilk, M.: An analysis of variance test for normality (complete samples)†, Biometrika, 52, 591–611, https://doi.org/10.1093/biomet/52.3-4.591, 1965.

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, Geoscientific Model Development, 12, 1319–1350, https://doi.org/10.5194/gmd-12-1319-2019, https://www.geosci-model-dev.net/12/1319/2019/, 2019.

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.

Urban, D., Roberts, M. J., Schlenker, W., and Lobell, D. B.: Projected temperature changes indicate significant increase in interannual variability of U.S. maize yields: A Letter, Climatic Change, 112, 525–533, https://doi.org/10.1007/s10584-012-0428-2, 2012.

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.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 50, 7505–7514, 2014.

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.

Zabel, F., Delzeit, R., Schneider, J. M., Seppelt, R., Mauser, W., and Vàclavík, T.: Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity, Nature Communications, 10, 2844, https://doi.org/10.1038/s41467-019-10775-z, http://www.nature.com/articles/s41467-019-10775-z, 2019.

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.