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

James Franke[1,2], Christoph Müller[3], Joshua Elliott[2,4], Alexander Ruane[5], Jonas Jägermeyr[3,2,4,5], Abigail Snyder[6], Marie Dury[7], Pete Falloon[8], Christian Folberth[9], Louis François[7], Tobias Hank[10], Cesar Izaurralde[11,12], Ingrid Jacquemin[7], Curtis Jones[11], Michelle Li[2,13], Wenfeng Liu[14,15], Stefan Olin[16], Meridel Phillips[5,17], Thomas Pugh[18,19], Ashwan Reddy[11], Karina Williams[8], Ziwei Wang[1,2], Florian Zabel[10], and Elisabeth Moyer[1,2]

1. Department of the Geophysical Sciences, University of Chicago, Chicago, IL, USA.
2. Center for Robust Decision-making on Climate and Energy Policy (RDCEP), University of Chicago, Chicago, IL, USA.
3. Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany.
4. Department of Computer Science, University of Chicago.
5. NASA Goddard Institute for Space Studies, New York, NY, United States.
6. Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA.
7. Unité de Modélisation du Climat et des Cycles Biogéochimiques, UR SPHERES, Institut d'Astrophysique et de Géophysique, University of Liège, Belgium.
8. Met Office Hadley Centre, Exeter, United Kingdom.
9. Ecosystem Services and Management Program, IIASA, Laxenburg, Austria.
10. Department of Geography, Ludwig-Maximilians-Universität, Munich, Germany.
11. Department of Geographical Sciences, University of Maryland, College Park, MD, USA.
12. Texas Agrilife Research and Extension, Texas A&M University, Temple, TX, USA.
13. Department of Statistics, University of Chicago, Chicago, IL, USA.
14. EAWAG, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland.
15. Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
16. Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden.
17. Earth Institute Center for Climate Systems Research, Columbia University, New York, NY, USA.
18. School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK.
19. Birmingham Institute of Forest Research, University of Birmingham, Birmingham, UK.

Contents

1 Experiment simulation sampling in variable space

Simulation sampling across the defined variable space is not uniform tn the GGCMI Phase II experiment. Figure S1 compares sampling density in the models used in the emulator analysis.

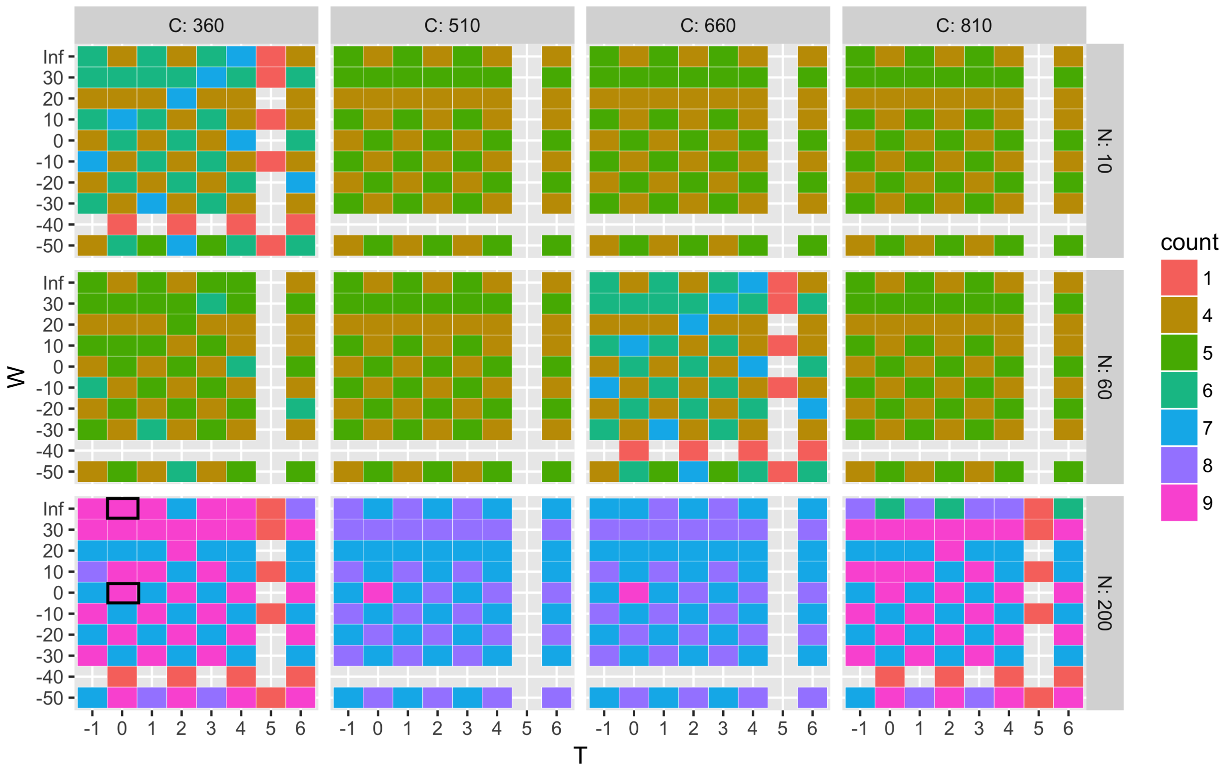

Figure S1: Heatmap illustrating the number of model simulations provided for each of the scenarios in CTWN variable space. The maximum number is 9, the number of models included in the emulator analysis. (That is, we exclude here the three GGCMI Phase II models not included in the emulator analysis.) Black boxes mark the "baseline" cases for rainfed and irrigated simulations. Normalized error calculations are run only over scenarios in which 9 models contribute simulations (pink boxes).

2 Cultivated areas

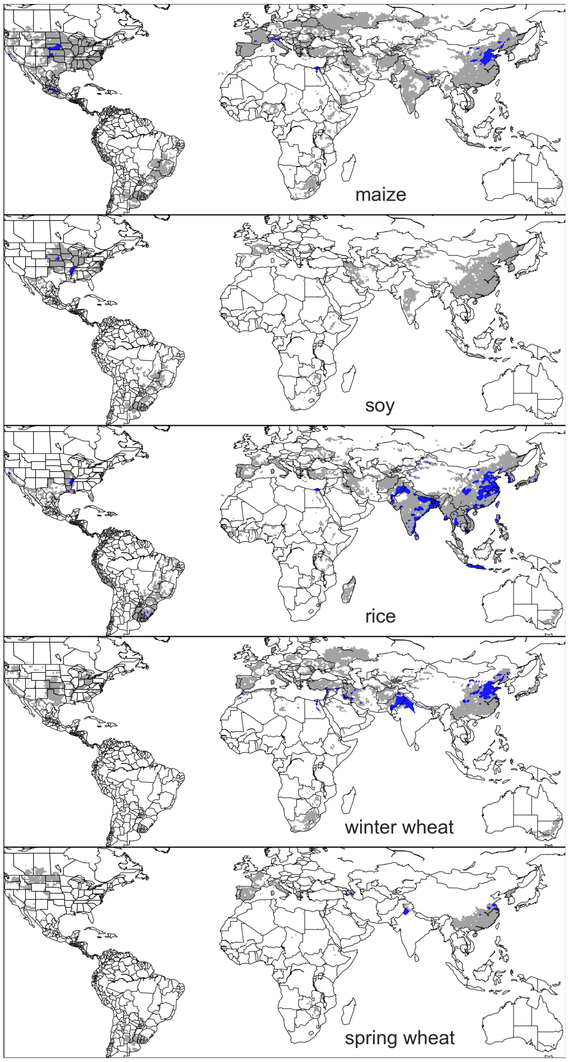

Figure S2: Presently cultivated area in the real world for irrigated (left) and rainfed (right) crops, from the MIRCA2000 dataset (**?**). Blue areas show grid-cells with more that 20,00 hectares of crop cultivated, and gray areas those with more that 10 hectares cultivated. Data are taken directly from the MIRCA2000 dataset for maize, rice, and soy. Winter and spring wheat areas are adapted from MIRCA2000 and sorted by growing season.

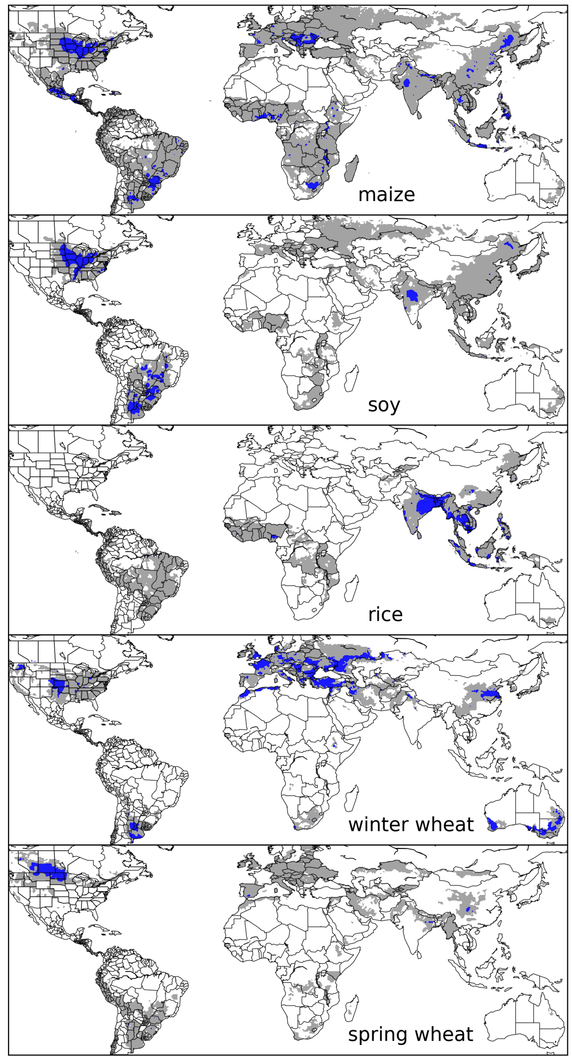

Figure S3: XX.

3 Variability changes in future climate projections

The GGCMI Phase II simulation dataset does not sample changes in climate variability. Large impacts to yields driven by changing variability would decrease the practical utility of the emulator for impacts assessments. Figure S3-S5 shows examples of changes in variability across representative climate model in the CMIP-5 archive. We use the HadGEM2 model for assessing the ability of GGCMI emulators to reproduce yield changes simulated under more realistic climate projections. Most crop models included in GGCMI phase II take daily minimum and maximum temperature (PROMET and JULES take sub-daily temperature) as inputs. Changes in variability are typically larger for maximum temperature than for minimum temperature in the climate models included here (some models show decreases in variability in minimum daily temperatures). Table S1 shows summaries for each crop and model weighted by production at the global scale.

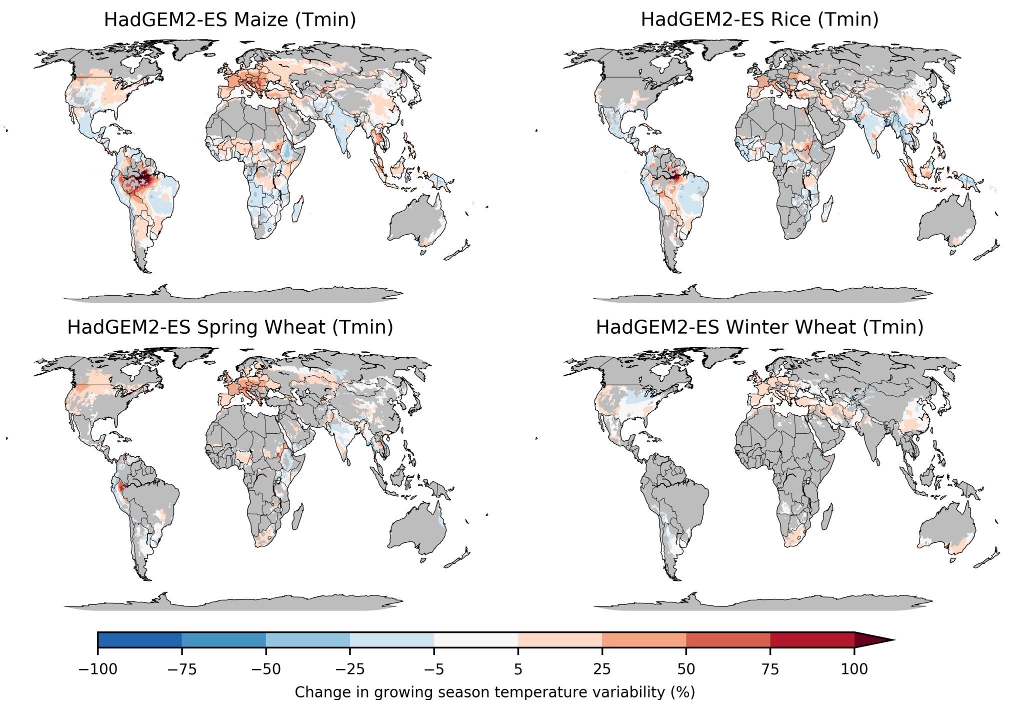

Figure S4: Change in growing season daily minimum temperature variability for the HadGEM2 model for selected crops under RCP 8.5 at the end of century. The heatmap shows the percentage change in standard deviation in daily minimum temperature relative to 1981-2010 using a 30-year mean of within-growing-season variability for 2070-2099. Blue indicates a decrease in variability and red an increase. Maps are masked to show only currently cultivated areas. Strong percentage increases in the tropics are typically due to very low variability in the baseline.

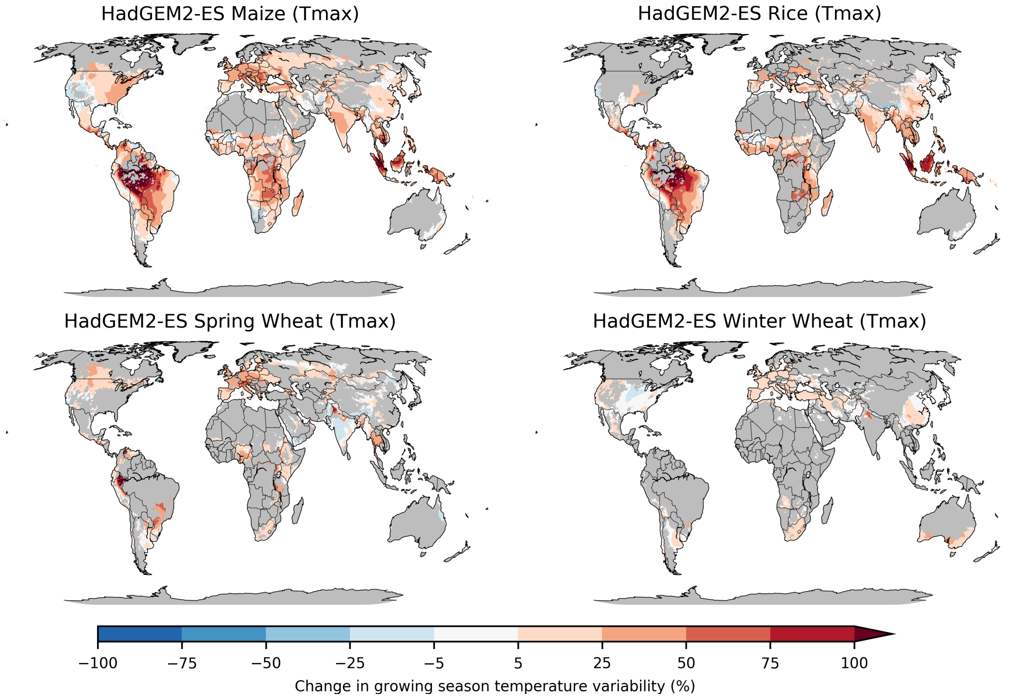

Figure S5: Convention same as S3 except now for daily maximum temperature. Changes in daily maximum temperature variability are generally higher than daily minimum temperature.

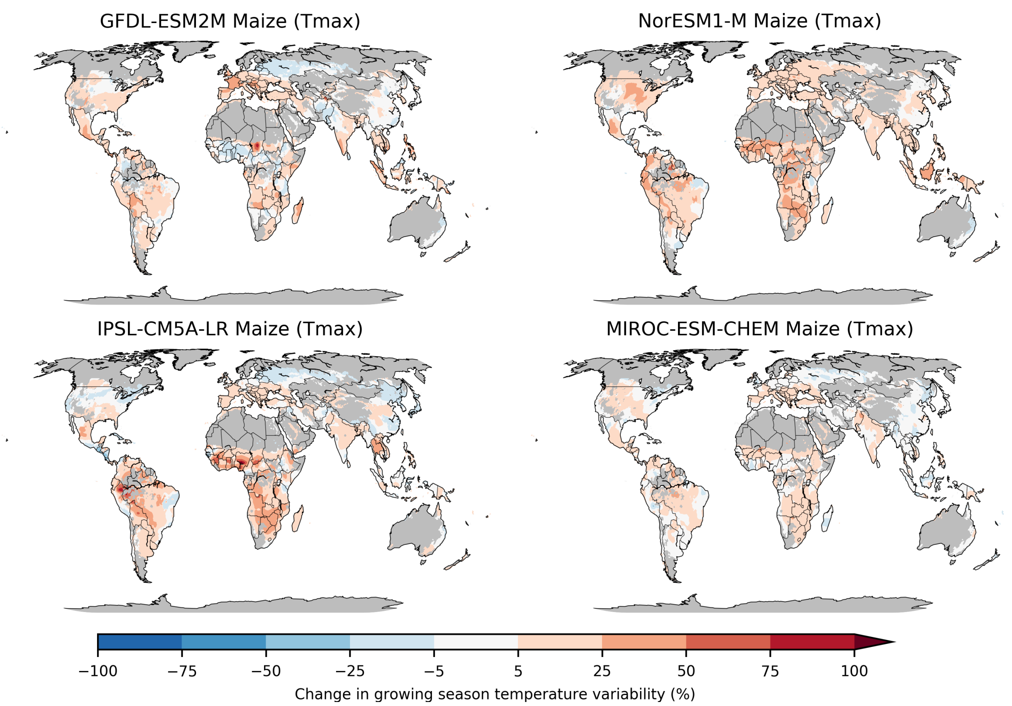

Figure S6: Convention same as S3 except now for daily maximum temperature for maize for other selected CMIP-5 models. HadGEM2 has the strongest change in variability for maize and the other crops.

Table 1: Global production-weighted change in growing season daily maximum temperature variability as a percentage of baseline variability for the five climate models included in the ISIMIP project. Mean within-growing season temperature standard deviation across 30 growing seasons for the baseline case (1981-2010) is compared to the end of century (2070-2099) under RCP8.5. LPJmL model simulated yields and current cultivation area (MIRCA) are used to weight values by production. Values in parenthesis are the change in variability by the same metric for daily minimum temperature within the growing season. The HadGEM2-ES model is highlighted in bold because we use this model for our emulator evaluation in the main text. We select the HadGEM2-ES model because it shows the highest changes in variability.

Model	**Maize %**	**Soybean %**	**Rice %**	**S. Wheat %**	**W. Wheat %**
HadGEM2-ES	**9.7 (2.1)**	**10.4 (-0.6)**	**10.1 (-3.3)**	**6.4 (4.7)**	**3.6 (1.7)**
GFDL-ESM2M	3.6 (0.9)	3.4 (0.6)	2.7 (-0.3)	1.2 (-0.3)	2.0 (1.0)
NorESM1-M	6.7 (-1.1)	6.5 (-4.0)	5.9 (-3.5)	4.5 (3.3)	2.4 (0.8)
IPSL-CM5A-LR	3.3 (3.3)	3.4 (0.6)	3.4 (1.4)	1.4 (2.3)	1.2 (1.3)
MIROC-ESM	4.0 (1.2)	3.1 (0.1)	0.4 (-5.1)	2.2 (-0.3)	2.6 (2.3)

4 Yield response for A1 (growing season adaptation) simulations

This section shows illustrations of emulator ability to capture yield changes in A1 simulations; compare to main text Figures 5 and 6 showing A0 simulations. Responses to CWN factors are similar in both but responses to T are substantially weaker in A1 simulations, in which growing season length does not contract in warmer future conditions.

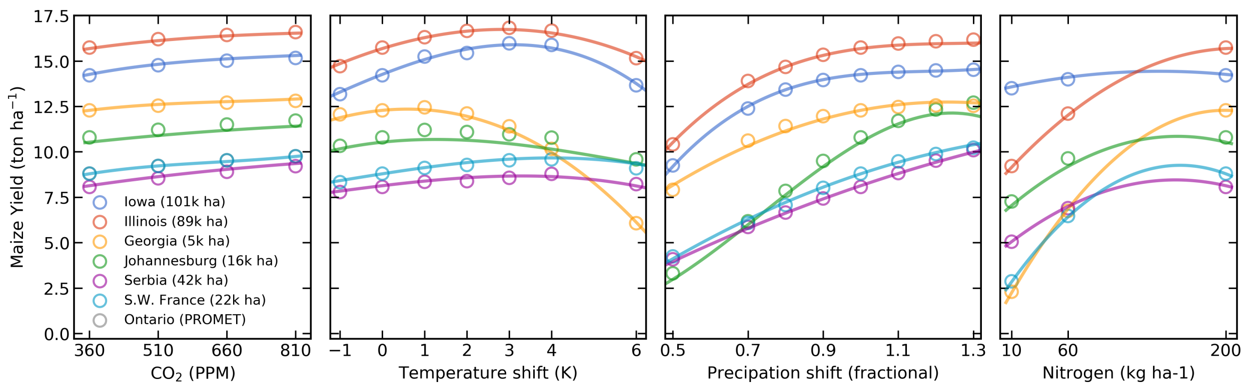

Figure S7: Illustration of spatial variations in yield response, which are successfully captured by the emulator for the A1 simulations. Panels show simulations (points) and emulations (lines) of rainfed maize in the pDSSAT model in six example locations selected to represent high-cultivation areas around the globe. Legend includes hectares cultivated in each selected grid cell. Each panel shows variation along a single variable, with others held at baseline values.

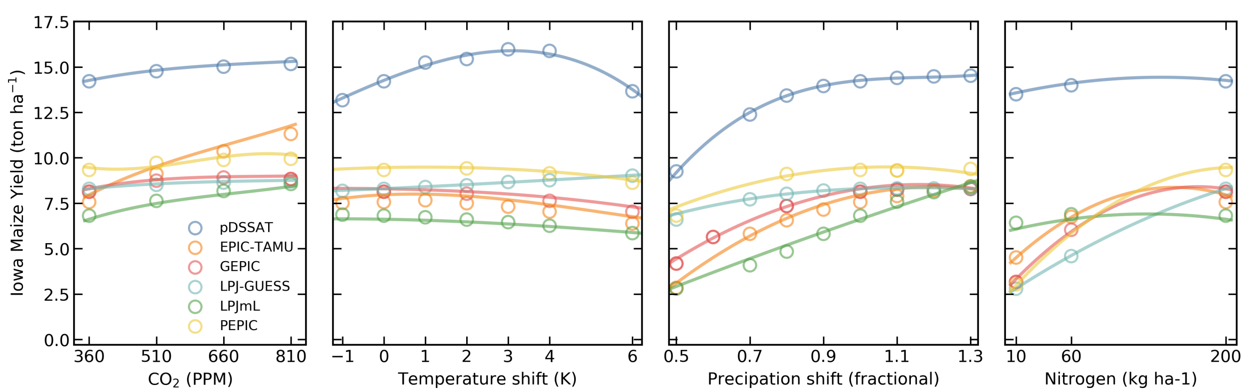

Figure S8: Illustration of variations in yield response across models for A1 simulations, again successfully captured by the emulator. Panels show simulations and emulations from six representative GGCMI models for rainfed maize in the same Iowa grid cell shown above, with the same plot conventions. Three models (PROMET, JULES, and CARAIB) that do not simulate the nitrogen dimension are omitted for clarity.

5 Normalized error for other cases

In manuscript Figure 7 we show normalized error for the A0 emulators over all rainfed crops, models, and T and W values for baseline CO_2 and nitrogen levels (360 ppm and 200 kg ha-1). Here we show normalized error in some alternate cases for comparison: *Figure S6:* A0 emulators of rainfed crops at higher CO_2, *Figure S7:* A1 emulators of rainfed crops at baseline values, *Figure S8):* A0 emulators of irrigated crops at baseline values. Results are generally similar, with a few exceptions. Normalized errors at higher CO_2 are generally lower because model disagreement is larger, lowering the denominator. Some model emulators for irrigation water demand are under-performing: LPJ-GUESS and CARAIB for some crops. A1 errors are larger than A0 errors for several crops and models: LPJmL rice, pDSSAT spring wheat, and PROMET winter wheat.

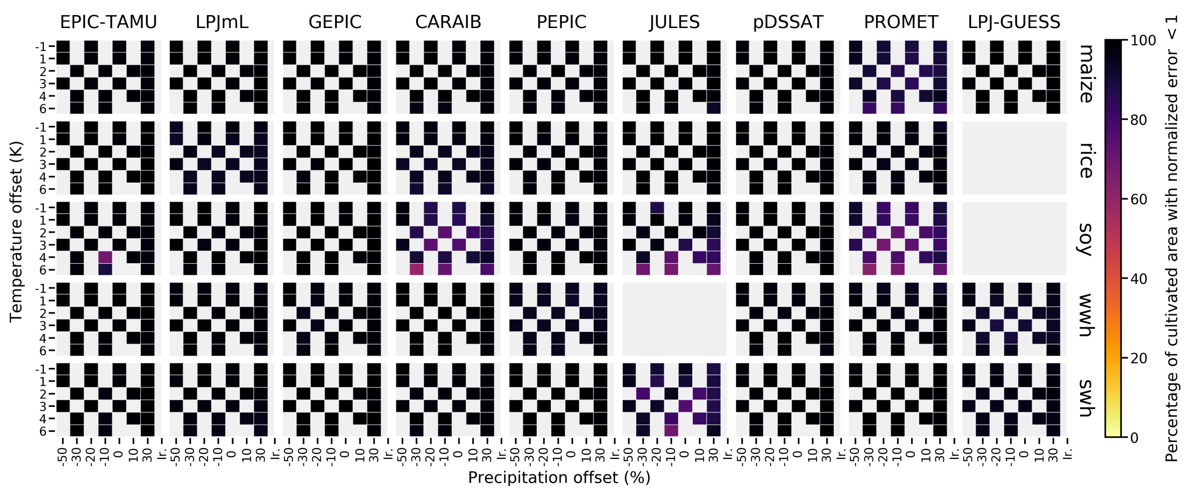

Figure S9: Fraction of currently cultivated hectares with normalized emulation error less than 1 for the CO_2=810 ppm and 200 kg N ha^{-1} yr^{-1} case for the temperature and precipitation perturbations scenarios provided by all 9 models included in the emulator analysis. Figure convention as in main text Figure 7. The yield response is generally easy to emulate over currently cultivated areas (black regions).

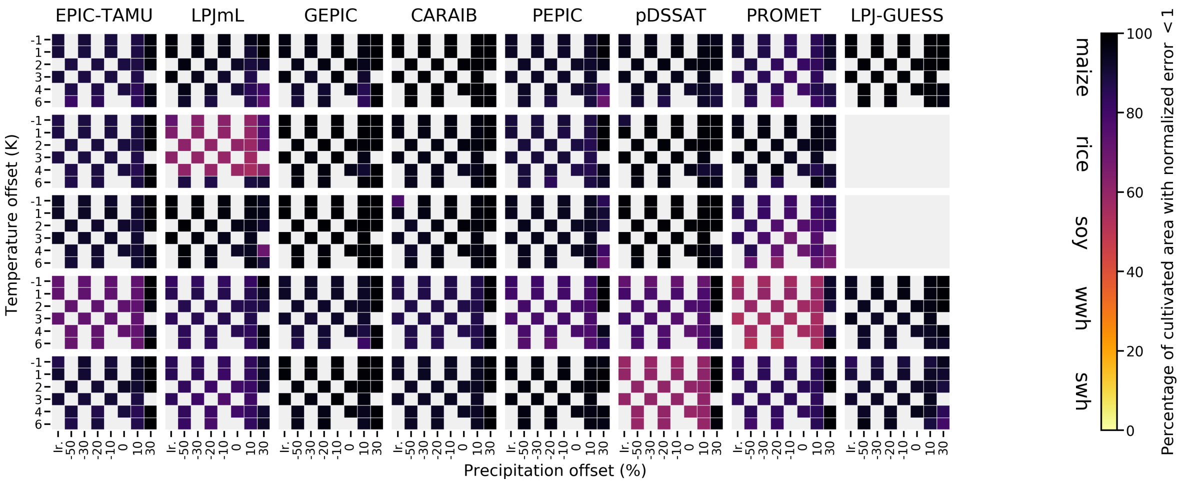

Figure S10: Fraction of currently cultivated hectares with normalized emulation error less than 1 for A1 yield emulation for CO_2=310 ppm and 200 kg N ha^{-1} yr^{-1} case. Figure convention as in main text Figure 7.

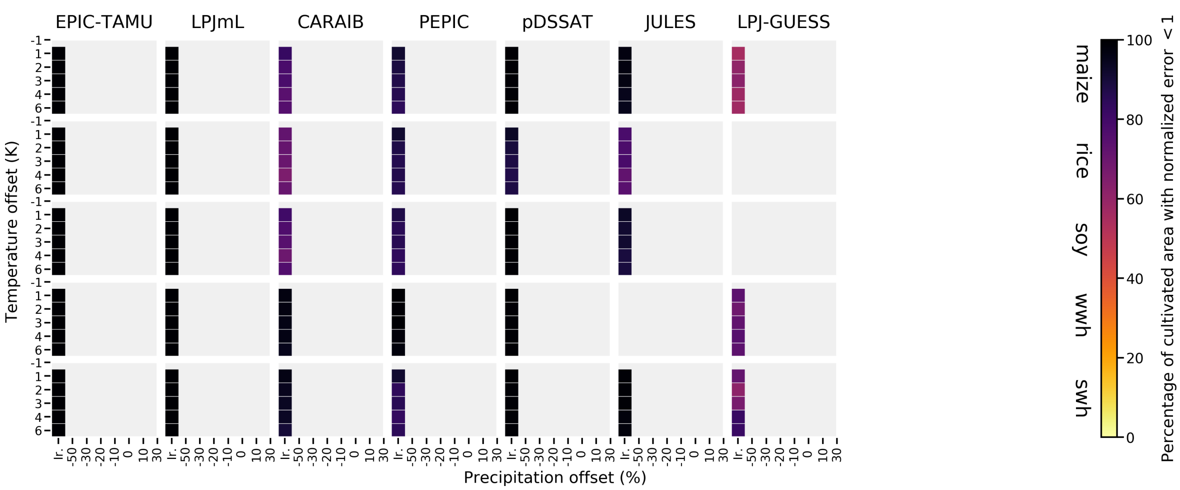

Figure S11: Fraction of currently cultivated hectares with normalized emulation error less than 1 for irrigated water demand emulation for CO_2=310 ppm and 200 kg N ha^{-1} yr^{-1} case. Figure convention as in main text Figure 7.

6 Comparison to transient climate run simulation at high latitude

In manuscript Section 4.3 we test the emulator against crop model simulations driven by a more realistic future climate projection to evaluate the impact of future variability changes that are not captured by the emulator. Figure S12 below isolates the mid- and high latitudes; compare to manuscript Figure 9 that shows global currently cultivated land. Results are generally unchanged by the restriction in latitude except for rice, which is typically grown in tropics and subtropics: 20% of global rice production is grown north of 30N and 1% north of 45N, with even less in the Southern hemisphere, only 0.8% south of 30S and none south of 45S.

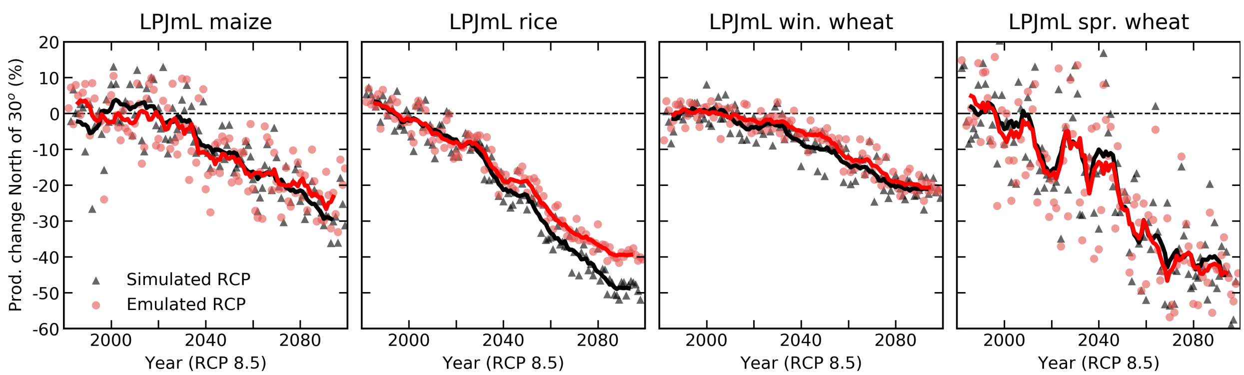

Figure S12: Illustration of the ability of the emulator to capture a more realistic future climate simulation, as in main text Figure 9 but here restricted to latitudes north of 30N.

7 Emulator products

This section amplifies on manuscript Section 5 with additional figures analogout to manuscript Figures 10 and 11.

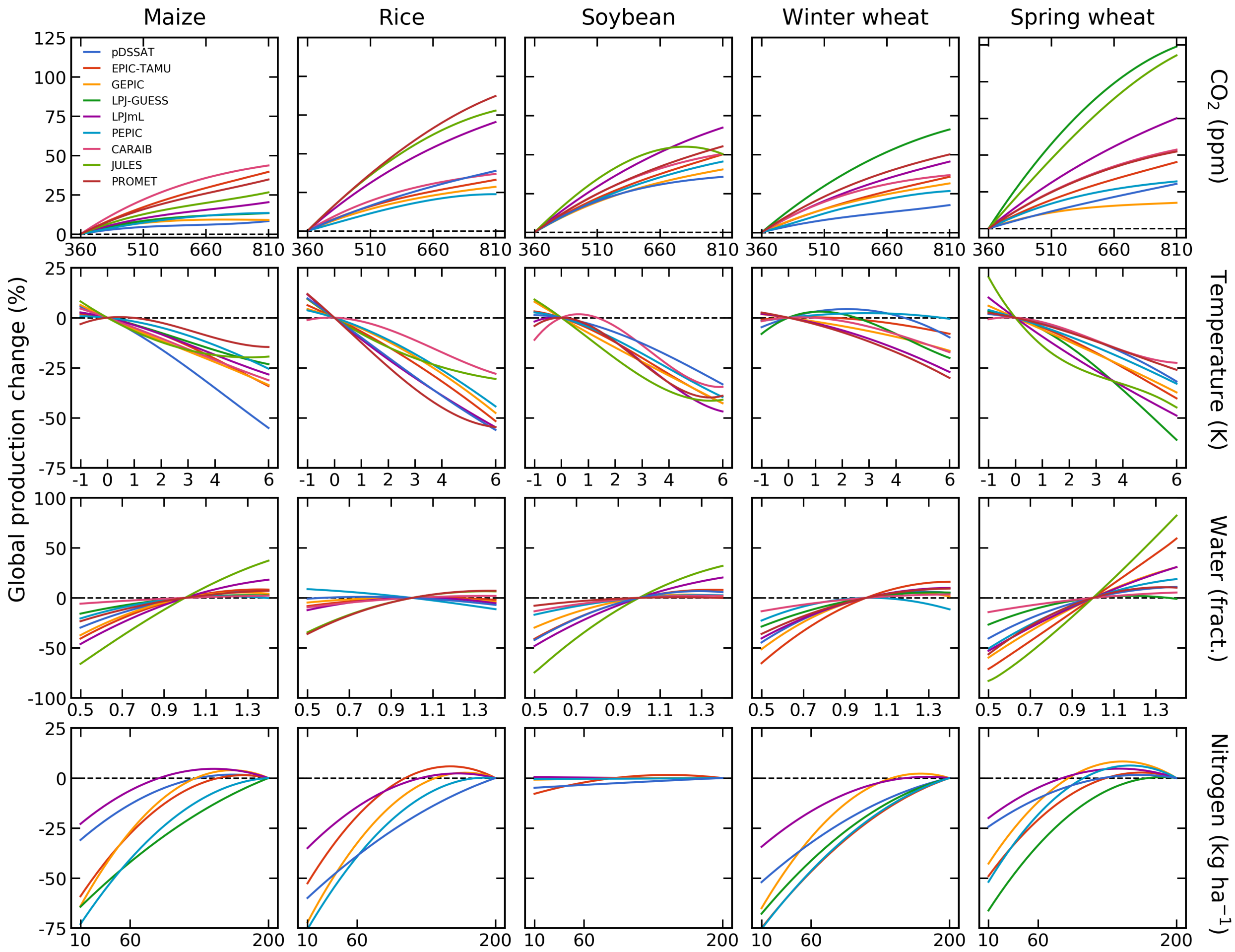

Figure S13: As in main text Figure 10 except now each model is shown in color. Not all models simulate every crop (JULES does not simulate winter wheat, and LPJ-GUESS does not simulate rice or soy) and three do not simulate the N response (CARAIB, JULES, and PROMET). For crops simulated, JULES and LPJ-GUES are often the outliers in strong CO_2 responses. LPJmL has the weakest N response for all crops, and pDSSAT has the strongest T response for maize.

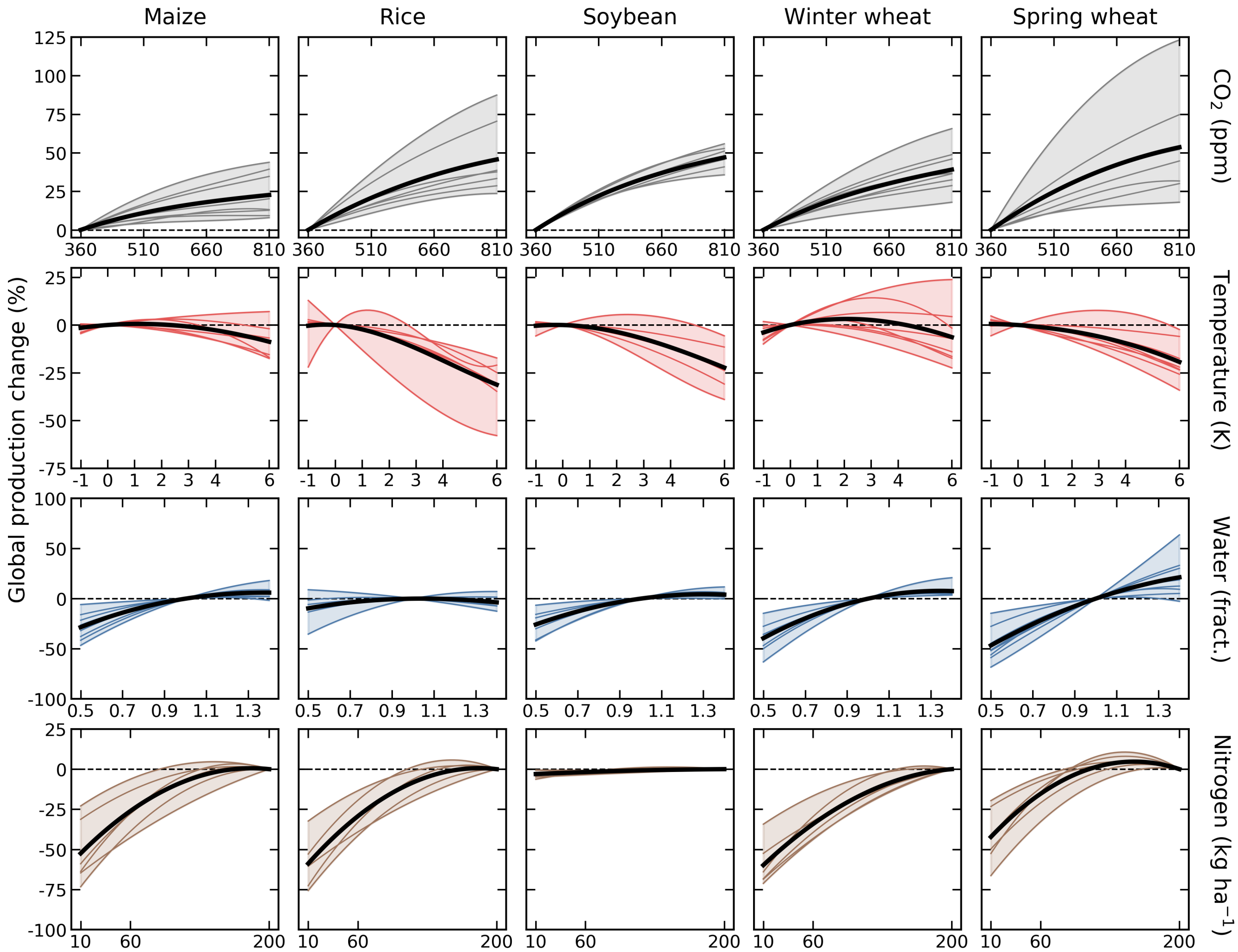

Figure S14: Emulated damage functions for rainfed crops for A1 (growing season adaptation) simulations, with conventions as in main text Figure 10 showing A0. Temperature responses are generally flatter than for A0 simulations, but responses to other factors are similar. Note that JULES does not provide A1 simulations.

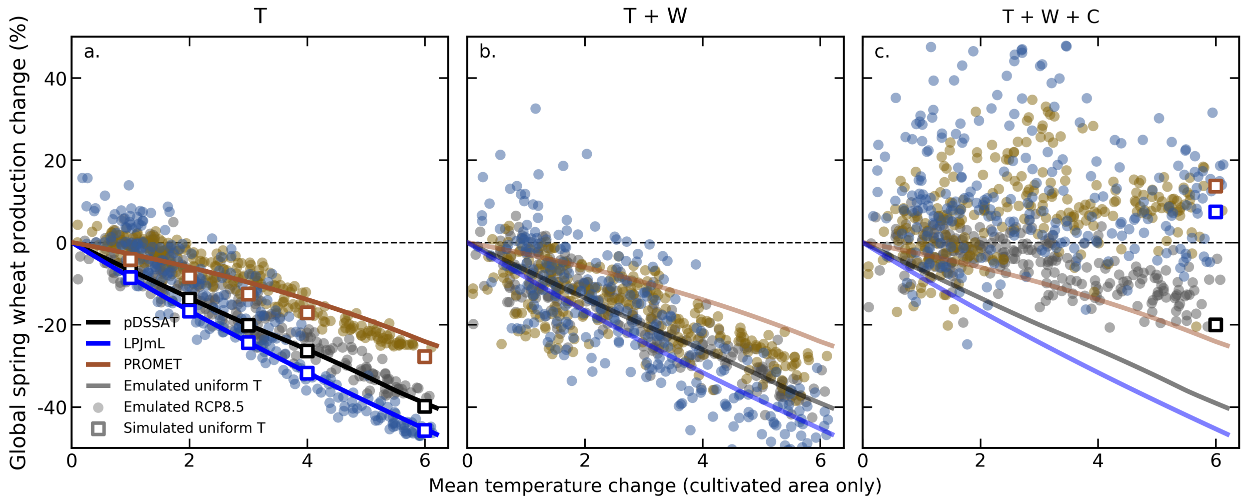

Figure S15: Illustration of the factors affecting yields in more realistic climate scenarios for rainfed and irrigated (current mix) spring wheat. Conventions as in main text Figure 11. Large emulator errors in PROMET spring wheat temperature response (panel **a**, compare open squares to line) are driven by Southern China, where discontinuities in yield responses make emulation problematic. (See Supplemental Material Section S11).

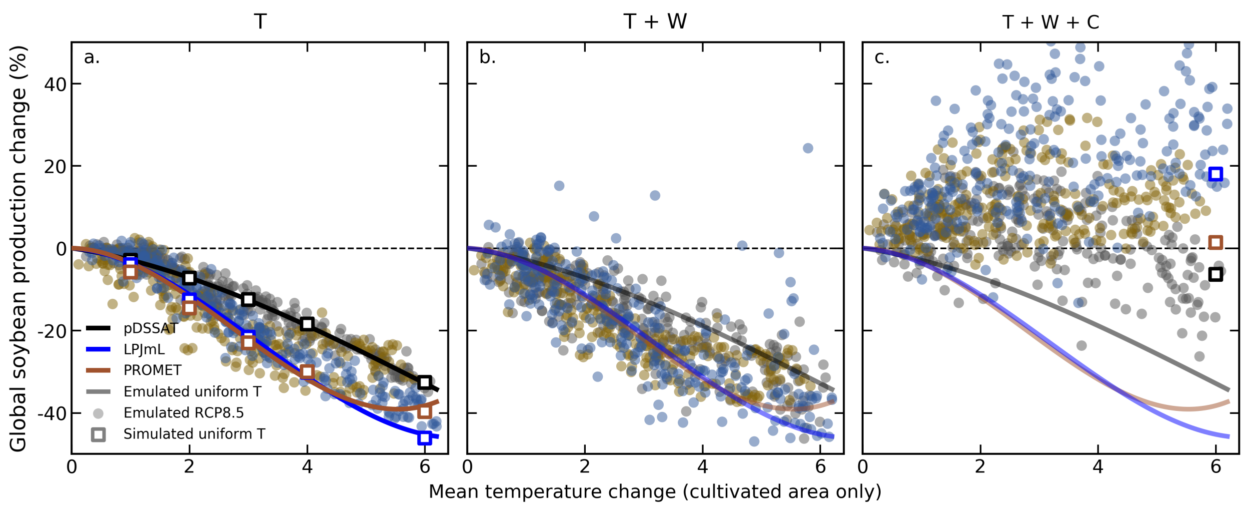

Figure S16: Illustration of the factors affecting yields in more realistic climate scenarios for rainfed and irrigated (current mix) soy. Conventions as in main text Figure 11. The split in PROMET soybean temperature response (panel **a**, note distinct groups of points) results from the model's sensitivity to differences in spatial patterns of temperature change across climate models.

8 Reduced specification (23-term) emulator examples

In this section we present analogous figures to those in the main text for the reduced-form (23-term) emulator. Issues with the reduced-form model are most prominent in PROMET for rice and soy, and JULES soy and spring wheat. We identify several potential factors that may in some way contribute to these models showing qualitatively different responses that require additional terms for emulation.

- PROMET and JULES do not allow nitrogen variation. (However, CARAIB also cannot vary N and is readily emulatable with the 23-term specification.)
- Both JULES and PROMET models are land system process models, originally developed with a broad focus, which have been adapted for managed vegetation (agriculture) only recently (2015). (CARAIB, by contrast, was originally developed as a vegetation model in the early 90's and has a longer history of agricultural focus.)
- Both PROMET and JULES have anomalously strong responses to individual factors in those crops problematic to emulate. PROMET is the most sensitive model of all the models for rice in C, T, and W, and JULES for soybeans in C, T, and W. For spring wheat, JULES is a high outlier in C, the most sensitive model in W and T, and shows an extra inflection point in the global temperature response not seen in any of the other models.
- PROMET is the quantitatively lowest-performing model for soybeans when compared to the historical FAO data for the top 10 producing countries.

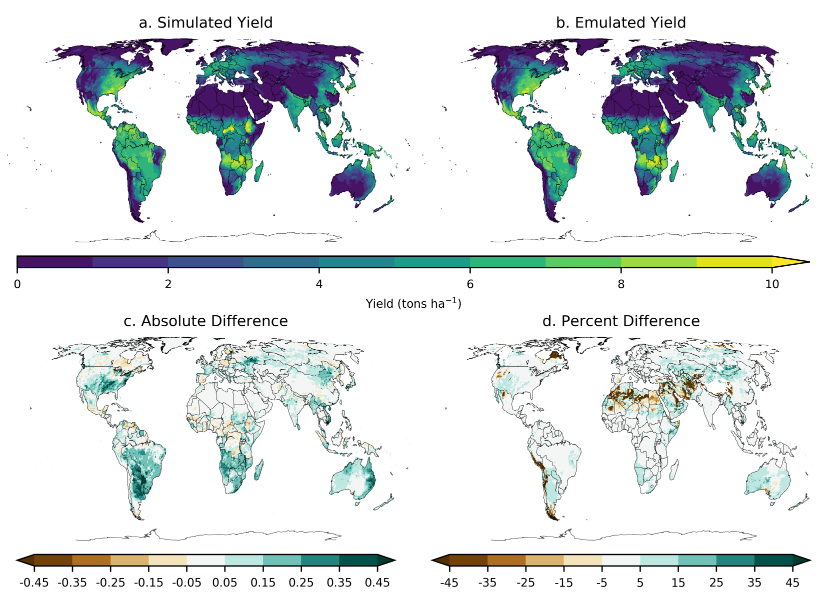

Figure S17: As in manuscript Figure 4, simulated (a.) and emulated (b.) yield under historical conditions for rainfed LPJmL maize, but here for the reduced (23-term) emulator specification. Emulator performance is worse primarily where crops are not currently grown.

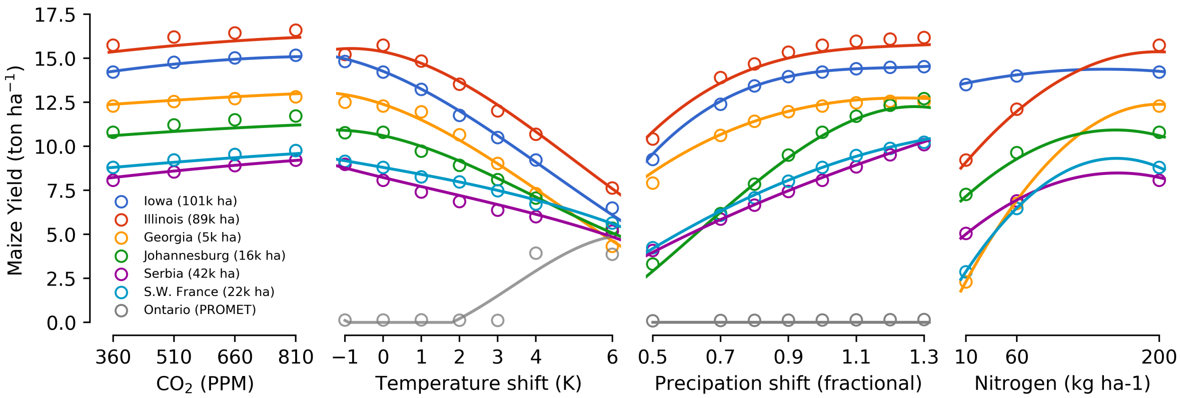

Figure S18: As in manuscript Figure 5, emulator performance in selected high-yield regions for rainfed pDSSAT maize (and one region for PROMET), but now with the reduced (23-term) emulator specification. Emulator performance is similar.

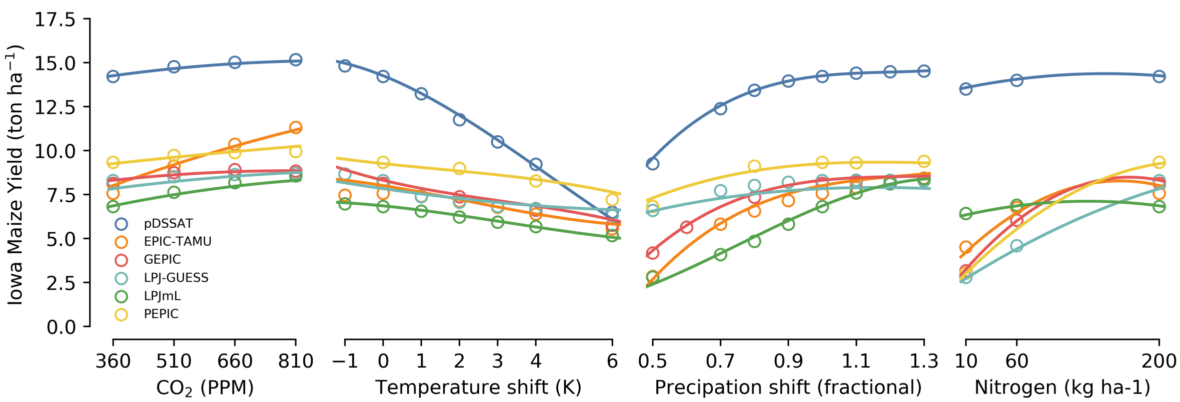

Figure S19: As in manuscript Figure 6, emulator performance across models for rainfed maize in one grid cell in Iowa, but now with the 23-term emulator specification. Note that JULES and PROMET are not shown.

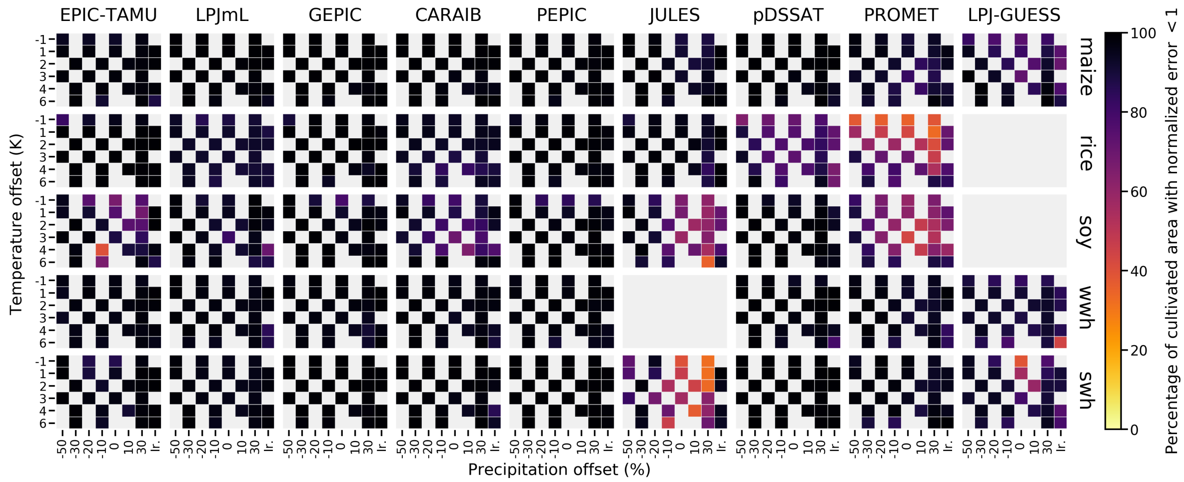

Figure S20: As in manuscript Figure 7, normalized error of all 9 models emulated on currently cultivated land, over all crops and all sampled T and W inputs, with CO_2 and nitrogen held fixed at baseline values, now with the reduced (23-term) emulator specification. Degradation of performance is most evident in JULES soy and spring wheat and PROMET rice and soy.

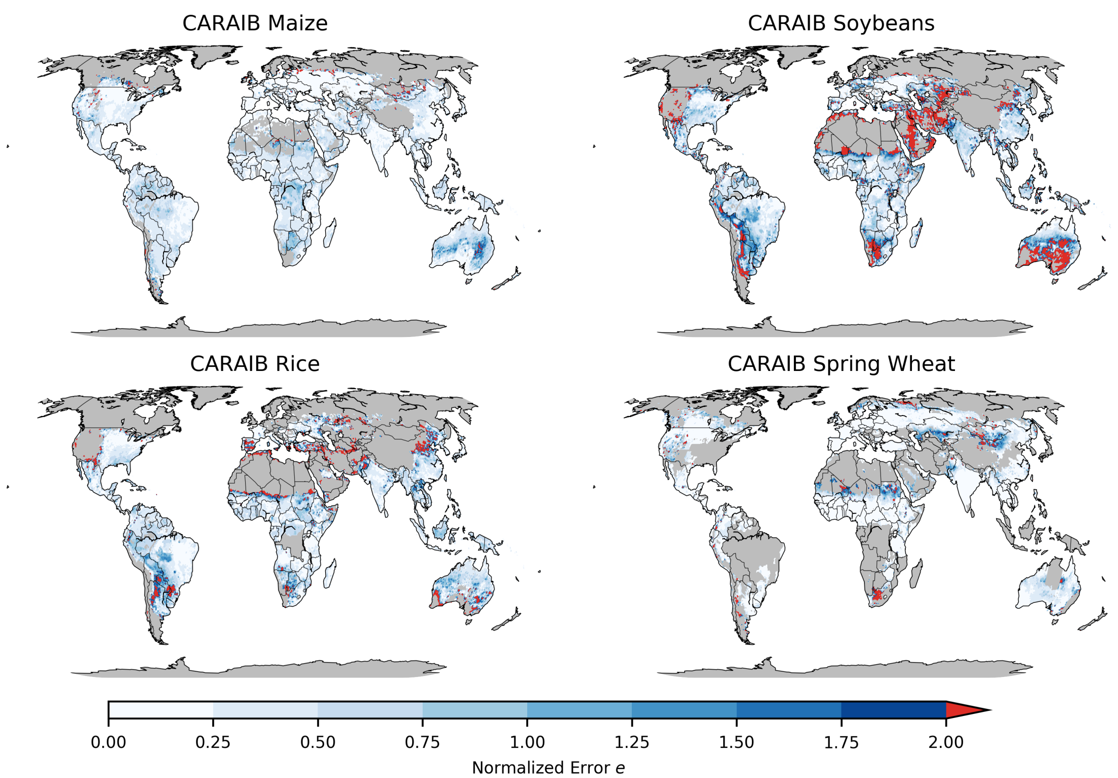

Figure S21: As in manuscript Figure 8, normalized error for rainfed crops in CARAIB for the T+4 scenario, but here with the reduced (23-term) emulator specification. Degradation of performance is most evident in marginal lands where crops are not currently grown.

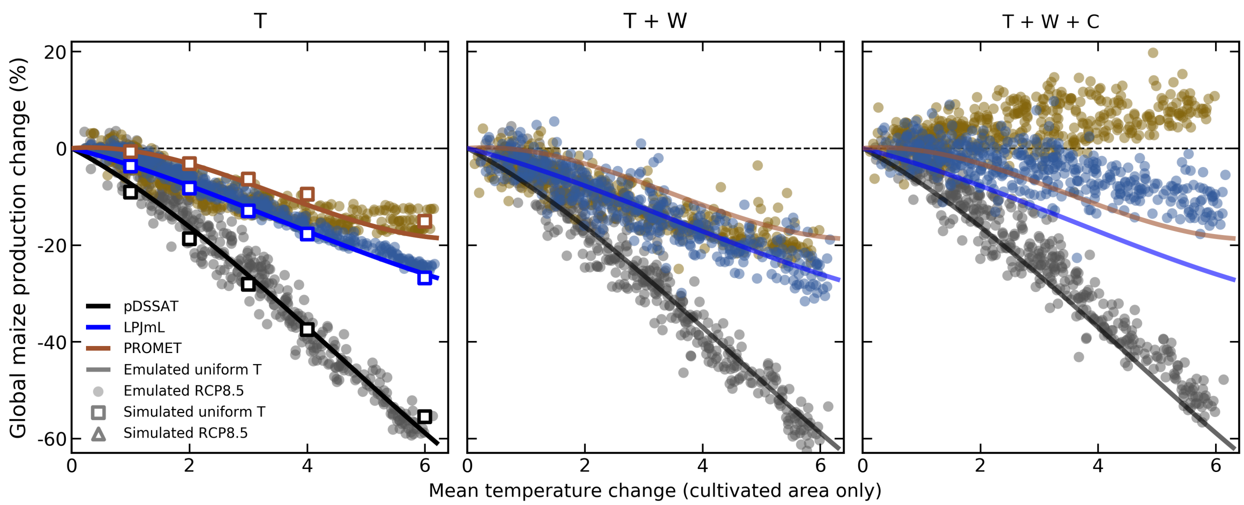

Figure S22: As in manuscript Figure 11, rainfed maize on currently cultivated land, but here with the reduced (23-term) emulator specification. Note that strong C response for PROMET is different here than with the full-form emulator, because higher order C (C^3, $C^2 * T$ ...) interaction terms are needed for accurate emulation.

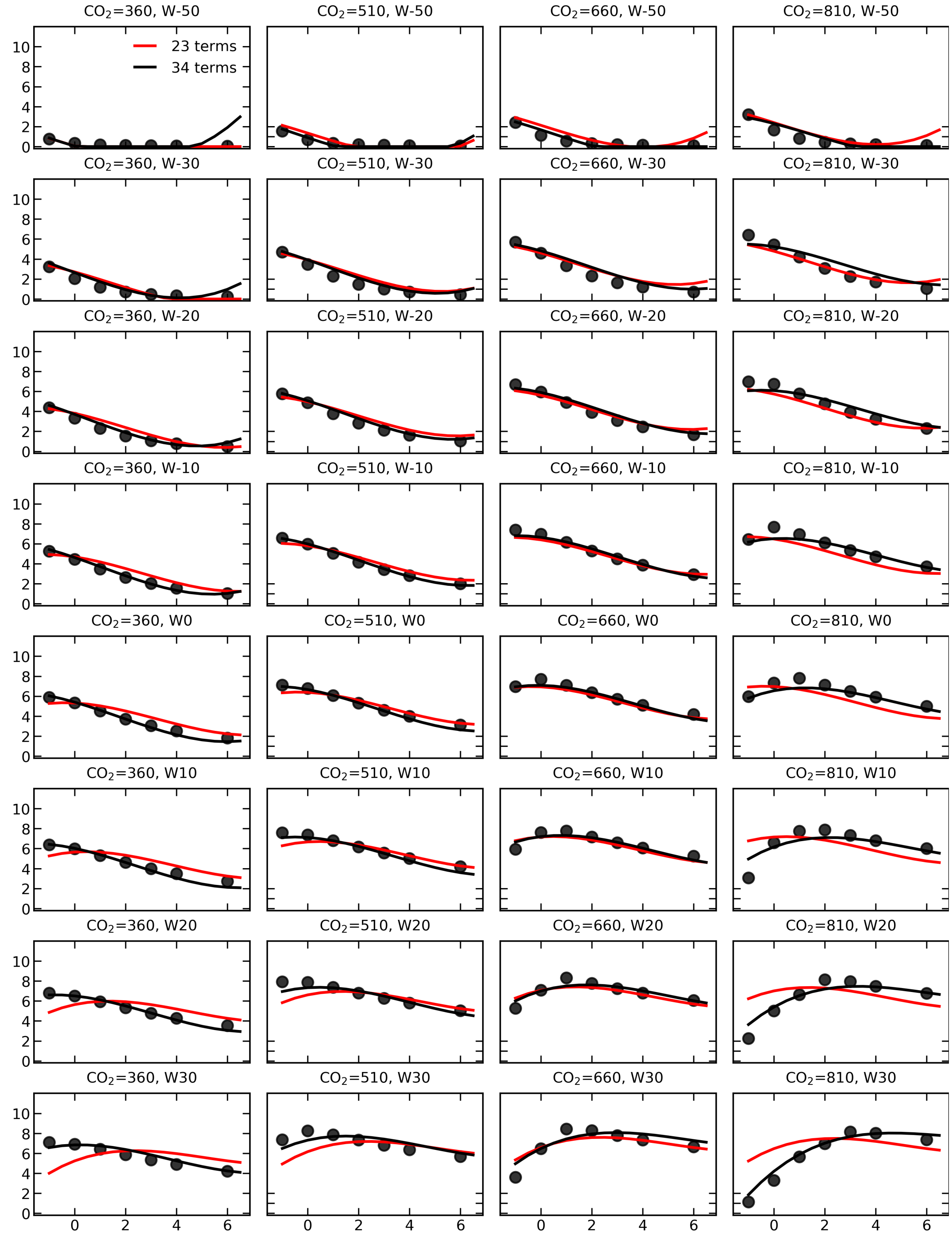

Figure S23: Example of emulator failure, showing failure induced becuase of strongly interacting terms. Simulated and emulated values for JULES soybeans in Southern Germany. RMSE = 41% of baseline yield for the reduced form (23-term) emulator. The downturn in yields as C and W increase can only be captured by the higher order C interaction terms of the full 34-term specification.

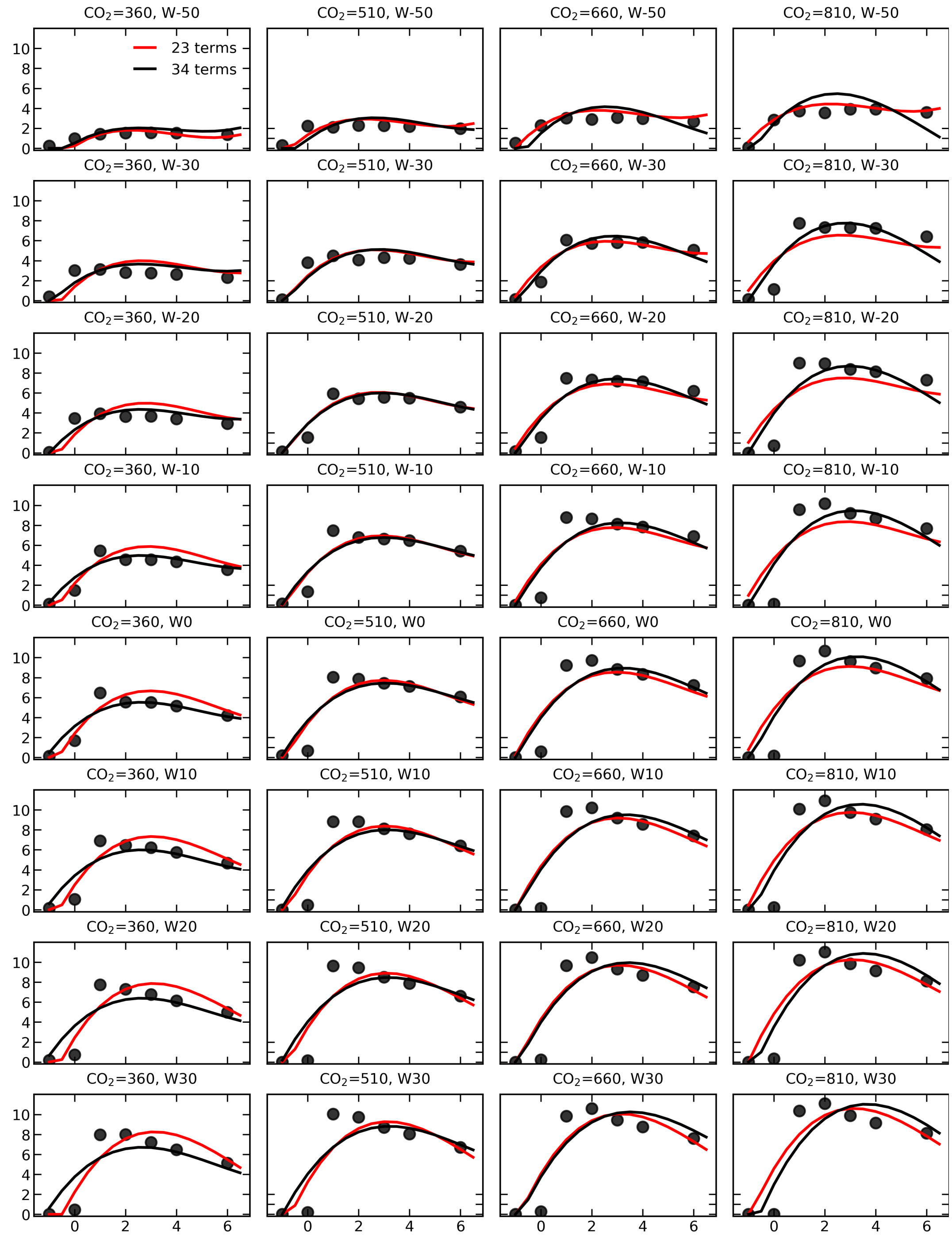

Figure S24: Example of emulator failure, showing failure induced by abrupt changes in yields. Simulated and emulated values for PROMET rice in India (Arunachal Pradesh). RMSE = 132% of baseline yield for red (reduced fit). The step change in the yields around 0 K at higher water specifications cannot be captured by any third order polynomial. Both 23- and 34- emulator specifications fail in this example.

9 Yield Responses for other crops and models

Spatial patterns of yields are well captured for all crops and models. Manuscript Figure 4 illustrated this using LPJmL maize; for reference, we show here yield response spatial patterns for other crops and models.

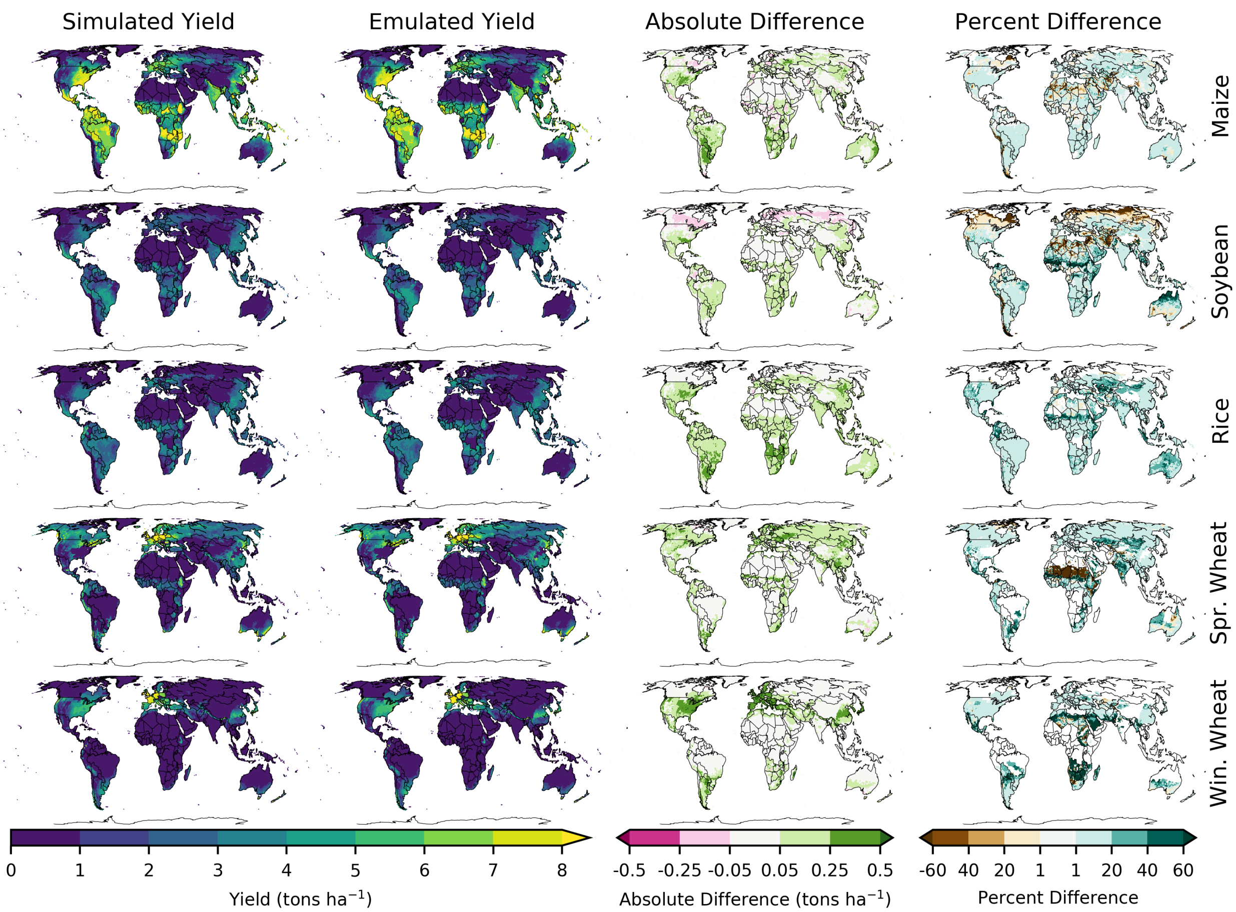

Figure S25: Spatial yield response and emulator error for LPJmL for all 5 GGCMI Phase II crops. Convention as in manuscript Figure 4.

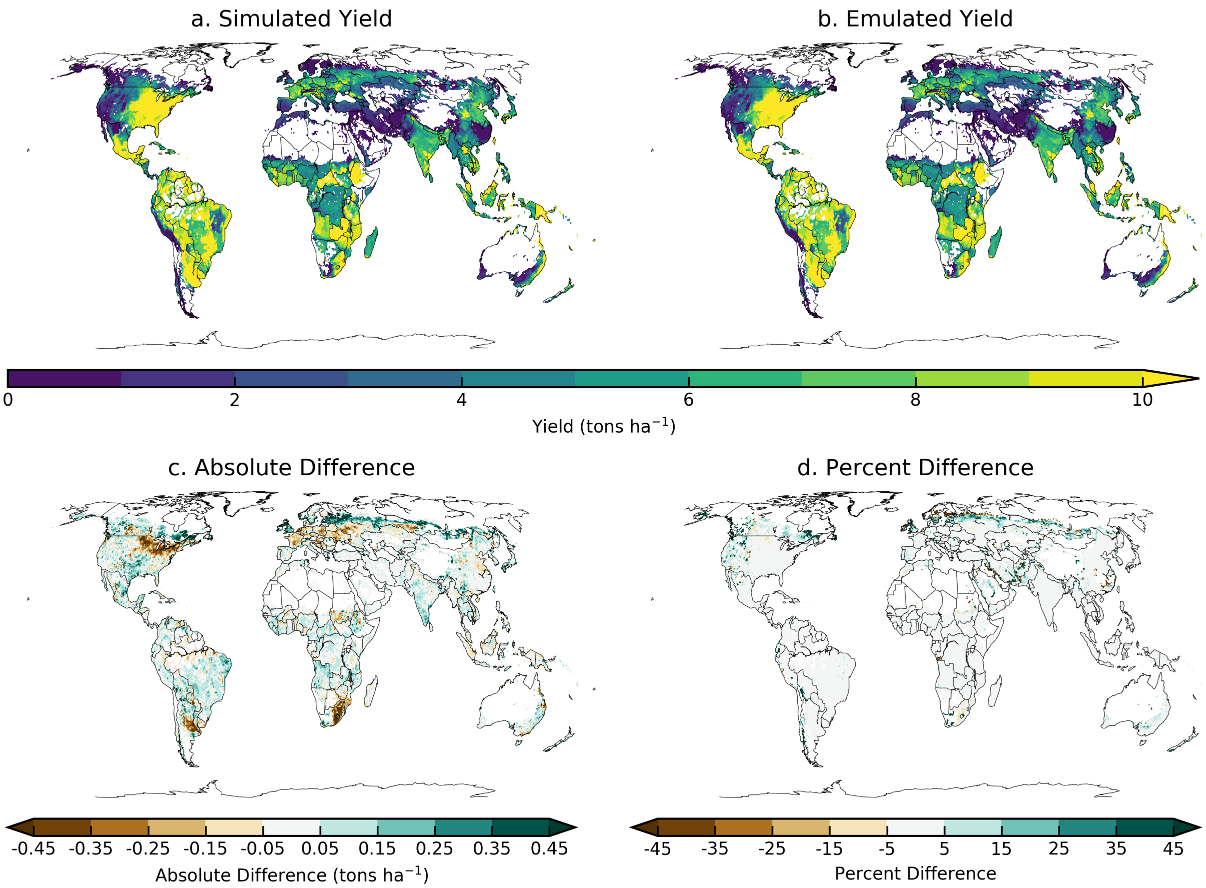

Figure S26: Spatial yield response and emulator error for pDSSAT for maize. Convention as in manuscript Figure 4. pDSSAT absolute yields are significantly higher than those in LPJmL but spatial patterns are similar.

10 Cross validation error for all models

In this section we present maps of cross validation error (values found in main text Table 3 are aggregated up from the grid cell level). Errors are generally low as a percentage of yield change in each grid cell. Errors can be above 10% of yield change for the out-of-sample test in some regions for some crops, most noatably spring wheat in southern China in the PROMET model.

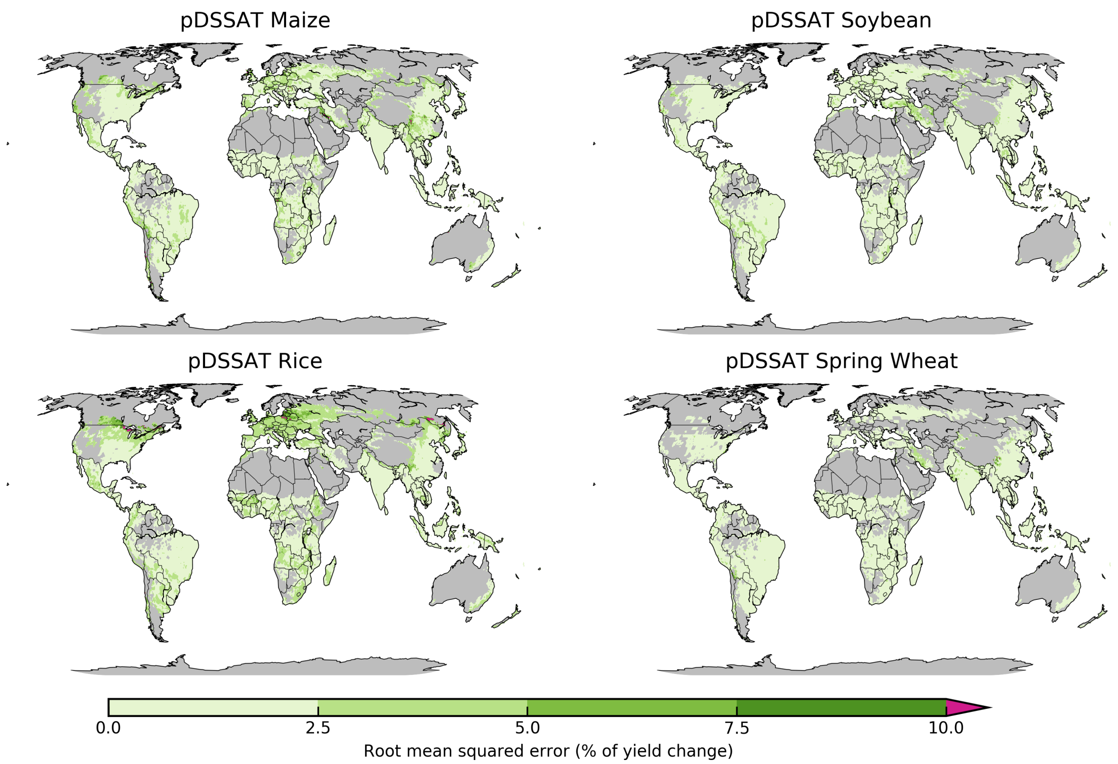

Figure S27: Root mean squared error for three-fold cross validation for the pDSSAT model for rainfed crops. Values shown as a percentage of yield change in each grid cell.

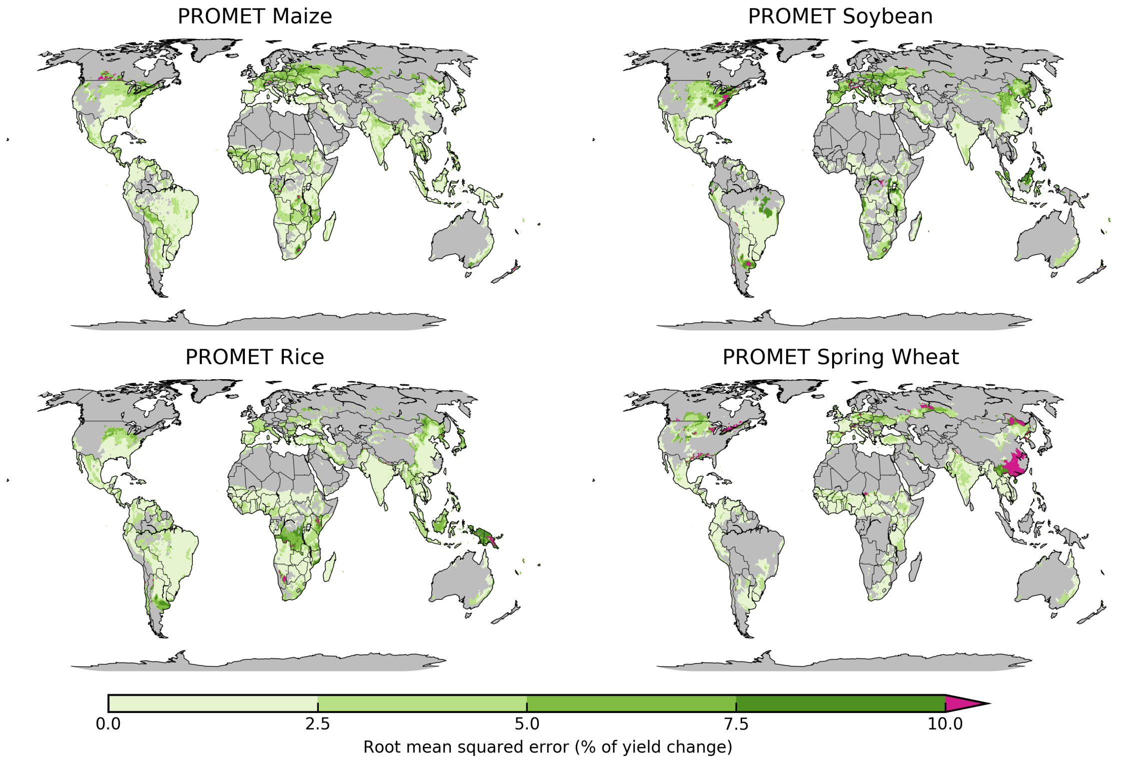

Figure S28: Map of root mean squared error for three fold cross validation process for the PROMET model for rainfed crops. Values shown as a percentage of yield change in each grid cell.

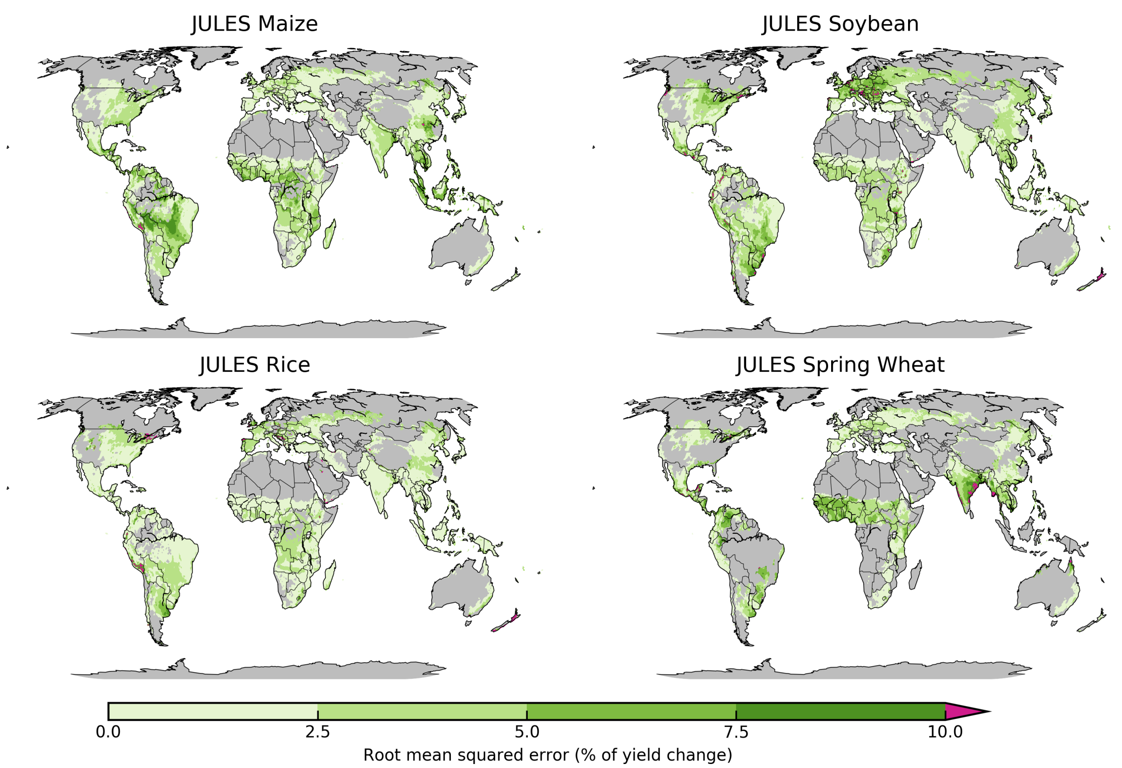

Figure S29: Map of root mean squared error for three fold cross validation process for the JULES model for rainfed crops. Values shown as a percentage of yield change in each grid cell.

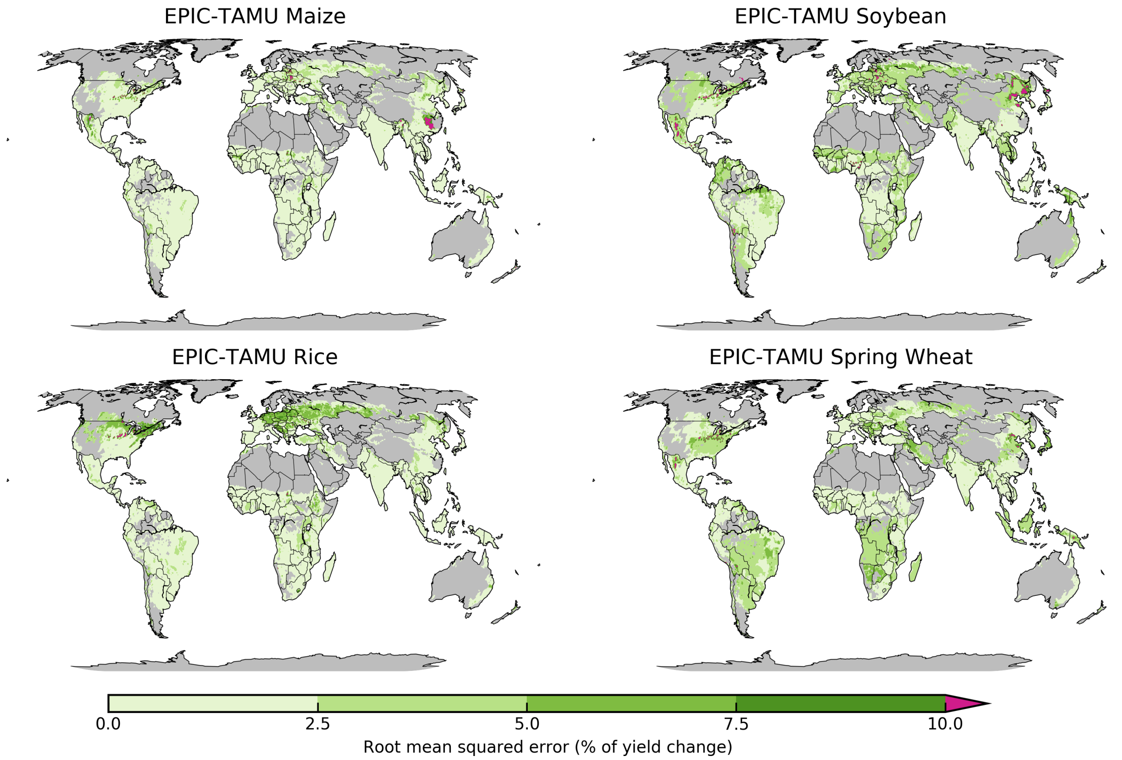

Figure S30: Map of root mean squared error for three fold cross validation process for the EPIC-TAMU model for rainfed crops. Values shown as a percentage of yield change in each grid cell.

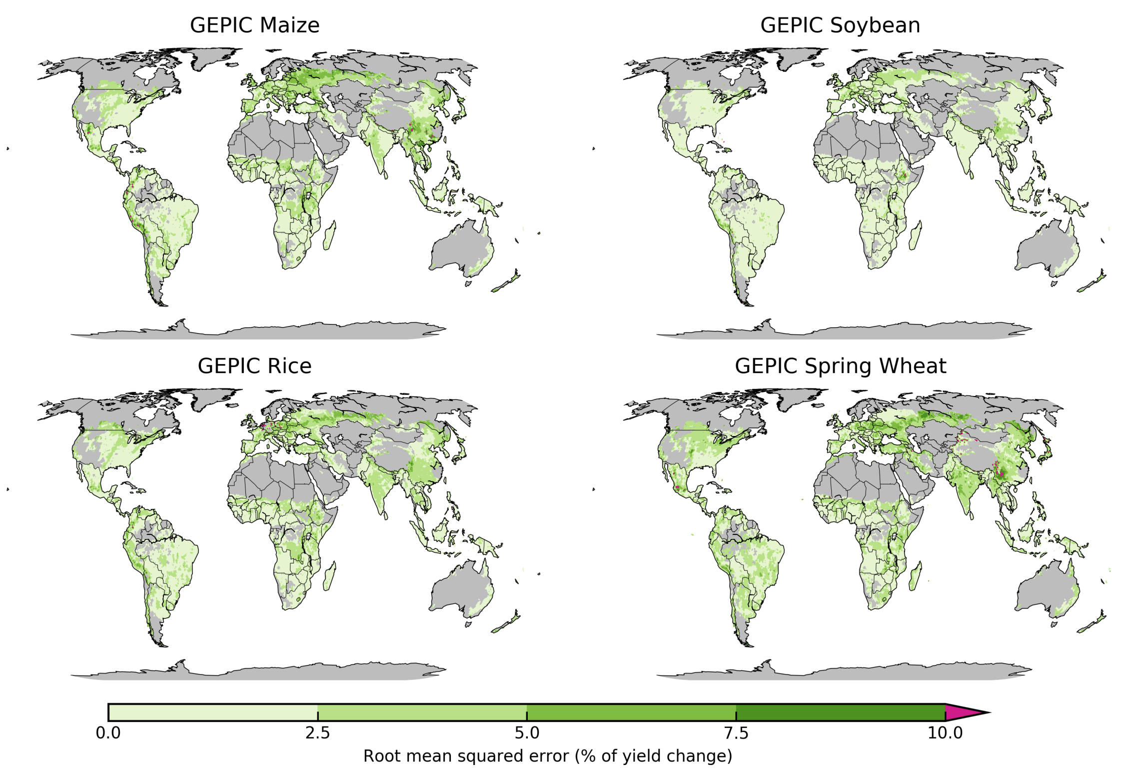

Figure S31: Map of root mean squared error for three fold cross validation process for the GEPIC model for rainfed crops. Values shown as a percentage of yield change in each grid cell.

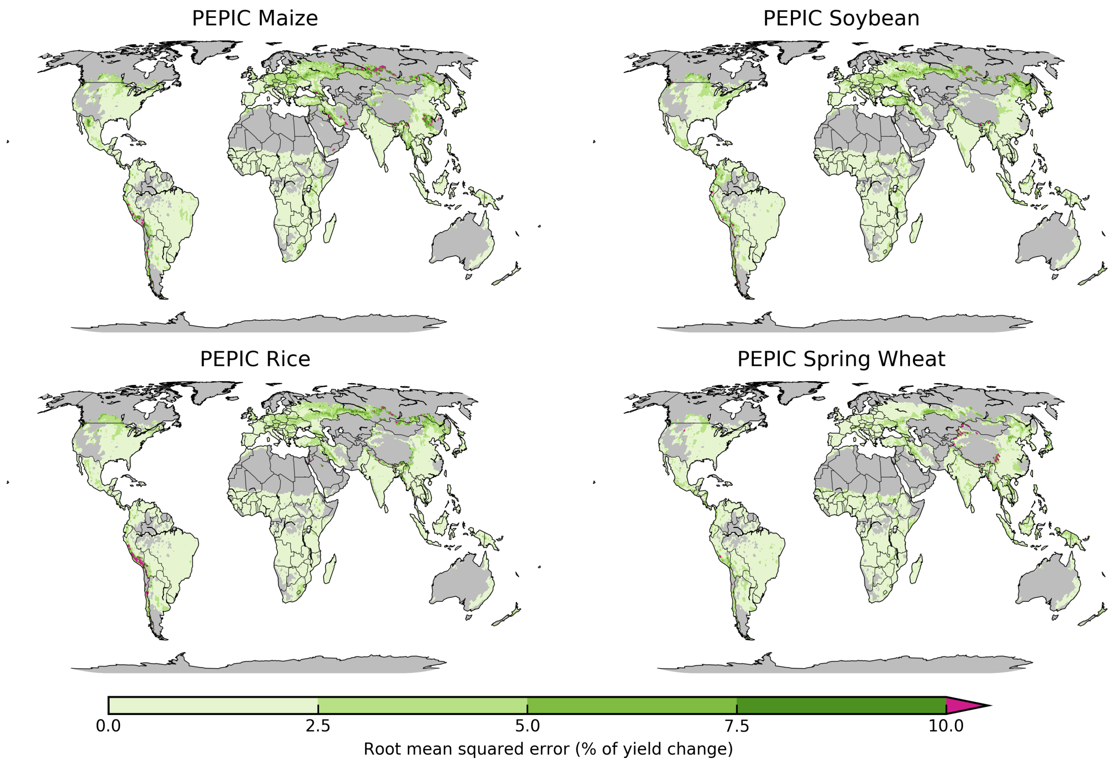

Figure S32: Map of root mean squared error for three fold cross validation process for the PEPIC model for rainfed crops. Values shown as a percentage of yield change in each grid cell.