https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

A data-driven method for identifying climate drivers of agricultural yield failure from daily weather data

Lily-belle Sweet[1,2], Christoph Müller[3], Jonas Jägermeyr[3,4,5], and Jakob Zscheischler[1,2,6]

[1]Department of Compound Environmental Risks, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
[2]Department of Hydro Sciences, TUD Dresden University of Technology, Dresden, Germany
[3]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany
[4]Columbia University, Climate School, New York, NY, USA
[5]NASA Goddard Institute for Space Studies (GISS), New York, NY, USA
[6]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden-Leipzig, Germany

Correspondence: Lily-belle Sweet (lily-belle.sweet@ufz.de)

Abstract. Climate-related impacts, such as agricultural yield failure, often occur in response to a range of specific weather conditions taking place across different time periods, such as during the growing season. Identifying which weather conditions and timings are most strongly associated with a certain impact is difficult because of the overwhelming number of possible predictor combinations from different aggregation periods. Here we address this challenge and introduce a method for identifying a small number of climate drivers of an impact from high-resolution meteorological data. Based on the principle that causal drivers should generalize across different environments, our proposed two-stage approach systematically generates, tests, and discards candidate features using machine learning and then generates a set of robust drivers. We evaluate the method using simulated US maize yield data from two process-based global gridded crop models and rigorous out-of-sample testing (using approximately 30 years of early 20th-century climate and yield data for training and over 70 years of subsequent data for testing). The climate drivers identified align with crop model mechanisms and consistently use only the weather variables that are taken as input by the respective models. Logistic regression models using ten drivers as predictors show strong predictive performance on the held-out test period even under shifting climatic conditions, achieving correlations of 0.70-0.85 between predicted and true annual proportions of grid cells experiencing yield failure. This approach circumvents the limitations of post-hoc interpretability in black-box machine learning models, allowing researchers to use parsimonious statistical models to explore relationships between climate and impacts, while still harnessing the predictive power of high-resolution, multivariate weather data. We demonstrate this method in the context of agricultural yield failure, but it is also applicable for studying other climate-related impacts such as forest die-off, wildfire incidents, landslides, or flooding.

1 Introduction

Climate impacts to human and natural systems are often highly dependent on the weather conditions experienced during specific time periods. For instance, annual crop yields are most influenced by the climatic conditions during the growing season, and may be more sensitive to certain weather conditions at different growth stages. For maize, up to 60% of the variance in

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

US yields at national or county level has been explained by climate variables (Frieler et al., 2017; Mistry et al., 2017), and process-based crop models, which encapsulate the day-by-day biophysical plant response to weather conditions, have been able to explain up to 79% of US maize yield variation (Müller et al., 2017). However, current models have been found to exhibit very different responses to climate drivers (Müller et al., 2024) and vary widely in their projections of climate change impacts (Jägermeyr et al., 2021; Wang et al., 2024). Crop models generally include heat and drought stress effects, as well as other temperature and moisture responses, but have been shown to underestimate the impact of such extremes (Heinicke et al., 2022). Researchers have argued that models do not consider the influence of compound stressors on yields due to knowledge gaps (Webber et al., 2022), and that improved scientific understanding of these relationships is needed to inform and constrain the next generation of crop models (Kim et al., 2025; Nóia Júnior et al., 2025).

The impacts of weather on agricultural yields are complex and nonlinear (Siebert et al., 2017a). High temperatures during growing seasons in the US have been consistently linked with negative maize yield effects (Schlenker and Roberts, 2009; Schauberger et al., 2017; Zhao et al., 2017), but the impact is reduced by irrigation (Troy et al., 2015; Siebert et al., 2017b) and can be mitigated or exacerbated by the co-occurring level of precipitation, soil moisture or evapotranspiration (Lobell et al., 2013; Jin et al., 2017; Matiu et al., 2017; Mistry et al., 2017; Rigden et al., 2020; Haqiqi et al., 2021; Lesk et al., 2021; Ting et al., 2023). Additionally, as plant phenology develops, their sensitivity to weather conditions changes. For example, the effect of short-term heatwaves on yields is more severe during the reproductive or grain-filling period of maize development than the vegetative period (Siebers et al., 2017; Zhu et al., 2019).

Machine learning (ML) models, which can capture highly complex and nonlinear relationships and are well-suited for high-dimensional tasks, are also being explored for agricultural yield forecasting and modelling (Sweet et al., 2025). A wide range of approaches have been able to predict US maize yields with good performance (Crane-Droesch, 2018; Martínez-Ferrer et al., 2020; Lin et al., 2020; Liu et al., 2022). However, just like statistical models, ML models are vulnerable to confounding effects, and their increased flexibility allows them to more easily overfit to spatiotemporal dependencies. This leads to poor predictive performance in years or regions outside of the training data (Ferraciolli et al., 2019; Shahhosseini et al., 2019; Silva et al., 2023; Sweet et al., 2023; Filippi et al., 2024), suggesting that the relationships learned by the models are not robust, and therefore unlikely to reflect the underlying physical processes. 'Knowledge-guided' or 'hybrid' ML, in which known process understanding is embedded into the model structure or training procedure, can help ameliorate this, but can also potentially degrade model performance if the constraints used are not well-specified (Moon et al., 2023; Han et al., 2025). 'Post hoc' interpretable or explainable techniques, in which the structure of a trained model or its response to interventions on input variables is analysed, have also been used to study the relationships learned by complex ML models and thereby improve process understanding (Shahhosseini et al., 2019; Hoffman et al., 2020; Ryo, 2022; Jiang et al., 2024a, b). However, in studies employing ML to analyse relationships between climate and impacts, variables are used, in general, if it is plausible that they could play a causal role. This means that spurious relationships learned by the model, or inaccurate interpretations, are difficult to detect. This could be a substantial issue given that studies have shown commonly-used methodologies to give contradictory results (Lischeid et al., 2022; Mamalakis et al., 2022), return spurious or physically-implausible relationships (Wadoux et al., 2020; Sweet et al., 2023), or even fail to out-perform random guessing (Bilodeau et al., 2024).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Statistical models, on the other hand, are often inherently interpretable, and have been frequently used to study the relationships between weather and crop yield (Vogel et al., 2019, 2021; Heilemann et al., 2024). Before fitting these models, a parsimonious set of predictive variables must be chosen, which can significantly impact model behaviour (Chen et al., 2024). In most studies that aim to predict climate-related yield variability, climate conditions are approximated by the annual or growing-season mean precipitation or temperature, or by indicators selected based on expert knowledge, such as the number of growing-degree days or extreme degree days, the maximum or minimum temperature experienced, and the average diurnal range. However, this can obscure relationships that are not captured by the chosen aggregate variables, such as the effect of short-term weather events at different phenological stages. Selecting a parsimonious set of features that reflect the underlying biophysical relationships requires a high level of domain expertise, and does not allow for the discovery of unexpected or more specific climate drivers.

We propose a data-driven method for obtaining simple, human-understandable climate drivers of yield failure, inspired by the idea that causal drivers should generalise to unseen environments (Peters et al., 2016; Richens and Everitt, 2024). This allows researchers to use parsimonious, interpretable models to explore relationships between weather and impact, while still making use of the information available in high-dimensional, multivariate climate data, rather than using non-robust post hoc interpretation methods to approximate the functionality of trained black-box ML models. To validate the method's performance and robustness, we make use of simulated US maize yields from global gridded process-based crop models (pDSSAT and LPJmL). This allows us to compare the identified drivers to known model mechanisms, and to test the predictive skill of models using these drivers under conditions that reflect the challenges inherent to observational studies, such as correlation between predictive variables, sampling bias from the increased availability of data in locations where climate conditions favour higher yields, and distribution shifts induced by the changing climate. Our testbed consists of simulations driven by climate reanalysis spanning 1902-2016, but we use only the first 30% of the timeseries to identify drivers and train models and then test their performance over the last 70%. This train-test split is more rigorous than in most ML studies and allows us to examine the suitability of the method for conducting similar analyses using observational data under the influence of climate change. We then demonstrate this by applying the method to county-level reported US maize yield and meteorological data.

2 Data

We use gridded simulations of US rainfed maize yield from two process-based crop models: LPJmL (von Bloh et al., 2018; Lutz et al., 2019) and pDSSAT (Jägermeyr et al., 2021). LPJmL is a dynamic global vegetation model, designed to capture terrestrial carbon and water cycles of natural and agricultural systems. pDSSAT (Jones et al., 2003; Elliott et al., 2014), on the other hand, has been developed to simulate field-scale processes, with a focus on the interactions between crop development, soil, atmosphere and management factors.

Both crop models were driven by daily climate data spanning 1901-2019, at 0.5 degree spatial resolution, based on reanalysis from W5E5 v2.0 (Cucchi et al., 2020; Lange et al., 2021) and GSWP3 v1.09 (Dirmeyer et al., 2006; Hyungjun Kim, 2017). The data was homogenised to W5E5 for 1901-1978 using the ISIMIP3BASD v2.5.0 bias adjustment method (Lange, 2019, 2021).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

The simulations were produced according to the ISIMIP3a protocol and all forcing data, as well as the corresponding simulated yields, are publicly available from the ISIMIP data repository (Lange et al., 2022). Sowing dates are based on a combination of observational data products (Jägermeyr et al., 2021) and vary spatially but not year-to-year. The models were also given annual atmospheric CO_2 concentrations and soil parameters, which we do not consider for this study. We restrict our study area to current maize cropping areas of the continental US, which we define as grid cells containing at least ten harvested hectares according to the MIRCA2000 data product (Portmann et al., 2010).

We use ten climate variables at daily resolution: near-surface relative humidity ($hurs$, %), near-surface specific humidity ($huss$, kgkg^{-1}), precipitation (pr, mm), surface air pressure (ps, Pa), surface downwelling longwave radiation ($rlds$, Wm^{-2}) and shortwave radiation ($rsds$, Wm^{-2}), near-surface windspeed ($sfcwind$, ms^{-1}), near-surface air temperature (tas, °C), daily minimum ($tasmin$, °C) and maximum near-surface air temperature ($tasmax$, °C). However, each crop model only considers a subset of those variables; both models take pr, $tasmin$, $tasmax$, $rsds$, and $rlds$ as input, and LPJmL additionally uses $sfcwind$ and tas. pDSSAT does not use tas, but internally estimates hourly temperatures based on daily $tasmin$ and $tasmax$ values.

Annual county-level observed maize yields are obtained from the United States Department of Agriculture National Agricultural Statistics Service (USDA NASS) for 1983 to 2018 (United States Department of Agriculture (USDA), 2025), and county-level boundaries are obtained from the US Census Bureau's 2018 TIGER/Line shapefiles (U.S. Census Bureau, 2018). Yields are detrended by subtracting a five-year rolling mean, and counties are removed from the dataset if, after detrending, fewer than ten years of yield data are available. We limit the dataset to counties in which maize harvested areas are predominantly rainfed and at least 100 hectares are harvested (according to MIRCA2000), resulting in 551 total counties. In the majority of counties, yields are not reported separately for rainfed or irrigated areas. For counties in which yields are reported separately, rainfed yields are used only if available for all years. Otherwise, and for all other counties, overall reported yields are used.

Observed daily meteorological data over the same time period, at 4 km spatial resolution, is obtained from the PRISM climate group (PRISM Group, 2018) and aggregated to county-level using a weighted average of overlapping grid cells based on enclosed cropping area. Variables used consist of daily minimum and maximum temperature ($^\circ C$), minimum and maximum vapour pressure deficit (hPa), precipitation (mm) and dewpoint temperature ($^\circ C$). Daily mean temperature and vapour pressure deficit is estimated by averaging the minimum and maximum values, resulting in a total of eight meteorological variables.

3 Methods

3.1 Setup

Each datapoint in our compiled dataset consists of a binary target variable and the corresponding daily multivariate climate input for one growing season at one grid cell. The target variable is maize yield failure, which we define as any yield below the 10th percentile observed during the training period at that grid cell. We test our method on both non-detrended and detrended yields (based on subtracting a seven-year rolling mean, meaning that we discard datapoints from the first six years). The dataset

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

is split into a training set (the first 30% of growing seasons, covering 1902-1937 for non-detrended yields and 1907-1940 for detrended) and test set (the last 70% of growing seasons, covering 1938-2016 and 1941-2016). Note that, for non-detrended LPJmL yields, the 10th percentile over the training years was zero for 167 grid cells. These low-yielding areas are removed from the training and test set. After processing, the training sets are made up of 62,000 to 78,000 datapoints and the test sets consist of 145,000 to 173,000.

Our predictive variables derive from the same daily multivariate climate data that was used to drive the crop models. For each grid cell and year, we take 90 days of each climate variable before the planting date (to consider e.g. lagged soil moisture effects) and 240 days after, resulting in a 3300-dimensional dataset (as we have ten climate variables). To use an interpretable model with this data, a smaller number of climate drivers must first be defined (generally, by applying aggregation functions over specified time intervals - for instance, the monthly maximum temperatures or number of growing degree days during the growing season). This feature selection step is normally based on expert knowledge. Instead, our proposed method extracts a small number of simple climate drivers of yield failure from daily multivariate weather data based on their robust predictive power in held-out time periods.

3.2 Approach

The method for identifying climate drivers consists of three steps (Fig. 1): first, we generate multiple pools of candidate features by using a set of chosen aggregations over the daily climate variable data from 30 randomly-sampled time intervals (with minimum duration of two weeks); second, we select ten features from each pool using sequential forward feature selection, based on the predictive performance of ML models on held-out time periods; finally, we collect the selected features from the pools, and extract a set of ten condensed climate drivers from that collection using agglomerative clustering, based on the aggregation methods, variables and periods of the growing season which are selected most frequently.

When selecting features from each pool (step 2), we use an extremely randomised trees model (extra trees). This model consists of an ensemble of decision trees, similar to the more widely-known random forest, but rather than identifying the optimal split value for each feature at each node, the thresholds are selected at random. Extra trees models are therefore faster to train, and can result in smoother functional relationships when predictions are averaged across trees, leading to improved out-of-sample performance (Geurts et al., 2006). Like other tree-based models, they handle nonlinear relationships well, are robust to outliers and skewed distributions, are computationally efficient, and can be trained in parallel. We do not tune hyperparameters; all models used consist of 50 trees of maximum depth 30. At each split, half of the predictive variables are considered, and each leaf node must contain a minimum weighted fraction of 0.0001 of all samples (with samples weighted inversely proportionally to class frequency). Features are selected sequentially based on optimising model performance using ten-fold temporal cross-validation, reflecting our training and test split as described above, with each test fold consisting of a held-out period of three or four consecutive years.

To obtain a final set of ten climate drivers of yield failure we apply agglomerative clustering to the selected features from 100 pools of candidates (step 3). Features are only clustered together if they substantially overlap in time and use the same aggregation method and climate variable. The ten largest clusters are then condensed into single drivers by using the 10th

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

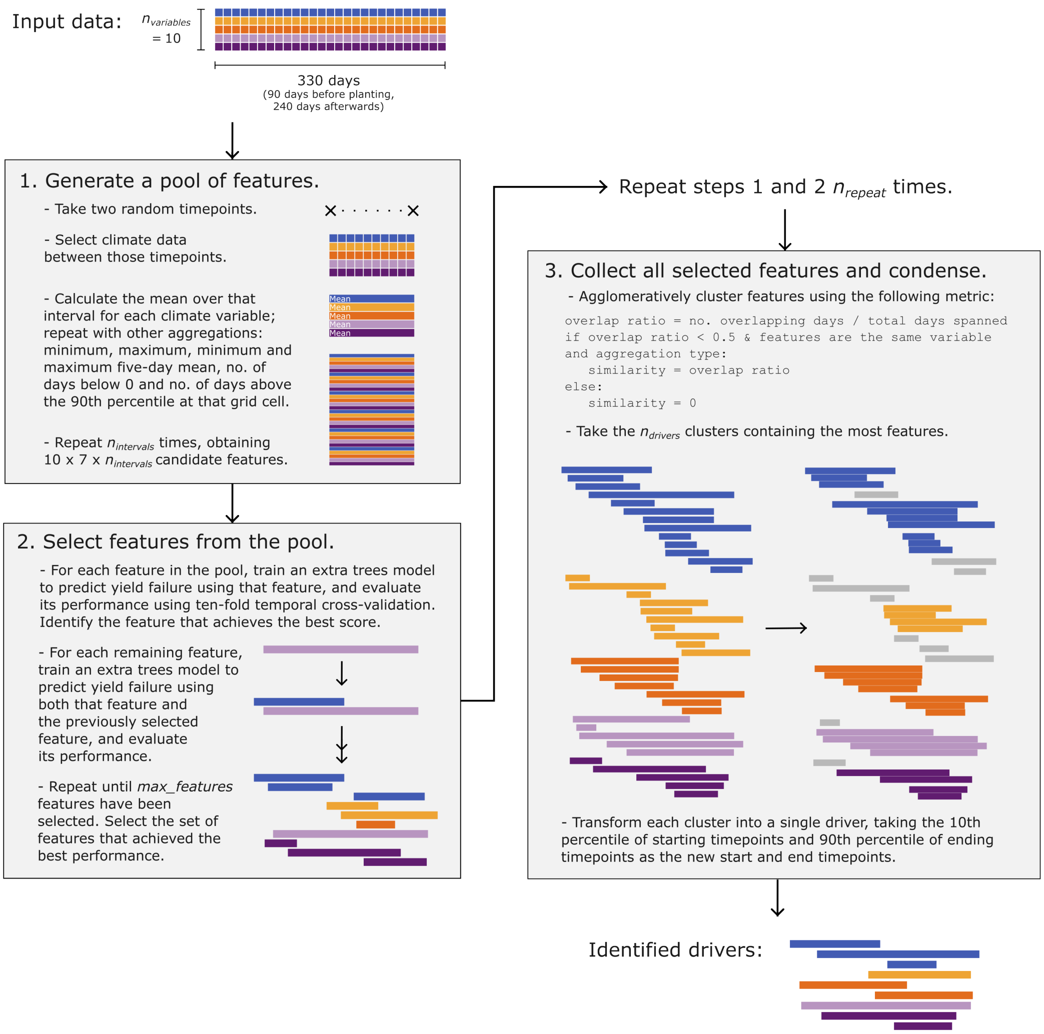

Figure 1. Flowchart illustrating our proposed method to identify a small number of simple climate drivers from multivariate time series weather data. The method requires some parameters to be chosen: $n\_intervals$, the number of random time intervals to sample for each pool of candidate features; $max\_features$, the maximum number of features to select from each pool and $n\_drivers$, the number of condensed drivers desired. Additionally, the cross-validation strategy used when selecting features from each pool must be specified.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

percentile of starting timepoints and 90th percentile of ending timepoints within the cluster as the new start and end timepoints. To create simple and interpretable models, we then fit logistic regression models using the ten identified drivers. Climate drivers are identified and models are trained using the training data only. We perform this procedure (identify ten features, fit logistic regression model) for both crop models.

3.3 Model evaluation and sensitivity analyses

We evaluate model performance and the robustness of our approach to different parameter choices in several ways. The ability of the final logistic regression models to predict yield failure over the entire test period is assessed using four performance metrics: i) the area under the receiver operating characteristic curve (ROC AUC), a ranking metric that reflects the ability of the model to distinguish between binary classes at a range of classification thresholds; ii) the average precision, which measures the tradeoff between precision and recall at different classification thresholds; iii) the Brier score loss, which measures the mean squared difference between each predicted probability and the true (binary) outcome; iv) and the logistic loss (log loss), also known as the cross-entropy loss, which measures the negative log-likelihood of each predicted probability given the true outcome. Note that the prediction task is an imbalanced classification problem, with non-failure years occurring approximately nine times as frequently as yield-failure years. Additionally, we investigate the ability of the model to capture temporal patterns in yield failures over the entire region studied by calculating the percentage of grid cells each year predicted to be experiencing yield failure and comparing this to the simulated ‘ground truth’. As predictions from the logistic regression model range between 0 and 1, we select a threshold to define yield failure that achieves the maximal f-score (the harmonic mean of the precision and recall) over the training set.

We assess if drivers are plausible by checking whether they only include the variables used as input by the crop models (i.e., ps, $hurs$ and $huss$ should not appear, nor should $sfcwind$ for pDSSAT). We also compare their distributions between yield failure and non-yield failure years and assess their consistency with associations captured by interpretable models and expert judgement.

We further test the robustness of the approach by conducting 100 bootstrap repeats, sampling 20 pools (with replacement) from the 100 pools described above. For each repeat, we extract sets of ten climate drivers from the 20 pools, and evaluate the extent to which these sets consist of the variables actually taken as input by the crop models by counting the variables used by the features collected from each pool (in Step 2 of the driver identification method), and from the condensed drivers over all bootstrap repeats. Second, for each of the repeats, we construct models with 5, 10 or 15 identified drivers and evaluate their predictive performance over the test period, in comparison with commonly-used baseline feature sets.

The baselines used consist of mean climate conditions aggregated over different time intervals: either the growing-season (defined as 180 days after planting), or on a quarterly (three consecutive ninety-day intervals after planting) or monthly (six consecutive thirty-day intervals after planting) basis. These are calculated for a subset of the climate variables (pr, tas and $rsds$) or for all ten. In addition, we include five extreme indicators that are often used in data-driven agricultural yield models (Vogel et al., 2019, 2021). These are calculated over the first 180 days after planting, and consist of the minimum and maximum temperatures experienced, the maximum five-day precipitation sum, the number of days in which the temperature drops below

0 °C, and the number of days in which temperatures exceed the 90th percentile (as defined during the training period for each grid cell). This results in six baseline feature sets containing between 8 and 65 weather features.

As the number of predictors can affect model performance, to facilitate a fair comparison between the baselines and the sets of climate drivers identified using our approach, we use Lasso logistic regression models, which uses L1 regularisation to force the coefficients of less important features to 0, effectively performing feature selection. Predictors are first scaled by removing the median and scaling to the 10th and 90th quantile range (based on training-set statistics), and the value of the hyperparameter α, which controls the level of regularization, is selected using five-fold temporal cross-validation over the training years. We also compare results when including second-order interaction terms and when using more complex extra trees models (with the same hyperparameters as previously described).

Finally, we assess the robustness of the approach to a number of parameter and methodology choices. The cross-validation strategy chosen when selecting features from each pool is important (Sweet et al., 2023; Meyer et al., 2019, 2018). We explore the impact this choice has on the results by comparing model performance scores when using three other cross-validation strategies (random sampling, spatial clusters and clusters in feature-space) in addition to the temporal strategy described above. We also test the impact of using 5, 10 or 15 cross-validation folds. Furthermore, we conduct sensitivity tests in which several parameters of our method are varied both individually and in combination: the number of time intervals sampled in each pool of candidate features, the maximum number of features selected from each pool, and the number of drivers identified.

3.4 Demonstration on observational data

Observational data is split into training years (1983 - 2007) and test years (2008 - 2017). County-level yield failure years are determined based on falling below the tenth percentile of (detrended) yields over the training period. Using maize planting dates from WorldCereal (Franch et al., 2022), daily multivariate climate data is obtained for 90 days before planting and 240 subsequent for each year and county.

After processing, the method is applied on the training data to obtain four climate drivers in the same manner as described above. 100 pools of candidate features are obtained based on 20 sampled time intervals per pool, and ten features are selected from each using five-fold temporal cross-validation and extra trees models. The obtained drivers are then used to fit a logistic regression model to the training data, and performance is evaluated on data from the test years.

4 Results

4.1 Crop model simulations

Maize yields simulated by pDSSAT are more variable in space and time than LPJmL, but both simulations exhibit lower average yields in the western regions of the US and declining annual yields until the late 1930s, followed by an upward trend (Fig. 2a-c). As management and genetic changes over time are not considered in these simulations, this behaviour is the result of changing climate conditions, and in fact these trends coincide with opposing shifts in average growing-season temperatures

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

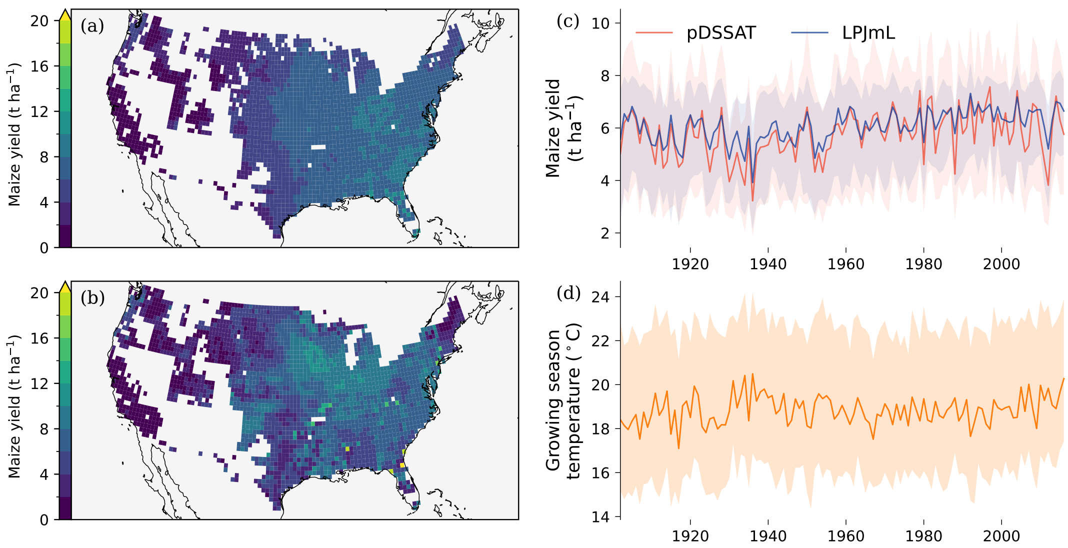

Figure 2. Overview of the rainfed maize yield simulations from both crop models. (a), (b) Average yields over the entire time period studied, for LPJmL and pDSSAT respectively, for grid cells included in the dataset. (c) Distributions of yields from year to year, with darker lines denoting the median yield across all grid cells for each crop model, and shaded regions covering the interquartile ranges. (d) Distribution of annual growing season temperatures (here defined as the average daily mean temperature in the first 180 days after planting) in the corresponding climate forcing data.

(Fig. 2d). The high spatial variability in simulated yields and temperatures suggests that a space-for-time substitution may be able to improve the ability of data-driven models to extrapolate to warmer (future) years. Despite increasing temperatures in later decades, most locations experiencing an unprecedentedly-warm growing season are unlikely to be outside of the range of previously-experienced conditions at all grid cells until that year.

Composite plots of the growing-season daily climate conditions for normal and yield-failure years suggest that non-causal climate variables are associated with yield failure in these simulations for both crop models (Fig. 3, with further variables in Fig. A1). For example, near-surface relative humidity is not used as input for LPJmL or pDSSAT, but yield failure years tend to have drier growing seasons in both crop models (Fig. 3g,h). Composite plots of growing-season windspeed show an association between lower winds and yield failure for both models (Fig. A1i,j), despite the fact that windspeed is an input variable for LPJmL but not pDSSAT. Yield-failure years tend to have higher growing-season temperatures, but there is a wide overlap in temperature conditions between failure and non-failure years (Fig. 3c,d). Similarly, periods of low rainfall in the first four months after sowing are associated with yield failure, but many non-failure years also experience periods of low rainfall at that time (Fig. 3e,f). Furthermore, while simulations from the two crop models differ in the proportion of grid cells experiencing yield failure by up to 10% in some years (Fig. 3a,b), relationships suggested by all composite plots are

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

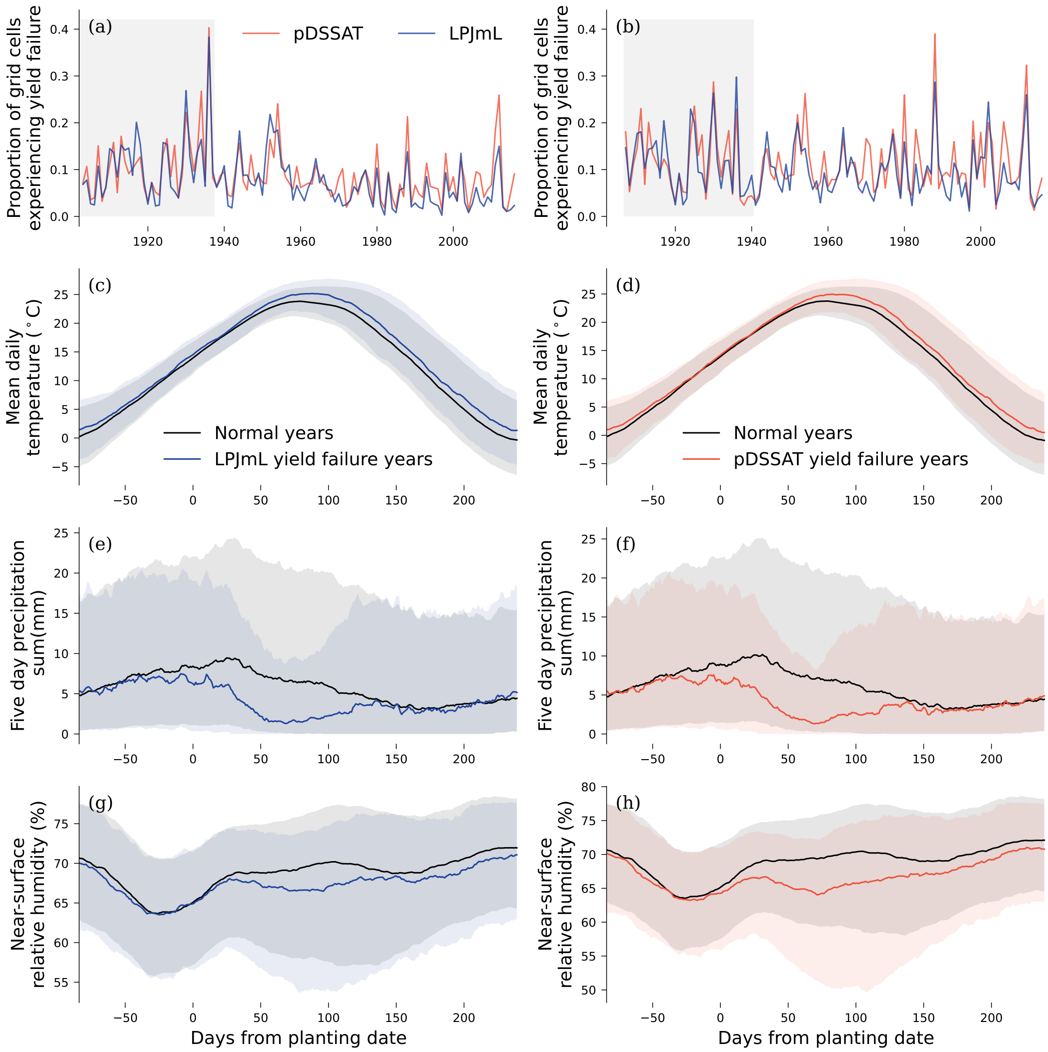

Figure 3. (a) Proportion of grid cells experiencing yield failure for each year of the studied period. Pale grey shaded years denote the training period used for identifying drivers and training predictive models. All other years are held out and used to evaluate model performance. (b) Same as (a) after yields are detrended based on subtracting the seven-year rolling average. (c)-(h) Composite plots of mean daily temperature, five day precipitation sums and near-surface humidity ((c), (e), and (g) for LPJmL and (d), (f) and (h) for pDSSAT). Solid lines denote the median daily climate conditions over all years and locations, for either normal or yield failure years, and shaded regions indicate the interquartile range. Mean daily temperature and near-surface relative humidity are smoothed by taking the seven-day rolling mean.

visually similar. Composite plots of growing-season daily climate conditions for normal and yield-failure years when yields are detrended are very similar (Figs. B1, B2).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

4.2 Resulting models perform well on held-out years and are interpretable

A logistic regression model fit on the training data using ten identified climate drivers of pDSSAT yield failure as predictors is able to achieve a ROC AUC of 0.84, average precision of 0.36, Brier score of 0.059 and log loss of 0.21 on the test years. Additionally, the model captures the year-to-year variability in yield failure well (Fig. 4a), with a Pearson correlation of 0.84 between the true and predicted annual proportions of grid cells experiencing yield failure. Notably, model performance does not appear to deteriorate over time: the year in which the highest average precision is achieved is more than half a century later than the end of the training period. Maps of predictions and ground truth in the year where the model achieves the highest average precision (2011; Fig. 4b,c) show general agreement in spatial variability. In the year with lowest average precision (1997; 4d,e), the model is able to accurately predict the (very low) proportion of grid cells that experience yield failure in that year. A logistic regression model using ten climate drivers for LPJmL is able to achieve similarly high performance, with test ROC AUC of 0.83, average precision of 0.35, Brier score of 0.051 and log loss of 0.20 (Fig. A2). A Pearson correlation of 0.79 is achieved between true and predicted annual proportions of grid cells experiencing yield failure. In 1952, the year in which the model achieves highest average precision (Fig. A2b,c), predictions largely captures the spatial variability of the ground truth, but the proportion of grid cells experiencing yield failure (using a threshold defined based on training set performance) is overestimated. The annual yield failure prevalence in general appears to be overestimated for LPJmL, in contrast to pDSSAT. As the Brier scores achieved by the models are similar, this could be attributed to the threshold used to define yield failure. Similar levels of model performance are achieved for both crop models when yields are detrended before defining yield failure (Figs. B3, B4).

The ten climate drivers of yield failure identified by our method for each crop model consist of variables that are taken as input by those respective models, with the exception of one driver based on mean daily temperature for pDSSAT (Fig. 5). However, as pDSSAT takes as input the daily minimum and maximum temperatures and uses them to estimate hourly average temperatures for use in model processes, it could be argued that mean daily temperature is a plausible driver. For both crop models, four drivers are aggregates of daily precipitation and, in combination, span a time period starting before planting and ending four months afterwards, with overlap between drivers. In terms of aggregation methods selected, precipitation drivers identified measure the number of days above the 90th percentile, which can be interpreted as the number of rainy days, or the mean daily precipitation during a time interval, which is proportional to the cumulative rainfall in that period.

Temperature is more strongly represented in climate drivers identified for pDSSAT (five drivers) than LPJmL (two drivers). No temperature data before two months after planting is considered in drivers identified for LPJmL, but two drivers identified for pDSSAT use minimum daily temperatures starting at (approximately) planting. Shortwave radiation is included in the set of drivers for both crop models, and longwave radiation and windspeed is also considered for LPJmL. No identified drivers take into account the climate conditions later than six months from planting, and the only variable considered before planting is precipitation.

Analysis of the relationships between identified climate drivers and yield failure is made possible by the use of simple, interpretable models. Odds ratios associated with the climate drivers for fitted logistic regression models are reported in Table

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

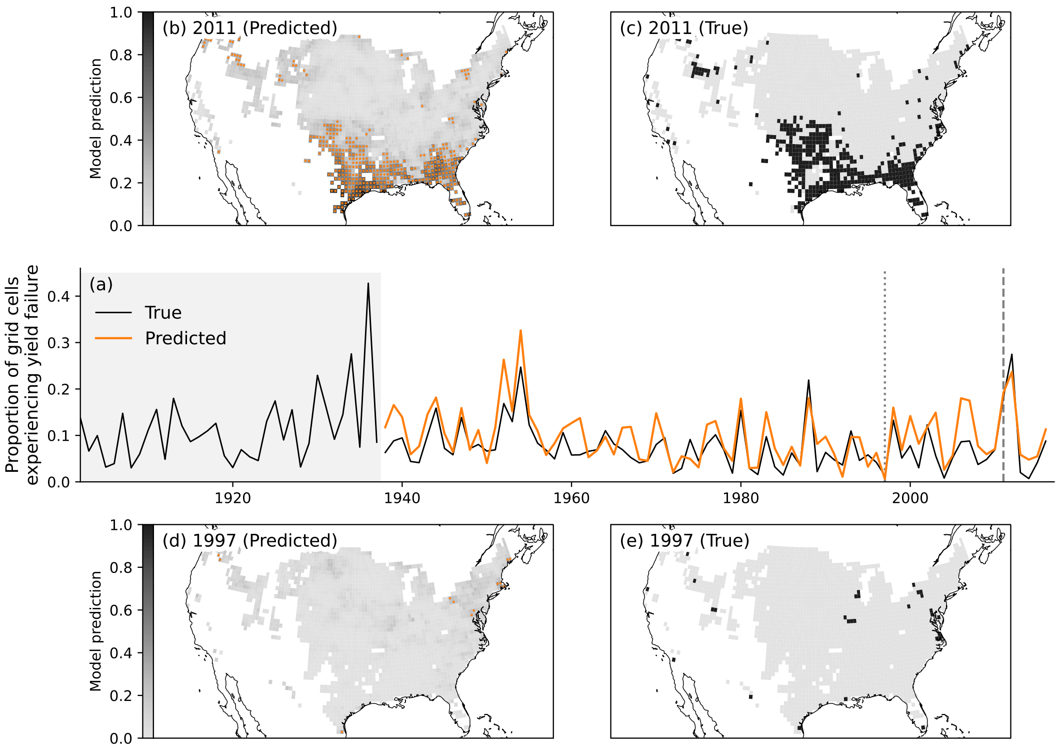

Figure 4. Temporal and spatial predictions of a logistic regression model using ten identified pDSSAT drivers. (a) True (black) and predicted (orange) proportion of grid cells experiencing yield failure each year, using a threshold of 0.243 to define a yield failure prediction (selected based on achieving the maximal f-score of 0.438 over the training years, which are marked by the grey shaded area). The dashed vertical line denotes the year in which the highest average precision is achieved, 2011. Model predictions for this year are illustrated in panel (b), with orange dots marking grid cells where the prediction exceeds the threshold and therefore yield failure is predicted. The corresponding ground truth is shown in panel (c), with black areas indicating locations experiencing yield failure. Similarly, the dotted vertical line in panel (a) marks the year in which the lowest average precision is achieved (1992), and the corresponding model predictions and ground truth for this year are shown in panels (d) and (e) respectively.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Table 1. Odds ratios of maize yield failure associated with a unit increase in each predictor, based on the coefficients of logistic regression models fit using 10 identified climate drivers for pDSSAT and LPJmL. Predictors are labelled with the climate variable used, the aggregation method, and the beginning and end day (in relation to the planting date) of the time interval over which the aggregation is applied. Predictors are sorted by magnitude of the odds ratio.

Crop model	Predictor	Odds ratio
pDSSAT	tasmin_mean_d4_d98	1.070
	tas_Days>90p_d58_d173	1.040
	tasmax_Days>90p_d51_d180	1.022
	tasmin_min_d69_d122	0.992
	rsds_mean_d70_d135	0.977
	tasmin_Days>90p_d0_d133	0.970
	pr_Days>90p_d-13_d70	0.968
	pr_Days>90p_d42_d91	0.935
	pr_mean_d64_d109	0.714
	pr_mean_d26_d80	0.582
LPJmL	tasmax_Days>90p_d56_d183	1.029
	rlds_mean_d91_d136	1.018
	rsds_mean_d8_d89	0.994
	rsds_max_5d_mean_d82_d134	0.980
	pr_Days>90p_d42_d92	0.974
	pr_Days>90p_d-41_d73	0.971
	tasmin_mean_d57_d125	0.942
	sfcwind_mean_d5_d83	0.939
	pr_mean_d37_d74	0.682
	pr_mean_d59_d104	0.526

1. For both crop models, the two most influential climate drivers are based on mean precipitation, with increased rainfall strongly associated with lower yield failure probability. Additionally, the climate driver with the strongest positive association with yield failure is temperature-related (for pDSSAT, the mean minimum daily temperature in the first three months after planting, and for LPJmL, the number of days where the maximum temperature exceeds the 90th percentile between two and six months after planting). For both crop models, however, both positive and negative associations with yield failure are identified for temperature-related drivers, pointing towards nonlinear relationships between growing-season temperature and yield failure probability.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

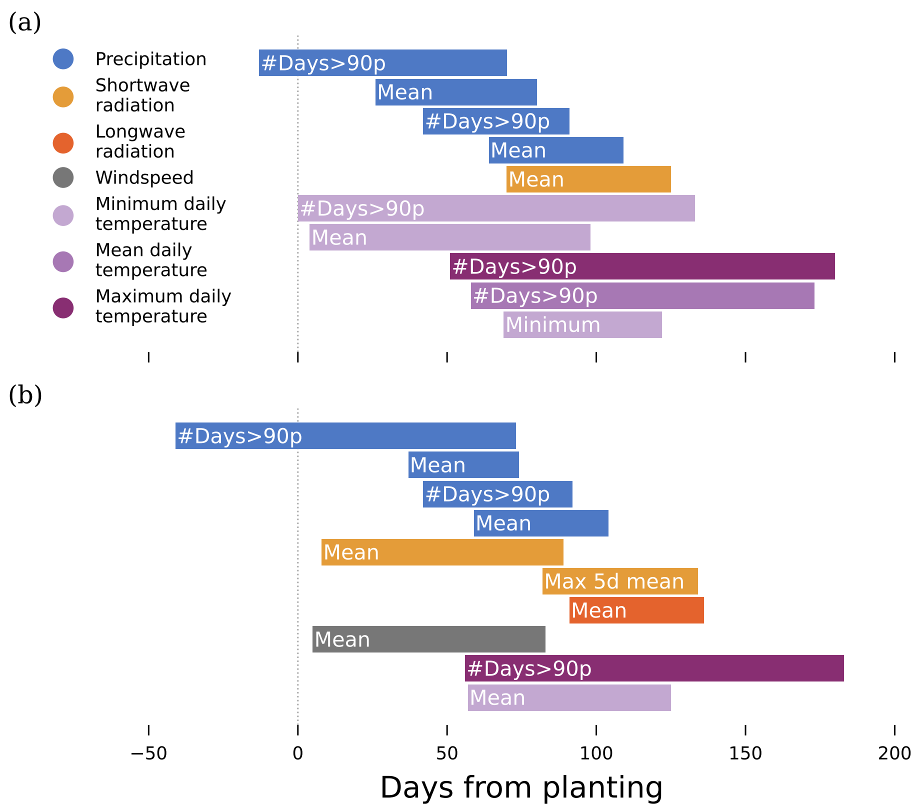

Figure 5. Identified climate drivers of maize yield failure for (a) pDSSAT and (b) LPJmL, based on using 30 sampled time periods to generate 100 pools of candidate features, using ten-fold temporal cross-validation to select ten features from each pool, and condensing the resulting 1000 features to ten drivers. Each climate driver is denoted by a coloured horizontal bar, and consists of an aggregated daily climate variable over a time interval which is defined relative to the planting date (illustrated by the bar's length and position). The caption overlaid on each bar describes the aggregation method used (i.e., 'mean' indicates that the corresponding daily climate values are averaged over the time period selected; '#Days>90p' means the number of days in which the respective variable exceeds the 90th percentile for that location; 'Max 5d mean' takes the maximum value of the five-day rolling mean of the respective climate variable over the selected time period). Drivers are ordered by variable and start date.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

When yields are detrended before applying the method, the precipitation- and temperature-related drivers identified are similar to those identified without detrending for both crop models (Fig. B5). Precipitation is the only climate variable considered before planting by any driver, and climate conditions later than six months after planting are ignored. However, when yields are detrended, windspeed and long-wave radiation is not used by any of the climate drivers identified for LPJmL, nor is shortwave radiation for pDSSAT.

Histograms of some identified climate drivers in yield failure and non-yield failure years for each model are largely in agreement with the associations implied by the odds ratios (Fig. 6). For both crop models, yield failure years are more likely to have less rainfall between 8 and 16 weeks after planting (Fig. 6a,b) and fewer rainy days between 6 and 12 weeks after planting (Fig. 6e,f). Normal years are more likely to have fewer days during the period two and six months after planting in which the maximum temperature exceeds the 90th percentile (Fig. 6c,d). For LPJmL, mean windspeeds in the first three months after planting of over 3.5ms^{-1} associated with decreased chances of yield failure (Fig. 6g). For pDSSAT, the yield failure distribution of the number of warm days (in which minimum daily temperatures exceed the 90th percentile) in the first five months after planting is slightly higher than normal years (Fig. 6h). However, the associated odds ratio from a fitted logistic regression model (0.970; Table 1) suggests that increasing warm days in this period is linked to decreasing odds of yield failure.

4.3 Our method robustly identifies plausible climate drivers that outperform many baseline feature sets

To evaluate the robustness of the approach we now turn to the results for the 100 bootstrap repeats, where in each repeat 20 pools are sampled from 100 total pools of candidate features. The original pools, by definition, contain an equal number of features based on each of the ten climate variables (Step 1). After selecting ten features from each pool based on cross-validation predictive performance (Step 2) we find that the majority of features selected consist of variables that are used as input by the crop models (Figs. A3a, A4a). However, some variables not used by the crop models are also selected, albeit less frequently (surface pressure, windspeed, and specific humidity for pDSSAT and surface pressure and specific and relative humidity for LPJmL).

After condensing the features selected from twenty pools to final sets of ten climate drivers (Step 3), variables not used as input by the crop models are eliminated for pDSSAT in all 100 bootstrap repeats, and in 96 out of 100 repeats for LPJmL (Figs. A3b, A4b). This behaviour is consistent when yields are detrended (Figs. B6a,b, B7a,b). This shows that our approach consistently and robustly identifies climate drivers based on variables that are used as input by the crop models and are therefore more plausibly causally-related to crop failure.

The duration of the time intervals over which climate conditions are aggregated in the selected features before condensing (Step 2) displays three peaks: at two to four weeks, at two to three months, and at approximately four to five months. This third peak gradually tapers off, with some selected features based on nine months of climate data. The time intervals over which condensed drivers are aggregated (Step 3) show two distinct peaks: at approximately eight weeks, and at four to five months. Drivers never, or rarely, consider more than six months of climate data. These behaviours are largely consistent across both crop models (Figs. A3c, A4c) and also when yields are detrended (Figs. B6c, B7c).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

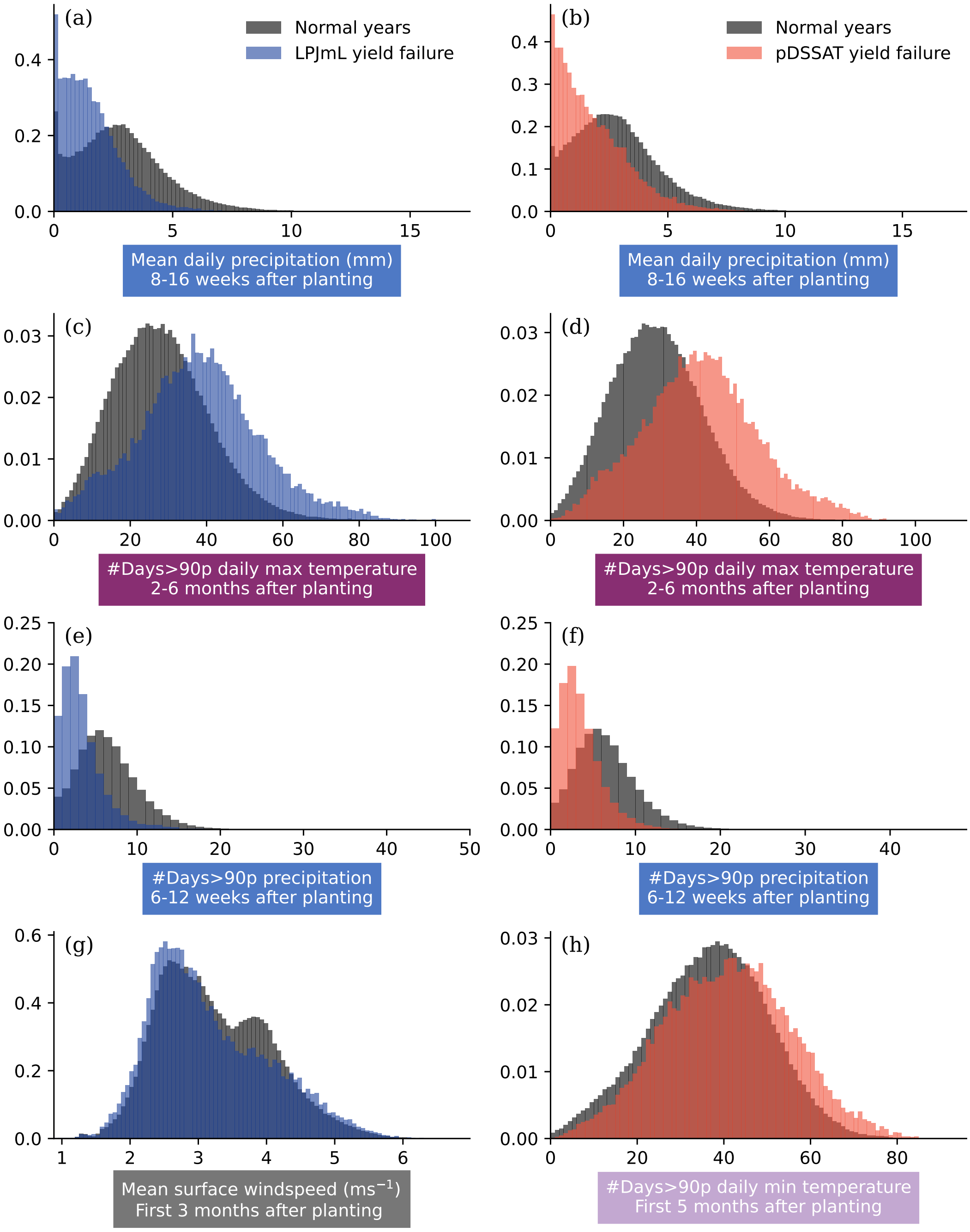

Figure 6. Distributions for normal and yield-failure years of some of the identified climate drivers of maize yield failure shown in Fig. 5, for LPJmL (panels (a), (c), (e) and (g)) and pDSSAT (panels (b), (d), (f) and (h)).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

For both crop models, Lasso logistic regression models using sets of 5, 10 or 15 identified climate drivers, fitted on training data only, are able to achieve evaluation scores on the test years consistently higher than those using average growing-season climate features and five extreme indicators (Fig. 7). Models using only five identified climate drivers consistently perform better in terms of ROC AUC than all baseline feature sets tested for pDSSAT, and models using 15 drivers consistently outperform all baseline feature sets for both crop models. For pDSSAT, using ten or fifteen climate drivers consistently outperforms all baseline feature sets in terms of average precision, and achieves similar performance to baseline feature sets consisting of monthly climate variables and extreme indicators for LPJmL. Results are similar for Brier score (Fig. A7c,d) and log loss (Fig. A8c,d).

If second-order interaction terms are included, models using fifteen identified climate drivers can often outperform, or perform similarly to, models using monthly climate averages, and consistently outperform all other feature sets tested, for all four evaluation metrics calculated (Figs. A5-A8). Using five climate drivers is sufficient to achieve similar performance to baseline features sets using quarterly climate variables and five extreme indicators. Results are similar when using more complex extra trees models, and for both crop models and all models tested when yields are detrended (Figs. B8-B11).

We find that the results of this methodology is sensitive to the cross-validation method used, both in terms of climate variables selected and the test performance of models using sets of identified drivers. Using temporal cross-validation strategy results in the highest proportion of potentially-causal variables used in features selected from pools (Step 2) and in identified drivers (Step 3), and this is consistent for both crop models (Figs. C1, C2) and also when yields are detrended (Figs. C3, C4). Spatial cross-validation and feature-cluster cross-validation results in a lower proportion of potentially-causal variables across all settings, and random cross-validation the lowest. Similarly, identified driver sets selected using temporal cross-validation in general achieve the highest test scores and those selected using random cross-validation the lowest (Fig. C5). This general behaviour is robust to the evaluation metric, the model used (Lasso logistic regression with and without second-order interactions or extra trees), the crop model and if yields are detrended.

In comparison to the choice of cross-validation strategy, the performance of the identified driver sets is relatively insensitive to all other parameters tested. Increasing the number of time intervals sampled per pool slightly improves model performance with resulting sets of drivers for pDSSAT, but not LPJmL. Interestingly, selecting fewer features from each pool leads to increased performance of identified driver sets in Step 3.

4.4 Meaningful and predictive drivers of yield failure are obtained by applying the method to observational data

Using the same methodology for observational US county-level data and daily meteorological data from eight variables, we obtain four climate drivers of maize yield failure (Fig. 8). The only variables used are precipitation, maximum vapour pressure deficit and maximum temperature, and the drivers cover a time period starting approximately one month before planting until around six months afterwards. Precipitation is relevant for a three-month period starting one month prior to planting, and the number of days in which the maximum temperature exceeds the 90th percentile are relevant from one month from planting until six months from planting. Two drivers are based on vapour pressure deficit conditions in two distinct phases of the growing

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

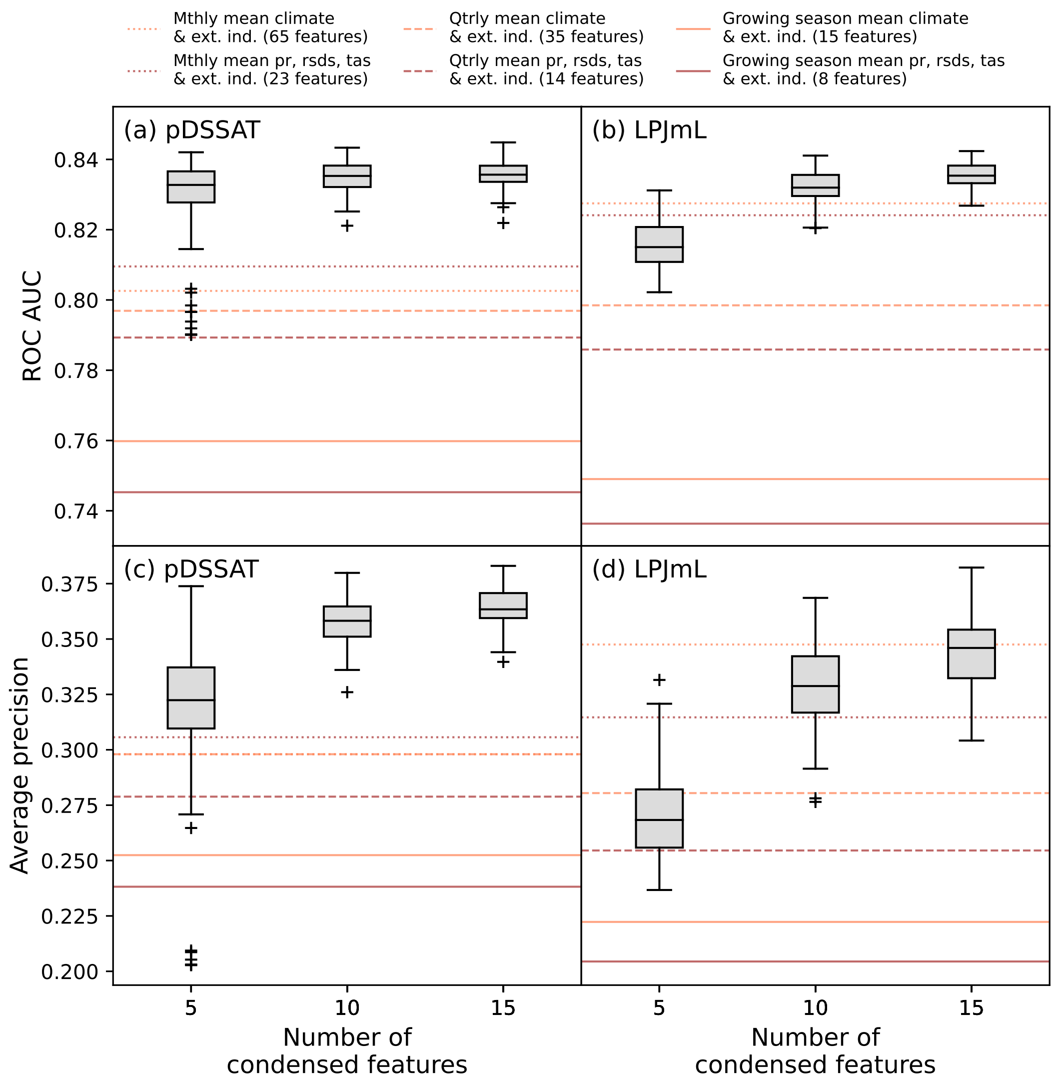

Figure 7. Comparison of the test-set performance of Lasso models for pDSSAT ((a) and (c)) and LPJmL ((b) and (d)). Box plots show the scores for models using sets of five, ten or fifteen drivers using our proposed method, and horizontal lines denote the scores for the same models using baseline feature sets as predictors (made up of mean aggregate climate variables over the growing season or in monthly/quarterly intervals as well as five extreme indicators).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

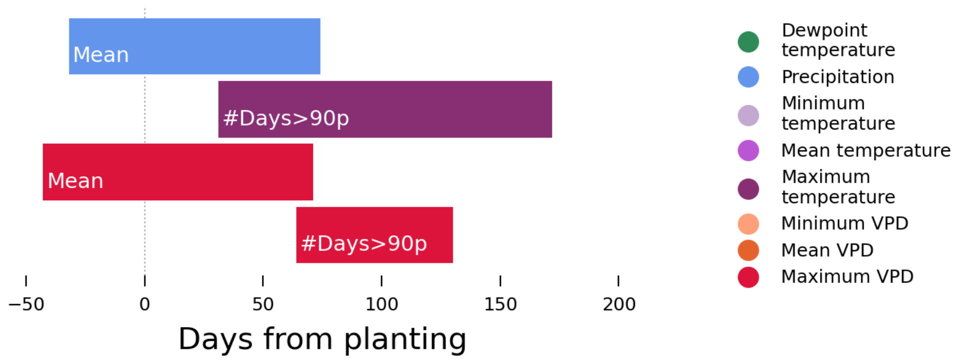

Figure 8. Four identified climate drivers of US county-level maize yield failure, for predominantly-rainfed counties, only based on using 20 sampled time periods to generate 100 pools of candidate features, using ten-fold temporal cross-validation to select ten features from each pool, and condensing the resulting 1000 features.

season and based on different aggregation methods, with the number of days of extreme vapour pressure deficit conditions most relevant for a two-month period starting approximately two months after planting.

A logistic regression model with the identified four drivers as predictors, fitted to data from the training years, achieves a test ROC AUC score of 0.85, average precision of 0.54, Brier score of 0.073 and log loss of 0.26, and the Pearson correlation of predicted versus ground truth annual proportions of counties experiencing yield failure is 0.98 (Fig. 9). Notably, the test years include a year which was heavily impacted by drought (2012), which is well-captured by the model (achieving an average precision in that year of 0.83).

5 Discussion and Conclusions

Identifying simple climate drivers from high frequency meteorological data that are associated with impacts is a challenging task. Most studies so far rely on selecting from a large set of predefined aggregate features in combination with statistical models (Ben-Ari et al., 2018; Vogel et al., 2021; Anand et al., 2024b; Heilemann et al., 2024). The choice of aggregation strategy and climate features used can significantly impact model performance and behaviour (Chen et al., 2024). Other studies use complex machine learning approaches to derive a mapping from meteorological conditions to impacts, but can struggle to obtain useful interpretations (Wolanin et al., 2020; Anand et al., 2024a). Our approach addresses this challenge in an efficient way. We generate many candidate features based on random time slices of the meteorological data, and use a robust, two-stage selection and synthesis process to extract a small set of drivers. While the method exploits the flexibility of ML models and the information contained in high-resolution, multivariate data, the resulting drivers are simple and consist of single climate variables aggregated over defined time intervals. We are then able to construct parsimonious logistic regression models using these drivers that are easy to interpret and have very high accuracy.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

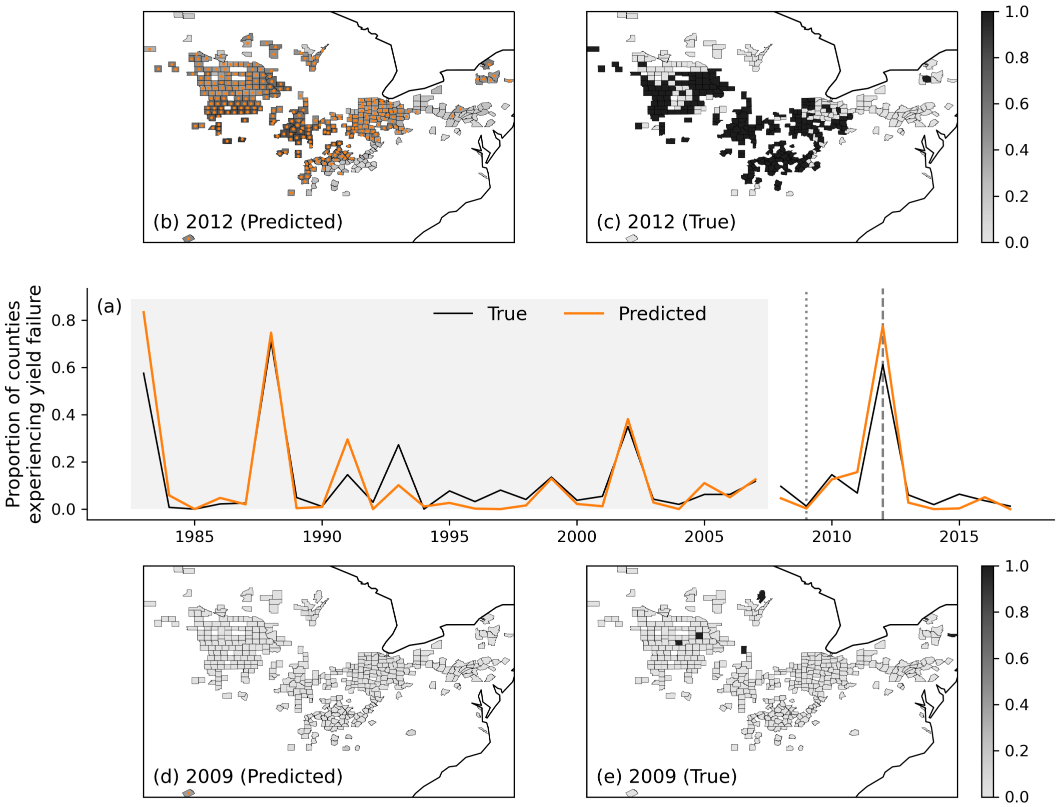

Figure 9. Temporal and spatial predictions of a logistic regression model using four identified drivers using observational county-level data. (a) True (black) and predicted (orange) proportion of counties experiencing yield failure each year. The dashed vertical line denotes the year in which the highest average precision is achieved, 2012, and model predictions for this year are illustrated in panel (b). Orange dots mark counties where yield failure is predicted by the model based on a threshold determined over the training years, and the corresponding ground truth is shown in panel (c). The dotted vertical line in panel (a) marks the year in which the lowest average precision is achieved, 2009, and the corresponding predictions and ground truth are shown in panels (d) and (e) respectively.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Our method identifies robust drivers in two stages: first, features are selected from large pools of candidates based on cross-validation performance; second, the selected features from all pools are clustered based on similarity and drivers are extracted from each of the largest clusters. The first stage is similar to methods that have been shown to improve model transferability in previous studies (Ludwig et al., 2023; Sweet et al., 2023). Our results suggest that the second stage, a novel contribution of this study, is also needed to reliably obtain climate drivers using only variables that are taken as input by the crop models. Interestingly, it also dramatically reduces the proportion of features lasting less than one month or more than six months. This second stage could be interpreted as obtaining drivers that are useful in combination with varying sets of other predictors, and that are robust to variation in start and end dates. Lischeid et al. (2022) found that machine learning models of yield variability using different predictor sets were able to achieve equally-good performance scores, and concluded that expert judgement was needed to identify the most reliable models. Our approach is, perhaps, a way of overcoming this challenge.

The predictive performance of models on held-out years and the consistency of identified relationships in different time periods can be used as evidence for a causal link, as has been done for instance by Schlenker and Roberts (2009) to show nonlinear temperature effects on different crop types. Our approach is motivated by a similar argument: climate indices that robustly improve model performance at predicting an impact in held-out years are more likely to be causal drivers. Using yield simulations and their corresponding forcing data allows us to test this in a situation where some climate variables, such as humidity, are not taken as input by the crop models (and therefore do not have a causal effect on yields), but are still associated with the target variable due to dependencies between climate variables (as shown in composite plots). Additionally, each crop model uses a different subset of variables. The fact that the drivers identified by our method consistently consider only the climate variables taken as input by each crop model, despite the presence of strong associations between other variables and yield failure, is encouraging.

Many variables which influence agricultural yield are correlated in space and among themselves, which can lead to errors in empirically-estimated relationships. Climate change causes warmer temperatures and shifts in precipitation patterns, strongly affecting agricultural regions (Lobell and Di Tommaso, 2025). Data-driven models have difficulties extrapolating outside of the observed distribution, which could impede their use for projecting impacts under future climate change scenarios. The robustness of our results, despite the challenging split between training and test years used in this study, show that our approach is successful in tackling these challenges. Data-driven emulators of process-based crop models have been proposed to efficiently downscale simulations and facilitate the creation of large ensembles of agricultural projections in different climate scenarios. Published emulators of crop models have used predictors such as the mean and standard deviation of temperature and precipitation over the growing season or per month, and measures of climate extremes such as the number of extreme degree days or the maximum consecutive days with no rainfall (Folberth et al., 2019; Liu et al., 2023). However, due to the presence of spatiotemporal autocorrelation in the data used for training these models, emulator skill has been found to drop sharply when tested on held-out years or locations (Liu et al., 2023; Sweet et al., 2023). Our approach, on the other hand, identifies a set of climate features that are robust drivers of a specific impact or metric and can be used to create a lightweight, interpretable model that achieves high performance on datapoints many decades after the training period. Such emulators or metamodels

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

could also be used to analyse the processes leading to specific model outcomes which might emerge from the interaction of multiple complex mechanisms and therefore not be well-understood.

Our method is robust to the majority of parameters tested, with the exception of cross-validation strategy used for selecting features from each pool (Step 2). This is in alignment with previous findings that feature selection and hyperparameter tuning is sensitive to this choice (Meyer et al., 2018; Sweet et al., 2023). It is noteworthy that, while in general we find that driver sets identified using temporal cross-validation perform better in the test years, and those using random cross-validation perform the worst, the difference in performance is quite small in comparison to the drastic difference in the climate variables used in features selected and sets of identified drivers. Random cross-validation has recently been used to identify parsimonious sets of weather drivers for modelling the effect of climate change on crop production accounting for adaptation (Hultgren et al., 2025). Given that the sets of weather drivers used in data-driven models can impact resulting agricultural projections (Chen et al., 2024), our results suggest that results from such methodologies may be sensitive to the cross-validation strategy used.

This study uses crop model simulations, which allow us to validate the predictive performance of drivers under challenging conditions and confirm that the variables used in identified drivers correspond to those used by the crop models, and are plausible based on our knowledge of their underlying mechanisms. For LPJmL, precipitation matters most in the juvenile phase of crop development, where leaf area index increases most steeply. Water stress during this phase can lead to permanent reduction in light interception, and thus yield penalty. Windspeed impacts nitrogen loss, which has a similar effect in the early growth period. Hot days (which are represented in our identified drivers counting the days above the 90th percentile of maximum daily temperature) can damage photosynthetic capacity if they are higher than an optimal threshold. Long-wave radiation is considered in the calculated of atmospheric water demand, and so can contribute to water stress in the crop model, but it is unclear why the later growing season has been selected by our approach. However, these interpretations should be considered with caution. Both of the crop models used consist of multiple mechanisms that can interact in surprising ways, and the process by which we transform yields to yield failures is dependent on the grid cell and time period selected. This means that, even in this simulated setup, confirming that climate drivers identified by our approach are ‘accurate’ is not straightforward.

When moving to observational data, some additional challenges need to be considered. There might be sampling biases, for instance because farmers choose to plant crops in climatically-suitable regions or reduce harvested areas in years with adverse weather conditions. Furthermore, changing management practices, irrigation and crop breeding strongly affect crop yields, often exceeding trends related to climate change. Disentangling the influence of weather is also made more challenging as adaptation occurs in response to and in anticipation of climate change. However, despite these difficulties, by applying the method to reported county-level maize yields and eight corresponding meteorological variables, we obtain four climate drivers of maize yield failure that are physically plausible, and a simple logistic regression model using those predictors is able to capture the effects of the drought conditions of 2012. Vapour pressure deficit is identified as particularly relevant, with elevated levels later in the growing season found to be particularly relevant. This corresponds with the findings of Lobell et al. (2014), in which monthly aggregates of precipitation, vapour pressure deficit and minimum and maximum daily temperatures were used as candidate predictors for multivariate adaptive regression splines, and vapour pressure deficit in the third month after sowing identified as the most influential variable. High temperatures during the growing season (after the first month of plant

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

development) are also identified as a driver of yield failure using our method. Increased temperatures and elevated vapour pressure deficit conditions both contribute to yield loss through separate mechanisms, but disentangling their effects can be challenging as the two factors are highly correlated and their impacts depend on precipitation levels (Hsiao et al., 2019). It is noteworthy that the logistic regression model is unable to capture the effects of the heavy rainfall conditions in 1993. This may be due to the inability of the variables used to capture the effects of soil moisture on yields. For example, Rigden et al. (2020) identified negative impacts to maize yields in the US Midwest when either vapour pressure deficit or soil moistures reached high levels. Future work could include soil moisture and other variables to explore these relationships further.

While model performance using the identified drivers over the studied period is good, and the identified drivers appear plausible, it is difficult to assess the robustness of the methodology on observational data. Crop yields are affected by many types of weather conditions during different times of the growing season, making it challenging to argue that any obtained drivers are not plausible in some way. Furthermore, data is only available over a limited period of time, during which agricultural practices have evolved, making robust model evaluation challenging. Therefore, while the strength of our results suggest that this approach could be applied to observation data, we advise conducting careful sanity checks and validating that the results are in agreement with scientific understanding.

In conclusion, here we present an efficient and flexible approach to identify climate drivers associated with impacts. The identified drivers can be used as predictors in simple, interpretable models for further analysis, or could be used to guide experimental design for studying the effect of multiple stressors and different types of compound events, which otherwise may suffer from the curse of high dimensionality (Zscheischler et al., 2020; Webber et al., 2022; Kim et al., 2025). We test the approach on simulated crop yield data, show that it is robust to different parameter choices, and, finally, demonstrate the use of the method on observational data. While we illustrate the use of this approach for identifying drivers of maize yield failure, it could also be applied to other climate impacts that are influenced by weather events with different timings and durations, such as forest mortality, wildfire occurrence or flooding.

Code and data availability. Maize yield simulations used, as well as the corresponding daily climate forcing data, are available from the ISIMIP data repository https://doi.org/10.48364/ISIMIP.982724.1 (Lange et al., 2022). Observed county-level maize yield data is publically available from USDA NASS (United States Department of Agriculture (USDA), 2025) and daily meteorological variables are available from PRISM (PRISM Group, 2018). Code used to download and process the data, perform the experiments and generate the results and figures in this manuscript, as well as further sensitivity analyses, can be accessed at https://doi.org/10.5281/zenodo.15725041 (Sweet, 2025).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Appendix A: Supporting figures for non-detrended data

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

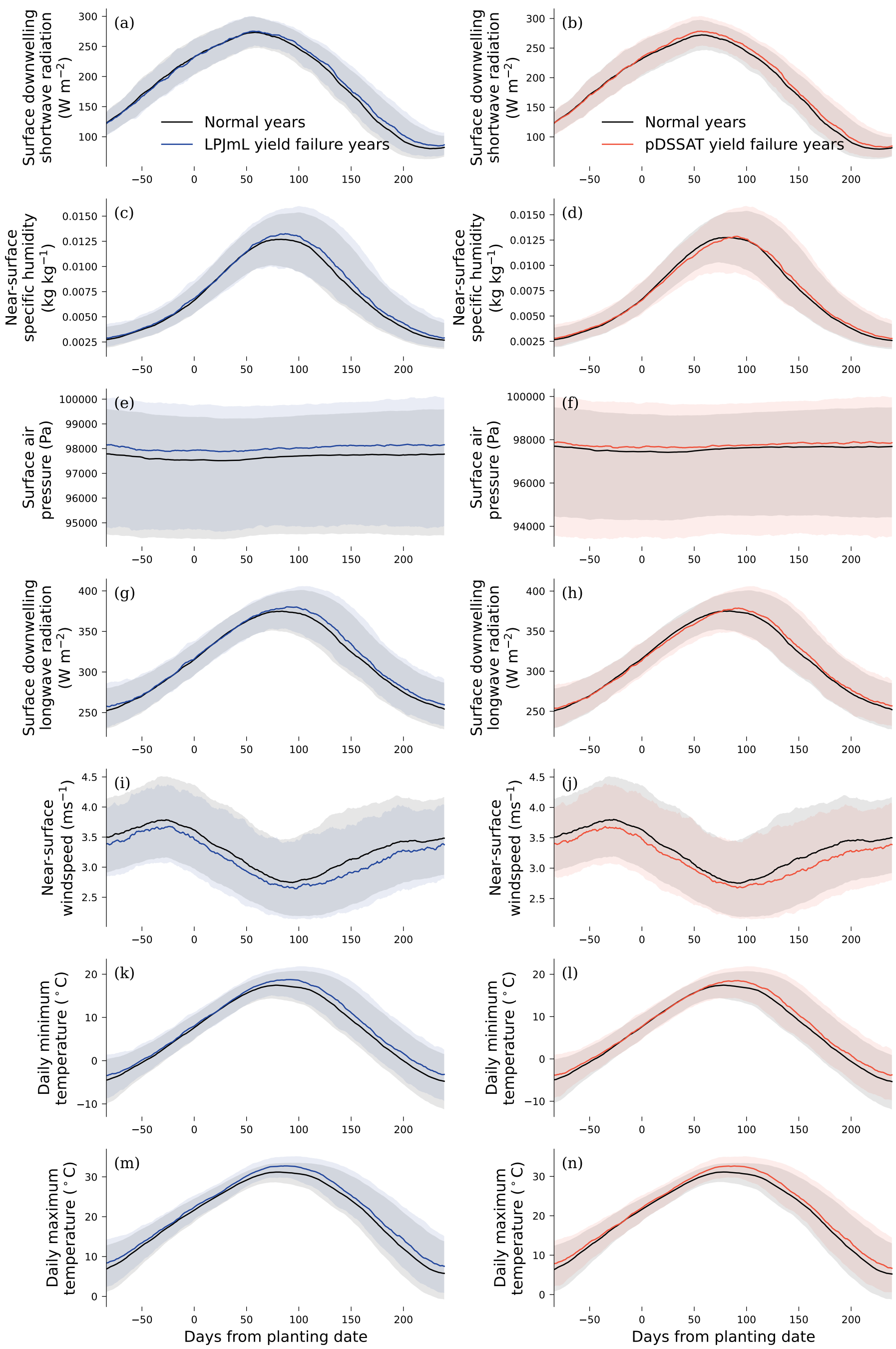

Figure A1. As in Fig. 3, composite plots of the seven-day rolling mean of surface downwelling shortwave radiation, near-surface specific humidity, surface air pressure, surface downwelling longwave radiation, near-surface wind speed, daily minimum temperature, and daily maximum temperature for LPJmL (panels (a), (c), (e), (g), (i), (k), (m)) and pDSSAT (panels (b), (d), (f), (h), (j), (l), (n)) for either normal years or yield failure years, without detrending.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

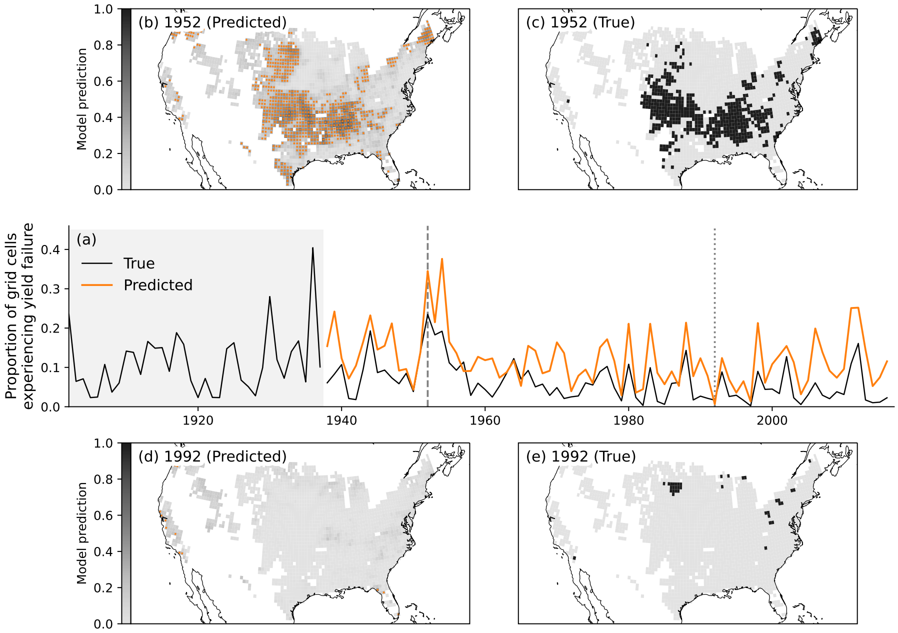

Figure A2. As Fig. 4 but for LPJmL. The threshold used for predicting yield failure years is 0.225, which achieves an f-score on the training years of 0.471.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

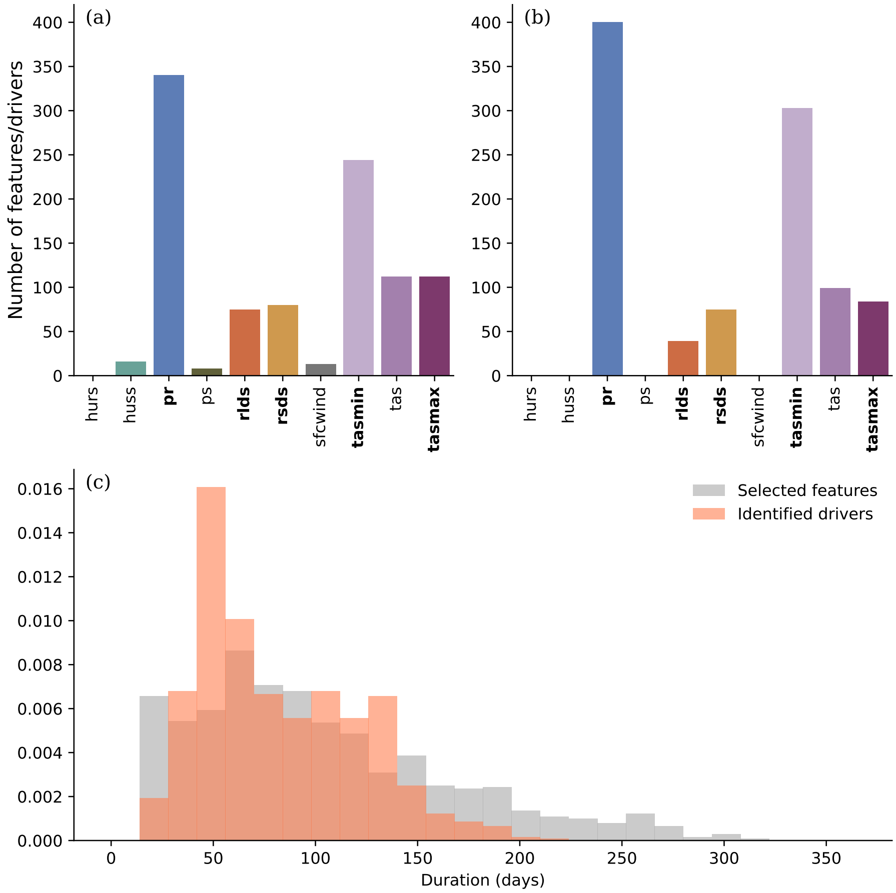

Figure A3. Climate variables used in features selected after Step 2 of our method (a) and in drivers identified from condensing those features in Step 3 (b), for pDSSAT without detrending. Bold text indicates variables used as input by the crop model: precipitation (pr, mm), surface downwelling longwave radiation ($rlds$, Wm^{-2}) and shortwave radiation ($rsds$, Wm^{-2}), daily minimum ($tasmin$, °C) and maximum near-surface air temperature ($tasmax$, °C). Not used by the crop model are near-surface relative humidity ($hurs$, %), near-surface specific humidity ($huss$, kgkg^{-1}), surface air pressure (ps, Pa), near-surface windspeed ($sfcwind$, ms^{-1}) and mean daily near-surface air temperature (tas, °C). However, pDSSAT does use minimum and maximum daily temperatures which are then downscaled to hourly values. The density histogram in panel (c) shows the duration in days of the resulting features or drivers at each stage.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

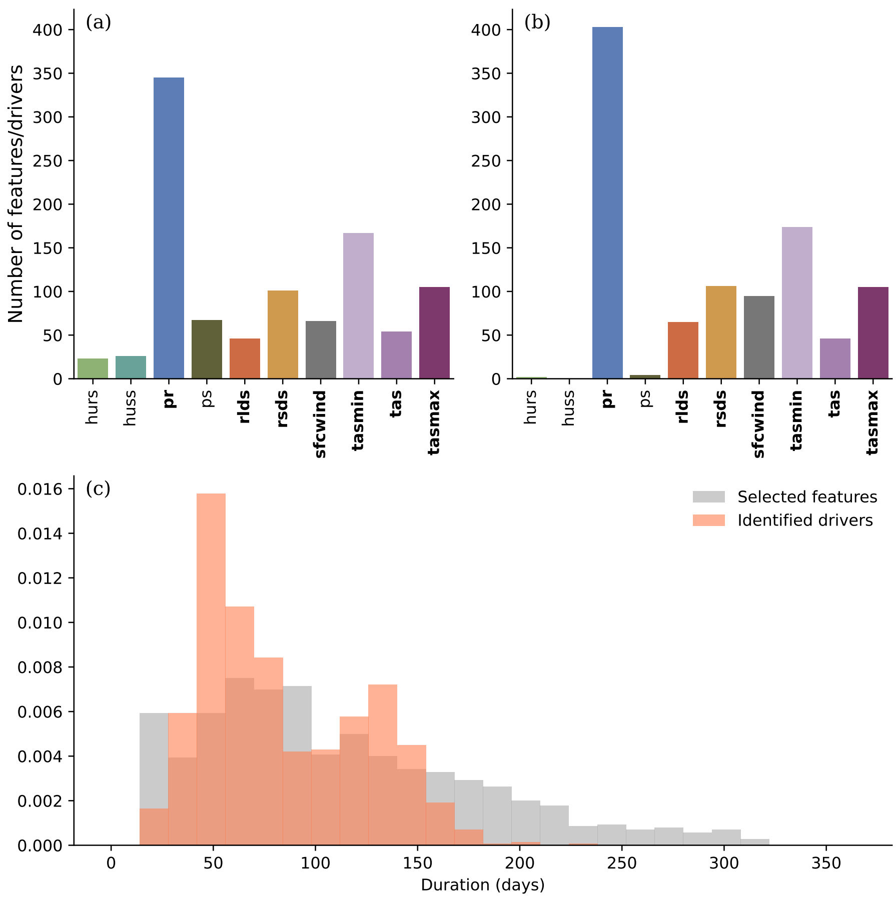

Figure A4. As Fig. A3 but for LPJmL.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

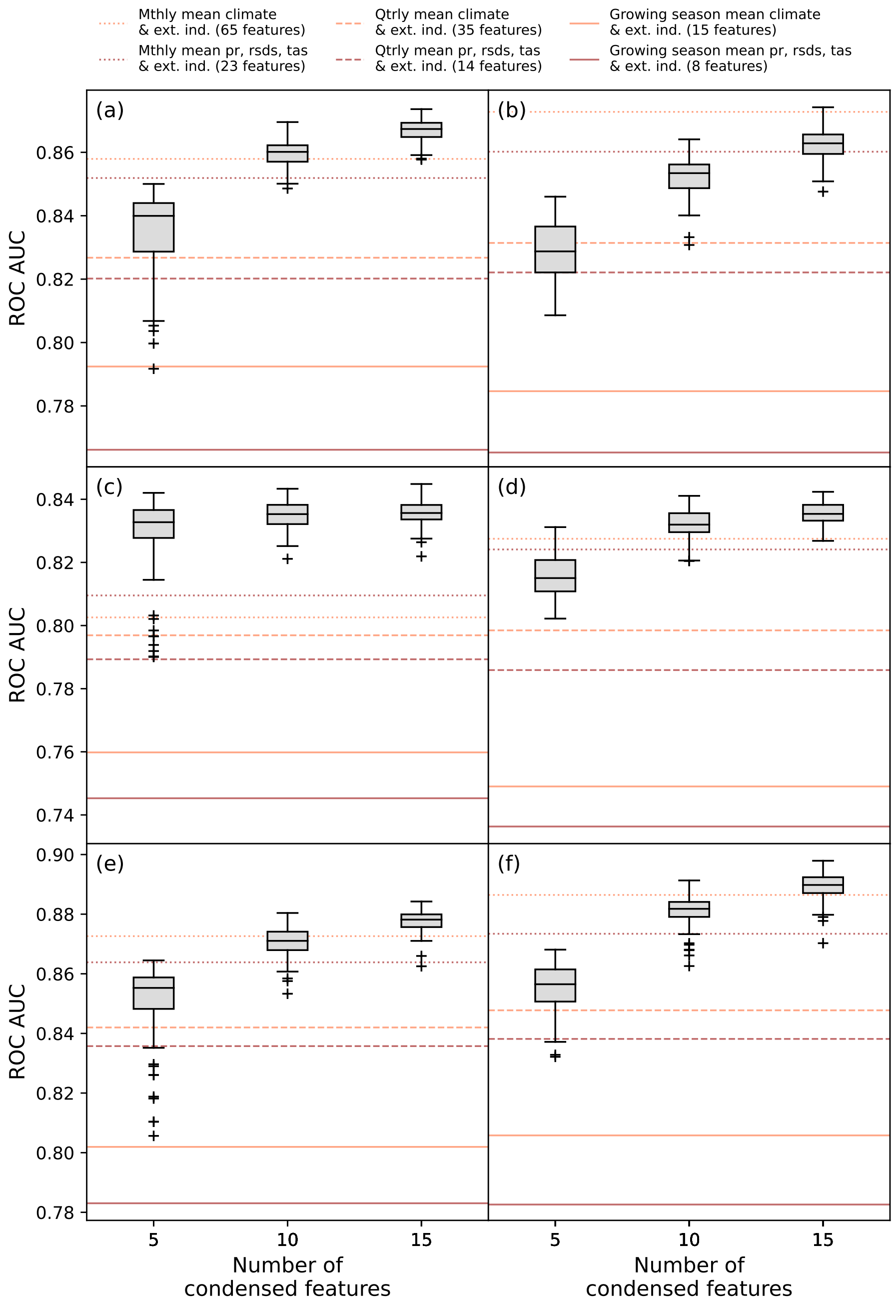

Figure A5. ROC AUC scores on test set for models trained on training set, using feature sets derived from training set data (higher is better). (a), (c) and (e) show scores for pDSSAT and (b), (d) and (f) show LPJmL. (a) and (b) use LASSO logistic regression and include interaction terms; (c) and (d) use LASSO without interaction terms, (e) and (f) use extra trees.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

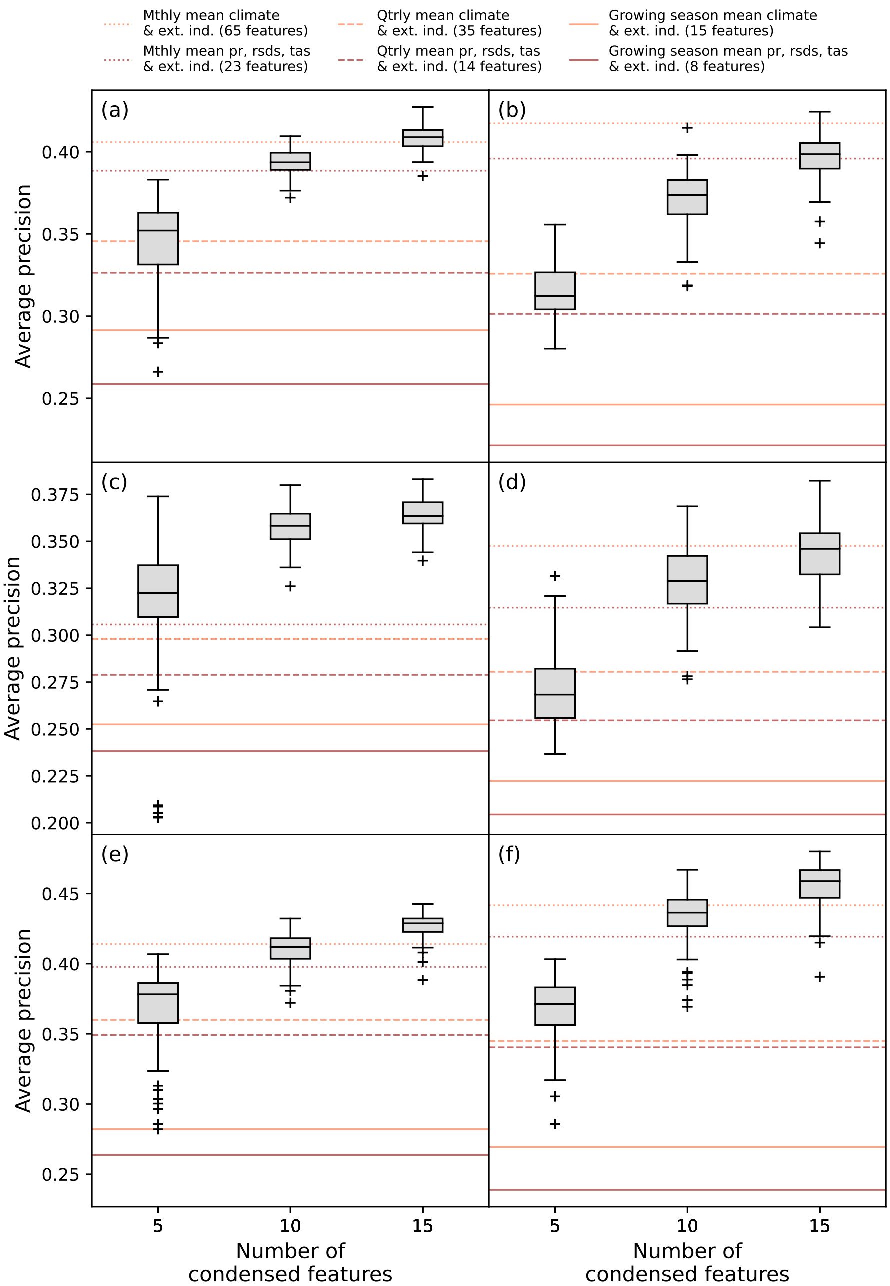

Figure A6. As in Fig. A5 but for average precision (higher is better).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

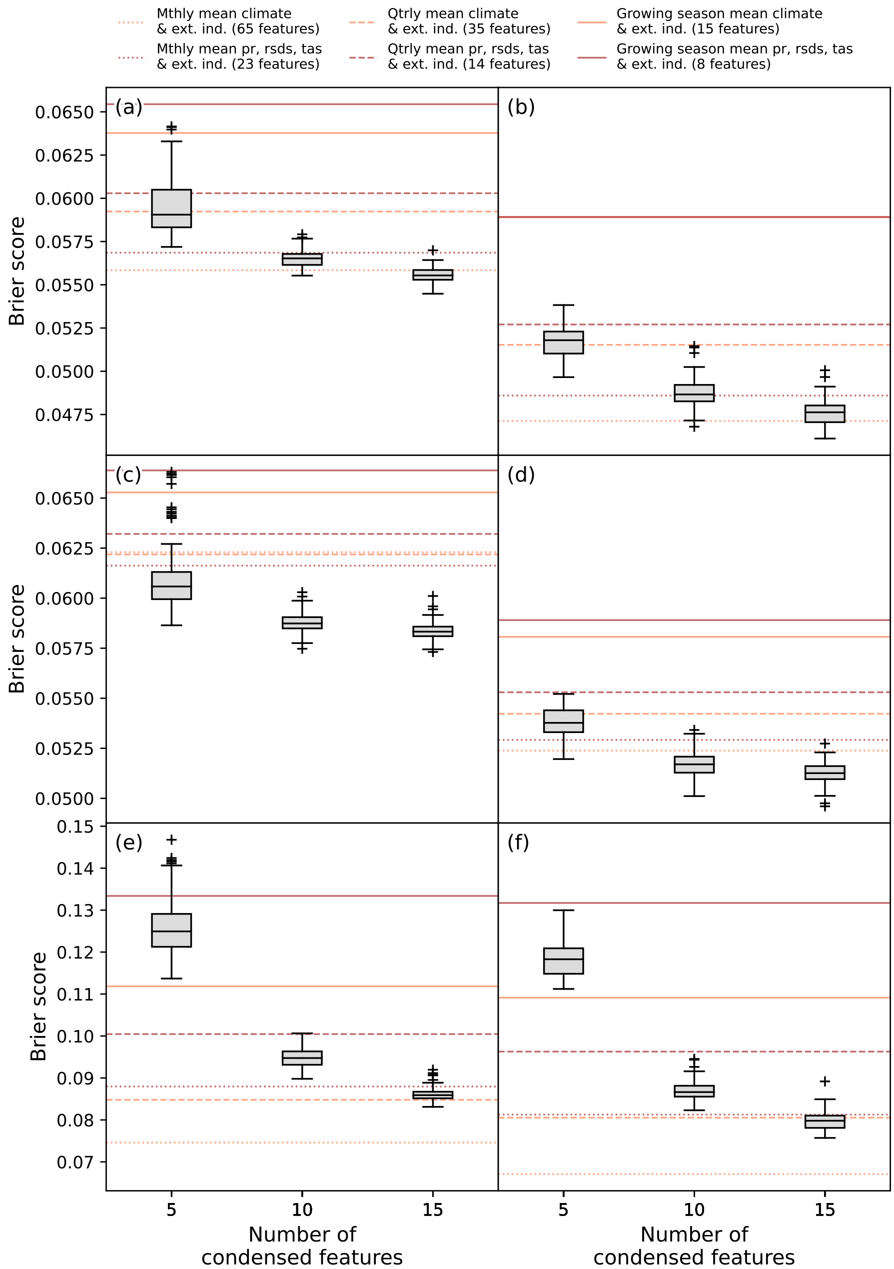

Figure A7. As in Fig. A5 but for Brier score (lower is better).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

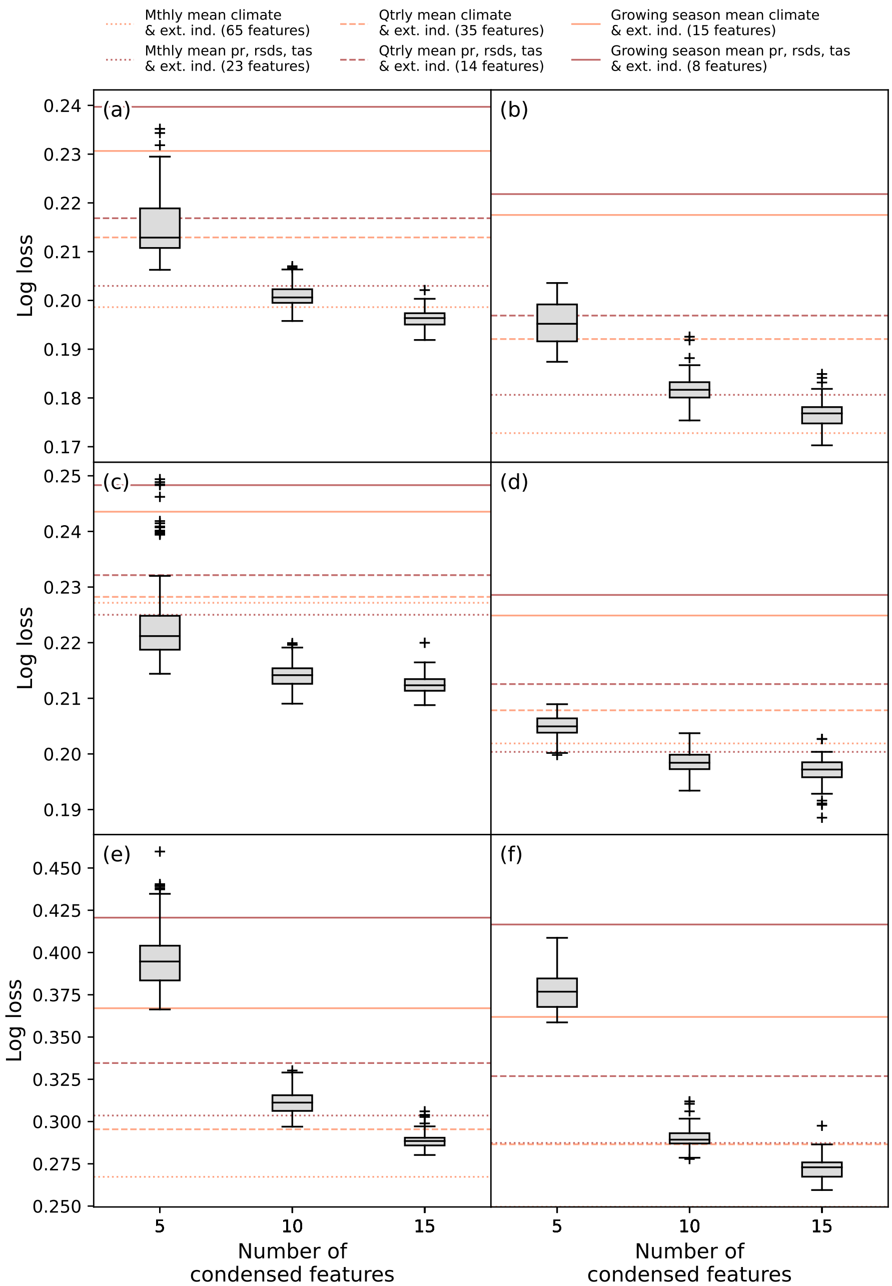

Figure A8. As in Fig. A5 but for log loss (lower is better).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

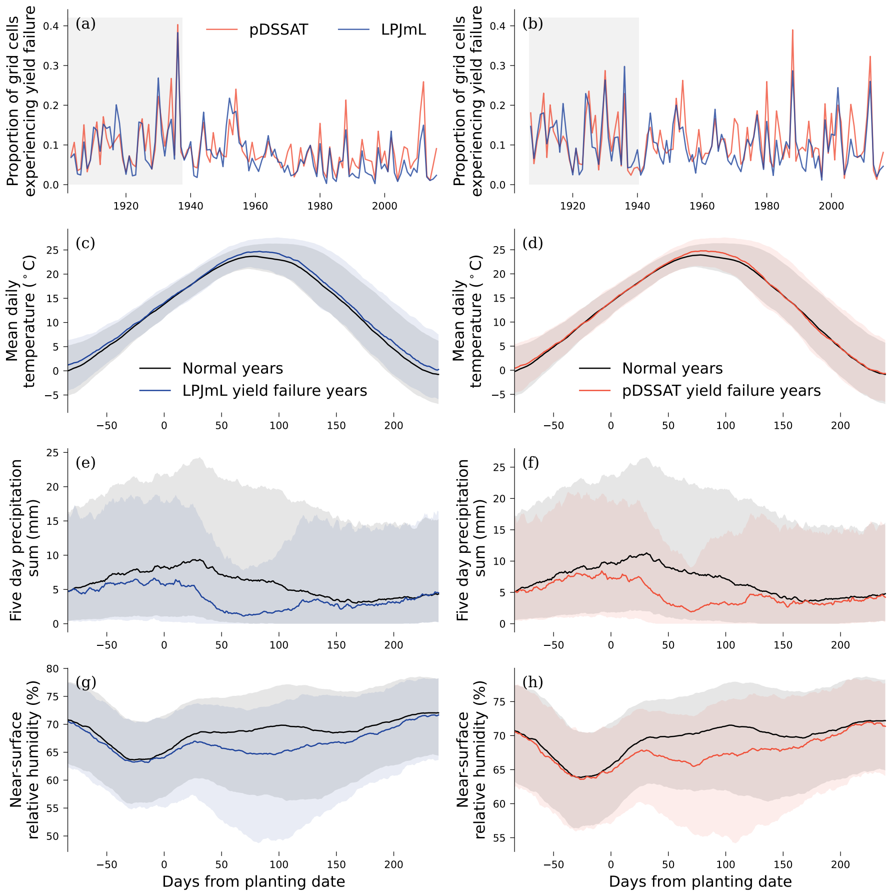

Figure B1. As Fig. 3 but with yield failure years defined using detrended yields.

Appendix B: Supporting figures for detrended data

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

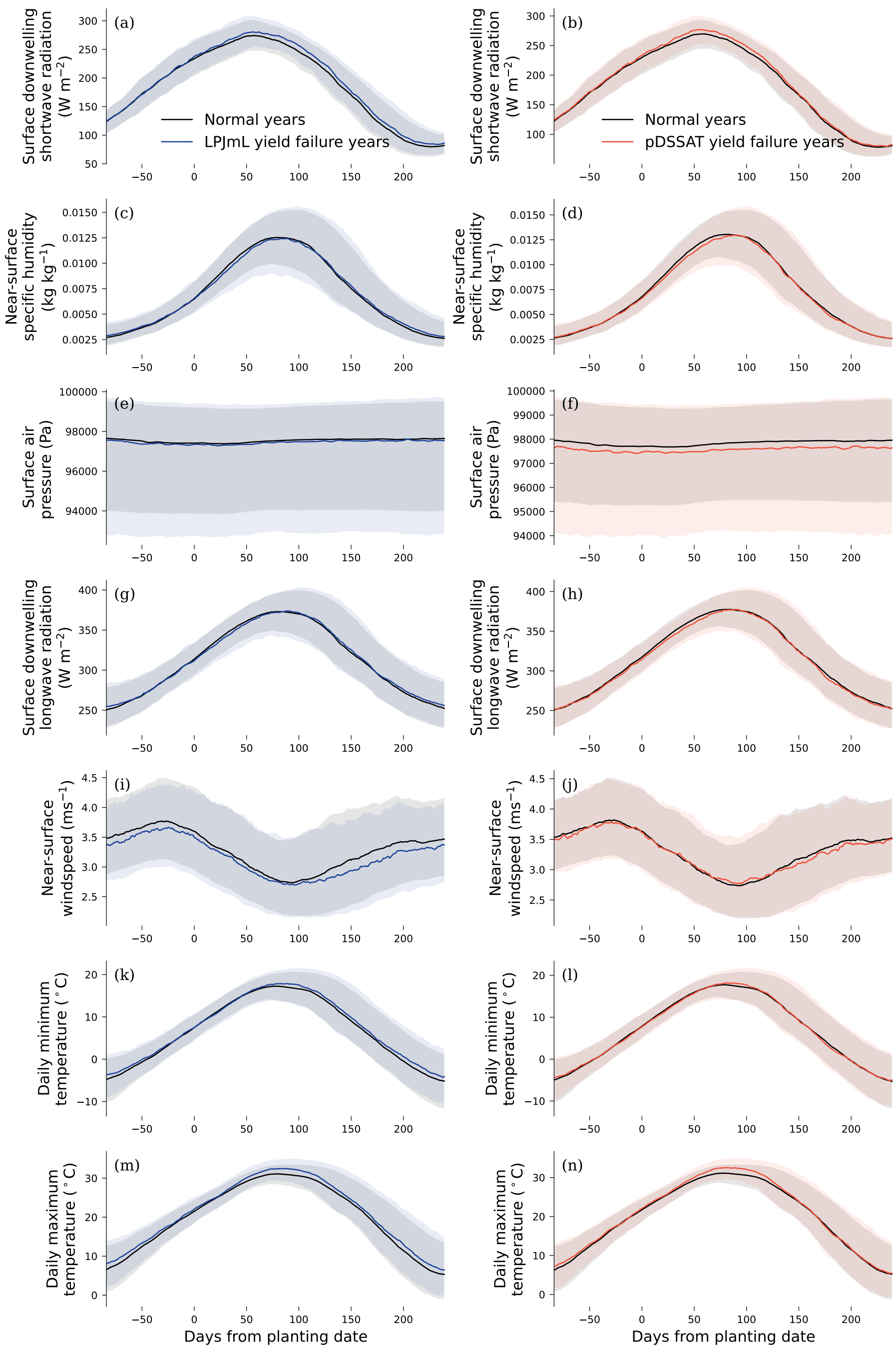

Figure B2. As Fig. A1 but for detrended yields.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

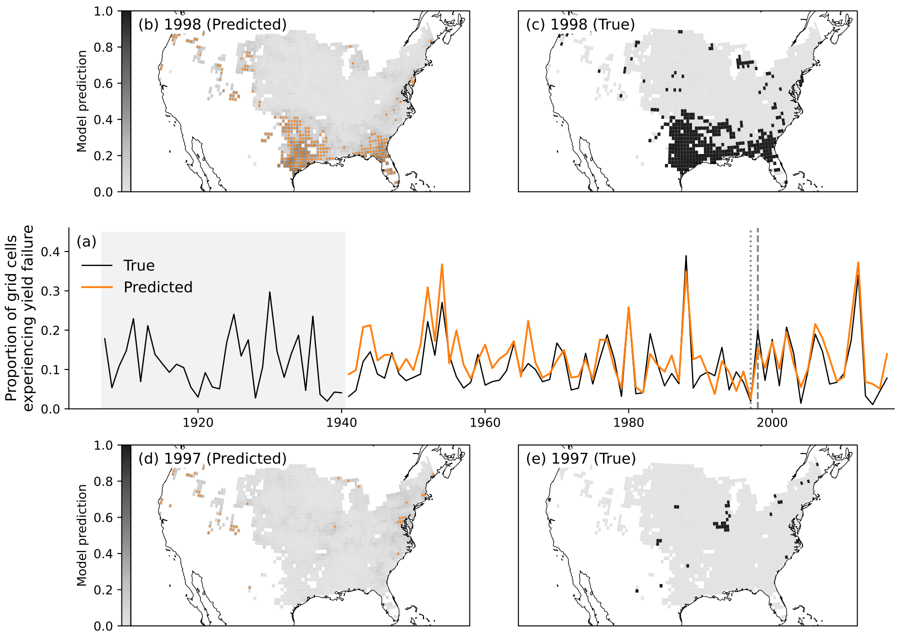

Figure B3. As Fig. 4, for pDSSAT after yields are detrended by removing the seven-year rolling average. The model achieves a ROC AUC of 0.82, average precision of 0.35, Brier score of 0.08 and log loss of 0.28 over the test period. The threshold used for predicting yield failure years is 0.195, which results in an f-score on the training years of 0.404. The Pearson correlation of predicted and true annual proportions of grid cells experiencing yield failure is 0.85.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

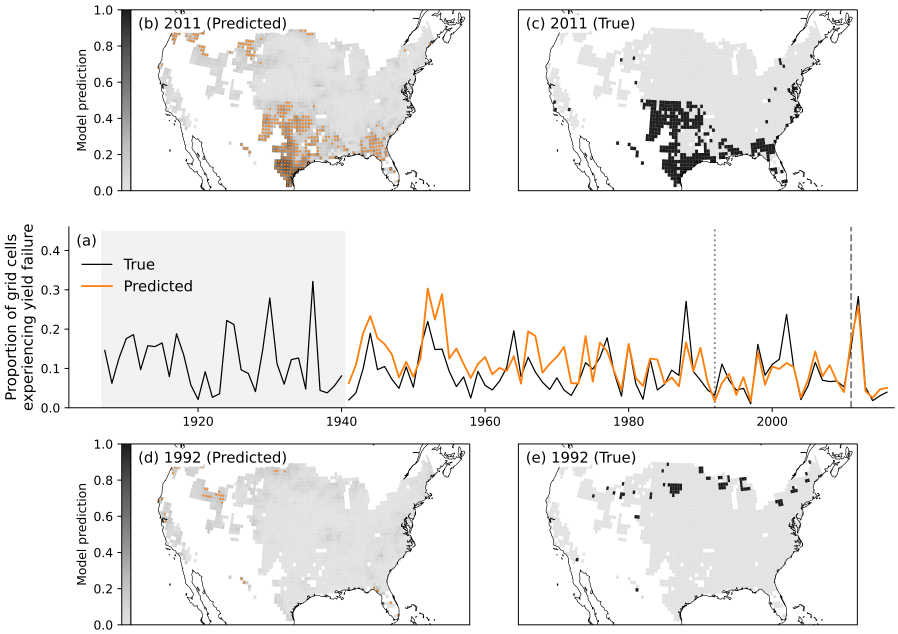

Figure B4. As Fig. 4, for LPJmL after yields are detrended. The model achieves a ROC AUC of 0.82, average precision of 0.33, Brier score of 0.07 and log loss of 0.24 over the test period. The threshold used for predicting yield failure years is 0.209, giving an f-score on the training years of 0.419. The Pearson correlation of predicted and true annual proportions of grid cells experiencing yield failure is 0.70.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

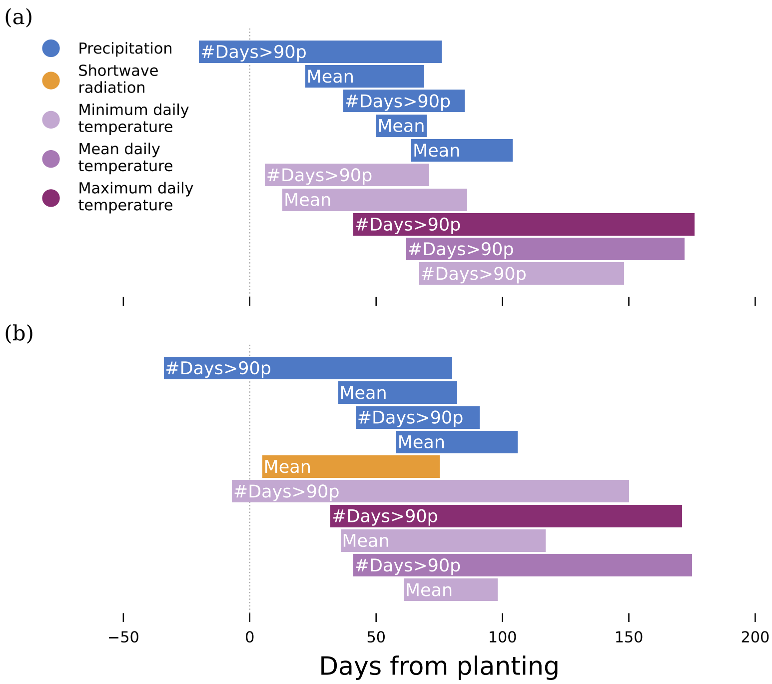

Figure B5. As Fig. 5, identified climate drivers of maize yield failure for (a) pDSSAT and (b) LPJmL when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

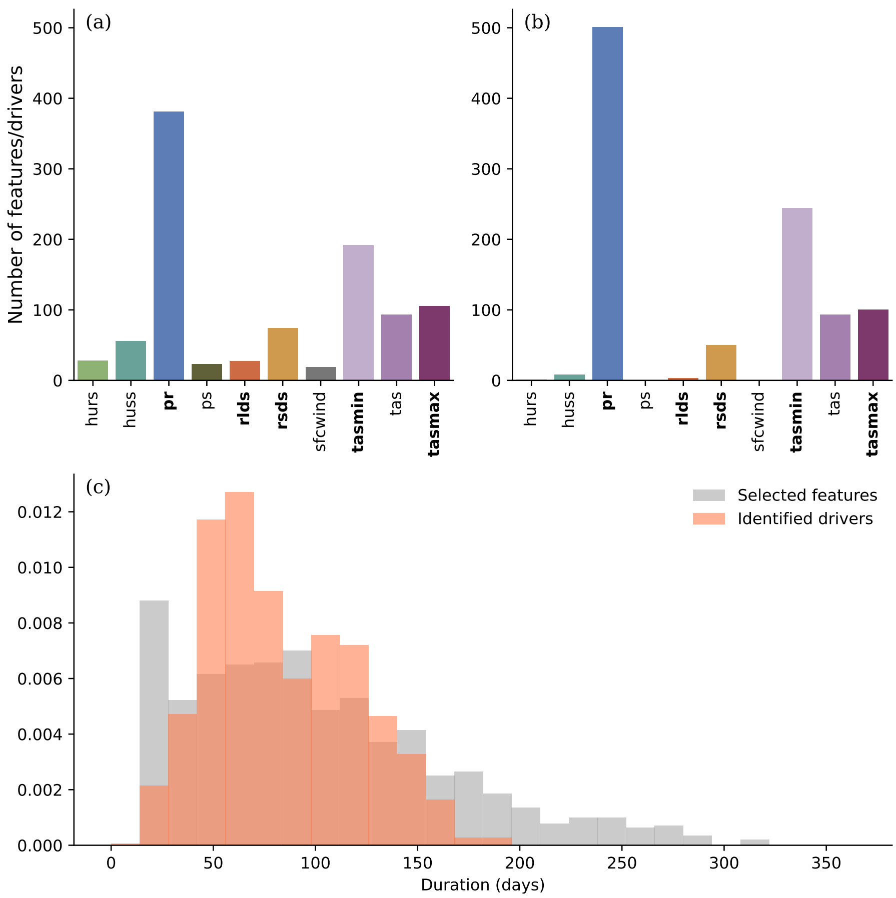

Figure B6. As Fig. A3, for pDSSAT when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

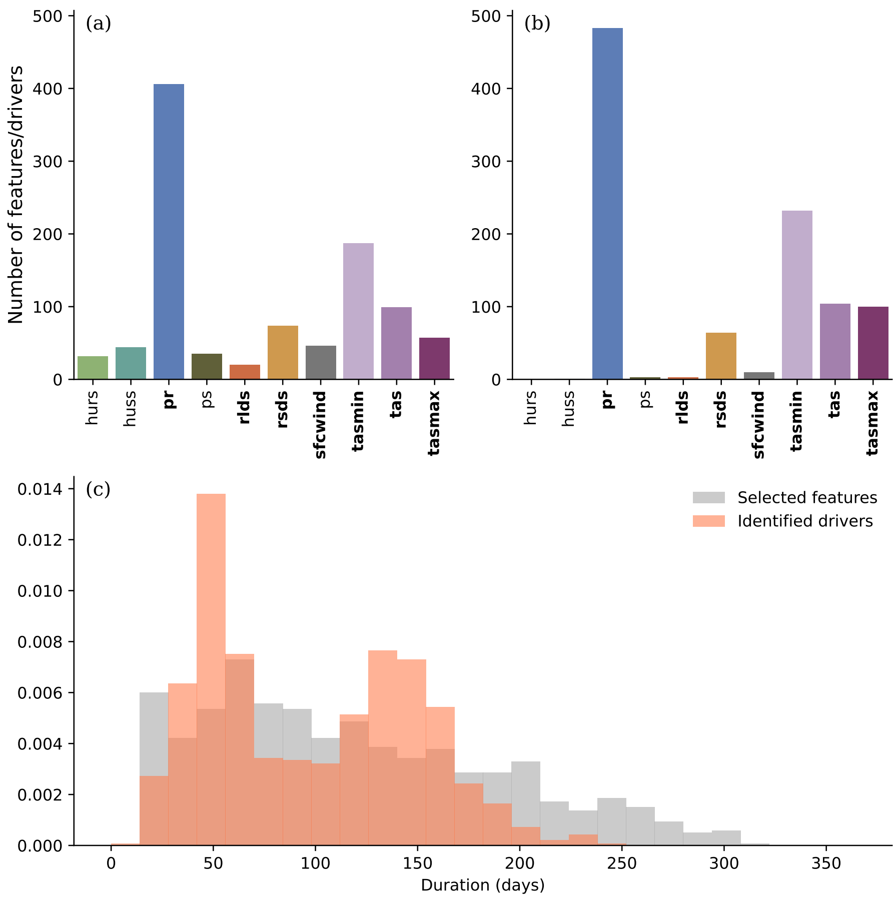

Figure B7. As Fig. A4, for LPJmL when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

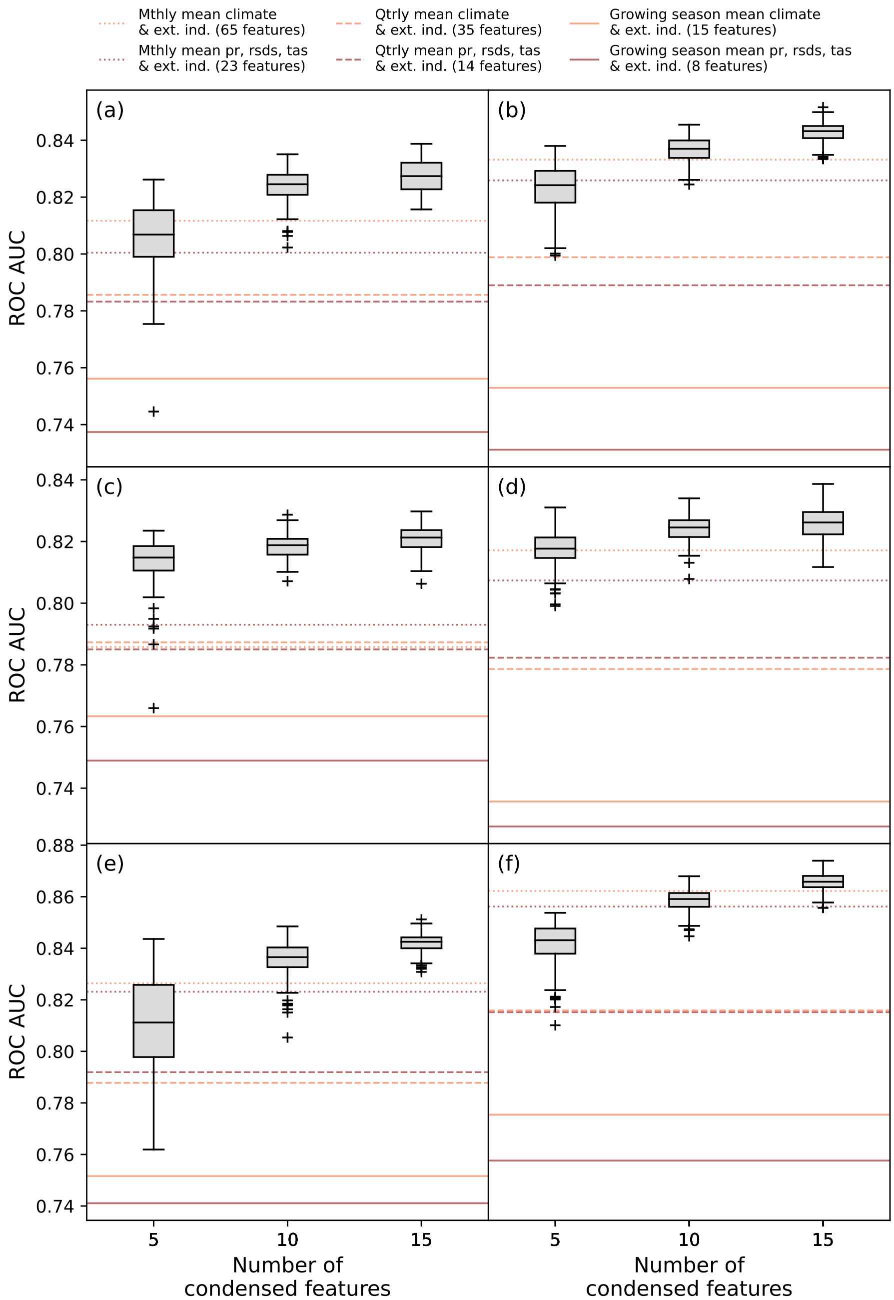

Figure B8. As Fig. A5, when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

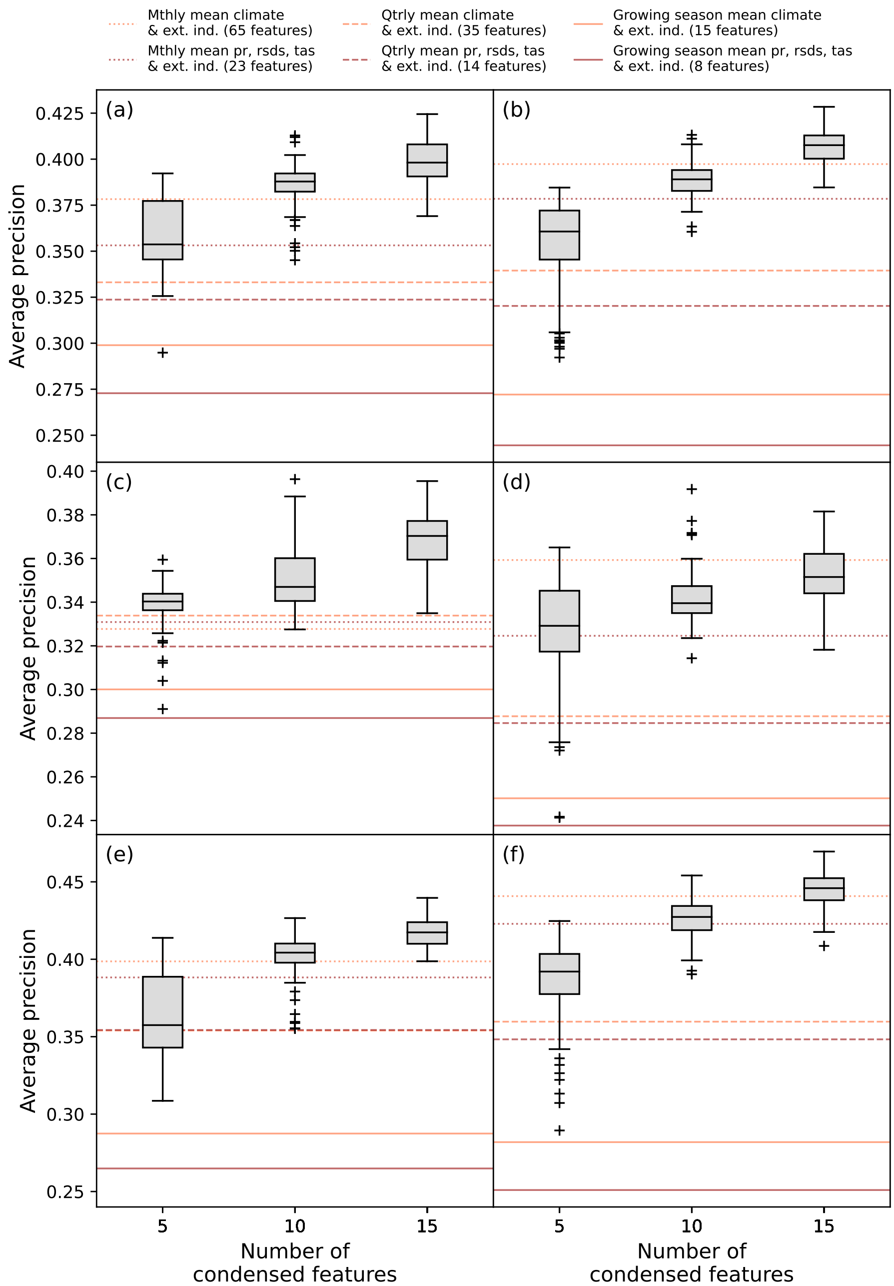

Figure B9. As Fig. A6, when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

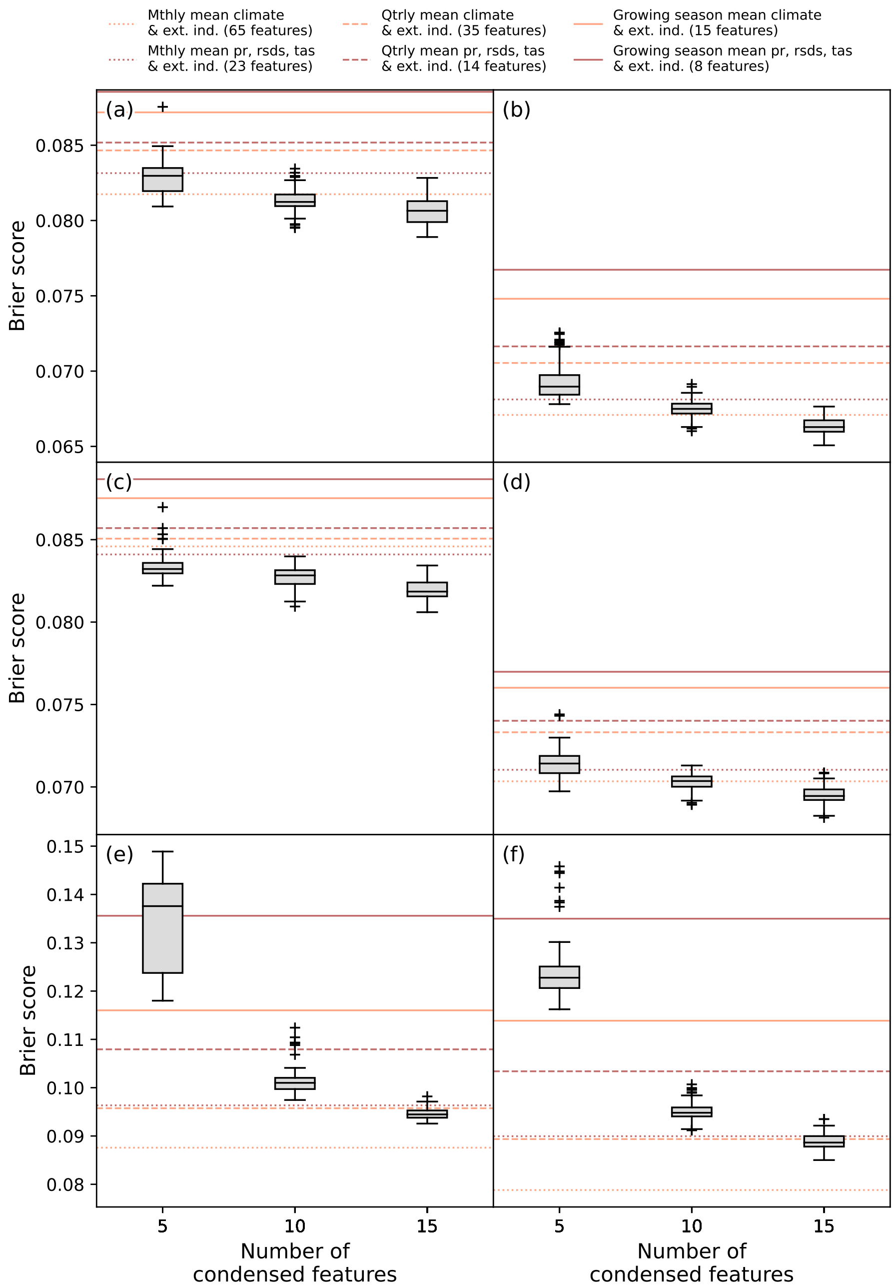

Figure B10. As Fig. A7, when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

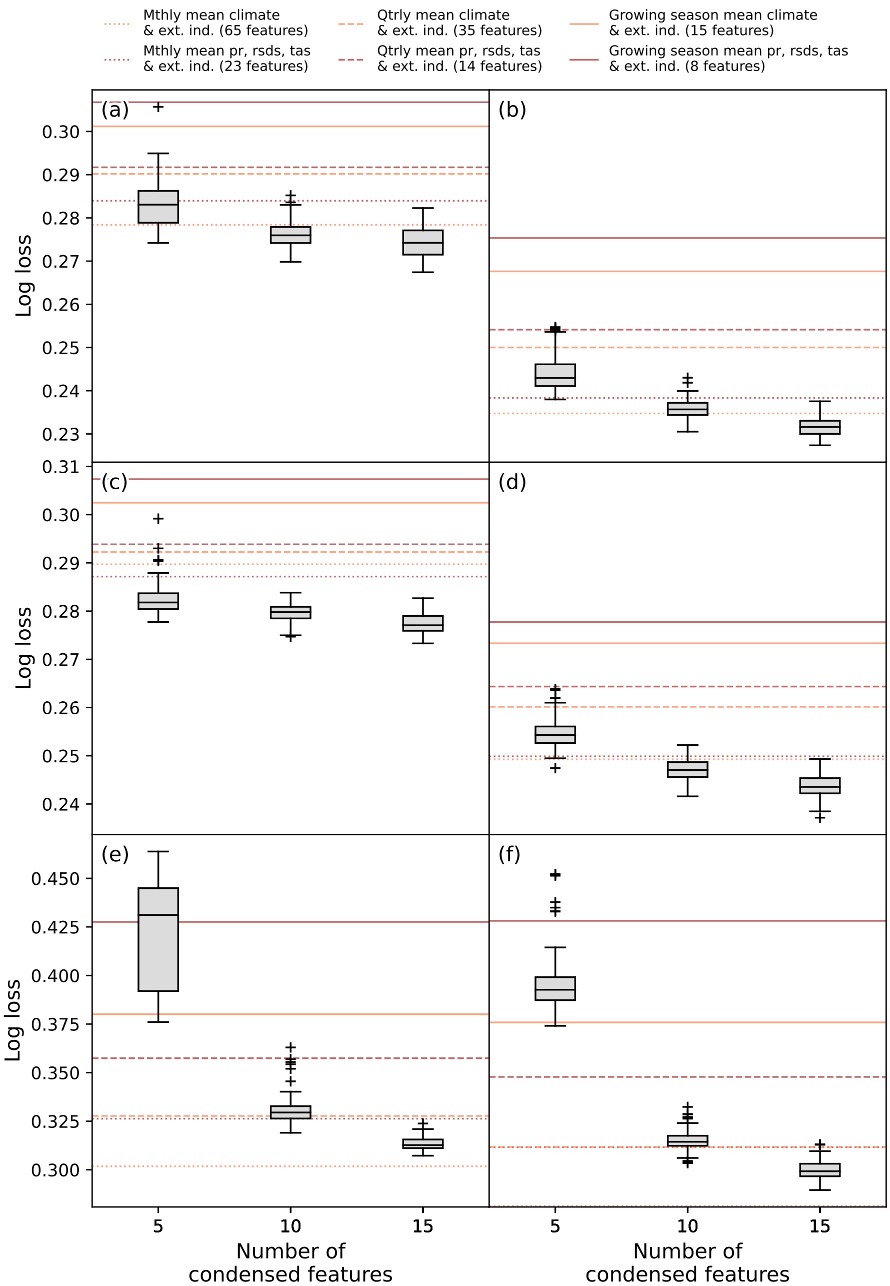

Figure B11. As Fig. A8, when yields are detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Appendix C: Sensitivity tests using different cross validation strategies

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

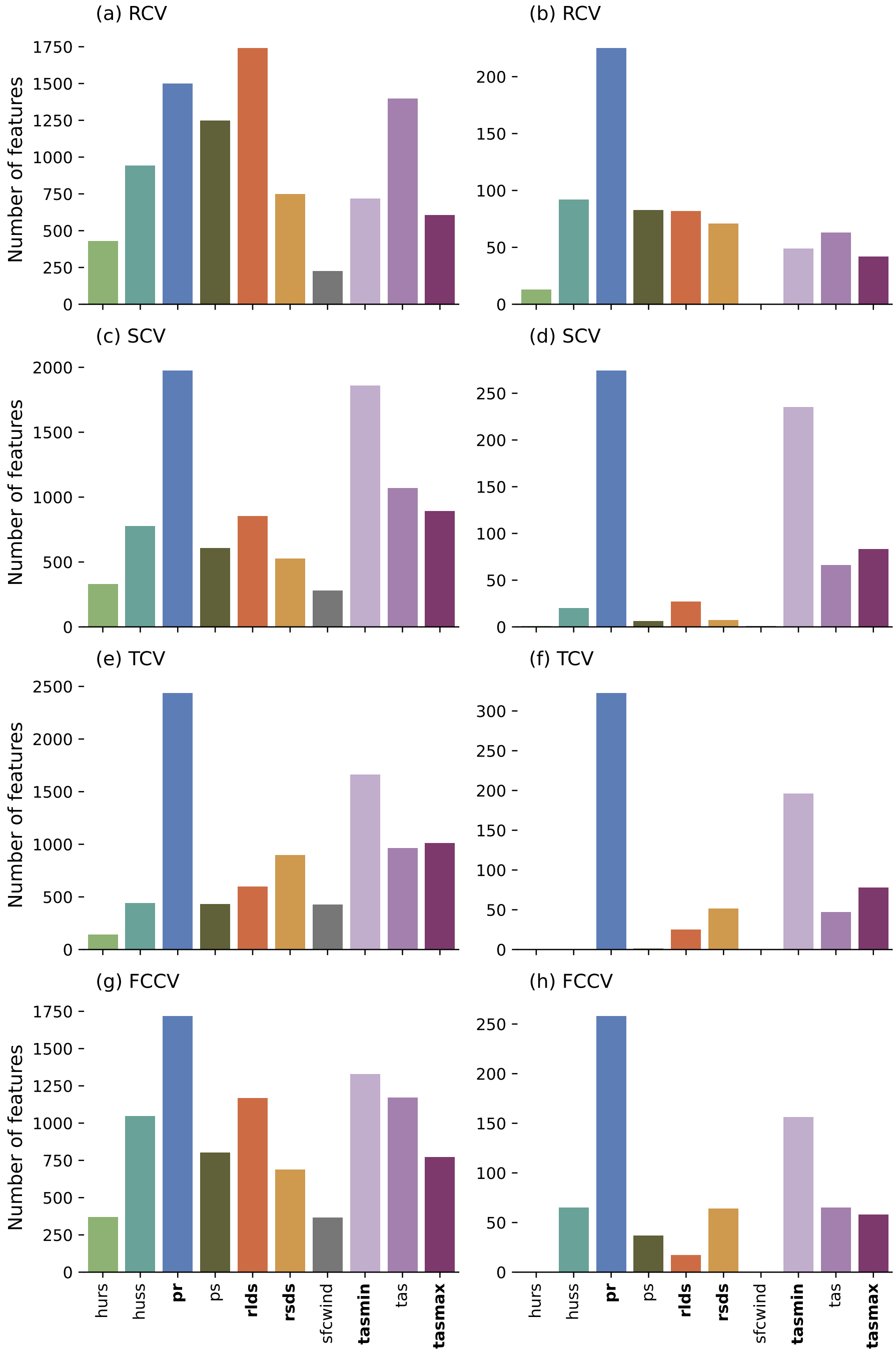

Figure C1. Climate variables used by features selected in Step 2 (a,c,e,g) and in drivers identified in Step 3 (b,d,f,h) for pDSSAT without detrending, depending on cross-validation strategy used to select features: random (RCV), spatial clusters (SCV), temporal (TCV) and feature-based clusters (FCCV). An exhaustive grid parameter sweep is performed (sampling 10, 20 or 30 time intervals in each pool; selecting 10, 20 or 30 features from each pool; and using 5, 10 or 15 cross-validation folds), using 20 pools per unique parameter combination, and sets of 5, 10 and 15 drivers are extracted. Features and drivers resulting from all parameter combinations are grouped by variable and counted.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

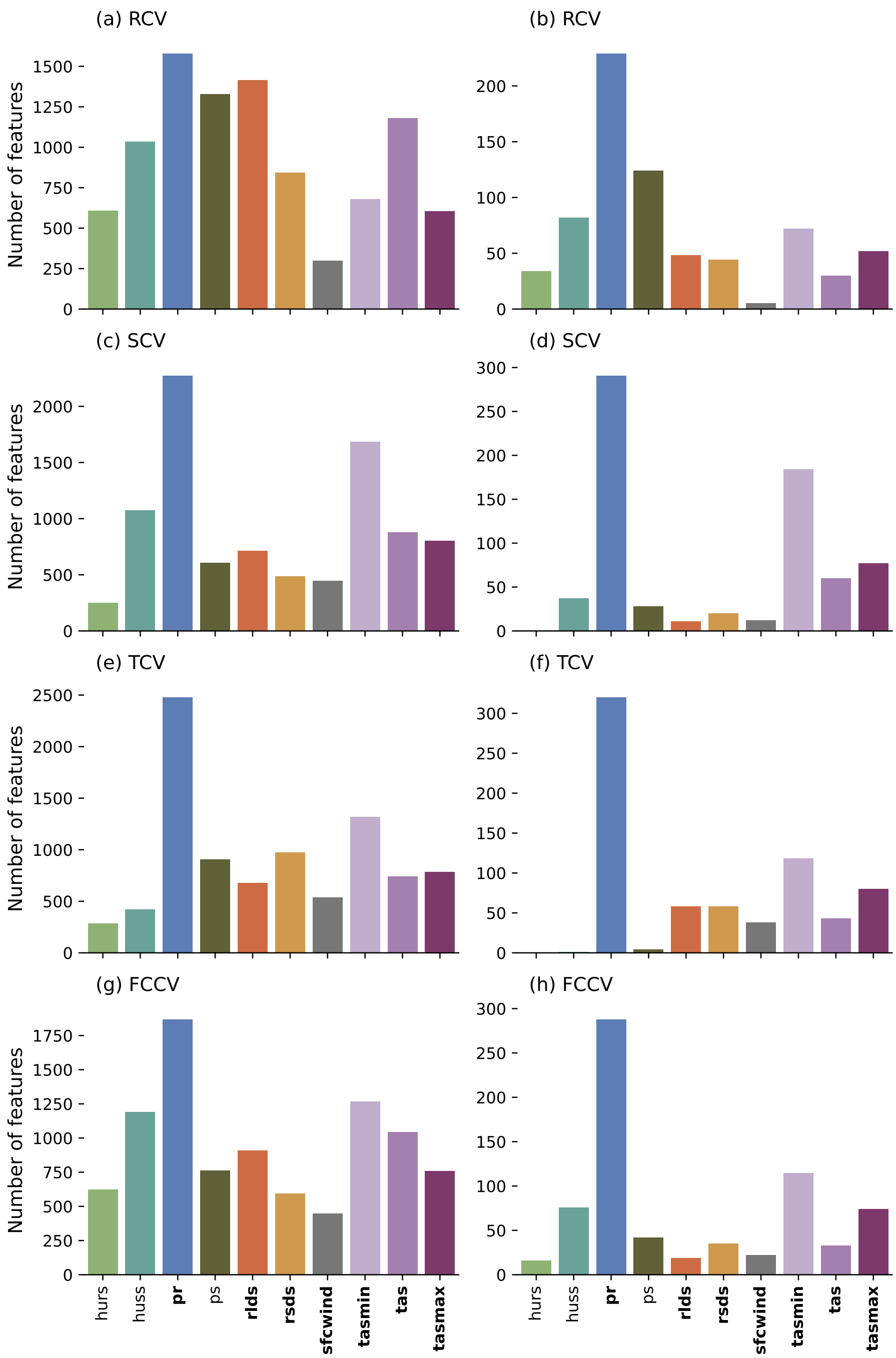

Figure C2. As in Fig. C1, but for LPJmL without detrending.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

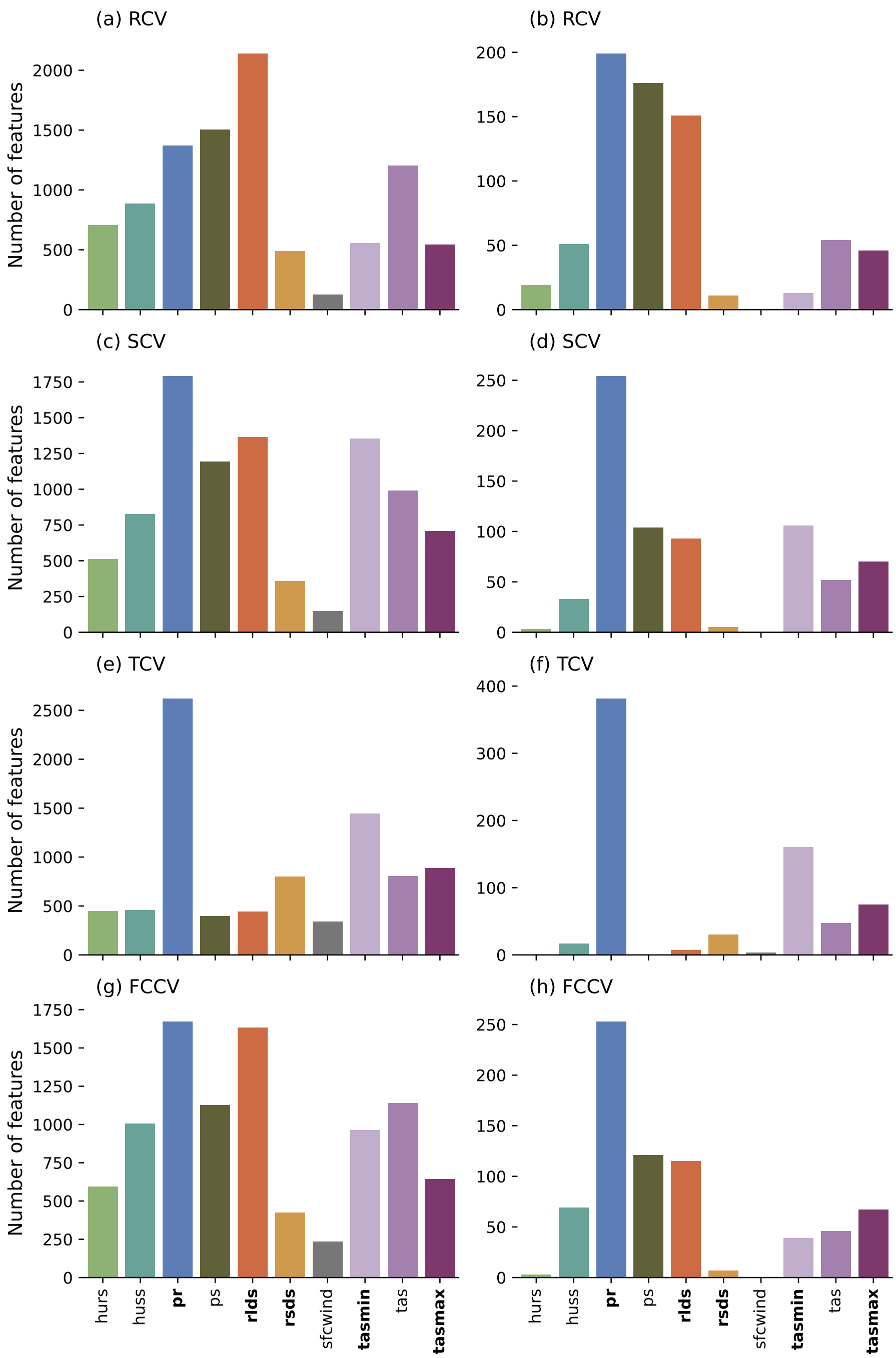

Figure C3. As in Fig. C1, but for pDSSAT with yields detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

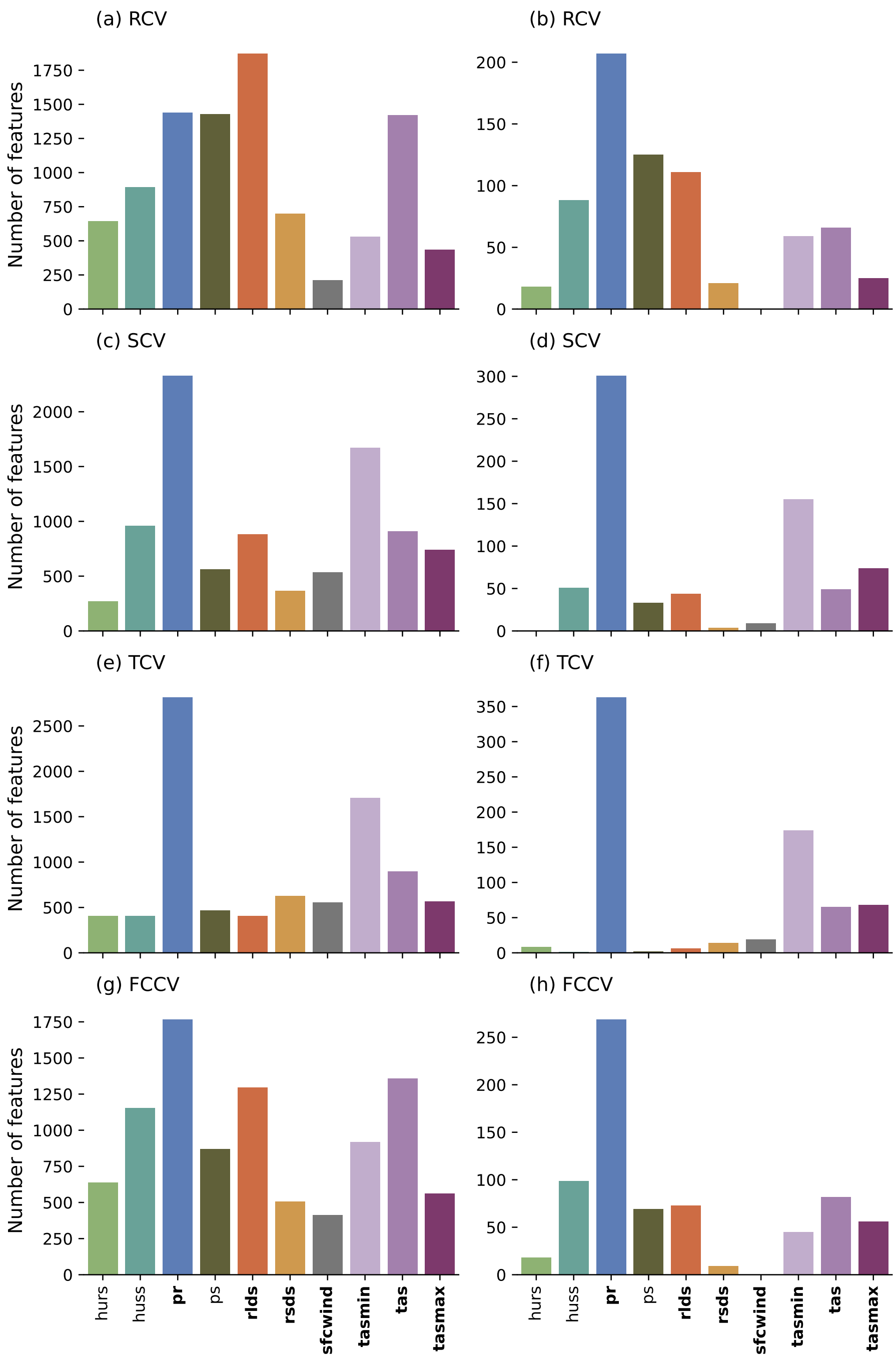

Figure C4. As in Fig. C1, but for LPJmL with yields detrended.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

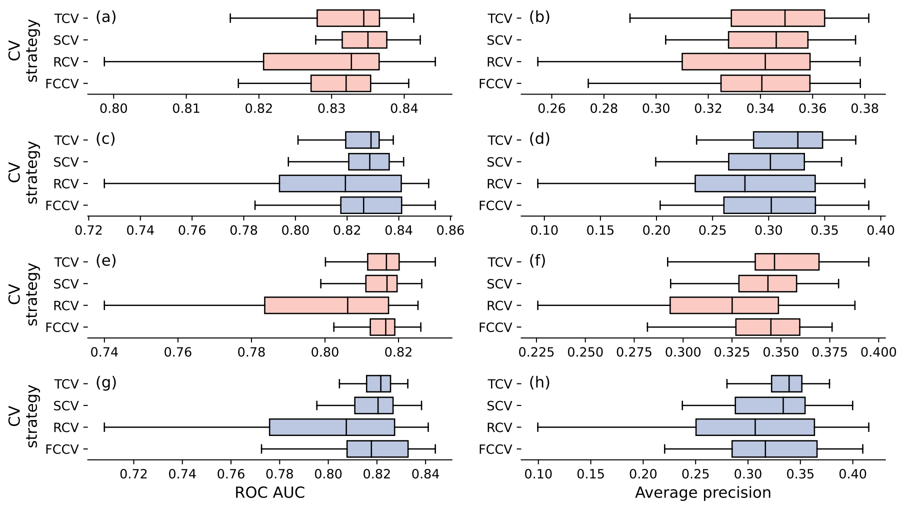

Figure C5. Performance on the test years of Lasso logistic regression models using identified driver sets as predictors, using different cross-validation strategies, based on the parameter sweeps described in Fig. C1. Panels (a) and (b) show results for non-detrended pDSSAT (red); panels (c) and (d) show non-detrended LPJmL (blue); panels (e) and (f) show detrended pDSSAT (red) and panels (g) and (h) show results for detrended LPJmL (blue).

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Author contributions. LS conceptualised the approach, developed the methodology, performed the experiments and prepared the initial draft with input and supervision from JZ. CM and JJ provided input and guidance on the crop yield simulations and corresponding climate forcing data and interpretation of the results. All authors contributed to the review and revision of the manuscript.

Competing interests. CM is a member of the editorial board of Geoscientific Model Development. The authors also have no other competing interests to declare.

Acknowledgements. LS and JZ acknowledge the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, grant agreement VH-NG-1537). We also thank the WKDV team at the UFZ for their work maintaining the EVE cluster on which the experiments and analysis in this manuscript were conducted.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

References

Anand, M., Bohn, F. J., Camps-Valls, G., Fischer, R., Huth, A., Sweet, L.-b., and Zscheischler, J.: Identifying compound weather drivers of forest biomass loss with generative deep learning, Environmental Data Science, 3, e4, 2024a.

Anand, M., Hamed, R., Linscheid, N., Silva, P. S., Andre, J., Zscheischler, J., Garry, F. K., and Bastos, A.: Winter Climate Preconditioning of Summer Vegetation Extremes in the Northern Hemisphere, Environmental Research Letters, 19, 094 045, https://doi.org/10.1088/1748-9326/ad627d, 2024b.

Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van der Velde, M., and Makowski, D.: Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France, Nature Communications, 9, 1627, 2018.

Bilodeau, B., Jaques, N., Koh, P. W., and Kim, B.: Impossibility theorems for feature attribution, Proceedings of the National Academy of Sciences, 121, e2304406 120, https://doi.org/10.1073/pnas.2304406120, 2024.

Chen, M., Guilpart, N., and Makowski, D.: Comparison of methods to aggregate climate data to predict crop yield: an application to soybean, Environmental Research Letters, 19, 054 049, https://doi.org/10.1088/1748-9326/ad42b5, 2024.

Crane-Droesch, A.: Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, Environmental Research Letters, 13, 114 003, https://doi.org/10.1088/1748-9326/aae159, 2018.

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

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2: Multimodel Analysis and Implications for Our Perception of the Land Surface, Bulletin of the American Meteorological Society, https://doi.org/10.1175/BAMS-87-10-1381, 2006.

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., and Foster, I.: The parallel system for integrating impact models and sectors (pSIMS), Environ. Model. Software, 62, 509–516, https://doi.org/10.1016/j.envsoft.2014.04.008, 2014.

Ferraciolli, M. A., Bocca, F. F., and Rodrigues, L. H. A.: Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models, Computers and Electronics in Agriculture, 161, 233–240, https://doi.org/10.1016/j.compag.2018.09.003, 2019.

Filippi, P., Han, S. Y., and Bishop, T. F.: On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies, Precision Agriculture, 26, 8, https://doi.org/10.1007/s11119-024-10212-2, 2024.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M.: Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning, Agricultural and Forest Meteorology, 264, 1–15, https://doi.org/10.1016/j.agrformet.2018.09.021, 2019.

Franch, B., Cintas, J., Becker-Reshef, I., Sanchez-Torres, M. J., Roger, J., Skakun, S., Sobrino, J. A., Van Tricht, K., Degerickx, J., Gilliams, S., Koetz, B., Szantoi, Z., and Whitcraft, A.: Global crop calendars of maize and wheat in the framework of the WorldCereal project, GIScience & Remote Sensing, 59, 885–913, https://doi.org/10.1080/15481603.2022.2079273, 2022.

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

Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees, Machine Learning, 63, 3–42, https://doi.org/10.1007/s10994-006-6226-1, 2006.

Han, J., Shi, L., Yang, Q., Yu, J., and Athanasiadis, I. N.: Knowledge-guided machine learning with multivariate sparse data for crop growth modelling, Field Crops Research, 328, 109 912, https://doi.org/10.1016/j.fcr.2025.109912, 2025.

Haqiqi, I., Grogan, D. S., Hertel, T. W., and Schlenker, W.: Quantifying the impacts of compound extremes on agriculture, Hydrology and Earth System Sciences, 25, 551–564, https://doi.org/10.5194/hess-25-551-2021, 2021.

Heilemann, J., Klassert, C., Samaniego, L., Thober, S., Marx, A., Boeing, F., Klauer, B., and Gawel, E.: Projecting impacts of extreme weather events on crop yields using LASSO regression, Weather and Climate Extremes, 46, 100 738, https://doi.org/10.1016/j.wace.2024.100738, 2024.

Heinicke, S., Frieler, K., Jägermeyr, J., and Mengel, M.: Global gridded crop models underestimate yield responses to droughts and heatwaves, Environmental Research Letters, 17, 044 026, https://doi.org/10.1088/1748-9326/ac592e, 2022.

Hoffman, A. L., Kemanian, A. R., and Forest, C. E.: The response of maize, sorghum, and soybean yield to growing-phase climate revealed with machine learning, Environmental Research Letters, 15, 094 013, https://doi.org/10.1088/1748-9326/ab7b22, 2020.

Hsiao, J., Swann, A. L. S., and Kim, S.-H.: Maize yield under a changing climate: The hidden role of vapor pressure deficit, Agricultural and Forest Meteorology, 279, 107 692, https://doi.org/10.1016/j.agrformet.2019.107692, 2019.

Hultgren, A., Carleton, T., Delgado, M., Gergel, D. R., Greenstone, M., Houser, T., Hsiang, S., Jina, A., Kopp, R. E., Malevich, S. B., McCusker, K. E., Mayer, T., Nath, I., Rising, J., Rode, A., and Yuan, J.: Impacts of climate change on global agriculture accounting for adaptation, Nature, 642, 644–652, https://doi.org/10.1038/s41586-025-09085-w, 2025.

Hyungjun Kim: Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1), https://doi.org/doi:10.20783/DIAS.501, version 1, 2017.

Jiang, S., Sweet, L.-b., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., and Zscheischler, J.: How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences, Earth's Future, 12, e2024EF004 540, https://doi.org/10.1029/2024EF004540, 2024a.

Jiang, S., Tarasova, L., Yu, G., and Zscheischler, J.: Compounding effects in flood drivers challenge estimates of extreme river floods, Science Advances, 10, eadl4005, 2024b.

Jin, Z., Zhuang, Q., Wang, J., Archontoulis, S. V., Zobel, Z., and Kotamarthi, V. R.: The combined and separate impacts of climate extremes on the current and future US rainfed maize and soybean production under elevated CO2, Global Change Biology, 23, 2687–2704, https://doi.org/10.1111/gcb.13617, 2017.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., and Ritchie, J. T.: The DSSAT cropping system model, Eur. J. Agron., 18, 235–265, https://doi.org/10.1016/S1161-0301(02)00107-7, 2003.

Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., Faye, B., Foster, I., Folberth, C., Franke, J. A., Fuchs, K., Guarin, J. R., Heinke, J., Hoogenboom, G., Iizumi, T., Jain, A. K., Kelly, D., Khabarov, N., Lange, S., Lin, T.-S., Liu, W., Mialyk, O., Minoli, S., Moyer, E. J., Okada, M., Phillips, M., Porter, C., Rabin, S. S., Scheer, C., Schneider, J. M., Schyns, J. F., Skalsky, R., Smerald, A., Stella, T., Stephens, H., Webber, H., Zabel, F., and Rosenzweig, C.: Climate impacts on global agriculture emerge earlier in new generation of climate and crop models, Nature Food, 2, 873–885, https://doi.org/10.1038/s43016-021-00400-y, 2021.

Kim, Y.-U., Ruane, A. C., Finger, R., and Webber, H.: Robust assessment of climatic risks to crop production, Nature Food, pp. 1–2, https://doi.org/10.1038/s43016-025-01168-1, 2025.

Lange, S.: Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0), Geoscientific Model Development, 12, 3055–3070, https://doi.org/10.5194/gmd-12-3055-2019, 2019.

Lange, S.: ISIMIP3BASD, https://doi.org/10.5281/zenodo.4686991, 2021.

Lange, S., Menz, C., Gleixner, S., Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Schmied, H. M., Hersbach, H., Buontempo, C., and Cagnazzo, C.: WFDE5 over land merged with ERA5 over the ocean (W5E5 v2.0), https://doi.org/10.48364/ISIMIP.342217, version Number: 2.0, 2021.

Lange, S., Mengel, M., Treu, S., and Büchner, M.: ISIMIP3a atmospheric climate input data, https://doi.org/10.48364/ISIMIP.982724.1, version Number: 1.1, 2022.

Lesk, C., Coffel, E., Winter, J., Ray, D., Zscheischler, J., Seneviratne, S. I., and Horton, R.: Stronger temperature–moisture couplings exacerbate the impact of climate warming on global crop yields, Nature Food, 2, 683–691, https://doi.org/10.1038/s43016-021-00341-6, 2021.

Lin, T., Zhong, R., Wang, Y., Xu, J., Jiang, H., Xu, J., Ying, Y., Rodriguez, L., Ting, K. C., and Li, H.: DeepCropNet: a deep spatial-temporal learning framework for county-level corn yield estimation, Environmental Research Letters, 15, 034 016, https://doi.org/10.1088/1748-9326/ab66cb, 2020.

Lischeid, G., Webber, H., Sommer, M., Nendel, C., and Ewert, F.: Machine learning in crop yield modelling: A powerful tool, but no surrogate for science, Agricultural and Forest Meteorology, 312, 108 698, https://doi.org/10.1016/j.agrformet.2021.108698, 2022.

Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., and Wang, G.: Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model, Artificial Intelligence for the Earth Systems, 1, https://doi.org/10.1175/AIES-D-22-0002.1, 2022.

Liu, W., Ye, T., Müller, C., Jägermeyr, J., Franke, J. A., Stephens, H., and Chen, S.: The statistical emulators of GGCMI phase 2: responses of year-to-year variation of crop yield to CO_2, temperature, water, and nitrogen perturbations, Geoscientific Model Development, 16, 7203–7221, https://doi.org/10.5194/gmd-16-7203-2023, 2023.

Lobell, D. B. and Di Tommaso, S.: A half-century of climate change in major agricultural regions: Trends, impacts, and surprises, Proceedings of the National Academy of Sciences, 122, e2502789 122, https://doi.org/10.1073/pnas.2502789122, 2025.

Lobell, D. B., Hammer, G. L., McLean, G., Messina, C., Roberts, M. J., and Schlenker, W.: The critical role of extreme heat for maize production in the United States, Nature Climate Change, 3, 497–501, https://doi.org/10.1038/nclimate1832, 2013.

Lobell, D. B., Roberts, M. J., Schlenker, W., Braun, N., Little, B. B., Rejesus, R. M., and Hammer, G. L.: Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest, Science, 344, 516–519, https://doi.org/10.1126/science.1251423, 2014.

Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., and Meyer, H.: Assessing and improving the transferability of current global spatial prediction models, Global Ecology and Biogeography, 32, 356–368, https://doi.org/10.1111/geb.13635, 2023.

Lutz, F., Herzfeld, T., Heinke, J., Rolinski, S., Schaphoff, S., von Bloh, W., Stoorvogel, J. J., and Müller, C.: Simulating the effect of tillage practices with the global ecosystem model LPJmL (version 5.0-tillage), Geoscientific Model Development, 12, 2419–2440, https://doi.org/10.5194/gmd-12-2419-2019, 2019.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I.: Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience, Artificial Intelligence for the Earth Systems, 1, https://doi.org/10.1175/AIES-D-22-0012.1, 2022.

Martínez-Ferrer, L., Piles, M., and Camps-Valls, G.: Crop Yield Estimation and Interpretability With Gaussian Processes, IEEE Geoscience and Remote Sensing Letters, pp. 1–5, https://doi.org/10.1109/LGRS.2020.3016140, 2020.

Matiu, M., Ankerst, D. P., and Menzel, A.: Interactions between temperature and drought in global and regional crop yield variability during 1961-2014, PLOS ONE, 12, e0178 339, https://doi.org/10.1371/journal.pone.0178339, 2017.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation, Environmental Modelling & Software, 101, 1–9, https://doi.org/10.1016/j.envsoft.2017.12.001, 2018.

Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T.: Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction, Ecological Modelling, 411, 108 815, https://doi.org/10.1016/j.ecolmodel.2019.108815, 2019.

Mistry, M. N., Wing, I. S., and Cian, E. D.: Simulated vs. empirical weather responsiveness of crop yields: US evidence and implications for the agricultural impacts of climate change, Environmental Research Letters, 12, 075 007, https://doi.org/10.1088/1748-9326/aa788c, 2017.

Moon, T., Kim, D., Kwon, S., and Son, J. E.: Process-Based Crop Modeling for High Applicability with Attention Mechanism and Multitask Decoders, Plant Phenomics, 5, 0035, https://doi.org/10.34133/plantphenomics.0035, 2023.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S., Iizumi, T., Izaurralde, R. C., Jones, C., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T. A. M., Ray, D. K., Reddy, A., Rosenzweig, C., Ruane, A. C., Sakurai, G., Schmid, E., Skalsky, R., Song, C. X., Wang, X., de Wit, A., and Yang, H.: Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications, Geoscientific Model Development, 10, 1403–1422, https://doi.org/10.5194/gmd-10-1403-2017, 2017.

Müller, C., Jägermeyr, J., Franke, J. A., Ruane, A. C., Balkovic, J., Ciais, P., Dury, M., Falloon, P., Folberth, C., Hank, T., Hoffmann, M., Izaurralde, R. C., Jacquemin, I., Khabarov, N., Liu, W., Olin, S., Pugh, T. A. M., Wang, X., Williams, K., Zabel, F., and Elliott, J. W.: Substantial Differences in Crop Yield Sensitivities Between Models Call for Functionality-Based Model Evaluation, Earth's Future, 12, e2023EF003 773, https://doi.org/10.1029/2023EF003773, 2024.

Nóia Júnior, R. d. S., Asseng, S., Müller, C., Deswarte, J.-C., Cohan, J.-P., and Martre, P.: Negative impacts of climate change on crop yields are underestimated, Trends Plant Sci., https://doi.org/10.1016/j.tplants.2025.05.002, 2025.

Peters, J., Bühlmann, P., and Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78, 947–1012, https://doi.org/10.1111/rssb.12167, 2016.

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

PRISM Group: PRISM Climate Group Data, http://prism.oregonstate.edu, accessed May 2025, 2018.

Richens, J. and Everitt, T.: Robust agents learn causal world models, in: The Twelfth International Conference on Learning Representations, 2024.

Rigden, A. J., Mueller, N. D., Holbrook, N. M., Pillai, N., and Huybers, P.: Combined influence of soil moisture and atmospheric evaporative demand is important for accurately predicting US maize yields, Nature Food, 1, 127–133, https://doi.org/10.1038/s43016-020-0028-7, 2020.

Ryo, M.: Explainable artificial intelligence and interpretable machine learning for agricultural data analysis, Artificial Intelligence in Agriculture, 6, 257–265, https://doi.org/10.1016/j.aiia.2022.11.003, 2022.

Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., Pugh, T. A. M., Rolinski, S., Schaphoff, S., Schmid, E., Wang, X., Schlenker, W., and Frieler, K.: Consistent negative response of US crops to high temperatures in observations and crop models, Nature Communications, 8, 13 931, https://doi.org/10.1038/ncomms13931, 2017.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Schlenker, W. and Roberts, M. J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change, Proceedings of the National Academy of Sciences, 106, 15 594–15 598, https://doi.org/10.1073/pnas.0906865106, 2009.

Shahhosseini, M., Martinez-Feria, R. A., Hu, G., and Archontoulis, S. V.: Maize yield and nitrate loss prediction with machine learning algorithms, Environmental Research Letters, 14, 124 026, https://doi.org/10.1088/1748-9326/ab5268, 2019.

Siebers, M. H., Slattery, R. A., Yendrek, C. R., Locke, A. M., Drag, D., Ainsworth, E. A., Bernacchi, C. J., and Ort, D. R.: Simulated heat waves during maize reproductive stages alter reproductive growth but have no lasting effect when applied during vegetative stages, Agriculture, Ecosystems & Environment, 240, 162–170, https://doi.org/10.1016/j.agee.2016.11.008, 2017.

Siebert, S., Webber, H., and Rezaei, E. E.: Weather impacts on crop yields - searching for simple answers to a complex problem, Environmental Research Letters, 12, 081 001, https://doi.org/10.1088/1748-9326/aa7f15, 2017a.

Siebert, S., Webber, H., Zhao, G., and Ewert, F.: Heat stress is overestimated in climate impact studies for irrigated agriculture, Environmental Research Letters, 12, 054 023, https://doi.org/10.1088/1748-9326/aa702f, 2017b.

Silva, J. V., Heerwaarden, J. v., Reidsma, P., Laborte, A. G., Tesfaye, K., and Ittersum, M. K. v.: Big data, small explanatory and predictive power: Lessons from random forest modeling of on-farm yield variability and implications for data-driven agronomy, Field Crops Research, 302, 109 063, https://doi.org/10.1016/j.fcr.2023.109063, 2023.

Sweet, L.-b.: A data-driven method for identifying climate drivers of agricultural yield failure from daily weather data, https://doi.org/10.5281/zenodo.15725041, 2025.

Sweet, L.-b., Müller, C., Anand, M., and Zscheischler, J.: Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models, Artificial Intelligence for the Earth Systems, 2, https://doi.org/10.1175/AIES-D-23-0026.1, 2023.

Sweet, L.-b., Athanasiadis, I. N., Bree, R. v., Castellano, A., Martre, P., Paudel, D., Ruane, A. C., and Zscheischler, J.: Transdisciplinary coordination is essential for advancing agricultural modeling with machine learning, One Earth, 8, https://doi.org/10.1016/j.oneear.2025.101233, 2025.

Ting, M., Lesk, C., Liu, C., Li, C., Horton, R. M., Coffel, E. D., Rogers, C. D. W., and Singh, D.: Contrasting impacts of dry versus humid heat on US corn and soybean yields, Scientific Reports, 13, 710, https://doi.org/10.1038/s41598-023-27931-7, 2023.

Troy, T. J., Kipgen, C., and Pal, I.: The impact of climate extremes and irrigation on US crop yields, Environmental Research Letters, 10, 054 013, https://doi.org/10.1088/1748-9326/10/5/054013, 2015.

United States Department of Agriculture (USDA): Quick Stats 2.0, https://quickstats.nass.usda.gov/, 2025.

U.S. Census Bureau: TIGER/Line Shapefile, 2018, Current State and Equivalent National, https://catalog.data.gov/dataset/tiger-line-shapefile-2018-nation-u-s-current-state-and-equivalent-national, 2018.

Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., Meinshausen, N., and Frieler, K.: The effects of climate extremes on global agricultural yields, Environmental Research Letters, 14, 054 010, https://doi.org/10.1088/1748-9326/ab154b, 2019.

Vogel, J., Rivoire, P., Deidda, C., Rahimi, L., Sauter, C. A., Tschumi, E., van der Wiel, K., Zhang, T., and Zscheischler, J.: Identifying meteorological drivers of extreme impacts: an application to simulated crop yields, Earth System Dynamics, 12, 151–172, https://doi.org/10.5194/esd-12-151-2021, 2021.

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

Wadoux, A. M. J.-C., Samuel-Rosa, A., Poggio, L., and Mulder, V. L.: A note on knowledge discovery and machine learning in digital soil mapping, European Journal of Soil Science, 71, 133–136, https://doi.org/10.1111/ejss.12909, 2020.

https://doi.org/10.5194/egusphere-2025-3006
Preprint. Discussion started: 26 August 2025

Wang, B., Jägermeyr, J., O'Leary, G. J., Wallach, D., Ruane, A. C., Feng, P., Li, L., Liu, D. L., Waters, C., Yu, Q., Asseng, S., and Rosenzweig, C.: Pathways to identify and reduce uncertainties in agricultural climate impact assessments, Nature Food, pp. 1–7, https://doi.org/10.1038/s43016-024-01014-w, 2024.

Webber, H., Rezaei, E. E., Ryo, M., and Ewert, F.: Framework to guide modeling single and multiple abiotic stresses in arable crops, Agriculture, Ecosystems & Environment, 340, 108 179, https://doi.org/10.1016/j.agee.2022.108179, 2022.

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., and Guanter, L.: Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt, Environmental Research Letters, 15, 024 019, https://doi.org/10.1088/1748-9326/ab68ac, 2020.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J.-L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., and Asseng, S.: Temperature increase reduces global yields of major crops in four independent estimates, Proceedings of the National Academy of Sciences, 114, 9326–9331, https://doi.org/10.1073/pnas.1701762114, 2017.

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

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., Maraun, D., Ramos, A. M., Ridder, N. N., Thiery, W., and Vignotto, E.: A typology of compound weather and climate events, Nature Reviews Earth & Environment, 1, 333–347, https://doi.org/10.1038/s43017-020-0060-z, 2020.