GGCMI global gridded crop model emulators of yield and irrigation water demand

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, Nesw 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 uniform 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. XXXXXXXX

1 Introduction

Improving crop yield projections under future climate change is critical for global food security in the twenty-first century. Process-based crop simulation models (numerical models that simulate the process of photosynthesis and the biology and phenology of individual crops) provide many advantages for projecting the impacts of climate change on agriculture but are difficult to directly integrate into larger economic assessments due to computational constraints. Statistical crop models (models developed from historical weather and yield data) lack some of the important features of process-based models including future atmospheric CO_2 fertilization and potential changes in agricultural management, but typically take analytical forms that are easily implemented by downstream impact modelers. Both types of crop models models continue to be used, and comparative studies have concluded that when done carefully, both approaches can provide similar yield estimates (e.g. Lobell and Burke, 2010; Moore et al., 2017; Roberts et al., 2017; Zhao et al., 2017).

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

Interest is rising in applying statistical emulation to crop models as 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 (for multiple crops, using multiple scenarios as a training set). More recently, emulators have begun to be used in the context of multi-model intercomparisons, with Blanc and Sultan (2015); Blanc (2017); Ostberg et al. (2018) and Mistry et al. (2017) using them to analyze the five crop models of the Inter-Sectoral Impacts Model Intercomparison Project (ISIMIP) (Warszawski et al., 2014), which simulated yields for maize, soy, wheat, and rice. Choices differ: Blanc and Sultan (2015) and Blanc (2017) base their emulation on historical simulations and three climate scenarios for one Representative Concentration Pathway (RPC8.5), which represents a high level of global warming; and use local weather variables and yields in their regression across soil

regions; Ostberg et al. (2018) use global mean temperature change (and CO_2) as regressors then pattern-scale to emulate local yields; while Mistry et al. (2017) compare emulated and observed historical yields, using local weather data and a historical crop simulation. These efforts do share important common features: all emulate annual crop yields across the entire scenario or scenarios, and when future scenarios are considered, they are non-stationary, i.e. their input climate parameters evolve over the course of the simulations.

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

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

In this paper we describe a set of crop model emulators developed from a new simulation output dataset as of the Global Gridded Crop Model Intercomparison (GGCMI) effort under the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013, 2014). The emulators are developed from a training set consisting of the largest-yet parameter sweep run by globally gridded crop models for maize, rice, soy and wheat. Globally-gridded crop model emulators allow representing process-based crop model responses in economic assessments in a computationally-cheap fashion and assist in crop model intercomparison and improvement efforts. In the following sections we describe the training dataset, the model statistical specification, and some evidence of model reliability.

2 Methods

2.1 Training dataset

The simulation output dataset that serves as the training set for the crop model emulators presented here are part of the GGCMI phase II project. The GGCMI Phase II simulations are described in detail in Franke et al (this issue), but we summarize briefly

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

Input variable	**Tested range**	**Unit**
[CO_2] (C)	**360**, 510, 660, 810	ppm
Temperature (T)	-1, **0**, 1, 2, 3, 4, 6	°C
Precipitation (W)	-50, -30, -20, -10, **0**, 10, 20, 30, (and W_{inf})	%
Applied nitrogen (N)	10, 60, **200**	kg ha^{-1}
Adaptation (A)	**A0: none**, A1: new cultivar to maintain original growing season length	-

here. The training dataset consists of the outputs from nine process-based crop models for five different crops (maize, rice, soybean, spring wheat, and winter wheat), each simulated globally at the 0.5 degree latitude and longitude resolution. Each simulation scenario is run for 30 years over the historic weather variability for the period of 1981-2010. Simulations are run for both rainfed and irrigated crops across a variety of climate and management input values (Table 1). In each case, the temperature and/or precipitation is offset from the historical values by either an additive mean shift or a fractional multiplier. Atmospheric CO_2 and applied nitrogen fertilizer varied as well and take uniform values globally for each 30-year simulation case.

We construct our emulator of 30-year climatological mean yields because most economic models project impacts at some aggregate temporal scale (decades or larger) or utilize some temporally aggregated climate projection as their input, and because the GGCMI Phase II training dataset is structured in a way that allows temporally aggregation. The year-over-year responses 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. In the example Figure 1, responses to year-over-year temperature variations are 100% larger than those to long-term perturbations in the baseline case, and larger still under warmer conditions, rising to nearly 200% more in the T+6 case. The stronger year-over-year response under warmer conditions also manifests as a wider distribution of yields (Figure 2). The same general principal applies to the precipitation dimension however in this case it is more manifestly an artifact of sampling range in the historical period (Figure 3). 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). Note that the GGCMI Phase II datasets do not capture one climatological factor, potential future distributional shifts, because all simulations are run with fixed offsets from the historical

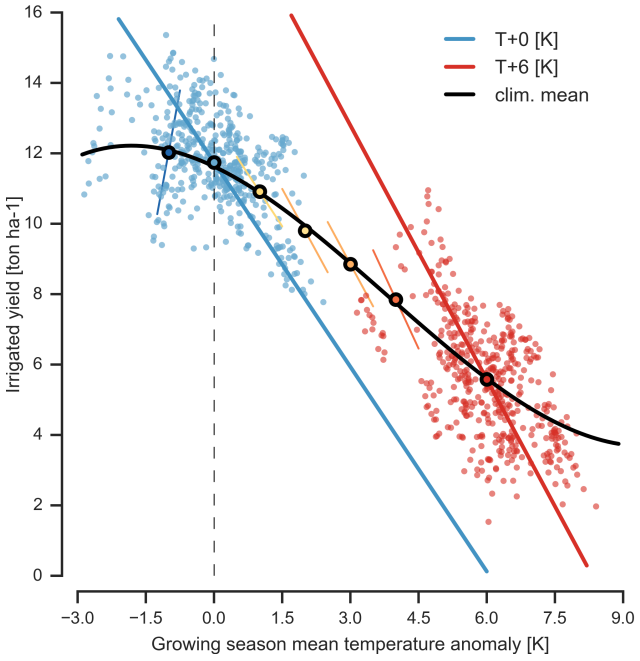

Figure 1. Example showing distinction between crop yield responses to year-to-year and climatological mean temperature shifts. Figure shows irrigated maize for a representative high-yield region (nine adjacent grid cells in northern Iowa) from the pDSSAT model, for the baseline 1981-2010 historical climate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

climatology. Prior work has suggested that mean changes are the dominant drivers of climatological crop yield shifts in non-arid regions (e.g. Glotter et al., 2014).

We develop crop model emulators separately for yield and irrigation water demand with no growing season adaptation, and under a growing season adaptation methodology where crop phenology is allowed to vary to maintain relatively constant growing season length under the warming scenarios (so-called A0 and A1 scenarios, see Franke et al. for more details). Emulators for irrigated simulations are separate from rainfed simulations and therefore irrigation water demand is only provided for irrigated simulations. Therefore, each crop model included consists of six emulators.

2.2 Emulator model selection

Emulation involves fitting individual regression models for each crop, simulation model, and 0.5 degree geographic pixel from the GGCMI Phase II dataset; the regressors are the applied constant perturbations in CO_2, temperature, water, and nitrogen (C,T, W, N). We regress 30-year climatological mean yields against a third-order polynomial in C, T, W, and N with interaction

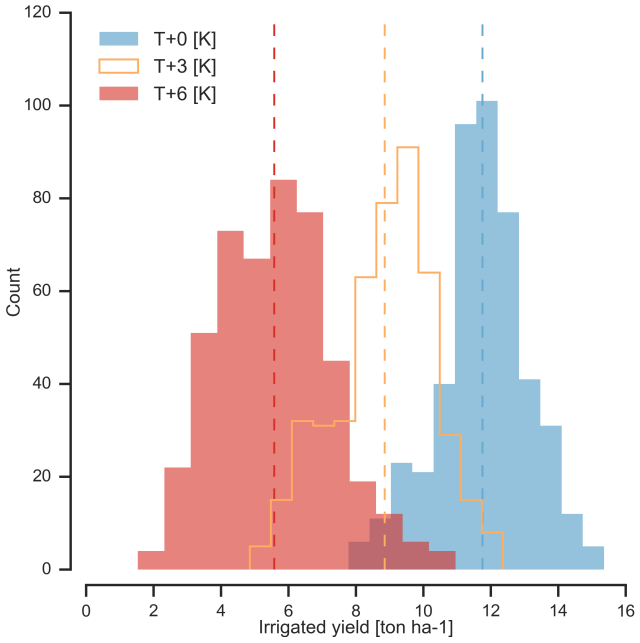

Figure 2. Example showing climatological mean yields and distribution of yearly yields for three 30-year scenarios. Figure shows irrigated maize for nine adjacent high-yield grid cells of Figure 1 from the pDSSAT model, for the baseline 1981-2010 historical climate (blue) and for scenarios with temperature shifted by T+3 (orange) and T+6 K (red), with other variables held at baseline values. The stronger year-over-year temperature response with higher temperatures seen in Figure 1 is manifested here as larger variance in annual yields even though the variance in climate drivers is identical. In this work we emulate not the year-over-year distributions but the climatological mean response (dashed vertical lines).

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

We do not focus on comparing different functional forms in this study, and instead choose a relatively simple parametric model with a polynomial basis function that allows for some interpretation of resultant parameter weights. Some prior studies have used other statistical specifications in crop model emulation, e.g. 39 term fractional polynomial in Blanc and Sultan (2015) and Blanc (2017), who borrow information across space by fitting grid points simultaneously across soil region in a

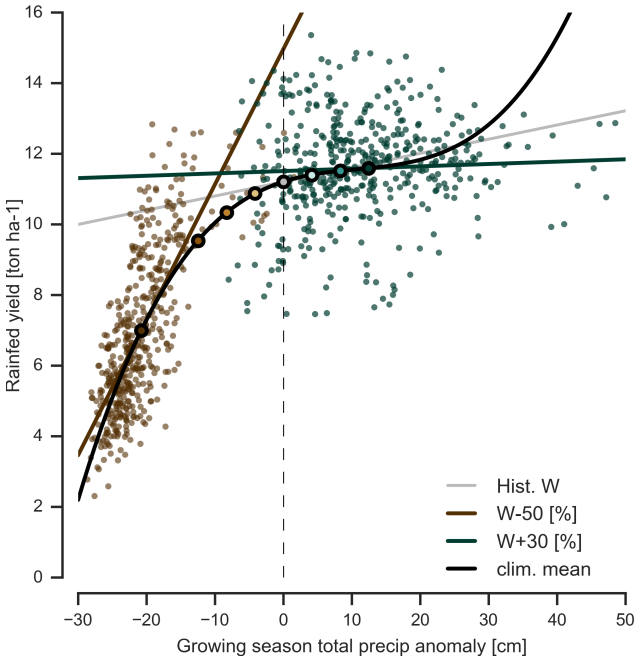

Figure 3. Example showing distinction between crop yield responses to year-to-year and climatological mean precipitation shifts. Figure shows rainfed maize for a representative high-yield region (nine adjacent grid cells in northern Iowa) from the pDSSAT model, for the baseline 1981-2010 historical climate (gray) and for the scenario of maximum and minimum precipitation change (brown and green). Other covariates are held at baseline values. Open black circles mark climatological mean yield values for all seven precipitation scenarios (W -50%, -30%, -20%, -10%, W, +10%, +20%, +30%). Colored lines show total least squares linear regressions of year-over-year variations in each scenario. Bold Black line shows the emulator fit through the climatological mean values.

panel regression. The relativity simple parametric form used here allows fast model emulation at the grid cell level as opposed to the global or large regional level. The emulation therefore indirectly includes any yield response to geographically distributed factors such as soil type, insolation, and the baseline climate, and preserves the spatial resolution of the parent models. We hold the functional form constant in space, across all crops, and models to facilitate potential parameter-by-parameter simulation model comparison.

Although the GGCMI Phase II sampled variable space is relatively large, it is still sufficiently limited that use of the full polynomial basis function described above can be problematic. We therefore reduce the number of terms through a feature selection cross-validation process in which terms in the polynomial are tested for importance. In this procedure higher-order and interaction terms are added successively to the regression model one by one. We then calculate an aggregate mean absolute error with each increasing terms and eliminate those terms that do not contribute significant reductions (top row of Figure 4). 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 (i.e. temperature cubed will not be included without the temperature squared term, and the linear temperature term). We select terms by applying the feature selection process to three example models: two that provided the complete

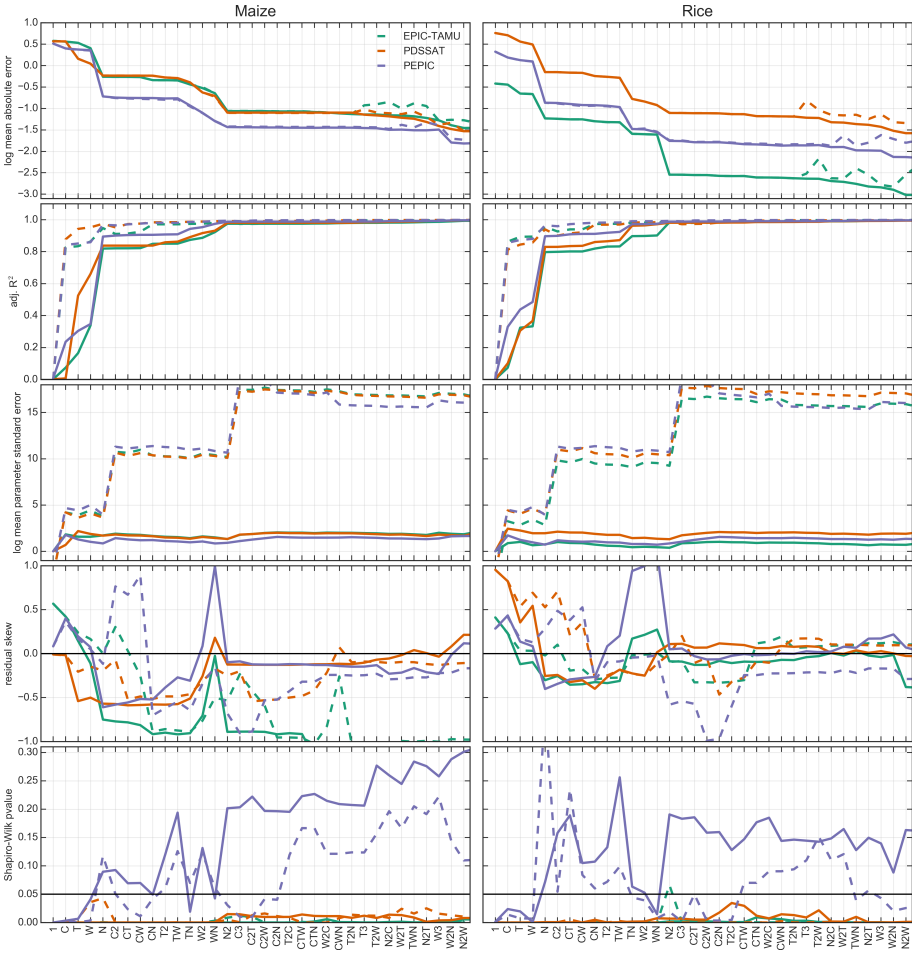

Figure 4. Summary results from polynomial feature selection process. The top row illustrates log mean absolute error between emulated yield and simulated values calculated with a three fold cross validation process, where the emulator is trained on two thirds of the data and predicts the remaining third. 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 and dashed lines indicated standard ordinary least squares. Colors indicated three different crop models.

set of 672 rainfed simulations (pDSSAT, EPIC-TAMU, and one that provided the smallest training set (130, 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 air 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 4). 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 4). The feature selection process results in a final polynomial in 23 terms, with 11 terms eliminated. We omit the N^3 term, which cannot be fitted because we sample only three nitrogen levels. We eliminate many of the C terms: the cubic, the CT, CTN, and CWN interaction terms, and all higher order interaction terms in C. Finally, we eliminate two 2nd-order interaction terms in T and one in W. Implication of this choice include that nitrogen interactions are complex and important, and that water interaction effects are more nonlinear than those in temperature. The resulting statistical model (Equation 1) is used for all grid cells, models, and rainfed crops. (The regressions for irrigated crops do not contain the W terms and the models that do not sample the nitrogen levels omit the N terms).

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

To fit the parameters K, we use a Bayesian Ridge regularization method (MacKay, 1991), which reduces volatility in parameter estimates when the sampling is sparse, by weighting parameter estimates towards zero. This results in a reduction in mean absolute error for some of th high-order interaction terms in the model (top row of Figure 4) and drastically reduces standard parameter error in the model by stabilizing the estimates. The Bayesian Ridge method is necessary over the OLS to maintain a consistent functional form across all models and locations (see Table 2). The Bayesian Ridge 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.

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

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 4 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 that provided less than 50 simulations. The lowest number of simulations emulated across the full parameter space is then 130 (for the PEPIC model). The yield output for a single GGCMI Phase II model that simulates all scenarios and all five crops is ∼12.5 GB; the emulator is ∼20 MB, a reduction of nearly three orders of magnitude.

3 Results

3.1 Model Validation

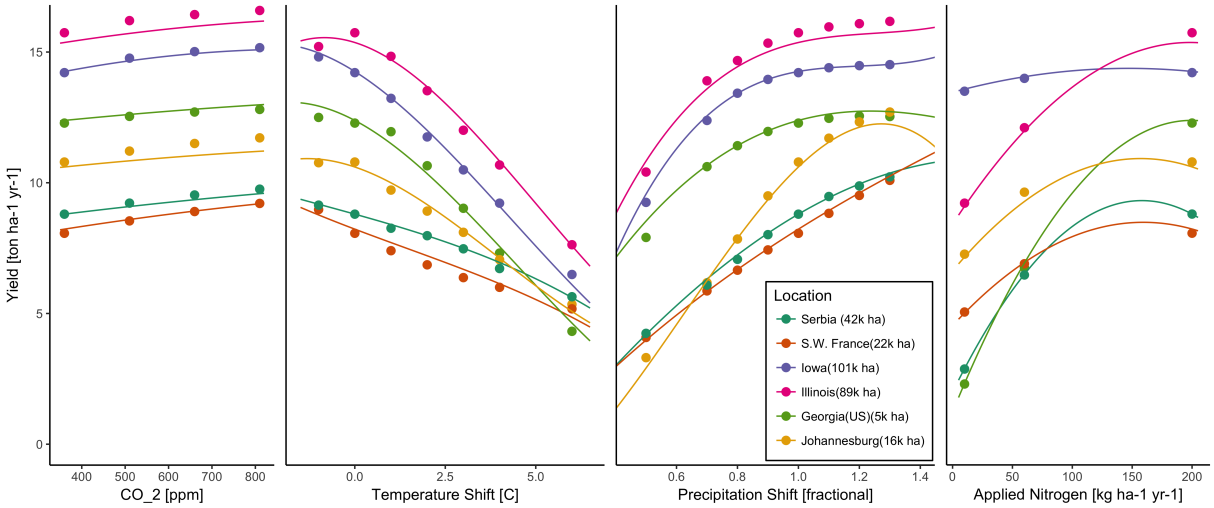

Figure 5. Illustration of spatial variations in yield response successfully captured by the emulator. We show rainfed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares cultivated in each selected grid cell. Each panel shows variation along a single variable, with others held at baseline values. Dots show climatological mean yields and lines the results of the full 4D emulator of Equation 1. In general the climatological response surface is sufficiently smooth that it can be represented within the sampled variable space by the simple polynomial used in this work. Extrapolation can however produce misleading results. Nitrogen fits in some cases may not be realistic at intermediate values given limited sampling. For more detailed emulator assessment, see Appendix B.

Model emulation with the parametric method proposed here is only possible when crop yield responses are sufficiently smooth and continuous to allow fitting with a relatively simple functional form, but this condition largely holds in the GGCMI Phase II simulations. Responses are quite diverse across locations, crops, and models, but in most cases local responses are

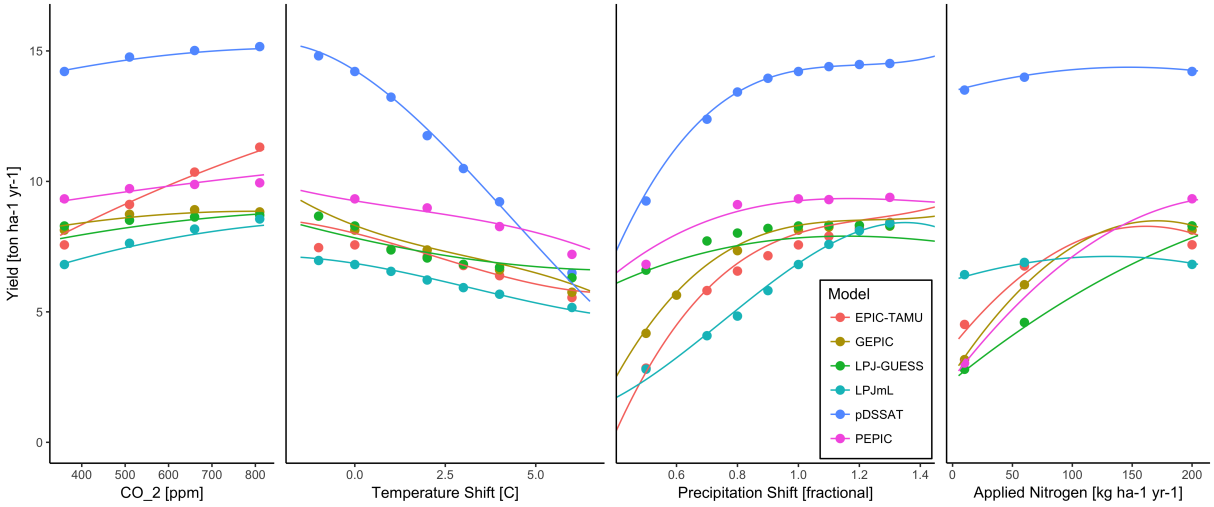

Figure 6. Illustration of across-model variations in yield response successfully captured by the emulator. Figures shows simulations and emulations from six models for rainfed maize in the same Iowa grid cell shown in Figure 5, with the same plot conventions. Models that do not simulate the nitrogen 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. For more detailed emulator assessment, see Appendix B. As in Figure 5, extrapolation out of the sample space is potentially problematic.

regular enough to permit emulation. 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. Each panel in Figure 5 shows simulated yield output from scenarios varying only along a single dimension (CO_2, temperature, precipitation, or nitrogen addition), with other inputs held fixed at baseline levels, compared to the full 4D emulation across the parameter space. Yields evolve smoothly across the space sampled, and the polynomial fit captures the climatological response to perturbations. Crop yield responses generally follow similar functional forms across models, though with a large spread in magnitude partly due to the lack of calibration. 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.

While the nitrogen dimension is important, it is also the most problematic 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 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

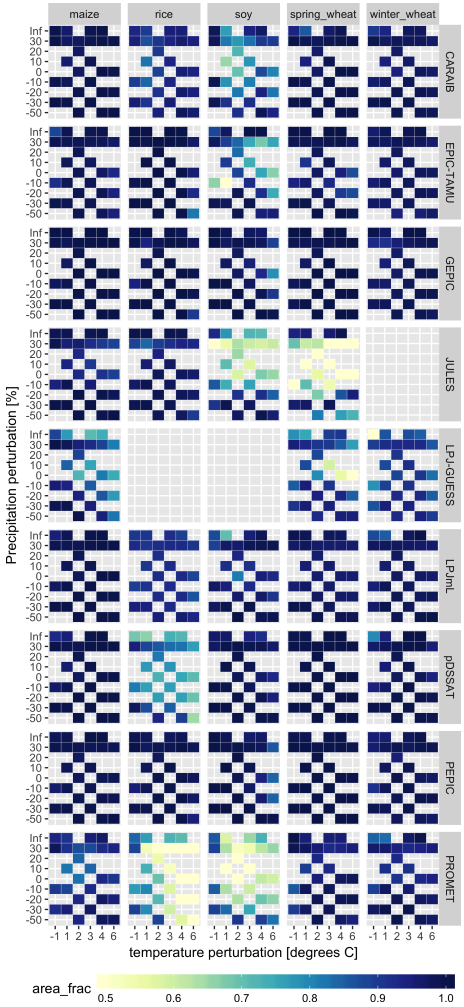

Figure 7. Assessment of emulator performance over currently cultivated areas based on normalized error (Equations 4, 3). We show performance of all 9 models emulated, over all crops and all sampled T and P inputs, but with CO_2 and nitrogen held fixed at baseline values. Large columns are crops and large 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 S3) 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.

Ridge estimator tends to mitigate the 'peak-decline effect' compared to ordinary least squares. In addition, the polynomial fit

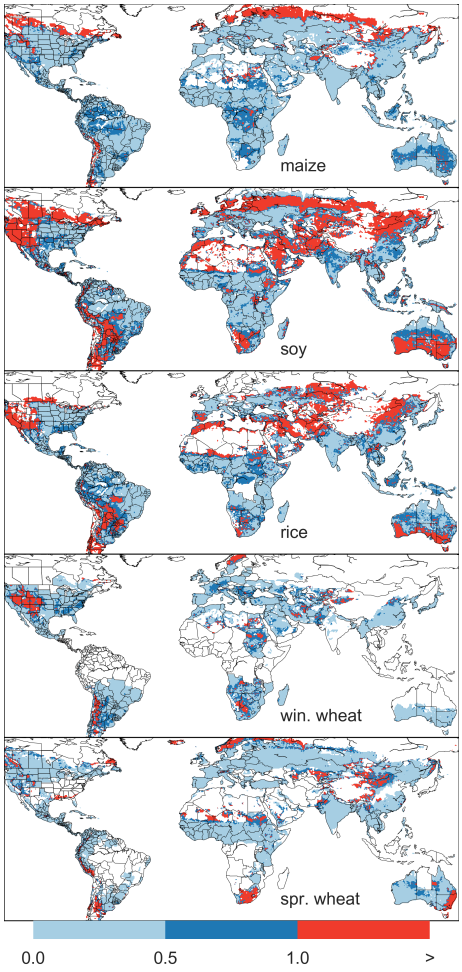

Figure 8. Illustration of our test of emulator performance, applied to the CARAIB model for the T+4 scenario for rainfed crops. Contour colors indicate the normalized emulator error e, where $e > 1$ means that emulator error exceeds the multi-model standard deviation. White areas are those where crops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

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.

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

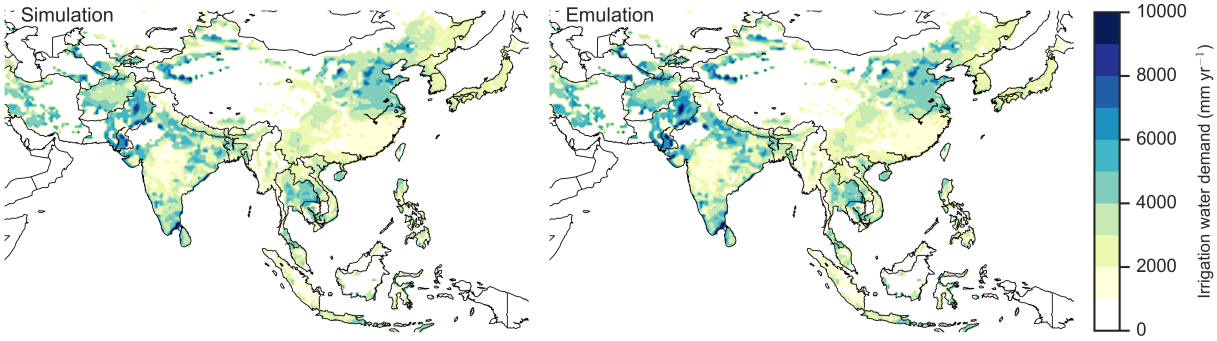

Figure 9. Illustration of the spatial pattern of irrigation water demand.

error e for each scenario as the difference between the fractional yield change from the emulator and that in the original simulation, divided by the standard deviation of the multi-model spread (Equations 3 and 4):

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

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

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 normalized error e is the difference between the emulated fractional change in yield and that actually simulated, normalized by σ_{sim}, the standard deviation in simulated fractional yields change $F_{sim,\,scn.}$ across all models. The emulator is fitted across all available simulation outputs for each grid cell, model, and crop, and then the error is calculated across the each of the simulation scenarios provided by all nine models (Figure S3).

This metric implies that emulation is generally satisfactory, with several distinct exceptions. Almost all model-crop combination emulators have normalized errors less than one over nearly all currently cultivated hectares (Figure 9), but some individual model-crop combinations are difficult to emulate (e.g. PROMET for rice and soy, JULES for soy and spring wheat, Figures S24-S25). Problems with emulating PROMET for rice and soy may have to do with the parametrization of the phenology for those crops which lengthens the growing season in some cases. Normalized errors for soy are somewhat higher across all models not because emulator fidelity is worse but because models agree more closely on yield changes for soy than for other crops (see Figure S18), lowering the denominator. Emulator performance often degrades in geographic locations where crops are not currently cultivated. For example, emulator performance may be satisfactory over cultivated areas for all crops, but uncultivated regions may show some problematic areas (Figure 9 shows a CARAIB model case, see also Figure S26).

This first 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

Table 2. Mean absolute error of emulator representation of a simulation as a percentage of baseline yield for the cross-validation process. A 3-fold stratified k-fold cross validation scheme is utilized where the model is trained on two-thirds of the data and validated on the held-out remaining third (repeated three times). The split does not represent a uniform number of samples in each location or in each model because simulation sampling extent in variable space is heterogeneous. The calculation only includes grid cells with at least 1 % of surface area cultivated with a specific crop (approximately 1000 grid cells in each case). The table displays area weighted mean ('WM') shows the mean error weighted by hectares grown in each grid cell (Portmann et al., 2010) and 'MD' shows the unweighted median across grid cell values. * Indicates cases where the OLS linear model fails.

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

zero. Second, the performance metric scales emulator fidelity not by the magnitude of yield changes but by the inter-model spread in those changes. The normalized error e for a model depends not only on the fidelity of its emulator in reproducing a given simulation but on the particular suite of models considered in the intercomparison exercise. Where models differ more widely, the standard for emulators becomes less stringent. This effect is readily seen when comparing assessments of emulator performance in simulations at baseline CO_2 (Figure 7) with those at higher CO_2 levels (Figure S27) because models disagree on the magnitude of CO_2 fertilization. The rationale for this choice of assessment metric is to relate the fidelity of the emulation to an estimate of true uncertainty, which we take as the multi-model spread. We therefore do not provide a formal parameter uncertainty analysis, but note that the GGCMI Phase II dataset is well-suited to statistical exploration of emulation approaches and quantification of emulator fidelity. More rigorous emulator assessments that could be preformed in future work include: testing other statistical specifications including non-parametric models and calculating standard error on emulator parameters.

We also provide a more stringent test of emulator performance; a three-fold cross validation (out-of-sample validation). Here the training data is split and the model is trained on two thirds of the data and tested on the held out portion (the process is then repeated three times to cover all data in the training set). We normalize the error in each grid cell by dividing by the yield in that grid cell in the baseline (T+0, W+0, C=360, N=200) case and show aggregations by grid cell and weighted by area cultivated per grid cell. Errors are generally low as a percentage of yield –even for this strict protocol– and when weighted by area, essentially zero in most cases (Table 2). Note that the cross validation process often does not include "edge" simulations

in the training set (i.e. those at the highest or lowest value in that dimension) on one or more folds of the training set split. The "edge" cases are then predicted during the prediction phase of cross validation. 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.

3.2 Emulator Products

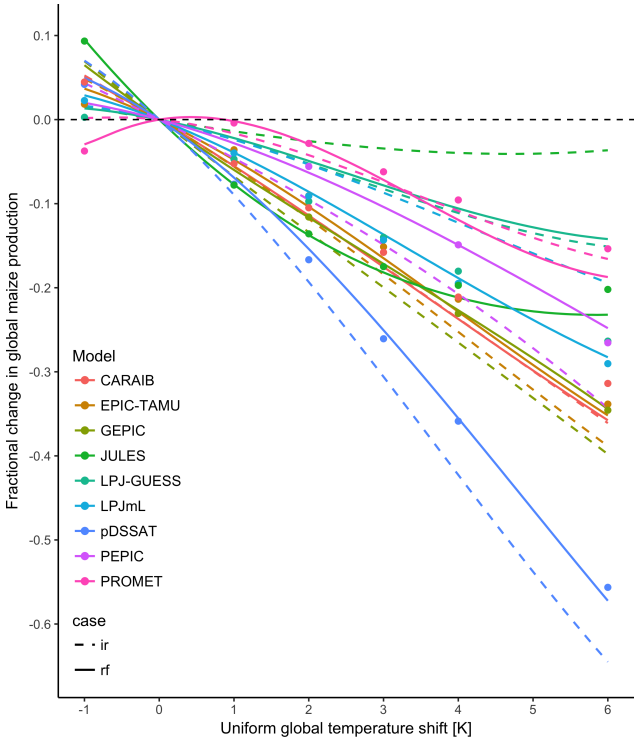

Figure 10. Global emulated damages for maize on currently cultivated lands for the GGCMI Phase II models emulated, for uniform temperature shifts with other inputs held at baseline. (The damage function is created from aggregating up emulated values at the grid cell level, not from a regression of global mean yields.) Lines are emulations for rainfed (solid) and irrigated (dashed) crops; for comparison, dots are the simulated values for the rainfed case. For most models, irrigated crops show a sharper reduction than rainfed because of the locations of cultivated areas: irrigated crops tend to be grown in warmer areas where impacts are more severe for a given temperature shift (the exceptions are PROMET, JULES, and LPJmL, see Franke et al. this issue for more details). For other crops and scenarios see Figures S18-21 in the supplemental material.

Because the emulator or "surrogate model" transforms the discrete simulation sample space into a continuous response surface at any geographic scale, it can be used for a variety of applications, including construction of continuous damage functions. As an example, we show a damage function constructed from the 4D emulation, aggregated to global yield, with

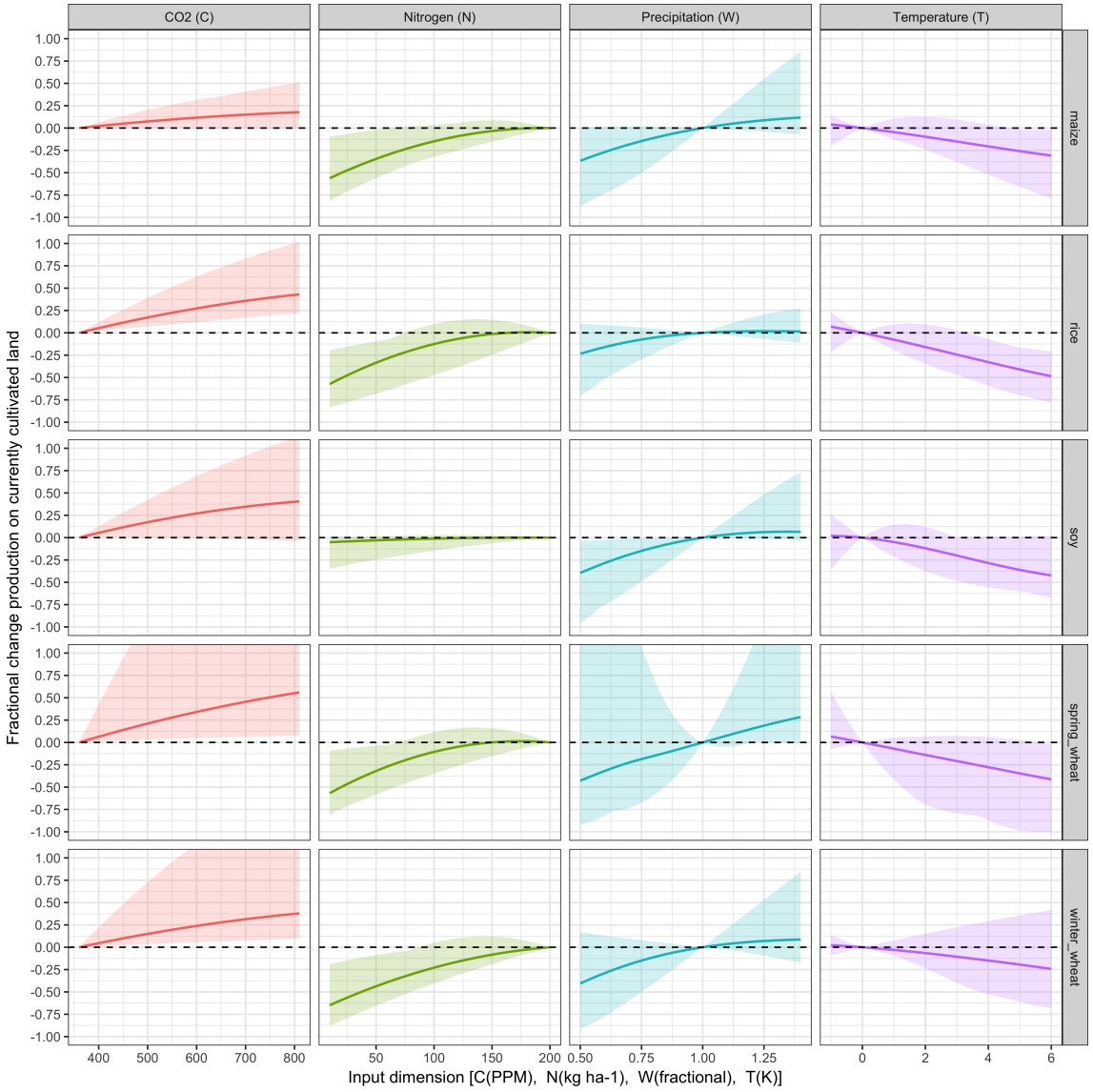

Figure 11. Multi-model ensemble spread in sensitivity to changes in all four dimensions for crops at the global level. Minimum, MME mean and maximum model emulated damage function for currently cultivated areas at the global level. All other covariates held constant in each case and the fractional change in production is relative to the baseline case (T+0K, W+0%, C=360ppm, and N=200kg ha-1).

simulated values shown for comparison (Figure 10, which shows maize on currently cultivated land; see Figures S18-S21 for other crops and dimensions). The emulated values closely match simulations even at this aggregation level.

Finally, we present global damage functions across all four dimensions tested in this study (Figure 11). 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 for 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, and rice is not generally grown in water-limited conditions so shows the lowest response to changes in precipitation.

Note that these functions are presented only as examples and do not represent true global projections, because they are developed from simulation data with a uniform temperature shift while increases in global mean temperature should manifest non-uniformly in space and distributions (Sippel et al., 2015, 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.

4 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.

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.

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.

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.

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.

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.

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

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

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

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

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.

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.

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.

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.

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.

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.

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.