Check for updates

Earth's Future

RESEARCH ARTICLE
10.1029/2025EF007457

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[3,4,5], Atul K. Jain[6], Tzu-Shun Lin[7], Oleksandr Mialyk[8], Christoph Müller[4], Alex C. Ruane[5], Florian Zabel[9], Juraj Balkovič[1], Chenzhi Wang[10], Babacar Faye[11], Jose R. Guarin[5,12], Toshichika Iizumi[13], Nikolay Khabarov[1], Wenfeng Liu[14], Masashi Okada[15], Sam S. Rabin[7], Clemens Scheer[16], Rastislav Skalský[1], and Christian Folberth[1,17]

[1]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, [2]University of Florida, Gainesville, FL, USA, [3]Climate School, Columbia University, New York, NY, USA, [4]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany, [5]NASA Goddard Institute for Space Studies (GISS), New York, NY, USA, [6]University of Illinois Urbana-Champaign, Urbana, IL, USA, [7]NSF National Center for Atmospheric Research (NCAR), Boulder, CO, USA, [8]University of Twente, Enschede, The Netherlands, [9]Department of Environmental Sciences, University of Basel, Basel, Switzerland, [10]Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF) e.V., Müncheberg, Germany, [11]Department of Environment, Biodiversity and Sustainable Development, University of Sine Saloum El Hadj Ibrahima NIASS, Kaolack, Senegal, [12]Autonomic Integra LLC, Bethesda, MD, USA, [13]National Agriculture and Food Research Organization (NARO), Tsukuba, Japan, [14]China Agricultural University, Center for Agricultural Water Research in China, Beijing, China, [15]National Institute for Environmental Studies (NIES), Tsukuba, Japan, [16]IMK-IFU, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [17]Centre for Climate Resilience, University Augsburg, Augsburg, Germany

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

Supporting Information:

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

Correspondence to:

T. Oberleitner,
oberleitner@iiasa.ac.at

Citation:

Oberleitner, T., Baklanov, A., Ferreira, T. B., Hoogenboom, G., Jägermeyr, J., Jain, A. K., et al. (2026). Explaining climatic drivers of yield anomalies in global crop models through metamodel-based attribution. *Earth's Future*, *14*, e2025EF007457. https://doi.org/10.1029/2025EF007457

Received 9 OCT 2025
Accepted 17 SEP 2026

Author Contributions:

Conceptualization: Thomas Oberleitner, Christian Folberth
Data curation: Thomas Oberleitner
Formal analysis: Thomas Oberleitner
Funding acquisition: Christian Folberth
Investigation: Thomas Oberleitner, Christian Folberth
Methodology: Thomas Oberleitner, Artem Baklanov, Christian Folberth
Project administration: Christian Folberth
Software: Thomas Oberleitner
Supervision: Christian Folberth
Validation: Thomas Oberleitner

© 2026 The Author(s). Earth's Future published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

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. The comparison with climate sensitivities derived from a global yield reference data set indicates that no single GGCM performs best across all climate regions. Individual GGCMs may respond well to one climatic stressor but not others, underscoring that bulk benchmarking of GGCMs against observations can obfuscate processes requiring targeted improvement. Our findings demonstrate that evaluating opportunistic data, that is, 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.

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 data sets can be used to explain why crop models disagree, without running new, resource-intensive experiments.

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Visualization: Thomas Oberleitner
Writing – original draft: Thomas Oberleitner, Artem Baklanov, Christian Folberth
Writing – review & editing: Thomas Oberleitner, Thiago Berton Ferreira, Gerrit Hoogenboom, Jonas Jägermeyr, Atul K. Jain, Tzu-Shun Lin, Oleksandr Mialyk, Christoph Müller, Alex C. Ruane, Florian Zabel, Juraj Balkovič, Chenzhi Wang, Babacar Faye, Jose R. Guarin, Toshichika Iizumi, Nikolay Khabarov, Wenfeng Liu, Masashi Okada, Sam S. Rabin, Clemens Scheer, Rastislav Skalský, Christian Folberth

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, 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). In parts of the world, there is often limited agreement even on the direction of change.

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 toward 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 & 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 (a) transient effects (e.g., sum of growing season precipitation) or (b) 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

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 yields and consequently anomalies have frequently been shown not to be driven by weather only (Cottrell et al., 2019; Iizumi et al., 2025; Vogel et al., 2019; Wei et al., 2023) we include observations in our evaluation to put our findings in context and perform a local evaluation of the methodology in the US Corn Belt as an intensively cultivated, well-documented region.

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 & 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 & 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 Table S1 of Supporting Information S1. 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° × 0.5° (approx. 55 × 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., 2024). We focus on rainfed crop yields (specifier noirr in the data repository) unless otherwise stated. Simulated crop yields assuming full irrigation (specifier firr) were used as a counterfactual that eliminates the influence of water deficit. 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 Section 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, IIASA, ISRIC, ISS-CAS, JRC, 2012; Volkholz & Müller, 2020). To allow for a fair comparison of results between GGCMs as well as the observation data set (GDHY), a common land-mask is applied. that is, only pixels are selected for which yield data is available across all data sets. Furthermore, we select only pixels in which the fraction of irrigated areas is <10% to exclude areas equipped for irrigation that may distort climate sensitivities. For the fully irrigated scenario, we apply the same land mask (i.e., pixels with <10% irrigated area under rainfed conditions) to maintain consistency with the rainfed analysis, as this scenario serves as a counterfactual that eliminates water deficit impacts. Accordingly, the comparison with yield observations was solely performed for rainfed crop yields.

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 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 Supporting Information S1.

While we focus on evaluating feature importances of GGCMs, we include a global reference data set of reported and spatially disaggregated crop yields, the Global Data set of Historical Yields (GDHY) for major crops, spanning the time period 1982–2016 (Iizumi & Sakai, 2020). As opposed to simulated yields, these data are subject to potential bias due to spatial attribution of crops, technological advances, changes in crop management

10.1029/2025EF007457

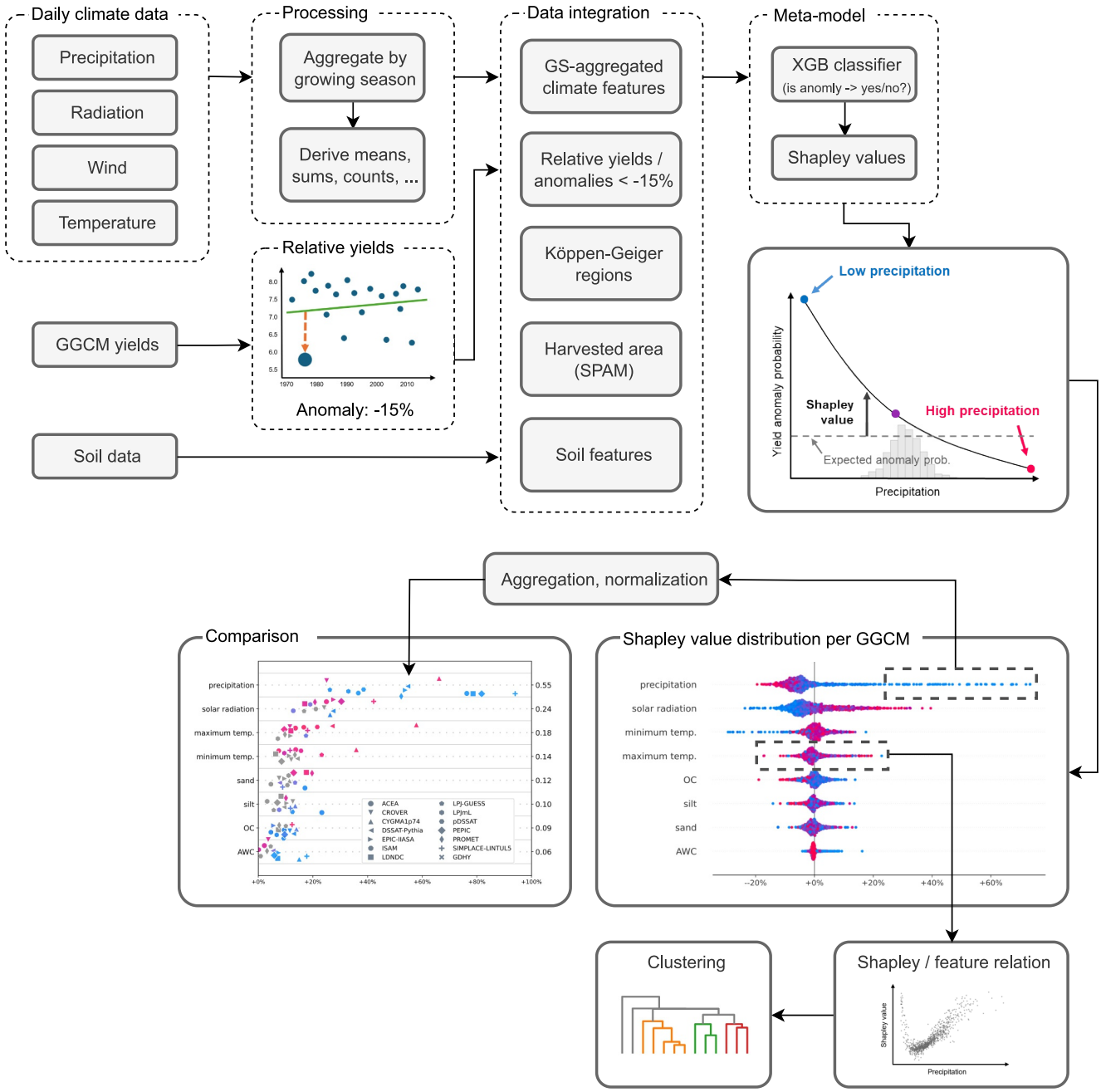


Figure 1. Schematic of the study design. Starting from climate and soil data used as forcings in GGCM crop yield simulations, features are derived for use in machine learning models. These are integrated with masks for climate domains and harvested areas to train XGBoost classifiers that predict yield 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 Section 2.1 for details. Abbreviations and symbols used throughout the manuscript are summarized in Supporting Information S1.

over time, and quality in data reporting, among others. These affect both trends and variability in yields (Iizumi et al., 2025) rendering it more challenging to robustly identify climate-driven yield anomalies.

To illustrate and evaluate the proposed importance attribution approach in a most robust setting, we apply the analysis to the US corn belt—a region with high data quality and well-studied anomalies. Crop yields in the region have shown a clear linear upward trend over the past decades due to continuously improved management and germplasm (Grassini et al., 2015), allowing for robust identification of climate-driven yield anomalies. The goal

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of the study is to evaluate the effectiveness of different parts of the data and modeling pipeline. Specifically, the following questions are addressed:

- Are known anomalies captured by the relative yields and threshold across the ensemble?
- How effective are GGCMs in simulating different kinds of yield anomalies?
- Do metamodels reliably reproduce GGCM anomalies?
- Are explanations provided by SHAP values biophysically meaningful?
- Does feature importance agree with global results overall in a less noisy subset of the data?

To identify a suitable set of pixels for this analysis, the data resulting from the pipeline described above is masked by the administrative boundaries of the US states Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Ohio, and Wisconsin. This constitutes the training data for the metamodels. Within these data, the pixel with the strongest average relative yield anomaly over the 1982–2014 period in the observation data set (GDHY) is selected for detailed analysis. The area and timeframe investigated contain two very prominent yield anomalies with contrasting weather patterns: (a) a heat-drought in the year 2012 and (b) a cold-wet season in 1993 (Bundy et al., 2022; Ortez et al., 2023).

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, that is, the expected value.

$$\bar{a}_{i,t} = \frac{y_{i,t} - \mu_{i,t}}{\mu_{i,t}}. \tag{1}$$

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, for example, number of wet days. The latter type is more common in the analysis of extreme events (Mistry, 2019) and is 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. For GGCMs, annual growing season dates were provided as simulations outputs in GGCMI phase3a. For the present study these were adopted from the public repository (Jägermeyr et al., 2024). To estimate growing seasons for the observational yield data set, the growing season was estimated using the GDD-based approach as implemented in the EPIC crop model by accumulating daily heat units from planting date until crop maturity, whereas the first day of maturity is selected as the harvest date. If maturity is not reached, the harvest date is set to 21 days after the calendar-based harvest date. Details are provided in Folberth et al. (2025). Climate feature data is then merged with soil, site, and management features and further reduced to avoid correlations (see Text S1 in Supporting Information S1). Table 1 provides an overview of all features considered herein.

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
Overview of Features Used in Importance Attribution

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 > 1 mm)
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 ≥ 10 mm)
Consecutive wet days	Number of consecutive wet days (precipitation > 1 mm)
Consecutive dry days	Number of consecutive dry days (precipitation ≤ 1 mm)
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 [%]

Note. 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).

Results for the extreme feature set are provided in Figure S3 of Supporting Information S1. 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 by 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, that is, 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).

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 constructs a regularized ensemble of decision trees (Chen & 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, parameter values for the best estimators identified during cross-validation, and discriminative performance measured by the AUROC (area under the receiver operating characteristic curve), as well as Brier Skill Scores as measures of probabilistic prediction performance, are provided in Tables S2–S5 of Supporting Information S1. Calibration curves for all classifiers are shown in Figure S31 of Supporting Information S1 and indicate no consistent evidence of strong systematic miscalibration across the GGCM classifiers. The largest deviations occur for the observational GDHY data set, consistent with the greater noise expected in observational compared with simulated yield data. 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, that is, 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 & 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:

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

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 $f_x(z' \setminus i)$ 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:

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

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

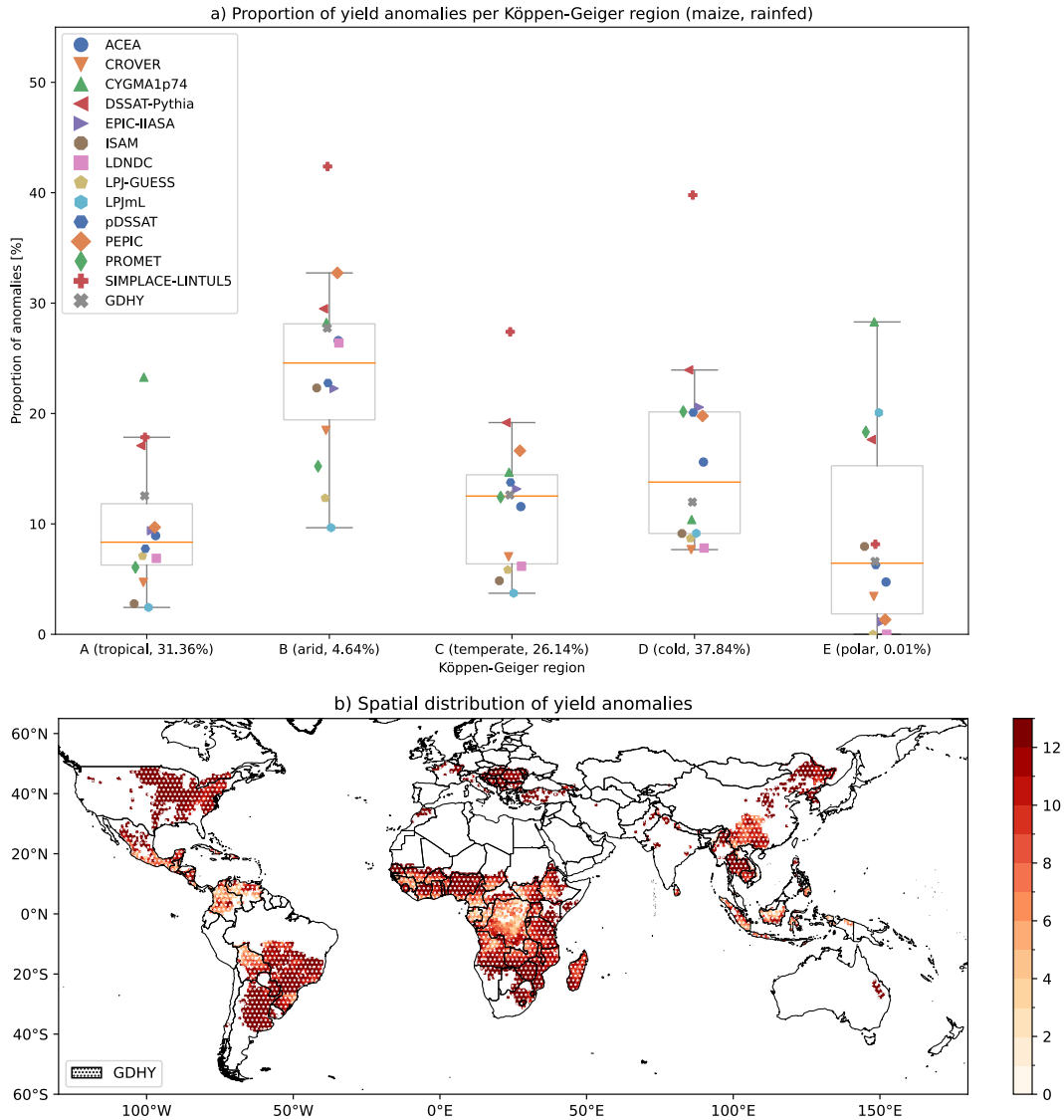


Figure 2. (a) Proportion of maize yield anomalies (≤−15% from the expected value) per GGCM and Köppen-Geiger region. The proportion of harvested area for rainfed maize that falls within each region is shown in brackets. GDHY refers to the yield observation data set by (Iizumi & Sakai, 2020). Only locations 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 further masked by areas where the proportion of irrigation is below 10%, according to SPAM, and all GGCMs, as well as GDHY, provide data. Pixels hatched in white indicate the occurrence of yield anomalies in the GDHY data set.

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

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i. \tag{4}$$

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.

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A problem that could potentially affect explanations is the presence of features with low cardinality, that is, a small number of discrete values compared to other features in the sample. Soil properties can be categorized as such. To prevent misleading explanations, we use SHAP's interventional perturbation mode, which marginalizes over a background distribution, here derived from the full data for a Köppen-Geiger region. This yields explanations that are robust against local differences in feature discretization.

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 (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. that is, it is 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 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):

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

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. Sensitivities can accordingly be considered agnostic toward 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:

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

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 4 are an aggregation of the top 5% of importance scores per feature. Therefore, they do not sum to one across features.

10.1029/2025EF007457

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

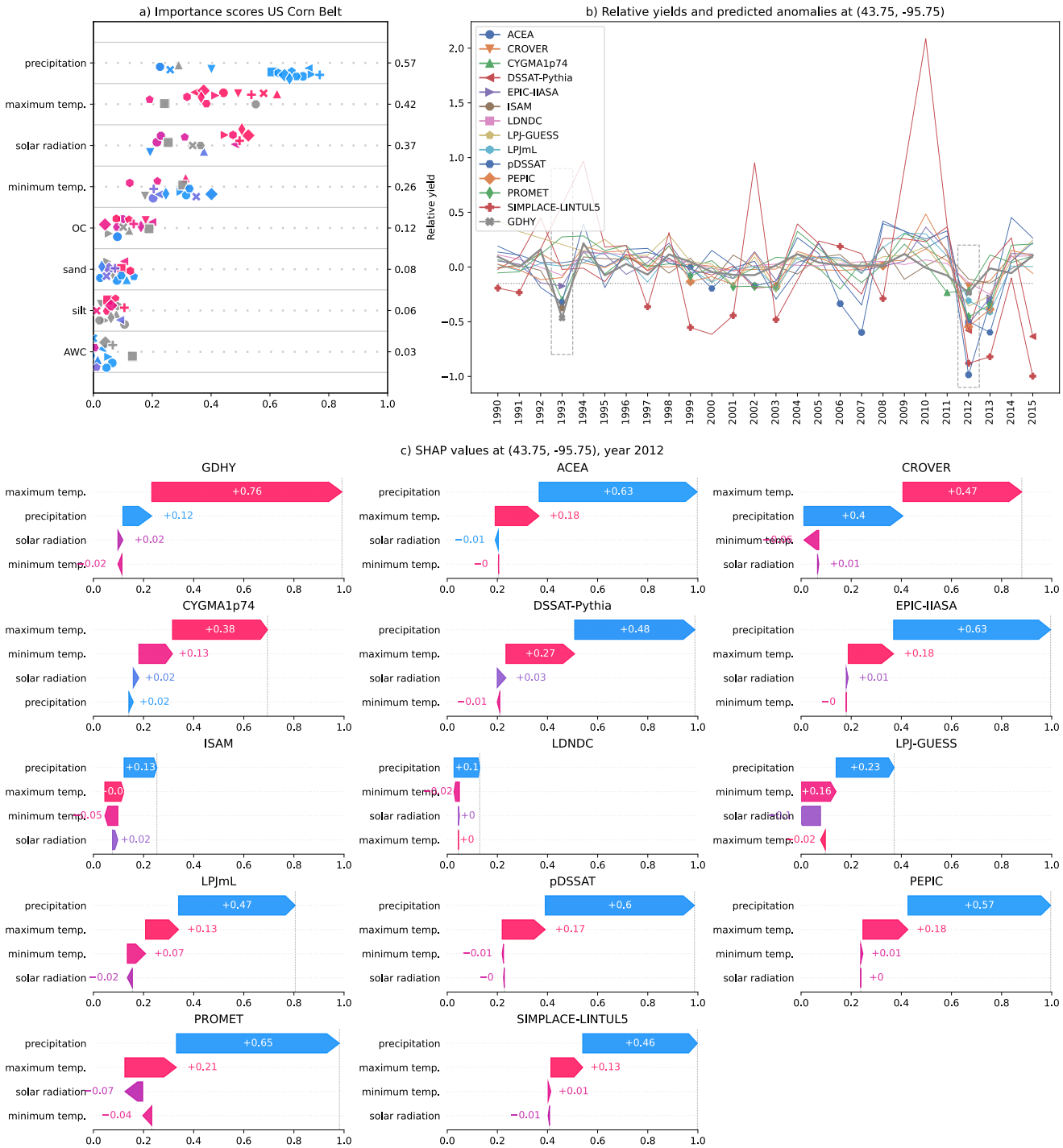


Figure 3. Method validation results performed on US Corn Belt data. Panel (a) shows the top importance scores for the region, analogous to Figure 4. Panel (b) shows a section of the time-series of relative yields for the test pixel. The lines represent relative yields in the training data for GGCM metamodels or the GDHY observed yields. A symbol indicates whether a GGCM metamodel predicts an anomaly in a given year, that is, the predicted anomaly probability of the classifier exceeds 50%. Panel (c) contains the SHAP values for each GGCM and GDHY, whereas soil properties are omitted as they were assigned SHAP values of zero throughout.

2.8. Clustering Method

To highlight similarities between GGCMs as well as outliers in terms of feature importance, we present clustering results in Section 3.5. 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

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Proportion of Pixel-Years in Which Anomalies From GGCMs Match Observations (GDHY) [%]

KG	ACEA	CROVER	CYGMA	DSSAT-Py	EPIC-IIASA	ISAM	LDNDC	LPJ-GUESS	LPJmL	pDSSAT	PEPIC	PROMET	SIMPLACE
A	81 (7)	85 (3)	71 (20)	76 (14)	81 (7)	86 (2)	83 (5)	80 (6)	86 (2)	83 (6)	80 (8)	83 (5)	75 (15)
B	62 (18)	67 (11)	60 (19)	60 (19)	62 (15)	62 (15)	62 (17)	68 (8)	68 (6)	63 (15)	59 (21)	66 (10)	56 (28)
C	79 (10)	83 (5)	76 (12)	74 (16)	78 (10)	84 (4)	83 (4)	75 (5)	85 (3)	78 (11)	76 (13)	79 (10)	67 (24)
D	83 (10)	87 (3)	84 (7)	77 (16)	81 (13)	84 (6)	87 (4)	78 (6)	86 (5)	80 (14)	81 (13)	80 (13)	66 (31)
E	80 (3)	83 (2)	74 (25)	77 (14)	78 (0)	72 (6)	82 (0)	83 (0)	67 (16)	82 (5)	77 (0)	78 (15)	74 (6)

Note. False positives, that is, pixel-years in which a GGCM but not GDHY produces an anomaly, are shown in brackets. False positives are pixel-years where a GGCM shows an anomaly but GDHY does not; this may also result from observational under-reporting rather than model error. KG: Köppen-Geiger regions (A = tropical, B = arid, C = temperate, D = cold, E = polar).

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).

For visualization, dendrogram edges are colored according to subtree membership using the default MATLAB/SciPy heuristic threshold (70% of the maximum linkage distance). This highlights major branches in the hierarchy and provides a qualitative indication of grouping structure.

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. for example, 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.

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 largely follow the distribution of anomalies, except in smaller sub-regions where the data set lacks data, such as the DRC.

The number of anomalies per GGCM, pixel, and year is compared to observations in Table 2. Only locations and years that exist in both data sets are considered when calculating the percentages. The highest agreement across the ensemble with observations occurs in cold regions (*D*, mean 81%), closely followed by the tropics (A, 80%), while (semi-)arid climates show the lowest agreement (B, 63%). Among GGCMs, the marginal agreement across all climates is highest for CROVER (80%) and lowest for SIMPLACE (67%). Note that these comparisons are performed on potentially noisy observations and do not necessarily reflect model skill. Accordingly, the agreement rates should solely be interpreted as consistency between models and observations. Results for pixel-years where anomalies are reported by GGCMs but not GDHY (i.e., false positives) are shown in brackets. For most GGCMs, these have a high occurrence in arid (B) climates, where the true positive rate is lowest, indicating an inherently different response between reference yields and GGCMs in this climate. This is most pronounced for CYGMA, PEPIC, and SIMPLACE. Yet, for some GGCMs, their occurrence can also be substantial (≥20%) in other climates, such as the tropics (A) and alpine (E) climates for CYGMA, and temperate (C) and cold (E) climates for SIMPLACE.

10.1029/2025EF007457

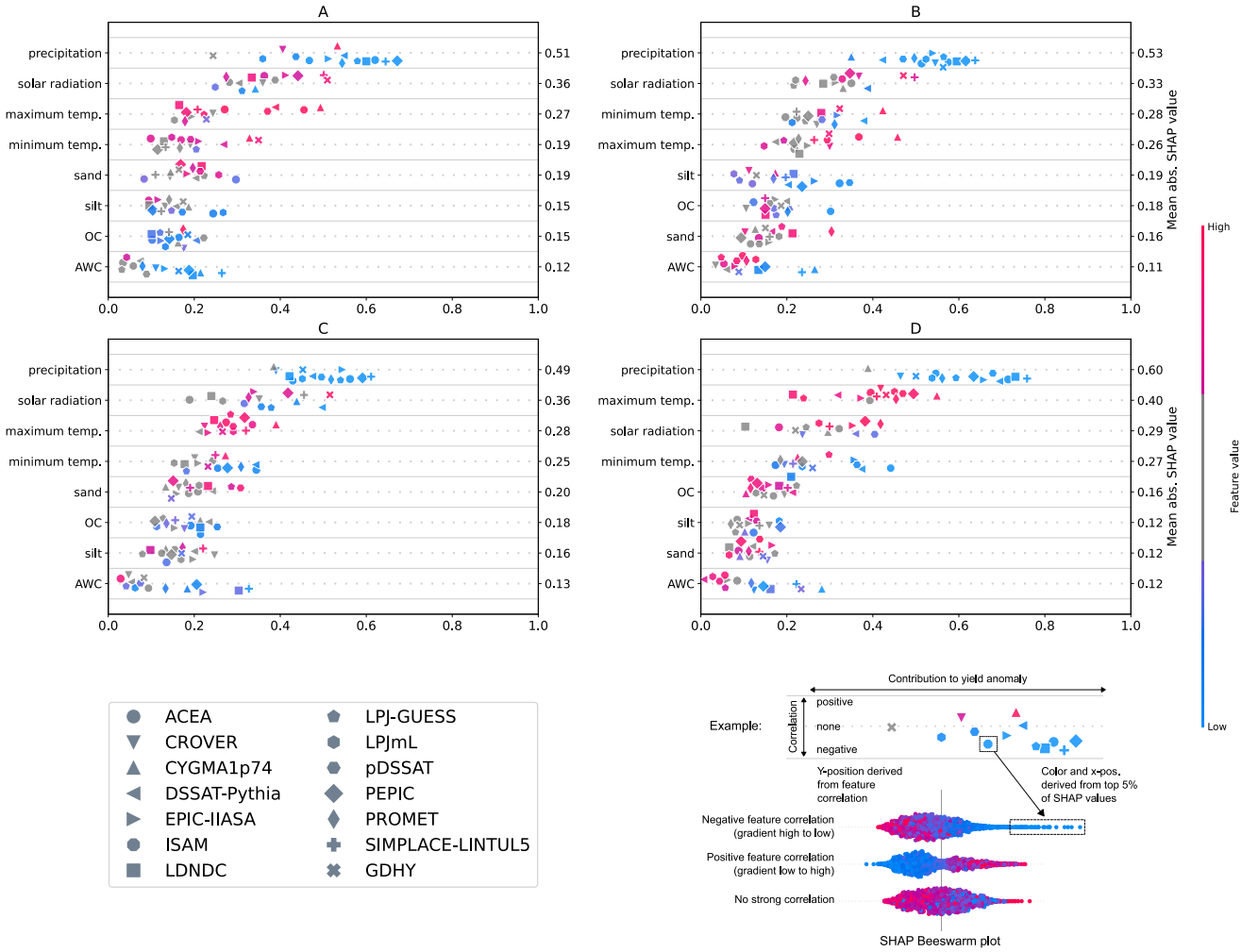


Figure 4. Drivers of yield anomalies per GGCM (symbols) and across the four major Koeppen-Geiger climate regions (panels) A (tropical), B (arid), C (temperate), and D (cold). Individual features are listed on the *y*-axis per panel and sorted by mean absolute SHAP value. The *x*-axis shows the mean of 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 (gray). 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. The legend in the bottom-right panel illustrates the interpretation of symbol color and vertical position using schematic example relationships between feature values and SHAP importance scores. See methods Section 2.7 for defaults on the importance score, Section 2.3 for feature descriptions, and Text S2 in Supporting Information S1 for technical details about the plot.

3.2. Location-Specific Evaluation

To evaluate the effectiveness of the proposed approach, two known crop failures are investigated in the US corn belt. The first one is a major drought in 2012 that resulted in reduced yields in most parts of the corn belt. The second is an exceptionally cold and wet season in 1993 that affected mainly the northwestern region (Bundy et al., 2022; Ortez et al., 2023). A test pixel was selected that covers both events to illustrate observations, relative yield in the training data, and importances for a single location.

The results for the 2012 heatwave are summarized in Figure 3. Panel (a) illustrates that the subregion results closely agree with the global results shown in Figure 4d. Panel (D) should be used for comparison as it corresponds to the dominant Köppen-Geiger class in the US Corn Belt. The timeseries data of relative yields for the selected test pixel are shown in panel (b), and GGCM symbols indicate whether the metamodel predicts a yield anomaly in a specific year. This is clearly the case for the 2012 drought, where most GGCMs reproduce the anomaly. Yield failures due to a cold and wet growing season are harder to capture for GGCMs. However, for those models where the event is reflected in the training data as reduced relative yield, metamodels correctly

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

predict the 1993 anomaly. This is the case for GDHY, CYGMA, EPIC-IIASA, ISAM, and pDSSAT. Panel (c) contains the metamodel-based explanations of the 2012 drought for the test pixel. Overall, GGCMs agree on the high importance of low precipitation and high max. temperature, which is even true for models that do not exceed the 50% yield anomaly probability threshold used by the classification model to indicate an anomaly (ISAM, LDNDC, and LPJ-GUESS). Note that importances of the top two features are flipped for the metamodel trained on observations, which ranks temperature over precipitation, and that this agrees with the importance ranking of CROVER and CYGMA also seen in the global results.

SHAP values for the 1993 event are provided in Figure S14 of Supporting Information S1. For the models predicting an anomaly, importances differ clearly from those in 2012, and low temperatures rank the highest, followed by high precipitation and low radiation (EPIC-IIASA, ISAM, and pDSSAT). Interestingly, GDHY values low radiation as the most important feature, and ranks high precipitation second.

Going back to the objectives of this validation study, the importance score patterns for the region investigated match those derived from the global analysis. While subject to the underlying training data, that is, GGCM simulated yields and detrending procedure, metamodels were largely capable of reproducing anomalies. This especially applies to those prevalent in climate data, such as droughts. For those models where the wet and cold event was captured in the training data, anomalies were predicted, and SHAP values clearly reflected biophysical relationships different from those of the 2012 drought. This manifests in flipped high/low values in precipitation and temperatures, as well as a contribution from solar radiation.

3.3. Global 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 4). 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, the sensitivities underscore the heterogeneity of GGCMs in the ensemble. Different climatic drivers have varying importances and signal strengths in yield simulations, 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: (a) as the key driver photosynthesis where low radiation causes low biomass overall, and (b) 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 underscores 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 partially B. 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 (D). For minimum temperature, the picture is less clear, with a larger number of GGCMs showing either mixed

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Absolute Differences of Importance Scores Between GGCMs and GDHY

KG	ACEA	CROVER	CYGMA	DSSAT-Pythia	EPIC-IIASA	ISAM	LDNDC	LPJ-GUESS	LPJmL	pDSSAT	PEPIC	PROMET	SIMPLACE	Σ
A	1.21	0.65	0.83	0.89	0.71	1.18	0.99	0.94	0.97	0.81	0.88	0.92	0.86	11.84
B	0.6	0.27	0.88	0.58	0.45	0.5	0.53	0.58	0.64	0.55	0.54	0.64	0.5	7.26
C	0.65	0.4	0.45	0.38	0.5	0.34	0.78	0.63	0.51	0.46	0.51	0.5	0.61	6.72
D	0.61	0.31	0.44	0.85	0.67	0.57	0.77	0.65	0.5	0.76	0.61	0.5	0.47	7.71
Σ	3.08	1.63	2.6	2.7	2.34	2.59	3.08	2.79	2.62	2.57	2.54	2.57	2.44	

Note. Results per climate feature are provided in Table S6 of Supporting Information S1. KG: Köppen-Geiger regions (A = tropical, B = arid, C = temperate, D = cold, E = polar).

responses or responding to low values. This pattern scales with cooler climates from A over C to D. (Semi-)arid 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.

Overall, soil properties contribute little to yield anomalies, and there is no matching order of sand, silt, and organic carbon (OC) across regions. Available Water Capacity (AWC) consistently scores lowest in the list of features investigated. Also, the directionality of feature values (low, high) varies greatly, indicating strong divergence in how they affect the soil water balance or affect yield anomaly occurrence otherwise.

Polar climates (E; including here foremost high mountain areas; Figure S2 in Supporting Information S1) 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 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.

Results for the reference yield data set GDHY show that the score gradient (i.e., high or low values causing anomalies) mostly corresponds to the ensemble majority for top-ranking features. An exception are polar climates (E), where observations are likely too sparse and no clear direction can be identified (see Figure S2 in Supporting Information S1). 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.

Still, we provide an indicator for how well GGCMs align with observations for contextualization. Each cell in Table 3 shows the sum of absolute differences of scores, that is, the difference between GDHY and GGCM on the *x*-axis in Figure 4, across all features for a GGCM. Sums are provided in the margins to show the overall agreement across Köppen-Geiger regions and GGCMs. Overall, CROVER aligns best with observations, while ACEA and LDNDC show the largest difference. Anomaly drivers match observations best in temperate (C) and worst by a large margin in arid (B) regions.

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 (Figures S4–S6 in Supporting Information S1). 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 photosynthesis. For maximum temperatures, the picture is more mixed. Especially (semi-)arid climates show a divide among GGCMs with responsiveness to high

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

values among some members and to low values or with mixed signals among others. Notably, the latter have an overall low tendency to produce anomalies under conditions of sufficient water supply (Figure S7 in Supporting Information S1), while the former 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 for 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 waterlogging, among others, cannot readily be interpreted without targeted experiments.

As a supplementary analysis, facilitating and mitigating effects between climate features were investigated. SHAP enables the attribution of importance to feature interactions, producing scores that are independent of individual feature contributions. Because the method for calculating interactions is strongly affected by even minor correlations and due to the volume of results, a highly condensed report is provided as Text S3 in Supporting Information S1.

3.4. 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 in Supporting Information S1) shows highly comparable results, but a more pronounced picture where extremes may be explicitly represented in GGCMs. To show this similarity numerically, SHAP values for GS-aggregated and extreme features were pooled into generalized (precipitation, temperature, soil) vectors and compared using cosine-similarity. Across GGCMs and climate regions, cosine similarities were consistently high (mean = 0.93, range = 0.82–0.99; Table S8 in Supporting Information S1), indicating that the dominant process-level attribution patterns were largely preserved across feature sets. However, solar radiation, a feature with high relevance in the standard set was not included in this comparison due to lack of an equivalent in the extreme feature set.

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 are 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.

3.5. 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 5). The clustering is performed on the full sample of 1000 data points used for importance attribution. SHAP and feature values are treated as two dimensions of a bivariate distribution to calculate Wasserstein distances between GGCMs (see Section 2.8 for details). These distances are the inputs for agglomerative, hierarchical clustering, and the results are presented as dendrograms to group similar behaviors of the feature-SHAP relationships. Furthermore, the distribution is illustrated in scatterplots in (Figures S21–S24 of Supporting Information S1). As a basic indicator of discrete clusters, edges to GGCMs belonging to the same group share the same color in the dendrograms, with GGCMs not belonging to a cluster shown in gray. This is only provided as a rough visual guide, as the result of the clustering algorithm is a matrix of distances rather than a discrete cluster assignment. Any distance threshold can be chosen to indicate group association in the dendrogram. Here, we retained the default threshold, as it yields a visually interpretable number of color groups in many applications (see Section 2). The clustering results are also subject to uncertainty arising from finite sample variability and stochastic sampling effects. We therefore performed a bootstrap-based clustering stability analysis to verify that the presented dendrogram structures are representative Text S3, Figure S32 in Supporting Information S1. In addition to cluster analysis, scatterplots provided in the SI can be consulted for further detail.

For precipitation in region A, two clusters emerge. One group shows high importance scores at low values, strong decay, and flattening as precipitation increases (ACEA, EPIC-IIASA, LDNDC, LPJ-GUESS, pDSSAT, PEPIC,

10.1029/2025EF007457

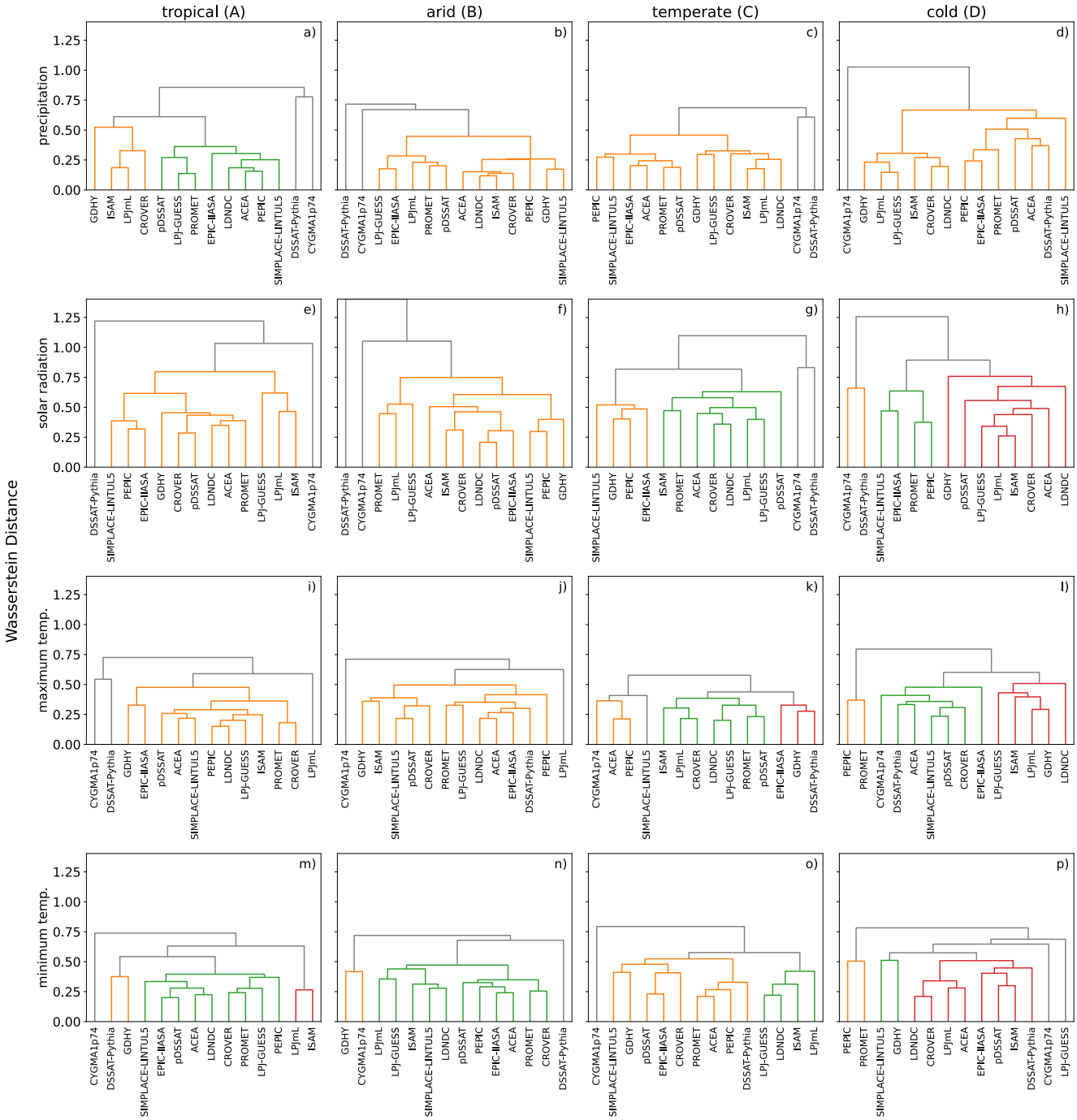


Figure 5. Dendrograms of GGCM responses to individual climate features per climate region. Panels (a–d), (e–h), (i–l), and (m–p) show precipitation, solar radiation, maximum temperature, and minimum temperature, respectively, across the tropical, arid, temperate, and cold regions. Distances between models are quantified as Wasserstein distances and include both SHAP and feature values (see Section 2.8 for details). The *y*-axis shows the distance, that is, 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 Figures S21–S24 in Supporting Information S1 for scatter plots of SHAP versus climate feature values.

PROMET, and SIMPLACE). Importances in the other group are low or zero throughout (CROVER, ISAM, LPJmL, GDHY, see Figure S21 in Supporting Information S1). CYGMA and DSSAT-Pythia are exempt due to a more hyperbolic response and a rebound in importance with the increase in total precipitation. In regions B and C, responses agree across the ensemble; however, CYGMA and DSSAT-Pythia are singled out again, here due to a

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

lack of feature values in the higher ranges, which could be caused by differences in their crop calendars for the latter (see discussion below). In regions C and D, CYGMA is the only model that attributes equally high importance to high precipitation as to low.

Similar to precipitation, CYGMA and DSSAT-Pythia are not assigned to the main group for solar radiation in regions A and B. In region A, importances for the GGCM ensemble and observations show hardly any trend and are clustered around zero, with scattering of Shapley values for some GGCMs at high radiation values. CYGMA exhibits a clear decreasing trend with feature values in a small range. DSSAT-Pythia has similarly low radiation values and an agglomerate of positive Shapley values at very low radiation. Similar patterns occur for regions B and C, where the feature range of CYGMA and DSSAT-Pythia is smaller and shows trends different from the rest of the ensemble. Notably, importances for DSSAT-Pythia decrease as radiation increases and indicate a strong mitigating effect at around 2,000 and 3,000 MJ/m^2 in the two climates. Two clusters are distinguished in region C: one with flat or slightly decreasing importances (ACEA, CROVER, ISAM, LDNDC, LPJ-GUESS, LPJmL, pDSSAT, and PROMET) and one with an increasing trend (EPIC-IIASA, PEPIC, SIMPLACE, and GDHY). In region D, SHAP values for all GGCMs are clustered into a small range, and distributions are more varied. Three groups emerge containing GGCMs with a flat response (ACEA, CROVER, ISAM, LDNDC, LPJ-GUESS, LPJmL, pDSSAT, and GDHY), with an increasing trend (EPIC-IIASA, PEPIC, PROMET, and SIMPLACE), or with differing feature ranges and/or decreasing trend (CYGMA, DSSAT-Pythia).

SHAP values for temperature responses are generally lower, and clustering is noisier. Therefore, the coloring and separation into distinct groups should be interpreted with particular caution in this case. In region A, importances are low throughout the temperature range and quite similar across most of the ensemble. Foremost, CYGMA and DSSAT-Pythia show substantially increasing importance with increasing temperatures, while few GGCMs show a pronounced response only around 35°C, which hardly occurs. In all regions, CYGMA shows a unique response with a quadratically increasing trend as the most temperature-sensitive GGCM. LPJmL assigns little importance to maximum temperature in regions A to C. While three groups are distinguished algorithmically in region C, differences appear subtle when looking at the scatterplots and the distances in the dendrograms. ACEA, CYGMA, pDSSAT, PEPIC, and SIMPLACE show a slightly increasing trend compared to the rest of the ensemble. However, they are not placed in the same group and therefore colored differently. Results are clearer in region D, with a group of GGCMs responding with a flat trend (ISAM, LDNDC, LPJ-GUESS, LPJmL, and GDHY), another with a parabolic response (ACEA, CROVER, CYGMA, DSSAT-Pythia, EPIC-IIASA, pDSSAT, and SIMPLACE), and a third group with feature values distributed over a slightly lower range (PEPIC, PROMET).

Patterns and clusters for minimum temperature are largely similar to those of maximum temperature. SHAP values are low or zero across the minimum temperature range for all GGCMs other than CYGMA, which again shows quadratically increasing importances in all regions. While also centered around zero, a slight increase in the highest 10% of minimum temperature values can be observed in region B, where DSSAT-Pythia is excluded from the main group because it is the only model with a strong, decreasing trend in importance. This decreasing trend in importance for DSSAT-Pythia can also be seen in region D.

SHAP values from the observation data set GDHY are generally embedded in the main clusters and do not present outliers for any of the regions or climate features. To identify which GGCMs are most strongly aligned with GDHY, the cophenetic distance is calculated, that is, the distance depicted in the dendrograms (Table S7 in Supporting Information S1). When comparing the mean distance across climate regions for each feature, results for precipitation are largely aligned with those in Table 3. Essentially, CROVER and ISAM are the GGCMs found in proximity to GDHY, while the distance is greatest for CYGMA. For radiation, SIMPLACE is closest to GDHY and DSSAT-Pythia the farthest. ISAM aligns best with GDHY for maximum temperature, and DSSAT-Pythia and SIMPLACE are the best matches for minimum temperatures. Therefore, based on clustering results and the analysis of cophenetic distance, no GGCM aligns with observations consistently. Notably, there are clear discrepancies between climate sensitivities and the agreement in reproducing observed anomalies (see Section 3.8).

Under fully irrigated conditions, clustering results are similar, and the overall patterns of GGCMs included and excluded from the main groups generally match those of the rainfed scenario (Figure S6 in Supporting Information S1). Note that the observation data set GDHY was omitted here as it predominantly reflects rainfed conditions (see Section 2.1).For precipitation, which is notably of low importance under irrigated conditions (Figure S4 in Supporting Information S1), the DSSAT models form a cluster of increasing SHAP values in the tropics, whereas no clear trend can be seen for the other models (not shown). This is also true in regions B-D,

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where only DSSAT models and CYGMA attribute some importance to precipitation and show a slightly decreasing trend. In region D, PROMET shows the same pattern. For solar radiation, the most important feature in this scenario, essentially the same clusters are assigned as in the rainfed maize scenario, with DSSAT-Pythia and CYGMA being singled out in all regions. PEPIC is excluded in the polar region due to higher radiation values in the sampled data. For maximum temperature, CYGMA is singled out due to the exceptionally high importance of high temperatures in all climates but cold regimes. In region A, the other models are clustered into a group of GGCMs that attribute some importance to higher temperature values (DSSAT-Pythia, ISAM, pDSSAT, and SIMPLACE) and the remainder of the ensemble that attributes none. The pattern is repeated in region C, with EPIC-IIASA joining the former group. Other than CYGMA, DSSAT-Pythia, and pDSSAT are the only GGCMs attributing importance to maximum temperature in the polar region. For minimum temperature, CYGMA and DSSAT-Pythia show elevated scores for higher temperatures in region A, whereas scores for the remainder of the ensemble are close to zero. This strong increasing trend can also be seen in region B for CYGMA, and the other models are split into a group that attributes some importance (CROVER, DSSAT-Pythia, EPIC-IIASA, ISAM, pDSSAT, PEPIC, PROMET, and SIMPLACE) and the rest that attributes none. Data for region C is noisy, and colors are assigned due to slight changes in the SHAP distribution. From the scatterplots, it can be seen that the only models attributing some importance to minimum temperature are CYGMA and the DSSAT models. In region D, CYGMA, DSSAT-Pythia, EPIC-IIASA, ISAM, and pDSSAT attribute importance to low minimum temperature values, and scores for the residual GGCMs are close to zero.

Across all analyses, 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, defined by 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 (Figures S25a–S5d in Supporting Information S1). While we expect this not to affect the robustness of the importance analysis per se, it affects the clustering in particular (see also Section 3.8).

3.6. Differences Among Model Crops

Results for soybean show patterns comparable to those for maize (Figures S9–S13 in Supporting Information S1). 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 the ensemble members by their response shapes indicates a more homogenous majority with similar GGCMs showing rather unique responses, for example, 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. Quantified Crop-Climate Sensitivities in the Context of Earlier Research

Regarding the impacts of high temperatures, Liu, Yang, et al. (2016) found that GGCM simulations, point-based simulations, and empirical models based on observations predicted similar yield losses under global warming. However, GGCMs almost uniformly showed the lowest impact among the three methods. Notably, different empirical models resulted in highly different impact estimates underpinning that it is not trivial to generalize crop-climate relations based on observations. Wang et al. (2017) investigated highly heterogeneous temperature response functions in 29 field-scale crop models and found that improved response functions can substantially reduce uncertainty and improve skill within selected ensemble members (none of which is part of the present study). Herein, we still see large variations in temperature sensitivities among GGCMs, ranging from very flat

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

responses to very pronounced impacts apparently driven by temperature thresholds (see Section 3.5; Figure S23 in Supporting Information S1). This suggests that there is still a substantial lag in the adoption and evaluation of findings from field-scale crop model studies into GGCMs, and the latter may profit from closer collaboration with the former.

Recent studies have addressed the skills and limitations of crop models to reproduce impacts of excess water on crop growth as a long-neglected impact pathway (Garcia-Vila et al., 2025; Liu et al., 2023). Garcia-Vila et al. (2025) in particular provide a detailed overview of relevant processes within >20 crop models, and find highly heterogeneous process representations, from impacts on nutrient cycling over photosynthesis to root aeration. Only a few of the GGCMs here are based on crop models included in that study - namely EPIC, LDNDC, AquaCrop, DSSAT, and SIMPLACE. While all of them have some of the relevant processes implemented, they show hardly any response to excess water herein. Also, in the cold-wet anomaly example for the Corn Belt (Figure S14 in Supporting Information S1), only one of the four GGCMs capturing the anomaly reacts to excess precipitation directly. The other three indicate an anomaly response to low temperature and, in part, to radiation. One reason could be that processes are first implemented and evaluated in crop models in highly controlled field-scale experiments in which water accumulation in the soil profile can be enforced (Nóia Júnior et al., 2023). In GGCM studies, in contrast, conditions of excess water hardly occur. This is because accumulation of water at the landscape scale is not considered, soil data hardly provide information on impermeable layers, the use of mean slopes typically causes later outflow of water, and groundwater table interactions are not accounted for.

Climate impacts on crops are often and increasingly compounding, as is the case most prominently for hot droughts (Rezaei et al., 2023). For example, a study combining observations, GGCM simulations, and empirical modeling suggests that high temperature impacts have mostly been causing yield declines through water stress in the Corn Belt in the past (Schauberger et al., 2017). Herein, we find that in most GGCMs, low precipitation slightly surpasses maximum temperature as a driver for yield anomalies in that region. Yield observations in turn are more sensitive to temperatures and, in part, solar radiation than low precipitation. The extent to which high temperatures cause plant stresses in GGCMs directly or through high PET estimates will need to be further investigated in studies disentangling underlying processes.

Finally, solar radiation has not been subject to crop model sensitivity analyses or large-scale evaluations of crop-climate relations to our best knowledge. For some GGCMs, this climate variable shows high importance, which indicates a contribution to impacts through PET estimation, as GGCMs commonly do not account for radiative damage. This should motivate further research into drivers of drought from a water demand side, as has recently been initiated in a study on Vapor Pressure Deficit (VPD) effects on drought in crop models (Webber et al., 2025).

3.8. Synthesis of Yield Anomaly Occurrence and Climate Responses

The prior sections revealed (a) heterogeneous agreement between observed and GGCMs' yield anomalies and (b) similarly heterogeneous climate sensitivities of GGCMs in producing yield anomalies that (c) hardly show consistent groupings of GGCMs across climate regions. Drawing on the findings of the prior sections, it is remarkable that there is no strong agreement between skills in matching observed anomalies and the drivers of anomalies in GGCMs. Contrasting the matching of anomalies and the response patterns underlying the clustering of GGCMs, expressed as cophenetic distance, underpins this (Figure S29 in Supporting Information S1). Notably, SIMPLACE shows the overall lowest match in reproducing yield anomalies derived from GDHY (Table 2; not considering E climates that are largely excluded from analyses due to limited sample numbers) but does not necessarily show the largest distances to GDHY's response to individual climate features. In fact, it has the closest matching response to precipitation in B climates, where its discrepancy to GDHY appears foremost driven by its temperature response (Figure S21b in Supporting Information S1). The other two GGCMs showing comparably low agreement with GDHY, namely CYGMA and DSSAT-Pythia, show for solar radiation, precipitation, and in part temperatures, also the largest distances to the responses of the yield observation data set. The former in particular exhibits the largest distance to GDHY across the ensemble and climate regions, showing that solar radiation, a climate variable thus far underappreciated in crop model sensitivity analyses, deserves more scrutiny. At high values, it can contribute to the occurrence of droughts, as it is part of most PET estimation methods' driving terms, or it may have compensatory effects in the context of yield anomalies if models' photosynthesis processes are overproportionately responsive to solar radiation. Apart from this, no clear patterns emerge from the comparison of anomaly matches and climate responses. Notably, even if a GGCM shows very high skill in

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reproducing anomalies, as is the case for LPJmL in most climates, it may have a lower correspondence to observed crop-climate responses across the spectrum of feature values than other GGCMs, as is the case for this model in B climates. This underpins that solely using crop yield data to evaluate model skills—as has long been the most common practice—should be complemented with climate response analyses as performed herein to develop a sound understanding of GGCM ensembles' behavior and inform interpretations that truly reflect responses to change in individual climate quantities.

3.9. Limitations and Outlook

The feature importances presented herein characterize the input-output relationships of highly heterogeneous models, and they are derived by applying a novel 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 features 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 Table S4 in Supporting Information S1), indicating that feature effects are well captured on average with either approach. Future work could develop data-driven metamodels that are better aligned with the process models' structures, for example, accumulate biomass and yield formation in daily or sub-seasonal time steps. This would reduce generalization and conserve more detail of the processes. While not necessarily more accurate and computationally more expensive, such models are expected to allow for better interpretability (Ljung, 1999). However, defining sub-seasonal features for a crop model ensemble may not be trivial, as models conceptualize and implement phenological development and crop growth stages quite differently. The feasibility of using sub-seasonal features may therefore need to first be tested for well-defined case studies.

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, here climate reanalysis for the recent past with a business-as-usual management. This provides important insights into GGCMs' sensitivities for the evolving domain of yield anomalies. In turn, results should not be used directly to interpret results from climate projections, where CO_2 effects can affect outcomes substantially (Jägermeyr et al., 2021; Toreti et al., 2020) and climate patterns may exceed the domain investigated herein. As our method for the first time allows a straightforward way to quantify drivers of crop-climate relations alongside experiments, we suggest applying the code published alongside this study for targeted assessments of GGCM sensitivities in such experiments.

Related to the problem of reducing inputs to a set of aggregated features is the inherent complexity and heterogeneity of crop models. Not only do GGCMs use vast sets of parameters specific to individual models, but they also employ different biophysical processes to simulate crop phenology, evapotranspiration (ET), or growth stresses, among others. This complicates the attribution of input effects to specific structural differences that affect biomass and yield formation. Excess water, for example, can affect root aeration, causing root oxygen stress and/or increasing nutrient deficits through amplified nutrient leaching. Where both processes are present in a model, further diagnostic simulation outputs directly reflecting outcomes of these processes will be required to identify the dominating underlying processes, ideally at sub-seasonal to daily time steps and within well-controlled experiments. Such efforts are presently ongoing in various field-scale ensemble studies of AgMIP. At global scales, in turn, even GGCMs from the same model family or using the same core model, such as EPIC or DSSAT, often show large discrepancies in both crop yields (Folberth et al., 2019; Müller et al., 2017, 2024) and externalities (Wartenburger et al., 2018), indicating that model setups and parameterizations can outweigh structural differences among GGCMs. Our findings herein reaffirm this. We hence suggest that GGCM ensembles may substantially profit from engaging in field-scale studies to thoroughly evaluate responses at the process-level, while our results can inform what additional climatic drivers should receive more scrutiny.

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Finally, crop cultivars and maturation implemented in the simulations vary substantially across GGCMs (Figure S25 in Supporting Information S1), 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 analyses 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.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

Data sets 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, IIASA, ISRIC, ISSCAS, JRC (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 Oberleitner (2026).

Acknowledgments

T.O., C.F., A.B., J.B., N.K., and R.S. were supported by the Austrian Science Fund (FWF) (Grant 10.55776/P36220). C.F. and T.O. received funding from the Future of Life Institute (project ANFOS). A.C.R. and J.R.G. received support from the NASA Earth Sciences Division via the GISS Impacts Group. M.O. was supported by the Environment Research and Technology Development Fund JPMEERF20242001 of the Environmental Restoration and Conservation Agency provided by the Ministry of the Environment of Japan.

References

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

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. *Scientific Data*, *5*(1), 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. *Nature Communications*, *9*(1), 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).

Bundy, L. R., Gensini, V. A., & Russo, M. S. (2022). Insured corn losses in the United States from weather and climate perils. *Journal of Applied Meteorology and Climatology*, *61*(8), 969–988. https://doi.org/10.1175/JAMC-D-21-0245.1

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., et al. (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., et al. (2020). WFDE5: Bias-adjusted ERA5 reanalysis data for impact studies [Dataset]. *Earth System Science Data*, *12*(3), 2097–2120. https://doi.org/10.5194/essd-12-2097-2020

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., et al. (2015). The global gridded crop model intercomparison: Data and modeling protocols for phase 1 (V1.0). *Geoscientific Model Development*, *8*(2), 261–277. https://doi.org/10.5194/gmd-8-261-2015

FAO, IIASA, ISRIC, ISSCAS, JRC. (2012). Harmonized world soil database (version 1.2) [Dataset]. Retrieved from https://www.sidalc.net/search/Record/dig-wur-nl-wurpubs-500737/Description

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., et al. (2021). POT: Python optimal transport. *Journal of Machine Learning Research*, *22*, 1–8.

Folberth, C., Baklanov, A., Khabarov, N., Oberleitner, T., Balkovič, J., & Skalský, R. (2025). CROMES v1.0: A flexible CROp model emulator suite for climate impact assessment. *Geoscientific Model Development*, *18*(17), 5759–5779. https://doi.org/10.5194/gmd-18-5759-2025

Folberth, C., Elliott, J., Müller, C., Balkovič, J., Chryssanthacopoulos, J., Izaurralde, R. C., et al. (2019). Parameterization-induced uncertainties and impacts of crop management harmonization in a global gridded crop model ensemble. *PLoS One*, *14*(9), e0221862. https://doi.org/10.1371/journal.pone.0221862

Franke, J., Müller, C., Elliott, J., Ruane, A. C., Jagermeyr, J., Balkovic, J., et al. (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 (GMDD)*, 1–30. https://doi.org/10.5194/gmd-2019-237

Franke, J. A., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Balkovic, J., et al. (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*(5), 2315–2336. https://doi.org/10.5194/gmd-13-2315-2020

Frieler, K., Schauberger, B., Arneth, A., Balkovič, J., Chryssanthacopoulos, J., Deryng, D., et al. (2017). Understanding the weather signal in national crop-yield variability. *Earth's Future*, *5*(6), 605–616. https://doi.org/10.1002/2016EF000525

Frieler, K., Volkholz, J., Lange, S., Schewe, J., Mengel, M., del Rocío Rivas López, M., et al. (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

Garcia-Vila, M., dos Santos Vianna, M., Harrison, M. T., Liu, K., De, S., Nóia-Júnior, R., et al. (2025). Gaps and strategies for accurate simulation of waterlogging impacts on crop productivity. *Nature Food*, *6*(6), 1–10. https://doi.org/10.1038/s43016-025-01179-y

Gebrechorkos, S. H., Sheffield, J., Vicente-Serrano, S. M., Funk, C., Miralles, D. G., Peng, J., et al. (2025). Warming accelerates global drought severity. *Nature*, 1–8. https://doi.org/10.1038/s41586-025-09047-2

Grassini, P., Specht, J. E., Tollenaar, M., Ciampitti, I., & Cassman, K. G. (2015). High-yield maize–soybean cropping systems in the US corn belt. In *Crop physiology* (pp. 17–41). Elsevier. https://doi.org/10.1016/B978-0-12-417104-6.00002-9

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7

Iizumi, T., & Sakai, T. (2020). The global dataset of historical yields for major crops 1981–2016 [Dataset]. *Scientific Data*, *7*(1). https://doi.org/10.1038/s41597-020-0433-7.97.

Iizumi, T., Sakai, T., Masaki, Y., Oyoshi, K., Takimoto, T., Shiogama, H., et al. (2025). Assessing the capacity of agricultural research and development to increase the stability of global crop yields under climate change. *PNAS Nexus*, *4*, pgaf099. https://doi.org/10.1093/pnasnexus/pgaf099

International Food Policy Research Institute. (2020). Global spatially-disaggregated crop production statistics data for 2010 version 2.0 [Dataset]. https://doi.org/10.7910/DVN/PRFF8V

Jägermeyr, J., Lin, T.-S., Rabin, S., Balkovic, J., Elliott, J. W., Faye, B., et al. (2024). ISIMIP3a simulation data from the agriculture sector [Dataset]. https://doi.org/10.48364/ISIMIP.370868.1

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

Jägermeyr, J., Robock, A., Elliott, J., Müller, C., Xia, L., Khabarov, N., et al. (2020). A regional nuclear conflict would compromise global food security. *PNAS*, *117*(13), 7071–7081. https://doi.org/10.1073/pnas.1919049117

Jiang, S., Sweet, L., Blougouras, G., Brenning, A., Li, W., Reichstein, M., et al. (2024). How interpretable machine learning can benefit process understanding in the geosciences. *Earth's Future*, *12*(7), e2024EF004540. https://doi.org/10.1029/2024EF004540

Lange, S. (2019). WFDE5 over land merged with ERA5 over the ocean (W5E5) [Dataset]. 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*(7), 2325–2337. https://doi.org/10.1111/gcb.14628

Liu, B., Asseng, S., Müller, C., Ewert, F., Elliott, J., Lobell, D. B., et al. (2016). Similar estimates of temperature impacts on global wheat yield by three independent methods. *Nature Climate Change*, *6*(12), 1130–1136. https://doi.org/10.1038/nclimate3115

Liu, W., Yang, H., Folberth, C., Wang, X., Luo, Q., & Schulin, R. (2016). Global investigation of impacts of PET methods on simulating crop-water relations for maize. *Agricultural and Forest Meteorology*, *221*, 164–175. https://doi.org/10.1016/j.agrformet.2016.02.017

Liu, K., Harrison, M. T., Yan, H., Liu, D. L., Meinke, H., Hoogenboom, G., et al. (2023). Silver lining to a climate crisis in multiple prospects for alleviating crop waterlogging under future climates. *Nature Communications*, *14*(1), 765. https://doi.org/10.1038/s41467-023-36129-4

Ljung, L. (1999). *System identification: Theory for the user, 2nd ed. Prentice hall information and system sciences series*. Prentice Hall PTR.

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., et al. (2015). Multimodel ensembles of wheat growth: Many models are better than one. *Global Change Biology*, *21*(2), 911–925. https://doi.org/10.1111/gcb.12768

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. *Nature Geoscience*, *16*(5), 410–415. https://doi.org/10.1038/s41561-023-01177-4

Mistry, M. (2019). A high-resolution global gridded historical dataset of climate extreme indices. *Data*, *4*(1), 41. https://doi.org/10.3390/data4010041

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Molina Bacca, E. J., Stevanović, M., Bodirsky, B. L., Karstens, K., Chen, D. M.-C., Leip, D., et al. (2023). Uncertainty in land-use adaptation persists despite crop model projections showing lower impacts under high warming. *Communications Earth & Environment*, *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., et al. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond explainable AI, lecture notes in computer science* (pp. 39–68). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_4

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

Müller, C., Elliott, J., Kelly, D., Arneth, A., Balkovic, J., Ciais, P., et al. (2019). The global gridded crop model intercomparison phase 1 simulation dataset. *Scientific Data*, *6*, 50. https://doi.org/10.1038/s41597-019-0023-8

Müller, C., Franke, J., Jägermeyr, J., Ruane, A. C., Elliott, J., Moyer, E., et al. (2021). Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. *Environmental Research Letters*, *16*(3), 034040. https://doi.org/10.1088/1748-9326/abd8fc

Müller, C., Jägermeyr, J., Franke, J. A., Ruane, A. C., Balkovic, J., Ciais, P., et al. (2024). Substantial differences in crop yield sensitivities between models call for functionality-based model evaluation. *Earth's Future*, *12*(3), e2023EF003773. https://doi.org/10.1029/2023EF003773

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. https://doi.org/10.48550/ARXIV.1109.2378

Nóia Júnior, R. D. S., Asseng, S., García-Vila, M., Liu, K., Stocca, V., Dos Santos Vianna, M., et al. (2023). A call to action for global research on the implications of waterlogging for wheat growth and yield. *Agricultural Water Management*, *284*, 108334. https://doi.org/10.1016/j.agwat.2023.108334

Oberleitner, T. (2026). Global gridded crop model feature attribution pipeline (version v1.0.0) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.22910743

Orlov, A., Jägermeyr, J., Müller, C., Daloz, A. S., Zabel, F., Minoli, S., et al. (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

Ortez, O. A., Lindsey, A. J., Thomison, P. R., Coulter, J. A., Singh, M. P., Carrijo, D. R., et al. (2023). Corn response to long-term seasonal weather stressors: A review. *Crop Science*, *63*(6), 3210–3235. https://doi.org/10.1002/csc2.21101

Rezaei, E. E., Webber, H., Asseng, S., Boote, K., Durand, J. L., Ewert, F., et al. (2023). Climate change impacts on crop yields. *Nature Reviews Earth & Environment*, *4*(12), 1–16. https://doi.org/10.1038/s43017-023-00491-0

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

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., et al. (2013). The agricultural model intercomparison and improvement project (AgMIP): Protocols and pilot studies. *Agricultural and Forest Meteorology, 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., et al. (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. Wiley paperback series*. Wiley.

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

Schewe, J., Gosling, S. N., Reyer, C., Zhao, F., Ciais, P., Elliott, J., et al. (2019). State-of-the-art global models underestimate impacts from climate extremes. *Nature Communications*, *10*(1), 1005. https://doi.org/10.1038/s41467-019-08745-6

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

Schmidt-Traub, G., Obersteiner, M., & Mosnier, A. (2019). Fix the broken food system in three steps. *Nature*, *569*(7755), 181–183. https://doi.org/10.1038/d41586-019-01420-2

Sweet, L., Athanasiadis, I. N., van Bree, R., Castellano, A., Martre, P., Paudel, D., et al. (2025). Transdisciplinary coordination is essential for advancing agricultural modeling with machine learning. *One Earth*, *8*(4), 101233. https://doi.org/10.1016/j.oneear.2025.101233

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

Villani, C. (2009). The Wasserstein distances. In *Optimal Transport, Grundlehren Der Mathematischen Wissenschaften* (pp. 93–111). Springer. 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., et al. (2019). The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, *14*(5), 054010. https://doi.org/10.1088/1748-9326/ab154b

Volkholz, J., & Müller, C. (2020). ISIMIP3 soil input data [Dataset]. https://doi.org/10.48364/ISIMIP.942125

Wang, E., Martre, P., Zhao, Z., Ewert, F., Maiorano, A., Rötter, R. P., et al. (2017). The uncertainty of crop yield projections is reduced by improved temperature response functions. *Nature Plants*, *3*(8), 17102. https://doi.org/10.1038/nplants.2017.102

Wartenburger, R., Seneviratne, S. I., Hirschi, M., Chang, J., Ciais, P., Deryng, D., et al. (2018). Evapotranspiration simulations in ISIMIP2a—Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets. *Environmental Research Letters*, *13*(7), 075001. https://doi.org/10.1088/1748-9326/aac4bb

Webber, H., Cooke, D., Wang, C., Asseng, S., Martre, P., Ewert, F., et al. (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., 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. *Nature Sustainability*, *6*(10), 1–9. https://doi.org/10.1038/s41893-023-01152-2

Williams, J. R., Jones, C. A., Kiniry, J. R., & Spanel, D. A. (1989). The EPIC crop growth model. *Transactions of the ASAE*, *32*(2), 0497–0511. https://doi.org/10.13031/2013.31032

Xiao, L., Liu, L., Asseng, S., Xia, Y., Tang, L., Liu, B., et al. (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., et al. (2020). A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps [Dataset]. *Earth System Science Data*, *12*(4), 3545–3572. https://doi.org/10.5194/essd-12-3545-2020

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Zhao, N., Yu, J. Y., Dzieciolowski, K., & Bui, T. (2024). Error analysis of shapley value-based model explanations: An informative perspective. In G. Avni, M. Giacobbe, T. T. Johnson, G. Katz, A. Lukina, N. Narodytska, et al. (Eds.), *AI verification, lecture notes in computer science* (pp. 29–48). Springer Nature. 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*(5), e2020EF001815. https://doi.org/10.1029/2020EF001815

References From the Supporting Information

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

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* (pp. 1–8). ACM. https://doi.org/10.1145/2484762.2484814

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 System Dynamics*, *11*(3), 641–652. https://doi.org/10.5194/esd-11-641-2020

Haas, E., Klatt, S., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., et al. (2013). LandscapeDNDC: A process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. *Landscape Ecology*, *28*(4), 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*(4), 3934–3965. https://doi.org/10.3390/rs70403934

Iizumi, T., Furuya, J., Shen, Z., Kim, W., Okada, M., Fujimori, S., et al. (2017). Responses of crop yield growth to global temperature and socioeconomic changes. *Scientific Reports*, *7*(1), 7800. https://doi.org/10.1038/s41598-017-08214-4

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., et al. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, *18*(3–4), 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7

Jägermeyr, J., & Frieler, K. (2018). Spatial variations in crop growing seasons pivotal to reproduce global fluctuations in maize and wheat yields. *Science Advances*, *4*(11), eaat4517. https://doi.org/10.1126/sciadv.aat4517

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. *Nature Communications*, *6*(1), 8946. https://doi.org/10.1038/ncomms9946

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. *Hydrology and Earth System Sciences*, *26*(4), 923–940. https://doi.org/10.5194/hess-26-923-2022

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*(9), 1207–1220. https://doi.org/10.1029/2017EF000763

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*, *10*(6), 621–637. https://doi.org/10.1046/j.1466-822x.2001.00256.x

Sultan, B., Defrance, D., & Iizumi, T. (2019). Evidence of crop production losses in West Africa due to historical global warming in two crop models. *Scientific Reports*, *9*(1), 12834. https://doi.org/10.1038/s41598-019-49167-0

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

Webber, H., Ewert, F., Olesen, J. E., Müller, C., Fronzek, S., Ruane, A. C., et al. (2018). Diverging importance of drought stress for maize and winter wheat in Europe. *Nature Communications*, *9*(1), 4249. https://doi.org/10.1038/s41467-018-06525-2

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. *Nature Communications*, *10*(1), 2844. https://doi.org/10.1038/s41467-019-10775-z

23284277, 2026, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025EF007457 by CochraneAustria, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License