Supplementary Information to ”Global gridded crop models underestimate yield losses from climatic extremes

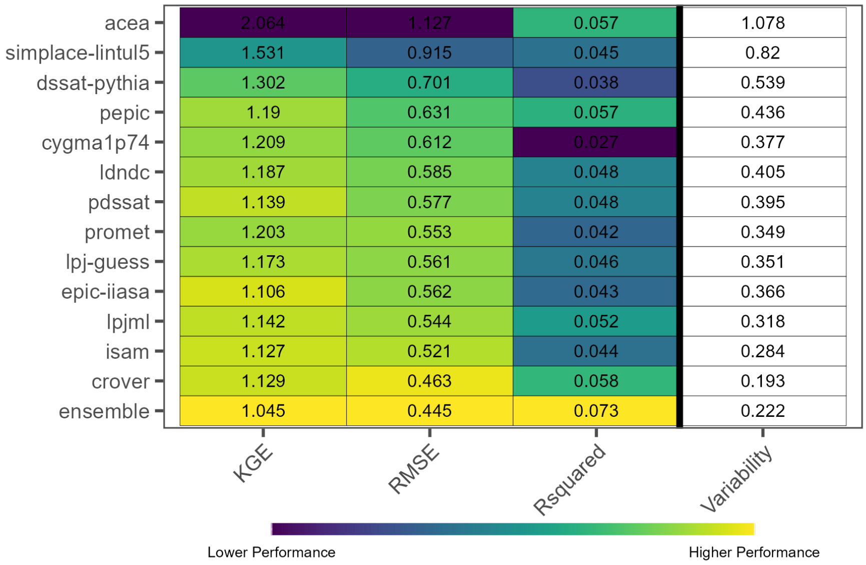
Cornelia Auer[1,20], Kobe De Maeyer[2,20], Christoph Müller[1], Michael Hoehle[3], Jacob Schewe[1], Jonas Jaegermeyr[4,5,1], Juraj Balkovic[6], Thiago Berton Ferreira[7], Babacar Faye[8], Christian Folberth[6], Jose R. Guarin[9], Stefanie Heinicke[1], Gerrit Hoogenboom[7], Toshichika Iizumi[10], Atul K. Jain[11], Tzu-Shun Lin[12], Wenfeng Liu[13,14], Oleksandr Mialyk[15], Masashi Okada[16], Sam S. Rabin[12], Chenzhi Wang[17], Heidi Webber[17,18], Florian Zabel[19]

[1] *Potsdam Institute for Climate Impact Research, Potsdam, Germany*
[2] *Utrecht University, Netherlands*
[3] *University of Greifswald, Germany*
[4] *Columbia University, Climate School, New York, NY 10025, USA*
[5] *NASA Goddard Institute for Space Studies, New York, NY 10025, USA*
[6] *Biodiversity and Natural Resources Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria*
[7] *University of Florida, USA*
[8] *Universite du Sine Saloum El Hadj Ibrahima NIASS (USSEIN), Kaolack, Senegal*
[9] *NASA Goddard Institute for Space Studies, New York, NY 10025, USA*
[10] *National Agriculture and Food Research Organization, Tsukuba, Japan*
[11] *Department of Climate, Meteorology and Atmospheric Sciences (CLiMAS), University of Illinois, Urbana-Champaign*
[12] *NSF National Center for Atmospheric Research, Boulder, CO, USA*
[13] *State Key Laboratory of Efficient Utilization of Agricultural Water Resources, China Agricultural University, Beijing 100083, China*
[14] *Center for Agricultural Water Research in China, College of Water Resources and Civil Engineering, China Agricultural University, Beijing 100083, China*
[15] *University of Twente, the Netherlands*
[16] *National Institute for Environmental Studies, Tsukuba, Japan*
[17] *Leibniz Centre for Agricultural Landscape Research (ZALF), Germany*
[18] *BTU Cottbus, Cottbus, Germany*
[19] *Department of Environmental Sciences, University of Basel, Switzerland*
[20] *These authors contributed equally: Cornelia Auer, Kobe De Maeyer*

July 21, 2025

1 General model performance

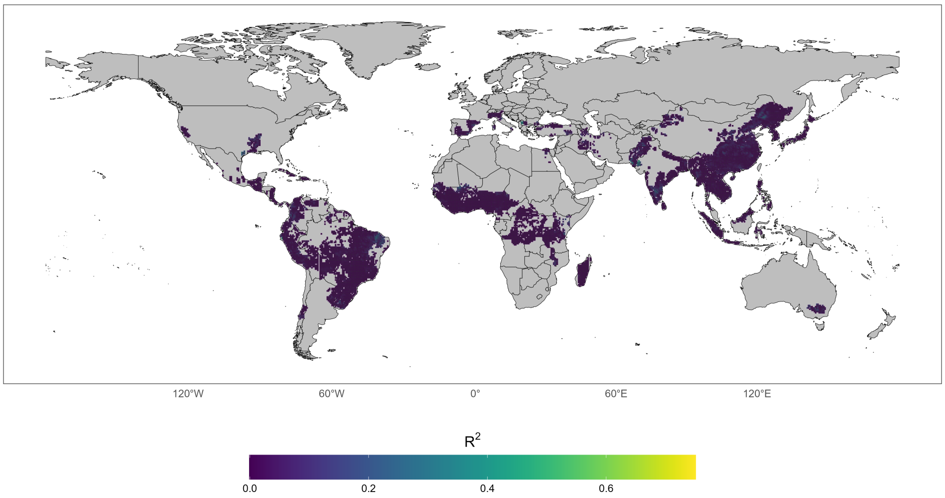
As an alternative to the general model performance table shown in Figure 1 in the main text, where the metric values were averaged using a harvest area weighted mean, we provide here the same table using an unweighted mean (see Supplementary Figure 1). This approach does not account for the differences in the harvest area between grid cells and results in significantly lower R^2 values. This indicates that especially low producing areas (e.g. on the African continent but not only) are poorly represented and lead to lower overall performance.



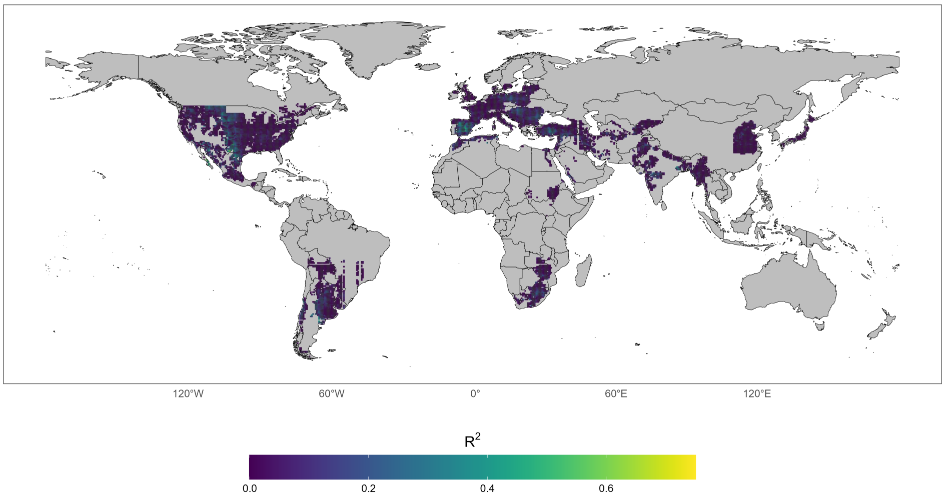
	KGE	RMSE	Rsquared	Variability
acea	2.064	1.127	0.057	1.078
simplace-lintul5	1.531	0.915	0.045	0.82
dssat-pythia	1.302	0.701	0.038	0.539
pepic	1.19	0.631	0.057	0.436
cygma1p74	1.209	0.612	0.027	0.377
ldndc	1.187	0.585	0.048	0.405
pdssat	1.139	0.577	0.048	0.395
promet	1.203	0.553	0.042	0.349
lpj-guess	1.173	0.561	0.046	0.351
epic-iiasa	1.106	0.562	0.043	0.366
lpjml	1.142	0.544	0.052	0.318
isam	1.127	0.521	0.044	0.284
crover	1.129	0.463	0.058	0.193
ensemble	1.045	0.445	0.073	0.222

Supplementary Figure 1: Global assessment of overall crop model performance by comparing the Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) against model results across the time period 1981-2016. Results are aggregated over all four crops (soy, maize, winter wheat, and first growing season rice) and based on the absolute detrending. Table comparing the individual models and model ensemble median for multiple metrics: R^2, Root Mean Square Error (RMSE), Kling-Gupta Efficiency (KGE), and variability (measured by standard deviation). The metrics are computed per grid cell and then averaged into a single value by taking an unweighted mean. Models with low variability tend to exhibit favorable KGE and RMSE values alongside moderate R^2 scores (relative to models with higher variability). Please note, the same table is provided in the main text, without being weighted by harvest area. This suggests that low-yielding regions, often coinciding with regions in income countries, e.g. on the African continent, are inadequately represented in the models, which contributes to lower overall performance.

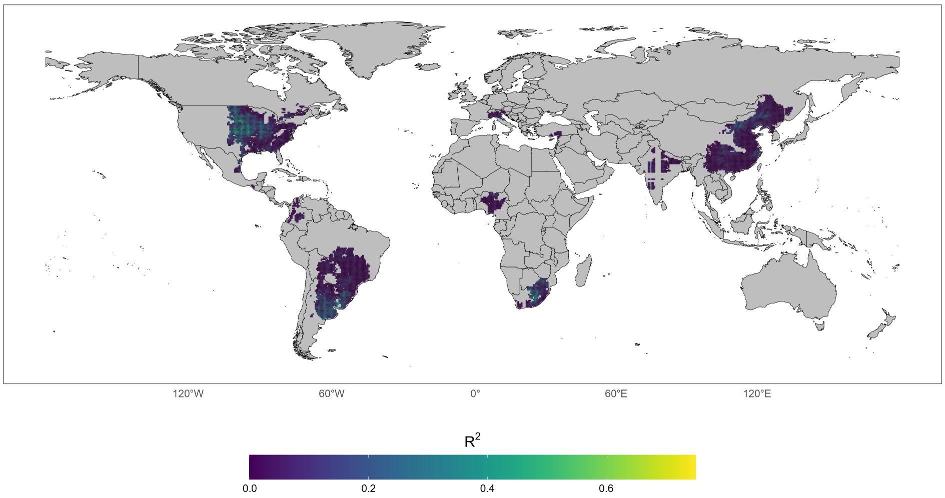
Additionally we provide here the global performance maps for each main crop type- soy, maize, winter wheat, and first growing season rice- separately instead of aggregated over crop types as presented in Figure 1 from the main analysis.



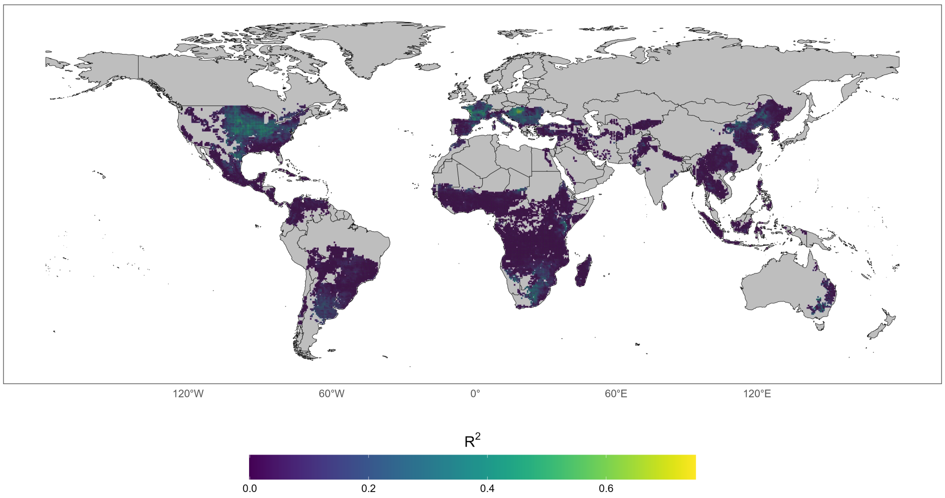
Supplementary Figure 2: RICE (first growing season) – Global assessment of general crop model performance (1981–2016), comparing Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]), using absolute detrended yield values (see Eq. 6). Global R^2 map for the model ensemble median and benchmark data.



Supplementary Figure 3: WINTER WHEAT– Global assessment of geneal crop model performance (1981–2016), comparing Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]), using absolute detrended yield values (see Eq. 6). Global R^2 map for the model ensemble median and benchmark data.



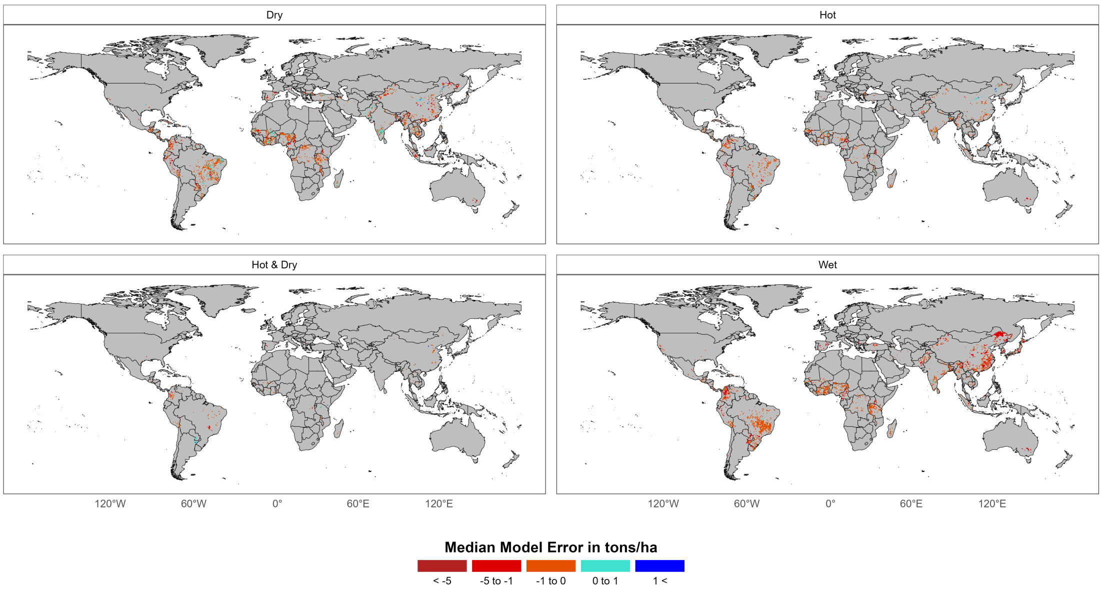
Supplementary Figure 4: SOY – Global assessment of general crop model performance (1981–2016), comparing Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]), using absolute detrended yield values (see Eq. 6). Global R^2 map for the model ensemble median and benchmark data.



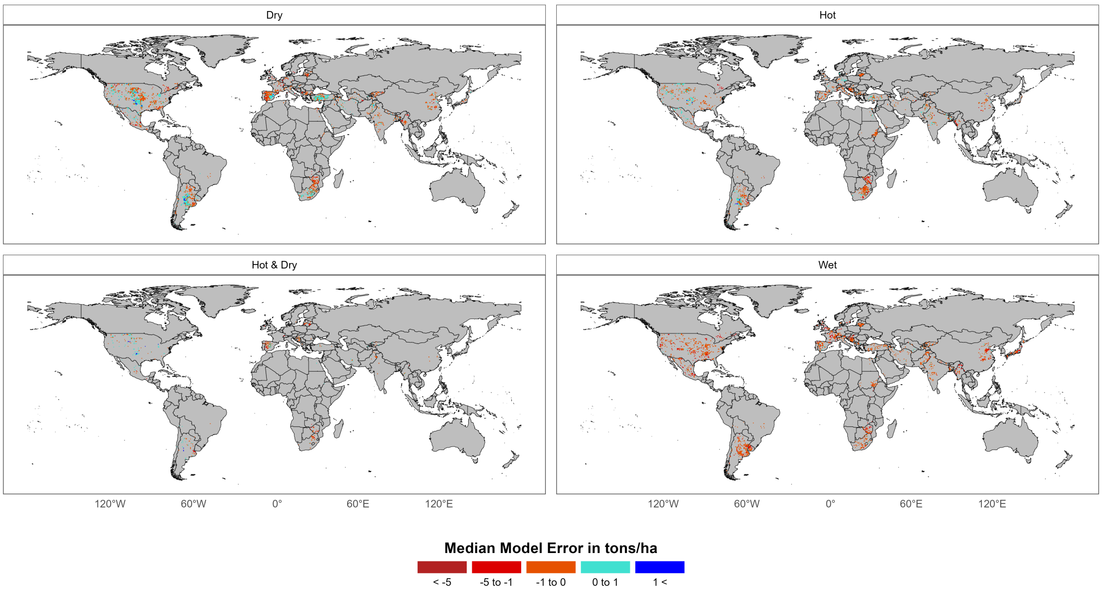
Supplementary Figure 5: MAIZE – Global assessment of general crop model performance (1981–2016), comparing Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]), using absolute detrended yield values (see Eq. 6). Global R^2 map for the model ensemble median and benchmark data.

2 Climatic extremes spatial performance

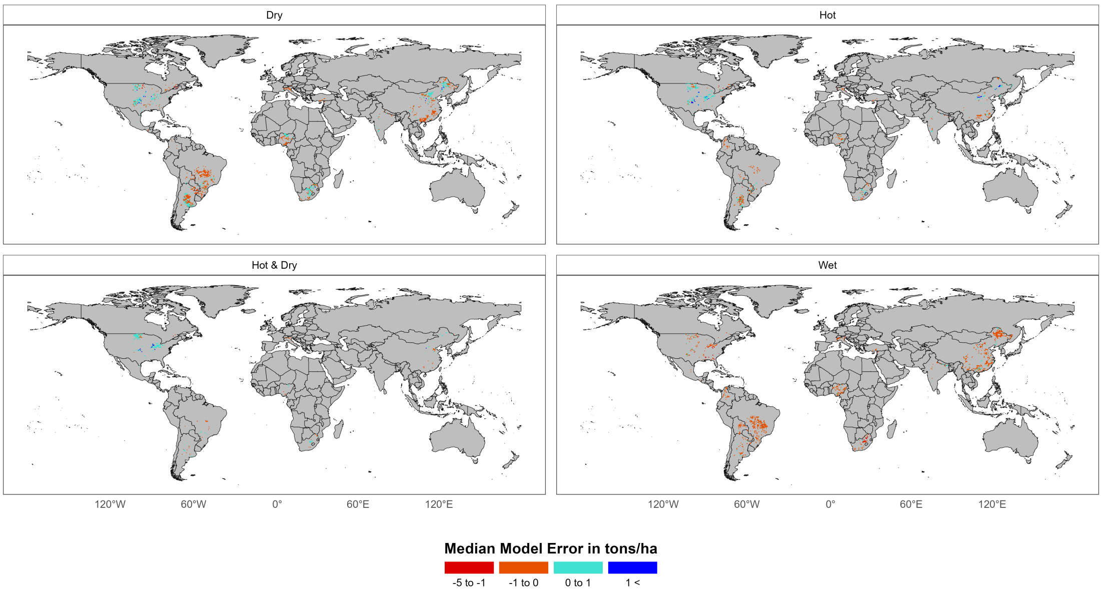
Here we assess the spatial distribution of median errors of the model ensemble for different climatic extremes per crop instead of aggregated over crops as presented in Figure 2 from the main analysis.



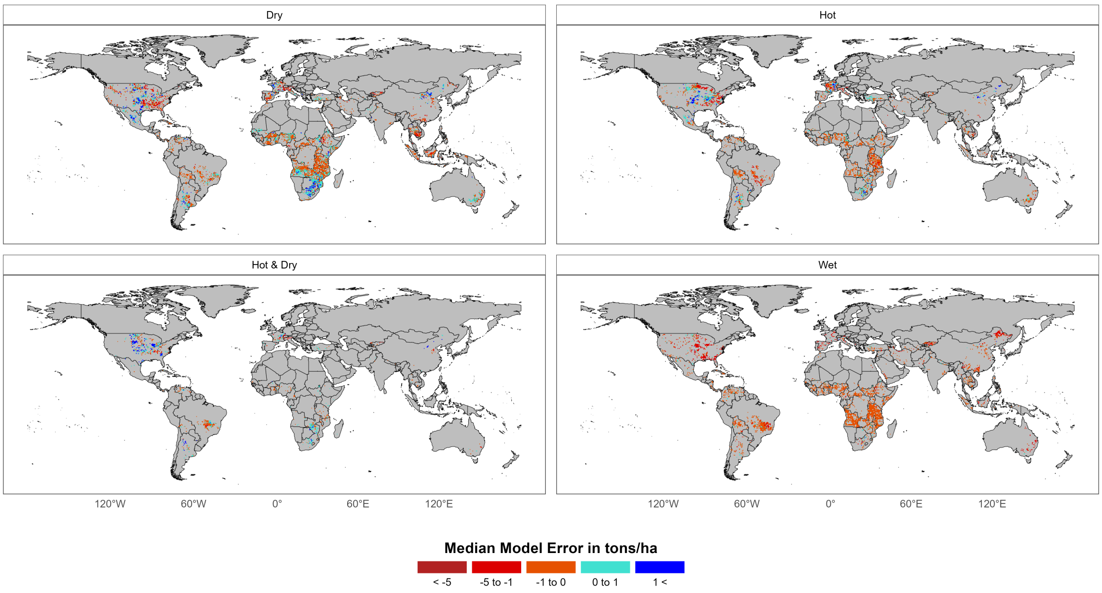
Supplementary Figure 6: RICE (first growing season) – Spatial distribution of model ensemble median errors under different extreme weather conditions: dry, hot, compound hot & dry, and wet. Results are based on detrended yield anomalies (Eq. 6). Colors indicate the median absolute difference between the model ensemble median and benchmark data from GDHY (Iizumi and Sakai [2020]), with orange to red denoting underestimation and turquoise to blue indicating overestimation.



Supplementary Figure 7: WINTER WHEAT – Spatial distribution of model ensemble median errors under different extreme weather conditions: dry, hot, compound hot & dry, and wet. Results are based on detrended yield anomalies (Eq. 6). Colors indicate the median absolute difference between the model ensemble median and benchmark data from GDHY (Iizumi and Sakai [2020]), with orange to red denoting underestimation and turquoise to blue indicating overestimation.



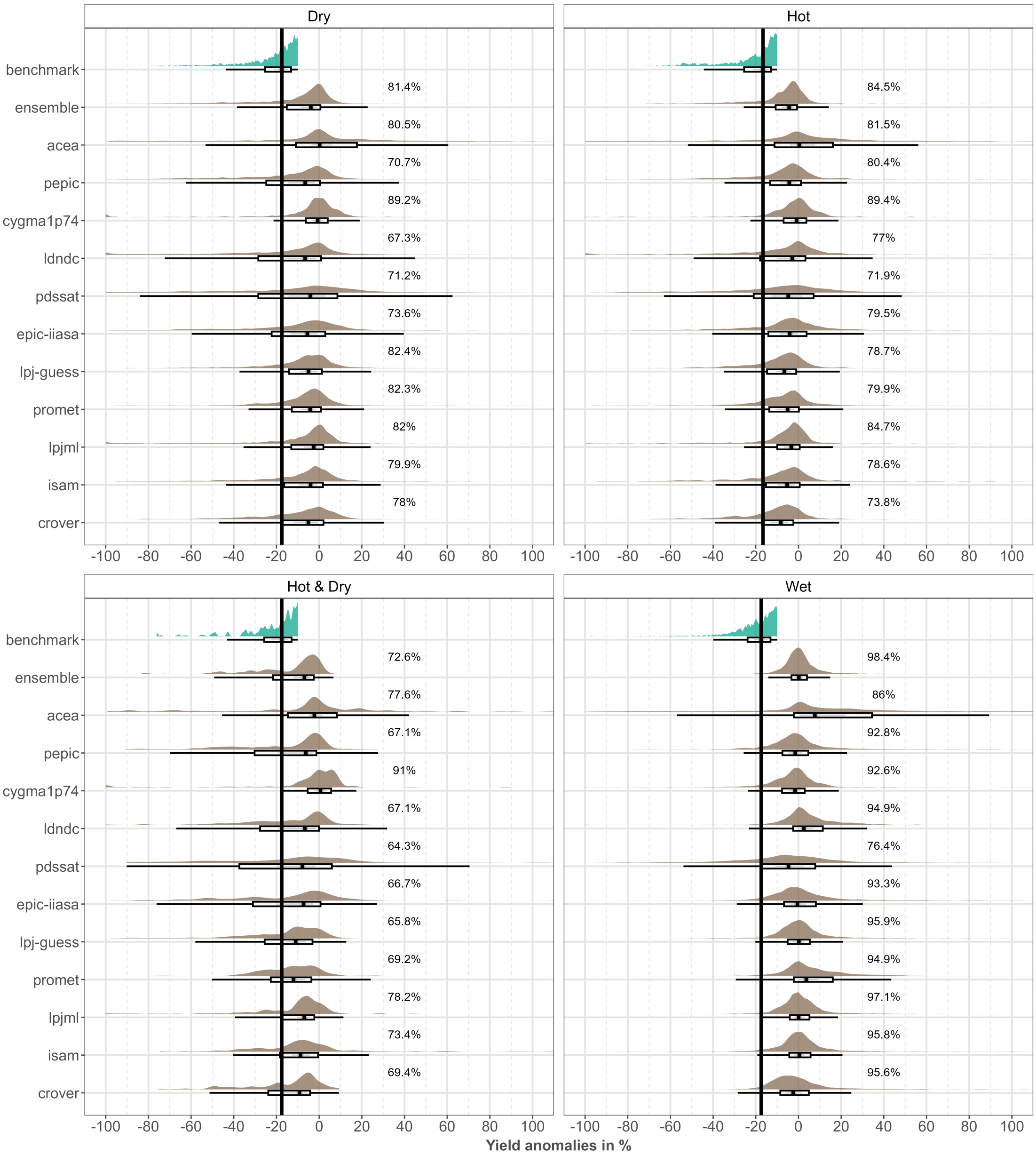
Supplementary Figure 8: SOY – Spatial distribution of model ensemble median errors under different extreme weather conditions: dry, hot, compound hot & dry, and wet. Results are based on detrended yield anomalies (Eq. 6). Colors indicate the median absolute difference between the model ensemble median and benchmark data from GDHY (Iizumi and Sakai [2020]), with orange to red denoting underestimation and turquoise to blue indicating overestimation.



Supplementary Figure 9: MAIZE– Spatial distribution of model ensemble median errors under different extreme weather conditions: dry, hot, compound hot & dry, and wet. Results are based on detrended yield anomalies (Eq. 6). Colors indicate the median absolute difference between the model ensemble median and benchmark data from GDHY (Iizumi and Sakai [2020]), with orange to red denoting underestimation and turquoise to blue indicating overestimation.

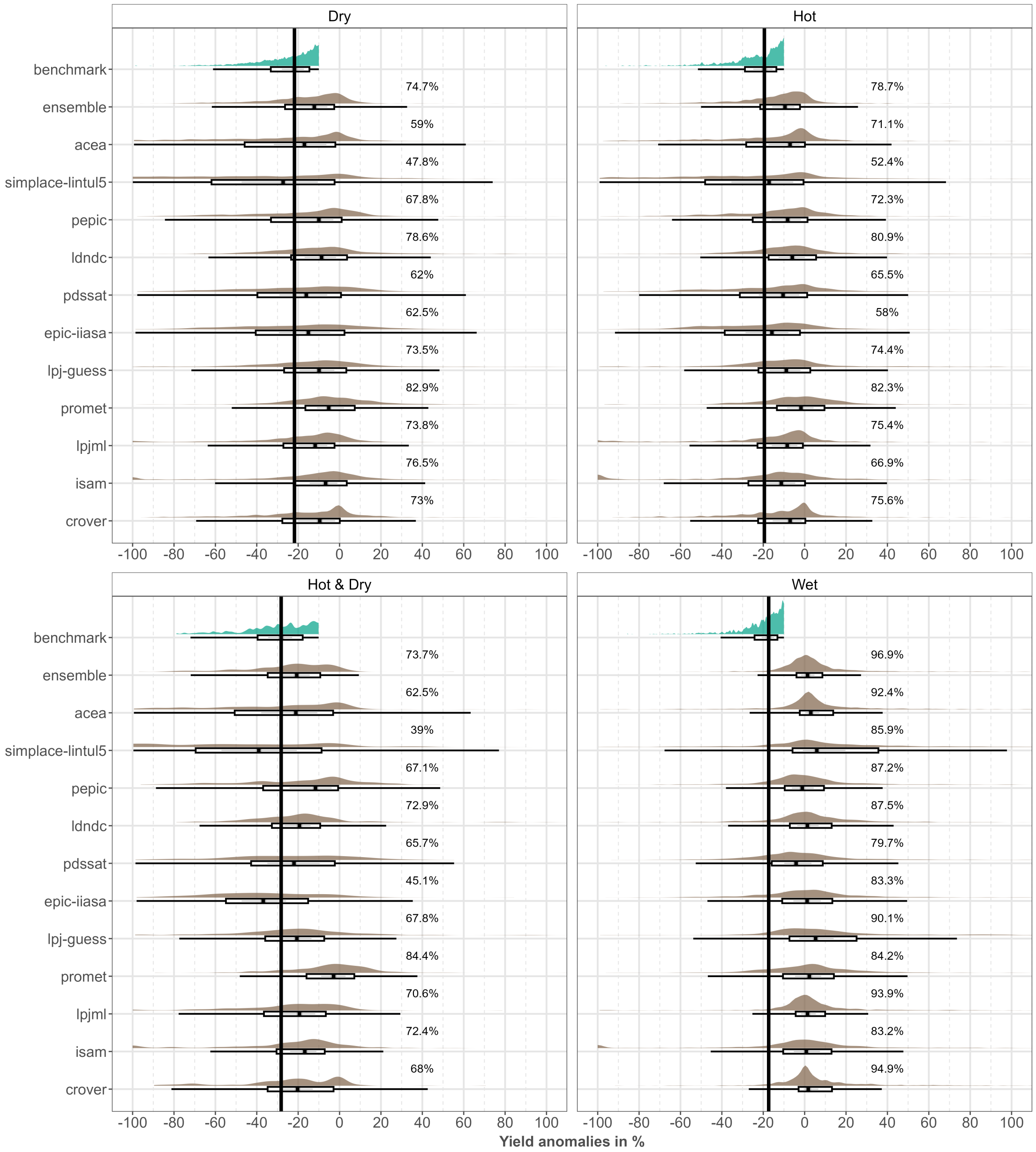
3 Raincloud plots: relative detrended yield anomalies

Four main crop types - soy, maize, winter wheat, and first growing season rice- are assessed individually below. The Supplementary Figures 10, 11, 12, 13, depict raincloud plots of the distribution of relative detrended yield anomalies (in %) of the Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) (red) and of individual and model ensemble (blue) across the extremes hot, dry, compound hot & dry, and wet. Positive yield anomalies show a model misinterpretation, where extremes interpreted to having positive impacts on crop growth.) Boxplots below the distribution plots depict the median and interquartile range (IQR). These plots are the pendant to the Figure 3 in the main text, there aggregated over all crop types.

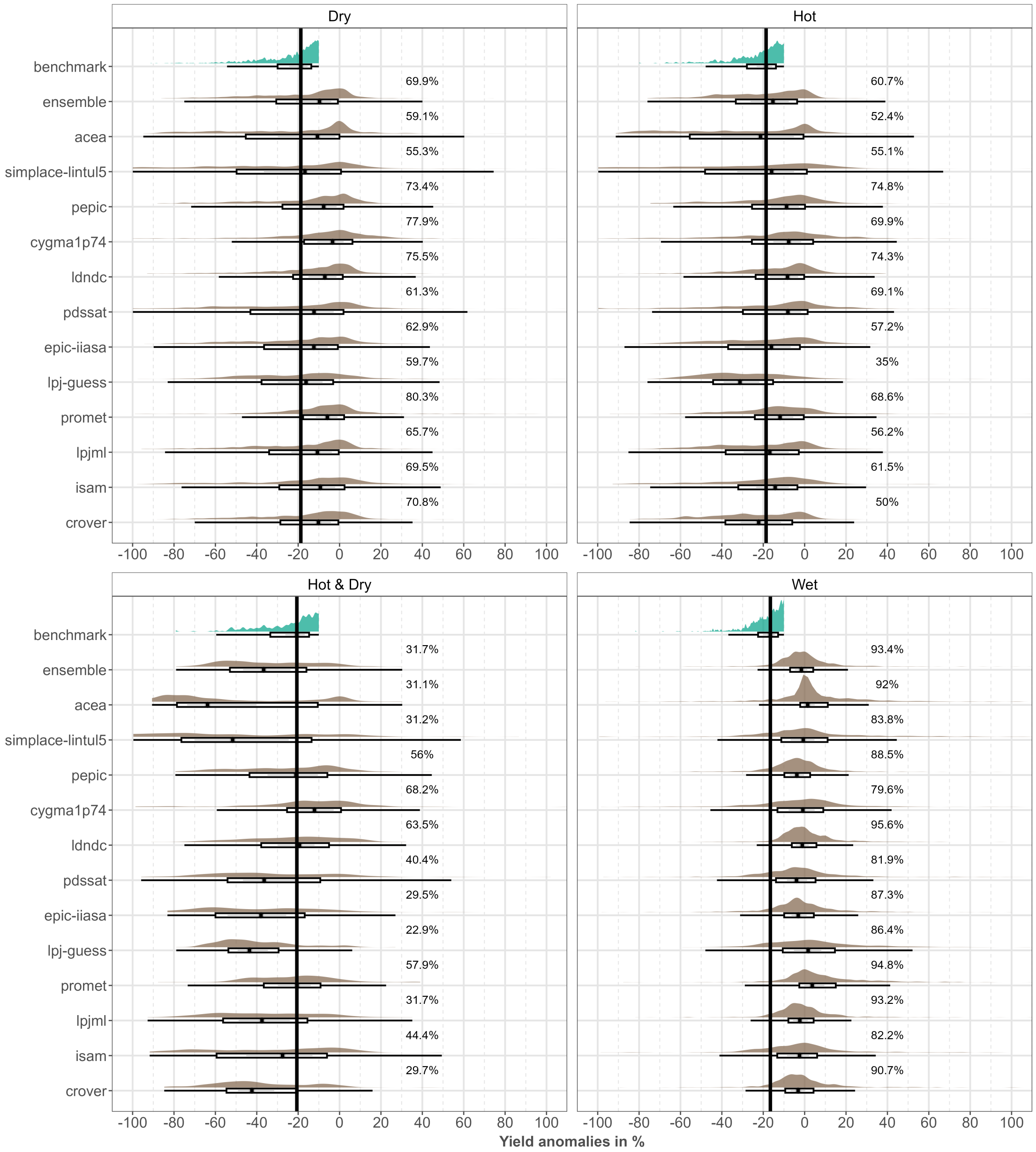


Supplementary Figure 10: RICE (first growing season) – Raincloud plots for relative detrended yield anomalies (red=benchmark, gray=models) across extremes. Boxplots show median/IQR. Extreme conditions for rice seems to be a specific challenge. Anomaly distributions show greater leftward shifts for hot than dry extremes, hot & dry extremes are better represented. Wet event distributions typically centre around zero, except crover and pdssat models better capture wet impacts.

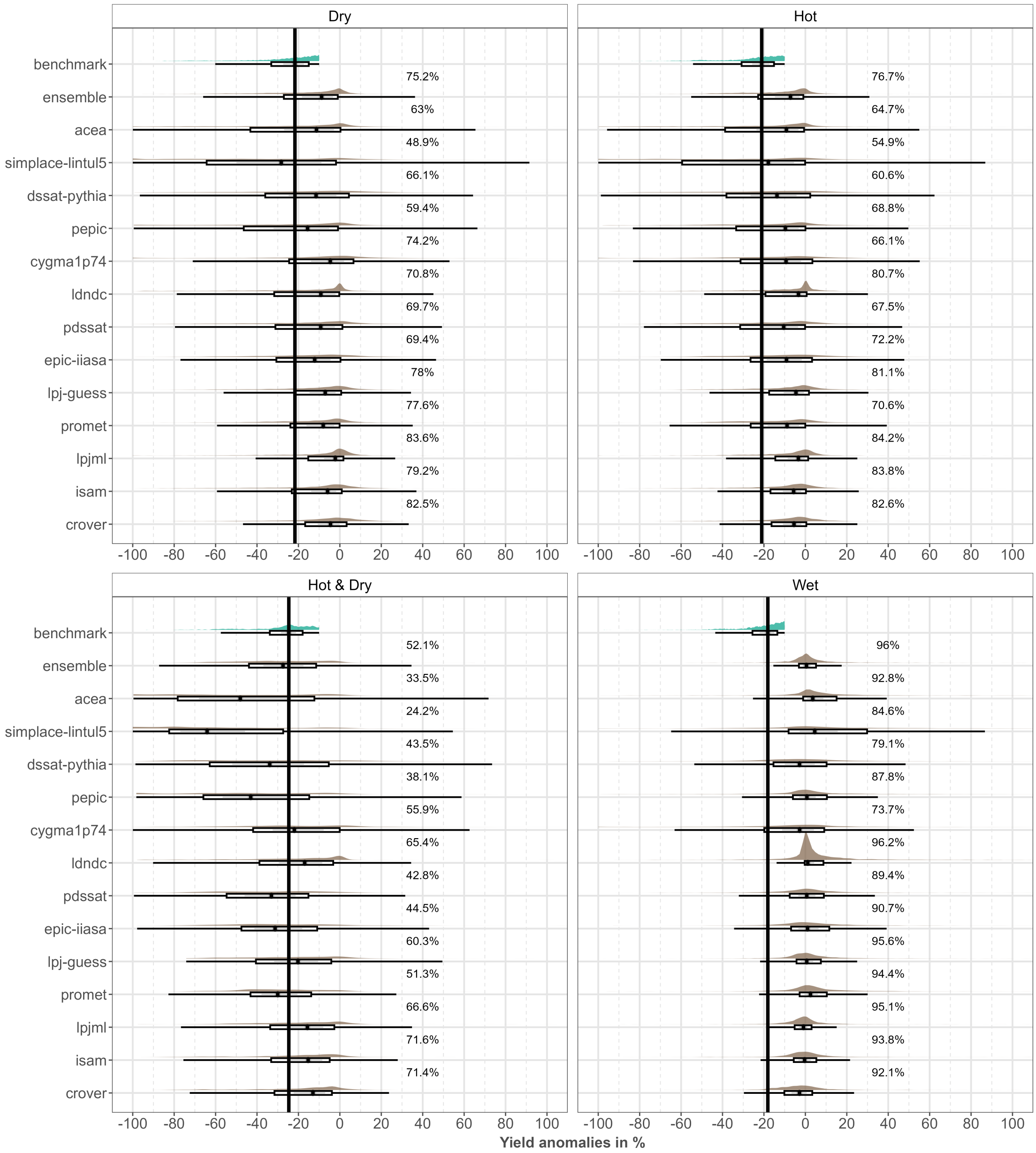
Note: medians may be skewed due to considerable number of outliers.



Supplementary Figure 11: WINTER WHEAT – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) across extremes. Boxplots show median/IQR. Models much better catch the signal of the extremes, except for wet extremes. For dry, hot, dry & hot mostly all models have their interquartile range in the negative area. The models show high variability. For hot & dry two models even overestimate the impact (simplace-lintul5 and epic-iiasa). Wet event distributions typically centre around zero, except pdssat model better captures the wet impact. lpj-guess and simplace-lintul5 misinterpret the signal and even show positive impact (median) by the extreme. Note: medians may be skewed due to considerable number of outliers.



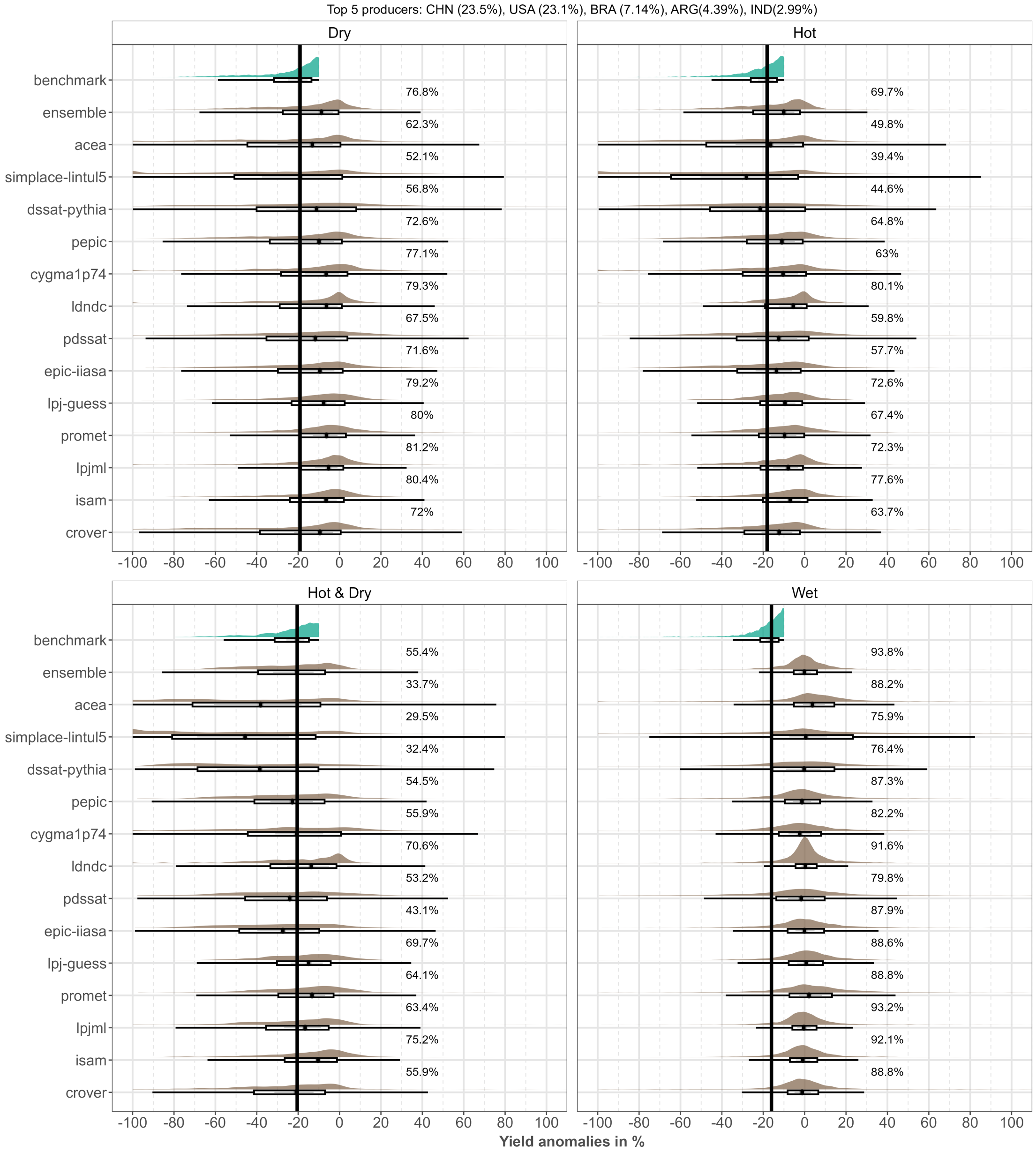
Supplementary Figure 12: SOY – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) across extremes. Boxplots show median/IQR. Models show exhibit exceptional behaviour among each other for soy. This is especially expressed in the hot & dry case, where median values are broadly distributed over the negative range. While for dry and hot extremes underestimation dominates, in the case of compound hot & dry extremes most models overestimate the impact (when looking at the median). Wet event distributions typically centre around zero, except pdssat model better captures the wet impact.



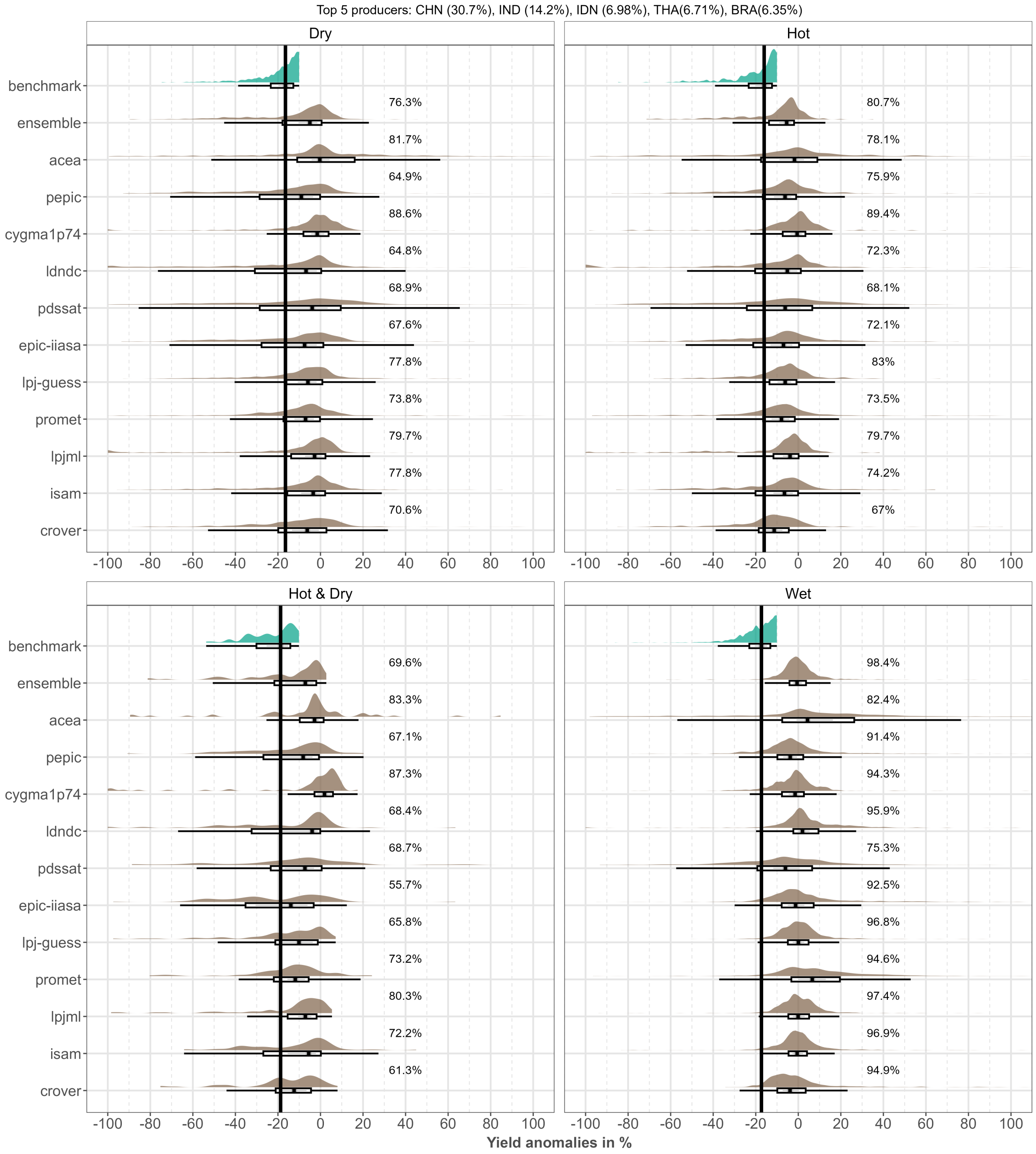
Supplementary Figure 13: MAIZE – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) across extremes. Boxplots show median/IQR. Maize has very few data points, still some trends can be detected. Dry and hot extremes are underestimated, but the negative signal is often detected. For compound hot & dry extremes more than half of the models have a median showing higher impacts than the benchmark data. Wet event distributions typically centre around zero, and models hardly catch the signal.

4 Raincloud plots: relative detrended yield anomalies for top five producers

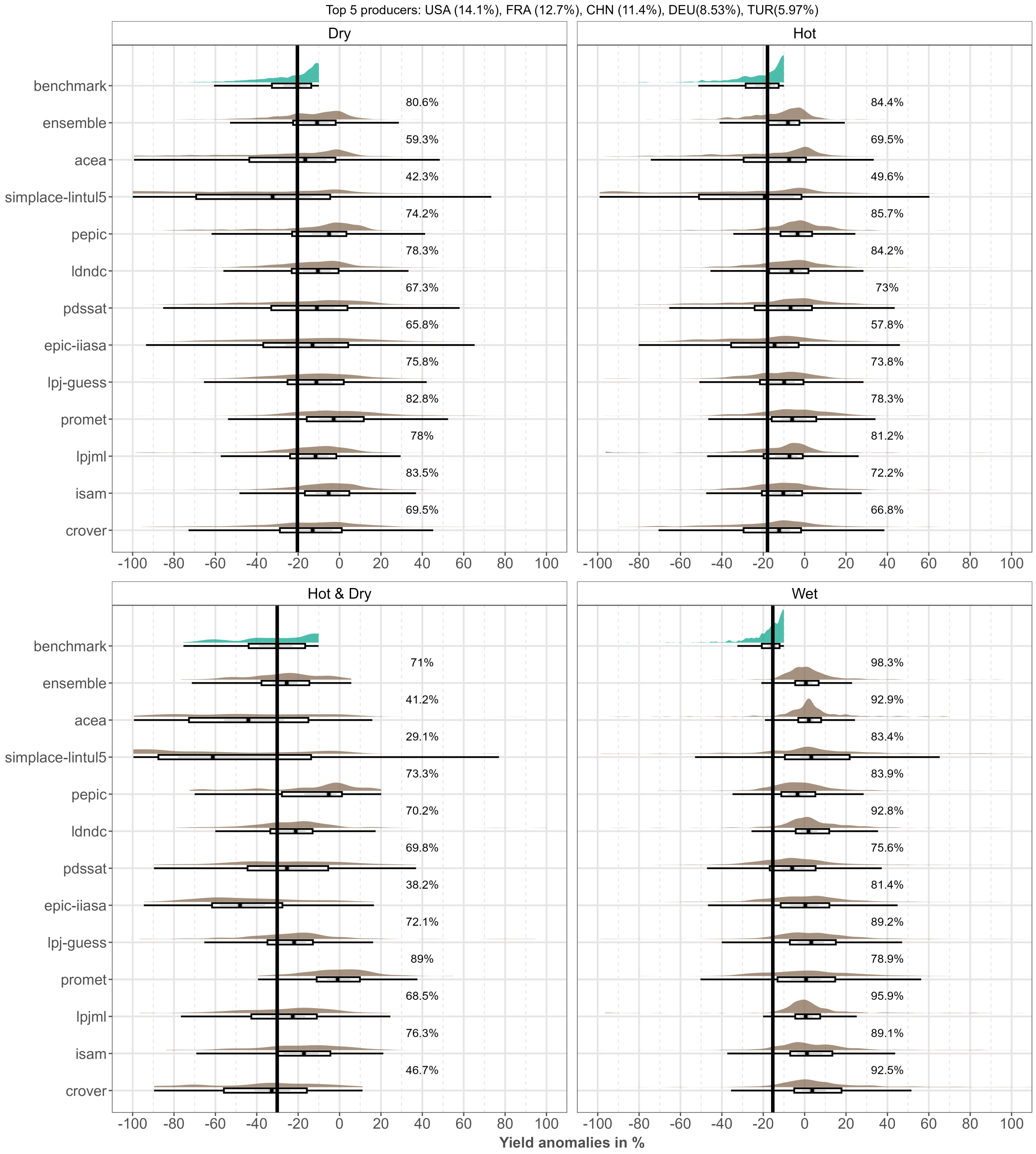
Four main crop types - soy, maize, winter wheat, and first growing season rice- are assessed individually below. Instead of using all data as done earlier, we only look at extremes occurring in the top five producers for each crop respectively here. The Supplementary Figures 14, 15, 16, 17, 18, depict raincloud plots of the distribution of relative detrended yield anomalies (in %) of the Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) (red) and of individual and model ensemble (blue) across the extremes hot, dry, compound hot & dry, and wet. Positive yield anomalies show a model misinterpretation, where extremes interpreted to having positive impacts on crop growth.) Boxplots below the distribution plots depict the median and interquartile range (IQR). These plots are the pendant to the Figure 3 in the main text, there aggregated over all crop types and for all data points.



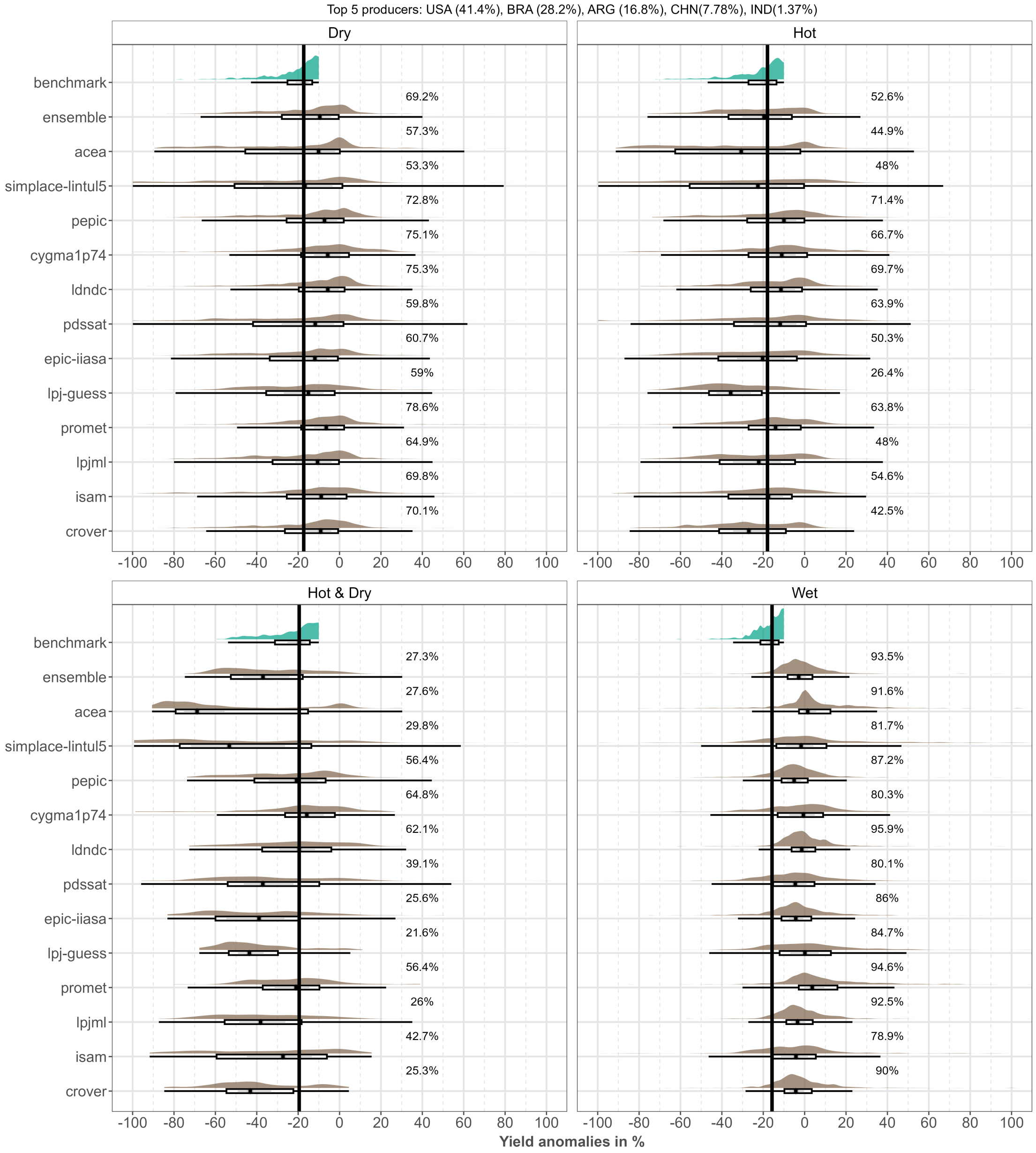
Supplementary Figure 14: Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) of aggregated crop types of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median for climate extremes. Boxplots show median/IQR. The percentages shown alongside the distributions indicate the proportion of underestimation cases for each model and the ensemble. Only extreme events occurring in the top five producing countries—China (23.5% of global production), USA (23.1%), Brazil (7.14%), Argentina (4.39%), and India (2.99%)—are included in this figure. Consistent with the main findings, models tend to underestimate the effects of dry, hot, and wet extremes even when restricted to the top-producing countries.



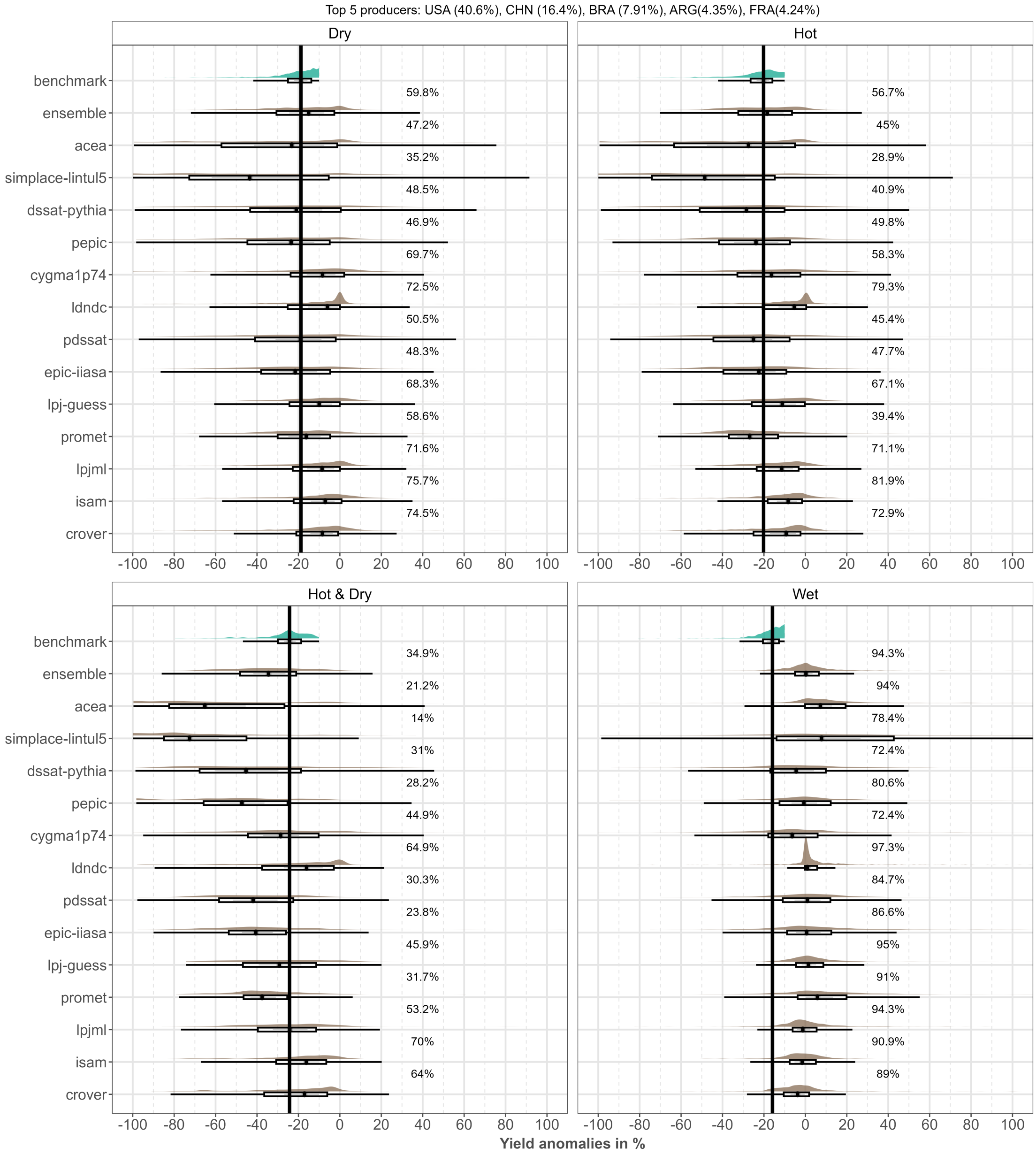
Supplementary Figure 15: RICE (first growing season) – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median for climate extremes. Boxplots show median/IQR. The percentages shown alongside the distributions indicate the proportion of underestimation cases for each model and the ensemble. Only extreme events occurring in the top five producing countries—China (30.7% of global production), India (14.2%), Indonesia (6.98%), Thailand (6.71%), and Brazil (6.35%)—are included in this figure. Consistent with the main findings, models tend to underestimate the effects of climate extremes even when restricted to the top-producing countries.



Supplementary Figure 16: WINTER WHEAT – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median for climate extremes. Boxplots show median/IQR. The percentages shown alongside the distributions indicate the proportion of underestimation cases for each model and the ensemble. Only extreme events occurring in the top five producing countries—USA (14.1% of global production), France (12.7%), China (11.4%), Germany (8.53%), Türkiye (5.97%)—are included in this figure. Consistent with the main findings, models tend to underestimate the effects of dry, hot, and wet extremes even when restricted to the top-producing countries.



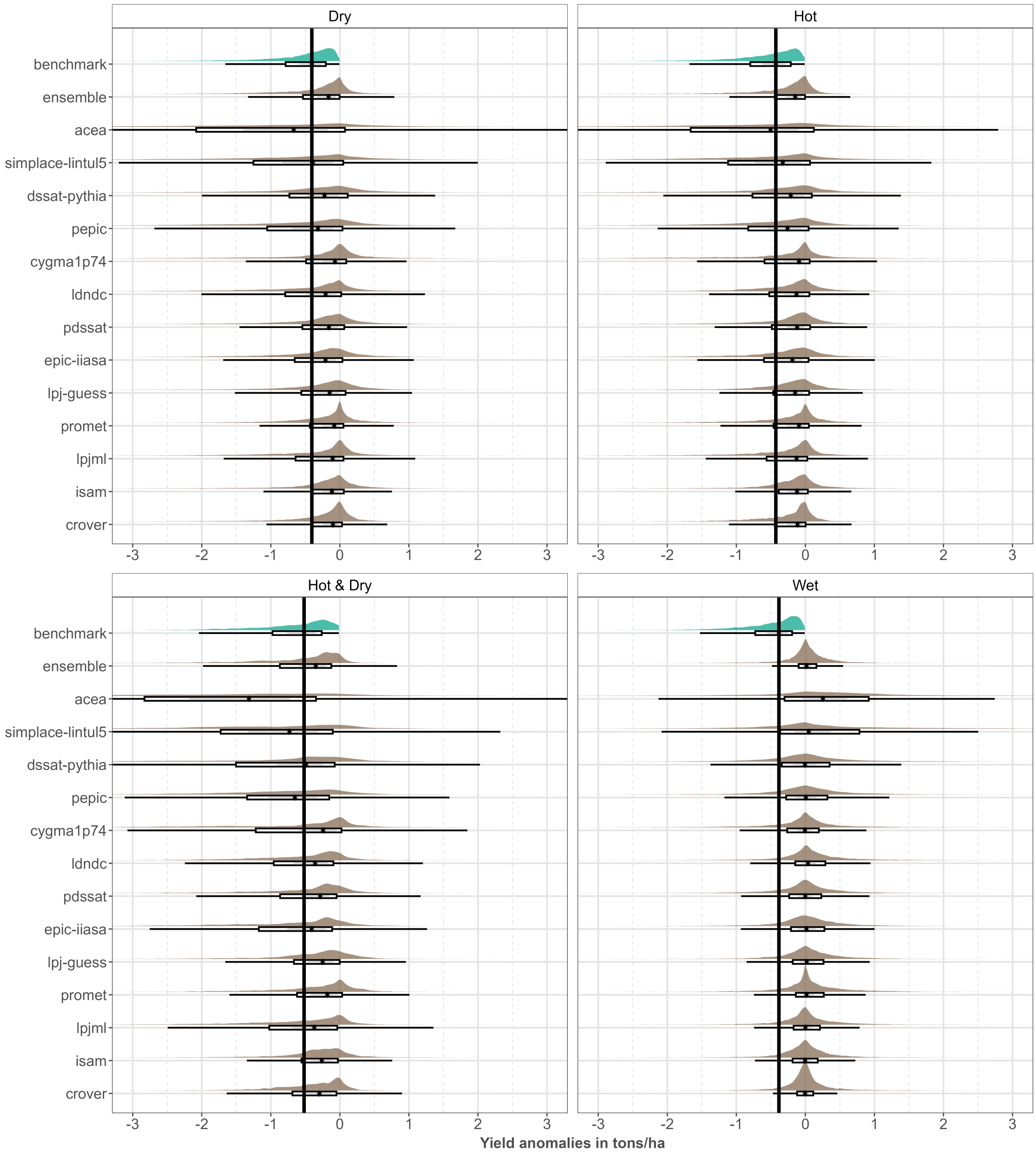
Supplementary Figure 17: SOY – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median for climate extremes. Boxplots show median/IQR. The percentages shown alongside the distributions indicate the proportion of underestimation cases for each model and the ensemble. Only extreme events occurring in the top five producing countries—USA (41.4% of global production), Brazil (28.2%), Argentina (16.8%), China (7.8%), and India (1.37%)—are included in this figure. Consistent with the main findings, models tend to underestimate the effects of dry and wet extremes even when restricted to the top-producing countries. For hot and compound hot and dry extremes, some models generally overestimate the impacts instead.



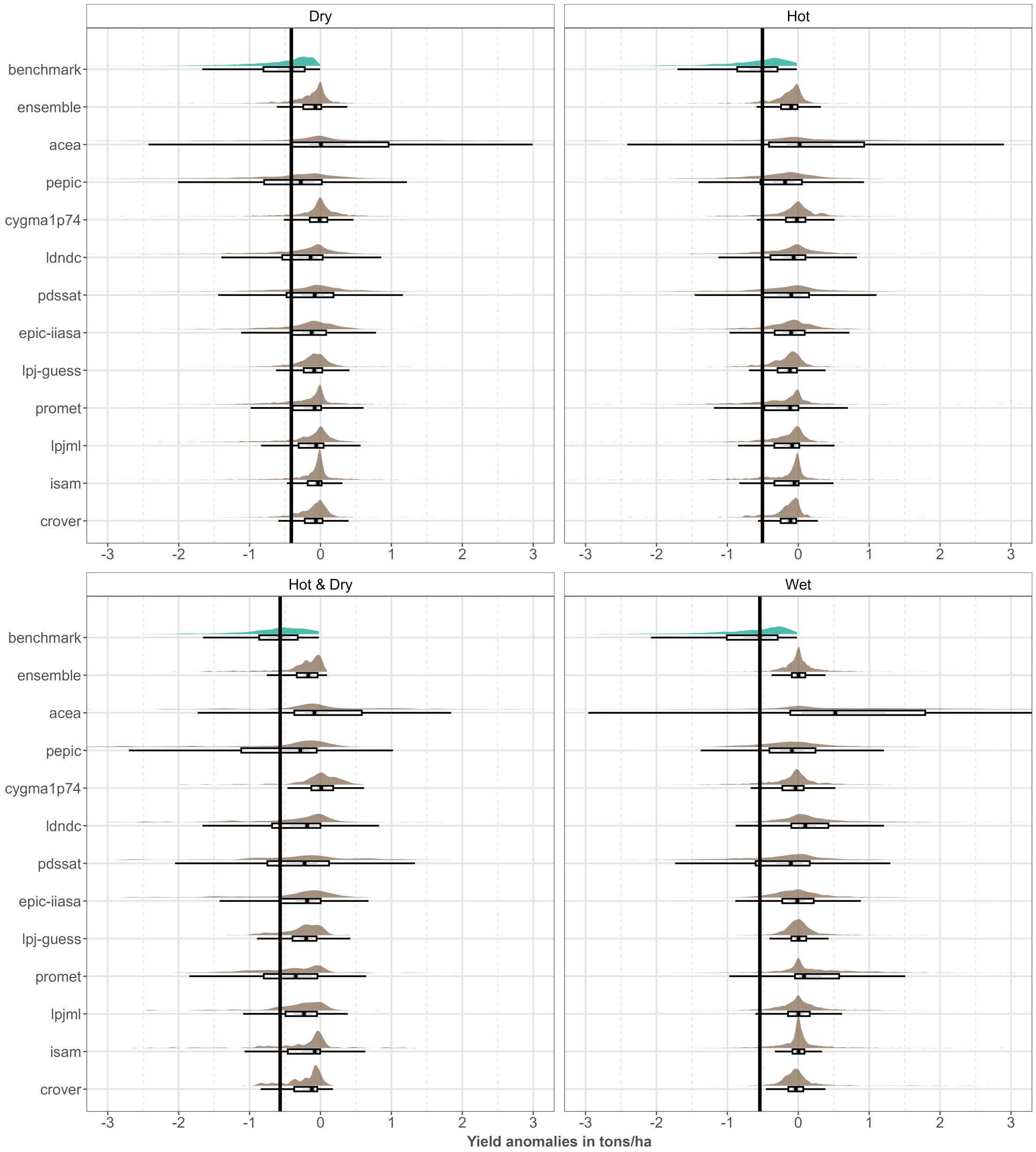
Supplementary Figure 18: MAIZE – Raincloud plots for relative detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median for climate extremes. Boxplots show median/IQR. The percentages shown alongside the distributions indicate the proportion of underestimation cases for each model and the ensemble. Only extreme events occurring in the top five producing countries—USA (40.6% of global production), China (16.4%), Brazil (7.91%), Argentina (4.35%), and France (4.24%)—are included in this figure. The models show a mixed response: while some generally underestimate the effects of climatic extremes, others tend to overestimate them—except in the case of wet extremes, where all models systematically underestimate the impacts.

5 Raincloud plots: absolute detrended yield anomalies

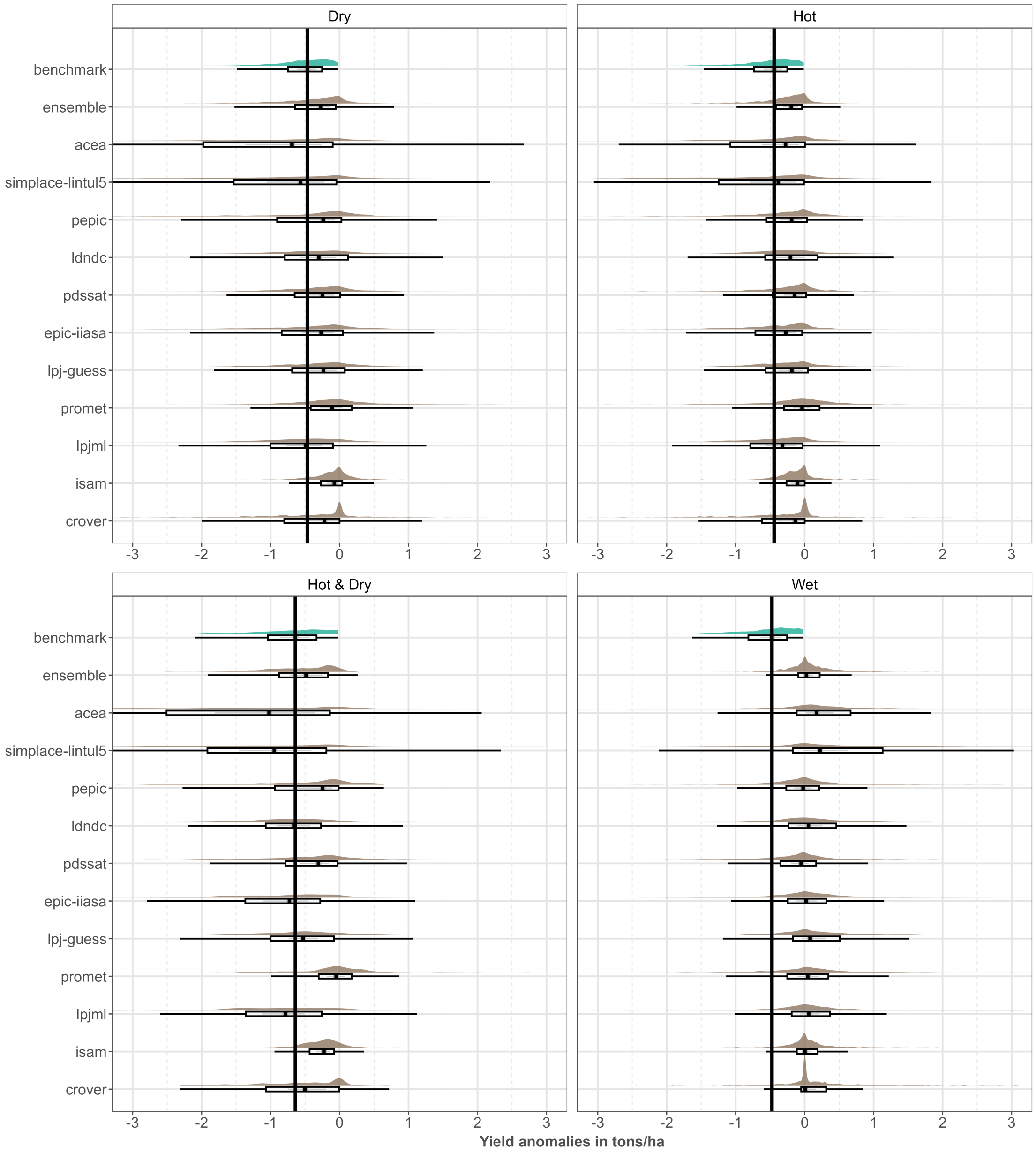
Four main crop types - first growing season rice, winter wheat, soy, and maize - are assessed together and individually below. The Supplementary Figures 19, 20, 21, 22,23 depict raincloud plots of the absolute detrended distribution of yield anomalies (in tons/ha) of the Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) (red) and of individual and model ensemble (blue) across the extremes hot, dry, compound hot & dry, and wet. Positive yield anomalies show a model misinterpretation, where extremes are assigned to having positive impacts on crop growth.) Boxplots below the distribution plots depict the median and interquartile range (IQR).



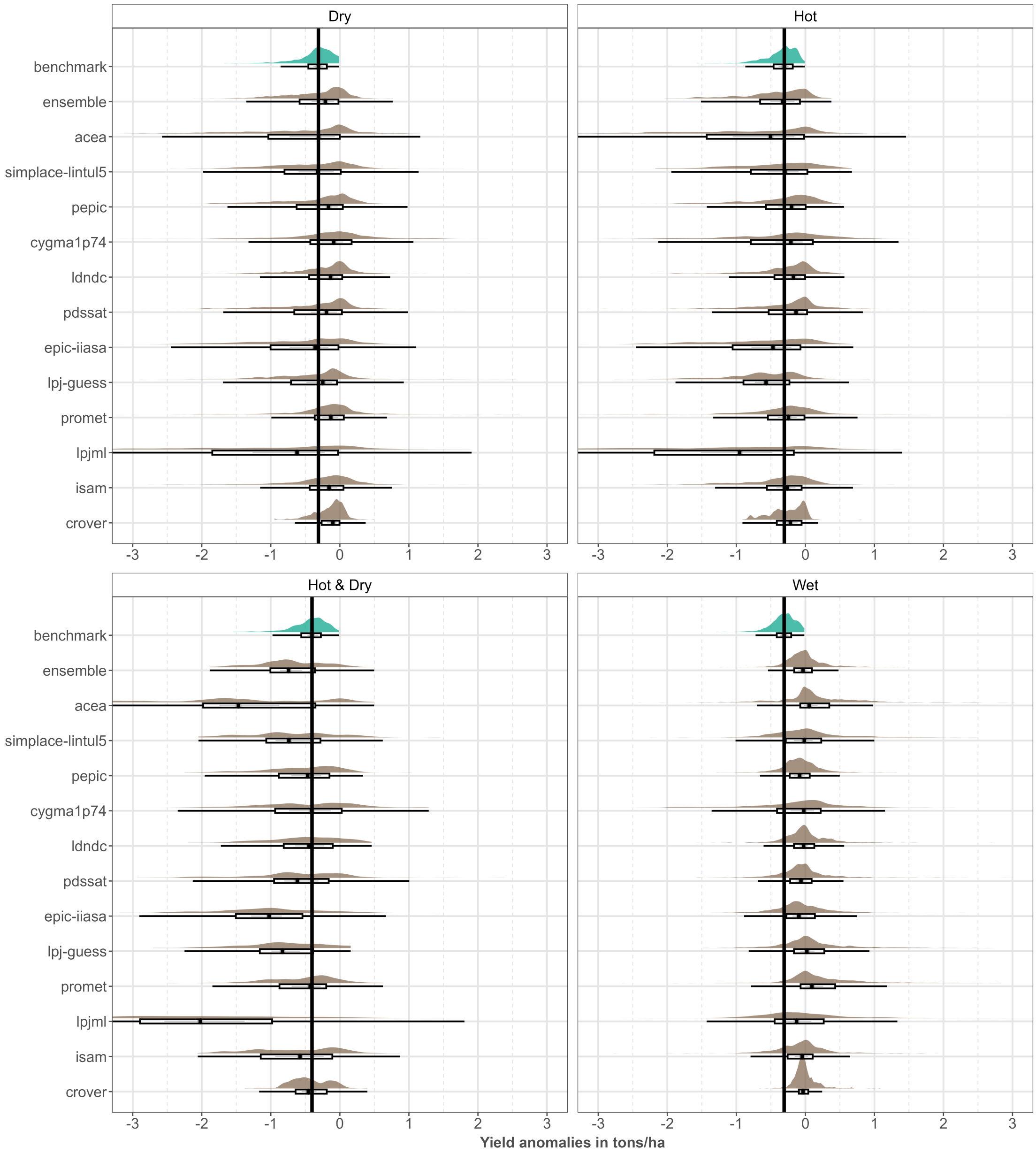
Supplementary Figure 19: Raincloud plots for absolute detrended yield anomalies (red=benchmark, blue=models) of aggregated crop types of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median (tons/ha) for climate extremes. Boxplots show median/IQR. Although the median shows that in most cases the models are able to capture a negative signal by the respective extreme they are seldomly able to simulate the magnitude.



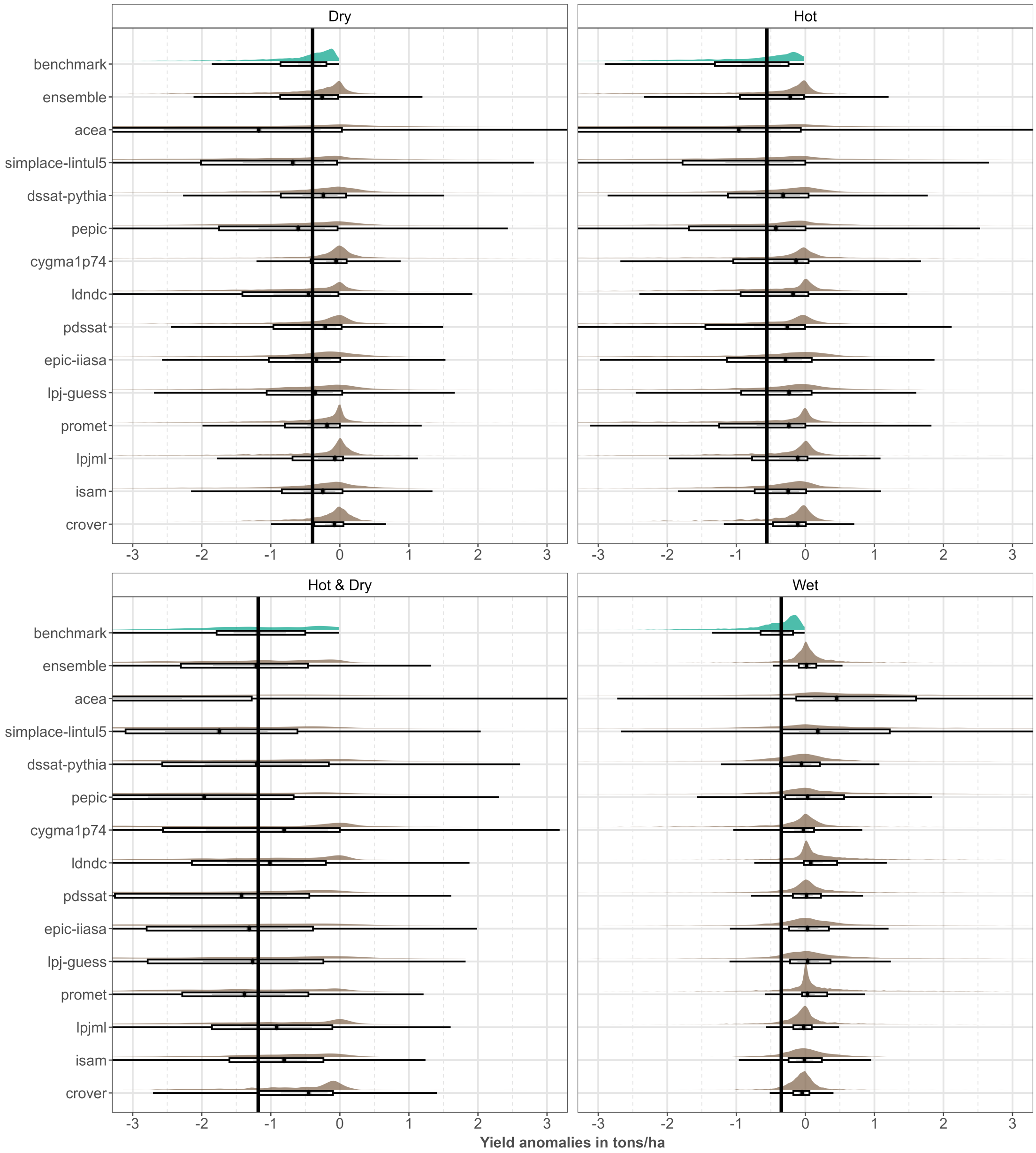
Supplementary Figure 20: RICE (first growing season) – Raincloud plots for absolute detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median (tons/ha) for climate extremes. Boxplots show median/IQR. For dry and hot extremes the anomaly distributions except for pepic mostly centre around zero. For wet as well, with the exception of acea which misinterprets the signal as something positive. For hot & dry combined the distribution of nearly half of the models shifts to the left and thus, better represents the negative impact of the signal.



Supplementary Figure 21: WINTER WHEAT – Raincloud plots for absolute detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median (tons/ha) for climate extremes. Boxplots show median/IQR. Some models exhibit strong outliers for wheat under all extreme types. Median values range from over to underestimation, with lpjml performing having close agreement of its median with the median of the benchmark for dry, hot, hot & dry. Wet extremes are consequently underestimated and anomalies centre again for most models around zero.



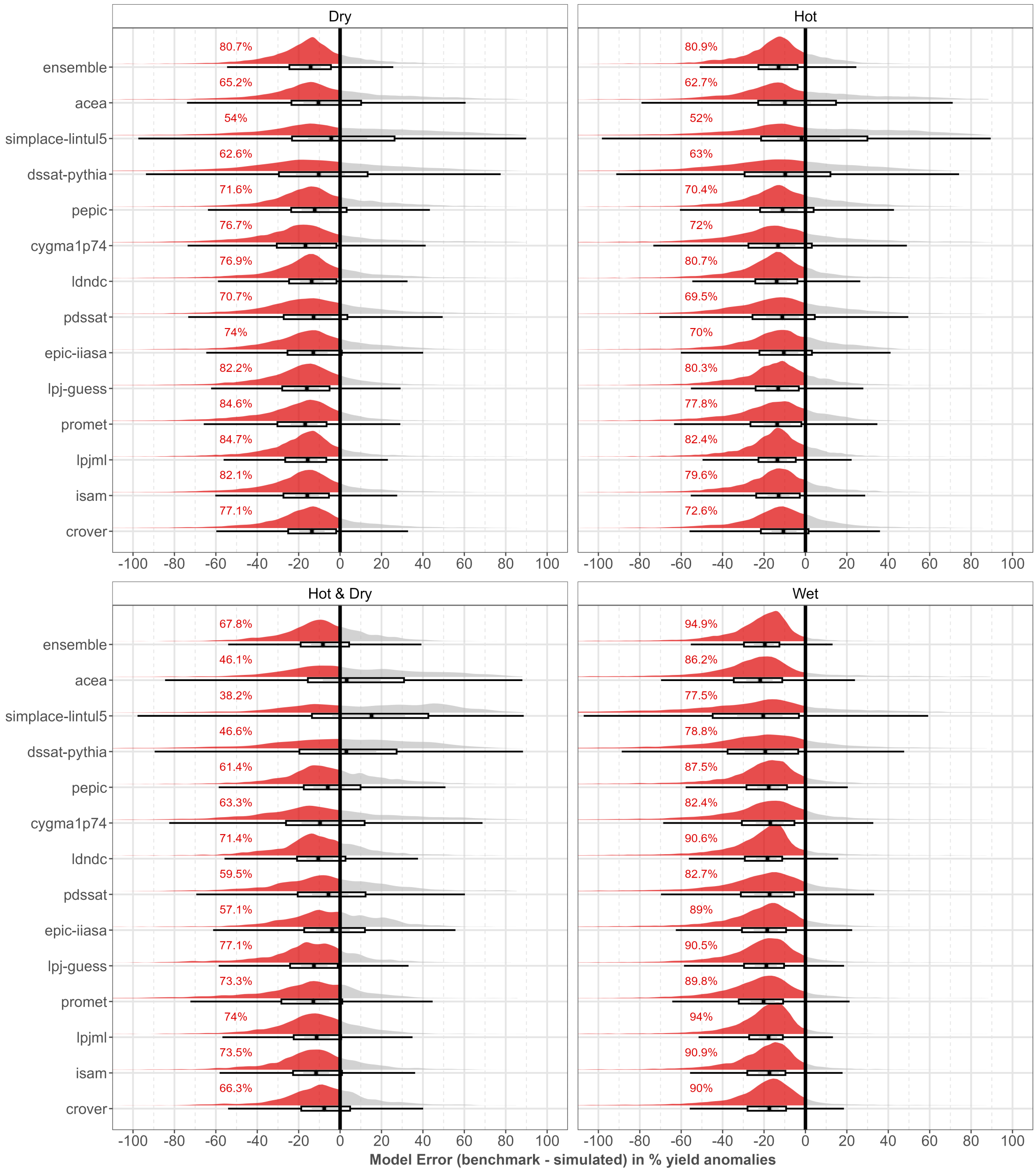
Supplementary Figure 22: SOY – Raincloud plots for absolute detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median (tons/ha) for climate extremes. Boxplots show median/IQR. The heterogeneity we saw for the relative anomalies above in Figure 12 is not as expressed for dry and hot extremes here with an absolute error. Many models have a median that is very close to the median of the benchmark data. lpjml consequently overestimates the impact (except for wet) and has an unusual high variance. Hot & dry has a many models where the median even exceeds the benchmark median, and again for wet extremes the distribution of the models mostly centres around zero, except this time for lpjml that has tendency to better detect the negative signal.



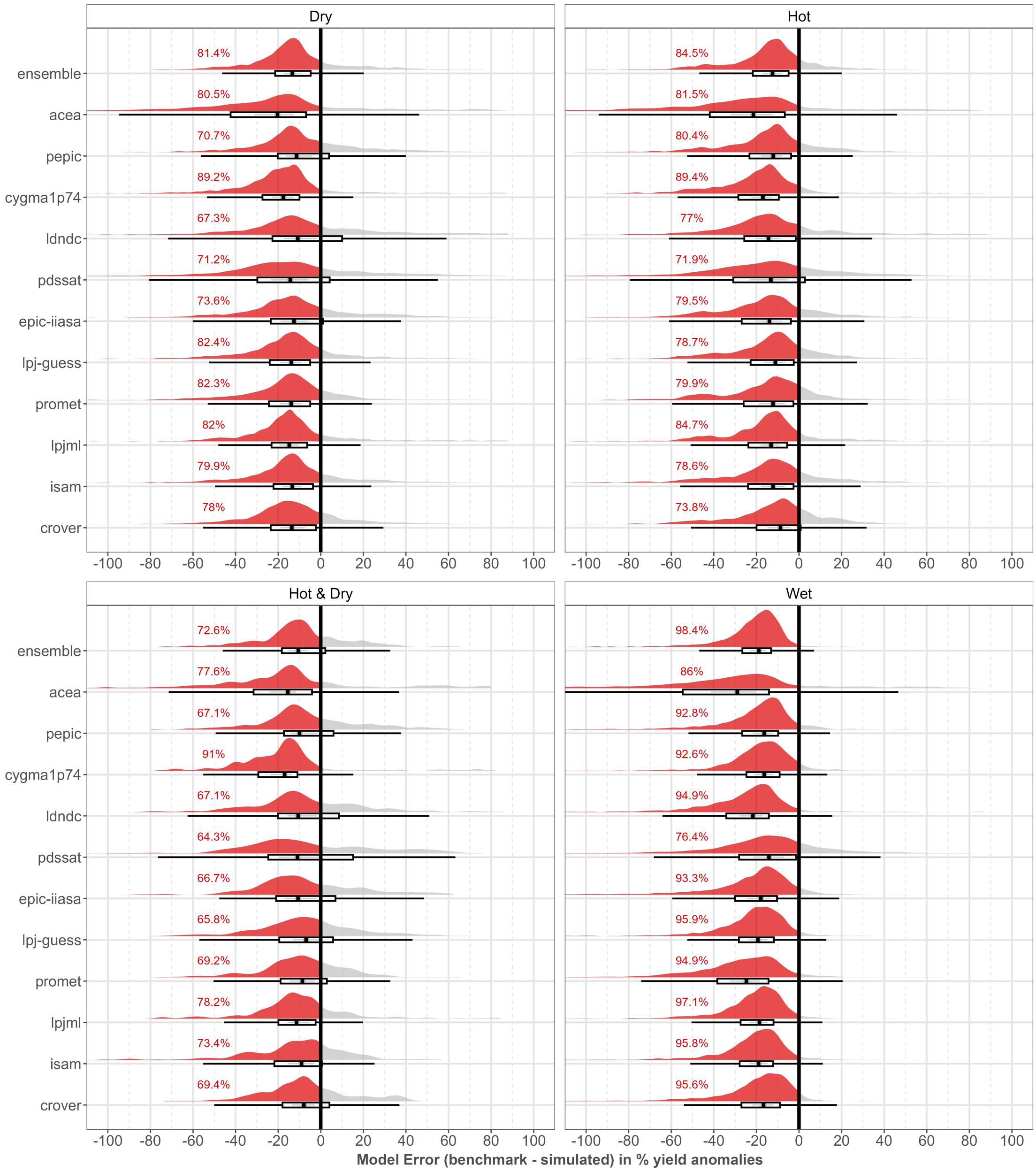
Supplementary Figure 23: MAIZE – Raincloud plots for absolute detrended yield anomalies (red=benchmark, blue=models) of Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020]) and individual models or model ensemble median (tons/ha) for climate extremes. Boxplots show median/IQR. Compared to the other crops we see that already the benchmark data for maize has a much broader range in terms of absolute anomalies. We also see that the models exhibit a broad range of outliers across dry, hot, compound hot & dry extremes. For many models the median in the case of dry extremes is close to that one of the benchmark data. Less so for hot extremes. For compound hot & dry extremes the models exhibit high variance and strong outliers in the positive and negative direction. For wet extremes the distribution of the models mostly centres around zero, except for acea and simplace-lintul5 misinterpreting the signal in the positive direction.

6 Raincloud plots: relative model error

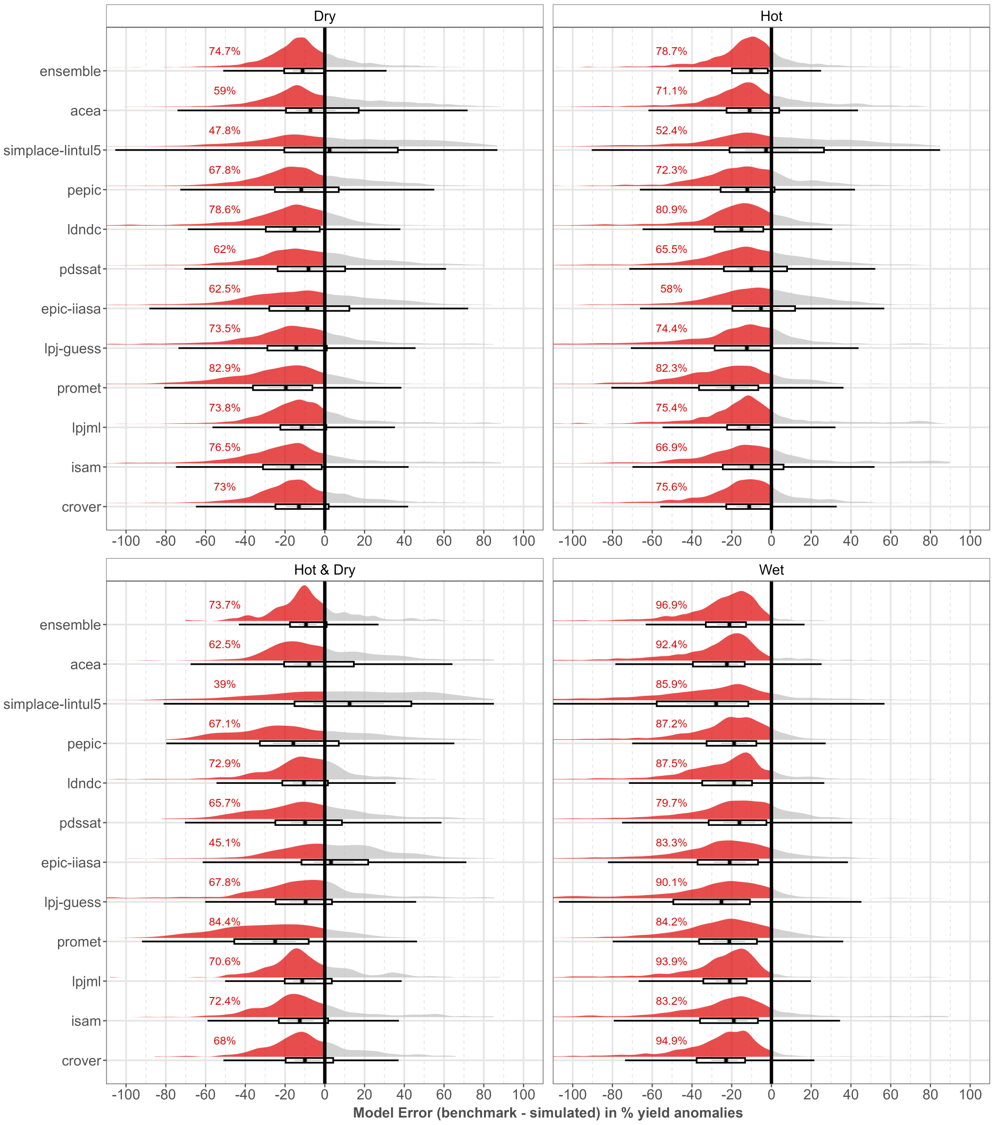
To also depict the extent of the error, we display for four main crop types - first growing season rice, winter wheat, soy, and maize - the distribution of the individual and ensemble relative model error in the Supplementary Figures 24, 25, 26, 27,28. The relative model error is quantified as the difference between observed (Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020])) and simulated yield anomalies in percent (i.e. based on relative detrending) across the extremes hot, dry, compound hot & dry, and wet. Positive relative model errors represent an overestimation of the negative impact. Boxplots below the distribution plots depict the median and interquartile range (IQR).



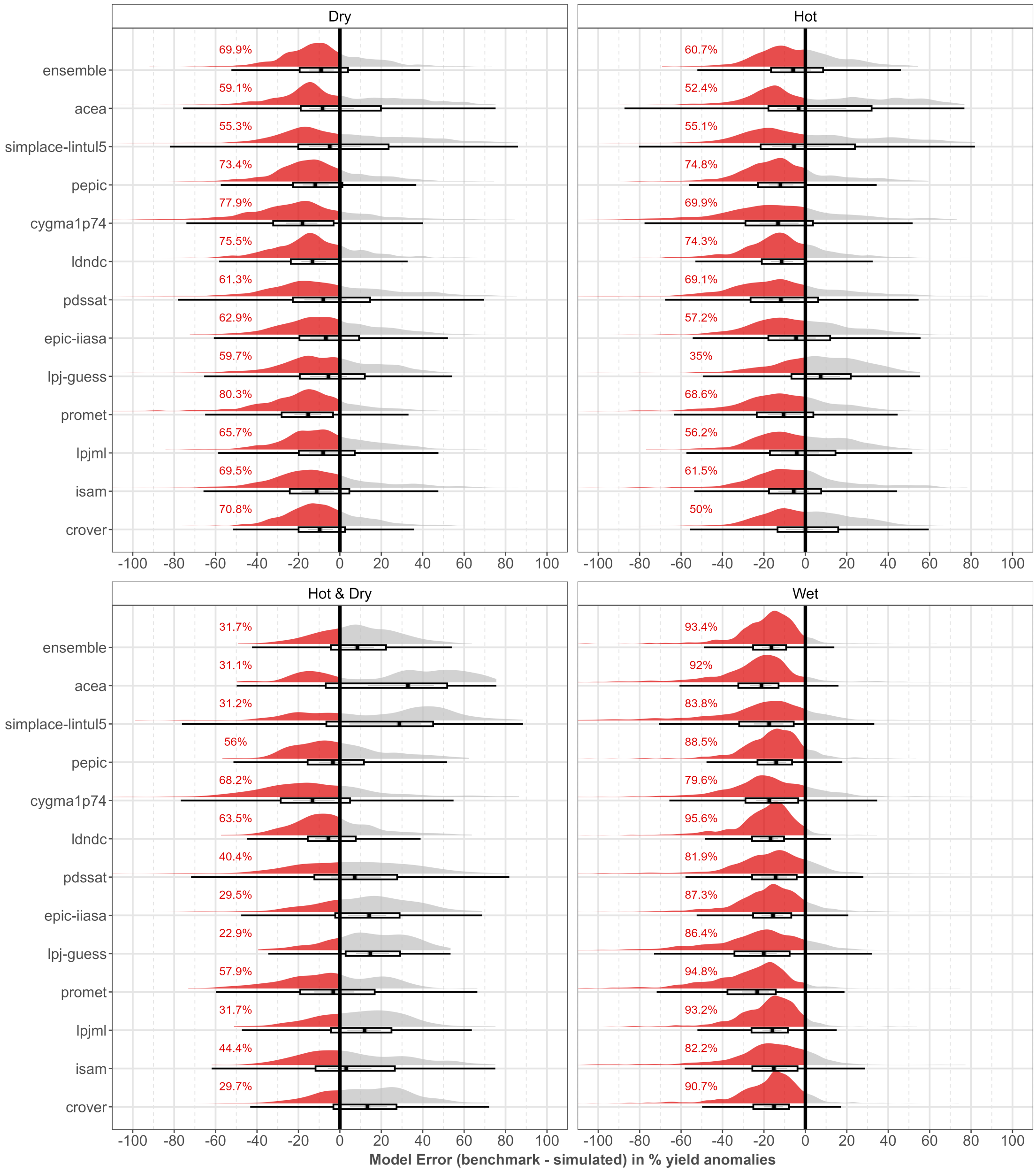
Supplementary Figure 24: Raincloud plot illustrating the distribution of individual and ensemble relative model error values across different extreme events, aggregated over four main crop types: first growing season rice, winter wheat, soy, and maize. The relative model error is quantified as the difference between observed and simulated yield anomalies in percent (i.e. based on relative detrending). Red areas indicate underestimation by the model (percentage denoted in red), while grey areas represent overestimation. Red percentage numbers indicate data points with underestimation. Boxplots below the distribution plots depict the median and interquartile range. For dry, hot, and wet extremes, all models underestimate the impact more frequently than they overestimate it. Notably, wet extremes are the least accurately captured by the models. The ensemble consistently ranks among the lower-performing results across all event types. For compound hot & dry extremes, all models demonstrate reduced underestimation. Models with high variance exhibit a median error close to zero, however, still coupled with strong overestimation.



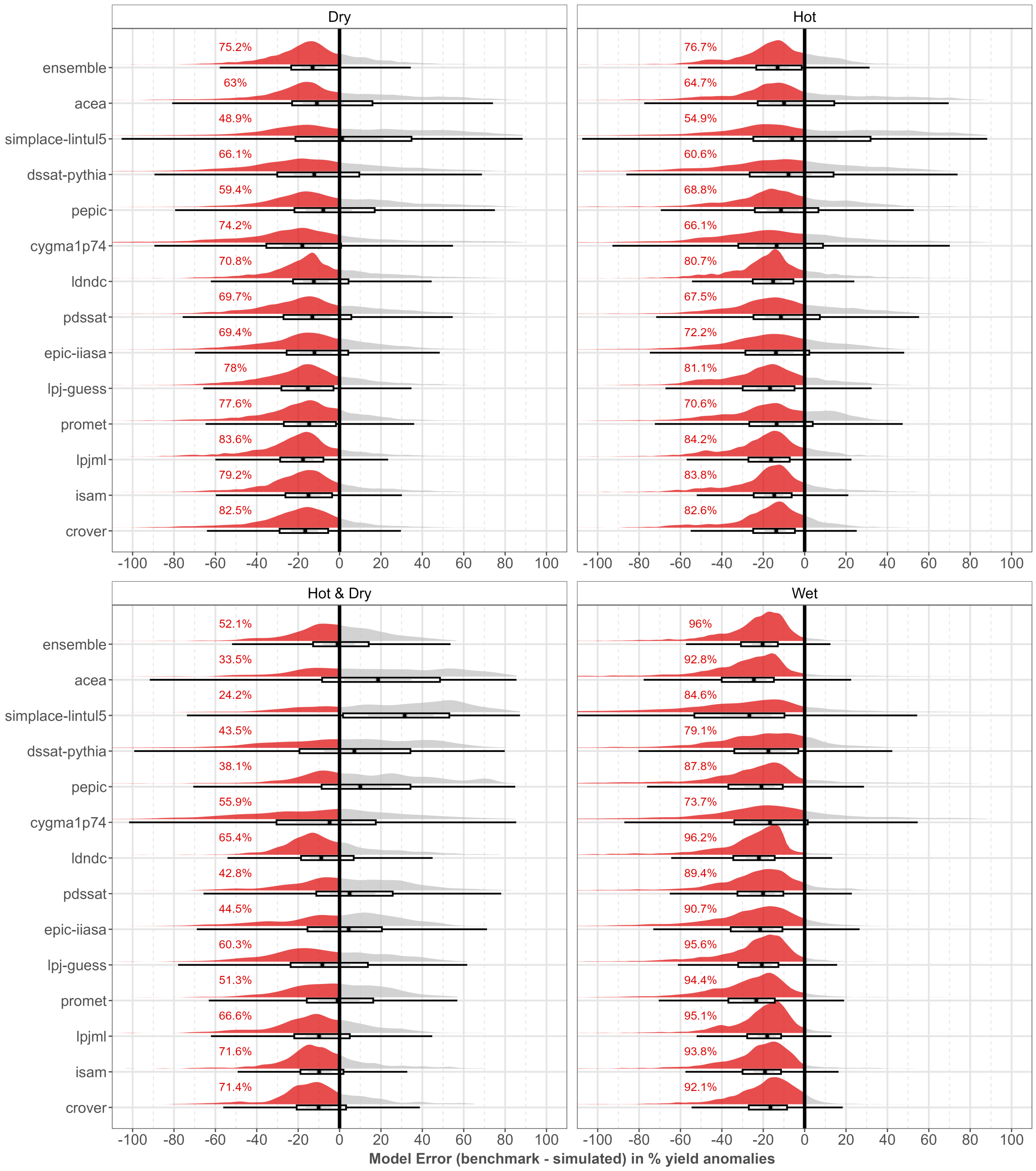
Supplementary Figure 25: RICE (first growing season) – Raincloud plots for model error across extremes (relative detrended anomaly difference: benchmark - model, in %), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. All extreme types lead to a rather homogeneous picture of underestimation. .



Supplementary Figure 26: WINTER WHEAT – Raincloud plots for model error across extremes (relative detrended anomaly difference: benchmark - model, in %), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. Dry, hot, and compound hot & dry extremes lead to a rather homogeneous picture of underestimation. Only simplace-lintul5 and epic-iiasa overestimate more often for compound hot & dry extremes. For wet extremes we get a rather homogeneous picture of underestimation again.



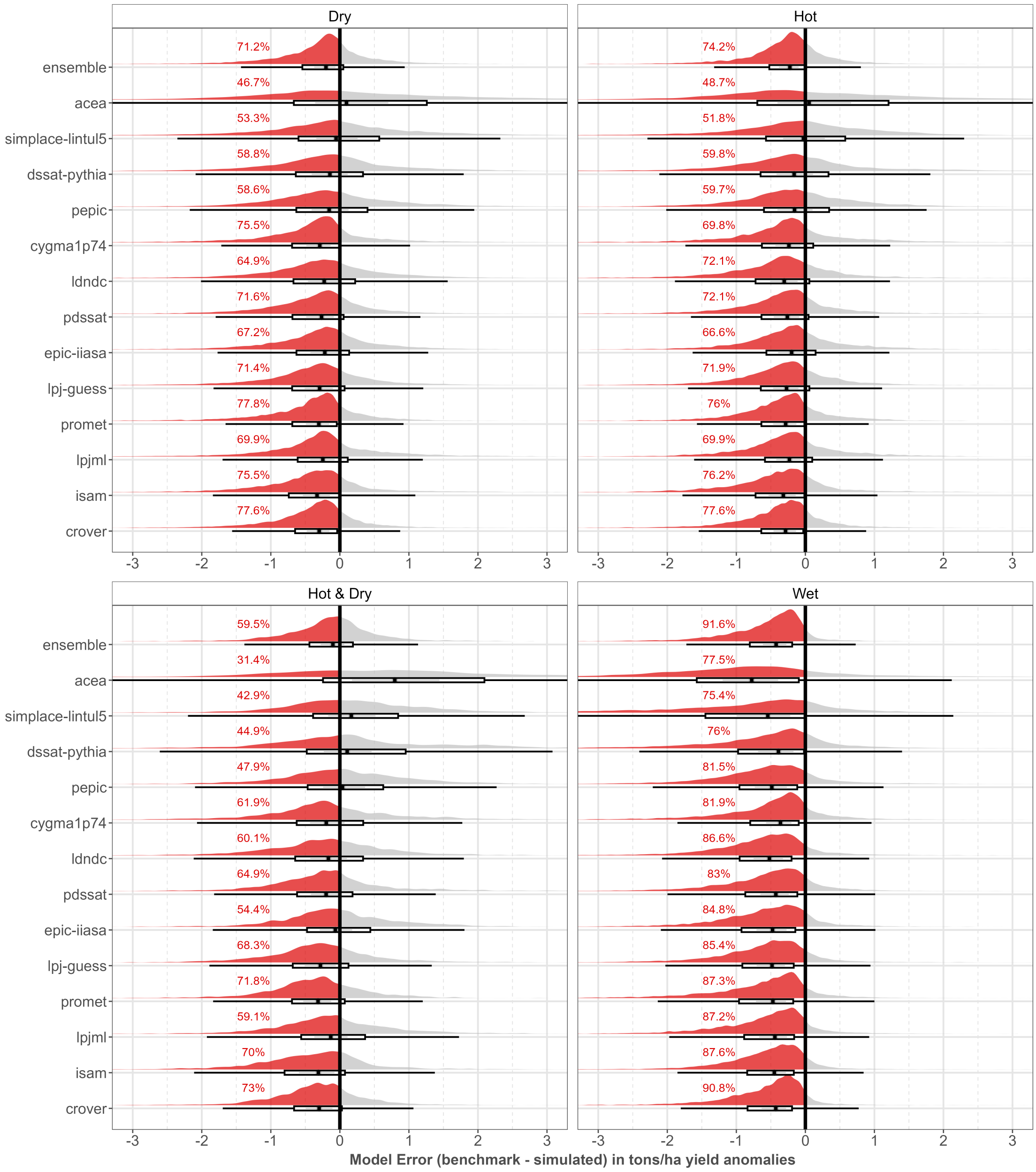
Supplementary Figure 27: SOY – Raincloud plots for model error across extremes (relative detrended anomaly difference: benchmark - model, in %), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. In general, the models show smaller proportions of underestimation. For compound hot & dry extremes, the models more often overestimated than underestimated. For wet extremes we get a rather homogeneous picture of underestimation.



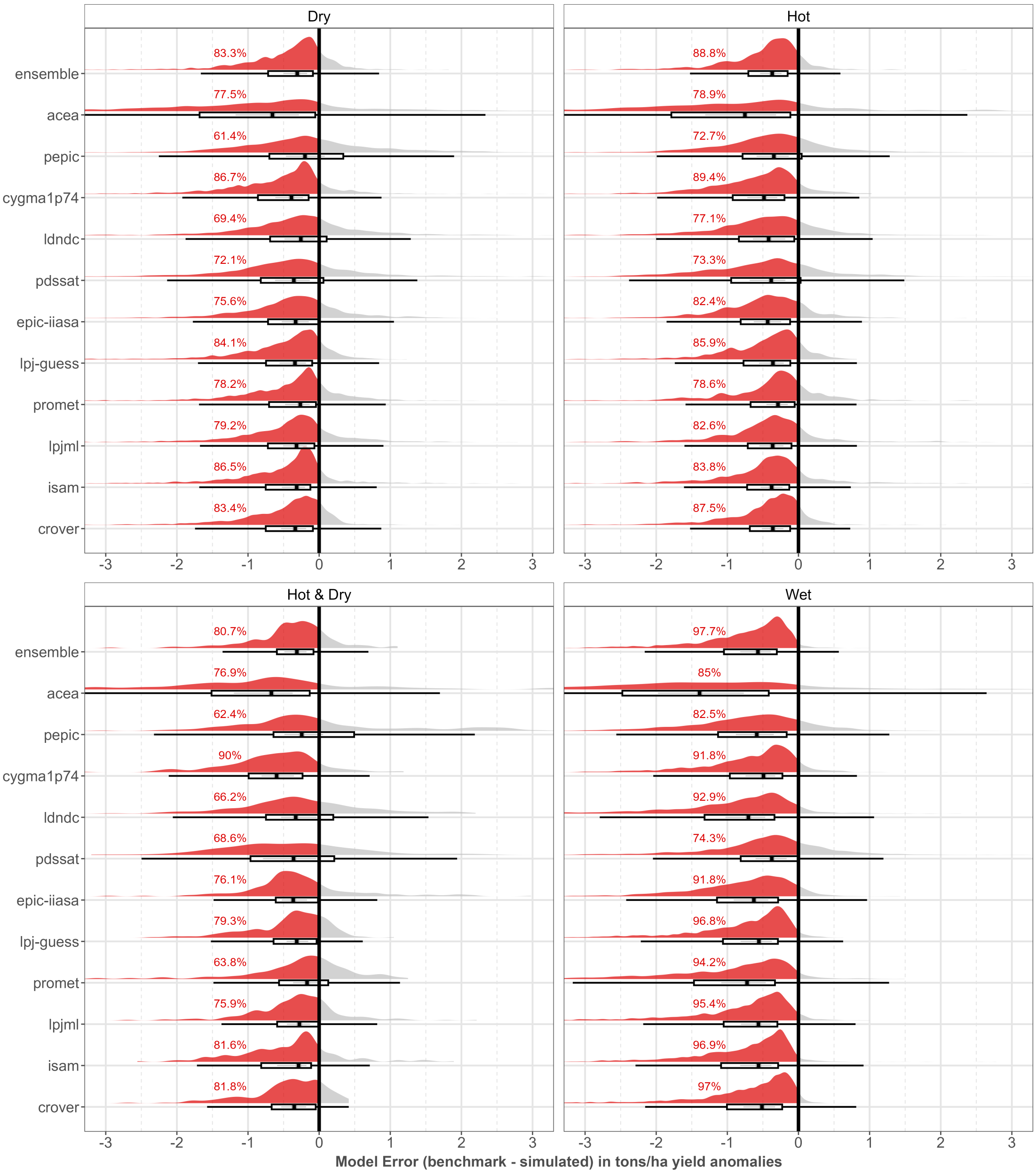
Supplementary Figure 28: MAIZE – Raincloud plots for model error across extremes (relative detrended anomaly difference: benchmark - model, in %), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. dry extremes lead to slightly less underestimated data points than hot extremes. Compound hot & dry extremes lead to a very mixed picture - while some models keep the trend to underestimate the impact others even overestimate it (simplace lintul5, acea, pepic, dssat-pythia, epic-iiasa). The ensemble nearly overestimates (52.1%) as often as it underestimates (47.9%). For wet extremes the amount of underestimated data points ranks from 73.7% for cygma1p74 to 96.2% for the ensemble.

7 Raincloud plots: absolute model error

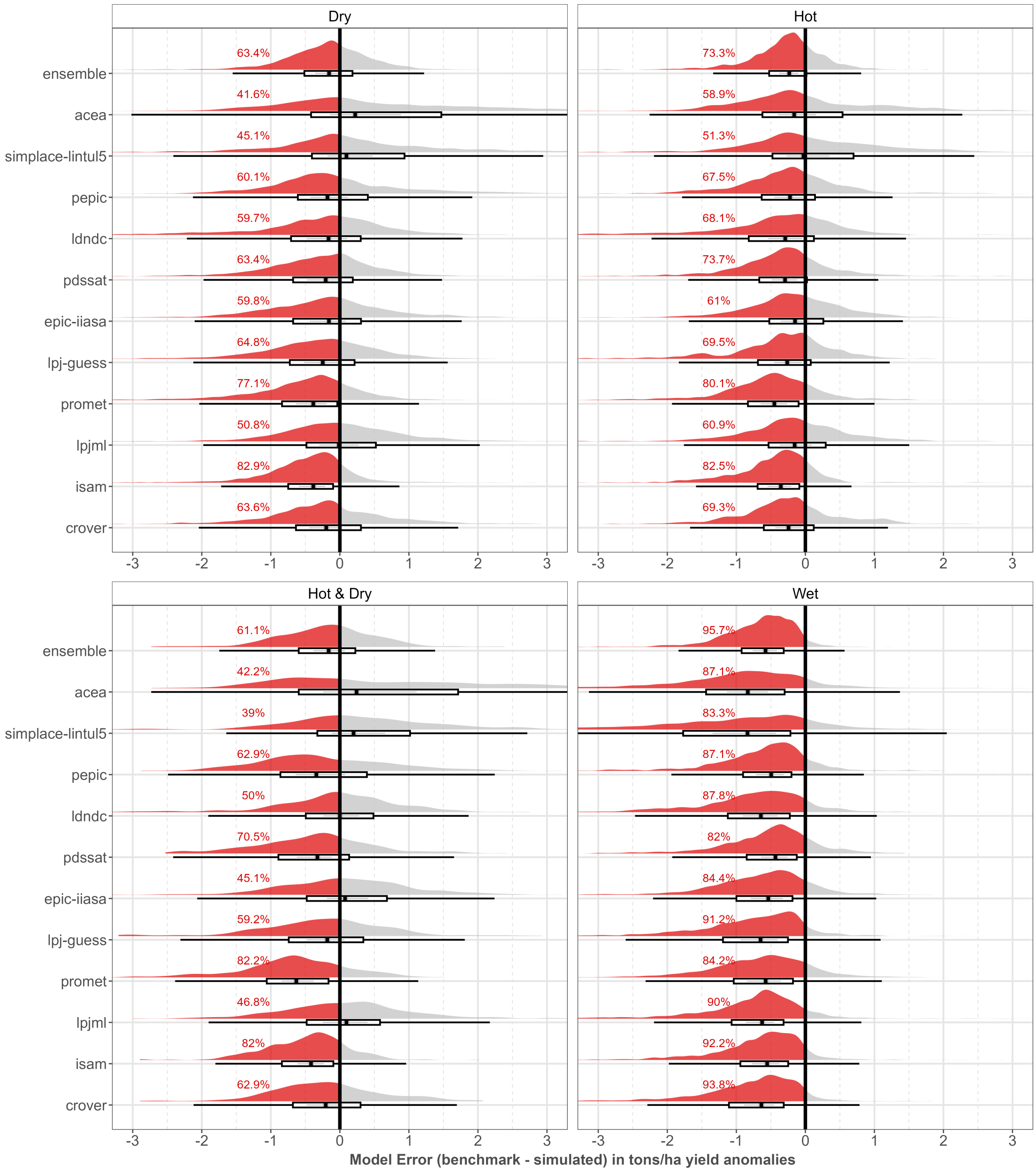
To also depict the extent of the error, we display for four main crop types - first growing season rice, winter wheat, soy, and maize - the distribution of the individual and ensemble absolute model error in the Supplementary Figures 29, 30, 31, 32,33. The absolute model error is quantified as the difference between observed (Iizumi et al. benchmark data GDHY (Iizumi and Sakai [2020])) and simulated yield anomalies in tons/ha (i.e. based on absolute detrending) across the extremes hot, dry, compound hot & dry, and wet. Positive absolute model errors represent an overestimation of the negative impact. Boxplots below the distribution plots depict the median and interquartile range (IQR).



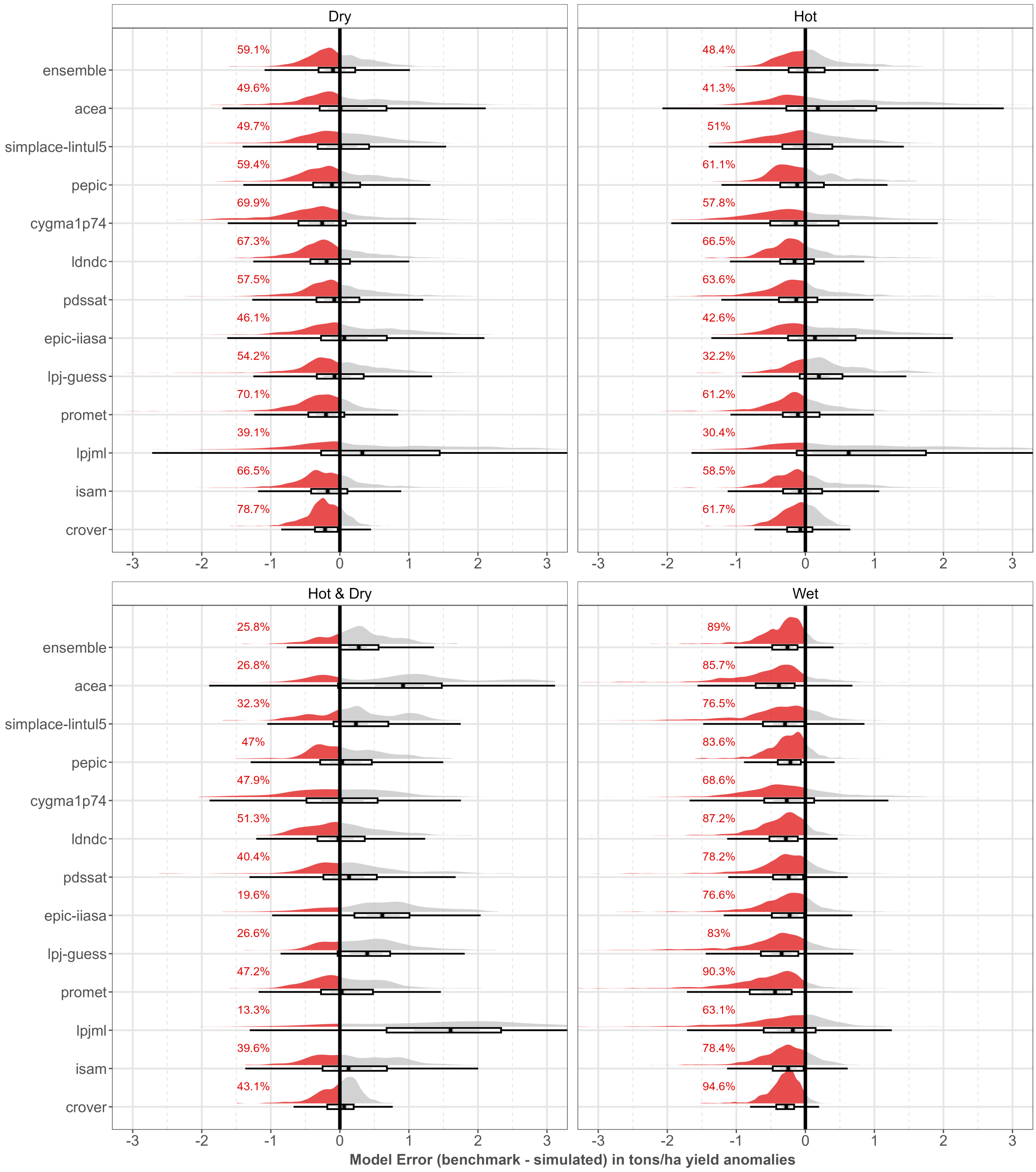
Supplementary Figure 29: Raincloud plots for model error across extremes aggregated for four main crop types soy, maize, winter wheat, and first growing season rice (absolute detrended anomaly difference: benchmark - model, in tons/ha), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. Interestingly, the median absolute error is quite close in case of dry and hot extremes for most models. Simplace-lintul5 is close to zero and acea overestimates more than it underestimates in the median. For compound hot & dry extremes the results are more heterogeneous: some models shift with the median error close to zero (simplace-lintul5, dssat-pythia, pepic, cygma1p74, ldndc, pdssat, epic-iiasa, lpjml).



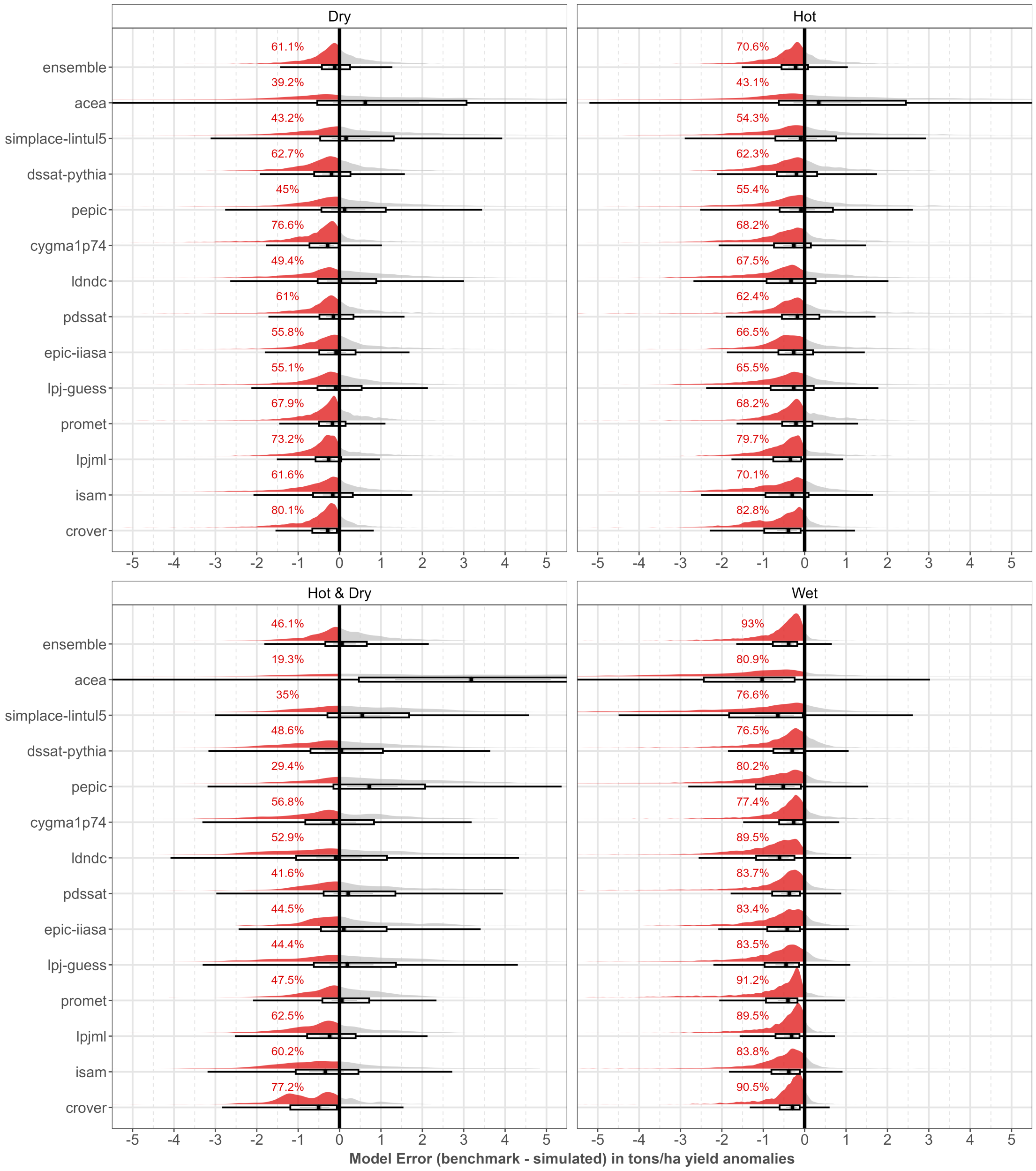
Supplementary Figure 30: RICE (first growing season) – Raincloud plots for model error across extremes (absolute detrended anomaly difference: benchmark - model, in tons/ha), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. For all four extremes the models exhibit a homogeneous picture of underestimation, especially pronounced again for wet extremes. Acea is special out as with a wide range of underestimation errors (IQR) and a higher median error.



Supplementary Figure 31: WINTER WHEAT – Raincloud plots for model error across extremes (absolute detrended anomaly difference: benchmark - model, in tons/ha), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. Compared to the other crops, wheat results look more balanced, and many models have reduced data points of underestimation for dry and hot extremes, which flips for compound hot & dry, here acea, simplace-lintul5, epic-iiasa and lpjml even overestimate the loss more often than they underestimate it. Ldndc is perfectly in the middle. Wet extremes are again consequently underestimated.



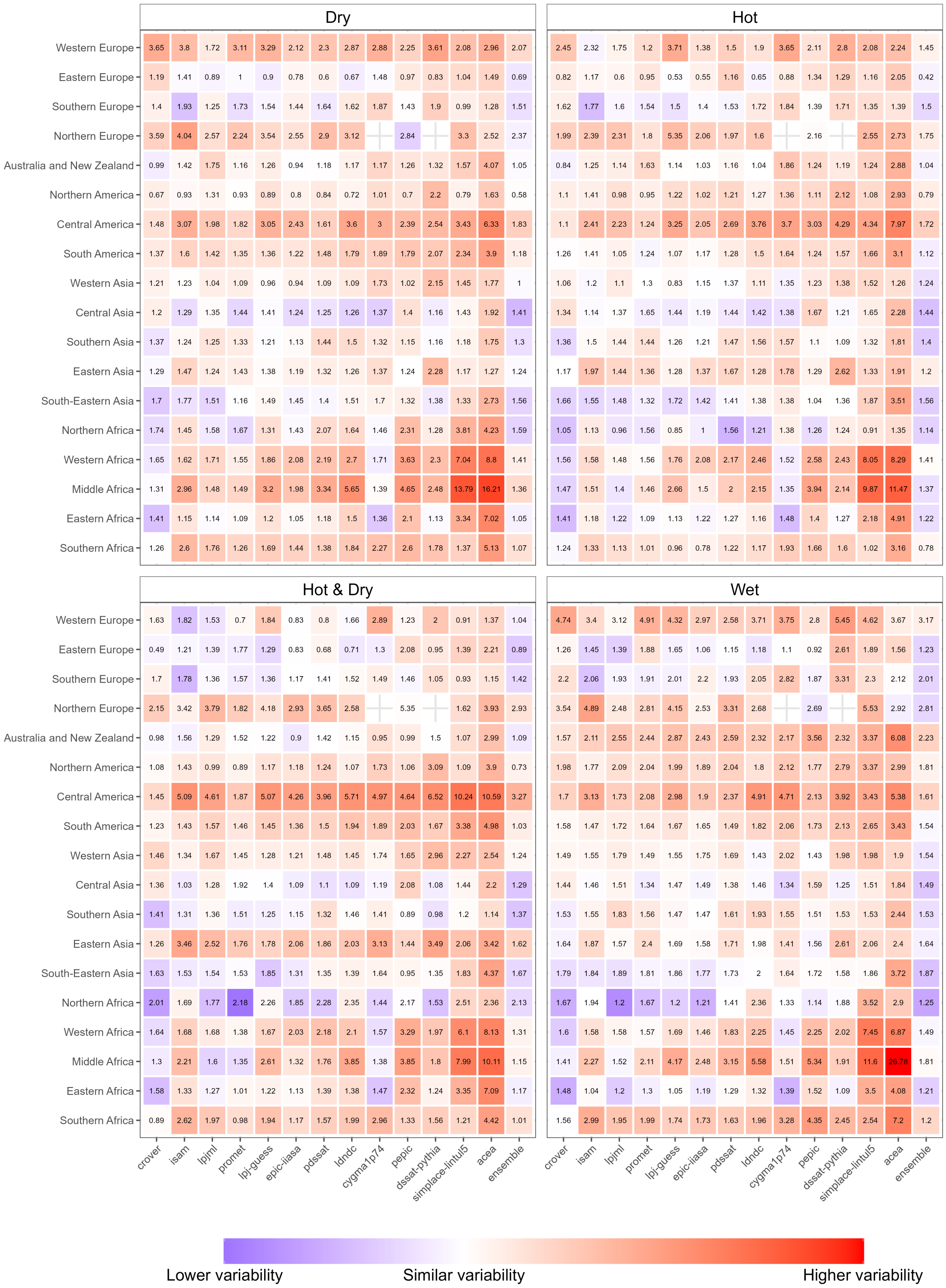
Supplementary Figure 32: SOY – Raincloud plots for model error across extremes (absolute detrended anomaly difference: benchmark - model, in tons/ha), with red=underestimation, gray=overestimation. Boxplots show median/IQR, red percentage numbers indicate data points with underestimation. For soy under dry or hot extremes models have a much lower tendency to underestimate the impact. In the case of compound hot & dry only ldndc underestimates the impact. Wet extremes are again consequently underestimated.



Supplementary Figure 33: MAIZE – Raincloud plots for model error across extremes (absolute detrended anomaly difference: benchmark - model, in tons/ha), with red=underestimation, gray=overestimation. Boxplots show median/IQR. Although nearly all models underestimate dry or hot extremes, over and underestimation are more balanced. In the case of compound hot & dry extremes more models overestimate the extreme. However, the distributions become here exceptionally wide and uncertainty seems high. Wet extremes are again consequently underestimated.

8 Decomposing the Kling-Gupta Efficiency

As the Kling-Gupta efficiency (KGE, Eq. 17) consists of three components: alpha (variability), beta (bias) and r (correlation), we decompose the composite values (see Figure 4 in the Main text) plotting regional heatmaps for alpha (Supplementary Figure 34), beta (Supplementary Figure 35), and r (Supplementary Figure 36) separately. Further, we provide the regional heatmap for the hit rate (Eq. 18) metric (Supplementary Figure 37). Note that all metrics here are based on absolute detrending (Eq. 6). We also create a global map showing to which region each grid cell is allocated (Figure 38).



Dry

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	3.65	3.8	1.72	3.11	3.29	2.12	2.3	2.87	2.88	2.25	3.61	2.08	2.96	2.07
Eastern Europe	1.19	1.41	0.89	1	0.9	0.78	0.6	0.67	1.48	0.97	0.83	1.04	1.49	0.69
Southern Europe	1.4	1.93	1.25	1.73	1.54	1.44	1.64	1.62	1.87	1.43	1.9	0.99	1.28	1.51
Northern Europe	3.59	4.04	2.57	2.24	3.54	2.55	2.9	3.12		2.84		3.3	2.52	2.37
Australia and New Zealand	0.99	1.42	1.75	1.16	1.26	0.94	1.18	1.17	1.17	1.26	1.32	1.57	4.07	1.05
Northern America	0.67	0.93	1.31	0.93	0.89	0.8	0.84	0.72	1.01	0.7	2.2	0.79	1.63	0.58
Central America	1.48	3.07	1.98	1.82	3.05	2.43	1.61	3.6	3	2.39	2.54	3.43	6.33	1.83
South America	1.37	1.6	1.42	1.35	1.36	1.22	1.48	1.79	1.89	1.79	2.07	2.34	3.9	1.18
Western Asia	1.21	1.23	1.04	1.09	0.96	0.94	1.09	1.09	1.73	1.02	2.15	1.45	1.77	1
Central Asia	1.2	1.29	1.35	1.44	1.41	1.24	1.25	1.26	1.37	1.4	1.16	1.43	1.92	1.41
Southern Asia	1.37	1.24	1.25	1.33	1.21	1.13	1.44	1.5	1.32	1.15	1.16	1.18	1.75	1.3
Eastern Asia	1.29	1.47	1.24	1.43	1.38	1.19	1.32	1.26	1.37	1.24	2.28	1.17	1.27	1.24
South-Eastern Asia	1.7	1.77	1.51	1.16	1.49	1.45	1.4	1.51	1.7	1.32	1.38	1.33	2.73	1.56
Northern Africa	1.74	1.45	1.58	1.67	1.31	1.43	2.07	1.64	1.46	2.31	1.28	3.81	4.23	1.59
Western Africa	1.65	1.62	1.71	1.55	1.86	2.08	2.19	2.7	1.71	3.63	2.3	7.04	8.8	1.41
Middle Africa	1.31	2.96	1.48	1.49	3.2	1.98	3.34	5.65	1.39	4.65	2.48	13.79	16.21	1.36
Eastern Africa	1.41	1.15	1.14	1.09	1.2	1.05	1.18	1.5	1.36	2.1	1.13	3.34	7.02	1.05
Southern Africa	1.26	2.6	1.76	1.26	1.69	1.44	1.38	1.84	2.27	2.6	1.78	1.37	5.13	1.07

Hot

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	2.45	2.32	1.75	1.2	3.71	1.38	1.5	1.9	3.65	2.11	2.8	2.08	2.24	1.45
Eastern Europe	0.82	1.17	0.6	0.95	0.53	0.55	1.16	0.65	0.88	1.34	1.29	1.16	2.05	0.42
Southern Europe	1.62	1.77	1.6	1.54	1.5	1.4	1.53	1.72	1.84	1.39	1.71	1.35	1.39	1.5
Northern Europe	1.99	2.39	2.31	1.8	5.35	2.06	1.97	1.6		2.16		2.55	2.73	1.75
Australia and New Zealand	0.84	1.25	1.14	1.63	1.14	1.03	1.16	1.04	1.86	1.24	1.19	1.24	2.88	1.04
Northern America	1.1	1.41	0.98	0.95	1.22	1.02	1.21	1.27	1.36	1.11	2.12	1.08	2.93	0.79
Central America	1.1	2.41	2.23	1.24	3.25	2.05	2.69	3.76	3.7	3.03	4.29	4.34	7.97	1.72
South America	1.26	1.41	1.05	1.24	1.07	1.17	1.28	1.5	1.64	1.24	1.57	1.66	3.1	1.12
Western Asia	1.06	1.2	1.1	1.3	0.83	1.15	1.37	1.11	1.35	1.23	1.38	1.52	1.26	1.24
Central Asia	1.34	1.14	1.37	1.65	1.44	1.19	1.44	1.42	1.38	1.67	1.21	1.65	2.28	1.44
Southern Asia	1.36	1.5	1.44	1.44	1.26	1.21	1.47	1.56	1.57	1.1	1.09	1.32	1.81	1.4
Eastern Asia	1.17	1.97	1.44	1.36	1.28	1.37	1.67	1.28	1.78	1.29	2.62	1.33	1.91	1.2
South-Eastern Asia	1.66	1.55	1.48	1.32	1.72	1.42	1.41	1.38	1.38	1.04	1.36	1.87	3.51	1.56
Northern Africa	1.05	1.13	0.96	1.56	0.85	1	1.56	1.21	1.38	1.26	1.24	0.91	1.35	1.14
Western Africa	1.56	1.58	1.48	1.56	1.76	2.08	2.17	2.46	1.52	2.58	2.43	8.05	8.29	1.41
Middle Africa	1.47	1.51	1.4	1.46	2.66	1.5	2	2.15	1.35	3.94	2.14	9.87	11.47	1.37
Eastern Africa	1.41	1.18	1.22	1.09	1.13	1.22	1.27	1.16	1.48	1.4	1.27	2.18	4.91	1.22
Southern Africa	1.24	1.33	1.13	1.01	0.96	0.78	1.22	1.17	1.93	1.66	1.6	1.02	3.16	0.78

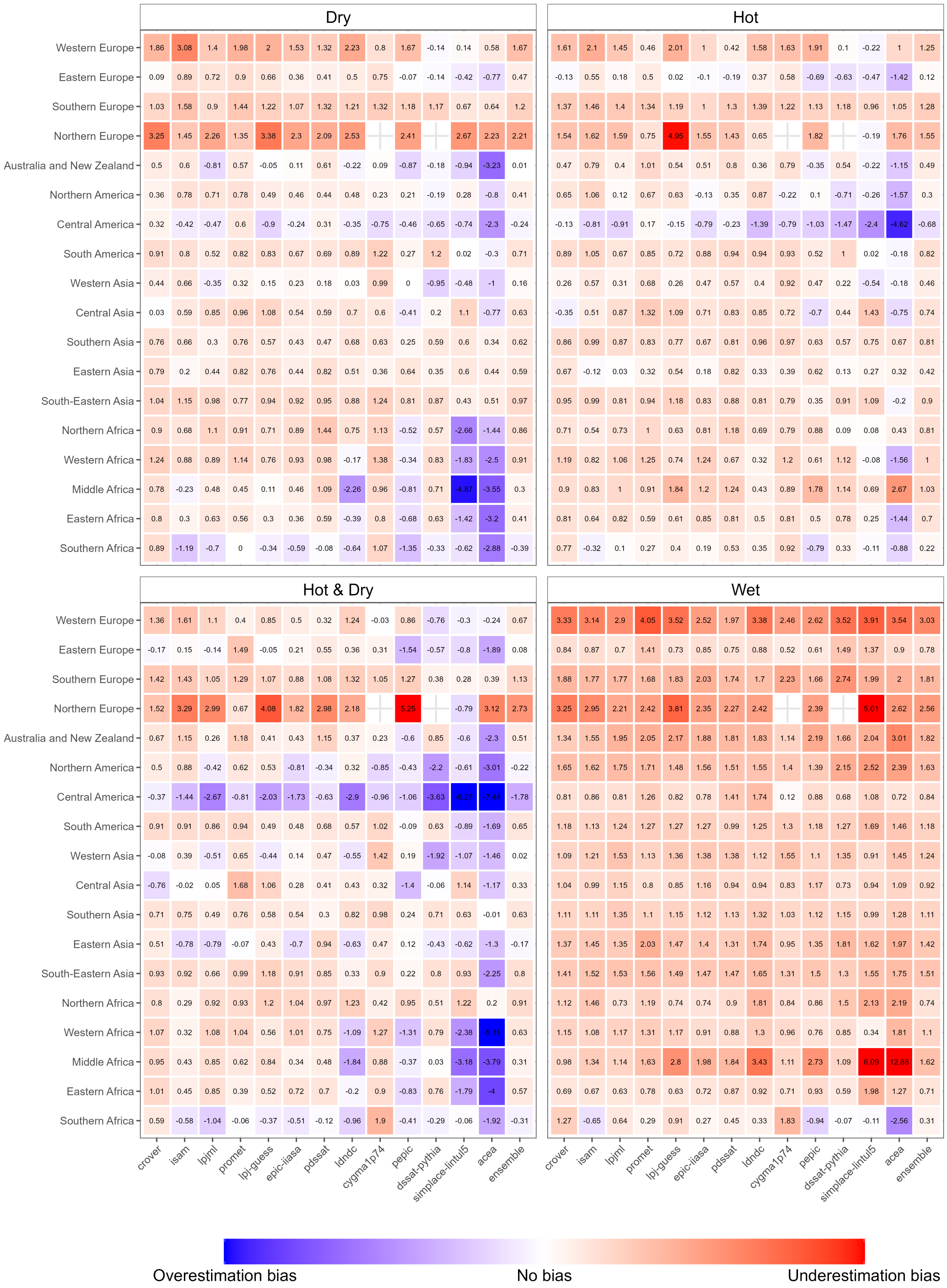
Hot & Dry

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	1.63	1.82	1.53	0.7	1.84	0.83	0.8	1.66	2.89	1.23	2	0.91	1.37	1.04
Eastern Europe	0.49	1.21	1.39	1.77	1.29	0.83	0.68	0.71	1.3	2.08	0.95	1.39	2.21	0.89
Southern Europe	1.7	1.78	1.36	1.57	1.36	1.17	1.41	1.52	1.49	1.46	1.05	0.93	1.15	1.42
Northern Europe	2.15	3.42	3.79	1.82	4.18	2.93	3.65	2.58		5.35		1.62	3.93	2.93
Australia and New Zealand	0.98	1.56	1.29	1.52	1.22	0.9	1.42	1.15	0.95	0.99	1.5	1.07	2.99	1.09
Northern America	1.08	1.43	0.99	0.89	1.17	1.18	1.24	1.07	1.73	1.06	3.09	1.09	3.9	0.73
Central America	1.45	5.09	4.61	1.87	5.07	4.26	3.96	5.71	4.97	4.64	6.52	10.24	10.59	3.27
South America	1.23	1.43	1.57	1.46	1.45	1.36	1.5	1.94	1.89	2.03	1.67	3.38	4.98	1.03
Western Asia	1.46	1.34	1.67	1.45	1.28	1.21	1.48	1.45	1.74	1.65	2.96	2.27	2.54	1.24
Central Asia	1.36	1.03	1.28	1.92	1.4	1.09	1.1	1.09	1.19	2.08	1.08	1.44	2.2	1.29
Southern Asia	1.41	1.31	1.36	1.51	1.25	1.15	1.32	1.46	1.41	0.89	0.98	1.2	1.14	1.37
Eastern Asia	1.26	3.46	2.52	1.76	1.78	2.06	1.86	2.03	3.13	1.44	3.49	2.06	3.42	1.62
South-Eastern Asia	1.63	1.53	1.54	1.53	1.85	1.31	1.35	1.39	1.64	0.95	1.35	1.83	4.37	1.67
Northern Africa	2.01	1.69	1.77	2.18	2.26	1.85	2.28	2.35	1.44	2.17	1.53	2.51	2.36	2.13
Western Africa	1.64	1.68	1.68	1.38	1.67	2.03	2.18	2.1	1.57	3.29	1.97	6.1	8.13	1.31
Middle Africa	1.3	2.21	1.6	1.35	2.61	1.32	1.76	3.85	1.38	3.85	1.8	7.99	10.11	1.15
Eastern Africa	1.58	1.33	1.27	1.01	1.22	1.13	1.39	1.38	1.47	2.32	1.24	3.35	7.09	1.17
Southern Africa	0.89	2.62	1.97	0.98	1.94	1.17	1.57	1.99	2.96	1.33	1.56	1.21	4.42	1.01

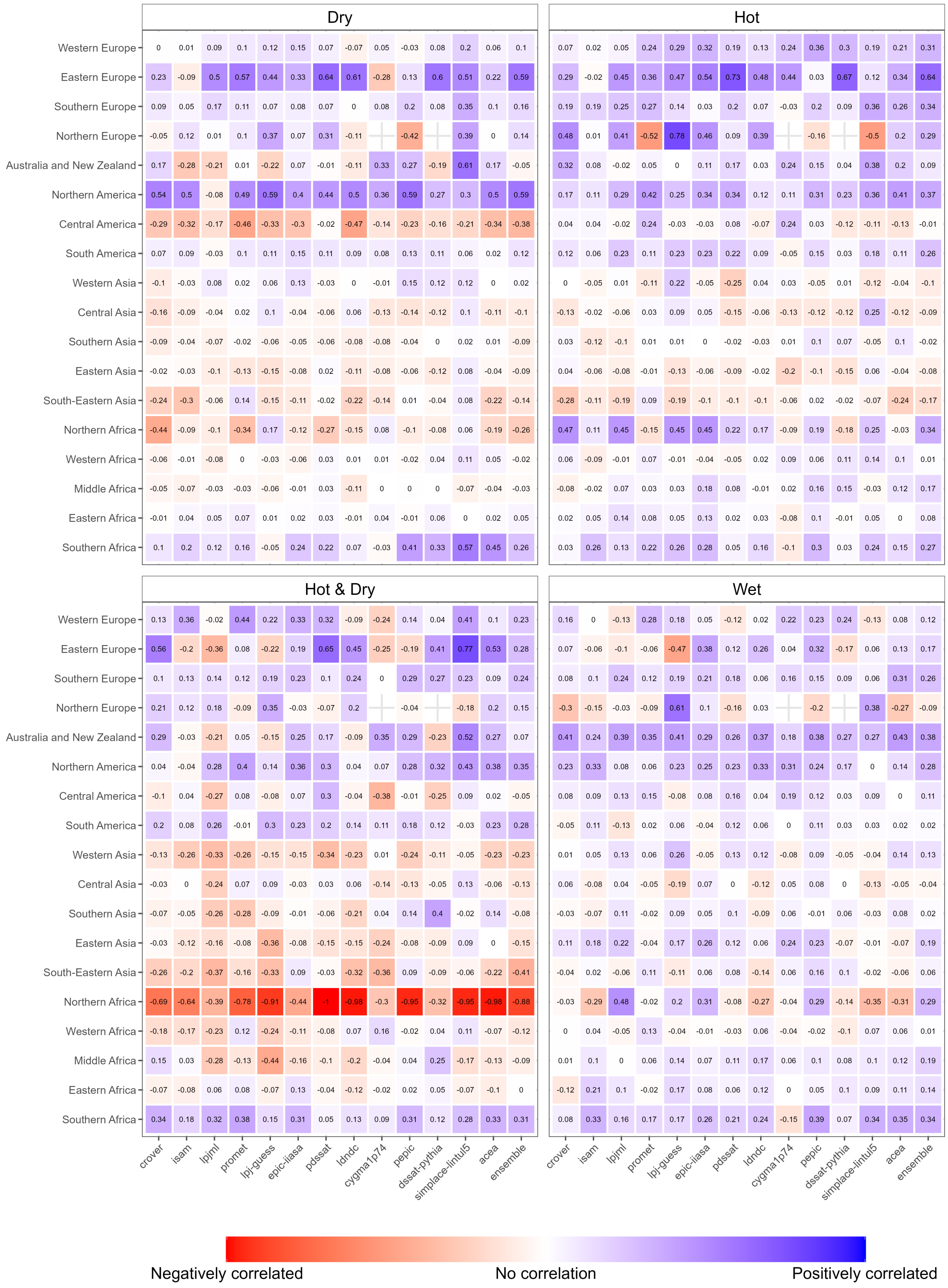
Wet

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	4.74	3.4	3.12	4.91	4.32	2.97	2.58	3.71	3.75	2.8	5.45	4.62	3.67	3.17
Eastern Europe	1.26	1.45	1.39	1.88	1.65	1.06	1.15	1.18	1.1	0.92	2.61	1.89	1.56	1.23
Southern Europe	2.2	2.06	1.93	1.91	2.01	2.2	1.93	2.05	2.82	1.87	3.31	2.3	2.12	2.01
Northern Europe	3.54	4.89	2.48	2.81	4.15	2.53	3.31	2.68		2.69		5.53	2.92	2.81
Australia and New Zealand	1.57	2.11	2.55	2.44	2.87	2.43	2.59	2.32	2.17	3.56	2.32	3.37	6.08	2.23
Northern America	1.98	1.77	2.09	2.04	1.99	1.89	2.04	1.8	2.12	1.77	2.79	3.37	2.99	1.81
Central America	1.7	3.13	1.73	2.08	2.98	1.9	2.37	4.91	4.71	2.13	3.92	3.43	5.38	1.61
South America	1.58	1.47	1.72	1.64	1.67	1.65	1.49	1.82	2.06	1.73	2.13	2.65	3.43	1.54
Western Asia	1.49	1.55	1.79	1.49	1.55	1.75	1.69	1.43	2.02	1.43	1.98	1.98	1.9	1.54
Central Asia	1.44	1.46	1.51	1.34	1.47	1.49	1.38	1.46	1.34	1.59	1.25	1.51	1.84	1.49
Southern Asia	1.53	1.55	1.83	1.56	1.47	1.47	1.61	1.93	1.55	1.51	1.53	1.53	2.44	1.53
Eastern Asia	1.64	1.87	1.57	2.4	1.69	1.58	1.71	1.98	1.41	1.56	2.61	2.06	2.4	1.64
South-Eastern Asia	1.79	1.84	1.89	1.81	1.86	1.77	1.73	2	1.64	1.72	1.58	1.86	3.72	1.87
Northern Africa	1.67	1.94	1.2	1.67	1.2	1.21	1.41	2.36	1.33	1.14	1.88	3.52	2.9	1.25
Western Africa	1.6	1.58	1.58	1.57	1.69	1.46	1.83	2.25	1.45	2.25	2.02	7.45	6.87	1.49
Middle Africa	1.41	2.27	1.52	2.11	4.17	2.48	3.15	5.58	1.51	5.34	1.91	11.6	26.78	1.81
Eastern Africa	1.48	1.04	1.2	1.3	1.05	1.19	1.29	1.32	1.39	1.52	1.09	3.5	4.08	1.21
Southern Africa	1.56	2.99	1.95	1.99	1.74	1.73	1.63	1.96	3.28	4.35	2.45	2.54	7.2	1.2

Supplementary Figure 34: Alpha of Kling Gupta Efficiency indicating model variability. Values ≥ 1 indicate a higher variability in the respective model simulation data than the variability in the benchmark data. The closer to 1, the better the model performance as the variabilities are more equal. We note that the models generally show higher variability than the benchmark data across all regions and extreme types. For Middle Africa, Central America, Western Europe, and Southern Africa we see that many models have a variability that is about 3 times or sometimes even up to 10 times higher than the benchmark data.



Supplementary Figure 35: Beta of Kling-Gupta Efficiency indicating model bias in t/ha. Values close to 0 indicate lower model bias, positive values indicate an underestimation of climate extreme impacts on crops, negative values indicate overestimation. Generally the models tend to underestimate, especially for extreme wetness conditions and except for hot & dry, where we see a more balanced picture of over-and underestimation. We note that across all extremes, the models systematically underestimate climate extremes in (Western) Europe with especially large biases under extreme dry and extreme wet conditions. In contrast with the other models, Acea shows more frequent overestimation than underestimation. Almost all models overestimate in Central America except under wet extremes. In Southern Africa, biases are generally lower, especially under wet extremes the contrast with other regions is remarkable, with some models even overestimating the impacts.



Dry

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	0	0.01	0.09	0.1	0.12	0.15	0.07	-0.07	0.05	-0.03	0.08	0.2	0.06	0.1
Eastern Europe	0.23	-0.09	0.5	0.57	0.44	0.33	0.64	0.61	-0.28	0.13	0.6	0.51	0.22	0.59
Southern Europe	0.09	0.05	0.17	0.11	0.07	0.08	0.07	0	0.08	0.2	0.08	0.35	0.1	0.16
Northern Europe	-0.05	0.12	0.01	0.1	0.37	0.07	0.31	-0.11		-0.42		0.39	0	0.14
Australia and New Zealand	0.17	-0.28	-0.21	0.01	-0.22	0.07	-0.01	-0.11	0.33	0.27	-0.19	0.61	0.17	-0.05
Northern America	0.54	0.5	-0.08	0.49	0.59	0.4	0.44	0.5	0.36	0.59	0.27	0.3	0.5	0.59
Central America	-0.29	-0.32	-0.17	-0.46	-0.33	-0.3	-0.02	-0.47	-0.14	-0.23	-0.16	-0.21	-0.34	-0.38
South America	0.07	0.09	-0.03	0.1	0.11	0.15	0.11	0.09	0.08	0.13	0.11	0.06	0.02	0.12
Western Asia	-0.1	-0.03	0.08	0.02	0.06	0.13	-0.03	0	-0.01	0.15	0.12	0.12	0	0.02
Central Asia	-0.16	-0.09	-0.04	0.02	0.1	-0.04	-0.06	0.06	-0.13	-0.14	-0.12	0.1	-0.11	-0.1
Southern Asia	-0.09	-0.04	-0.07	-0.02	-0.06	-0.05	-0.06	-0.08	-0.08	-0.04	0	0.02	0.01	-0.09
Eastern Asia	-0.02	-0.03	-0.1	-0.13	-0.15	-0.08	0.02	-0.11	-0.08	-0.06	-0.12	0.08	-0.04	-0.09
South-Eastern Asia	-0.24	-0.3	-0.06	0.14	-0.15	-0.11	-0.02	-0.22	-0.14	0.01	-0.04	0.08	-0.22	-0.14
Northern Africa	-0.44	-0.09	-0.1	-0.34	0.17	-0.12	-0.27	-0.15	0.08	-0.1	-0.08	0.06	-0.19	-0.26
Western Africa	-0.06	-0.01	-0.08	0	-0.03	-0.06	0.03	0.01	0.01	-0.02	0.04	0.11	0.05	-0.02
Middle Africa	-0.05	-0.07	-0.03	-0.03	-0.06	-0.01	0.03	-0.11	0	0	0	-0.07	-0.04	-0.03
Eastern Africa	-0.01	0.04	0.05	0.07	0.01	0.02	0.03	-0.01	0.04	-0.01	0.06	0	0.02	0.05
Southern Africa	0.1	0.2	0.12	0.16	-0.05	0.24	0.22	0.07	-0.03	0.41	0.33	0.57	0.45	0.26

Hot

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	0.07	0.02	0.05	0.24	0.29	0.32	0.19	0.13	0.24	0.36	0.3	0.19	0.21	0.31
Eastern Europe	0.29	-0.02	0.45	0.36	0.47	0.54	0.73	0.48	0.44	0.03	0.67	0.12	0.34	0.64
Southern Europe	0.19	0.19	0.25	0.27	0.14	0.03	0.2	0.07	-0.03	0.2	0.09	0.36	0.26	0.34
Northern Europe	0.48	0.01	0.41	-0.52	0.78	0.46	0.09	0.39		-0.16		-0.5	0.2	0.29
Australia and New Zealand	0.32	0.08	-0.02	0.05	0	0.11	0.17	0.03	0.24	0.15	0.04	0.38	0.2	0.09
Northern America	0.17	0.11	0.29	0.42	0.25	0.34	0.34	0.12	0.11	0.31	0.23	0.36	0.41	0.37
Central America	0.04	0.04	-0.02	0.24	-0.03	-0.03	0.08	-0.07	0.24	0.03	-0.12	-0.11	-0.13	-0.01
South America	0.12	0.06	0.23	0.11	0.23	0.23	0.22	0.09	-0.05	0.15	0.03	0.18	0.11	0.26
Western Asia	0	-0.05	0.01	-0.11	0.22	-0.05	-0.25	0.04	0.03	-0.05	0.01	-0.12	-0.04	-0.1
Central Asia	-0.13	-0.02	-0.06	0.03	0.09	0.05	-0.15	-0.06	-0.13	-0.12	-0.12	0.25	-0.12	-0.09
Southern Asia	0.03	-0.12	-0.1	0.01	0.01	0	-0.02	-0.03	0.01	0.1	0.07	-0.05	0.1	-0.02
Eastern Asia	0.04	-0.06	-0.08	-0.01	-0.13	-0.06	-0.09	-0.02	-0.2	-0.1	-0.15	0.06	-0.04	-0.08
South-Eastern Asia	-0.28	-0.11	-0.19	0.09	-0.19	-0.1	-0.1	-0.1	-0.06	0.02	-0.02	-0.07	-0.24	-0.17
Northern Africa	0.47	0.11	0.45	-0.15	0.45	0.45	0.22	0.17	-0.09	0.19	-0.18	0.25	-0.03	0.34
Western Africa	0.06	-0.09	-0.01	0.07	-0.01	-0.04	-0.05	0.02	0.09	0.06	0.11	0.14	0.1	0.01
Middle Africa	-0.08	-0.02	0.07	0.03	0.03	0.18	0.08	-0.01	0.02	0.16	0.15	-0.03	0.12	0.17
Eastern Africa	0.02	0.05	0.14	0.08	0.05	0.13	0.02	0.03	-0.08	0.1	-0.01	0.05	0	0.08
Southern Africa	0.03	0.26	0.13	0.22	0.26	0.28	0.05	0.16	-0.1	0.3	0.03	0.24	0.15	0.27

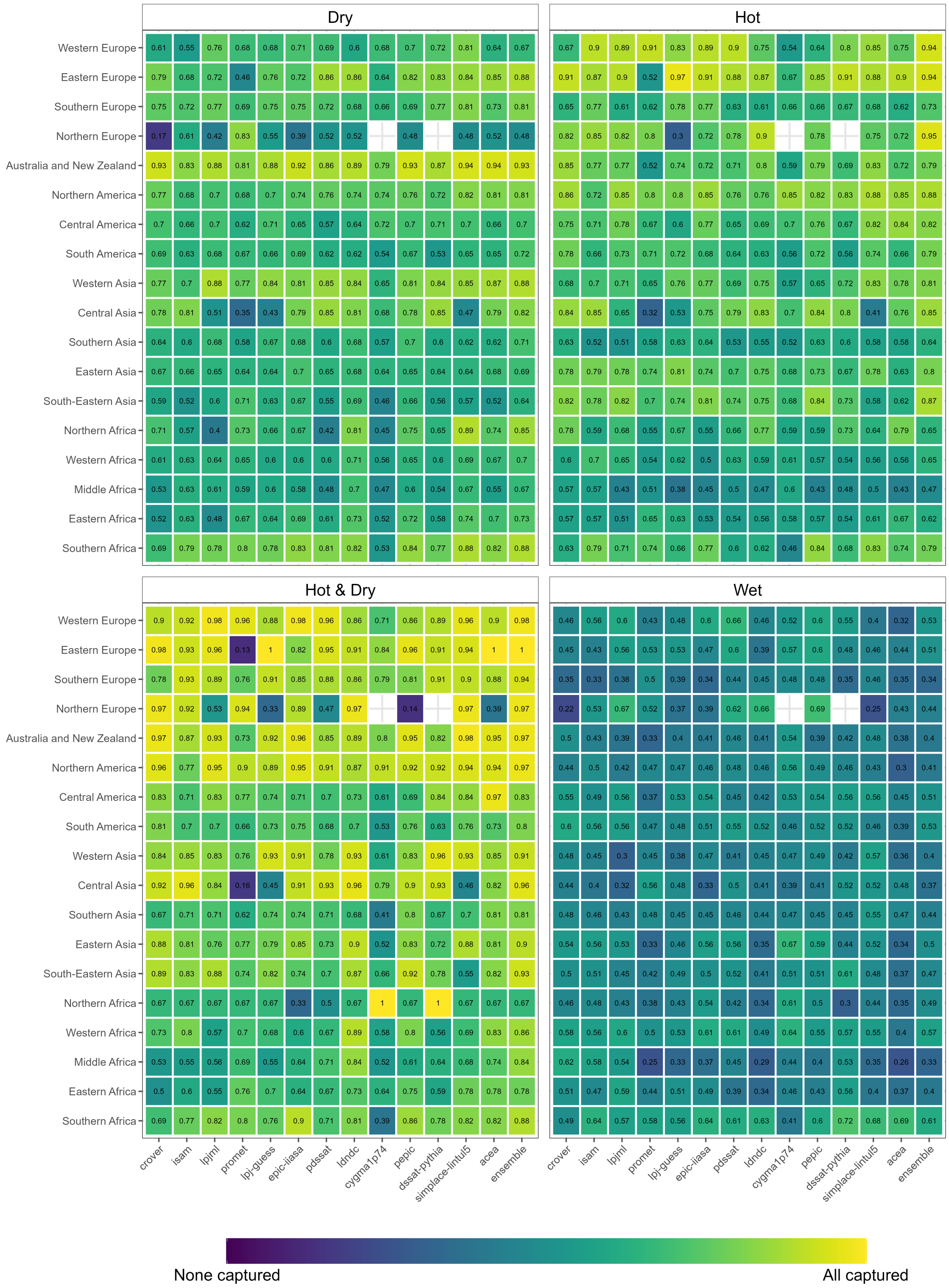
Hot & Dry

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	0.13	0.36	-0.02	0.44	0.22	0.33	0.32	-0.09	-0.24	0.14	0.04	0.41	0.1	0.23
Eastern Europe	0.56	-0.2	-0.36	0.08	-0.22	0.19	0.65	0.45	-0.25	-0.19	0.41	0.77	0.53	0.28
Southern Europe	0.1	0.13	0.14	0.12	0.19	0.23	0.1	0.24	0	0.29	0.27	0.23	0.09	0.24
Northern Europe	0.21	0.12	0.18	-0.09	0.35	-0.03	-0.07	0.2		-0.04		-0.18	0.2	0.15
Australia and New Zealand	0.29	-0.03	-0.21	0.05	-0.15	0.25	0.17	-0.09	0.35	0.29	-0.23	0.52	0.27	0.07
Northern America	0.04	-0.04	0.28	0.4	0.14	0.36	0.3	0.04	0.07	0.28	0.32	0.43	0.38	0.35
Central America	-0.1	0.04	-0.27	0.08	-0.08	0.07	0.3	-0.04	-0.38	-0.01	-0.25	0.09	0.02	-0.05
South America	0.2	0.08	0.26	-0.01	0.3	0.23	0.2	0.14	0.11	0.18	0.12	-0.03	0.23	0.28
Western Asia	-0.13	-0.26	-0.33	-0.26	-0.15	-0.15	-0.34	-0.23	0.01	-0.24	-0.11	-0.05	-0.23	-0.23
Central Asia	-0.03	0	-0.24	0.07	0.09	-0.03	0.03	0.06	-0.14	-0.13	-0.05	0.13	-0.06	-0.13
Southern Asia	-0.07	-0.05	-0.26	-0.28	-0.09	-0.01	-0.06	-0.21	0.04	0.14	0.4	-0.02	0.14	-0.08
Eastern Asia	-0.03	-0.12	-0.16	-0.08	-0.36	-0.08	-0.15	-0.15	-0.24	-0.08	-0.09	0.09	0	-0.15
South-Eastern Asia	-0.26	-0.2	-0.37	-0.16	-0.33	0.09	-0.03	-0.32	-0.36	0.09	-0.09	-0.06	-0.22	-0.41
Northern Africa	-0.69	-0.64	-0.39	-0.78	-0.91	-0.44	-1	-0.98	-0.3	-0.95	-0.32	-0.95	-0.98	-0.88
Western Africa	-0.18	-0.17	-0.23	0.12	-0.24	-0.11	-0.08	0.07	0.16	-0.02	0.04	0.11	-0.07	-0.12
Middle Africa	0.15	0.03	-0.28	-0.13	-0.44	-0.16	-0.1	-0.2	-0.04	0.04	0.25	-0.17	-0.13	-0.09
Eastern Africa	-0.07	-0.08	0.06	0.08	-0.07	0.13	-0.04	-0.12	-0.02	0.02	0.05	-0.07	-0.1	0
Southern Africa	0.34	0.18	0.32	0.38	0.15	0.31	0.05	0.13	0.09	0.31	0.12	0.28	0.33	0.31

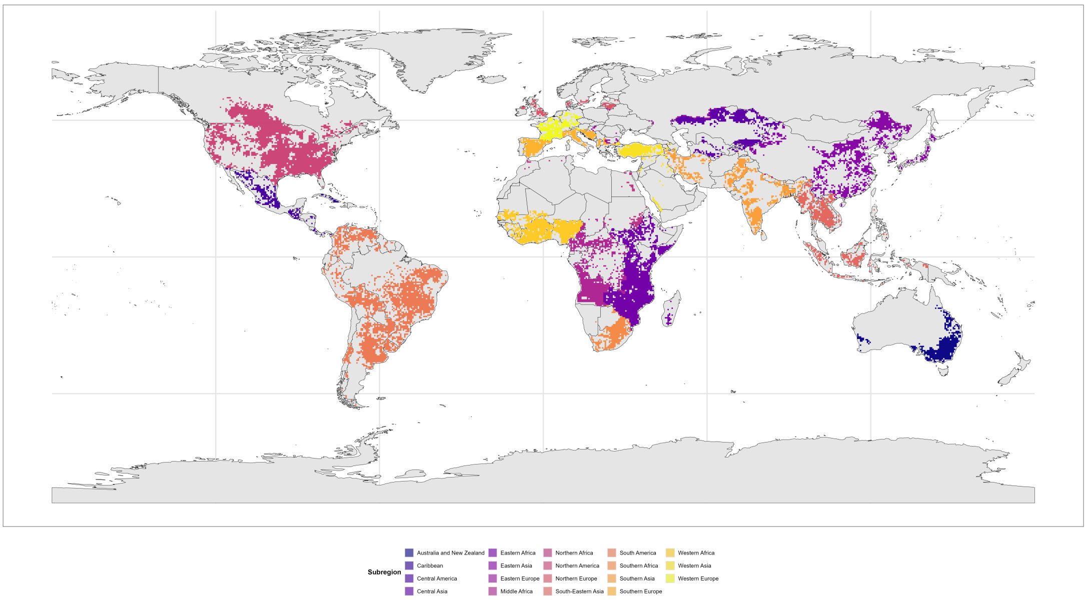
Wet

	crover	isam	lpjml	promet	lpj-guess	epic-iiasa	pdssat	ldndc	cygma1p74	pepic	dssat-pythia	simplace-lintul5	acea	ensemble
Western Europe	0.16	0	-0.13	0.28	0.18	0.05	-0.12	0.02	0.22	0.23	0.24	-0.13	0.08	0.12
Eastern Europe	0.07	-0.06	-0.1	-0.06	-0.47	0.38	0.12	0.26	0.04	0.32	-0.17	0.06	0.13	0.17
Southern Europe	0.08	0.1	0.24	0.12	0.19	0.21	0.18	0.06	0.16	0.15	0.09	0.06	0.31	0.26
Northern Europe	-0.3	-0.15	-0.03	-0.09	0.61	0.1	-0.16	0.03		-0.2		0.38	-0.27	-0.09
Australia and New Zealand	0.41	0.24	0.39	0.35	0.41	0.29	0.26	0.37	0.18	0.38	0.27	0.27	0.43	0.38
Northern America	0.23	0.33	0.08	0.06	0.23	0.25	0.23	0.33	0.31	0.24	0.17	0	0.14	0.28
Central America	0.08	0.09	0.13	0.15	-0.08	0.08	0.16	0.04	0.19	0.12	0.03	0.09	0	0.11
South America	-0.05	0.11	-0.13	0.02	0.06	-0.04	0.12	0.06	0	0.11	0.03	0.03	0.02	0.02
Western Asia	0.01	0.05	0.13	0.06	0.26	-0.05	0.13	0.12	-0.08	0.09	-0.05	-0.04	0.14	0.13
Central Asia	0.06	-0.08	0.04	-0.05	-0.19	0.07	0	-0.12	0.05	0.08	0	-0.13	-0.05	-0.04
Southern Asia	-0.03	-0.07	0.11	-0.02	0.09	0.05	0.1	-0.09	0.06	-0.01	0.06	-0.03	0.08	0.02
Eastern Asia	0.11	0.18	0.22	-0.04	0.17	0.26	0.12	0.06	0.24	0.23	-0.07	-0.01	-0.07	0.19
South-Eastern Asia	-0.04	0.02	-0.06	0.11	-0.11	0.06	0.08	-0.14	0.06	0.16	0.1	-0.02	-0.06	0.06
Northern Africa	-0.03	-0.29	0.48	-0.02	0.2	0.31	-0.08	-0.27	-0.04	0.29	-0.14	-0.35	-0.31	0.29
Western Africa	0	0.04	-0.05	0.13	-0.04	-0.01	-0.03	0.06	-0.04	-0.02	-0.1	0.07	0.06	0.01
Middle Africa	0.01	0.1	0	0.06	0.14	0.07	0.11	0.17	0.06	0.1	0.08	0.1	0.12	0.19
Eastern Africa	-0.12	0.21	0.1	-0.02	0.17	0.08	0.06	0.12	0	0.05	0.1	0.09	0.11	0.14
Southern Africa	0.08	0.33	0.16	0.17	0.17	0.26	0.21	0.24	-0.15	0.39	0.07	0.34	0.35	0.34

Supplementary Figure 36: r of Kling-Gupta Efficiency indicating correlation between benchmark and simulated data. In Northern America, (Eastern) Europe, and Southern Africa we observe the most positive correlation values. The simulation data are negatively correlated with the benchmark observations in Northern Africa, especially under compound hot & dry extremes.



Supplementary Figure 37: Hit rate (dimensionless fraction) indicating the frequency of climate extremes captured by the respective model i.e. events where the model shows at least a negative detrended yields. The models typically capture about 70% of the dry or hot events, while less than 50% of the wet events. In general, the models seem to capture hot events slightly better than dry events. For compound hot & dry events we notice even higher hit rates close to 100%, except for regions close to the equator such as South America, Middle Africa, and Southern Asia. In general, the models show large regional differences with some models (e.g. ldndc under dry events) having the highest hit rates for a certain region (Western and Eastern Africa) but the lowest for another region (Western Europe). Regions where all models show consistently high hit rates are Australia & New Zealand for dry events and Northern America for compound hot & dry events.



Supplementary Figure 38: Regions used for the regional heatmaps

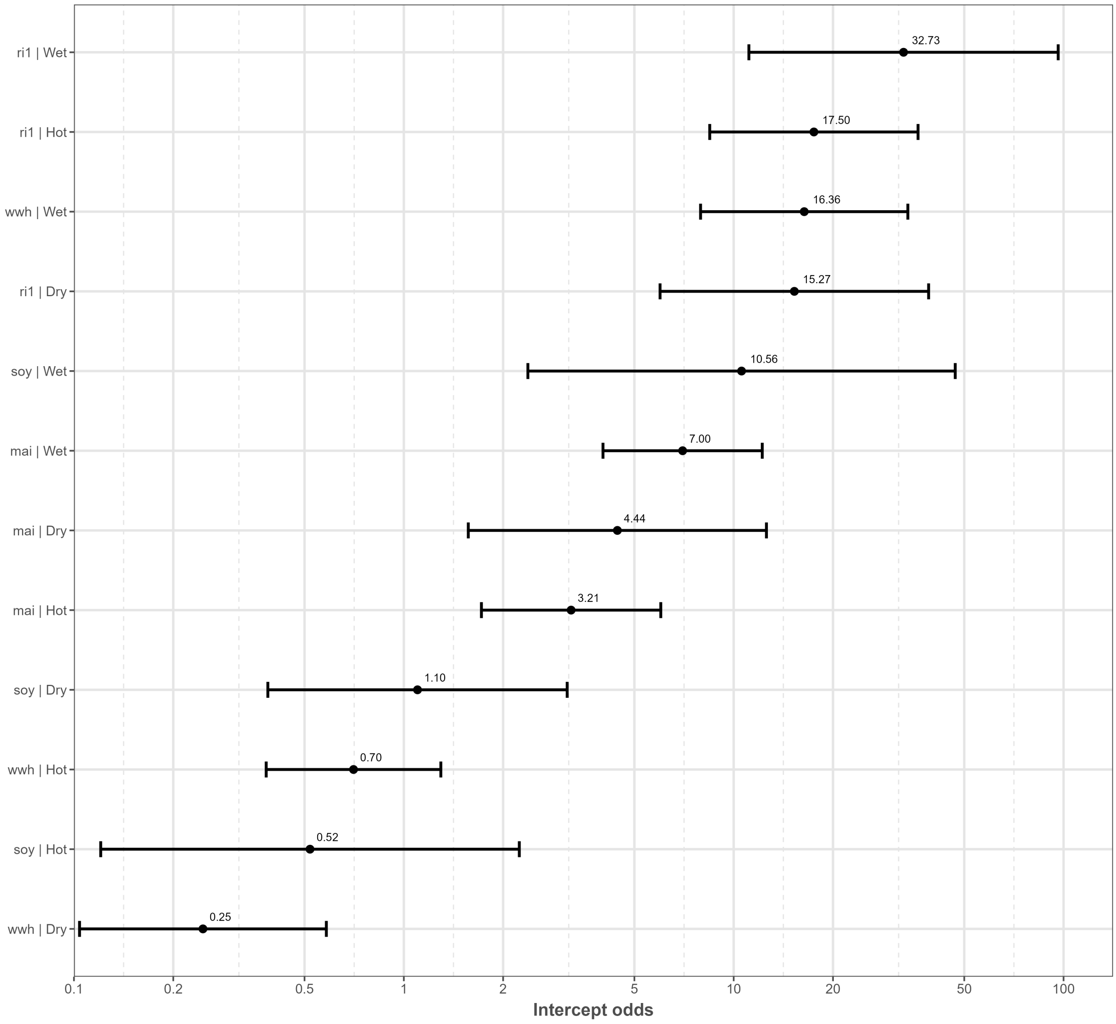
9 Model characteristics analysis: baseline (intercept) odds from GLMMs

In Supplementary Figure 39 we visualize the baseline (intercept) odds of underestimating yield losses under climatic extremes estimated from the Generalized Linear Mixed Models (GLMMs) we fitted separately for each crop (maize, rice, wheat, and soy) and different climatic extreme types (dry, wet, hot). See Eq. 22 for more details on how we defined the GLMM. The visualized intercept odds should be used along with the odds ratios shown in Figure 5 in the Main text to infer the absolute effects of the model characteristics on the odds of underestimation. From the intercept odds, one can calculate the probability of underestimation per crop and extreme type under the hypothetical absence of any model characteristic. For instance, for soybean under extreme wet conditions, a crop model without any of the analyzed characteristics is estimated to have an odds of underestimation of 10.56. This corresponds to a probability of underestimation of:

$$p_{\text{baseline}} = \frac{10.56}{10.56 + 1} = 91.3\%$$

Now suppose the presence of a model characteristic, for example, the inclusion of soil carbon and nitrogen cycling, which has an estimated odds ratio of 0.39. This means the odds decrease by a factor of 0.39. The new probability of underestimation is then:

$$p_{\text{soilCN}} = \frac{10.56 \times 0.39}{1 + (10.56 \times 0.39)} = 80.5\%$$



Supplementary Figure 39: Forest plot of the intercept (baseline) odds of underestimating the yield losses under specific climatic extremes inferred from the Generalized Linear Mixed Model (GLMM, Eq. 22). The errorbars represent the 90% confidence intervals and the numbers above the dot the mean point estimate. The odds values indicate the probability of underestimation relative to the probability of not underestimating the impacts of climatic extremes in the hypothetical absence of any of the crop model characteristics analyzed in this study. The odds ratios presented in Figure 5 from the Main analysis, indicate the estimated factors by which these baseline odds would change when including the corresponding model characteristic. The odds of underestimation is highest for rice under extreme wet conditions, while wheat under extreme dry conditions shows the lowest odds and hence probability of underestimation.

10 Methods

11 Model characteristics

Table 1 summarizes model characteristics relevant for simulating crop development and yield under climate change following e.g. tables in Müller et al. [2017] S1-S4. In this paper, the information on the characteristics is coded as follows: ‘1‘ indicates the characteristic is implemented, ‘0‘ means it is not included, and ‘-1‘ denotes the characteristic is not relevant for the model.

Soil carbon and nitrogen cycling

Simulating soil carbon and nitrogen (CN) cycling is crucial for representing nutrient availability and long-term soil fertility, which directly influence crop growth and yield.

- 1: Soil carbon and nitrogen cycling is explicitly simulated, allowing for dynamic nutrient feedbacks on crop growth.

- 0: Soil carbon and nitrogen cycling is not simulated; nutrient availability is statically or externally prescribed.

Waterlogging stress
Waterlogging occurs when excess soil water limits oxygen availability to roots, leading to reduced crop growth and potential yield loss.

- 1: Waterlogging stress is simulated, reducing crop growth under excessive soil moisture conditions.
- 0: Waterlogging is not simulated; crops do not experience stress from excess water.

Explicit photosynthesis
We distinguish two approaches of modeling light utilization: Photosynthesis-based (PS) models simulate physiological processes directly, while radiation use efficiency (RUE) models use empirical relationships.

- 1: Light utilization is modeled through explicit photosynthesis processes, capturing detailed physiological responses.
- 0: Light utilization is modeled using empirical radiation use efficiency relationships.
- -1: Not applicable; the model does not simulate this process.

Complex temperature stress
Temperature stress can impact crops in multiple ways, from basic reductions in photosynthesis and biomass to more complex effects on reproduction, phenology, and yield components.

- 1: Complex temperature stress is simulated, affecting multiple processes such as reproduction, phenology, and harvest index.
- 0: Only simple temperature stress is simulated, impacting photosynthesis and biomass accumulation.

Variable roots
Root development can be sensitive to temperature, influencing the plant's ability to access water and nutrients. Further, water availability (under drought or waterlogging conditions) can limit root expansion and function. Both, temperature and water stress are key drivers of root growth.

- 1: Root development responds dynamically to temperature or water stress. These signals affect root growth rates, distribution and function.
- 0: Root development is influenced by other stresses, but not by temperature or water stress.
- -1: Root development is not explicitly modeled in relation to stress.

Dynamic harvest modeling
The harvest index (HI) determines the proportion of biomass allocated to yield, and can be fixed or dynamically adjusted in response to environmental and management factors.

- 1: Harvest index and biomass allocation are dynamically simulated, responding to environmental and management changes.
- 0: Harvest index is fixed.
- -1: Not explicitly modeled.

Model	Soil carbon & nitrogen cycling	Waterlogging stress	Explicit photosynthesis	Complex temperature stress	Variable roots	Dynamic harvest modeling
acea	0	1	-1	1	1	1
crover	0	0	1	0	-1	0
cygma1p74	0	1	0	0	0	1
dssat-pythia	1	1	0	0	1	1
epic-iiasa	1	1	0	0	1	1
isam	1	0	1	1	1	1
ldndc	1	0	1	0	0	1
lpj-guess	1	0	1	0	0	1
lpjml	1	0	1	0	0	0
pepic	1	1	0	0	1	1
promet	0	0	1	1	1	1
simplace-lintul5	1	0	0	0	1	1
pdssat	1	0	1	1	1	1

Supplementary Table 1: Model comparison across key characteristics, as described in Subsection 11, for the applied ISIMIP3a model setup ISIMIP. Values indicate: 1 = characteristic is present; 0 = alternative representation; −1 = characteristic not applicable for the model.

References

Toshichika Iizumi and Toru Sakai. The global dataset of historical yields for major crops 1981–2016. *Scientific Data*, 7(1), March 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0433-7. URL http://dx.doi.org/10.1038/s41597-020-0433-7.

Christoph Müller, Joshua Elliott, James Chryssanthacopoulos, Almut Arneth, Juraj Balkovic, Philippe Ciais, Delphine Deryng, Christian Folberth, Michael Glotter, Steven Hoek, Toshichika Iizumi, Roberto C. Izaurralde, Curtis Jones, Nikolay Khabarov, Peter Lawrence, Wenfeng Liu, Stefan Olin, Thomas A. M. Pugh, Deepak K. Ray, Ashwan Reddy, Cynthia Rosenzweig, Alex C. Ruane, Gen Sakurai, Erwin Schmid, Rastislav Skalsky, Carol X. Song, Xuhui Wang, Allard de Wit, and Hong Yang. Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. *Geoscientific Model Development*, 10(4):1403–1422, April 2017. ISSN 1991-959X. doi: 10.5194/gmd-10-1403-2017. URL https://gmd.copernicus.org/articles/10/1403/2017/. Publisher: Copernicus GmbH.