Emulating the GGCMI Phase II experiment: global gridded emulators of crop model responses to changes in CO2, temperature, nitrogen, and water (protocol version 1.0)

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

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

Correspondence: James Franke (jfranke@uchicago.edu)

Abstract. Statistical emulation of process-based crop models provides the opportunity to combine some of the 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. 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. The GGCMI Phase II systematic parameter sweep

protocol allows disentangling the climate-driven mean response from year-over-year variations; we show that the two responses have very different relationships to standard climate metrics such as mean growing season temperature. The climatological mean yield response can be readily represented with a simple polynomial in almost all locations where crops are currently grown, permitting a tool that captures model responses in a lightweight, computationally tractable form. Crop model emulation should therefore facilitate both model comparison and integrated assessment of climate impacts.

1 Introduction

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 larger 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).

Statistical emulation allows combining some of the advantageous features of both statistical and process-based models. Crop model emulators allow representing process-based crop model responses in a computationally-inexpensive form appropriate for economic assessments, and assist in crop model intercomparison and improvement efforts. 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. Emulation allows producing 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), who used a machine learning approach to predict Mediterranean wheat yields. Studies developing single-model emulators include Holzkämper et al. (2012) for the CropSyst model, Ruane et al. (2013) for the CERES wheat model, and Oyebamiji et al. (2015) for the LPJmL model. More

recently, emulators have begun to be used in the context of multi-model intercomparisons, 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 Impacts Model Intercomparison Project (ISIMIP). ISIMIP offers a relatively large traning 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 (e.g. CO_2 and temperature). such as those of ISIMIP. 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 (Fronzek et al., 2018; Snyder et al., 2018; Ruiz-Ramos et al., 2018). All three 2018 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 offer global coverage, but 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. (2018) analyzes four crops (maize, wheat, rice, soy) but only in one crop model (GCAM) (Calvin et al., 2019).

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 temperature, appreciation, CO_2, and applied nitrogen. 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

Table 1. Crop models included in GGCMI Phase II emulators and the number of CTWN-A 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. (2019). The three simulation models not included in the emulators have been removed for clarity.

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. (2016a, b)	X	X	X	X	X	X	149 / 121
PROMET, Hank et al. (2015); Mauser et al. (2015)	X	X	X	X	X	X	261 / 232

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

The GGCMI Phase II simulations are described in detail in Franke et al. (2019), but we summarize briefly here. The experiment involves nine different globally gridded crop models, each simulating multiple crops (maize, rice, soy, and spring and winter wheat) across a parameter sweep of as many as 1400 scenarios, each driven by a historical climate timeseries with systematic perturbations to CO_2, temperature, precipitation, and nitrogen application (CTWN). Table 1 shows the participating models

Table 2. GGCMI Phase II input levels for the parameter sweep. Temperature and precipitation values are the perturbations from the historical climatology. W-percentage does not apply to the irrigated (W_∞) simulations, which are all simulated at the maximum beneficial levels of water. Bold font indicates the 'baseline' historical level. One model provided simulations at the T + 5 level. The full protocol samples across all parameter combinations for a total of 756 cases. See Figure SX in the supplement for number of simulations associated with each combination of input levels. Table repeated from Franke et al. (2019)

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	-

and the scenarios that each provides, and Table 2 shows the specified 4 levels of atmospheric CO_2, 7 of temperature, 8 of precipitation, 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 four times, for irrigated and rainfed crops, and for fixed growing seasons and growing seasons shortening in warmer climates. (Fixed and variable growing-season scenarios are termed A1 and A0, respectively.) The complete protocol for a modeling group involves up to 19,440 years of global simulated output for rainfed crop and growing-season case; or 42,120 years across all simulations for a single modeling group. Because the computational demand is high, modeling groups were allowed to submit at various specified levels of participation, with the lowest level of participation consisting of 20% of the maximum possible simulations; the mean participation level is 65% of max potential runs.

Each individual crop model simulation is run for 30 years over historic weather for the period of 1981-2010, 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). Temperature perturbations are applied as additive mean shifts (from -1 to +6 degrees C), precipitation as fractional multipliers (from -50% to +30%), and CO_2 and nitrogen application as fixed values. (See Table 2 for all values.) Models provide global output at 0.5 degree latitude and longitude resolution for each simulation year.

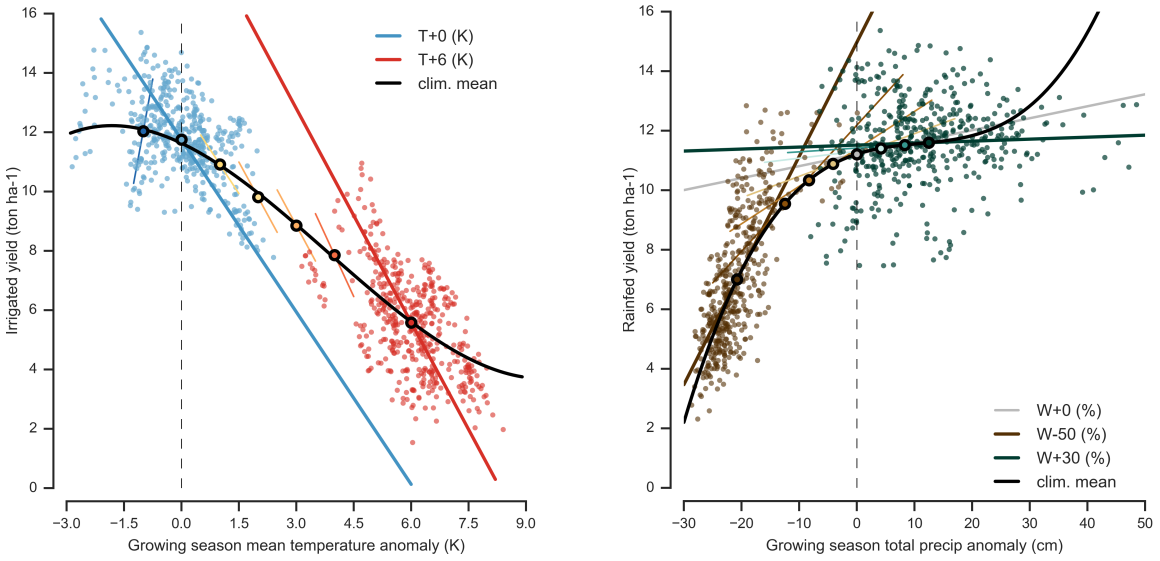

Figure 1. Example showing distinction between crop yield responses to year-to-year and climatological mean temperature and precipitation shifts. Figures shows maize for a representative high-yield region (nine adjacent grid cells in northern Iowa) from the pDSSAT model *Left*: irrigated maize, climatological mean values for all temperature cases (T-1, +0, +1, +2, +3, +4, +6) with other variables held at baseline values, and individual years for T+0, blue, and T+6 K, red. *Right*: rainfed maize, all precipitation cases (W -50%, -30%, -20%, -10%, W, +10%, +20%, +30%), with individual years shown for W-50%, grown, and W+50%, green. Open black circles mark climatological mean yields for each case and short colored lines the linear regression to individual yields in each case. Bold black lines show 3rd order polynomial fits through climatological mean values. In the temperature case, year-over-year yield responses are very different from the response to longer-term climate perturbations, and rise under warmer conditions. In the precipitation case, year-over-year responses resemble the climatological mean response, but the response is highly nonlinear.

2.2 Climatological vs. year-to-year response

you're saying that the year-over-year would add noise but not change your fit. But, I think this is why you talk about the nonlinearity of the year-over-year response, because in some cases it WOULD add bias. right? If the year-over-year response is the same as the mean response it's fine. If it's different from the mean response and not changing, I think even that adds bias, because your "noise" is then not stationary - its relationship to the mean response differs depending on temperature. If it's different from the mean response and changing, then it's definitely not stationary.

so I think the organizational flow is 1) we emulate the climatological mean response, because that's what we care about, 2) the year-over-year response is not the same as the climatological response (Figure 2), 3) if we included the year-over-year response it's both noisier, making emulation harder, and can also actually bias the emulated response

then you close by saying the GGCMI dataset lets you use a relatively simple form that converges well. That transitions you to the emulation section

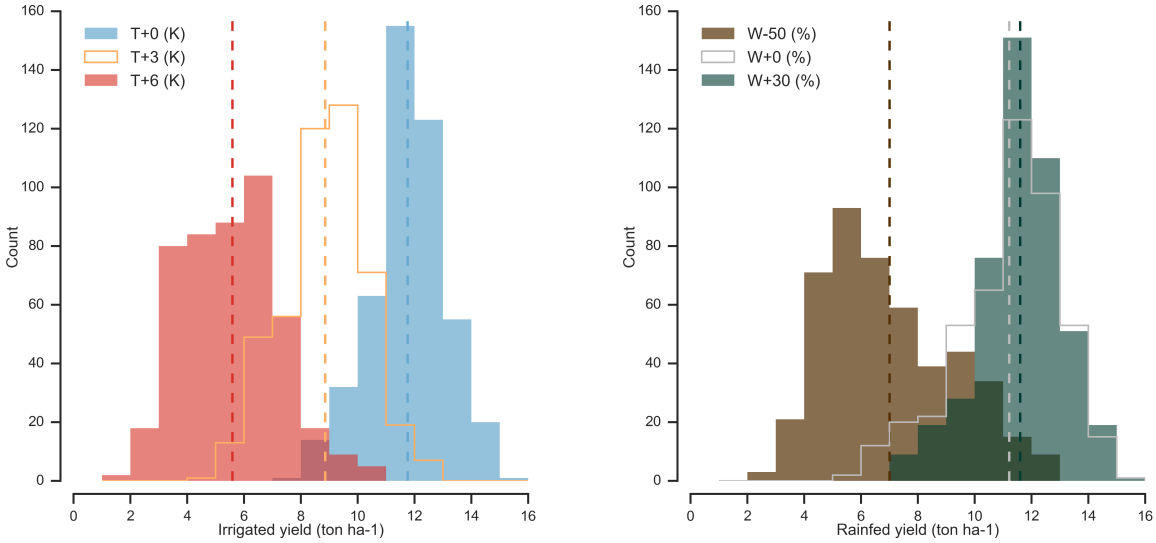

Figure 2. Example showing results of increased crop yield sensitivity to year-over-year climate variations under climate stress. Figure shows distributions of yields from examples of Figure 1, of irrigated (left) and rainfed (right) maize in Iowa in scenarios of altered temperature (left) and precipitation (right). Under large warming (T+6) or drying (P-50%), increased sensitivity means that distributions of year-over-year crop yields widen relative to present-day simulations, even though input climate has identical variance in climate drivers.

We emulate the yield response to C, T, W, and N at the 30-year climatological mean level. Three important distinctions are made between the year-to-year yield response and the climatological-mean yield response. First, the year-over-year responses to weather are generally quantitatively distinct from (and, in the case of temperature, larger than) the climatological mean responses in the GGCMI Phase II simulation output dataset at each level. In the example in Figure 1 for irrigated maize in Iowa, responses to year-over-year temperature variations are 100% larger than those to long-term perturbations in the baseline case, and larger still under warmer conditions, rising to nearly 200% more in the T+6 case. The same general principal applies to a lesser degree the precipitation dimension however in this case it is more manifestly an artifact of sampling range in the historical period (Figure 1, right). Year-over-year and climatological responses can differ for many reasons including memory in the crop model, lurking covariates, and differing associated distributions of daily growing-season daily weather (e.g. Ruane et al., 2016). Crops are less sensitive to short-term precipitation changes, since they care about soil moisture and the soil integrates. Previous work has shown that in all but very arid regions, crop responses in models are not sensitive to how the precipitation is distributed within a month (Glotter et al., 2014).

Second, the year-over-year response is not constant across different levels of climate stress. For example, the stronger year-over-year response to temperature under warmer conditions also manifests as a wider distribution of yearly yields (Figure 2) within a 30-year simulation even though the variance in input temperates is unchanged. Likewise for precipitation changes, strong decreases in precipitation results in a widening of the distribution in yields even though the variance in precipitation is decreasing (from approximately 60 cm yr-1 to 30 cm yr-1 in this example). In both cases show in the Figure (Figure

2), the year-over-year response is nonlinear, getting stronger with more severe mean climate shifts. The strengthening of crop sensitivity under severe climate shifts causes distribution of modeled crop yields becomes wider, even when underlying climate variability is fixed. For temperature, that strengthening year-over-year response adds a complication to emulating when using a non-stationary climate scenario where mean climate gradually warms. In contrast for precipitation, the year over year response is locally similar to the climatological response. Note that since we emulate mean climatological-yields, we do not capture any of these higher-order moment changes in yield in this study.

Finally, the climatological-mean yield response is considerably smoother than the year-to-year yield response and makes emulation relatively simple. Note that the GGCMI Phase II training dataset does not capture potential changes in climate variability, because all simulations are run with fixed offsets from the historical climatology. However, prior work has suggested that mean changes are the dominant drivers of climatological crop yield shifts in all but arid regions (e.g. Glotter et al., 2014). Critically for emulation, the mean-climatological yield can easily be related to the mean-climatological shift in temperature or precipitation in the GGCMI Phase II dataset.

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 irrigation water demand in both the A0 and A1 growing season, meaning that each model and crop combination results in six regressions. (See Supplementary Material XX 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 includes 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 34 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 10 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 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.

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

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, PEPIC). Feature importance is not uniform due to spatial heterogeneity across models and crops, so we weight the loss function by current cultivation area 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 indicated by the terms where the error is reduced and where additional terms provide no predictive benefit (line slopes match in Figure 3). The feature selection process results in a final polynomial in 23 terms, with 11 terms eliminated. We omit the N^3 term, which cannot be fitted because we sample only three nitrogen levels. We eliminate many of the C terms: the cubic, the CT, CTN, and CWN interaction terms, and all higher order interaction terms in C. Finally, we eliminate two 2nd-order interaction terms in T and one in W. Implication of this choice include that nitrogen interactions are complex and important, and that water interaction effects are more nonlinear than those in temperature.

3.3 Model fitting

To fit the parameters K, we use a Bayesian Ridge regularization method (MacKay, 1991), which reduces volatility in parameter estimates when the sampling is sparse, by weighting parameter estimates towards zero. The choice results in a reduction in mean absolute error for some of th high-order interaction terms in the model (top row of Figure 3) and drastically reduces standard parameter error in the model by stabilizing the estimates. The estimation method scores relatively lower on adjusted R^2 (equation 2, where n is the number of samples and k is the number of features) for the simplest parameter specifications, but reaches parity with the OLS at the number of terms included in this study. The Bayesian Ridge method is necessary over the standard OLS to maintain a consistent functional form across all models and locations (see Table 3).

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

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 one of the crop models shown in Figure 3 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.

We use the implementation of the Bayesian Ridge estimator from the scikit-learn package in Python (Pedregosa et al., 2011). In the GGCMI Phase II experiment, the most problematic fits are those for models that provided a limited number of cases or for low-yield geographic regions where some modeling groups did not run all scenarios. We do not attempt to emulate models

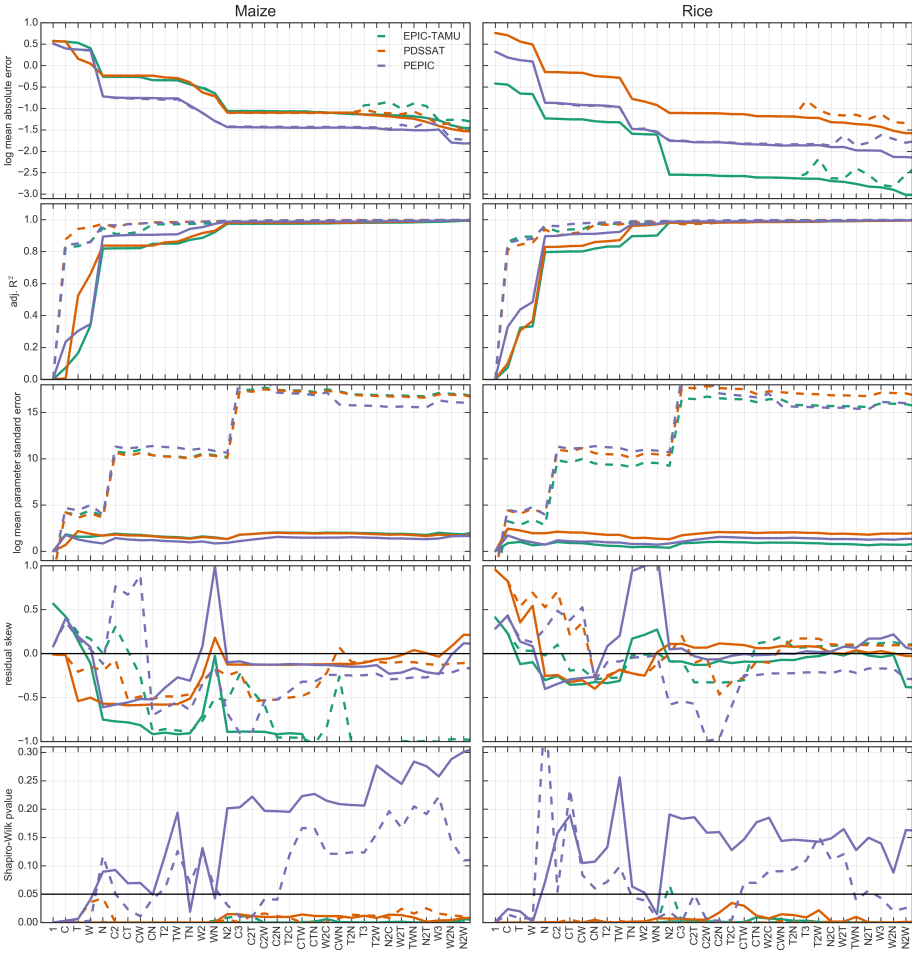

Figure 3. Summary results from polynomial feature selection process. The top row illustrates log mean absolute error between emulated yield and simulated values calculated with a three fold cross validation process, where the emulator is trained on two thirds of the data and predicts the remaining third. The second row illustrates the adjusted R^2 score for the fit at each model specification where additional terms are penalized. The third row illustrates the log mean standard parameter error and the forth and fifth rows illustrate the distribution of the residuals. The X- axis indicates terms included in the model at each step progressively where T = temperature, T2 = temperature^2, TW = temperature * water and so on. The terms that did not reduce the aggregate error (horizontal lines) are not included in the final model. Solid lines indicate Bayesian Ridge regression results, dashed lines indicated standard ordinary least squares and colors indicated three different crop models.

that provided less than 50 simulations. The lowest number of simulations emulated across the full parameter space is then 120 (for the PEPIC model).

4 Emulator evaluation

In the following section we present some illustrations of the yield response to drivers and the emulator performance. In general, model emulation with the parametric method used here is only possible when the crop yield responses are sufficiently smooth and continuous to allow fitting with a relatively simple functional form, but this condition largely holds in the GGCMI Phase II simulations. However, the principle of no-free-lunch applies, so losses between the emulator and simulations are great enough to be problematic in certain cases, especially for models with limited sampling or in geographic locations where crops are currently not grown. The errors between the simulation and the emulation are generally small compared to the difference across different crop models or across the response to different climate model inputs.

4.1 Yield response

Yield patterns in the GGCMI Phase II archive are geographically diverse. The emulator is able to successfully capture the spatial pattern of yields for a single simulation case and emulation errors are low in regions currently under cultivation (Figure 4). In this example, nearly all grid cells have errors below 0.5 ha^{-1} for the baseline case. However, emulation errors as a percentage of baseline yield do reach high values in some areas with no cultivation in the real world. Many of these regions are not currently viable for agriculture (and therefore have very low simulated baseline yields) and may never become viable even under extreme climate change. Some differences in spatial skill exist across models and crops, with maize being the qualitatively easiest to emulate across all models. See supplemental Figures SXX-SXX for more crops and models examples.

Yield responses to the four main drivers considered here (C, T, W, and N) are quite diverse across locations, crops, and models, but in most cases the local climatological mean responses are smooth enough to permit emulation with the functional form used here. Geographic diversity is high within a single crop and model (Figure 5, rainfed maize in pDSSAT); this heterogeneity supports the choice of emulating at the grid cell level. Yields evolve smoothly across the space sampled, and the polynomial fit captures the climatological-mean response to perturbations. Crop yield responses generally follow similar functional forms across models, though with a spread in magnitude partly due to the lack of calibration. Inter-model diversity for a single crop and location is also high (Figure 6, rainfed maize in northern Iowa, also shown in Figure 5). Differences in response shape can lead to differences in the fidelity of emulation, though comparison here is complicated by the different simulation experiment sampling regimes across models. Note that models are most similar in their responses to temperature perturbations.

The emulator fails when the yield response has a discontinuity or is irregular in some regard. For example, some simulation models report no yield under current conditions (too cold) and continue to report no yield until a simulation case with a significant amount of warming. Under a warming of several degrees, agriculture is now viable in this model and non-zeros yields are returned (Figure 5, PROMET in gray). Under these conditions, the 3rd order polynomial cannot fit the data, and errors are high. Baring simulation model errors, theses locations are almost exclusively confined to areas with little to no cultivation in the real world. This is quantified in more detail in the following section.

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 three nitrogen levels (10, 60 and 200 kg N

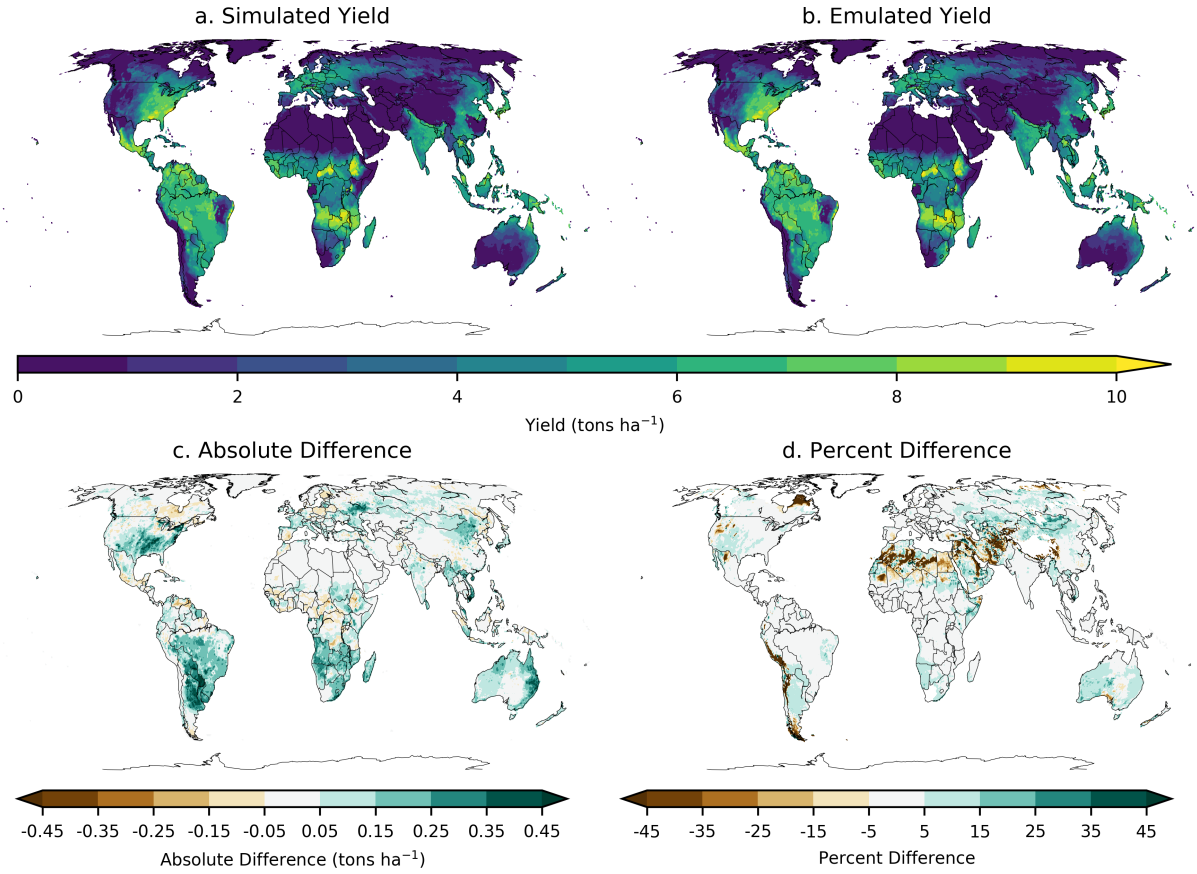

Figure 4. Illustration of spatial pattern in baseline yield successfully captured by the emulator. Simulated and emulated yield under historical (1981-2010) conditions for rainfed maize from the LPJmL model. Absolute yield differences are less that 0.5 ton ha^{-1} in almost all (99.8%) grid cells across the globe. Percent differences (from simulated baseline) 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 supplemental for more examples.

y^{-1} ha^{-1}), so a third-order fit would be over-determined but a second-order fit can result in potentially unphysical results. Steep and nonlinear declines in yield with lower nitrogen levels mean that some regressions imply a peak in yield between the 100 and 200 kg N y^{-1} ha^{-1} levels. While it is possible that over-application of nitrogen at the wrong time in the growing period could lead to reduced yields, these relative strength of this feature is are potentially an artifact of the fit. The Bayesian Ridge estimator (shown in Figure 5) tends to mitigate the 'peak-decline effect' in the nitrogen dimension compared to ordinary least squares. In addition, the polynomial fit cannot capture the well-documented saturation effect of nitrogen application (e.g. Ingestad, 1977) as accurately as would be possible with a non-parametric model.

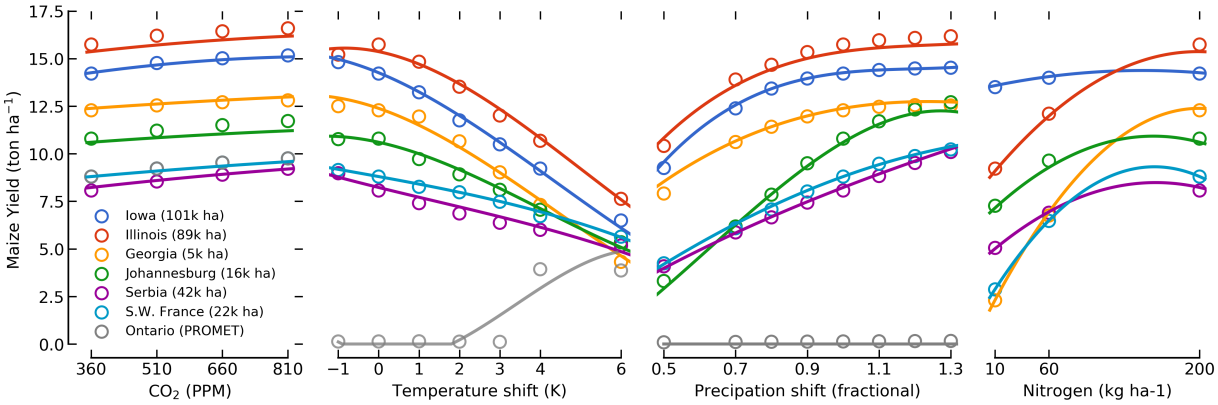

Figure 5. Illustration of spatial variations in yield response successfully captured by the emulator. Figures show rainfed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares cultivated in each selected grid cell. Each panel shows variation along a single variable, with others held at baseline values. Dots show climatological mean yields and lines the results of the full 4D emulator of Equation 1. In general the climatological response surface is sufficiently smooth that it can be represented within the sampled variable space by the simple polynomial used in this work. Extrapolation can however produce misleading results. The rainfed maize response in north-central Ontario is shown for the PROMET model as a good example of an area where the emulator fails.

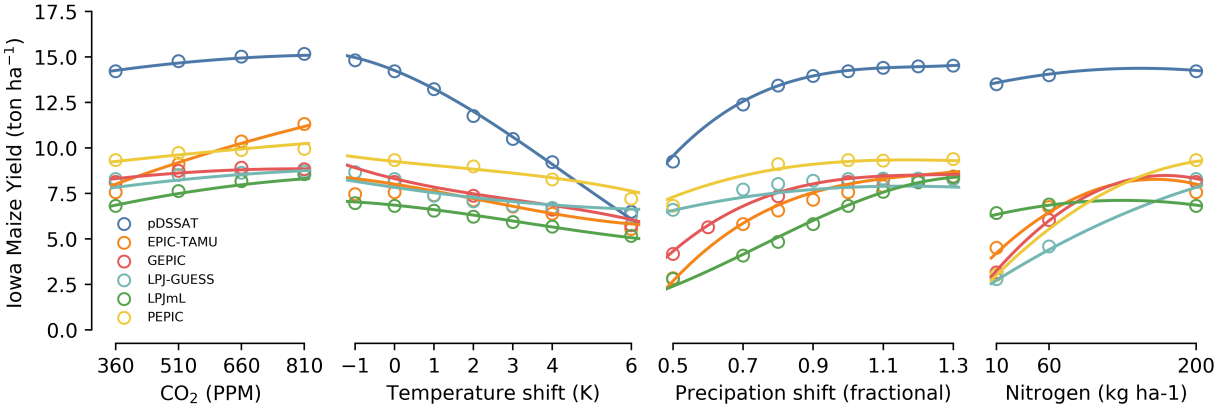

Figure 6. Illustration of across-model variations in yield response successfully captured by the emulator. Figures shows simulations and emulations from six models for rainfed maize in the same Iowa grid cell shown in Figure 5, with the same plot conventions. Models that do not simulate the nitrogen dimension are omitted for clarity. Note that models are uncalibrated, increasing spread in absolute yields. While most model responses can readily emulated with a simple polynomial, some response surfaces diverge slightly from the polynomial approach (e.g. LPJ-GUESS here) and lead to emulation error, though error generally remains small relative to inter-model uncertainty.

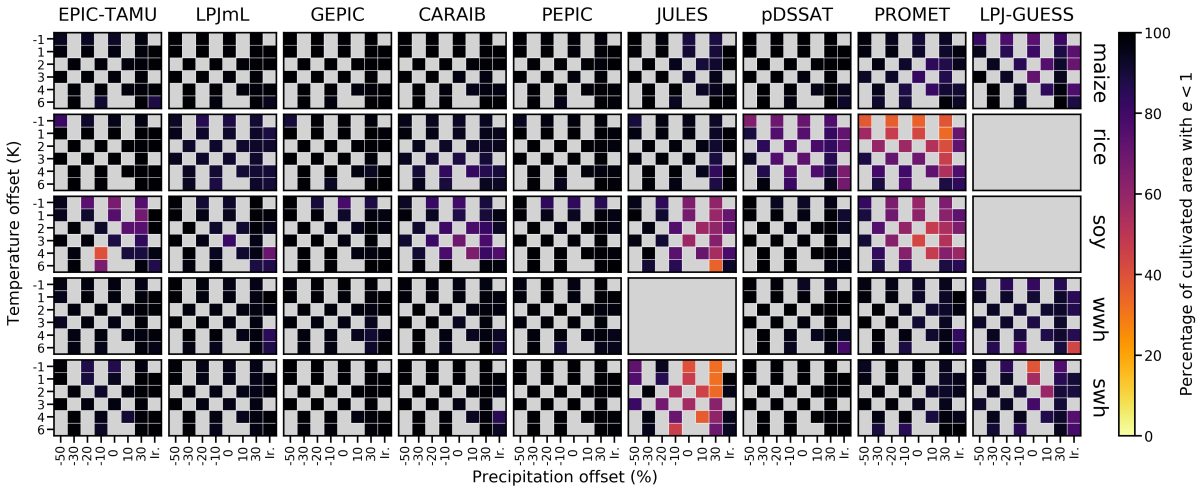

Figure 7. Assessment of emulator performance over currently cultivated areas based on normalized error (Equations 4, 3). We show performance of all 9 models emulated, over all crops and all sampled T and P inputs, but with CO_2 and nitrogen held fixed at baseline values. Large columns are crops and large rows models; squares within are T, P scenario pairs. Colors denote the fraction of currently cultivated hectares ("area frac") for each crop with normalized area e less that 1 indicating the the error between the emulation and simulation less that one standard deviation of the ensemble simulation spread. Of the 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 soy. Emulator performance is generally satisfactory, with some exceptions. Emulator failures (significant areas of poor performance) occur for individual crop-model combinations, with performance generally degrading for hotter and wetter scenarios.

4.2 Emulator performance metrics

Our emulators consist of nearly 3 million individual regressions, so presenting concise performance metrics poses a challenge. Additionally, no general agreed upon criteria exist for defining an acceptable crop model emulator, so we present two different metrics: one operational and one more stringent. First, for a multi-model comparison exercise like GGCMI Phase II, one reasonable criterion is what we term the "normalized error", which compares the fidelity of an emulator for a given model and scenario to the inter-model uncertainty. Second, we show a standard out-of-sample cross-validation aggregate mean error for each model independent of the other members of the ensemble.

For the first metric, we define the normalized error e for each scenario. The normalized error e is the difference between the emulated fractional change in yield and that actually simulated, normalized by the standard deviation in simulated fractional yield changes across all models for that scenario (Equations 3 and 4).

$$F_{scn.} = \frac{Y_{scn.} - Y_{baseline}}{Y_{baseline}} \tag{3}$$

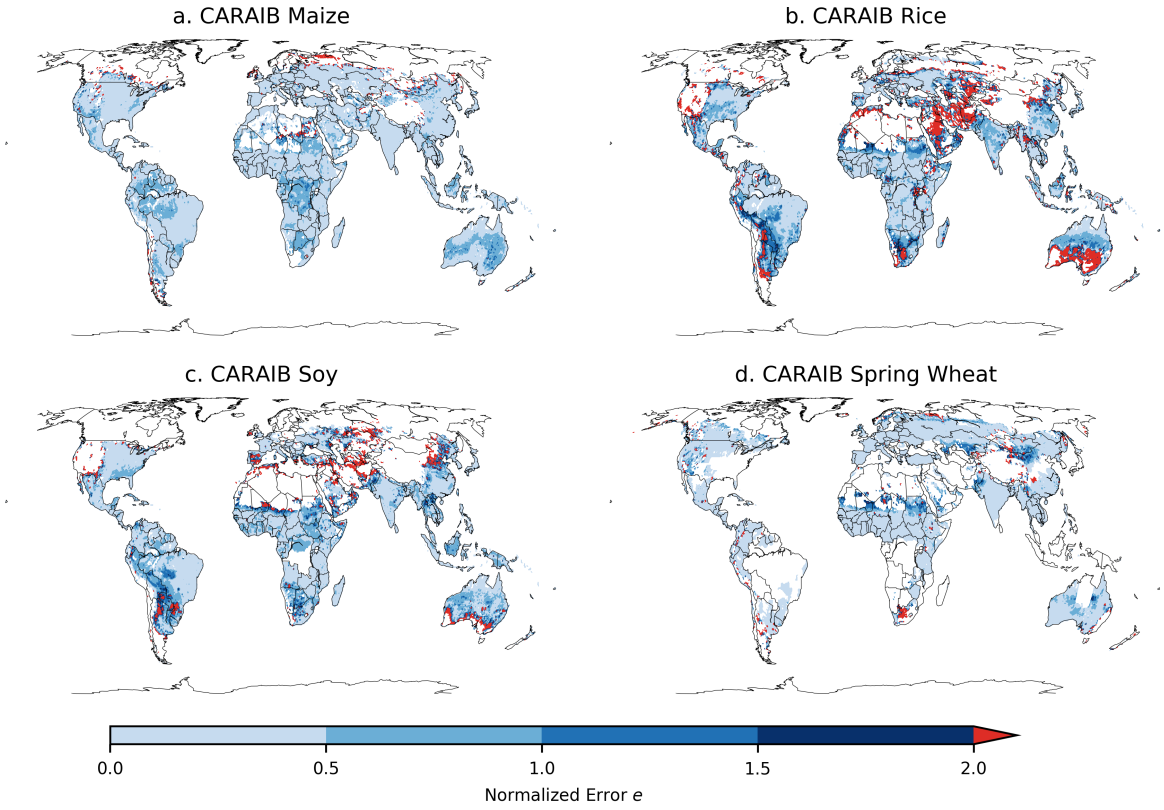

Figure 8. Illustration of our test of emulator performance, applied to the CARAIB model for the T+4 scenario for rainfed crops. Contour colors indicate the normalized emulator error e, where $e > 1$ means that emulator error exceeds the multi-model standard deviation. White areas are those where crops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

$$e_{scn.} = \frac{F_{em,\,scn.} - F_{sim,\,scn.}}{\sigma_{sim,\,scn.}} \tag{4}$$

Here $F_{scn.}$ is the fractional change in a model's mean emulated or simulated yield from the defined historical baseline, in a certain setting or scenario (scn.) in C, T, W, and N space; $Y_{scn.}$ and $Y_{baseline}$ are the absolute emulated or simulated mean yields. The emulator is fitted across all available simulation outputs for each grid cell, model, and crop, and then the error is calculated across the each of the simulation scenarios provided by all nine models (See Figure SX for number of simulations in each case).

A normalized error $e < 1$ indicates that the emulator is closer to the simulation than 1 standard deviation of the simulations in that scenario. This metric implies that emulation is generally satisfactory, with almost all model-crop combination emulators

Table 3. Mean absolute error of emulator representation of a simulation as a percentage of baseline simulated yield for the cross-validation process for rainfed crops. A 4-fold stratified k-fold cross validation scheme is utilized where the model is trained on 75% of the data and validated on the held-out 25% (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 mean error (as a percentage of baseline yield) weighted by hectares grown in each grid cell (Portmann et al., 2010). * Indicates cases where the OLS linear model fails.

Model	Maize (%)	Soy (%)	Rice (%)	S. Wheat (%)	W. Wheat (%)
CARAIB	1.15	4.20	2.33	0.36	0.53
EPIC-TAMU	2.46	3.35	0.56	2.10*	1.38
JULES	3.99	20.1	2.89	9.90	NA
GEPIC	2.97	0.66	2.17	1.60	2.04
LPJ-GUESS	0.57	NA	NA	xxxx	0.18
LPJmL	2.04	1.52	1.09	0.51	0.65
pDSSAT	2.91	1.95	2.23	0.44	1.79
PROMET	4.41	7.48	6.06	16.8	7.07
PEPIC	1.30*	0.72*	1.16*	0.63*	1.65*

have normalized errors less than one over nearly all currently cultivated hectares (Figure 8, dark colors indicate good performance and light colors poor performance). Exceptions include a few individual model-crop combinations, which are difficult to emulate with the polynomial used here including PROMET (as previously mentioned, see also Figure 5) for rice and soy, JULES for soy and spring wheat. Problems with emulating PROMET for rice and soy may have to do with the parametrization of the phenology for those crops which lengthens the growing season in some cases. Another reasons why emulation can be problematic in some models has to do with a saturation in yield reduction under temperature increases that cannot be easily captured with the 3rd order fit. Emulator performance also often degrades in general in geographic locations where crops are not currently cultivated. For example, emulator performance may be satisfactory over currently cultivated areas for all crops from a model, but uncultivated regions may show some problematic areas (Figure 8 shows a CARAIB model case, see also Figure SXX-SXX for other models).

The normalized error assessment procedure is relatively forgiving for several reasons. First, each emulation is evaluated against the simulation actually used to train the (in-sample validation). Had we used a spline interpolation the error would necessarily be zero. Second, the performance metric scales emulator fidelity not by the magnitude of yield changes but by the inter-model spread in those changes. The normalized error e for a model depends not only on the fidelity of its emulator in reproducing a given simulation but on the particular suite of models considered in the intercomparison exercise. Where models differ more widely, the standard for emulators becomes less stringent. For example, normalized errors for soy are somewhat higher across all models not because emulator fidelity is worse but because models agree more closely on yield changes for soy than for other crops (see Figure SXX), lowering the denominator. The rationale for this choice of assessment metric is to relate the fidelity of the emulation to an estimate of true uncertainty, which we take as the multi-model spread.

We also provide a second, more stringent test of emulator performance: namely a four-fold cross validation (also termed out-of-sample validation). In this test the training data is split and the model is trained on 75% of the data and tested on the held out 25%. The mean absolute error is the calculated between the emulated (predicted) and simulated ("ground truth") values across all cases in the held out 25% of the simulations. The process is then repeated four times to cover all data in the training set. Finally, we normalize the mean absolute error in each grid cell by dividing by the simulated yield in that grid cell in the baseline case (T+0, W+0, C=360, N=200) and report a single aggregated mean absolute error values that are weighted by area cultivation area in the real world. Geographic locations with very low baseline yields are especially problematic in this metric because minor differences result in high percent errors. For this reason we mask all areas with less that 0.1 ton ha-1 in the baseline simulation.

Errors are generally low as a percentage of yield, even for this relatively strict protocol, below 5% of baseline yield for most crop model combinations (Table 3). The notable exceptions are the JULES model for soy and spring wheat and the PROMET model for soy, rice, and the wheats as was seen before in the previous performance metric. Note that under the conditions of this test, the training set often does not include "edge" simulations (i.e. those at the highest or lowest value in that dimension). Predictions in the test phase are therefore extrapolating out to these edge values (e.g. predicted a T+6 case that was not included in the training set). Such an extrapolation during cross validation is not representative based on the intended use of the emulator, which should only be used within the sample space of the overall training set. Maps showing the spatial patterns of errors can seen in supplemental Figures SXX-SXX.

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 9) 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 soy. On the other side, muted responses include soy, an efficient atmospheric nitrogen-fixer, is relatively insensitive to nitrogen, rice is not generally grown in 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 9 only as examples and do not represent true global projections, because they are developed from simulation data with a uniform temperature shift while increases in global mean temperature should manifest non-uniformly in space and distributions (Sippel et al., 2015, e.g). 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.

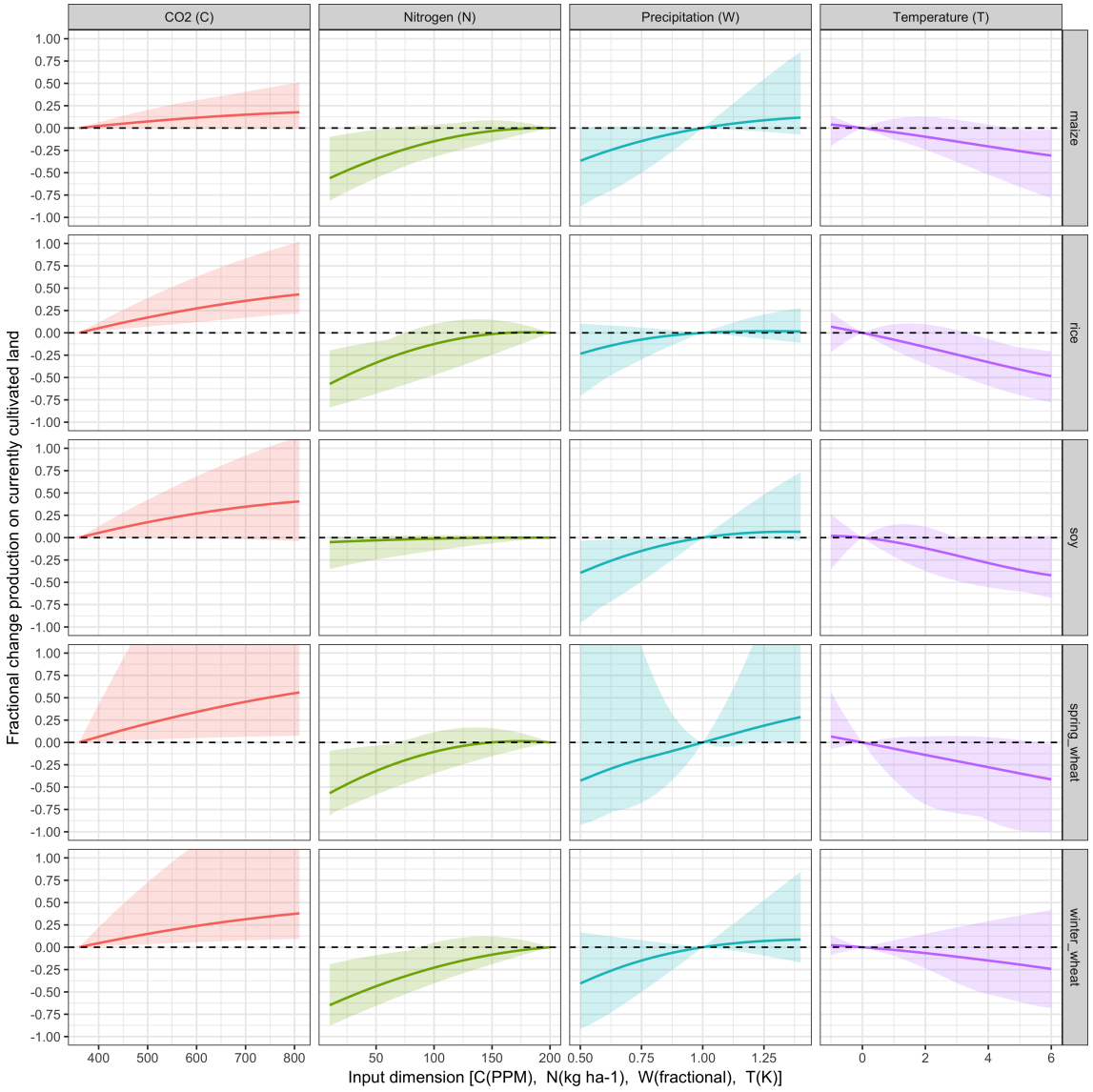

Figure 9. Emulated global damage functions for the five crops included in GGCMI Phase II, from the multi-model mean, for the four dimensions varied: CO_2, temperature, water, and nitrogen (collectively "CTWN"). Solid line shows the multi-model ensemble median and shaded area shows the high and low model projection. All other covariates held constant at baseline values (T+0K, W+0%, C = 360ppm, and N = 200kg ha-1). Damages are reported as fractional change in production relative to the baseline case over currently cultivated land (Portmann et al., 2010).

The emulator can also be utilized for investigating the contributions of the different major climate drivers to production outcomes. The emulated crop model yield responses to business-as-usual climate change (Representative Concentration Pathway (RCP) 8.5) are shown in Figure 10 for 30 climate models from the CMIP-5 archive (Taylor et al., 2012) at the decadal scale.

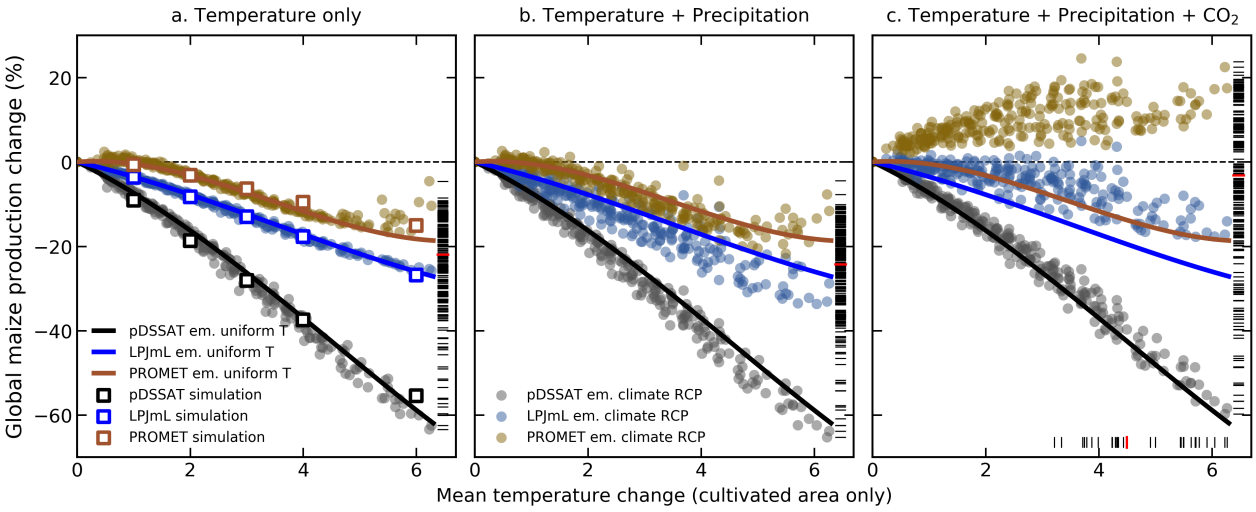

Figure 10. Global emulated damages for maize on currently cultivated lands for three example GGCMI Phase II models. Scatter points show the emulated production change from the 1980-2010 mean value for each crop model using inputs from 30 climate models from the CMIP-5 archive (Taylor et al., 2012) at the decadal timescale for RCP 8.5. The x-axis represents the mean temperature shift over all grid cells where crops are grown (no weighted by cultivation area). Open square markers show the Phase II simulated values at each temperature level with other inputs held at baseline values (W+0%, C = 360ppm, and N = 200kg ha-1). Bold lines show emulated values with uniform global temperature shifts applied and precipitation and CO_2 held constant. Panels a-c show CMIP-5 temperature only, temperature and precipitation, and all three effects respectively. Bold lines are repeated in panels b and c for reference. Rug plots show end of century (2090-2100 mean) values for temperature change for each CMIP-5 model (panel c. onlys) and the resultant maize production change for each crop and climate model combination (including the additional six crop model emulators not shown in full). Red rug shows the ensemble median value. In all cases nitrogen is fixed at 200 (kg $ha^{-}1$) and global production values are aggregated up from the grid cell level based on the growing areas for rainfed and irrigated maize (Portmann et al., 2010). Models not shown: CARAIB, EPIC-TAMU, JULES, GEPIC, LPJ-GUESS, and PEPIC.

The spread across climate model response increases considerably with precipitation changes and CO_2 effects included, with precipitation changes reducing production and CO_2 increasing it. CO2 response is more heterogeneous across crop models that precipitation changes. Note that no direct comparison to the simulations is possible in cases b or c.

The differences between the emulation and the simulations are similar are much lower than the differences across climate models. PROMET, the quantitatively most difficult model to emulate for maize is shown to illustrate that emulation error at the global scale is still small compared to the spread across models or the spread across climate projections when all factors are included. The spatial pattern of temperature change under a realistic climate scenario is relatively insignificant. The pattern of temperature change has the biggest impact for the PROMET model at the highest global temperature changes.

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. Additionally, because the simulation protocol was designed to focus on change in yield under climate perturbations and not on replicating real-world yields, the models are not formally calibrated so cannot be used for impacts projections 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. 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/10.5281/zenodo.2605374.

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

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.

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

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

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, 2019.

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^{o}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.

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.

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

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

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

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, 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.

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

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

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

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

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

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

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

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

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, Accepted for publication in Geoscientific Model Development, pp. 1–42, https://doi.org/10.5194/gmd-2018-195, in open review, 2018.

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

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

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

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

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

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.

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.