Explaining climatic drivers of yield anomalies in global crop models through metamodel-based attribution

Thomas Oberleitner[1]*; Artem Baklanov[1]; Thiago Berton Ferreira[2]; Gerrit Hoogenboom[2]; Jonas Jägermeyr[3]; Atul Jain[4]; Tzu-Shun Lin[5]; Oleksandr Mialyk[6]; Christoph Müller[7]; Alex C. Ruane[8]; Florian Zabel[9]; Juraj Balkovič[1]; Chenzhi Wang[10]; Babacar Faye[11]; Jose R. Guarin[8]; Toshichika Iizumi[12]; Nikolay Khabarov[1]; Wenfeng Liu[13]; Masashi Okada[14]; Sam S. Rabin[5]; Clemens Scheer[15]; Rastislav Skalský[1]; Christian Folberth[1]

[1] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[2] University of Florida, USA
[3] Columbia University, USA
[4] University of Illinois, Urbana-Champaign, USA
[5] NSF National Center for Atmospheric Research (NCAR), USA
[6] University of Twente, The Netherlands
[7] Potsdam Institute for Climate Impact Research (PIK), Germany
[8] Climate Impacts Group, NASA Goddard Institute for Space Studies (GISS), USA
[9] Universität Basel, Departement Umweltwissenschaften, Switzerland
[10] Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF) e.V., Germany
[11] University of Sine Saloum El Hadj Ibrahima NIASS, Department of Environment, Biodiversity and Sustainable Development, Senegal
[12] National Agriculture and Food Research Organization (NARO), Japan
[13] China Agricultural University, Center for Agricultural Water Research in China, China
[14] National Institute for Environmental Studies (NIES), Japan
[15] IMK-IFU, Karlsruhe Institute of Technology (KIT), Germany
* Corresponding author

Abstract

Global gridded crop models (GGCMs) are important tools for assessing climate impacts on agriculture, yet significant divergence in their projections limits interpretability, and impact studies often treat GGCMs as black boxes. Targeted ensemble sensitivity analyses are demanding and not transferable to different ensembles. Here, we comprehensively evaluate climatic and soil drivers of crop yield anomalies in a state-of-the-art GGCM ensemble, using maize as a representative crop. Gradient boosting classifiers detect anomalies, SHapley Additive exPlanations (SHAP) values quantify feature importance, and methods are applied to a recent GGCM experiment driven by reanalysis climate data. We find broadly similar climatic drivers across the ensemble, though feature importance distributions differ. Low precipitation dominates under rainfed conditions, while solar radiation typically ranks second, highlighting that drought impacts depend on atmospheric water demand often omitted from sensitivity analyses. In some GGCMs, excess rather than insufficient water drives anomalies. With irrigation, low solar radiation or adverse temperatures become the main drivers. In (semi-)arid regions, some GGCMs respond more to cool conditions, others to warm ones. Soil features usually rank lowest but can be moderately important in some models. Our findings demonstrate that evaluating opportunistic data—experiments produced for other purposes—yields vital insights into GGCM divergence in impact studies. Code is publicly available on GitHub to support future attribution analyses and inform broad audiences about drivers of observed results.

Key points

- We present a toolset for evaluating anomaly drivers in model ensembles based on SHAP value distributions and novel visualization methods.
- Results indicate that solar radiation, low temperatures, and excess water are thus far neglected climatic drivers in some regions and models.
- While some drivers dominate the ensemble, most GGCMs show characteristic feature importances for specific drivers and regions.

Plain Language Summary

Computer models are often used to study how climate affects crop production worldwide. These models, called global gridded crop models (GGCMs), sometimes give very different results, which makes it hard to understand and compare their predictions. Usually, studies do not explain why models disagree, and detailed sensitivity tests are hard to compute and specific to models. In this study, we analyzed an existing set of GGCM results for maize to determine which weather and soil conditions cause unusually low yields. We used machine-learning methods to detect these yield anomalies and to measure the importance of different climate and soil factors. We found that while most models agree on the general role of climate drivers, they differ in how strongly they weigh each one. For example, lack of rainfall is usually the main driver under rainfed conditions, followed by solar radiation, showing that drought is influenced not only by rainfall but also by atmospheric water demand. Conversely, in some models, too much water is the main problem. Soil properties usually matter less but can be important in certain models or regions. Our findings show that existing datasets can already be used to explain why crop models disagree, without running new, resource-intensive experiments.

1. Introduction

Global gridded crop models (GGCMs) are typically a combination of a process-based core model that estimates crop growth, yield formation, and a varying range of agro-ecosystem processes, and a spatial computational framework that provides input data for each pixel in a defined region. Similar approaches are implemented in ecosystem models (Müller et al., 2019). Over the past decades, GGCMs have become key tools in global and large-scale agricultural climate impact assessments (Balkovič et al., 2014; Frieler et al., 2017; Jägermeyr et al., 2021; Rosenzweig et al., 2014; Schewe et al., 2019; Schleussner et al., 2018), provide input data for agro-economic and land-use change studies (e.g., Molina Bacca et al., 2023; Orlov et al., 2024) and inform policy-making processes (Schmidt-Traub et al., 2019). They also fill a critical role within the Agricultural Model Intercomparison and Improvement Project (AgMIP) contributing to model intercomparison and climate impact ensemble studies (Rosenzweig et al., 2013; Ruane et al., 2017).

With a growing number of GGCMs being applied and apparent disagreement in their projections (Müller et al. 2021, Müller et al 2024), ensemble studies have emerged as an approach to harmonizing forcing data and scenarios and thereby rendering divergence in outcomes subject to differences in model processes and setups (Elliott et al., 2015; Folberth et al., 2019; Franke et al., 2019; Frieler et al., 2024; Müller et al., 2021). While such ensemble studies have been found to improve robustness in outcomes compared to observations (Martre et al., 2015), climate impact studies still show tremendous deviations among GGCM responses to increasingly altered climate and atmospheric conditions (Jägermeyr et al., 2020, 2021; Rosenzweig et al., 2014) with often limited agreement even on the direction of change in parts of the world.

Although atmospheric CO_2 concentration ($[CO_2]$) has been identified as a key driver in this divergence under high concentration scenarios, using counterfactual scenarios (Jägermeyr et al., 2021), various studies have found large discrepancies in GGCM responses to high temperatures, drought, or extreme wetness. Most often, they compared GGCM responses to observations as a form of benchmarking. For example, Schauberger et al. (2017) evaluated GGCM yield responses to high temperature against observed yields in the US and found overall good agreement but a large spread among models. Similarly, (Li et al., 2019) compared GGCM outcomes to US yield records for extreme precipitation impacts and found very mixed responses, but an overall underestimation. The most comprehensive evaluation of GGCM sensitivities to climate, $[CO_2]$, and nutrient supply, thus far, has been performed by (Müller et al., 2024). The authors used a cube of global systematic perturbations in temperature, precipitation, $[CO_2]$, and N fertilizer inputs, which revealed again a large divergence in GGCM sensitivities – even for GGCMs based on the same or a closely related core model. Yet, the fact that the experiment these evaluations were based on was highly demanding, with > 700 global simulations per crop for the full set of perturbations (Franke et al., 2020), and that it cannot be transferred to the latest developments in GGCMs, ensembles, and experiments (Jägermeyr et al., 2021), highlights that approaches to sensitivity analysis or feature importance attribution are required that can be applied ad hoc to GGCM experiments as these are performed.

In research on observation-based crop yield-weather relationships, a diverse range of methods has been applied over the years, with a recent shift towards explainable machine-learning approaches. (Ben-Ari et al., 2018) evaluated an extreme wheat yield shock in France, focusing on compound events, and used logistic regressions for the quantification of drivers. While cold shocks have received overall little attention, (Xiao et al., 2018) use linear regression to quantify the impact of spring frosts on wheat yield losses in China. More recently, (Zhu et al., 2021) trained Random Forest (RF) models as classifiers for wheat yield shocks in Europe and combined these with SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to identify their climatic drivers. This has also

been proposed as a more general approach to further process-understanding in geosciences (Jiang et al., 2024).

In this study, we assess the sensitivities of GGCMs to climatic and non-climatic features driving crop yield anomalies, using similar methodological approaches but applied to simulations rather than observations. Crop yield anomalies are defined as occurrences less than or equal -15% from the detrended mean. We train classifier models per GGCM to predict yield anomalies for major Köppen-Geiger climate regions and subsequently evaluate feature importances using SHAP values. We define sets of growing season climate features that may cause anomalies either through (I) transient effects (e.g., sum of growing season precipitation) or (II) extremes (e.g., fraction of heating degree days), and cover all types of adverse weather – hot, cold, dry, and wet. In this first assessment, we focus on the transient effects as these are agnostic to potentially model-specific thresholds and cover all relevant climate variables. For solar radiation, for example, no extreme indicator has been defined thus far, but it is key in understanding the role of atmospheric water demand for droughts (Gebrechorkos et al., 2025). We use the set of extreme features as a source for secondary evaluations to assess their importance for yield anomalies. Albeit global data on crop yield anomalies have frequently been shown not to be driven by weather only (Cottrell et al., 2019; Vogel et al., 2019; Wei et al., 2023) we include observations in our evaluation to put our findings in context.

2. Methods

2.1. Study design and data

The study design and analytical approach are presented in Figure 1. In short, we train eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) classifiers to predict yield anomalies occurring in crop yield simulations of GGCMs. We then evaluate the XGBoost models' feature importances and their interactions using SHAP (Lundberg and Lee, 2017) to identify drivers of crop yield anomalies. Further details are provided in the subsequent sections. Model versions and key references are provided in Supplementary Table S 1. For reproducibility and further use, the Python code corresponding to this pipeline is available on GitHub (https://github.com/iiasa/ggcm-feature-importance).

All data were obtained at, or harmonized to, a spatial resolution of 0.5° x 0.5° (approx. 55 km x 55 km near the equator) and for the period 1971-2015 in the case of simulated crop yields and climate data. We use climate and soil data for feature importance attribution that were used as forcings in GGCM simulations. Crop yield estimates were sourced from 13 GGCMs of the Global Gridded Crop Model Intercomparison Project (GGCMI) contributing to the phase 3a simulation ensemble (Jägermeyr et al., 2021). Historical reanalysis climate data (GSWP3-W5E5, Cucchi et al., 2020; Lange, 2019) were provided by the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) (Frieler et al., 2024). Precipitation, surface downwelling shortwave radiation, and minimum and maximum temperature were selected as climate variables ubiquitously used in all GGCMs and aggregated over the growing season (GS) to produce generic or extreme features (see sect. 2.3). Further explanatory features are soil attributes (sand, silt, organic carbon, and available water capacity) reflecting texture and hydrologic characteristics based on the Harmonized World Soil Database v1.2 (FAO et al., 2012; Volkholz and Müller, 2020).

Absolute yields were detrended for the time series per pixel, and relative yields below -15% from the detrended mean are defined as anomalies (see section 2.2). This serves as the training and test set for XGBoost classification models with classes *anomaly* or *no anomaly* (see section 2.4). The

resulting metamodels form the basis for the calculation of SHAP values for individual features (see section 2.5) as well as interactions (see section 2.6).

As crop yield-climate relationships and resulting anomalies can differ substantially between broad climate domains, metamodels were trained by major Köppen-Geiger classes, namely A (tropical), B (arid), C (temperate), D (cold), and E (polar), based on Beck et al. (2018). To concentrate the analysis on regions relevant for crop cultivation, we remove pixels without harvested area for a particular crop, based on the Spatial Production Allocation Model (SPAM) 2010 v2r0 (International Food Policy Research Institute, 2020; Yu et al., 2020) in line with earlier ensemble studies (Jägermeyr et al., 2021). Herein, we focus on maize as a ubiquitously grown model crop and include soybean as a contrasting crop in the supplementary information.

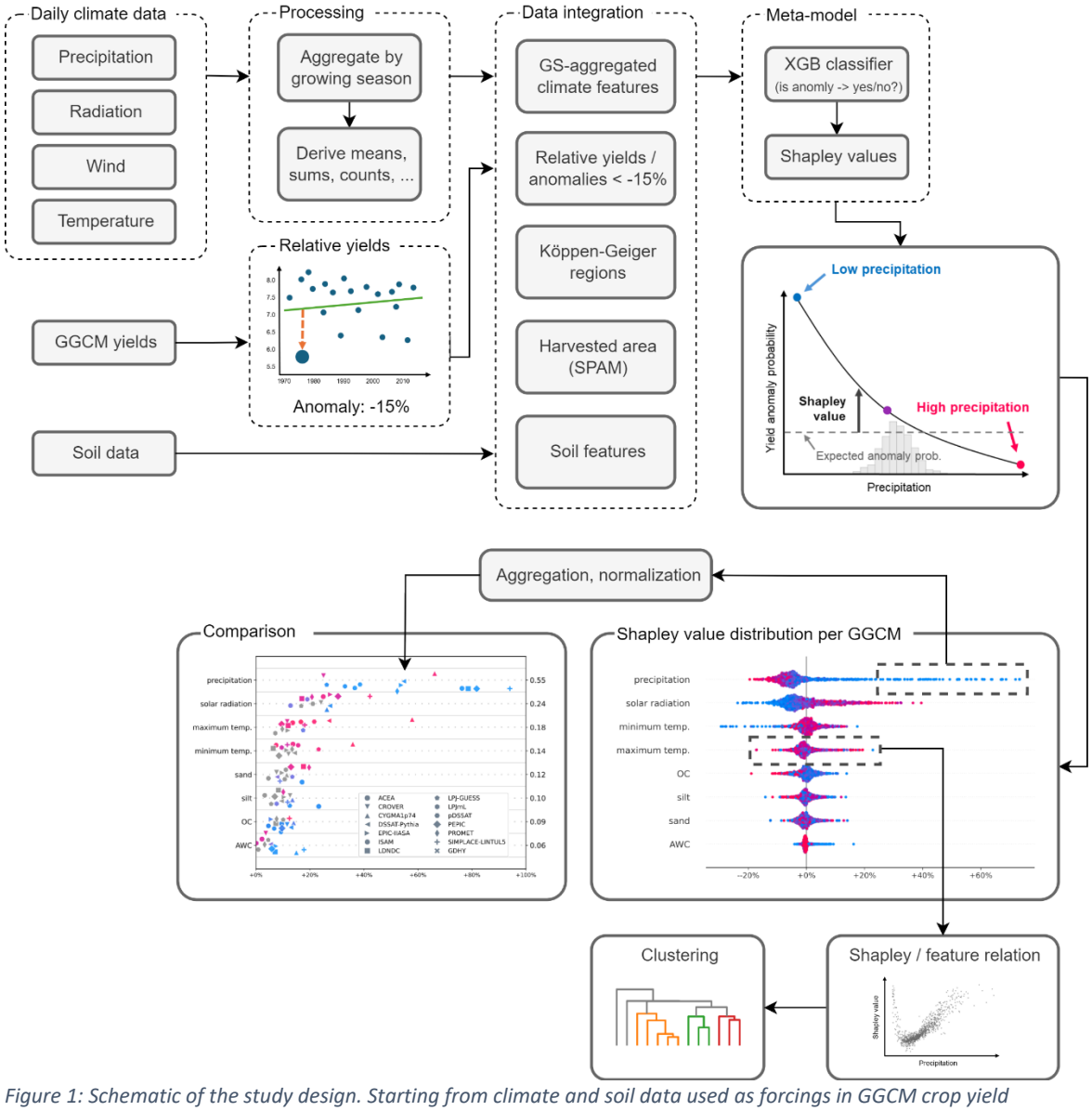

Figure 1: Schematic of the study design. Starting from climate and soil data used as forcings in GGCM crop yield simulations, features are derived for use in machine learning models. These are integrated with masks for climate domains and harvested areas to train XGBoost classifiers that predict yield anomalies. Subsequently, SHAP values are estimated for each feature as a measure of importance in predicting anomalies. We eventually analyze their distributions and response patterns across the ensemble. See sect. 2.1 for details.

While we focus on evaluating feature importances of GGCMs, we include a global reference dataset of reported and spatially disaggregated crop yields: the Global Dataset of Historical Yields (GDHY) for

major crops spanning the time period 1982-2016 (Iizumi and Sakai, 2020). As opposed to simulated yields, these data are subject to potential bias in spatial attribution of crops, changes in crop management over time, quality in data reporting, and other limitations. Therefore, we consider this comparison tentative and include it in the SI only.

2.2. Yield data detrending

To account for the effects of technological, management, and climate change, we apply Locally Weighted Scatterplot Smoothing (LOWESS) to the observational yield data and equally to those simulated by GGCMs, albeit these have static technology and management. LOWESS, or the almost identical LOESS (Locally Estimated Scatterplot Smoothing), is a common choice in yield detrending that can be parameterized to different timescales and expected fluctuations (Ben-Ari et al., 2018; Zhu et al., 2021). The relative yield is then expressed as the distance of the actual observation to the regression line, i.e., the expected value.

$$\bar{a}_{i,t} = \frac{y_{i,t} - \mu_{i,t}}{\mu_{i,t}} \qquad \text{(Equation 1)}$$

Where $\bar{a}$ is the relative yield anomaly, $y_{i,t}$ the observed yield and $\mu_{i,t}$ the expected value at the location i and time t. This formulation of relative yields is agnostic to the actual detrending procedure employed, as it only references the expected value $\mu_{i,t}$. The choice of detrending method is strongly related to the goal of the analysis. For example, if the goal is to determine the impact of climate change on yield anomalies, it would be counterproductive to remove the trend at this stage. We chose LOWESS with a 'fraction' parameter set to 0.5 and no reweighting, a choice for which visual inspection of randomly selected pixels showed good agreement with what should be considered a yield anomaly. Furthermore, the method and parameters were validated against known anomalies, such as the 2016 extreme yield loss on the French breadbasket (Ben-Ari et al., 2018).

2.3. Feature selection

As the selection of features is one of the most important choices in importance attribution, several versions and derivations of climate variables were investigated for suitability. Primarily, two types of features were compared: simple growing-season aggregates, such as average temperature, precipitation sum, etc., and features that are constructed by counting days within the growing season that satisfy certain criteria, e.g., number of wet days. The latter type is more common in the analysis of extreme events (Mistry, 2019) and they are expressed as fractions of growing season days to account for longer or shorter growing periods across the world, depending on GS temperatures. However, for their simplicity and improved interpretability across GGCMs, we chose GS-aggregates as our main feature set for the results presented in section 3. The climate feature data is then merged with soil, site, and management features and further reduced to avoid correlations (see Supplementary Text S2). Table 1 provides an overview of all features considered herein.

Table 1: Overview of features used in importance attribution. Climate features are grouped into a main set and an alternative set. Thresholds for features expressed as fractions of growing season days were sourced from earlier publications (McErlich et al., 2023; Mistry, 2019; Schauberger et al., 2017).

Short name	Description
Growing season aggregates	
solar radiation	Sum of solar radiation within the growing season [MJ m^{-2}]
max. temperature	Average maximum temperature [°C]
min. temperature	Average minimum temperature [°C]

precipitation	Sum of precipitation [mm]
Features expressed as fractions of growing season days	
wet days	Number of wet days (precipitation > 1mm)
heating degree days	Heating degree days (maximum daily temperature >= 30°C)
killing degree days	Killing degree days (maximum daily temperature >= 39°C)
frost days	Frost days (minimum daily temperature <= 0°C)
ice days	Ice days (maximum daily temperature <= 0°C)
heavy precipitation	Number of days with heavy precipitation (precipitation >= 10mm)
consecutive wet days	Number of consecutive wet days (precipitation > 1mm)
consecutive dry days	Number of consecutive dry days (precipitation <= 1mm)
Soil features	
sand	Sand content in topsoil [%]
silt	Silt content in topsoil [%]
AWC	Total plant available water capacity (AWC) [m^3 m^{-3}]
OC	Organic carbon (OC) content [%]

Results for the extreme feature set are provided in Figure S3. In designing the two types of feature sets, we account for potential impacts of hot, cold, wet, and dry weather on yield anomalies. Accordingly, GS-aggregated and GS-fraction features mostly express similar effects, but with different conceptualizations and quantifications. The choice of GS-aggregate features as the priority set was motivated by their robustness against model-specific thresholds and due to the inclusion of solar radiation.

2.4. Classifier model training

In its most direct form, accurate importance attribution is achieved by altering the inputs of a system and investigating how the change affects the output. Formally, this can be expressed globally, i.e., with a single metric per feature, and as a change in variance. This approach falls within the statistical domain of sensitivity analysis, which comprises a variety of methods that are adaptable to many different situations (Saltelli, 2008). However, we chose an alternative route for this analysis due to the following constraints and requirements.

1. Computation: The simulation of GGCM yields for the global timeseries data requires major work and computational efforts. Therefore, ad hoc analysis is difficult to perform on GGCMs directly.
2. Harmonization: Different models require different inputs and operate on different timescales. To enable a comparison, inputs need to be harmonized, here in the form of growing season aggregates.
3. Organization: We want to introduce a method that can be applied to an already existing data sample without the need for GGCM teams to opt into a specific project, run separate simulations, etc.
4. Explanation fidelity: While easy to communicate, global estimators, by definition, do not explain model behavior across the whole input domain and can be ambiguous and misleading, even for relatively simple models (Molnar et al., 2022).

To satisfy these constraints, we use a data-driven metamodel as a proxy for GGCMs, and SHAP values as local importance estimators, calculated for 1000 uniformly sampled data points across the input space (for details regarding SHAP, see next section). As the metamodel, we employ XGBoost, a highly scalable gradient boosting algorithm that yields a regularized random forest model (Chen and Guestrin, 2016). Based on best practices for threshold-based importance attribution, specifically in the binary case (Hastie et al., 2009), and previous work on attributing yield anomalies to climate variables (Ben-Ari et al., 2018), we train a binary classification model for each KG region and GGCM

on whether a GGCM output is considered a yield anomaly. Classification models are generally based on probabilities and thereby provide the additional benefit of quantifying some of the uncertainty involved in the prediction.

To avoid overfitting, hyperparameters (num. trees, learning rate, depth, min. child weight, gamma) were tuned using a randomized grid search and 10-fold cross-validation (CV) with stratified sampling that keeps the proportion of output classes constant across CV data slices. While predictive performance is not the primary interest here, overfitting to the training set can lead to inaccurate attribution of feature importance (Zhao et al., 2024). Training data size, the parameter values for the best estimators found in CV, and their discriminative performance in the form of the AUROC (Area under Receiver Operating Statistic curve) statistic are provided in supplementary tables Table S 2 - Table S 4. For interpretability and comparability of feature sensitivities (see section 2.7), it is important that the expected value of the classification model reflects the yield anomaly probability in the data, which is different per GGCM and region (Figure 2). Therefore, no reweighting of class probabilities is performed in model training. For the relative yield anomaly threshold, we choose a threshold of -15% from the expected yield, similar to previous work on yield anomalies (Ben-Ari et al., 2018). All data points less than or equal to that threshold are marked as yield anomalies. Furthermore, we only consider yield losses as anomalies, i.e., we do not investigate drivers of positive yield anomalies.

2.5. SHapley Additive exPlanations

Feature importances are derived from SHAP values and calculated by the Python module with the same name (Lundberg and Lee, 2017). SHAP was chosen for its expressiveness, mature formulation and implementation for machine-learning applications, and capacity to quantify feature interaction strength. Moreover, as a local estimator of importance, SHAP values facilitate an analysis that does not reduce feature importance to a single number. This is important for the analysis of complex input-output relationships, such as in the process-based models analyzed here. Shapley values are defined for a single data point x and feature *i* as:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|(M-|z'|-1)!}{M!} \left(f_x(z') - f_x(z' \backslash i) \right) \quad \text{(Equation 2)}$$

where ϕ_i denotes the Shapley value, $x' \in \{0,1\}$ the coalition vector indicating whether an element of x is included in the coalition or not, z' a subset of features within that coalition, M the maximum coalition size, $f_x(z')$ the model output, including feature i and $f_x(z' \backslash i)$ the model output excluding feature i. Note that Equation 2 requires all possible feature combinations to be exhausted for every data point, which makes computation challenging. SHAP provides several efficient approximations. For the analysis presented in this paper, we choose KernelSHAP as the approximation method, which reduces the combinatorial problem to a weighted least-squares fit that can be solved efficiently:

$$L(f,g,\pi_{x'}) = \sum_{z' \subseteq x'} \left[f\left(h_x(z')\right) - g(z') \right]^2 \frac{M-1}{(M \; choose \; |z'|)|z'|(M-|z'|)} \quad \text{(Equation 3)}$$

More intuitively, the SHAP value is the average change in model prediction resulting from including feature i, evaluated over all possible feature combinations, and herein indicates the contribution of feature values to the probability that a GGCM produces a yield anomaly.

One of the most important mathematical properties it satisfies is local accuracy, expressed as a linear, additive explanation model:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i' \quad \text{(Equation 4)}$$

Here, ϕ_0 is the Shapley value for including no features in the coalition, which is simply the expected value of model predictions for the data of interest X, or $\phi_0 = E(f(x)|X)$. A consequence of this additive property is that SHAP values are expressed as an offset of the average prediction value.

2.6. Interaction importance

While two individual features might only have a minor effect on the anomaly probability, their combined effect can be substantial if one reinforces the other. Shapley interaction values quantify the contribution of interactions alone, independent of the individual contributions. In other words, two features with high Shapley values by themselves can have an interaction value of zero if they do not interact. One example of interactions in crop models is the handling of different climate stresses. In the EPIC crop model, both heat and water deficit affect biomass development individually, and high temperatures can increase atmospheric water demand, exacerbating droughts. But the mutual exclusiveness of stresses in the model can also cause one to outweigh the other if physiologic heat stress and drought occur simultaneously (J. R. Williams et al., 1989). This applies to the majority of crop models and stresses, with few exceptions that consider co-occurring stress, for example, through multiplicative functions (Webber et al., 2022).

Shapley values can be used to quantify the strength of feature interactions as the difference of the interaction effect to the sum of main effects, i.e., the effect on the model prediction of including both minus the effect of including either. This implies a Shapley value of zero for non-interacting features – a desirable property for analysis. The definition of SHAP values, described in the section 2.5, is easily extended to include the strength of interactions between two features i and j with $i \neq j$ (Lundberg et al., 2018):

$$\phi_{i,j}(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-2)!}{2(M-1)!} \left(f_x(z') - f_x(z' \setminus i) - f_x(z' \setminus j) + f_x(z' \setminus \{i,j\}) \right) \quad \text{(Equation 5)}$$

The right-hand term essentially expresses the difference between including the two features individually and including them simultaneously. If this difference is zero, it is assumed that there is no interaction between features that affects the model response, while nonzero values indicate that one feature either facilitates (positive) or impedes (negative) the other.

2.7. Normalization and Importance Score

To facilitate comparability between results for different GGCMs, some properties of models and data need to be considered. Inputs for metamodels, such as climate and soil, can be assumed to follow similar distributions and, as classification models, their output always expresses a probability. While this does not mean that metamodels necessarily share the same learned relationship between input and output, SHAP values are derived by evaluating contributions of all possible feature combinations across the input domain, and sensitivities can be considered agnostic towards the actual form of the relationship. This makes such feature contributions comparable in principle. However, because SHAP values are expressed as the offset from an expected value, here, the baseline anomaly probability in the data, another problem is introduced, as this baseline can differ per GGCM. To enable a comparison of feature sensitivities in section 3, regardless of the expected value, we define a normalized importance score as:

$$\tilde{\phi}_i = \frac{\phi_i}{f(x) - E[f(x)]} = \frac{\phi_i}{\sum_j^p |\phi_j|} \quad \text{(Equation 6)}$$

Where $\tilde{\phi}_i$ is the score for feature i, p the number of features and $E[f(x)]$ the expected value. The normalized value, bounded within $[-1, 1]$, expresses the importance of a feature as the fraction of all feature contributions for a single data instance and eliminates the impact of different baselines and probability scales across models. It can be interpreted as "What fraction of the total deviation from the baseline does feature i account for?". Note that this only holds true for a single data point. The symbols indicating the importance of a single GGCM in Figure 3 are an aggregation of the top 5% of importance scores per feature. Therefore, they do not sum to one across features.

2.8. Clustering method

To highlight similarities between GGCMs as well as outliers in terms of feature importance, we present clustering results in section 3.4. For the hierarchical, agglomerative clustering algorithm (Müllner, 2011), each GGCM is characterized by its SHAP and corresponding feature values. These are treated as a bivariate distribution to calculate 2D Wasserstein, or "earth movers" distances for all GGCM combinations. The metric is a measure of dissimilarity between probability distributions and essentially captures the cost of moving a source to a target distribution (Villani, 2009). For multivariate distributions, this is not a trivial task, and special considerations regarding computation and interpolation need to be considered (Bonneel et al., 2011). We use the package Python Optimal Transport (POT) to calculate the distance matrix for clustering (Flamary et al., 2021).

To provide a rough grouping of GGCMs per KG region, edges of "well-separated" clusters are colored differently in the dendrograms. This separation is determined by manually setting a distance threshold of 0.5, which was chosen because it yields a balanced number of groups for this particular dataset. GGCMs with edges joining at Wasserstein distances (y-axis) below this threshold are considered a group.

3. Results and Discussion

3.1. Occurrence of maize yield anomalies in GGCMs

The tendency of models to produce anomalies *per se* varies strongly within and across major Köppen-Geiger climate regions and GGCMs, ranging from 3% to almost 50% (Figure 2a). The highest median occurrence of yield anomalies is found for (semi-)arid climates, followed by polar, and the lowest for tropical climates. Among the ensemble members, CYGMA1p74 and SIMPLACE show the highest occurrence of yield anomalies, and LPJmL, LPJ-GUESS, and ISAM the lowest. Some GGCMs have higher or lower rates of yield anomalies in specific climates. E.g., rates for PROMET are high in polar and low in arid climates, while for LDNDC, anomalies are more frequent in arid climates but otherwise low. Reported yields from GDHY are mostly bracketed by the GGCM ensemble, suggesting that the ensemble as a whole does not systematically under- or overestimate anomaly occurrence.

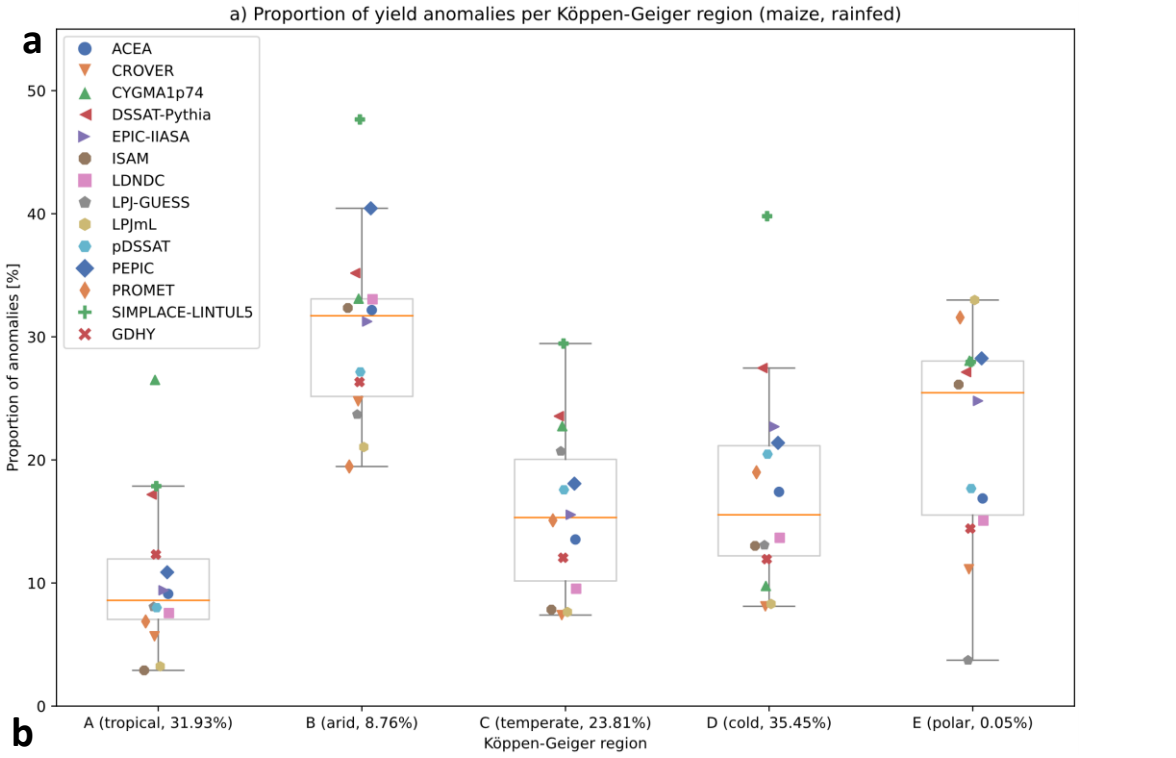

Figure 2: (a) Proportion of maize yield anomalies (≤ -15% from the expected value) per GGCM and Köppen-Geiger region. The proportion of harvested area for rainfed maize that falls within each region is shown in brackets. GDHY refers to the yield observation dataset by (Iizumi and Sakai, 2020). Only locations with rainfed maize harvested area according to the Spatial Production Allocation Model (SPAM) 2010 version 2r0 (Yu et al., 2020) are included. (b) Number of GGCMs per pixel for which at least one maize yield anomaly was detected in the time series. The data are masked by rainfed maize harvested area according to SPAM. Pixels hatched in white indicate the occurrence of yield anomalies in the GDHY dataset.

Spatially, the ensemble members tend to produce anomalies in the same geographic regions, with deviations mostly in climate regions that have overall low anomaly occurrence (Figure 2b). These include large parts of the tropics where pronounced drought and heat waves are less common, but also in the temperate climates of Southern and Central China. Observations don't necessarily follow the pattern of agreement among GGCMs, indicating anomalies, for example, in Southeast Asia and northern South America, but not in Eastern Europe, Russia, or India, among others.

3.2. Drivers of yield anomalies within the GGCMI ensemble

Across all climates and GGCMs, yield anomalies for rainfed maize can mostly be attributed to precipitation, expressed here as the sum over the growing season (Figure 3). Low precipitation is associated with higher yield anomaly probability, as indicated by the symbol colors and position along the x-axis. Also, their position at the lower half of the y-axis hints at a linear relationship between importance scores and feature values from low to high. CYGMAp74 presents an exception, where large volumes of growing season precipitation are associated with an increased anomaly probability in the tropics (A) and less so in cold (D) climates. This indicates that the model is more sensitive to excess water stress compared to the remainder of the ensemble, which is more sensitive to water limitations. CROVER shows a similar behavior but with a lower importance score. The clearest ensemble response is found for (semi-)arid climates (B), where excess precipitation hardly occurs by definition. Overall, sensitivities show strong variation in their contributions to yield anomalies across GGCMs, especially for high-ranking features where scores can vary by half of the available range and cover typically a quarter of it. While the overall ranking of features is fairly homogeneous, this underscores the heterogeneity of GGCMs in the ensemble when it comes to how different features are utilized in yield simulations and to what degree, which is of particular importance for extreme cases as seen here.

Shortwave solar radiation is the second most important factor in tropical (A) and temperate (C) climates, and ranks third elsewhere, following closely behind maximum temperature. The responses to solar radiation are already far less clear compared to precipitation, indicating substantial differences among the ensemble members. In all climate regions, about half of the ensemble shows anomalies in cases of high solar radiation, and the remainder in cases of low solar radiation or without a clear direction. While none of the ensemble members has a representation of solar radiation damage, radiation affects crop yields in two main ways, (I) as the key driver photosynthesis where low radiation causes low biomass overall and (II) as a driving term of atmospheric water demand where high solar radiation causes high demand in those GGCMs that use potential evapotranspiration (PET) functions considering this variable (e.g., Penman-Monteith). Accordingly, the high ranking of high solar radiation underpins recent findings for wheat crop models and field experiments, suggesting that simulation of atmospheric water demand requires more attention in crop model evaluations (Webber et al., 2025).

Maximum temperature ranks either second or third among climate regions, and minimum temperature is the lowest ranking climate feature. This underpins that responses to water deficit, incl. atmospheric water demand through high radiation, are more important across the ensemble than temperatures. The GGCM most sensitive to excess temperatures is CYGMA, most evidently in the warmer climate regions A and partiallyB. High maximum temperatures prevail as drivers of yield anomalies across GGCMs and climate regions, with few mixed signals. Low values drive the response of LPJ-GUESS in climate regions A, B, and C, and that of PEPIC in cold climates. For minimum temperature, the picture is less clear with a larger number of GGCMs showing either mixed responses or responding to low values. This pattern scales with cooler climates from A over C to D, whereas B climates cover both hot and cold domains, as elaborated further below in a more detailed assessment. Notably, the responses per climate region are also subject to the prevailing climate regimes, and it is hence not surprising that cold weather impacts dominate in colder climate regions. As with maximum temperature, CYGMA shows a high response to minimum temperature across regions A to C.

Soil properties contribute little to yield anomalies but are notable for a few GGCMs and climates. Across the ensemble, soil texture is relevant foremost in (semi-)arid climates where it contributes to soil hydrology and drought sensitivity, but the directionality of silt and sand fractions across models

varies greatly, indicating strong divergence in how they affect the soil water balance. SIMPLACE is the most sensitive to low precipitation, with Available Water Capacity (AWC) as the dominant soil feature in all non-arid climates.

Polar climates (E; including here foremost high mountain areas; Supplementary Figure S 2) comprise very low sample numbers and small harvested areas, which is why they are excluded from the main analysis and figures. Climate features in these regions follow largely a similar ranking, but with a more pronounced contribution of low temperatures and low solar radiation. Minimum temperature ranks higher than maximum temperature and is the main driver for only a few GGCMs.

While we include observation-based gridded crop yield anomalies solely as a tentative reference due to limited comparability with simulation results (see sect. 2.2), we still compare their derived importance scores at large (Supplementary Figure S 2). This shows that the score gradient (i.e., high or low values causing anomalies) mostly corresponds to the ensemble majority for top-ranking features, except for polar climates (E), where observations are likely too sparse and no clear direction can be identified. In most cases, importances for the GDHY data are bracketed by the ensemble, indicating a rough agreement between simulations and observations. Exceptions are precipitation in the tropics that hardly cause yield anomalies in GDHY, and high scores for solar radiation in (semi-)arid and temperate climates. This suggests that droughts in observed yields are driven by atmospheric water demand more than by water supply. However, this interpretation needs to be treated with caution, especially in the tropics, where changes in farming practices, impacts of pests and diseases, and other factors not accounted for in the models limit comparability. Targeted benchmarking studies should hence make use of observations sourced from data that are free from such potential biases.

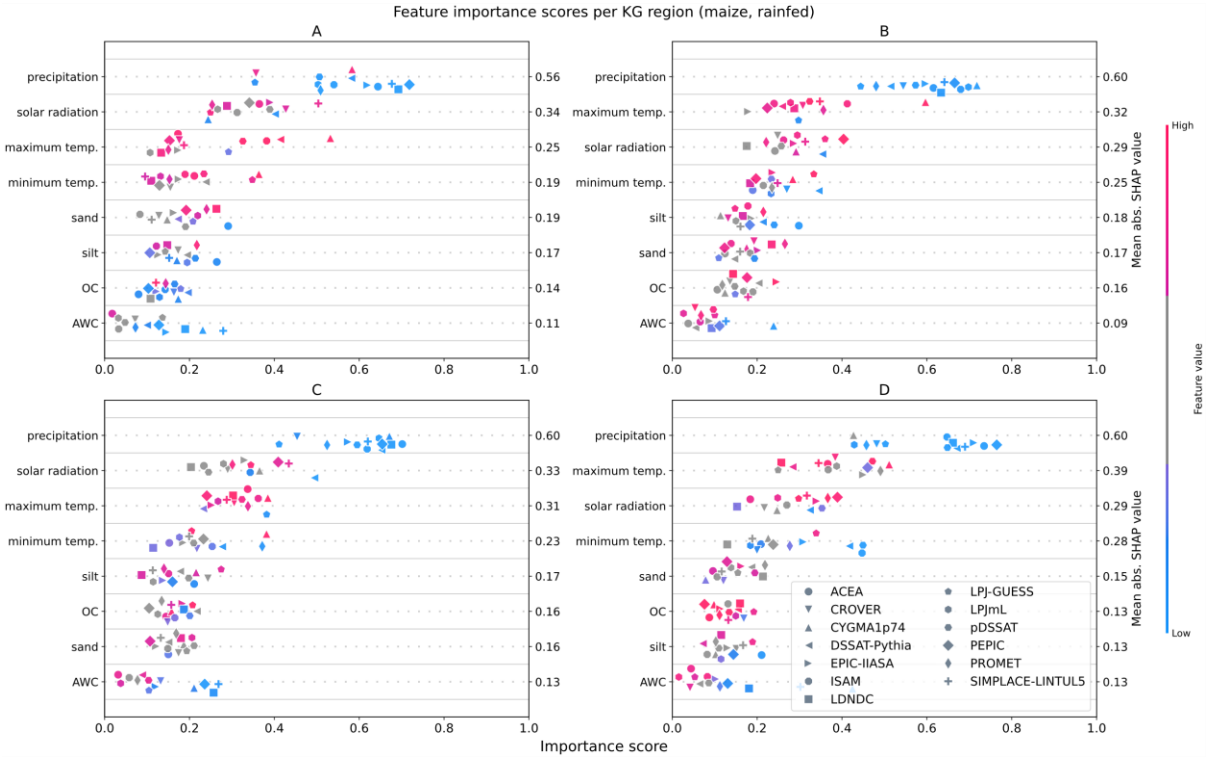

Figure 3: Drivers of yield anomalies per GGCM. Individual features are listed on the y-axis per panel and sorted by mean absolute SHAP value. The x-axis shows the mean of the top 5% of importance scores per feature (2.7). Symbol colors indicate whether low (blue) or high (red) feature values are associated with the anomaly, or if the top 5% of SHAP values do not capture a clear trend (grey). A symbol's vertical position within the feature band shows positive (top), negative (bottom), or no (middle) monotonic correlation between importance score and feature values, and roughly indicates the type and degree of linearity. See methods section 2.7 for defaults on the importance score, section 2.3 for feature descriptions, and Supplementary Text S1 for technical details about the plot.

A simulation scenario with sufficient irrigation virtually eliminates drought impacts, and thereby allows for a more nuanced evaluation of other climate variables under these particular conditions (Figure S4 - Figure S 6). In this set of simulations, solar radiation ranks first in tropical and temperate climates and third and second in (semi-)arid and cold climates, respectively. Furthermore, low radiation values tend to cause anomalies, which underpins that solar radiation may contribute to drought under rainfed (water-limited) conditions, but here predominantly drives anomalies through reduced radiation. For maximum temperatures, the picture is more mixed, and especially (semi-)arid climates show a divide among GGCMs with responsiveness to high values and others to low or mixed signals. Notably, the latter have an overall low tendency to produce anomalies under conditions of sufficient water supply (Figure S 7), or do so rather in cold arid regions, while the first have higher fractions of anomalies in hot arid regions. The most pronounced impact of high temperatures, on average, is again found for CYGMA, rendering it the most heat-sensitive ensemble member. Interestingly, AWC exhibits a high impact on anomalies for SIMPLACE across all climates and the EPIC-based GGCMs in the tropics (climate A). This shows that even with sufficient irrigation soil attributes are an important contributor to yield anomalies although the underlying processes that may relate to rooting depth, nutrient retention, or water logging among others cannot readily be interpreted without targeted experiments.

3.3. Alternative climate features

Using an alternative set of climate features that are based on fractions of days within the growing season exceeding specific thresholds (see Figure S3) shows highly comparable results, but a more pronounced picture where extremes may be explicitly represented in GGCMs. Interestingly, days with heavy precipitation emerge as the most important feature in all climates, which may be because this is the only feature reflecting specific volumes of precipitation. If these are low, the climate feature serves as a drought indicator. In most climate regions, other precipitation-related features rank second and third, showing that a single one of these extreme indicators may not be sufficient to explain yield anomalies. Heating degree days is the most important temperature-related feature, whereas killing degree days show limited impact. In cold climates (D), frost days become the most important feature for EPIC-IIASA and ISAM, and the overall most important feature in polar climates (E). Soil features show the same impacts as with the generic climate features, placed consistently in the lower half of the feature ranks.

3.4. Clustering of model responses

Results are grouped by their SHAP and corresponding feature values to identify patterns of response types within the ensemble and potential clusters (Figure 4). The clustering is performed on the full sample of 1000 data points used for importance attribution, whereas SHAP and feature values are treated as two dimensions of a bivariate distribution to calculate distances between GGCMs (see section 2.8 for details). This is done to provide a computational equivalent to a visual analysis of scatterplots (Figure S 20 - Figure S 23), which cannot fully capture all details of these distributions. This is especially true in the final assignment of clusters shown as different colors in the dendrograms. Therefore, we recommend consulting the scatterplots provided in the supplementary information, while the dendrograms below provide an aggregate overview of GGCM groupings.

For precipitation in region A, most GGCMs are placed in the same cluster showing high importance scores at low values, strong decay, and flattening as precipitation increases (Figure S 20). CROVER, CYGMA, DSSAT-Pythia, and SIMPLACE are exempt. Visual analysis of the scatterplots indicates a more hyperbolic response and a rebound in importance with the increase in total precipitation for CYGMA. This is less pronounced for CROVER, resulting in similar results for the aggregate top contributors shown in Figure 3A, while the overall response is more in line with the remainder of the

ensemble. In region B, responses agree well across the ensemble, though LPJ-GUESS is singled out by the clustering, which is in part due to a lower score for growing season precipitation. For region C, CYGMA and DSSAT-Pythia are placed in a separate group, which shows a slightly more abrupt decrease in importance for precipitation, as indicated in the scatterplots. Additionally, importances tend to increase for CYGMA with higher precipitation for a few data points. CROVER is not included in any group as it shows a rather flat response. Neither is LPJ-GUESS, where data points are clustered around the lower end of the precipitation scale. Responses in region D are split into three groups in the dendrogram. However, the split between the first and second group is close to the cutoff threshold, and visual inspection shows that both contain GGCMs with a parabolic response (ACEA, DSSAT-Pythia, pDSSAT, LDNDC, SIMPLACE in the first, EPIC-IIASA, PEPIC, and PROMET in the second). The third group is comprised of models with a rather flat shape (CROVER, ISAM, LPJmL). CYGMA and LPJ-GUESS are excluded from the groups as their distribution of precipitation data points is clustered toward the lower end of the scale compared to other models.

For solar radiation in region A, CYGMA, DSSAT-Pythia, LPJ-GUESS, and SIMPLACE are not assigned to the main group. CYGMA and DSSAT-Pythia show distinct response shapes with low variance (Figure S 21), while LPJ-GUESS contains some outliers with very high precipitation values in the scatterplot. Radiation importances for ISAM and LPJmL are very low throughout the range. Sensitivities for GGCMs in the main group are more spread out and oftentimes flat or with a slight upward trend. In regions B and C, scores for CYGMA are similarly condensed as in region A. DSSAT-Pythia shows a unique, substantially decreasing trend, indicating that only low or moderately low values contribute to anomalies. In region D, SHAP values for all GGCMs are clustered into a small range, and distributions are more varied, as indicated by the large number of models not assigned to a group. Here, the coloring due to the cutoff point at 0.5 is less informative, and the main branches of the dendrogram should be considered. CROVER, ISAM, DSSAT-Pythia, LDNDC, and pDSSAT are assigned to a group with data points that show no clear trend, and ACEA, LPJmL, EPIC-IIASA, PROMET, PEPIC, and SIMPLACE to a cluster with increasing importance. CYGMA and LPJ-GUESS are not assigned to any group based on their SHAP/feature distributions.

Sensitivities of temperature responses are quite similar across most of the ensemble. However, CYGMA shows a unique response with a quadratically increasing trend as the most temperature-sensitive GGCM. LPJ-GUESS shows partly inverse behavior to the ensemble in regions A-C, with a decreasing trend for maximum temperature. This is also true but less pronounced for DSSAT-Pythia in region C, where the two GGCMs are excluded from the main group. EPIC-IIASA, PEPIC, and PROMET show a clear hyperbolic response to maximum temperature in D climates and are assigned to a separate group. LPJ-GUESS is excluded from the clusters, probably due to outliers in the lower range of maximum temperatures.

Patterns and clusters for minimum temperature are largely similar to those of maximum temperature. CYGMA and LPJ-GUESS are excluded from the main groups in regions A and C due to their slightly increasing trend. In region B, two clusters emerge from the dendrogram: one with a decreasing trend, comprised of ACEA, CROVER, ISAM, LDNDC, LPJmL, pDSSAT, and PROMET, and one with a flat or increasing trend in CYGMA, EPIC-IIASA, PEPIC, LPG-GUESS, and SIMPLACE. DSSAT-Pythia is excluded because it is the only model with a strong, decreasing trend in importance. Again, region D shows the least agreement between GGCM distributions. A group of models with flat responses was identified, containing ACEA, CROVER, CYGMA, LPJmL, and LDNDC. EPIC-IIASA, pDSSAT, and ISAM show high importance of lower values, and DSSAT-Pythia shows a clear decreasing trend. PEPIC and PROMET were found to have a similar distribution, and LPJ-GUESS and SIMPLACE were excluded from any cluster.

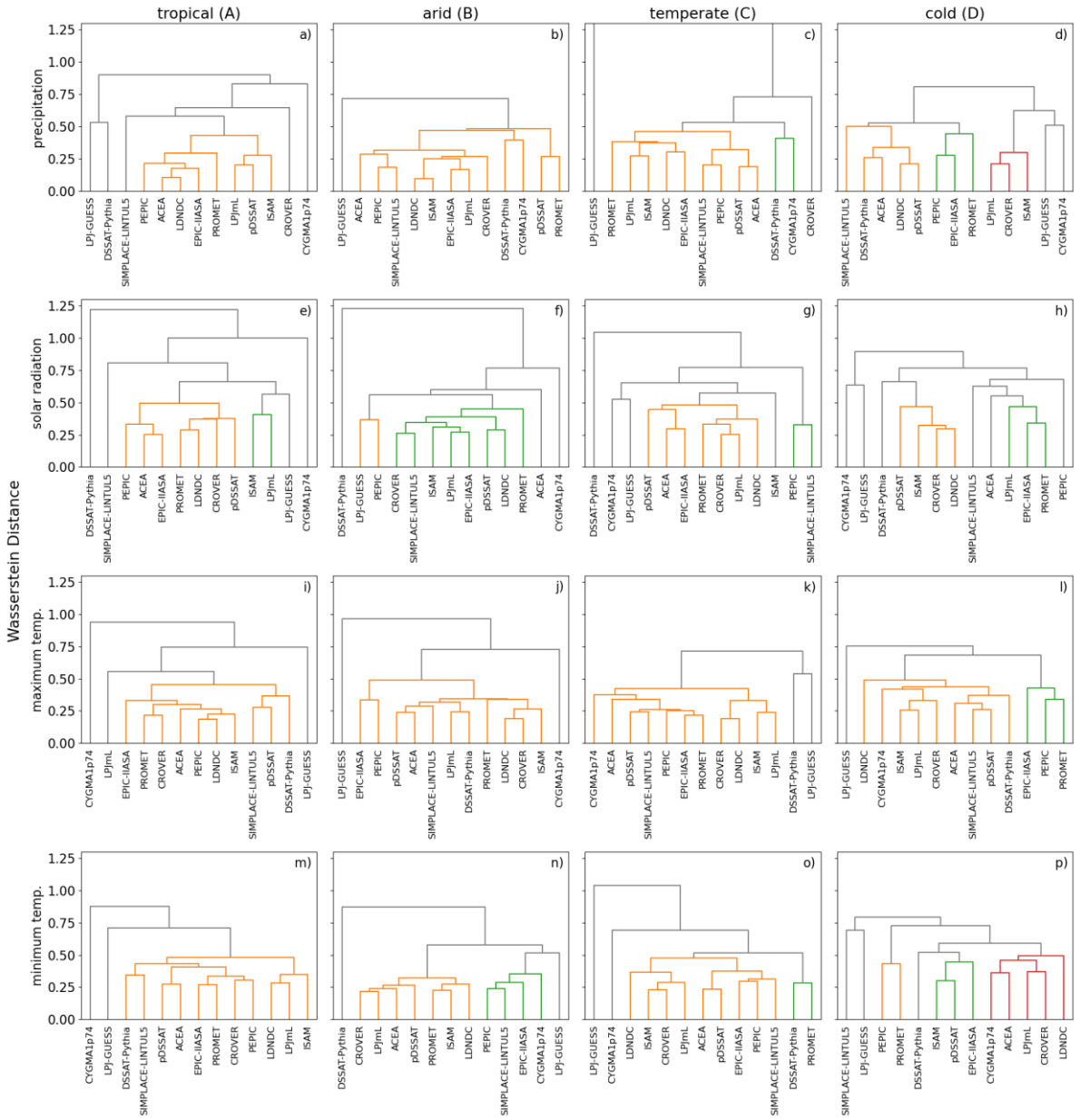

Figure 4. Dendrograms of GGCM responses to individual climate features per climate region. Distances between models are quantified as Wasserstein distances and include both SHAP and feature values (see sect. 2.8 for details). The y-axis shows the distance, i.e., the similarity between models. GGCMs with similar behavior merge at lower positions on the y-axis, while groups merging on top of the plots are less similar. The clustering is performed individually for each of the climatic drivers, and link lines are colored differently when groups can be clearly discriminated. See Supplementary Information X for scatter plots of SHAP vs. climate feature values.

Under fully irrigated conditions, clustering results are less clear, and the number of clusters below the distance threshold of 0.5 is generally higher (Figure S 6). However, for solar radiation, the most important feature in this scenario, the overall pattern of GGCMs included and excluded from main groups is similar: In regions A-C, DSSAT-Pythia and CYGMA are singled out, the same as in the results for rainfed maize. Results for the cold region (D) are partly different, with PEPIC, PROMET, and the two DSSAT GGCMs being separated from the ensemble.

For precipitation, which is notably of low importance in this scenario (Figure S4), the DSSAT models form a cluster of increasing importance in the tropics, whereas no clear trend can be seen for the other models (not shown). In (semi-)arid regions, the distribution of scores is quite varied. However, results of CYGMA, DSSAT-Pythia, and ISAM are of note as they are the only models with a decreasing

trend. CYGMA and DSSAT-Pythia form a cluster in temperate climates with slightly increased scores, whereas importances in region D are highly varied, and no clear clusters can be discerned.

For maximum temperatures, CYGMA is singled out due to the exceptionally high importance of high temperatures in all climates but cold regimes. There, a cluster of models shows high anomaly probability under low temperatures and virtually none under high, whereas the remainder of the ensemble tends towards a hyperbolic or flat response. DSSAT-Pythia shows similar behavior to CYGMA in region A, with increasing importance for higher temperatures. Because of some data points with very high and very low max. temperature, LPJ-GUESS is excluded from the main groups in region B-D.

For minimum temperature, CYGMA shows essentially the same behavior as for max temperature, with importances increasing for higher values in regions A and C, whereas scores for the remainder of the ensemble are close to zero. In semi-arid climates, DSSAT-Pythia emerges as the GGCM most sensitive to low temperatures (see also Supplementary Figure S4) and is also the model showing the largest fraction of anomalies (see Supplementary Figure S5). CYGMA and ISAM also indicate increased importance for higher values. In cold climate, the DSSAT models show the strongest response to minimum temperature, with values clustered around 5°C – 10°C.

Some of the differences between GGCMs identified in the clustering are due to different ranges of feature values. As feature values are growing season aggregates, they are highly affected by the growing season length, i.e., the planting and harvesting date of the crop. For example, for rainfed maize in KG region A, solar radiation for CYGMA and DSSAT-Pythia is more concentrated toward the lower end of values. Albeit planting dates and the duration of the growing season were harmonized in the underlying experiment, we find that maturity still differs greatly across the ensemble, with some models showing very early maturation in some regions and others delayed maturity (Figure S 24A-D). While we expect this not to affect the robustness of the importance analysis *per se*, it affects the clustering in particular (see also section 3.7).

3.5. Interactions among anomaly drivers

Incorporating SHAP interactions allows for evaluating interplays of features as another dimension of anomaly drivers. Due to their complexity and in part heterogeneity across the ensemble, only the top three are presented in Figure 5. Numeric results for all interactions are provided in Supplementary Table S 5 - Table S 6. The interaction between precipitation and solar radiation emerges clearly as the most dominant modulator of yield anomaly probability, both in terms of facilitating anomalies and mitigation. This is the case across all regions and GGCMs except CYGMA, where precipitation × maximum temperature interactions show higher importance scores. On average, the latter takes the second place, and precipitation × minimum temperature the third place. The order is flipped for some GGCMs and regions, such as for DSSAT-Pythia and EPIC-IIASA in region B or LDNDC in region D, where the interaction with minimum temperature clearly yields higher SHAP values. Overall, interactions with the maximum temperature are more sensitive, analogous to the results for standard SHAP values.

Note that the figure only shows the frequency, not the magnitude of interaction sensitivities. To present outlier-robust values for the average mitigating and facilitating effect, we report the first and third quartiles of the SHAP value distribution (Table S 7 - Table S 8). These can be substantial, ranging from -6.1% (mitigating) to +7.4% (facilitating), added to the overall anomaly probability. Finally, as discussed in section 2.6, computation of SHAP interaction values is highly susceptible to correlations, flagged here with an asterisk. While the key interactions are uncorrelated in most regions and models, they are in some instances and must be interpreted with caution. However, the

large number of uncorrelated occurrences suggests that the findings are sufficiently robust. Highly correlated interactions, such as minimum × maximum temperature, have been excluded from the analysis *a priori*.

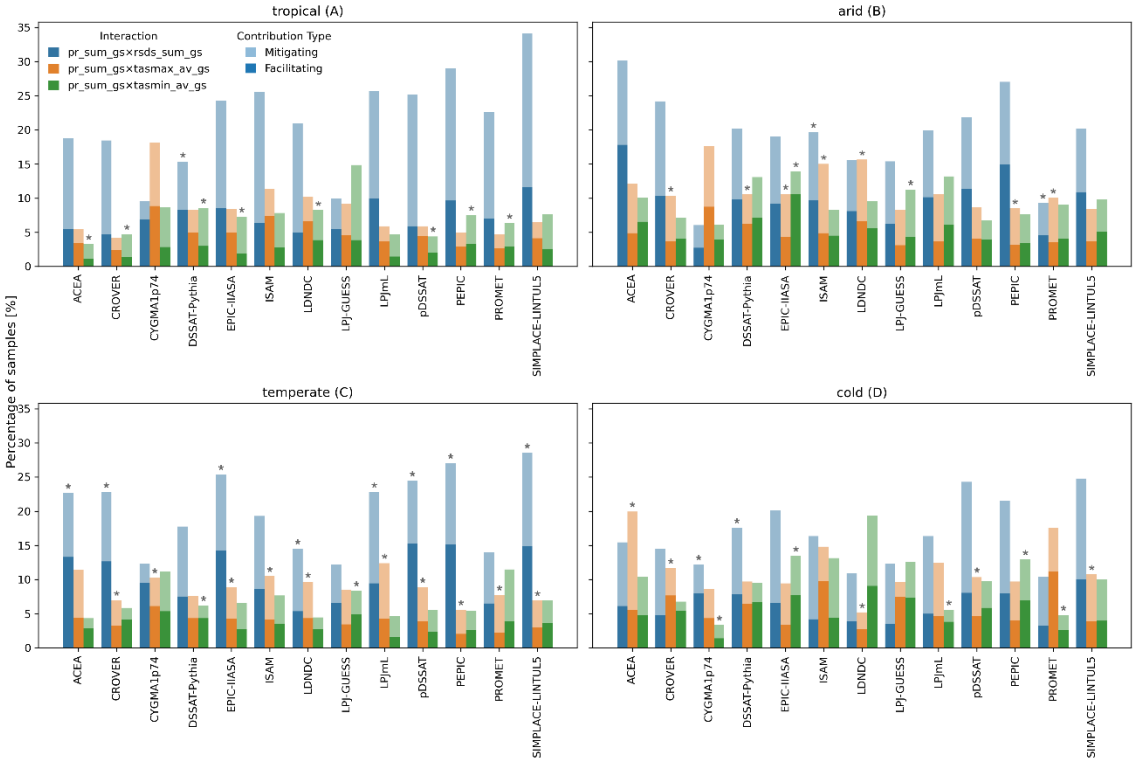

Figure 5. Percentage of the top three most dominant interactions in the set of 1000 sampled SHAP values. Here, both facilitating (contributing towards yield anomalies; dark hues) as well as mitigating (reducing the yield anomaly probability; light hues) sensitivities are shown as stacked bars. An asterisk above the bars indicates that the data for the GGCM x KG x interaction combination is subject to correlations above 0.1, which might affect the robustness of the SHAP interaction value.

3.6. Differences among model crops

Results for soybean show patterns comparable to those for maize (Supplementary Figure S9 to Figure S13). Differences in model responses may be partly due to differences in growing areas for maize and soybean, which are used for masking out regions in which growing a crop may be unsuitable. Furthermore, soybean is a C3 crop compared to C4 maize, which results in differences in photosynthesis, transpiration, and CO_2 responses, aside from general differences in crop physiology such as temperature thresholds for optimal growth.

In short, under rainfed conditions, the ranking of features remains largely the same as for rainfed maize with marginal shifts for solar radiation and maximum temperature. In contrast to maize, no increase in the importance of excess water in rainfed conditions was found. With sufficient irrigation, the impacts of high maximum temperature become pronounced for a wider range of GGCMs, but low maximum temperatures become more dominant in (semi-)arid and cold regions. Clustering of the ensemble members by their response shapes indicates a more homogenous majority with similar GGCMs showing rather unique responses, e.g., CYGMA, PROMET, LPG-GUESS, SIMPLACE. Note that for soybean, no simulation results are available for DSSAT-Pythia, which is why the model is excluded from the analyses.

3.7. Limitations and outlook

The feature importances presented herein characterize the input-output relationships of highly heterogeneous models, and they are derived by applying data-driven, metamodel-based importance attribution to an existing GGCM ensemble experiment. This approach is inherently subject to limitations due to assumptions and generalizations that have to be borne in mind when interpreting the results.

The machine-learning-based metamodels at the core of the analysis are trained on growing-season aggregates of crop model input and output data. Thereby, information is lost about how these quantities affect the accumulated daily timestep yield response in the original models. Thus, inputs that might strongly affect the yield in one part of the growing season but are compensated for in another are lost to aggregation, and so are other subtle effects. Also, weather conditions before the start of the growing season, specifically precipitation affecting soil moisture, are not captured by this approach (Sweet et al. 2025). Apart from aggregate features, we employ domain-specific feature engineering, such as the fraction of days within a growing season that exceed specific thresholds for temperature or precipitation. Both enable the training of expressive meta-models with good generalization performance (see Supplementary Table S4), indicating that feature effects are well captured on average with either approach. Future work could develop data-driven meta-models that are better aligned with the process model's structure, e.g., accumulate yield predictions in daily time steps. This would reduce generalization and conserve more detail of the processes. While not necessarily more accurate and computationally more expensive, such models generally allow for better interpretability (Ljung, 1999).

The data used in the ensemble experiment were not specifically designed to identify feature effects. They do not constitute a factorial or any common experimental design and are hence associated with the experiment they were derived from, i.e., climate reanalysis for the recent past with a business-as-usual management. While we trust that these already provide valuable insights into GGCMs' sensitivities for the evolving domain of yield anomalies, we stress that the results should not be used directly to interpret results from climate projections, where CO_2 effects can affect outcomes substantially (Toreti et al. 2020, Jägermeyr et al. 2021). We rather suggest applying the code published alongside this study for targeted assessments for GGCM sensitivities in such experiments.

In part, features herein are subject to correlations, which can distort the derived sensitivities (Aas et al., 2021), especially those of interactions. To limit this effect, correlated features were removed as far as possible in the analysis (see Supplementary Text S2) and otherwise flagged.

Finally, crop cultivars and maturation implemented in the simulations vary substantially across GGCMs (Figure S 24), and the resulting difference in growing season length has an impact on the distribution of climate features. These are derived from aggregating daily values by either calculating the mean (temperatures) or the sum (precipitation, solar radiation) across the growing season. Especially the latter is affected when this timeframe is particularly short, changing the mean of the distribution and potentially limiting comparability, especially for clustering. This needs to be considered when interpreting the findings in this article and other impact studies produced based on the ensemble. Further harmonizing the crop cultivars in future simulations could pave the way for a more precise comparison of ensemble members, albeit differences in the conceptualization of crop phenology in models are an intrinsic feature of the ensemble and a characteristic of uncertainty in process-implementation.

4. Conclusions

Our analysis identifies precipitation as the dominant variable affecting negative yield anomalies across a GGCM ensemble. Solar radiation and temperature-related features rank second. Radiation has thus far received limited attention in climate attribution, despite its central position in driving both photosynthesis and atmospheric water demand. We find that GGCMs vary strongly in their importance attribution to radiation as a driver of yield anomalies, highlighting its central role in understanding model-induced uncertainty in crop yield simulations. This should hence be accounted for in both factorial sensitivity analyses that could extend to the process-level for further insights on model response mechanisms and in the interpretation of impact studies. Similarly, we found highly varied responses to minimum and maximum temperatures, including higher anomaly probabilities for low values for a sub-group of models, which may inform temperature ranges for future analysis that have thus far focused on warming. Lastly, the presented method allows for identifying GGCMs with very specific responses under certain climatic conditions, which can inform further model development and the selection of ensemble members for specific applications. Overall, the results indicate that our method of importance attribution provides a means for quantitative evaluation of dominant GGCM features, using existing data from experiments such as reanalysis forcings or climate impact studies. Beyond earlier studies, it also includes interactions with non-climatic GGCM inputs, such as soil texture, that can catalyze climate impacts. Future research may seek to combine our method applied to an opportunistic sample herein with structured GGCM simulations as used in earlier studies that could eventually be reduced in volume with smart sampling approaches. We also expect our methodology to be of value in the analysis of GGCM experiments for which importance attribution is typically not performed, such as cooling and wetting from nuclear winter, geoengineering, and ocean current disturbance.

Acknowledgments

This research was supported by the Austrian Science Fund (FWF) (grant number 10.55776/P36220) and the Future of Life Institute (project ANFOS).

Open Research

Datasets for this research are available in these in-text data citation references: Jägermeyr et al., (2024); Iizumi and Sakai, (2020); Cucchi et al., (2020); Lange, (2019); FAO et al., (2012); Volkholz and Müller, (2020); International Food Policy Research Institute, (2020); Yu et al., (2020).

The code developed for this study is openly available at https://github.com/iiasa/ggcm-feature-importance.

Conflict of Interest Disclosure

The authors declare there are no conflicts of interest for this manuscript.

References

Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence 298, 103502. https://doi.org/10.1016/j.artint.2021.103502

Balkovič, J., van der Velde, M., Skalský, R., Xiong, W., Folberth, C., Khabarov, N., Smirnov, A., Mueller, N.D., Obersteiner, M., 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

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci Data 5, 180214. https://doi.org/10.1038/sdata.2018.214

Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van Der Velde, M., Makowski, D., 2018. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. Nat Commun 9, 1627. https://doi.org/10.1038/s41467-018-04087-x

Bonneel, N., Van De Panne, M., Paris, S., Heidrich, W., 2011. Displacement interpolation using Lagrangian mass transport, in: Proceedings of the 2011 SIGGRAPH Asia Conference. pp. 1–12.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cottrell, R.S., Nash, K.L., Halpern, B.S., Remenyi, T.A., Corney, S.P., Fleming, A., Fulton, E.A., Hornborg, S., Johne, A., Watson, R.A., Blanchard, J.L., 2019. Food production shocks across land and sea. Nature Sustainability 2, 130–137. https://doi.org/10.1038/s41893-018-0210-1

Cucchi, M., Weedon, G.P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., Buontempo, C., 2020. WFDE5: bias-adjusted ERA5 reanalysis data for impact studies. Earth System Science Data 12, 2097–2120. https://doi.org/10.5194/essd-12-2097-2020

Elliott, J., Kelly, D., Best, N., Wilde, M., Glotter, M., Foster, I., 2013. The parallel system for integrating impact models and sectors (pSIMS), in: Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery. Presented at the XSEDE '13: Extreme Science and Engineering Discovery Environment: Gateway to Discovery, ACM, San Diego California USA, pp. 1–8. https://doi.org/10.1145/2484762.2484814

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K.J., Büchner, M., Foster, I., Glotter, M., Heinke, J., Iizumi, T., Izaurralde, R.C., Mueller, N.D., Ray, D.K., Rosenzweig, C., Ruane, A.C., Sheffield, J., 2015. The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0). Geosci. Model Dev. 8, 261–277. https://doi.org/10.5194/gmd-8-261-2015

FAO, IIASA, ISRIC, ISSCAS, JRC, 2012. Harmonized World Soil Database (version 1.2).

Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N.T.H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D.J., Tavenard, R., Tong, A., Vayer, T., 2021. POT: Python Optimal Transport. Journal of Machine Learning Research 22, 1–8.

Folberth, C., Elliott, J., Müller, C., Balkovič, J., Chryssanthacopoulos, J., Izaurralde, R.C., Jones, C.D., Khabarov, N., Liu, W., Reddy, A., 2019. Parameterization-induced uncertainties and impacts of crop management harmonization in a global gridded crop model ensemble. PloS one 14, e0221862. https://doi.org/10.1371/journal.pone.0221862

Franke, J., Müller, C., Elliott, J., Ruane, A.C., Jagermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P., Folberth, C., Francois, L., Hank, T., Hoffmann, M., Izaurralde, R.C., Jacquemin, I., Jones, C.,

Khabarov, N., Koch, M., Li, M., Liu, W., Olin, S., Phillips, M., Pugh, T.A.M., Reddy, A., Wang, X., Williams, K., Zabel, F., Moyer, E., 2019. The GGCMI Phase II experiment: global gridded crop model simulations under uniform changes in CO_2, temperature, water, and nitrogen levels (protocol version 1.0). Geoscientific Model Development Discussions 1–30. https://doi.org/10.5194/gmd-2019-237

Franke, J.A., Müller, C., Elliott, J., Ruane, A.C., Jägermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P.D., Folberth, C., François, L., Hank, T., Hoffmann, M., Izaurralde, R.C., Jacquemin, I., Jones, C., Khabarov, N., Koch, M., Li, M., Liu, W., Olin, S., Phillips, M., Pugh, T.A.M., Reddy, A., Wang, X., Williams, K., Zabel, F., Moyer, E.J., 2020. 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, 2315–2336. https://doi.org/10.5194/gmd-13-2315-2020

Frieler, K., Schauberger, B., Arneth, A., Balkovič, J., Chryssanthacopoulos, J., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., Olin, S., Pugh, T.A.M., Schaphoff, S., Schewe, J., Schmid, E., Warszawski, L., Levermann, A., 2017. Understanding the weather signal in national crop-yield variability. Earth's Future 5, 605–616. https://doi.org/10.1002/2016EF000525

Frieler, K., Volkholz, J., Lange, S., Schewe, J., Mengel, M., del Rocío Rivas López, M., Otto, C., Reyer, C.P.O., Karger, D.N., Malle, J.T., Treu, S., Menz, C., Blanchard, J.L., Harrison, C.S., Petrik, C.M., Eddy, T.D., Ortega-Cisneros, K., Novaglio, C., Rousseau, Y., Watson, R.A., Stock, C., Liu, X., Heneghan, R., Tittensor, D., Maury, O., Büchner, M., Vogt, T., Wang, T., Sun, F., Sauer, I.J., Koch, J., Vanderkelen, I., Jägermeyr, J., Müller, C., Rabin, S., Klar, J., Vega del Valle, I.D., Lasslop, G., Chadburn, S., Burke, E., Gallego-Sala, A., Smith, N., Chang, J., Hantson, S., Burton, C., Gädeke, A., Li, F., Gosling, S.N., Müller Schmied, H., Hattermann, F., Wang, J., Yao, F., Hickler, T., Marcé, R., Pierson, D., Thiery, W., Mercado-Bettín, D., Ladwig, R., Ayala-Zamora, A.I., Forrest, M., Bechtold, M., 2024. Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP3a). Geoscientific Model Development 17, 1–51. https://doi.org/10.5194/gmd-17-1-2024

Gahlot, S., Lin, T.-S., Jain, A.K., Baidya Roy, S., Sehgal, V.K., Dhakar, R., 2020. Impact of environmental changes and land management practices on wheat production in India. Earth Syst. Dynam. 11, 641–652. https://doi.org/10.5194/esd-11-641-2020

Gebrechorkos, S.H., Sheffield, J., Vicente-Serrano, S.M., Funk, C., Miralles, D.G., Peng, J., Dyer, E., Talib, J., Beck, H.E., Singer, M.B., Dadson, S.J., 2025. Warming accelerates global drought severity. Nature 1–8. https://doi.org/10.1038/s41586-025-09047-2

Haas, E., Klatt, S., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., Grote, R., Breuer, L., Butterbach-Bahl, K., 2013. LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. Landscape Ecol 28, 615–636. https://doi.org/10.1007/s10980-012-9772-x

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, 3934–3965. https://doi.org/10.3390/rs70403934

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Iizumi, T., Furuya, J., Shen, Z., Kim, W., Okada, M., Fujimori, S., Hasegawa, T., Nishimori, M., 2017. Responses of crop yield growth to global temperature and socioeconomic changes. Sci Rep 7, 7800. https://doi.org/10.1038/s41598-017-08214-4

Iizumi, T., Sakai, T., 2020. The global dataset of historical yields for major crops 1981–2016. Sci Data 7, 97. https://doi.org/10.1038/s41597-020-0433-7

International Food Policy Research Institute, 2020. Global Spatially-Disaggregated Crop Production Statistics Data for 2010 Version 2.0. https://doi.org/10.7910/DVN/PRFF8V
J. R. Williams, C. A. Jones, J. R. Kiniry, D. A. Spanel, 1989. The EPIC Crop Growth Model. Transactions of the ASAE 32, 0497–0511. https://doi.org/10.13031/2013.31032
Jägermeyr, J., Frieler, K., 2018. Spatial variations in crop growing seasons pivotal to reproduce global fluctuations in maize and wheat yields. Sci. Adv. 4, eaat4517. https://doi.org/10.1126/sciadv.aat4517
Jägermeyr, J., Lin, T.-S., Rabin, S., Balkovic, J., Elliott, J.W., Faye, B., Folberth, C., Iizumi, T., Jain, A., Kiyoshi, T., Liu, W., Masashi, O., Mialyk, O., Müller, C., Stella, T., Wang, C., Webber, H., Yang, H., Zabel, F., Frieler, K., 2024. ISIMIP3a Simulation Data from the Agriculture Sector. https://doi.org/10.48364/ISIMIP.370868.1
Jägermeyr, J., Müller, C., Ruane, A.C., Elliott, J., Balkovic, J., Castillo, O., Faye, B., Foster, I., Folberth, C., Franke, J.A., Fuchs, K., Guarin, J.R., Heinke, J., Hoogenboom, G., Iizumi, T., Jain, A.K., Kelly, D., Khabarov, N., Lange, S., Lin, T.-S., Liu, W., Mialyk, O., Minoli, S., Moyer, E.J., Okada, M., Phillips, M., Porter, C., Rabin, S.S., Scheer, C., Schneider, J.M., Schyns, J.F., Skalsky, R., Smerald, A., Stella, T., Stephens, H., Webber, H., Zabel, F., Rosenzweig, C., 2021. Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. Nat Food 2, 873–885. https://doi.org/10.1038/s43016-021-00400-y
Jägermeyr, J., Robock, A., Elliott, J., Müller, C., Xia, L., Khabarov, N., Folberth, C., Schmid, E., Liu, W., Zabel, F., Rabin, S.S., Puma, M.J., Heslin, A., Franke, J., Foster, I., Asseng, S., Bardeen, C.G., Toon, O.B., Rosenzweig, C., 2020. A regional nuclear conflict would compromise global food security. PNAS 117, 7071–7081. https://doi.org/10.1073/pnas.1919049117
Jiang, S., Sweet, L., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., Zscheischler, J., 2024. How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences. Earth's Future 12, e2024EF004540. https://doi.org/10.1029/2024EF004540
Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European Journal of Agronomy 18, 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7
Lange, S., 2019. WFDE5 over land merged with ERA5 over the ocean (W5E5). https://doi.org/10.5880/PIK.2019.023
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, 2325–2337. https://doi.org/10.1111/gcb.14628
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
Ljung, L., 1999. System identification: theory for the user, 2nd ed. ed, Prentice Hall information and system sciences series. Prentice Hall PTR, Upper Saddle River, NJ.
Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30, 4765–4774.
Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent Individualized Feature Attribution for Tree Ensembles. https://doi.org/10.48550/ARXIV.1802.03888
Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rötter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J.,

2015. Multimodel ensembles of wheat growth: many models are better than one. Global Change Biology 21, 911–925. https://doi.org/10.1111/gcb.12768
Mauser, W., Klepper, G., Zabel, F., Delzeit, R., Hank, T., Putzenlechner, B., Calzadilla, A., 2015. Global biomass production potentials exceed expected future demand without the need for cropland expansion. Nat Commun 6, 8946. https://doi.org/10.1038/ncomms9946
McErlich, C., McDonald, A., Schuddeboom, A., Vishwanathan, G., Renwick, J., Rana, S., 2023. Positive correlation between wet-day frequency and intensity linked to universal precipitation drivers. Nat. Geosci. 16, 410–415. https://doi.org/10.1038/s41561-023-01177-4
Mialyk, O., Schyns, J.F., Booij, M.J., Hogeboom, R.J., 2022. Historical simulation of maize water footprints with a new global gridded crop model ACEA. Hydrol. Earth Syst. Sci. 26, 923–940. https://doi.org/10.5194/hess-26-923-2022
Mistry, M., 2019. A High-Resolution Global Gridded Historical Dataset of Climate Extreme Indices. Data 4, 41. https://doi.org/10.3390/data4010041
Molina Bacca, E.J., Stevanović, M., Bodirsky, B.L., Karstens, K., Chen, D.M.-C., Leip, D., Müller, C., Minoli, S., Heinke, J., Jägermeyr, J., Folberth, C., Iizumi, T., Jain, A.K., Liu, W., Okada, M., Smerald, A., Zabel, F., Lotze-Campen, H., Popp, A., 2023. Uncertainty in land-use adaptation persists despite crop model projections showing lower impacts under high warming. Commun Earth Environ 4, 1–13. https://doi.org/10.1038/s43247-023-00941-z
Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B., 2022. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models, in: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W. (Eds.), xxAI - Beyond Explainable AI, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 39–68. https://doi.org/10.1007/978-3-031-04083-2_4
Müller, C., Elliott, J., Kelly, D., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Hoek, S., Izaurralde, R.C., Jones, C.D., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T.A.M., Reddy, A., Rosenzweig, C., Ruane, A.C., Sakurai, G., Schmid, E., Skalsky, R., Wang, X., Wit, A. de, Yang, H., 2019. The Global Gridded Crop Model Intercomparison phase 1 simulation dataset. Scientific Data 6, 50. https://doi.org/10.1038/s41597-019-0023-8
Müller, C., Franke, J., Jägermeyr, J., Ruane, A.C., Elliott, J., Moyer, E., Heinke, J., Falloon, P.D., Folberth, C., Francois, L., Hank, T., Izaurralde, R.C., Jacquemin, I., Liu, W., Olin, S., Pugh, T.A.M., Williams, K., Zabel, F., 2021. Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. Environ. Res. Lett. 16, 034040. https://doi.org/10.1088/1748-9326/abd8fc
Müller, C., Jägermeyr, J., Franke, J.A., Ruane, A.C., Balkovic, J., Ciais, P., Dury, M., Falloon, P., Folberth, C., Hank, T., Hoffmann, M., Izaurralde, R.C., Jacquemin, I., Khabarov, N., Liu, W., Olin, S., Pugh, T.A.M., Wang, X., Williams, K., Zabel, F., Elliott, J.W., 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
Müllner, D., 2011. Modern hierarchical, agglomerative clustering algorithms. https://doi.org/10.48550/ARXIV.1109.2378
Okada, M., Iizumi, T., Sakamoto, T., Kotoku, M., Sakurai, G., Hijioka, Y., Nishimori, M., 2018. Varying Benefits of Irrigation Expansion for Crop Production Under a Changing Climate and Competitive Water Use Among Crops. Earth's Future 6, 1207–1220. https://doi.org/10.1029/2017EF000763
Orlov, A., Jägermeyr, J., Müller, C., Daloz, A.S., Zabel, F., Minoli, S., Liu, W., Lin, T.-S., Jain, A.K., Folberth, C., Okada, M., Poschlod, B., Smerald, A., Schneider, J.M., Sillmann, J., 2024. Human heat stress could offset potential economic benefits of CO2 fertilization in crop production under a high-emissions scenario. One Earth 7, 1250–1265. https://doi.org/10.1016/j.oneear.2024.06.012
Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid, E., Stehfest, E.,

Yang, H., Jones, J.W., 2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. PNAS 111, 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., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agricultural and Forest Meteorology, Agricultural prediction using climate model ensembles 170, 166–182. https://doi.org/10.1016/j.agrformet.2012.09.011
Ruane, A.C., Rosenzweig, C., Asseng, S., Boote, K.J., Elliott, J., Ewert, F., Jones, J.W., Martre, P., McDermid, S.P., Müller, C., Snyder, A., Thorburn, P.J., 2017. An AgMIP framework for improved agricultural representation in integrated assessment models. Environmental Research Letters 12, 125003. https://doi.org/10.1088/1748-9326/aa8da6
Saltelli, A. (Ed.), 2008. Sensitivity analysis, Paperback ed. ed, Wiley paperback series. Wiley, Chichester Weinheim.
Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., Pugh, T.A.M., Rolinski, S., Schaphoff, S., Schmid, E., Wang, X., Schlenker, W., Frieler, K., 2017. Consistent negative response of US crops to high temperatures in observations and crop models. Nature Communications 8, 13931. https://doi.org/10.1038/ncomms13931
Schewe, J., Gosling, S.N., Reyer, C., Zhao, F., Ciais, P., Elliott, J., Francois, L., Huber, V., Lotze, H.K., Seneviratne, S.I., Vliet, M.T.H. van, Vautard, R., Wada, Y., Breuer, L., Büchner, M., Carozza, D.A., Chang, J., Coll, M., Deryng, D., Wit, A. de, Eddy, T.D., Folberth, C., Frieler, K., Friend, A.D., Gerten, D., Gudmundsson, L., Hanasaki, N., Ito, A., Khabarov, N., Kim, H., Lawrence, P., Morfopoulos, C., Müller, C., Schmied, H.M., Orth, R., Ostberg, S., Pokhrel, Y., Pugh, T.A.M., Sakurai, G., Satoh, Y., Schmid, E., Stacke, T., Steenbeek, J., Steinkamp, J., Tang, Q., Tian, H., Tittensor, D.P., Volkholz, J., Wang, X., Warszawski, L., 2019. State-of-the-art global models underestimate impacts from climate extremes. Nature Communications 10, 1005. https://doi.org/10.1038/s41467-019-08745-6
Schleussner, C.-F., Deryng, D., Müller, C., Elliott, J., Saeed, F., Christian Folberth, Liu, W., Wang, X., Pugh, T.A.M., Thiery, W., Seneviratne, S.I., Rogelj, J., 2018. Crop productivity changes in 1.5 °C and 2 °C worlds under climate sensitivity uncertainty. Environ. Res. Lett. 13, 064007. https://doi.org/10.1088/1748-9326/aab63b
Schmidt-Traub, G., Obersteiner, M., Mosnier, A., 2019. Fix the broken food system in three steps. Nature 569, 181–183. https://doi.org/10.1038/d41586-019-01420-2
Smith, B., Prentice, I.C., Sykes, M.T., 2001. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. Global ecology and biogeography 621–637.
Sultan, B., Defrance, D., Iizumi, T., 2019. Evidence of crop production losses in West Africa due to historical global warming in two crop models. Sci Rep 9, 12834. https://doi.org/10.1038/s41598-019-49167-0
Villani, C., 2009. The Wasserstein distances, in: Optimal Transport, Grundlehren Der Mathematischen Wissenschaften. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 93–111. https://doi.org/10.1007/978-3-540-71050-9_6
Vogel, E., Donat, M.G., Alexander, L.V., Meinshausen, M., Ray, D.K., Karoly, D., Meinshausen, N., Frieler, K., 2019. The effects of climate extremes on global agricultural yields. Environmental Research Letters 14, 054010. https://doi.org/10.1088/1748-9326/ab154b
Volkholz, J., Müller, C., 2020. ISIMIP3 soil input data. https://doi.org/10.48364/ISIMIP.942125
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). Geosci. Model Dev. 11, 2789–2812. https://doi.org/10.5194/gmd-11-2789-2018

Webber, H., Cooke, D., Wang, C., Asseng, S., Martre, P., Ewert, F., Kimball, B., Hoogenboom, G., Evett, S., Chanzy, A., Garrigues, S., Olioso, A., Copeland, K.S., Steiner, J.L., Cammarano, D., Chen, Y., Crépeau, M., Diamantopoulos, E., Ferrise, R., Manceau, L., Gaiser, T., Gao, Y., Gayler, S., Guarin, J.R., Hunt, T., Jégo, G., Padovan, G., Pattey, E., Ripoche, D., Rodríguez, A., Ruiz-Ramos, M., Shelia, V., Srivastava, A.K., Supit, I., Tao, F., Thorp, K., Viswanathan, M., Weber, T., White, J., 2025. Wheat crop models underestimate drought stress in semi-arid and Mediterranean environments. Field Crops Research 332, 110032. https://doi.org/10.1016/j.fcr.2025.110032

Webber, H., Ewert, F., Olesen, J.E., Müller, C., Fronzek, S., Ruane, A.C., Bourgault, M., Martre, P., Ababaei, B., Bindi, M., Ferrise, R., Finger, R., Fodor, N., Gabaldón-Leal, C., Gaiser, T., Jabloun, M., Kersebaum, K.-C., Lizaso, J.I., Lorite, I.J., Manceau, L., Moriondo, M., Nendel, C., Rodríguez, A., Ruiz-Ramos, M., Semenov, M.A., Siebert, S., Stella, T., Stratonovitch, P., Trombi, G., Wallach, D., 2018. Diverging importance of drought stress for maize and winter wheat in Europe. Nat Commun 9, 4249. https://doi.org/10.1038/s41467-018-06525-2

Webber, H., Rezaei, E.E., Ryo, M., Ewert, F., 2022. Framework to guide modeling single and multiple abiotic stresses in arable crops. Agriculture, Ecosystems & Environment 340, 108179. https://doi.org/10.1016/j.agee.2022.108179

Wei, D., Gephart, J.A., Iizumi, T., Ramankutty, N., Davis, K.F., 2023. Key role of planted and harvested area fluctuations in US crop production shocks. Nat Sustain 1–9. https://doi.org/10.1038/s41893-023-01152-2

Xiao, L., Liu, L., Asseng, S., Xia, Y., Tang, L., Liu, B., Cao, W., Zhu, Y., 2018. Estimating spring frost and its impact on yield across winter wheat in China. Agricultural and Forest Meteorology 260–261, 154–164. https://doi.org/10.1016/j.agrformet.2018.06.006

Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A.K.B., Fritz, S., Xiong, W., Lu, M., Wu, W., Yang, P., 2020. A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps. Earth System Science Data 12, 3545–3572. https://doi.org/10.5194/essd-12-3545-2020

Zabel, F., Delzeit, R., Schneider, J.M., Seppelt, R., Mauser, W., Václavík, T., 2019. Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity. Nat Commun 10, 2844. https://doi.org/10.1038/s41467-019-10775-z

Zhao, N., Yu, J.Y., Dzieciolowski, K., Bui, T., 2024. Error Analysis of Shapley Value-Based Model Explanations: An Informative Perspective, in: Avni, G., Giacobbe, M., Johnson, T.T., Katz, G., Lukina, A., Narodytska, N., Schilling, C. (Eds.), AI Verification, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 29–48. https://doi.org/10.1007/978-3-031-65112-0_2

Zhu, P., Abramoff, R., Makowski, D., Ciais, P., 2021. Uncovering the Past and Future Climate Drivers of Wheat Yield Shocks in Europe With Machine Learning. Earth's Future 9, e2020EF001815. https://doi.org/10.1029/2020EF001815