ADVANCING EARTH AND SPACE SCIENCES

Earth's Future

RESEARCH ARTICLE
10.1029/2023EF003773

Substantial Differences in Crop Yield Sensitivities Between Models Call for Functionality-Based Model Evaluation

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

[1]Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany, [2]NASA Goddard Institute for Space Studies, New York, NY, USA, [3]Columbia Climate School, Center for Climate Systems Research, New York, NY, USA, [4]Department of Geophysical Sciences, University of Chicago, Chicago, IL, USA, [5]Biodiversity and Natural Resources Program, International Institute for Applied Systems Analysis, Laxenburg, Austria, [6]Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, Gif-sur-Yvette, France, [7]Unité de Modélisation du Climat et des Cycles Biogéochimiques, UR SPHERES, Institut d'Astrophysique et de Géophysique, University of Liège, Liège, Belgium, [8]Met Office Hadley Centre, Exeter, UK, [9]School of Biology, University of Bristol, Bristol, UK, [10]Department of Geography, Ludwig-Maximilians-Universität, Munich, Germany, [11]Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Georg-August-University Goettingen, Goettingen, Germany, [12]Department of Geographical Sciences, University of Maryland, College Park, MD, USA, [13]Advancing Systems Analysis Program, International Institute for Applied Systems Analysis, Laxenburg, Austria, [14]Center for Agricultural Water Research in China, College of Water Resources and Civil Engineering, China Agricultural University, Beijing, China, [15]State Key Laboratory of Efficient Utilization of Agricultural Water Resources, Beijing, China, [16]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden, [17]School of Geography, Earth & Environmental Science, University of Birmingham, Birmingham, UK, [18]Birmingham Insititute of Forest Research, University of Birmingham, Birmingham, UK, [19]Sino-French Institute for Earth System Science, College of Urban and Environmental Sciences, Peking University, Beijing, China, [20]Global Systems Institute, University of Exeter, Exeter, UK, [21]Department of Environmental Sciences, University of Basel, Basel, Switzerland, [22]Federation of American Scientists, Washington, DC, USA

Key Points:

- Crop models show strong differences in input sensitivities
- Standardized modeling experiments reveal differences in emergent functional relationships
- New standards in model evaluation are needed

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Müller,
christoph.mueller@pik-potsdam.de

Citation:

Müller, C., Jägermeyr, J., Franke, J. A., Ruane, A. C., Balkovic, J., Ciais, P., et al. (2024). Substantial differences in crop yield sensitivities between models call for functionality-based model evaluation. *Earth's Future*, *12*, e2023EF003773. https://doi.org/10.1029/2023EF003773

Received 2 MAY 2023
Accepted 21 FEB 2024

Author Contributions:

Conceptualization: Jonas Jägermeyr, Alex C. Ruane, Joshua W. Elliott
Data curation: Jonas Jägermeyr, Alex C. Ruane, Juraj Balkovic, Philippe Ciais, Marie Dury, Pete Falloon, Christian Folberth, Tobias Hank, Munir Hoffmann, R. Cesar Izaurralde, Ingrid Jacquemin, Nikolay Khabarov, Wenfeng Liu, Stefan Olin, Thomas A. M. Pugh, Xuhui Wang, Karina Williams, Florian Zabel, Joshua W. Elliott
Software: James A. Franke
Writing – review & editing: Jonas Jägermeyr, James A. Franke, Alex C. Ruane, Juraj Balkovic, Philippe Ciais, Marie Dury, Pete Falloon, Christian Folberth, Tobias Hank, Munir Hoffmann, R. Cesar Izaurralde, Ingrid Jacquemin, Nikolay Khabarov,

Abstract Crop models are often used to project future crop yield under climate and global change and typically show a broad range of outcomes. To understand differences in modeled responses, we analyzed modeled crop yield response types using impact response surfaces along four drivers of crop yield: carbon dioxide (C), temperature (T), water (W), and nitrogen (N). Crop yield response types help to understand differences in simulated responses per driver and their combinations rather than aggregated changes in yields as the result of simultaneous changes in various drivers. We find that models' sensitivities to the individual drivers are substantially different and often more different across models than across regions. There is some agreement across models with respect to the spatial patterns of response types but strong differences in the distribution of response types across models and their configurations suggests that models need to undergo further scrutiny. We suggest establishing standards in model evaluation based on emergent functionality not only against historical yield observations but also against dedicated experiments across different drivers to analyze emergent functional patterns of crop models.

Plain Language Summary Crop models are widely used to compute crop yields under future climate change. Yields are determined by many interacting processes. Simulated future crop yields often show a broad uncertainty range. We investigate the sensitivity of nine different crop models to individual model inputs (carbon dioxide, temperature, water, nitrogen) in a very large simulation data set and find that there are substantial differences. We conclude that crop model evaluation needs to include analyses of functional properties to avoid that very diverse model responses to drivers are not tracked if interacting processes cancel out in the historical evaluation period but not in future scenarios, leading to large differences between models.

1. Introduction

Crop models are often employed to project crop yields under changing conditions such as global warming and associated management change for adaptation (Jägermeyr et al., 2021). Multi-model ensembles are promoted to enhance the robustness of projections (Asseng et al., 2015; Martre et al., 2015), but questions remain on what

Wenfeng Liu, Stefan Olin, Thomas A. M. Pugh, Xuhui Wang, Karina Williams, Florian Zabel

causes often large differences between projections of individual models (e.g., Jägermeyr et al., 2021; Müller et al., 2021; E. Wang et al., 2022). Global Gridded Crop Models (GGCMs) are especially exposed to this question when applied for assessing climate change impacts (Jägermeyr et al., 2021; Schleussner et al., 2018), adaptation (Franke et al., 2022; Minoli et al., 2019; Zabel et al., 2021), or environmental impacts of agricultural production (W. Liu et al., 2018), because their results are used in downstream analyses, such as in integrated assessment (Ruane et al., 2017) or economic modeling for projecting future land-use change (Stevanović et al., 2016; Wiebe et al., 2015). Even though global gridded crop models are often based on detailed field-scale models or have implemented similar modeling principles in other ecosystem models (Müller et al., 2019) and show similar performance in evaluation against historical, national yield statistics (Franke et al., 2020a, 2020b; Müller et al., 2017), models are subject to substantial uncertainties from both model structure and parametrization (Folberth et al., 2019) as well as from calibration and input data quality (Ruane et al., 2021). This uncertainty shows most prominently in future projections under high-emission climate change scenarios, where models are exposed to driving data far outside the evaluation domain and results show large inter-model differences (Jägermeyr et al., 2021; Müller et al., 2021; Rosenzweig et al., 2014).

Climatic conditions (D. Liu et al., 2020) and soil properties (Qiao et al., 2022) determine yield potentials (Mauser et al., 2009; van Ittersum et al., 2013) and the suitability of different technologies, such as cultivars (Couëdel et al., 2021). Areas with similar climate and soil conditions show similar yield responses to variations in weather conditions, which can be monitored and reported using representative sites (Gommes et al., 2016). D. Liu et al. (2020) have identified the most limiting climate variable(s) across global crop production areas, finding that temperature has generally a higher impact on crop yields than precipitation for maize, rice, soybean, and wheat. Climate change is projected to alter climate conditions in many agricultural regions substantially (Franke et al., 2022; Jägermeyr et al., 2021; Ruane et al., 2018). Kummu et al. (2021), for example, find that substantial shares of these areas may be driven out of a climatic envelope suitable for agricultural production. Projections of future climate change demonstrate high levels of agreement on global mean temperature trajectories for given forcing scenarios, such as the "Shared Socioeconomic Pathways with Representative Concentration Pathways" (SSP-RCPs) framework (Tebaldi et al., 2021), but are subject to high levels of uncertainties when it comes to spatial and seasonal changes in temperatures and especially precipitation (e.g., Hawkins & Sutton, 2011; Monerie et al., 2020; Wu et al., 2022). Analyzing the sensitivity of cropping systems to changes in individual climate variables can thus help understand their fragility under changing climate.

Process-based crop models are widely accepted tools to project crop yields under changing climatic or management conditions and can help to inform decision making in direct or indirect ways. Crop models are employed at field to global scale and a large variety of crop models exists (e.g., Asseng et al., 2019; Müller et al., 2017). Model intercomparison projects, such as the Agricultural Model Intercomparison and Improvement Project AgMIP (Rosenzweig et al., 2013) have shed light on the inter-model uncertainty (Asseng et al., 2015; Palosuo et al., 2011; Rosenzweig et al., 2014; Ruane et al., 2017), leading to and following up on a call for a general overhaul of crop models (Rötter et al., 2011). Model development efforts since have led to various improvements of crop models (e.g., T. Li et al., 2017; Maiorano et al., 2017; Olin et al., 2015; von Bloh et al., 2018), disagreement between individual crop models remains high (Asseng et al., 2019; Jägermeyr et al., 2021; Kostková et al., 2021; Müller et al., 2021).

Local environmental conditions determine how individual crops are affected by changes in individual drivers. However, owing to the multiple interactions of drivers and processes in yield formation (Schauberger et al., 2016) and the incomplete implementation of processes in crop models (Boote et al., 2013), models can be expected to differ in crop yield projections and sensitivities to individual drivers. Still, regions with severe drought conditions should show substantial sensitivity to changes in water supply and regions with very little nitrogen availability should be sensitive to changes in nitrogen inputs. AgMIP's Global Gridded Crop Model Intercomparison (GGCMI) has set out to intercompare GGCMs in order to evaluate model performance, describe model uncertainties, identify inconsistencies within the ensemble and underlying reasons, and to ultimately improve models and modeling capacities (Elliott et al., 2015). The GGCMI Phase 2 experiment provides simulation data from a large, structured simulation experiment with regular perturbations of four different drivers of yield formation (atmospheric carbon dioxide concentrations (C), temperature (T), water (W), and nitrogen(N)), referred to as CTWN. The CTWN experiment is very well suited to study models' responses to changes in individual or combined driver dimensions. Modeled yield responses to such regular perturbations in drivers can be used to describe crop yield response types, which vary in space (water is a more important driver in arid environments

10.1029/2023EF003773

than in humid ones) and among models. If there were no model uncertainty, crop yield response types would determined by genotype, environment and management (G × E × M) characteristics of each farming system and could be identified with a single crop model. Under given model uncertainty, crop yield response types are, however, a function of the local cropping conditions, but also of model design, functionality, and parameterization (Folberth et al., 2019). Consequently, crop yield response types can describe differences in model behavior and spatial disagreement and can thus help identifying functional differences between models that can guide further model development. Tao et al. (2020) conducted a model intercomparison study with eight barley models for two sites and eight different simulation settings, combining offsets in air temperature, precipitation, irradiation and atmospheric CO_2. They find that the models' disagreement from different sensitivities to changes in temperatures and CO_2 was largest and could identify modeled dynamics of leaf area index as a process that is responsible for model divergence with respect to simulated evapotranspiration, above ground biomass, and yield. In this study, we are conducting a global analysis of GGCMs sensitivities to individual drivers of crop yields, deriving classes of model response types that allow for intercomparing models and regions, aiming to better understand sources of uncertainties in future crop yield projections with crop models.

2. Methods

2.1. The GGCMI Phase 2 Model Output Data Set

The GGCMI Phase 2 experiment is a structured input sensitivity test (Franke et al., 2020a, 2020b) with a modeling protocol that asked for up to 1,404 31-year global simulations at 0.5 arc-degree spatial resolution to assess models' sensitivities to changes in atmospheric carbon dioxide concentrations (C; 4 levels) temperature (T; 7 absolute offset levels, including zero), water supply (W; 9 relative offset levels, including zero), and nitrogen (N; 3 levels), the so-called CTWN experiment (see Table A1) (Franke et al., 2020a, 2020b). A fifth dimension in the CTWN Experiment on Adaptation (A) was not considered here, that is, we only used the simulation sets that assumed no change in cultivars (n = 756). Previous work has used emulators trained on the CTWN experiment (Franke et al., 2020a, 2020b) to explore the contribution from crop models to overall uncertainty in crop yield projections driven by climate change projections (Müller et al., 2021) and to explore the role of adaptation to future agricultural production (Zabel et al., 2021) and the latitudinal shifts in breadbasket regions (Franke et al., 2022). We also focused on rainfed growing conditions only, ignoring the settings with unlimited irrigation (W_{inf}). Of the 12 participating modeling groups, only four supplied all 756 simulation sets (EPIC-TAMU, LPJ-GUESS, LPJmL, and pDSSAT), but five additional modeling groups provided sufficient data to allow for emulation of their yield responses (CARAIB, GEPIC, JULES, PEPIC, and PROMET) and we used the emulators that were build on these simulations (Franke et al., 2020a, 2020b) to gap-fill missing simulation sets that were not provided. The remaining models (APSIM-UGOE, EPIC-IIASA, and ORCHIDEE-crop) are only shown here in the overview figure for completeness, but are not included in the following analyses. The scarcity of simulations provided by these modeling teams (33–44 of 756, see Table 1) does not allow for in-depth analysis and also led to exclusion of these models from the emulator training (Franke et al., 2020a, 2020b). The GGCMs' baseline simulations (C = 360, T = 0, W = 0 for rainfed and W = inf for irrigated, N = 200) have been used to evaluate model performance, even though uniform N application rates and atmospheric CO_2 concentrations of 360 ppm are not realistic representations of the current systems. Despite the uniform management assumption, Franke et al. (2020a, 2020b) show that the GGCMI Phase 2 model ensemble is capable of reproducing some of the observed yield variability, especially in top-producer countries (see reproduced Figure S2 in the Supplementary Information), but also at the global aggregation level (Figures S3–S6 in the Supplementary Information). GGCMs' skills in reproducing observed yield variability at the grid-cell level varies across GGCMs and crops but shows roughly similar skill to the GGCMI Phase 1 model ensemble tested by Müller et al. (2017).

2.2. Data Analysis

The analysis conducted here aims at understanding differences in models' sensitivities of simulated crop yield (y) to the CTWN drivers across crops and regions as well as understanding differences among models. We considered current crop-specific cropland extent, making use of the MIRCA2000 cropland data set Portmann et al. (2010). To avoid distortions of marginal production areas, we only considered grid cells (0.5° by 0.5° longitude/latitude, equivalent to 55 km by 55 km at the equator) with at least 200 ha of crop cultivation (rainfed and irrigated area). Spring and winter wheat are not separated in the MIRCA2000 data so we considered total wheat areas for both. MIRCA2000 data were also used for data aggregation to the global scale, using the provided crop-specific

10.1029/2023EF003773

Table 1

Number of Global Simulation Sets of Crop Yield (y) Included in the Global Gridded Crop Model Intercomparison Phase 2 Archive per Model and Crop for the Simulation Set

Model	Maize	Soybean	Rice	Winter wheat	Spring wheat	Nitrogen
CARAIB[a]	252	252	252	252	252	No
EPIC-TAMU[b]	756	756	756	756	756	Yes
JULES[c]	252	252	252	–	252	No
GEPIC[d]	430	430	430	430	430	Yes
LPJ-GUESS[e]	756	–	–	756	756	Yes
LPJmL[f]	756	756	756	756	756	Yes
pDSSAT[g]	756	756	756	756	756	Yes
PEPIC[h]	149	149	149	149	149	Yes
PROMET[i]	261	261	261	261	261	Yes
Totals	4,368	3,612	3,612	4,116	4,368	7
not included						
APSIM-UGOE[j]	44	44	44	–	44	Yes
EPIC-IIASA[k]	39	39	39	39	39	Yes
ORCHIDEE-crop[l]	33	–	33	33	–	Yes

Note. Some models do not account for nitrogen dynamics, as indicated in column "Nitrogen." Not all models provide data for all crops (indicated by "–" in the respective columns), but always supply the same simulation sets across all crops provided. [a]Dury et al. (2011). [b]Izaurralde et al. (2006). [c]Osborne et al. (2015), K. Williams and Falloon (2015), and K. Williams et al. (2017). [d]J. Liu et al. (2007) and Folberth et al. (2012). [e]Lindeskog et al. (2013) and Olin et al. (2015). [f]von Bloh et al. (2018). [g]Elliott et al. (2014) and Jones et al. (2003). [h]W. Liu, Yang, Folberth, et al. (2016) and W. Liu, Yang, Liu, et al. (2016). [i]Mauser and Bach (2015), Hank et al. (2015), and Mauser et al. (2009). [j]Keating et al. (2003) and Holzworth et al. (2014). [k]Balkovič et al. (2014). [l]Valade et al. (2014).

harvested areas as aggregation weights. Globally, there were 21,262 grid cells included for maize, 9,165 for soybean, 11,452 for rice, 17,032 for spring wheat, and 17,032 for winter wheat. With the sheer amount of data of the GGCMI Phase2 experiment (up to 4,368 global simulations, see Table 1), a visual representation of variations in model response is not helpful. We thus structured the analysis to condense the information in a meaningful way so that different response types can be identified and discussed.

2.2.1. Impact Response Surfaces

Impact Response Surfaces (IRSs) have been used to describe crop model behavior under changes in two driver dimensions (e.g., temperature and precipitation) (e.g., Pirttioja et al., 2015) and Fronzek et al. (2018) have used IRS to identify different model response types. Zabel et al. (2021) used IRSs to describe isolines for comparison of adapted and non-adapted global production systems. Here, we were interested in regional differences and thus constructed IRSs for each grid cell i, GGCM g, crop c, and each paired combination of two drivers $d1$ and $d2$ of the four CTWN dimensions (i.e., T ~ W, T ~ N, C ~ T, W ~ N, C ~ W, C ~ N). IRSs display yield changes ($\Delta y_{i,g,c}$) for any grid cell i or aggregation of grid cells for combination of any two drivers ($d1$ and $d2$) in relation to the average yield across all cases included in the IRS $\left(\bar{y}_{d1^*,d2^*,i,g,c}\right)$ as described by Equation 1, where $d1^*$ and $d2^*$ describe the full set of elements in $d1$ and $d2$ respectively. We used the average yield (of each respective IRS) rather than the yield at default conditions ($y_{i,g,c,C360,T0,W0,N200}$) as the default conditions were not always directly supplied by all models g.

$$\Delta y_{d1,d2,i,g,c} = \frac{y_{d1,d2,i,g,c}}{\bar{y}_{d1^*,d2^*,i,g,c}} * 100\% \tag{1}$$

The other two dimensions, not displayed in the IRS are kept at their default setting (C: 360 ppm, T: 0°C, W: 0%, N: 200 kg ha^{-1}). The atmospheric CO_2 concentration of 360 ppm refers to approximately the value of 1,995, the middle of the simulation period 1980–2010 of the GGCMI Phase2 experiment.

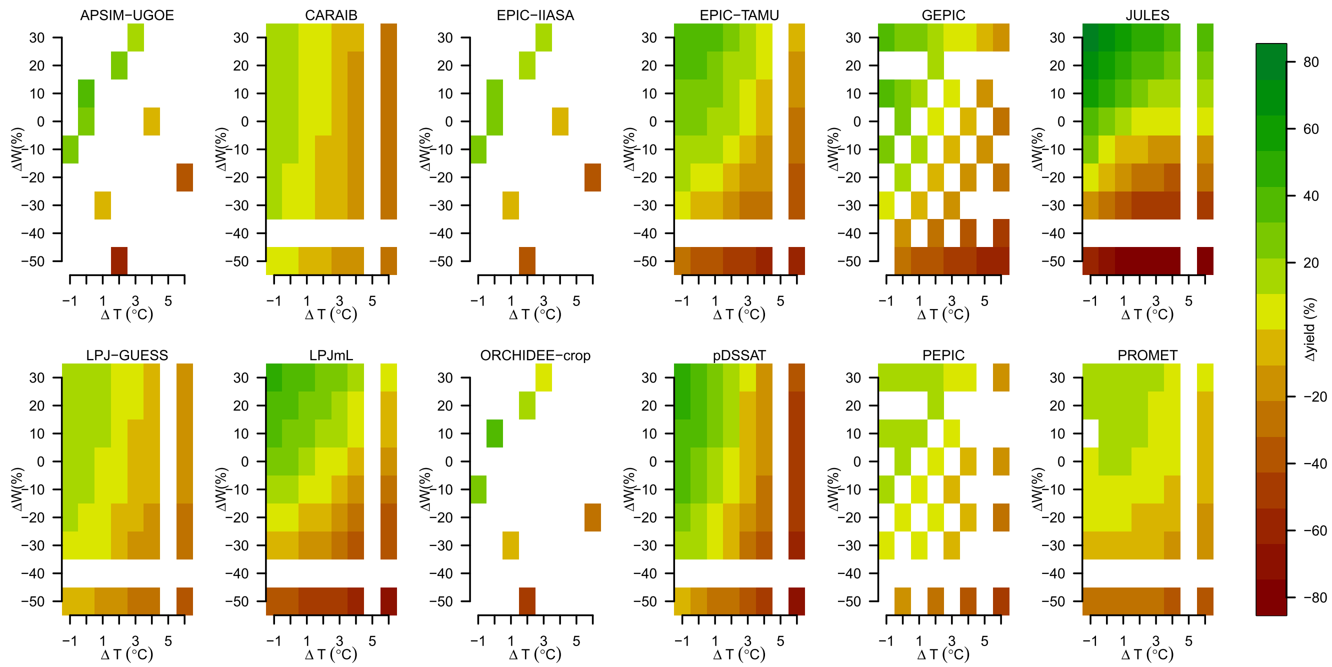

Figure 1. Illustrative example of crop model Impact Response Surfaces (IRS), here for maize and the temperature (T) and water (W) dimensions with atmospheric CO_2 (C) at 360 ppm and fertilizer input (N) at 200 kg ha^{-1}. All crop model simulations provided by the modeling groups are displayed as colored rectangles in the IRS. Colors indicate relative yield changes compared to the mean across all data points of the respective IRS. White spaces indicate missing simulation sets. The simulations for "W − 40%" and "T + 5" were not requested per protocol (Franke et al., 2020a, 2020b).

Depending on the global extent of cropland, 9,165 (soybean) to 21,262 (maize) of such IRS sets were constructed per CTWN dimension and crop, which cannot be displayed or interpreted as visuals. For illustrative purposes, Figure 1 shows IRS for the T ∼ W responses of globally aggregated maize yield.

2.2.2. Dominant Response Dimensions

IRSs show the response of projected yields for any two drivers ($d1$ and $d2$, e.g. T and W). The classification of IRS as proposed by Fronzek et al. (2018), which distinguishes nine cases of maximum yield location per IRS and the strength of the response per dimension, is still too complex for our purposes here, especially if extended from two (TW) to four (CTWN) dimensions of drivers. For a simpler metric to describe the characteristics of IRS, we identified the *dominant response dimension*, using response ratios (*RR*). Response ratios describe the relationship of the gradients along the two dimensions, based on minimum and maximum values, that is, ignoring the shape of these gradients (i.e., it does not matter if the minimum (or maximum) is at either end of the row or column). In contrast to the illustrative IRSs, the reference yield $\bar{y}$ cancels out in the computation of *RRs*, so we computed *RRs* based on actual yields (y) rather than yield changes (Δy). Any distortion that may be introduced by using the IRS' mean value rather than a standard simulation set thus does not affect any quantitative analysis here. In order to determine which of the two drivers dominates over the other, we selected the data slice from the CTWN cube that spans the full range of the drivers of interest ($d1$ and $d2$) at the default conditions of the other two drivers (e.g., $[T_{-1}.. T_{+6}]$ vs. $[P_{-50}.. P_{+30}]$ at C_{360} and N_{200}). Across that selected surface, we computed the range of simulated yields (y) for each grid cell i, model g, crop c for each element $j1$ of $d1$ across all elements $j2$ of $d2$, computing the average response to those drivers (e.g., R_T and R_W in Figure S1) by dividing by the number of elements n_{d1} and n_{d2}. The average response of the two drivers $d1$ and $d2$ are computed as described in Equations 2 and 3 and their combination to compute the response ratio $RR_{d1,d2,i,g,c}$ is described in Equation 4. *RR* ranges between 0 and 1 and describes the contribution of the first driver to the yield variation across both drivers. If these are perfectly balanced, *RR* is 0.5, if the first driver is the only driver of yield change, *RR* is 1, if it has no effect, *RR* is zero. All data processing and plotting was done in R, version 4.1.2 (R Core Team, 2021).

$$R_{d1,i,g,c} = \frac{\Sigma_{j2=1}^{n_{d2}} \max\left(y_{i,j2,g,c}\right) - \min\left(y_{i,j2,g,c}\right)}{n_{d2}} \tag{2}$$

$$R_{d2,i,g,c} = \frac{\Sigma_{j1=1}^{n_{d1}} \max\left(y_{i,j1,g,c}\right) - \min\left(y_{i,j1,g,c}\right)}{n_{d1}} \tag{3}$$

$$RR_{d1,d2,i,g,c} = \frac{R_{d1,i,g,c}}{R_{d1,i,g,c} + R_{d2,i,g,c}} \tag{4}$$

We describe the different *RR* values with the median value and the skewness of their distribution. Skewness was computed with R version 4.1.2 (R Core Team, 2021) with the *skewness* function of the *moments* R package, version 0.14.1, using Equation 5, with x for the data and n for the number of data points i in x and $\bar{x}$ for the mean of x.

$$skewness = \frac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2\right]^{(3/2)}} \tag{5}$$

Skewness values range between positive and negative infinity and values outside the [−0.4, 0.4] interval can be considered skewed, that is, data are distributed asymmetrically (Doane & Seward, 2011).

2.2.3. Cluster Analysis

RRs take continuous values in the interval [0, 1] and were computed for all six combinations of any two drivers of the CTWN data cube (T ∼ W, T ∼ N, C ∼ T, W ∼ N, C ∼ W, C ∼ N). In order to structure *RRs* into Crop Yield Response Types (*YRTs*), we use hierarchical clustering, separating *RR* combinations into clusters so that at least 90% of the overall variance in the total sample is explained by the separation into clusters. The resulting *YRT* describe differences across models and environments simultaneously. This allows for comparing regions and GGCMs with respect to their sensitivities to changes in the CTWN drivers under the full range of global crop growing conditions. In order to include all GGCMs with sufficient data provision, independent of their ability to provide data on responses to variation in N input (see Table 1), we also conducted the same analysis for the CTW data cube with 3 different combinations of any two drivers (T ∼ W, C ∼ T, C ∼ W), which we refer to as CTW-*YRT*. We used R version 4.1.2 (R Core Team, 2021) with R-package *Rclustercpp.hclust* (version 0.2.6) for large data sets with standard settings, that is, using *euclidean distances* and the *ward* method. For describing the characteristics of the individual clusters, we make use of the median and interquartile range of each *RRs* distribution within each cluster.

3. Results

3.1. Distribution of *RR*

The GGCMs show different distributions of *RR* across all crop-specific cropland. There are differences in the median values, but also in the shape — and skewness — of the distributions. Most *RR* values per GGCM are not normally distributed but highly skewed or bi-modal (see illustrative Figure 2). The differences in median values illustrate differences between models, as the distributions always refer to the same spatial sample (all grid cells with at least 200 ha crop-specific harvested area, according to Portmann et al. (2010)). Median values range substantially across GGCMs, but also across crops.

Maize yield simulations of CARAIB show very little response to changes in water supply in comparison to changes in temperature with a median $RR_{T,W}$ value of 0.84, which is in line with the vertical stripe pattern seen in the IRS for CARAIB in Figure 1. JULES maize yield simulations, on the other hand, show the opposite behavior with a median *RR* value of 0.26. Specific regional characteristics can also already be detected here. EPIC-TAMU, pDSSAT, and PROMET show a spike in the highest *RR* bin, indicating that there is a substantial number of grid cells (about 1,000 for EPIC-TAMU, >2,500 for pDSSAT, >2,100 for PROMET), in which changes in water supply have basically no effect in comparison to changes in temperature on simulated yields. JULES and LPJmL hardly have such maize-growing grid cells where water supply matters little in comparison to changes in

10.1029/2023EF003773

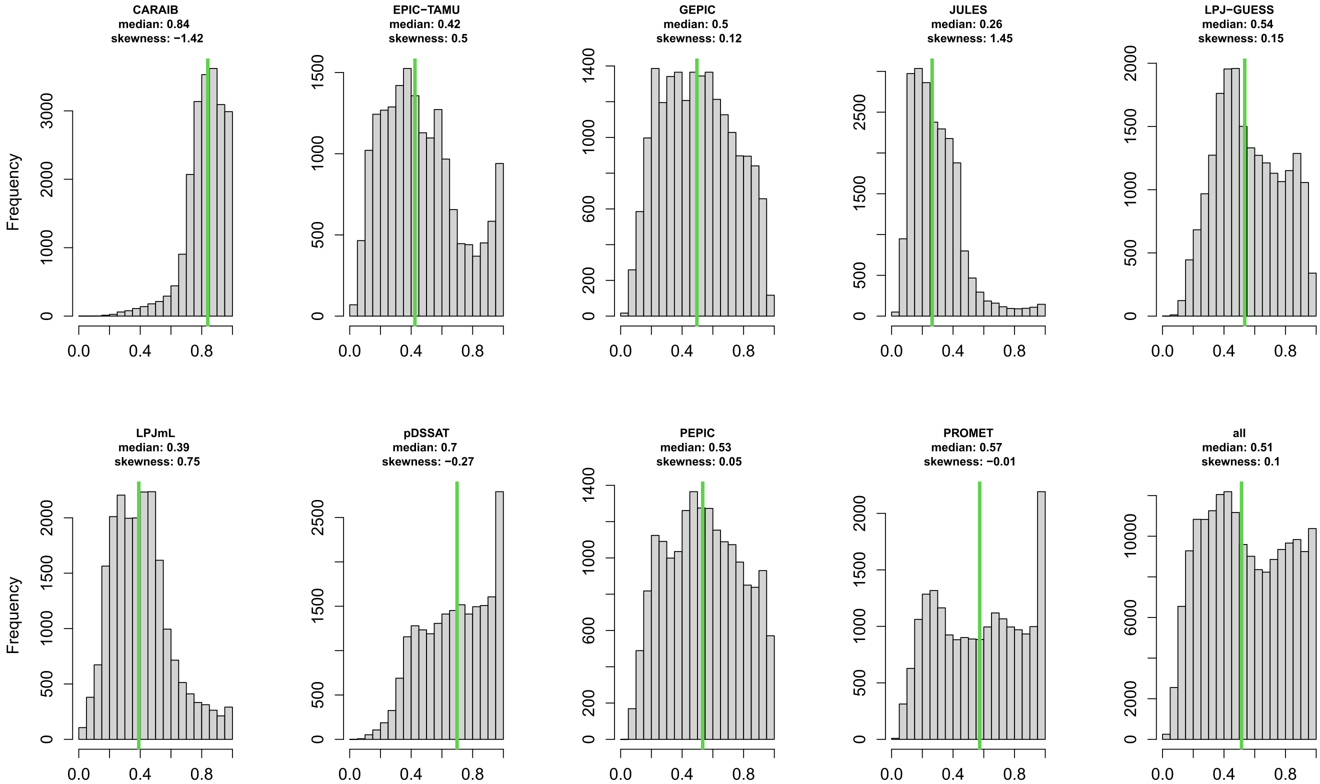

Figure 2. Distributions of maize Response Ratios for the temperature (T) versus water (W) domains ($RR_{T,W}$) of the nine different Global Gridded Crop Models (GGCMs), showing the importance of T in comparison to W at default C (360 ppm) and N (200 kg ha^{-1}). Values approaching unity indicate higher sensitivity to T than to W, while values approaching zero indicate higher sensitivity to W than to T (see Equation 4). Green vertical lines show the median value, which is also given in the title of each panel. Bottom right-hand panel shows the distribution across all GGCMs. Results are shown for currently cultivated maize cropland.

temperature. Some distributions are highly skewed or show bi-modal patterns, which is most prominent in the combined distribution across all nine GGCMs. There are too many *RR* and crop combinations to show all distributions as histograms in figures and we thus present results of variation in median and skewness values in Tables 2 and 3. Table 2 shows the range of median *RR* values across GGCMs, crops and driver combinations. Median values range between 0.06 (importance of C in comparison to N for maize in GEPIC data) and 1.0 (importance of C in comparison to N for soybean in GEPIC data). For all crops analyzed here, many *RRs* show a very broad range across GGCMs with differences between min and max values often well above 0.5 (Table 2). One exception is $RR_{C,T}$ of the C3 crops (other than spring wheat), where the range is only 0.3 or lower. Large differences between GGCM's *RRs* are particularly pronounced for $RR_{W,N}$ and $RR_{T,N}$ for all crops other than the N-fixing leguminous crop soybean.

3.2. Crop Yield Response Types

Identifying crop yield response types (*YRT*) can help to illustrate the similarities and differences between *RR* combinations across GGCMs and regions. The hierarchical clustering combines elements (individual data points (leaves) or clusters) by similarity and dendrograms illustrate the similarity of these elements (Figures S11–S20). Three (e.g., soybean, Figure S23) to six (e.g., winter wheat, Figure S27) clusters were needed to explain at least 90% of the overall variance in the global crop-specific simulation sets.

As already suggested by the GGCM-specific distributions of *RRs* (Figure 2, Tables 2 and 3), some GGCMs show substantially different *YRTs* than others, however, also the regional distribution of *YRTs* differs between individual GGCMs (see Figures 3 and 4 for maize and Figures S21–S28 for the other crops). Since the clusters are defined by similarity of *RR* combinations, the interpretation depends on the *RR* distributions within clusters, as displayed in Figure 3 for maize CTW responses (corresponding to Figure 4). As the C4 crop, maize sees no direct stimulation of photosynthesis through elevated atmospheric CO_2 concentrations, but only improvements in water-

Table 2
Median Values of RR Across Crops and Driver Combinations

Crop	Drivers	CARAIB	EPIC-TAMU	GEPIC	JULES	LPJ-GUESS	LPJmL	pDSSAT	PEPIC	PROMET	All	Min	Max
Maize	TW	0.84	0.42	0.5	0.26	0.54	0.39	0.7	0.53	0.57	0.51	0.26	0.84
	WN	NA	0.39	0.37	NA	0.18	0.75	0.56	0.22	0.76	0.47	0.18	0.76
	CW	0.81	0.38	0.14	0.19	0.43	0.25	0.19	0.32	0.57	0.32	0.14	0.81
	TN	NA	0.33	0.32	NA	0.21	0.64	0.76	0.23	0.82	0.47	0.21	0.82
	CT	0.48	0.42	0.14	0.35	0.5	0.33	0.08	0.26	0.37	0.33	0.08	0.5
	CN	NA	0.18	0.06	NA	0.09	0.3	0.16	0.11	0.92	0.16	0.06	0.92
Rice	TW	0.66	0.6	0.68	0.44	NA	0.69	0.75	0.72	0.56	0.64	0.44	0.75
	WN	NA	0.35	0.29	NA	NA	0.58	0.31	0.19	0.81	0.41	0.19	0.81
	CW	0.54	0.48	0.49	0.44	NA	0.66	0.49	0.5	0.5	0.51	0.44	0.66
	TN	NA	0.52	0.36	NA	NA	0.79	0.45	0.27	0.87	0.6	0.27	0.87
	CT	0.42	0.36	0.33	0.55	NA	0.46	0.37	0.37	0.44	0.41	0.33	0.55
	CN	NA	0.23	0.16	NA	NA	0.52	0.21	0.15	0.91	0.33	0.15	0.91
Soy	TW	0.84	0.49	0.62	0.37	NA	0.47	0.49	0.72	0.9	0.57	0.37	0.9
	WN	NA	0.9	0.98	NA	NA	0.99	0.9	0.89	0.56	0.93	0.56	0.99
	CW	0.75	0.4	0.42	0.37	NA	0.45	0.4	0.65	0.89	0.48	0.37	0.89
	TN	NA	0.91	0.99	NA	NA	0.99	0.9	0.96	0.9	0.95	0.9	0.99
	CT	0.44	0.41	0.33	0.53	NA	0.53	0.41	0.4	0.41	0.45	0.33	0.53
	CN	NA	0.81	1	NA	NA	0.99	0.9	0.9	0.91	0.93	0.81	1
Spring wheat	TW	0.68	0.31	0.45	0.35	0.74	0.47	0.42	0.51	0.39	0.47	0.31	0.74
	WN	NA	0.52	0.45	NA	0.22	0.78	0.61	0.31	0.85	0.51	0.22	0.85
	CW	0.72	0.25	0.18	0.35	0.7	0.45	0.32	0.41	0.47	0.43	0.18	0.72
	TN	NA	0.43	0.41	NA	0.46	0.77	0.57	0.3	0.83	0.53	0.3	0.83
	CT	0.58	0.42	0.24	0.51	0.69	0.5	0.39	0.38	0.56	0.49	0.24	0.69
	CN	NA	0.26	0.12	NA	0.38	0.59	0.38	0.19	0.93	0.37	0.12	0.93
Winter wheat	TW	0.65	0.31	0.45	NA	0.62	0.49	0.42	0.35	0.51	0.48	0.31	0.65
	WN	NA	0.39	0.44	NA	0.19	0.7	0.64	0.35	0.82	0.44	0.19	0.82
	CW	0.68	0.32	0.33	NA	0.68	0.47	0.2	0.35	0.54	0.45	0.2	0.68
	TN	NA	0.21	0.31	NA	0.33	0.68	0.55	0.22	0.84	0.38	0.21	0.84
	CT	0.54	0.51	0.44	NA	0.63	0.49	0.33	0.5	0.45	0.5	0.33	0.63
	CN	NA	0.15	0.17	NA	0.32	0.52	0.24	0.19	0.92	0.29	0.15	0.92

Note. CARAIB and JULES did not supply data for different N levels, LPJ-GUESS did not supply data for rice or soybean, JULES did not supply data for winter wheat. These missing data points are indicated by "NA" (not available).

use efficiency. Still, some GGCMs display substantial shares of maize growing areas where C is more dominant than changes in T and similar to changes in W (cluster#2; Figure 4). Temperature dominance (cluster#4, as well as clusters #1 and #3, in which T is dominant or on par with the other drivers) is particularly important in pDSSAT, GEPIC, PEPIC, and LPJmL, even though patterns differ (Figure 4).

For rice simulations, the distribution of different CTW-*YRTs* is more balanced across GGCMs (Figures S21 and S22), with JULES and PROMET showing little presence of cluster #4 (T dominance and C dominance over W, Figure S21) and LPJmL with little presence of cluster #2 (W dominance and balanced C vs. T response). Spatial patterns show some similarities with respect to cluster #4 (other than in JULES and PROMET) in the tropics and cluster #2 (other than LPJmL) in more arid regions of Asia, Africa, and south America.

Soybean CTW data are only clustered in three different CTW-*YRTs* (Figures S23 and S24), where JULES and to some lesser extent LPJmL are mostly characterized by cluster #2 (W dominates and C vs. T is balanced). CARAIB and PROMET show larger shares of cluster #3 (dominance of T and of C over W). There is larger

Table 3
Skewness of RR Distributions Across Crops and Driver Combinations

Crop	Drivers	CARAIB	EPIC-TAMU	GEPIC	JULES	LPJ-GUESS	LPJmL	pDSSAT	PEPIC	PROMET	All	Min	Max
Maize	TW	−1.42	0.5	0.12	1.45	0.15	0.75	−0.27	0.05	−0.01	0.1	−1.42	1.45
	WN	NA	0.39	0.68	NA	1.44	−0.53	0.01	1.14	−1.09	0.1	−1.09	1.44
	CW	−1.09	1.29	1.71	3.02	0.51	3.76	1.59	0.91	0.29	0.76	−1.09	3.76
	TN	NA	0.9	1.05	NA	2.13	−0.11	−0.45	1.32	−1.31	0.17	−1.31	2.13
	CT	−0.77	0.77	1.8	0.48	−0.39	0.28	2.11	0.94	0.27	0.33	−0.77	2.11
	CN	NA	1.42	2.15	NA	2.46	0.54	1.14	1.67	−2.14	0.88	−2.14	2.46
Rice	TW	−0.26	−0.1	−0.52	0.73	NA	−0.54	−0.66	−0.55	0.16	−0.21	−0.66	0.73
	WN	NA	0.53	1.13	NA	NA	−0.01	0.9	1.4	−2.24	0.28	−2.24	1.4
	CW	0.09	0.47	0.16	0.87	NA	−0.24	0.24	0.23	0.92	0.26	−0.24	0.92
	TN	NA	0.3	1.05	NA	NA	−0.29	0.63	1.35	−4.06	−0.07	−4.06	1.35
	CT	−0.07	1.34	1.41	−0.75	NA	−0.75	−0.76	1.19	0.58	0.32	−0.76	1.41
	CN	NA	1.14	1.54	NA	NA	0.42	1.82	1.38	−3.62	0.47	−3.62	1.82
Soy	TW	−0.87	0.43	−0.22	1.58	NA	0.45	0.34	−0.45	−1.07	0.11	−1.07	1.58
	WN	NA	−1.62	−2.45	NA	NA	−14.08	−1.45	−0.77	0.02	−1.45	−14.08	0.02
	CW	−0.46	1.13	0.7	1.53	NA	1.32	0.63	−0.12	−0.66	0.47	−0.66	1.53
	TN	NA	−2.12	−7.01	NA	NA	−3.27	−2	−4.03	−6.18	−3.52	−7.01	−2
	CT	−0.29	0.92	1.22	0.43	NA	−0.56	−0.05	0.71	−0.12	0.16	−0.56	1.22
	CN	NA	−1.72	−1.53	NA	NA	−5.54	−1.49	−2.98	−2.46	−3.07	−5.54	−1.49
Spring wheat	TW	−0.29	1.11	0.34	1.25	−0.61	0.51	0.52	0.15	0.82	0.33	−0.61	1.25
	WN	NA	0.02	0.4	NA	1.22	−0.78	−0.17	0.69	−2.77	0.02	−2.77	1.22
	CW	−0.34	1.91	1.26	2.12	−0.06	1.37	0.89	0.65	0.43	0.36	−0.34	2.12
	TN	NA	0.28	0.56	NA	0.46	−0.48	−0.04	1	−2.46	0.06	−2.46	1
	CT	−0.93	0.46	1.18	0.05	−1.69	0.21	0.4	0.64	−0.13	−0.06	−1.69	1.18
	CN	NA	0.64	1.41	NA	1.18	0.04	0.48	1.16	−2.27	0.45	−2.27	1.41
Winter wheat	TW	−0.3	0.93	0.2	NA	−0.1	0.1	0.39	0.73	0.08	0.18	−0.3	0.93
	WN	NA	0.39	0.51	NA	1.34	−0.41	−0.22	0.72	−1.37	0.28	−1.37	1.34
	CW	−0.21	1.88	0.98	NA	0	1.05	1.41	0.64	0.42	0.26	−0.21	1.88
	TN	NA	1.34	0.75	NA	0.84	−0.2	0.07	1.25	−1.88	0.41	−1.88	1.34
	CT	−0.46	−0.04	0.11	NA	−0.35	0.04	0.28	−0.1	0.38	−0.12	−0.46	0.38
	CN	NA	2.05	1.13	NA	1.27	0.18	0.73	1.2	−1.96	0.75	−1.96	2.05

Note. CARAIB and JULES did not supply data for different N levels, LPJ-GUESS did not supply data for rice or soybean, JULES did not supply data for winter wheat. These missing data points are indicated by "NA" (not available).

agreement ($n = 6$) on presence of cluster #2 CTW-*YRT* in Europe and parts of North America and moderate agreement for South America, and parts of Africa.

CTW-*YRTs* for spring wheat are more mixed (Figures S25 and S26). CARAIB, LPJ-GUESS and LPJmL show mostly clusters #1 (W with little importance and C vs. T balanced, Figure S25) and #2 (all balanced), but CARAIB has these two in approximately equal shares, while LPJ-GUESS has substantially more #1 and LPJmL substantially more #2. EPIC-TAMU, GEPIC and JULES are substantially more sensitive to W, JULES with mostly #3 (W dominates, C vs. T is balanced), GEPIC with mostly cluster #4 (C with little importance and T vs. W balanced). Spatial patterns are also mixed, with little pockets of multi-model agreement across all continents.

Also for winter wheat CTW-*YRTs* show a mixed picture with 6 distinct clusters (Figures S27 and S28). Here, a divide can be seen along the importance of C: only in four of the eight GGCMs (EPIC-TAMU, GEPIC, pDSSAT, PEPIC), cluster #6 can be found (in which C is of little importance and W dominates T), while cluster #2 (C dominates and T vs. W is balanced) is basically absent in these GGCMs. Cluster #2 is particularly widespread in CARAIB and LPJ-GUESS. Spatial patterns show little consistency across GGCMs.

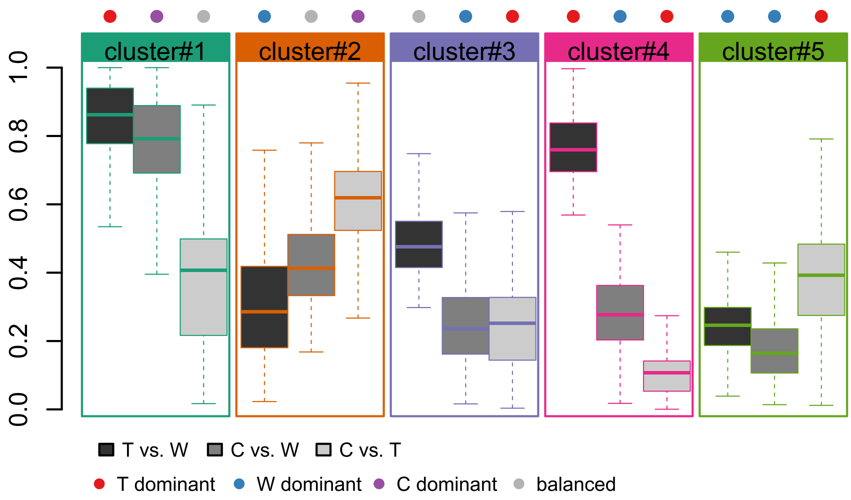

Figure 3. Distribution of *RRs* within CTW-*YRT* clusters. Within each cluster (colored boxes), three boxplots describe the distribution of *RRs* for T versus W (dark gray, left boxplot), C versus W (gray, middle boxplot), C versus T (light gray, right boxplot). Horizontal lines indicate the median value, boxes extent across the interquartile range (IQR). Whiskers extend to the most extreme value within 1.5 times the IQR, outliers beyond this threshold are omitted. Colored dots on top of each cluster box indicate what drivers dominates: red for T dominance, blue for W dominance, purple for C dominance, and gray for no dominance. We rate drivers as balanced (i.e., no dominance) if the median *RR* is between 0.4 and 0.6.

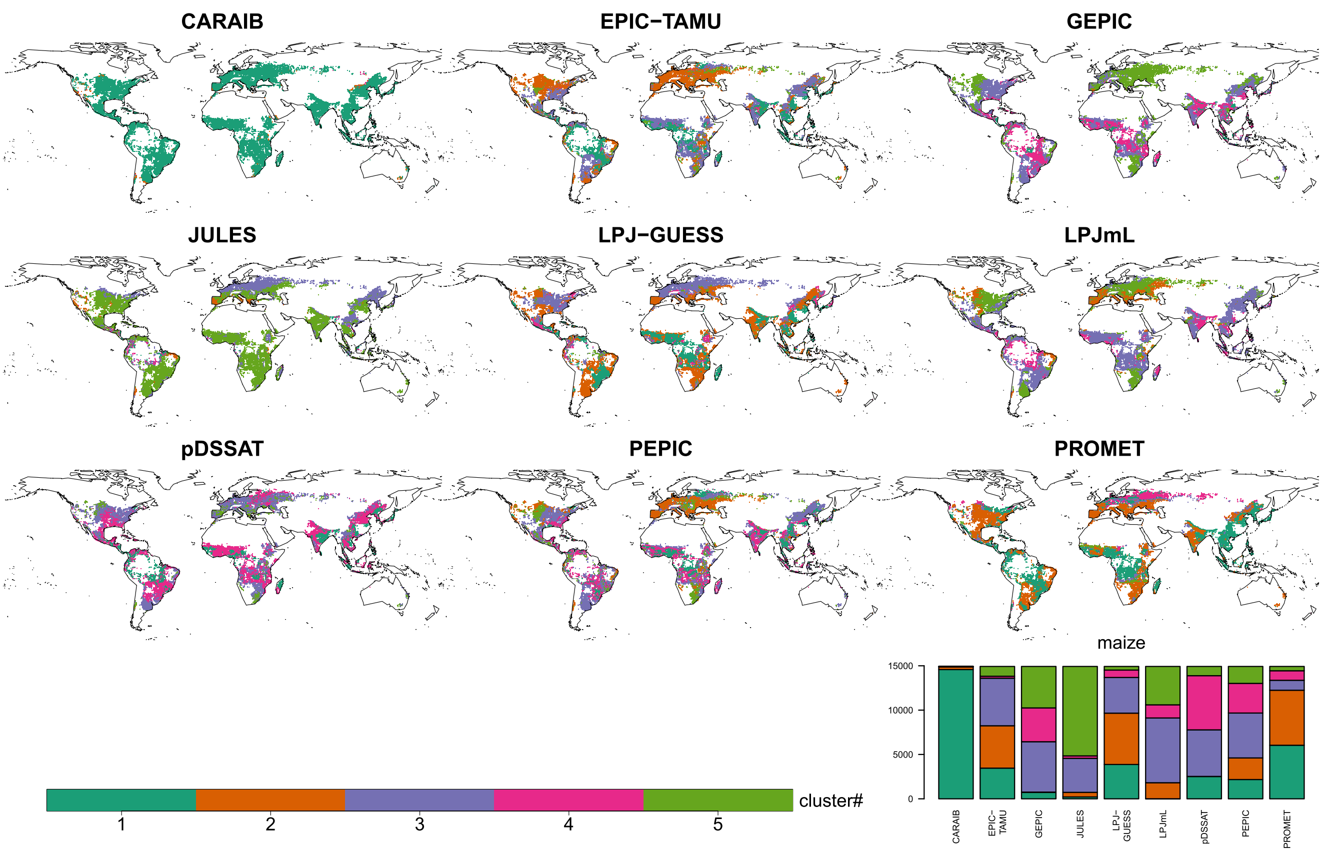

Figure 4. Spatial distribution of Crop Yield Response Types (*YRTs*) for each of the nine Global Gridded Crop Models (GGCMs) for maize, considering the C, T, and W dimensions, but without consideration of the N dimension, because this was not supplied by all GGCMs. The stacks in the bottom right-hand corner show the grid cell frequency distribution of the *YRT* clusters for each GGCM. The *RR* combinations characterizing each cluster are shown in Figure 3 above.

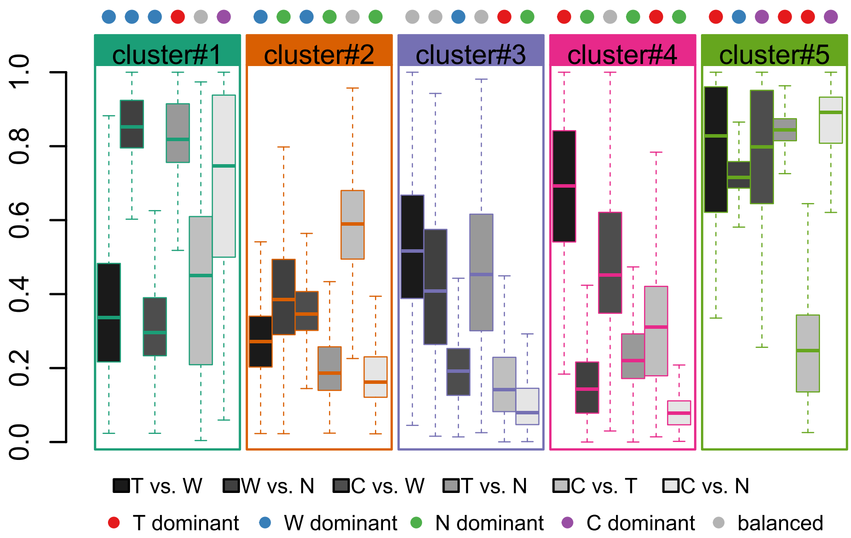

Figure 5. As Figure 3 but now for the full CTWN set with nitrogen and six combinations of any two drivers (gray shadings of boxplots, ordered from left to right).

If including the N dimension, the number of GGCMs is reduced to at most seven (Table 1), while the number of combinations of any two drivers increases to six. Still, the hierarchical clustering finds similar number of clusters with the threshold of 10% of overall variance within the clusters. The two models that show very little sensitivity of maize yields to water (CARAIB) or very high (JULES) did not provide any data along the N dimension, but within the reduced ensemble with N, there are again two models that show opposite behavior (Figures 5 and 6): LPJ-GUESS has a very strong response of maize yields to N either in combination with strong response to W (clusters #2) or with combination with strong response to T (cluster #4), while PROMET has little response to N either with strong sensitivity to W (cluster #1) or T (cluster #5). LPJmL and pDSSAT maize yields are dominated

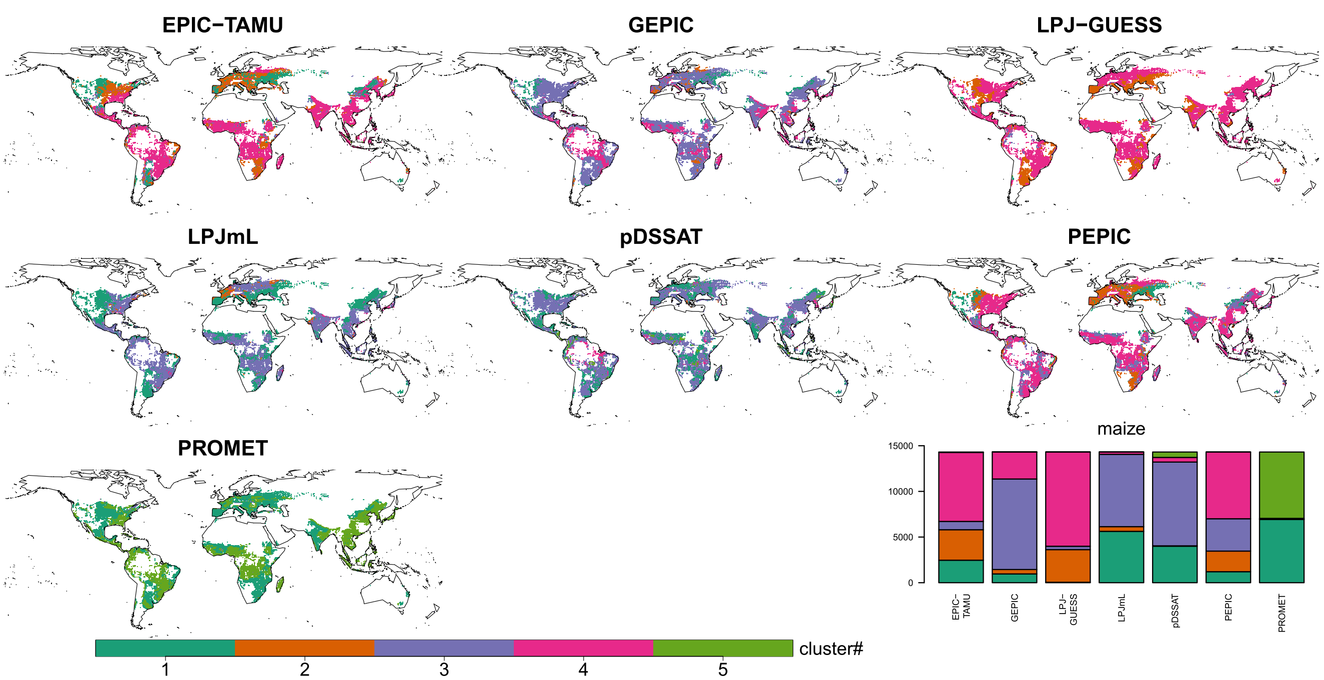

Figure 6. Spatial distribution of Crop Yield Response Types (*YRTs*) for each of the seven Global Gridded Crop Models (GGCMs) for maize including all four response dimensions (i.e., C, T, W, N). The stacks in the bottom right-hand corner show the grid cell frequency distribution of the *YRT* clusters for each GGCM. The *RR* combinations characterizing each cluster are shown in Figure 5 above.

10.1029/2023EF003773

by clusters #3 (with balanced responses, but T, W, and N all dominate C) and #1 (with mostly water dominance and little importance of N). Also GEPIC maize yields show large shares of cluster #3, but in combination with cluster #4. PROMET shows very little sensitivity to N also in rice yields (Figures S29 and S30) with almost all pixels being clustered in cluster #1, while all other GGCMs show strong importance of N in clusters #3 and #4, except for LPJmL, which has basically no occurrence of cluster #3 but of cluster #2 (where dimensions are more balanced but T dominates W and N), which is not very predominant in all other GGCMs. Spatial patterns of EPIC-TAMU, GEPIC, pDSSAT, and PEPIC rice sensitivities show some consistency, including LPJmL for Asia. For soybean, all six GGCMs that provided data, see little importance of N (with soybean being an N-fixing leguminous crop). PROMET soybean yield simulations are mostly in *YRT* cluster #3 (dominance of T and C), which is basically absent in GEPIC and LPJmL simulations (Figures S31 and S32). These two GGCMs find mostly clusters #2 (everything balanced, unless if compared to N) and #4 (T and C dominance). W and C dominance as in *YRT* cluster #1 is rare, but there is some cross-GGCM agreement on the regional occurrence of this *YRT* in SE-Europe and northern USA/Canada. Spring wheat *YRTs* are more mixed across GGCMs and regions. PROMET shows again little sensitivity to N with clusters #4 (W dominates and little importance of N otherwise) and #5 (T and C dominate). LPJmL also shows large shares of #4, but in combination with #1 (W and N dominate) and #2 (N and C dominate). *YRT* clusters #1 and #2 are also predominant for spring wheat *YRTs* of EPIC-TAMU, GEPIC, and PEPIC. For winter wheat *YRTs*, PROMET yield simulations shows also little sensitivity to N (clusters #3 with W dominance and #4 with T and C dominance), whereas LPJ-GUESS is most sensitive to N (cluster #1 with N dominance and all others balanced). Cluster #3 with W dominance is also found to some larger extent in LPJmL simulations, whereas all other GGCMs show large shares of cluster #2 with W and N dominance.

3.3. Emergent Functional Relationships

There are also different emergent functional relationships among GGCMs, that is, changes in functional responses that can be observed (emergent) but that we cannot attribute to actual model code structure or parameterization. Making use of the median *RRs*, we analyze how these change as a function of the other driver dimensions. Owing to the complexity of the data set, we constrain this analysis to median $RR_{T,W}$ responses to changes in C and N (Figure 7 for spring wheat, Figures S27–S40 for the other crops).

In some models and crops, the median $RR_{T,W}$ is hardly affected by increasing C (e.g., CARAIB, EPIC-TAMU, GEPIC, pDSSAT for spring wheat, Figure 7), whereas there are more pronounced changes in the median spring wheat $RR_{T,W}$ with changes in C for the other models. Similarly, the median $RR_{T,W}$ changes only little under different levels of N supply for some GGCMs (e.g., EPIC-TAMU, LPJ-GUESS, LPJmL, pDSSAT, PROMET for spring wheat, Figure 7) but more strongly in others. Also the direction of change varies across GGCMs. While some show an increasing importance of T versus W with increasing N supply (e.g., EPIC-TAMU, GEPIC, PEPIC for spring wheat, Figure 7), others see the opposite (decreasing importance of T vs. W with increasing N supply) or mixed cases. The combination of changes in C and N can lead to different emergent functional relationships, too: PEPIC spring wheat simulations show an increasing importance of T versus W with increasing C under high N supply, but a substantially lower importance of T versus W at low N supply and also a decreasing trend with higher C. Similar emergent functional relationships can be observed for the other crops analyzed here, but there are also crop-specific differences for some individual models. CARAIB shows always high median $RR_{T,W}$ values with little to no effect from changes in C across all five crops. EPIC-TAMU, GEPIC, and PEPIC all show very strong responses in median $RR_{T,W}$ to changes in N supply for rice (Figure S38), but much less so for winter wheat (Figure S40). LPJmL and PROMET see increasing median $RR_{T,W}$ with C for winter wheat (Figure S40), but less so for other crops (PROMET also increasing values for maize, Figure S37).

4. Discussion

We find that the crop models contributing to the GGCMI Phase 2 experiment (Franke et al., 2020a, 2020b) show substantial differences in yield responses to drivers along the carbon dioxide (C), temperature (T), water (W), and nitrogen (N) dimensions. These differences are caused by model structure and mechanisms as well as parameterization (Folberth et al., 2019). Because not all GGCMs provided the full set of CTWN simulations (Franke et al., 2020a, 2020b) (see Table 1), we used the emulators developed on these simulation sets (Franke et al., 2020a, 2020b) to gap-fill missing elements. Even though the emulators show generally good skill in reproducing model results, yield responses along the N dimension were particularly difficult to emulate with the low number of experiments in that dimension ($n = 3$, see Table A1). Also the number of simulations that needed to be

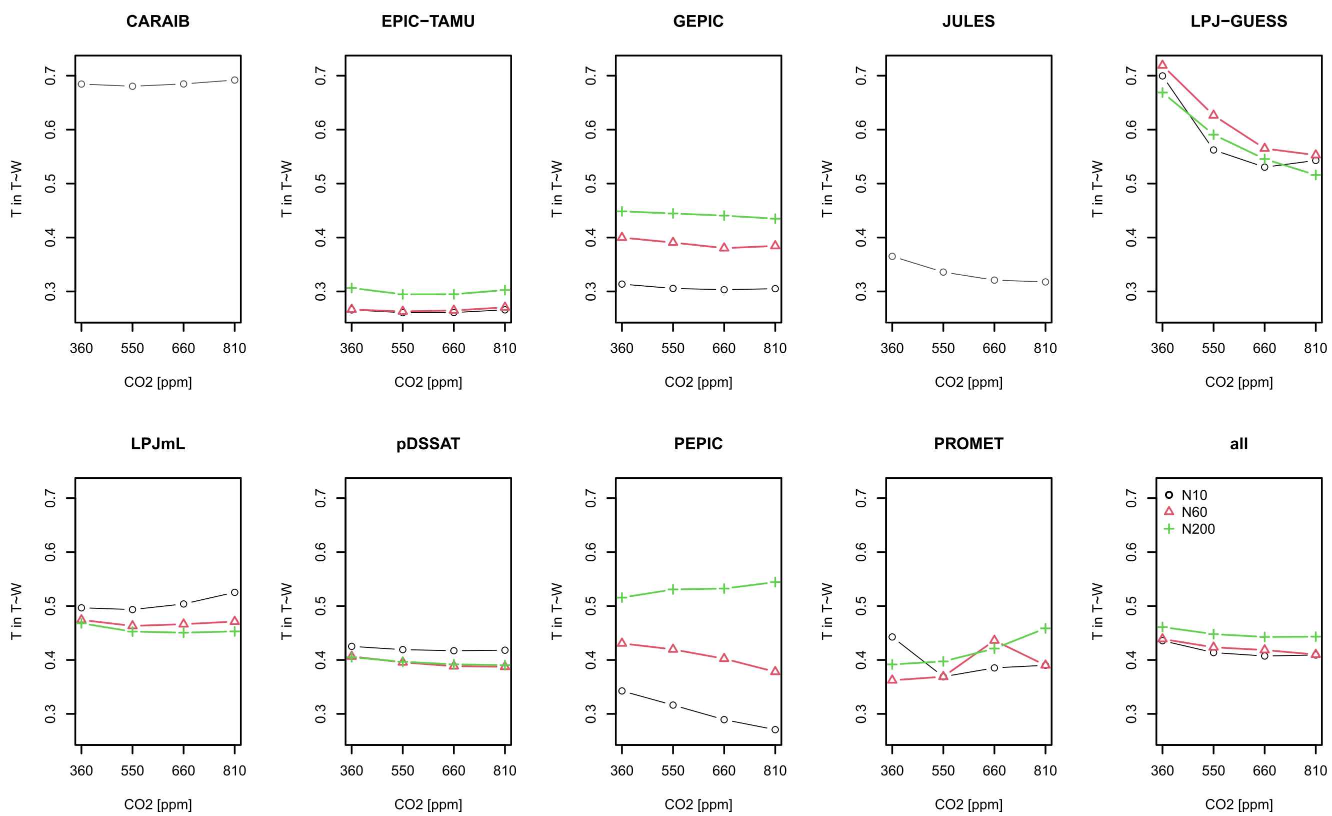

Figure 7. Spring wheat median $RR_{T,W}$ values under different levels of atmospheric CO_2 (x-axis) and N supply (colors). CARAIB and JULES did not supply data at different levels of nitrogen supply, their generic N response in shown in gray.

supplemented by emulated responses affects how well 'true' GGCM responses can be reproduced by the emulators. However, PEPIC, which supplied the smallest number of simulations of the ensemble considered here (Table 1) is in relatively good agreement with the other EPIC-based GGCMs considered here. This mechanism could also be a possible reason for the low N sensitivity of PROMET, which had also supplied only a small number of simulation sets for different N levels (Franke et al., 2020a, 2020b). The selection of the CTWN drivers does not cover the full range of climatic drivers of crop yield change (Schauberger et al., 2016) and albeit these are important, further research on additional drivers, such as irradiation as included in the study of Tao et al. (2020), would be helpful (Ruane et al., 2022).

Median *RRs* show a broad range of values, but some of this is expected. So is the role of CO_2 fertilization not very strong for maize as a C4 plant or the role of N inputs is relatively small for soybean, which can acquire atmospheric N via biological N fixation. The extreme soybean *RR* values for driver combinations with N (Table 2) can thus be explained by model design, but also indicate that more complex implementations should be implemented, as for example, done in a later LPJ-GUESS version than the one used here (Ma et al., 2022). Apart from such general responses of little C importance for maize yield simulations or little N importance for soybean yield simulations, large differences in the sensitivity to different drivers exist between models.

The skewness of *RR* distributions is in part determined by the environmental conditions of the spatial sample, that is, the actual crop-growing areas, yet model differences are also dominant here. In some cases, some models find highly negatively skewed distributions, whereas others find highly positive skewed distributions (e.g., −1.42 for CARAIB vs. 1.45 for JULES for the distribution of maize $RR_{T,W}$, or −1.69 for LPJ-GUESS vs. 1.18 for GEPIC for spring wheat $RR_{C,T}$). While we cannot expect normally distributed *RR* values as the cropland sample may not reflect normally distributed growing conditions, differences in skewness across GGCMs are only attributable to model functionality as they all use the same spatial sample here.

The clustering of *RR* values to *YRTs* illustrates that differences in response types can be larger between GGCMs than between regions. The arbitrary choice of leaving 10% of overall variance within clusters led to a small number of clusters ($n < 7$) that allow for qualitative description of their characteristics and interpretability. Even though this threshold does not follow any formal definition of the optimal number of clusters, we argue that it is important to only have a small number of clusters for discussing regional and GGCM-specific distributions of *YRTs*. The dendrograms (Figures S11–S20) show that the number of clusters is not very sensitive to smaller variations of the 10% threshold but that the number of clusters dramatically increases at thresholds <5%. Some models show consistent behavior across different crops (e.g., PROMET is typically not very sensitive to changes in N compared to changes in any other driver and CARAIB is not very sensitive to changes in W compared to other drivers). LPJ-GUESS shows greater sensitivity to C than other models for spring wheat, but greater sensitivity to N than any other model for winter wheat, where CARAIB shows the greatest sensitivity to C. PROMET also shows greatest sensitivity to C of winter wheat, but only from the ensemble that also supplied data on the N dimension, even though it tends to be rather insensitive to N in general. Similarity in spatial patterns of *YRTs* across some GGCMs and crops suggest that growing environments can be the dominant determinant of model sensitivities, as should be expected for perfect models. Differences in spatial patterns can stem from smaller differences along cluster borders that result in different clusters and suggest significant differences (classification problem) or differences in regional parameterizations, as applied by some GGCMs (Folberth et al., 2019), reflecting how sparsely the global diversity in farming systems (e.g., Jarvis et al., 2008) is reflected in crop models. Nonetheless, differences in spatial patterns across GGCMs suggest that differences in models' sensitivities to environmental drivers needs further attention from model development and application.

It can be expected that crop yields show interacting responses to simultaneous changes in CTWN drivers. If, for example, N limitation is lifted, W limitation may show more clearly and vice versa. The EPIC model has a maximum function approach and only considers the most severe from several stressors in daily biomass gains (Sharpley & Williams, 1990; J. R. Williams, 1990) and indeed, the EPIC-based GGCMs (EPIC-TAMU, GEPIC, PEPIC) show substantial differences in $RR_{T,W}$ under different N levels, except for soybean (as an N-fixing plant) and only to some limited extent for winter wheat. The GGCMs of the GGCMI Phase2 ensemble show a range of emergent functional relationships, varying between no effect (e.g., pDSSAT for soybean, Figure S39), layered but flat effects (e.g., GEPIC for maize, Figure S37), increasing (e.g., LPJmL for winter wheat, Figure S40) or decreasing (e.g., LPJ-GUESS for spring wheat, Figure 7). While it is quite possible that these emergent functional relationships should differ between crops, because of their physiological traits (e.g., C3 vs. C4 photosynthesis) and where they are grown, there should not be substantial differences in the overall *RR* level or in the direction of change under variations in additional drivers. We here only analyze highly aggregated data (global and temporal aggregation), but aggregation typically leads to more balanced responses with extremes canceling out so that even stronger differences can be expected at the more detailed level (individual sites and years).

We find that in a standard model evaluation, focusing on reproducing grain yield as for example, by Müller et al. (2017) (see Figures S2–S10), model differences are not as prominent as in response types and functional relationships. As such, crop models need to be evaluated not only with respect to reproduce observed yield dynamics, because final yields are affected by a multitude of processes and drivers (Schauberger et al., 2016) and Zhu et al. (2019) showed that error compensation in maize simulations can lead to accurate yield estimates. Different emergent functional relationships have been reported also for model intercomparison studies at site scale (Tao et al., 2020) and for other crops (e.g., E. Wang et al., 2022) and can originate from model parameterization (e.g., through different calibration methods), choice of subroutines (e.g., for potential evapotranspiration (Cammarano et al., 2016; Folberth et al., 2019)) or modeler choices (Albanito et al., 2022; Folberth et al., 2019; E. Wang et al., 2022). Fronzek et al. (2018) attempt to relate process implementation with IRS classes, identifying evapotranspiration models, soil water modules, and heat stress modules as important determinants of similarity between crop models.

Data availability for crop model evaluation on aspects other than yields is still a strong limitation, especially at large-scale applications. Remote sensing products may fill this gap to some extent (e.g., Cetin et al., 2023; Jiang et al., 2023; Jin et al., 2018; Yue et al., 2023).

Testing models for emergent functional properties, as also requested by Tao et al. (2020) could be an alternative approach to model evaluation, which requires knowledge on functional relationships and structured model experiments, such as the GGCMI Phase2 experiment (Franke et al., 2020a, 2020b), even though experimental design targeted to identifying specific functional relationships should drastically reduce the computation demand that was associated with the GGCMI Phase2 experiment. There are some examples of testing model for emergent functional properties (e.g., Schauberger et al., 2017), and dedicated efforts for model evaluation on aspects other than yield (e.g., Kimball et al., 2019, for maize evapotranspiration) but this is typically not integrated into standard model evaluation exercises. Schneck et al. (2022) provide an assessment of the emergent property *water-use efficiency* of their land surface model across different precipitation and temperature regimes (sampled from a global simulation rather than a stylized experiment design). The Earth System Modeling community has established standards for model evaluation (e.g., ESMValTool v2.0 Eyring et al., 2020), which can provide guidance from a technical and conceptual perspective. Yet, the climate system is described (and evaluated) by many different variables in contrast to the focus on the single end-of-season variable *yield* in crop modeling, limiting the comparability of evaluation standards in the two research domains. Horak et al. (2021) suggest a process-based evaluation of Intermediate Complexity Atmospheric Research Models that is based on stylized modeling experiments to help models become right for the right reasons. This approach is likely easier to transfer to crop modeling and we suggest that the idea of process-based model evaluation from targeted simulation experiments is pursued in future crop model evaluation efforts. The sensitivity of models to calibration (e.g., E. Wang et al., 2022) and parameterization (e.g., Folberth et al., 2019; E. Wang et al., 2017) indicates that model evaluation needs to be conducted continuously and cannot be substituted by references to model description papers or earlier evaluation efforts. The approach by Brown et al. (2018) to identify standard tests and include these into the *user interface* of the crop model *APSIM* to facilitate better model development is a promising approach. Such easy to access standard tests based on specific experiments can guide model development, but may be more limited for testing different case-specific model parametrizations. While the approach Brown et al. (2018) is model specific, such standard tests can be generalized to crop models in general, as demonstrated by efforts on general model benchmarking in global vegetation modeling (Kelley et al., 2013). Better efforts in the crop modeling community for model testing and evaluation are needed. At the same time, model deficiencies that have been identified by testing models, such as by Y. Li et al. (2019), who found that an ensemble of GGMCs (Müller et al., 2017, 2019) fails to reproduce the observed yield penalties under extremely wet conditions, need to incentivize model development. Global-scale crop model applications for high-end climate scenarios can expose crop models to input data combinations that have never been used during model development and may lead to unwanted model behavior. GGCMI data are thus tested for implausible data ranges (see quality check scripts at https://github.com/AgMIP-GGCMI/phase3_quality_check), but functional model test could test for plausibility in response to different drivers, such as.

- along a broad nitrogen supply gradient, crop yields should first increase and then level off;
- along a soil water content gradient from dry to very wet, crop yields should first increase and then decrease (e.g., Y. Li et al., 2019);
- exposure to canopy temperatures above lethal temperature thresholds should lead to an abort of crop growth.

5. Conclusions

The diversity in *RR* indicates that GGCMs have very different sensitivities to different climatic drivers and nitrogen supply. This has been discussed in the literature with a strong focus on the role of CO_2 on yield formation (Toreti et al., 2020), as many studies had presented results with and without CO_2 fertilization effects (e.g., Rosenzweig et al., 2014), bringing attention to this particular effect. We find that changes in temperatures, water or nitrogen supply yield similar strong differences when it comes to model sensitivities. Model evaluation should advance to including emergent functional relationships that may tell more about model plausibility and skill than only comparison with yield data from observations. For this, existing knowledge needs to be collected, tested for generalizability, and translated into simple tests that models can be subjected to. Modeling protocols need to be designed to enable such functionality tests rather than only comparisons with yield data. A community effort is needed to bring together knowledge collection and formalization, model experiment design and model testing in order to advance crop modeling toward reduced uncertainty in crop model applications.

Appendix A: Additional Tables

The appendix contains a table with the overview of the Phase 2 Experiment simulation setup (Table A1) and a list of DOIs for the Output Data Sets (Table A2).

Table A1
Global Gridded Crop Model Intercomparison Phase 2 Experiment Levels

Input variable	Label	Simulated levels	Unit
Atmospheric CO_2	C	360, 510, 660, 810	ppm
Temperature	T	−1, 0, 1, 2, 3, 4, 6	°C
Applied nitrogen	N	10, 60, 200	kg ha^{-1}

Note. Temperature and precipitation values indicate the perturbations from the historical climatology, atmospheric carbon dioxide (CO_2) and nitrogen values indicate absolute levels used in the simulations. Simulations with unlimited irrigation (W_{inf}) and adapted cultivars (A1) are shown for completeness only, but were not considered in this analysis. NA, not applicable.

Table A2
List of DOIs for Global Gridded Crop Model Intercomparison Phase 2 Output Data Sets (Franke et al., 2020a, 2020b)

Model	Maize	Soybean	Rice	Winter wheat	Spring wheat
APSIM-UGOE	2582531	2582535	2582533	2582537	2582539
CARAIB	2582499	2582508	2582504	2582516	2582522
EPIC-IIASA	2582453	2582461	2582457	2582463	2582465
EPIC-TAMU	2582349	2582367	2582352	2582392	2582418
JULES	2582543	2582547	2582545	–	2582551
GEPIC	2582247	2582258	2582251	2582260	2582263
LPJ-GUESS	2581625	–	–	2581638	2581640
LPJmL	2581356	2581498	2581436	2581565	2581606
ORCHIDEE-crop	2582441	–	2582445	2582449	–
pDSSAT	2582111	2582147	2582127	2582163	2582178
PEPIC	2582341	2582433	2582343	2582439	2582455
PROMET	2582467	2582488	2582479	2582490	2582492

Note. The data URL can be constructed by replacing "XX" in "https://doi.org/10.5281/zenodo.XX" with the values in the table for the data set of interest (e.g., https://doi.org/10.5281/zenodo.2582531 for maize data simulated by APSIM-UGOE). The GGCMI Phase 2 data archive had to be split in several archives, because data sets were too large for hosting as one data set.

Data Availability Statement

The simulation outputs of GGCMI Phase 2 output variables that we analyze here (Balkovic et al., 2019a, 2019b, 2019c, 2019d, 2019e; Dury et al., 2019a, 2019b, 2019c, 2019d, 2019e; Elliott, 2019a, 2019b, 2019c, 2019d, 2019e; Falloon & Williams, 2019a, 2019b, 2019c, 2019d; Folberth, 2019a, 2019b, 2019c, 2019d, 2019e; Hoffmann & Koch, 2019a, 2019b, 2019c, 2019d, 2019e; W. Liu, 2019a, 2019b, 2019c, 2019d, 2019e; Müller, 2019a, 2019b, 2019c, 2019d, 2019e; Pugh et al., 2019a, 2019b, 2019c; Reddy et al., 2019a, 2019b, 2019c, 2019d, 2019e; X. Wang & Ciais, 2019a, 2019b, 2019c; Zabel et al., 2019a, 2019b, 2019c, 2019d, 2019e) are available on zenodo.org. See Table A2 for data DOIs. Due to data size, the archive had to be split in several archives. All scripts used for processing the data in this analysis are available at zenodo.org (Müller, 2024).

Acknowledgments

J.J. was supported by the NASA GISS Climate Impacts Group and the Open Philanthropy Project. J.F. was supported by the NSF NRT program (Grant DGE-1735359) and the NSF Graduate Research Fellowship Program (Grant DGE-1746045). A.C.R. received support from the NASA Earth Sciences Division via the GISS Climate Impacts Group. S.O. and T. P. acknowledge 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). W.L. was supported by the National Natural Science Foundation of China (Grant 32361143871). X.W. acknowledges support from the National Natural Science Foundation of China (Grant 42171096). Open Access funding enabled and organized by Projekt DEAL.

References

Albanito, F., McBey, D., Harrison, M., Smith, P., Ehrhardt, F., Bhatia, A., et al. (2022). How modelers model: The overlooked social and human dimensions in model intercomparison studies. *Environmental Science & Technology*, *56*(18), 13485–13498. https://doi.org/10.1021/acs.est.2c02023

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., et al. (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, *5*(2), 143–147. https://doi.org/10.1038/nclimate2470

Asseng, S., Martre, P., Maiorano, A., Rötter, R. P., O'Leary, G. J., Fitzgerald, G. J., et al. (2019). Climate change impact and adaptation for wheat protein. *Global Change Biology*, *25*(1), 155–173. https://doi.org/10.1111/gcb.14481

Balkovic, J., Khabarov, N., & Skalsky, R. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-IIASA maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582453

Balkovic, J., Khabarov, N., & Skalsky, R. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-IIASA rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582457

Balkovic, J., Khabarov, N., & Skalsky, R. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-IIASA soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582461

Balkovic, J., Khabarov, N., & Skalsky, R. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-IIASA spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582465

Balkovic, J., Khabarov, N., & Skalsky, R. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-IIASA winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582463

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

Boote, K. J., Jones, J. W., White, J. W., Asseng, S., & Lizaso, J. I. (2013). Putting mechanisms into crop production models. *Plant, Cell and Environment*, *36*(9), 1658–1672. https://doi.org/10.1111/pce.12119

Brown, H., Huth, N., & Holzworth, D. (2018). Crop model improvement in APSIM: Using wheat as a case study. *European Journal of Agronomy*, *100*, 141–150. https://doi.org/10.1016/j.eja.2018.02.002

Cammarano, D., Rötter, R. P., Asseng, S., Ewert, F., Wallach, D., Martre, P., et al. (2016). Uncertainty of wheat water use: Simulated patterns and sensitivity to temperature and CO_2. *Field Crops Research*, *198*, 80–92. https://doi.org/10.1016/j.fcr.2016.08.015

Cetin, M., Alsenjar, O., Aksu, H., Golpinar, M. S., & Akgul, M. A. (2023). Estimation of crop water stress index and leaf area index based on remote sensing data. *Water Supply*, *23*(3), 1390–1404. https://doi.org/10.2166/ws.2023.051

Couëdel, A., Edreira, J. I. R., Pisa Lollato, R., Archontoulis, S., Sadras, V., & Grassini, P. (2021). Assessing environment types for maize, soybean, and wheat in the United States as determined by spatio-temporal variation in drought and heat stress. *Agricultural and Forest Meteorology*, *307*, 108513. https://doi.org/10.1016/j.agrformet.2021.108513

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? *Journal of Statistics Education*, *19*(2), 18. https://doi.org/10.1080/10691898.2011.11889611

Dury, M., Francois, L., & Jacquemin, I. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from CARAIB maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582499

Dury, M., Francois, L., & Jacquemin, I. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from CARAIB rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582504

Dury, M., Francois, L., & Jacquemin, I. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from CARAIB soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582508

Dury, M., Francois, L., & Jacquemin, I. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from CARAIB spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582522

Dury, M., Francois, L., & Jacquemin, I. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from CARAIB winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582516

Dury, M., Hambuckers, A., Warnant, P., Henrot, A., Favre, E., Ouberdous, M., & François, L. (2011). Responses of European forest ecosystems to 21st century climate: Assessing changes in interannual variability and fire intensity. *iForest: Biogeosciences and Forestry*, *4*(2), 82–99. https://doi.org/10.3832/ifor0572-004

Elliott, J. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from pDSSAT maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582111

Elliott, J. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from pDSSAT rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582127

Elliott, J. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from pDSSAT soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582147

Elliott, J. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from pDSSAT spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582178

Elliott, J. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from pDSSAT winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582163

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., et al. (2014). 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

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., et al. (2015). The Global Gridded Crop Model Intercomparison: Data and modeling protocols for Phase 1 (v1.0). *Geoscientific Model Development*, *8*(2), 261–277. https://doi.org/10.5194/gmd-8-261-2015

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – An extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Falloon, P., & Williams, K. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from JULES maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582543

Falloon, P., & Williams, K. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from JULES rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582545

Falloon, P., & Williams, K. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from JULES soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582547

10.1029/2023EF003773

Falloon, P., & Williams, K. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from JULES spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582551

Folberth, C. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from GEPIC maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582247

Folberth, C. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from GEPIC rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582251

Folberth, C. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from GEPIC soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582258

Folberth, C. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from GEPIC spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582263

Folberth, C. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from GEPIC winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582260

Folberth, C., Elliott, J., Müller, C., Balkovič, J., Chryssanthacopoulos, J., Izaurralde, R. C., et al. (2019). Parameterization-induced uncertainties and impacts of crop management harmonization in a global gridded crop model ensemble. *PLoS One*, *14*(9), 1–36. https://doi.org/10.1371/journal.pone.0221862

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

Franke, J. A., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Balkovic, J., et al. (2020a). The GGCMI Phase 2 experiment: Global gridded crop model simulations under uniform changes in CO_2, temperature, water, and nitrogen levels (protocol version 1.0). *Geoscientific Model Development*, *13*(5), 2315–2336. https://doi.org/10.5194/gmd-13-2315-2020

Franke, J. A., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Snyder, A., et al. (2020b). The GGCMI Phase 2 emulators: Global gridded crop model responses to changes in CO_2, temperature, water, and nitrogen (version 1.0). *Geoscientific Model Development*, *13*(9), 3995–4018. https://doi.org/10.5194/gmd-13-3995-2020

Franke, J. A., Müller, C., Minoli, S., Elliott, J., Folberth, C., Gardner, C., et al. (2022). Agricultural breadbaskets shift poleward given adaptive farmer behavior under climate change. *Global Change Biology*, *28*(1), 167–181. https://doi.org/10.1111/gcb.15868

Fronzek, S., Pirttioja, N., Carter, T. R., Bindi, M., Hoffmann, H., Palosuo, T., et al. (2018). 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

Gommes, R., Wu, B., Li, Z., & Zeng, H. (2016). Design and characterization of spatial units for monitoring global impacts of environmental factors on major crops and food security. *Food and Energy Security*, *5*(1), 40–55. https://doi.org/10.1002/fes3.73

Hank, T., Bach, H., & Mauser, W. (2015). Using a remote sensing-supported hydro-agroecological model for field-scale simulation of heterogeneous crop growth and yield: Application for wheat in central Europe. *Remote Sensing*, *7*(4), 3934–3965. https://doi.org/10.3390/rs70403934

Hawkins, E., & Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics*, *37*(1), 407–418. https://doi.org/10.1007/s00382-010-0810-6

Hoffmann, M., & Koch, M. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from APSIM-UGOE maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582531

Hoffmann, M., & Koch, M. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from APSIM-UGOE rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582533

Hoffmann, M., & Koch, M. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from APSIM-UGOE soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582535

Hoffmann, M., & Koch, M. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from APSIM-UGOE spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582539

Hoffmann, M., & Koch, M. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from APSIM-UGOE winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.2582537

Holzworth, D. P., Huth, N. I., de Voil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., et al. (2014). APSIM – Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling and Software*, *62*, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

Horak, J., Hofer, M., Gutmann, E., Gohm, A., & Rotach, M. W. (2021). A process-based evaluation of the Intermediate Complexity Atmospheric Research Model (ICAR) 1.0.1. *Geoscientific Model Development*, *14*(3), 1657–1680. https://doi.org/10.5194/gmd-14-1657-2021

Izaurralde, R., Williams, J., Mcgill, W., Rosenberg, N., & Quiroga Jakas, M. (2006). Simulating soil C dynamics with EPIC: Model description and testing against long-term data. *Ecological Modelling*, *192*(3–4), 362–384. https://doi.org/10.1016/j.ecolmodel.2005.07.010

Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., et al. (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nature Food*, *2*(11), 873–885. https://doi.org/10.1038/s43016-021-00400-y

Jarvis, D. I., Brown, A. H. D., Cuong, P. H., Collado-Panduro, L., Latournerie-Moreno, L., Gyawali, S., et al. (2008). A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(14), 5326–5331. https://doi.org/10.1073/pnas.0800607105

Jiang, J., Atkinson, P. M., Chen, C., Cao, Q., Tian, Y., Zhu, Y., et al. (2023). Combining UAV and Sentinel-2 satellite multi-spectral images to diagnose crop growth and N status in winter wheat at the county scale. *Field Crops Research*, *294*, 108860. https://doi.org/10.1016/j.fcr.2023.108860

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., & Wang, J. (2018). A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy*, *92*, 141–152. https://doi.org/10.1016/j.eja.2017.11.002

Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., et al. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, *18*(3), 235–265. https://doi.org/10.1016/s1161-0301(02)00107-7

Keating, B., Carberry, P., Hammer, G., Probert, M., Robertson, M., Holzworth, D., et al. (2003). An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, *18*(3), 267–288. https://doi.org/10.1016/s1161-0301(02)00108-9

Kelley, D. I., Prentice, I. C., Harrison, S. P., Wang, H., Simard, M., Fisher, J. B., & Willis, K. O. (2013). A comprehensive benchmarking system for evaluating global vegetation models. *Biogeosciences*, *10*(5), 3313–3340. https://doi.org/10.5194/bg-10-3313-2013

Kimball, B. A., Boote, K. J., Hatfield, J. L., Ahuja, L. R., Stockle, C., Archontoulis, S., et al. (2019). Simulation of maize evapotranspiration: An inter-comparison among 29 maize models. *Agricultural and Forest Meteorology*, *271*, 264–284. https://doi.org/10.1016/j.agrformet.2019.02.037

Kostková, M., Hlavinka, P., Pohanková, E., Kersebaum, K. C., Nendel, C., Gobin, A., et al. (2021). Performance of 13 crop simulation models and their ensemble for simulating four field crops in Central Europe. *The Journal of Agricultural Science*, *159*(1–2), 69–89. https://doi.org/10.1017/S0021859621000216

Kummu, M., Heino, M., Taka, M., Varis, O., & Viviroli, D. (2021). Climate change risks pushing one-third of global food production outside the safe climatic space. *One Earth*, *4*(5), 720–729. https://doi.org/10.1016/j.oneear.2021.04.017

Li, T., Angeles, O., Marcaida, M., Manalo, E., Manalili, M. P., Radanielson, A., & Mohanty, S. (2017). From ORYZA2000 to ORYZA (v3): An improved simulation model for rice in drought and nitrogen-deficient environments. *Agricultural and Forest Meteorology*, *237–238*, 246–256. https://doi.org/10.1016/j.agrformet.2017.02.025

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., & Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. *Global Change Biology*, *25*(7), 2325–2337. https://doi.org/10.1111/gcb.14628

Lindeskog, M., Arneth, A., Bondeau, A., Waha, K., Seaquist, J., Olin, S., & Smith, B. (2013). Implications of accounting for land use in simulations of ecosystem carbon cycling in Africa. *Earth System Dynamics*, *4*(2), 385–407. https://doi.org/10.5194/esd-4-385-2013

Liu, D., Mishra, A. K., & Ray, D. K. (2020). Sensitivity of global major crop yields to climate variables: A non-parametric elasticity analysis. *Science of the Total Environment*, *748*, 141431. https://doi.org/10.1016/j.scitotenv.2020.141431

Liu, J., Williams, J. R., Zehnder, A. J., & Yang, H. (2007). GEPIC - Modelling wheat yield and crop water productivity with high resolution on a global scale. *Agricultural Systems*, *94*(2), 478–493. https://doi.org/10.1016/j.agsy.2006.11.019

Liu, W. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PEPIC maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582341

Liu, W. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PEPIC rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582343

Liu, W. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PEPIC soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582433

Liu, W. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PEPIC spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582455

Liu, W. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PEPIC winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582439

Liu, W., Yang, H., Folberth, C., Müller, C., Ciais, P., Abbaspour, K. C., & Schulin, R. (2018). Achieving high crop yields with low nitrogen emissions in global agricultural input intensification. *Environmental Science & Technology*, *52*(23), 13782–13791. https://doi.org/10.1021/acs.est.8b03610

Liu, W., Yang, H., Folberth, C., Wang, X., Luo, Q., & Schulin, R. (2016). 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

Liu, W., Yang, H., Liu, J., Azevedo, L. B., Wang, X., Xu, Z., et al. (2016). 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

Ma, J., Olin, S., Anthoni, P., Rabin, S. S., Bayer, A. D., Nyawira, S. S., & Arneth, A. (2022). Modeling symbiotic biological nitrogen fixation in grain legumes globally with LPJ-GUESS (v4.0, r10285). *Geoscientific Model Development*, *15*(2), 815–839. https://doi.org/10.5194/gmd-15-815-2022

Maiorano, A., Martre, P., Asseng, S., Ewert, F., Müller, C., Rötter, R. P., et al. (2017). Crop model improvement reduces the uncertainty of the response to temperature of multi-model ensembles. *Field Crops Research*, *202*, 5–20. https://doi.org/10.1016/j.fcr.2016.05.001

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., et al. (2015). Multimodel ensembles of wheat growth: Many models are better than one. *Global Change Biology*, *21*(2), 911–925. https://doi.org/10.1111/gcb.12768

Mauser, W., & Bach, H. (2015). PROMET - Large scale distributed hydrological modelling to study the impact of climate change on the water flows of mountain watersheds. *Journal of Hydrology*, *376*(8946), 362–377. https://doi.org/10.1016/j.jhydrol.2009.07.046

Mauser, W., Klepper, G., Zabel, F., Delzeit, R., Hank, T., Putzenlechner, B., & Calzadilla, A. (2009). Global biomass production potentials exceed expected future demand without the need for cropland expansion. *Nature Communications*, *6*(3), 8946. https://doi.org/10.1038/ncomms9946

Minoli, S., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Zabel, F., et al. (2019). Global response patterns of major rainfed crops to adaptation by maintaining current growing periods and irrigation. *Earth's Future*, *7*(12), 1464–1480. https://doi.org/10.1029/2018EF001130

Monerie, P.-A., Wainwright, C. M., Sidibe, M., & Akinsanola, A. A. (2020). Model uncertainties in climate change impacts on Sahel precipitation in ensembles of CMIP5 and CMIP6 simulations. *Climate Dynamics*, *55*(5–6), 1385–1401. https://doi.org/10.1007/s00382-020-05332-0

Müller, C. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJmL maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581356

Müller, C. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJmL rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581436

Müller, C. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJmL soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581498

Müller, C. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJmL spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581606

Müller, C. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJmL winter wheat simulations. [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581565

Müller, C. (2024). Set of R scripts for GGCMI data processing as used in "Müller et al. 2023: Substantial differences in crop yield sensitivities between models call for functionality-based model evaluation" [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.10520392

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., et al. (2017). Global gridded crop model evaluation: Benchmarking, skills, deficiencies and implications. *Geoscientific Model Development*, *10*(4), 1403–1422. https://doi.org/10.5194/gmd-10-1403-2017

Müller, C., Elliott, J., Kelly, D., Arneth, A., Balkovic, J., Ciais, P., et al. (2019). The global gridded crop model intercomparison phase 1 simulation dataset. *Scientific Data*, *6*(1), 1–22. https://doi.org/10.1038/s41597-019-0023-8

Müller, C., Franke, J., Jägermeyr, J., Ruane, A. C., Elliott, J., Moyer, E., et al. (2021). Exploring uncertainties in global crop yield projections in a large ensemble of crop models and cmip5 and cmip6 climate scenarios. *Environmental Research Letters*, *16*(3), 034040. https://doi.org/10.1088/1748-9326/abd8fc

Olin, S., Schurgers, G., Lindeskog, M., Wårlind, D., Smith, B., Bodin, P., et al. (2015). Modelling the response of yields and tissue C: N to changes in atmospheric CO_2 and N management in the main wheat regions of western Europe. *Biogeosciences*, *12*(8), 2489–2515. https://doi.org/10.5194/bg-12-2489-2015

Osborne, T., Gornall, J., Hooker, J., Williams, K., Wiltshire, A., Betts, R., & Wheeler, T. (2015). JULES-crop: A parametrisation of crops in the Joint UK Land Environment Simulator. *Geoscientific Model Development*, *8*(4), 1139–1155. https://doi.org/10.5194/gmd-8-1139-2015

Palosuo, T., Kersebaum, K. C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J. E., et al. (2011). Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. *European Journal of Agronomy*, *35*(3), 103–114. https://doi.org/10.1016/j.eja.2011.05.001

Pirttioja, N., Carter, T. R., Fronzek, S., Bindi, M., Hoffmann, H., Palosuo, T., et al. (2015). Temperature and precipitation effects on wheat yield across a European transect: A crop model ensemble analysis using impact response surfaces. *Climate Research*, *65*, 87–105. https://doi.org/10.3354/cr01322

Portmann, F. T., Siebert, S., & Döll, P. (2010). MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, *24*(1). https://doi.org/10.1029/2008GB003435

Pugh, T. A. M., Olin, S., & Arneth, A. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJ-GUESS maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581625

Pugh, T. A. M., Olin, S., & Arneth, A. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJ-GUESS spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581640

Pugh, T. A. M., Olin, S., & Arneth, A. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from LPJ-GUESS winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2581638

Qiao, L., Wang, X., Smith, P., Fan, J., Lu, Y., Emmett, B., et al. (2022). Soil quality both increases crop production and improves resilience to climate change. *Nature Climate Change*, *12*(6), 574–580. https://doi.org/10.1038/s41558-022-01376-8

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. *Vienna, Austria*. Retrieved from https://www.R-project.org/

Reddy, A., Jones, C., & Izaurralde, R. C. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-TAMU maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582349

Reddy, A., Jones, C., & Izaurralde, R. C. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-TAMU rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582352

Reddy, A., Jones, C., & Izaurralde, R. C. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-TAMU soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582367

Reddy, A., Jones, C., & Izaurralde, R. C. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from EPIC-TAMU spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582418

Reddy, A., Jones, C., & Izaurralde, R. C. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from epic-tamu winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582392

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., et al. (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences*, *111*(9), 3268–3273. https://doi.org/10.1073/pnas.1222463110

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., et al. (2013). 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

Rötter, R. P., Carter, T. R., Olesen, J. E., & Porter, J. R. (2011). Crop–climate models need an overhaul. *Nature Climate Change*, *1*(4), 175–177. https://doi.org/10.1038/nclimate1152

Ruane, A. C., Phillips, M., Müller, C., Elliott, J., Jägermeyr, J., Arneth, A., et al. (2021). Strong regional influence of climatic forcing datasets on global crop model ensembles. *Agricultural and Forest Meteorology*, *300*, 108313. https://doi.org/10.1016/j.agrformet.2020.108313

Ruane, A. C., Phillips, M. M., & Rosenzweig, C. (2018). Climate shifts within major agricultural seasons for +1.5 and +2.0°C worlds: HAPPI projections and AgMIP modeling scenarios. *Agricultural and Forest Meteorology*, *259*, 329–344. https://doi.org/10.1016/j.agrformet.2018.05.013

Ruane, A. C., Rosenzweig, C., Asseng, S., Boote, K. J., Elliott, J., Ewert, F., et al. (2017). An AgMIP framework for improved agricultural representation in integrated assessment models. *Environmental Research Letters*, *12*(12), 125003. https://doi.org/10.1088/1748-9326/aa8da6

Ruane, A. C., Vautard, R., Ranasinghe, R., Sillmann, J., Coppola, E., Arnell, N., et al. (2022). The climatic impact-driver framework for assessment of risk-relevant climate information. *Earth's Future*, *10*(11), e2022EF002803. https://doi.org/10.1029/2022EF002803

Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., et al. (2017). Consistent negative response of US crops to high temperatures in observations and crop models. *Nature Communications*, *8*(1), 13931. https://doi.org/10.1038/ncomms13931

Schauberger, B., Rolinski, S., & Müller, C. (2016). A network-based approach for semi-quantitative knowledge mining and its application to yield variability. *Environmental Research Letters*, *11*(12), 123001. https://doi.org/10.1088/1748-9326/11/12/123001

Schleussner, C.-F., Deryng, D., Müller, C., Elliott, J., Saeed, F., Folberth, C., et al. (2018). Crop productivity changes in 1.5°C and 2°C worlds under climate sensitivity uncertainty. *Environmental Research Letters*, *13*(6), 064007. https://doi.org/10.1088/1748-9326/aab63b

Schneck, R., Gayler, V., Nabel, J. E. M. S., Raddatz, T., Reick, C. H., & Schnur, R. (2022). Assessment of JSBACHv4.30 as a land component of ICON-ESM-V1 in comparison to its predecessor JSBACHv3.2 of MPI-ESM1.2. *Geoscientific Model Development*, *15*(22), 8581–8611. https://doi.org/10.5194/gmd-15-8581-2022

Sharpley, A., & Williams, J. (1990). *EPIC-Erosion/Productivity Impact Calculator: 1. Model documentation*. (No. Technical Bulletin No. 1768). US Department of Agriculture. Retrieved from https://epicapex.tamu.edu/media/h2gkyznv/epicmodeldocumentation.pdf

Stevanović, M., Popp, A., Lotze-Campen, H., Dietrich, J. P., Müller, C., Bonsch, M., et al. (2016). The impact of high-end climate change on agricultural welfare. *Science Advances*, *2*(8), e1501452. https://doi.org/10.1126/sciadv.1501452

Tao, F., Palosuo, T., Rötter, R. P., Díaz-Ambrona, C. G. H., Inés Mínguez, M., Semenov, M. A., et al. (2020). Why do crop models diverge substantially in climate impact projections? A comprehensive analysis based on eight barley crop models. *Agricultural and Forest Meteorology*, *281*, 107851. https://doi.org/10.1016/j.agrformet.2019.107851

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., et al. (2021). Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6. *Earth System Dynamics Discussions*, *12*(1), 253–293. https://doi.org/10.5194/esd-12-253-2021

Toreti, A., Deryng, D., Tubiello, F. N., Müller, C., Kimball, B. A., Moser, G., et al. (2020). Narrowing uncertainties in the effects of elevated CO_2 on crops. *Nature Food*, *1*(12), 775–782. https://doi.org/10.1038/s43016-020-00195-4

Valade, A., Ciais, P., Vuichard, N., Viovy, N., Caubel, A., Huth, N., et al. (2014). Modeling sugarcane yield with a process-based model from site to continental scale: Uncertainties arising from model structure and parameter values. *Geoscientific Model Development*, *7*(3), 1225–1245. https://doi.org/10.5194/gmd-7-1225-2014 van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., & Hochman, Z. (2013). Yield gap analysis with local to global relevance—A review. *Field Crops Research*, *143*, 4–17. https://doi.org/10.1016/j.fcr.2012.09.009

von Bloh, W., Schaphoff, S., Müller, C., Rolinski, S., Waha, K., & Zaehle, S. (2018). Implementing the nitrogen cycle into the dynamic global vegetation, hydrology, and crop growth model LPJmL (version 5.0). *Geoscientific Model Development*, *11*(7), 2789–2812. https://doi.org/10.5194/gmd-11-2789-2018

Wang, E., He, D., Wang, J., Lilley, J. M., Christy, B., Hoffmann, M. P., et al. (2022). How reliable are current crop models for simulating growth and seed yield of canola across global sites and under future climate change? *Climate Change*, *172*(1), 1–22. https://doi.org/10.1007/s10584-022-03375-2

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R. P., et al. (2017). The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nature Plants*, *3*(17102), 1–13. https://doi.org/10.1038/nplants.2017.102

Wang, X., & Ciais, P. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from ORCHIDEE-crop maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582441

Wang, X., & Ciais, P. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from ORCHIDEE-crop rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582445

Wang, X., & Ciais, P. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from ORCHIDEE-crop winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582449

Wiebe, K., Lotze-Campen, H., Sands, R., Tabeau, A., van der Mensbrugghe, D., Biewald, A., et al. (2015). Climate change impacts on agriculture in 2050 under a range of plausible socioeconomic and emissions scenarios. *Environmental Research Letters*, *10*(8), 085010. https://doi.org/10.1088/1748-9326/10/8/085010

Williams, J. R. (1990). The erosion-productivity impact calculator (EPIC) model: A case history on JSTOR. *Philosophical Transactions: Biological Sciences*, *329*(1255), 421–428.

Williams, K., & Falloon, P. D. (2015). Sources of interannual yield variability in JULES-crop and implications for forcing with seasonal weather forecasts. *Geoscientific Model Development*, *8*(12), 3987–3997. https://doi.org/10.5194/gmd-8-3987-2015

Williams, K., Gornall, J., Harper, A., Wiltshire, A., Hemming, D., Quaife, T., et al. (2017). Evaluation of JULES-crop performance against site observations of irrigated maize from Mead, Nebraska. *Geoscientific Model Development*, *10*(3), 1291–1320. https://doi.org/10.5194/gmd-10-1291-2017

Wu, Y., Miao, C., Fan, X., Gou, J., Zhang, Q., & Zheng, H. (2022). Quantifying the uncertainty sources of future climate projections and narrowing uncertainties with bias correction techniques. *Earth's Future*, *10*(11), e2022EF002963. https://doi.org/10.1029/2022EF002963

Yue, J., Yang, H., Yang, G., Fu, Y., Wang, H., & Zhou, C. (2023). Estimating vertically growing crop above-ground biomass based on UAV remote sensing. *Computers and Electronics in Agriculture*, *205*, 107627. https://doi.org/10.1016/j.compag.2023.107627

Zabel, F., Hank, T., & Mauser, W. (2019a). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PROMET maize simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582467

Zabel, F., Hank, T., & Mauser, W. (2019b). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PROMET rice simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582479

Zabel, F., Hank, T., & Mauser, W. (2019c). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PROMET soybean simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582488

Zabel, F., Hank, T., & Mauser, W. (2019d). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PROMET spring wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582492

Zabel, F., Hank, T., & Mauser, W. (2019e). AgMIP's global gridded crop model intercomparison (GGCMI) phase II CTWN-A archive: Priority 1 outputs from PROMET winter wheat simulations [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.2582490

Zabel, F., Müller, C., Elliott, J., Minoli, S., Jägermeyr, J., Schneider, J. M., et al. (2021). Large potential for crop production adaptation depends on available future varieties. *Global Change Biology*, *27*(16), 3870–3882. https://doi.org/10.1111/gcb.15649

Zhu, P., Zhuang, Q., Archontoulis, S. V., Bernacchi, C., & Müller, C. (2019). Dissecting the nonlinear response of maize yield to high temperature stress with model-data integration. *Global Change Biology*, *25*(7), 2470–2484. https://doi.org/10.1111/gcb.14632