A data-driven framework for assessing climatic impact drivers in the context of food security

Marcos Roberto Benso¹, Roberto Fray Silva², Gabriela Chiquito Gesualdo^{1,5}, Antonio Mauro Saraiva², Alexandre Cláudio Botazzo Delbem⁴, Patricia Angélica Alves Marques³, José Antonio Marengo^{6,7,8}, and Eduardo Mario Mendiondo¹

¹São Carlos School of Engineering, University of São Paulo, São Carlos, SP, 13566-590, Brazil
²Institute of Advanced Studies, University of São Paulo, São Paulo, SP, 05508-050, Brazil
³Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, SP, 13418-900, Brazil
⁴Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, SP, 13566-590, Brazil
⁵Department of Geosciences, Pennsylvania State University, State College, PA 16801, USA
⁶National Center for Monitoring and Early Warning of Natural Disasters (Cemaden), São José dos Campos, SP, 12247-016, Brazil
⁷Graduate Program in Natural Disasters, São Paulo State University (UNESP)/Cemaden, São José dos Campos, SP, 12245-000, Brazil
⁸Graduate School of International Studies, Korea University, Seoul, South Korea

Correspondence: Marcos Roberto Benso (marcosbenso@gmail.com)

Received: 13 December 2023 – Discussion started: 21 February 2024 Revised: 28 November 2024 – Accepted: 10 February 2025 – Published: 10 April 2025

Abstract. Understanding how physical climate-related hazards affect food production requires transforming climate data into relevant information for regional risk assessment. Data-driven methods can bridge this gap; however, more development must be done to create interpretable models, emphasizing regions lacking data availability. The main objective of this article was to evaluate the impact of climate risks on food security. We adopted the climatic impact driver (CID) approach proposed by Working Group I (WGI) in the Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC). In this study, we applied the CID framework using a random forest model in a bootstrapping experiment to identify the most influential indices driving crop yield losses. We also used SHapley Additive exPlanations (SHAP) with the random forest model for explanatory analysis, enabling us to pinpoint critical thresholds for these indices-thresholds that, when exceeded, significantly increase the probability of impact. Additionally, we investigated the effects of two CID types (heat and cold and wet and dry) represented by categories of climate extreme indices on crop yields, with a particular focus on maize and soybeans in key agricultural municipalities in Brazil. We

found that mean precipitation is a highly relevant CID. However, there is a window in which crops are more vulnerable to a precipitation deficit. In many regions of Brazil, for example, soybeans face an increased risk of yield losses when precipitation falls below 100 mm per month in December, January and February – marking the end of the growing season in those areas. Nevertheless, including climate means remains highly relevant and recommended for studying the impact of climate risk on agriculture. Our findings contribute to a growing body of knowledge critical for informed decisionmaking, policy development and adaptive strategies in response to climate change and its impact on agriculture.

1 Introduction

Climate extremes, such as heat waves, droughts, floods and excessive precipitation, play a critical role in determining crop yield shortfalls (Vogel et al., 2019). Empirical evidence from multiple studies shows that models incorporating data from several weather variables are more accurate in explaining crop production variability than those relying solely on precipitation (Proctor et al., 2022; Ray et al., 2015) or single weather variables. Consequently, it is essential to assess agricultural production risk through extreme climate indices, which are key to monitoring hazards that impact food production and food security (Das et al., 2022; Schyns et al., 2015).

These natural hazards affecting society are often referred to as "impact drivers". As Ruane et al. (2022) argues, understanding impact drivers requires knowledge of the vulnerability and exposure of specific sectors. Sectoral information helps determine the magnitude of a driver's effect, which can be either beneficial or detrimental to its activities. This requires a co-creation process that aims to contextualize climate information for decision-making. This concept is embodied in the climatic impact driver (CID) framework, introduced by Working Group I (WGI) of the Intergovernmental Panel on Climate Change (IPCC) in its Sixth Assessment Report (AR6). The CID framework is still in its early development stages and aligns with the United Nations Sendai Framework for Disaster Risk Reduction (UNISDR) 2015–2030 (UNDRR, 2025), following the UNISDR hazard list definitions. However, the CID framework goes further by recognizing climate change as a significant hazard, which is not included in the UNISDR hazard list.

The formal definition of CIDs encompasses the physical conditions of the climate, including means, extremes and events. The CID framework emphasizes two critical aspects of risk assessment: defining indices for climatic impact drivers and identifying thresholds for climatic impact drivers (Ruane et al., 2022). These aspects reinforce the need to develop numerically computable indices that utilize one or a combination of climate variables to quantify the intensity and frequency of a CID. When these indices surpass certain thresholds, the risk of losses and damage increases. Although many indices have been used in the literature and summarized by Ranasinghe et al. (2021) for their relevance to the agricultural sector, it remains necessary to tailor approaches based on regional characteristics that bridge global insights with local solutions.

One way to understand the impact of climate change on agricultural production is through machine learning (ML) algorithms. ML algorithms can improve our understanding of how climate affects crop yields (Sidhu et al., 2023). Drawing on statistical learning theory (Vapnik, 1999), these algorithms can generalize patterns and make predictions from the available data. Several authors have applied ML algorithms to predict crop yields (Vogel et al., 2019; Sidhu et al., 2023; Han et al., 2019; Schierhorn et al., 2021; Silva Fuzzo et al., 2020).

Decision tree algorithms, such as random forest (RF) models, have been particularly effective in understanding the impact of weather extremes on crop yield variability (Vogel et al., 2019; Jeong et al., 2016; Schierhorn et al., 2021). The RF model combines tree predictors that are recursively split and used for predictions (Breiman, 2001). It provides insights into the importance of each feature in the model's overall performance. Studies by Vogel et al. (2019) and Schierhorn et al. (2021) have used this approach to explore the influence of extreme temperature and precipitation on predictive RF models for soybeans and maize. Both studies found that mean climate variables during growing seasons are the most relevant features for predicting crop yields. However, extreme weather indices, particularly those related to droughts and temperature extremes, can explain 18 %–43 % of crop yield variability (Vogel et al., 2019).

Despite efforts to improve the performance and interpretability of ML models, previous studies have relied on a somewhat limited selection of indices. As a result, other factors influencing crop yield variability may remain hidden and the underlying mechanisms of crop yield losses due to weather extremes could be overlooked. A generalized framework for variable and feature selection has been developed to enhance ML model performance, provide faster models and improve understanding of the underlying processes that generated the data (Guyon and Elisseeff, 2003). The backward recursive feature elimination (RFE) method, presented by Svetnik et al. (2004), leverages the RF algorithm's ability to generate variable importance as a variable reduction wrapper algorithm. Its applicability has been demonstrated in various studies, including those focused on crop selection (Wang and Li, 2023) and hyperspectral imaging for monitoring pasture quality (Pullanagari et al., 2018).

Although eliminating redundant features and variables can enhance our understanding of data structure, the challenge of interpretability remains. We propose using the modelagnostic explanation method introduced by Lundberg and Lee (2017), known as SHAP (SHapley Additive exPlanations). There are promising studies applying SHAP to environmental data (Wikle et al., 2023; Viana et al., 2021), with implications for soil moisture and evapotranspiration determination. The use of post hoc explanation algorithms for crop yields has been explored by Mariadass et al. (2022). However, to our knowledge, this method has not been specifically applied to predicting the impacts of extreme weather on food production.

In this context, we introduce a comprehensive modeling framework to enhance the interpretability of tree-based models that utilize climate data to predict crop yield losses. With the goal of assessing the impact of climate extremes on food production, this research employs the CID framework developed by IPCC Working Group I (IPCC WGI). The framework allows us to characterize climate extremes by creating numerically computable indices and determining relevant thresholds. The significance of this framework lies in its ability to provide a basis for incorporating climate information into studies, decision-making processes and policy development. By applying this framework to our research, we aim to provide valuable insights that can inform critical decisions, policies and strategies related to food production in the face of climate extremes.



Figure 1. Flowchart illustrating the methodology proposed for analyzing the impact of climate indices on crop yield.

2 Methodology

2.1 Modeling framework

We present a framework for investigating the impacts of weather and climate extremes on crop yields using ML, focusing on building a reproducible workflow, selecting features and producing explainable ML model outputs. A good feature of an ML algorithm is its *relevant* nature in explaining the target variable (in this study, crop yields) based on input variables. However, it should not be *redundant* regarding any other relevant predictor (Yu and Liu, 2003). In addition to these concepts widely applied in ML methods, we add the concepts of *explainable* and *operational* features. In ML, a feature is any variable that is used as an input variable for prediction. Therefore, this work will use the terms feature and variable interchangeably.

This framework (Fig. 1) consists of three steps. The first is data filtering; in this step, we removed highly correlated features (Pearson correlation greater than 0.9). Feature selection is a preprocessing step in ML models. The filtering process removes redundant features. In this study, the concepts of explainable and operational features are the motivation for our proposed methodology. We aim to achieve a balance between model performance, interpretability and practical applications. By focusing on explainable features, our objective is to create models that offer clear insights into decision-making processes, thereby promoting transparency and reliability. This interpretability is essential for stakeholders who must comprehend and validate the model's results.

The second step aims to select the most important variables. We use the abilities of the RF to generate the importance of variables to rank the most important variables. The third step is to define variable thresholds. To do that, we must apply another machine learning model to explain the first one.

The SHapley Additive exPlanations (SHAP) explanatory analysis is an explanation algorithm proposed by Štrumbelj and Kononenko (2014) and uses game theory to provide an efficient explanation of the predictions made by a ML algorithm. The SHAP method is used to explain how each variable was used to make each prediction.

In the second step, we focus on identifying the most important CIDs to assess their impact on food production. Different models are independently trained for each state being analyzed; a separate and unique machine learning model is developed and trained using data specific to that state. This implies that the analysis of CIDs on food production is customized to account for the unique characteristics, data and conditions present in each state, rather than being applied a single model uniformly across all states. The feature importance determined based on entropy is determined by calculating the reduction in entropy (information gain) each feature The creation of training, validation and testing subsets is crucial to avoiding overfitting and achieving reasonable estimates of model performance. The dataset was divided according to the chronological order of the data. The first 80% of the data, according to the timeline, were used to train the model, allowing it to learn and adjust its parameters. The remaining 20% were used for validation, meaning this portion was reserved for testing the model's predictions on data it had not seen during training. This approach, which incorporates a temporal aspect, is intended to simulate a real-world scenario where future data should be predicted. This method helps prevent overfitting by ensuring that the performance of the model is evaluated on new unseen data that come after the training period used, thus providing a realistic assessment of how the model will perform in practice.

To avoid temporal dependencies between data points from neighboring municipalities from being correlated within the same year, the best-fit model was selected using a leaveone-year-out cross-validation (LOYOCV) method and hyperparameters were chosen according to the mean CV performance in folds, following the recommendations of von Bloh et al. (2023). A fixed 10-year window was used for training, followed by 1 year as a test set. This process was repeated iteratively, leaving each year as the test set, while using the preceding 10 years for training. Performance metrics were calculated for each iteration, and scores were averaged to obtain an overall assessment of the performance of the model. The models were trained and optimized on the training dataset, and their performance was evaluated on the validation data to test the robustness of the models.

In the second step, we identify the most critical CIDs to assess their impact on food production. To achieve this, we used the RF model. Models were trained with three different crop yield datasets and different combinations of features, including precipitation means, temperature means, and combinations of means and extreme climate indices. The goal of this experiment was to identify the most important climate indices.

The third step builds on the results of the second step and uses the 10 most relevant climate indices employing RF explainability with the SHapley Additive exPlanations (SHAP) explanatory analysis. This approach aimed to provide a detailed understanding of how the model used these crucial indices and attempted to identify significant thresholds for these influential climate variables. The RF models were implemented using the R package ranger (Wright and Ziegler, 2017), and the SHAP results were implemented using the R package shapviz (Mayer, 2023).

The SHAP approach is based on explaining how each feature of the model was used to make a single prediction. The first step is to define a base prediction, which in this case is the expected value of the set X, which is the value of the adjusted and detrended crop yields E[f(X)]. Then, the RF algorithm is used to make a prediction f(x) for a single value of crop yield in a specific municipality and a particular year. The difference between the expected value and the prediction is called the SHAP value and preserves the unit of crop yields; therefore, here, it will be in metric tons per hectare.

In order to identify how each feature was used to generate the SHAP value, the algorithm recursively adds each feature and tests the importance of the feature for that prediction. However, the order in which the feature is added to the model is essential; that is why the principles of game theory introduced by Lloyd Stowell Shapley were used to solve the problem of allocating each feature's order and extracting its importance for the prediction. More details of the game theory used in SHAP results can be found in Strumbelj and Kononenko (2010).

The SHAP result was performed for each prediction using the R package treeshap (Komisarczyk et al., 2024). The package allows for generating a partial dependence plot, which is the relationship between the feature value and the contribution to the SHAP value. From this approach, we can find out which feature values are critical for crop yield losses, establish thresholds and contribute to the CID framework. Since the order of each feature is also evaluated, a second analysis is performed, which is used to evaluate the interaction between different features. In treeshap, the interaction between variables is determined by assessing the shared contribution of a pair of features to a model's prediction, beyond their separate effects. Initially, the SHAP values for each feature are computed to represent their individual impact on the model's output. For example, the interaction of features i and j is determined by subtracting the sum of their individual SHAP values from their combined contribution, i.e., interaction = $SHAP_{i, j} - (SHAP_i + SHAP_j)$. This method captures how the joint presence of two features affects the prediction compared to their independent contributions, uncovering synergistic or antagonistic interactions between the features.

2.2 Study area

Brazil is a significant producer of agricultural goods, as reported by the Food and Agriculture Organization (FAO) (FAO, 2025). The country is responsible for more than 10 % of the world's maize and more than 30 % of the global soybean production. Brazil is one of the four leading agricultural producers in the world, along with China, India and the United States, with a cultivated area of soybeans and maize of 58×10^6 ha. In Fig. 2, we show the delimitation of the study area. The map shows 452 selected municipalities that encompass the states of Rio Grande do Sul (RS), Santa Catarina (SC), Paraná (PR), São Paulo (SP), Mato Grosso do Sul (MS) and Minas Gerais (MG). The selection criteria will be explained in the subsection "Crop yield data". The map



Figure 2. Location of selected study municipalities with respect to observed cropland extent between from 2003 to 2019.

shows the percentage of cropland derived from the work of Potapov et al. (2022).

The growing season in the study area was defined using a global crop calendar for the second season of soybeans and maize determined by data from Sacks et al. (2010). We consider the planting dates to follow a normal distribution, with the mean date being the most probable date for farmers and the maximum and minimum dates being considered to be twice the standard deviation. For soybeans, sowing dates start in the middle of austral spring in October, peak in November and end in December. The harvest begins in late summer and extends to fall, from February to March. Since the second season of maize peaks in February, we consider the end of soybeans to be in February. For the second season of maize, planting begins at the end of January, after soybean harvest; peaks in February; and ends in the beginning of April. The harvest starts in June, peaks in August and ends in October.

2.3 Data collection and processing

In this section, we present the description of datasets used to analyze the impact of climate variables on soybean and maize crops in the state of Paraná. We used two criteria to select a dataset: (i) the data must comply with FAIR principles (i.e., data must be findable, accessible, interoperable, and reusable), and (ii) climate data must be updated frequently (ideally, with a minimum of daily update frequency).

We used three different datasets: (i) the statistical yearbooks of the state of Paraná (Parana, 2021), (ii) the Municipal Agricultural Production Survey by the Brazilian Institute of Geography and Statistics (IBGE) (de Geografia e Estatística, 2022) and (iii) the Global Dataset of Historical Yields (GDHY) (Iizumi and Sakai, 2020). For climate analysis, we used data from the fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5-Land reanalysis dataset (Muñoz-Sabater et al., 2021). We summarized the main characteristics of each dataset used in Fig. 1.

2.3.1 Crop yield data

Crop yield data are a vital component in understanding the impacts of climate on food production. Crop yields are generally made available at the municipal level. We used three datasets to analyze crop yields in Brazil. We collected crop yield data from the Brazilian Institute of Geography and Statistics (IBGE) at the municipal level (de Geografia e Estatística, 2022). In the study area, the double-cropping system is widely adopted. It, therefore, represents a potential bottleneck because the IBGE data from 1974 to 2022 are an annual aggregation of the total production of that crop within the municipality.

However, the Brazilian Institute of Geography and Statistics has been collecting maize in the first and second season since 2003. This matches the period during which the second maize season intensifies in Brazil. Data at the municipal level were filtered based on data availability. The missing years were removed from the dataset, and the municipalities with more than 2 years of missing data were disregarded. The selection resulted in 452 municipalities for soybeans that comprise the states of Rio Grande do Sul (RS), Santa Catarina (SC), Paraná (PR), São Paulo (SP), Mato Grosso do Sul (MS), Minas Gerais (MG) and Goiás (GO) and 216 municipalities for maize in the second season for Paraná (PR), São Paulo (SP), Mato Grosso do Sul (MS), Minas Gerais (MG) and Goiás (GO).

The Department of Rural Economy (Departamento de Economia Rural, Deral) of the state of Paraná, Brazil, is also responsible for collecting crop data at the municipal level. The method of collecting and processing data is similar to what is done by the Brazilian Institute of Geography and Statistics; therefore, a high level of redundancy is expected from these two datasets. This redundancy is necessary to validate data and remove outliers that might reduce the quality of a model. The same number of municipalities selected using IBGE data was used in data from Deral.

The Global Dataset of Historical Yields is a global annual time series of 0.5° grid-cell estimates for maize, rice, wheat and soybeans from 1981 to 2016. For each grid cell, crop yields are estimated in tha⁻¹ based on Food and Agriculture Organization (FAO) country-level yield statistics and then corrected using the remotely sensed leaf area index (LAI), the fraction of photosynthetically active radiation (FPAR) and crop-specific radiation use efficiency derived from reanalysis. Crop areas and crop calendars were derived from Sacks et al. (2010). More details about the dataset are described in Iizumi and Sakai (2020). The dataset was aggregated to the municipal level using zonal statistics in the terra package (Hijmans, 2023) in R Studio.

For statistical analysis, we removed the outliers of all crop yield datasets considering, for each year, neighboring munic-

Dataset	Variable	Spatial resolution	Temporal resolution	Time frame
Crop yields – IBGE (IBGE, 2023)	Soybeans	Municipal level	Annual	1974-2022
Crop yields – IBGE (IBGE, 2023)	Maize	Municipal level	Crop season	2003-2022
Crop yields – Paraná (Parana, 2021)	Soybeans, maize	Municipal level	Crop season	1997-2021
Crop yields – GDHY (Iizumi and Sakai, 2020)	Soybeans, maize	0.5° cell	Crop season	1981-2016
ERA5-Land (Muñoz-Sabater et al., 2021)	Precipitation, temperature	0.1° cell	Daily	1950-present

Table 1. Description of the datasets used in the case study.

ipalities using the interquartile range (IQR). For the outlier removal process, we defined "immediate regions" as clusters of municipalities geographically proximate to one another, as classified by the Brazilian Institute of Geography and Statistics. Crop yields within these regions exhibited a high degree of correlation, which was verified using the correlation index. To identify outliers, we applied the interquartile range (IQR) method for each year. Specifically, if the yield of a municipality in a given year deviated significantly from the yields of other municipalities in the same immediate region, it was classified as an outlier and excluded from the dataset. This approach ensured that only extreme and anomalous data points, not reflective of regional trends, were removed.

Changes in technology in seed production, fertilizers and land management, also known as technological trends (Liu and Ker, 2020), and other sources of trends such as climate change were removed by local polynomial regression fitting (LOESS) (Cleveland et al., 2017). Moreover, systematic changes in crop yields have also been associated with heteroskedasticity (Yang et al., 1992; Zhu et al., 2011; Ozaki et al., 2008). The residuals of the LOESS model were tested for heteroskedasticity. If heteroskedasticity was proved, it was removed using the method proposed by Ozaki et al. (2008). For further information on the preprocessing of crop yield data, please consult the Supplement.

2.3.2 ERA5-Land reanalysis dataset

Weather data were sourced from the fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5-Land reanalysis dataset (Muñoz-Sabater et al., 2021; Hersbach et al., 2020). The dataset has a $0.1^{\circ} \times 0.1^{\circ}$ latitude–longitude grid and was aggregated to the municipal area to match the spatial discretization of soybean crop yield (SBY). The data collection spanned 1980 to 2023, with daily observations. Weather variables included precipitation, maximum temperature and minimum temperature. We used different climate indices to evaluate multi-hazard risks. Since mean climate conditions of precipitation and temperature are the most relevant (Moriondo et al., 2011), we considered monthly precipitation, maximum and minimum temperatures, total precipitation, and mean temperature over growing seasons.

2.3.3 Indices for climatic impact drivers

The IPCC WGI has presented CIDs as a new approach to assessing climate data to analyze their effects on society. CIDs are represented by numerically computable indices and categorized into several types. In this paper, we considered wet and dry and hot and cold CIDs. To calculate the indices, we first considered the indices indicated by the Expert Team on Climate Change Detection and Indices (ETCCDI), which is supported by the World Meteorological Organization (WMO) Commission for Climatology, the Joint Commission for Oceanography and Marine Meteorology (JCOMM), and the research program on Climate Variability and Predictability (CLIVAR) Frich et al. (2002). A summary of the indices used according to the type of CID is shown in Table 2.

We also considered two drought-related indices, the Standardized Precipitation Index (SPI) (McKee et al., 1995) and the Standardized Precipitation and Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010). The SPI is based on the probability of monthly precipitation on different timescales, and it is recommended to be calculated with a time series of at least 30 years. The monthly time series must be fitted to a cumulative distribution function (CDF). We adopted the gamma distribution.

Then, the data are transformed to the standard normal distribution to calculate the SPI, a standardized value subtracting the transformed precipitation from the mean value and dividing by the standard deviation. The SPI can be calculated using different timescales representing previous meteorological conditions, typically 1 to 48 months. For agricultural applications, the 3-month SPI is most frequently used Kim et al. (2019).

The SPEI is a more recent index that incorporates temperature in the calculation of the SPI. A new step was added to the procedure, calculating the monthly potential evapotranspiration (PET) and the SPI using the same procedure described previously with the value of monthly precipitation minus monthly PET. PET was calculated using the Hargreaves method, which is calculated using maximum and minimum temperature and extraterrestrial radiation (RA) (Droogers and Allen, 2002).

The primary motivation for using distinct indices derived from the same fundamental data is to identify which features of the extremes are the most significant. Is it the magnitude of

M. R. Benso et al.: Data-driven assessment of climatic impact drivers

CID type	CID category	CID index abbreviation	CID index description	
Hot and M cold te	Mean air	temp	Monthly temperature mean	
	temperature	dtr	Daily temperature range: monthly mean difference between maximum and minimum daily temperature	
	Extreme heat	tx90p	Monthly percentage of days when maximum daily temperature is higher than the 90th percentile	
		tn90p	Monthly percentage of days when minimum daily temperature is higher than the 90th percentile	
		su	Number of summer days: monthly number of days when maximum daily temperature is higher than 25 °C	
_		tr	Number of tropical nights: monthly number of days when minimum daily temperature is higher than 20 °C	
		txx	Monthly maximum value of daily maximum temperature	
		tnx	Monthly maximum value of daily minimum temperature	
	Cold spell	tx10p	Monthly percentage of days when maximum daily temperature is lower than the 10th percentile	
		tn10p	Monthly percentage of days when minimum daily temperature is lower than the 10th percentile	
		tnn	Monthly minimum value of daily minimum temperature	
		txn	Monthly minimum value of daily maximum temperature	
	Mean precipitation	preptot	Monthly precipitation sum	
Wet and dry	Heavy precipitation	r10mm, r20mm	Monthly count of days when daily precipitation is higher than 10 and 20 mm	
		rx1day, rx5day	Monthly maximum 1 and 5 d precipitation	
	Agricultural and ecological drought	spei_3month, spei_6month	Standardized precipitation and evapotranspiration index for 3- and 6-month accumulations	
		spi_3month, spi_6month	Standardized precipitation index for 3- and 6-month accumulations	

Table 2. Description of the climatic impact drivers (CIDs) considered in this study and their respective indices.

an extreme, the length of time it lasts, or values that are either above or below a certain threshold? Research such as that by Vogel et al. (2019) has demonstrated the importance of extreme events in understanding the variability in crop yields. We summarize all the indices according to CID type and category in Table 2.

3 Results and discussion

The comparison of the datasets used in this study is important for evaluating the reliability of the data. High-quality crop yield data improve the calibration of crop growth models (Rosenzweig et al., 2014). However, they have a broader application in geosciences. Crop yield data are used to parameterize hydrological models in watersheds, especially in agricultural catchments, and improve soil moisture simulation (Sinnathamby et al., 2017).

We compared crop yields at the municipal level in Brazil. We observed that the IBGE and Paraná Deral data for soybeans and maize are highly correlated; however, outliers were detected in both datasets. The outlier removal process improved the agreement between the two datasets, suggesting that eliminating data improved the dataset's quality. Since Deral is only available in Paraná, for the other states of Brazil, the GDHY and IBGE data were compared. The Global Dataset of Historical Yields aggregated at the municipal level has a weak association with the other datasets. This result confirms what was reported by Iizumi and Sakai (2020). The GDHY data are based on satellite data collected from a fixed cropland map. In many regions of Brazil, there is a noticeable increase in croplands, which can influence the

Table 3. Correlation coefficients of three different crop yield datasets: (a) IBGE (n = 3845), Deral (n = 3120) and GDHY (n = 15411) for maize and (b) IBGE (n = 20629), Deral (n = 3432) and GDHY (n = 15406) for soybeans. Values represent the strength and direction of the relationships, with an asterisk (*) indicating statistically significant correlations of p < 0.001.

(a)	Maize – IBGE	Maize – Deral	Maize – GDHY
Maize – IBGE	1.000	0.910*	0.474*
Maize – Deral	0.910*	1.000	0.596*
Maize – GDHY	0.474*	0.596*	1.000
(b)	Soy – IBGE	Soy – Deral	Soy – GDHY
Soy – IBGE	1.000	0.968*	0.434*
Soy – Deral	0.968*	1.000	0.403*
Soy – GDHY	0.434*	0.403*	1.000

estimation of GDHY data. In addition, the exact location of the planted area within each municipality can vary from year to year.

3.1 Identifying key climate impact drivers

We tested a variety of indices that measure mean precipitation, mean temperature and extremes. We initially tested the ML technique using various inputs to demonstrate its ability to illustrate crop yield variability in the states examined. In Fig. 3, we demonstrate the model performance of the RF model considering different datasets. Taking the coefficient of determination, the climate variables explained the variability in soybean crop yields, on average, from 30 % to 40 % for IBGE, 25 % to 45 % for GDHY and 30 % to 50 % for Deral. The climate was explained for maize from 12 % to 15 % for IBGE and Deral and from 10 % to 45 % for GDHY.

The coefficient of determination quantified the proportion of the variance in the crop yield data (dependent or target variable) that the RF model can explain. The results are consistent with the values found in other similar studies; Ray et al. (2015) used municipal-level data to quantify the impact of climate variability on yields using regression models and determined that, in Brazil, climate variability explains 26 %– 34 % of soybean yields and 41 % of maize yields. Vogel et al. (2019) used the same dataset as Ray et al. (2015) considering South America and applied an RF model defining the values of 28 % for soybeans and 25 % for maize. It is important to note that none of these studies separated maize in the first and second season.

The model performance was generally higher for GDHY than for the other datasets for soybeans in the southern states (RS, SC and PR). The models based on precipitation means and the combination with temperature and extremes explain the variability in crop yield more than only the temperature means, which was observed in all three datasets. For maize in the second season, for MS, MG and GO, the models that combine mean temperature to mean precipitation and extremes explained the variability in crop yields more than only temperature, which was observed in GDHY and IBGE. The IBGE database showed that the maize model was much more effective in São Paulo than in any other state.

The RF models used in this study helped obtain the most relevant variables, and these variables were classified into CID types, i.e., wet and dry and hot and cold, and subcategories, as described in Table 2. The model demonstrates the importance of climate variables in explaining the variability in crop yields and allowed us to determine the critical CIDs considering each state and dataset. The following sections summarize the key CID categories for soybeans and maize.

Feature importance was summarized spatially and temporally. Figure 4 highlights similarities and dissimilarities regarding feature importance considering the different datasets.

In the Supplement, we show tables with the results of variable importance for all models. The analysis of variable importance for soybean datasets is shown in Table S1 in the Supplement. The analysis identifies extreme heat (tnx), the drought index (SPEI) and precipitation totals (preptot) as key variables that affect crop yields in different regions. These climate factors are crucial for predicting agricultural outcomes, with extreme heat and drought having a great impact on the results. The significance of these variables varies by region; for example, extreme heat and drought are critical in Paraná during February, while mean air temperature and extreme cold in January matter more in Minas Gerais. February and January are highlighted as pivotal months due to their association with significant climate events. In general, the results highlight the importance of addressing both heat and water stress in agricultural systems while taking into account spatial and temporal differences to enhance predictive accuracy.

For maize in the second season, the results are shown in Table S2. For Paraná (PR), key variables are April precipitation and May heat, which influence agricultural outcomes. Goiás (GO) is significantly affected by the April diurnal temperature range and the heat and precipitation of May. In Minas Gerais (MG), May precipitation is crucial, with the March temperatures and August heat also significant.

Mato Grosso do Sul (MS) deals with February temperatures and June heat and drought. Rio Grande do Sul (RS) sees the diurnal temperature range of May as vital, along with the March drought and the July–August precipitation. São Paulo (SP) contends with the heat of August and the rainfall of July, emphasizing the importance of temperature and precipitation on environmental and agricultural concerns.

3.1.1 Wet and dry

Changes in mean precipitation pose a threat to agricultural production. A precipitation deficit leads to the reduced availability of soil moisture, affecting plant development and reducing crop yields, and it is considered the most critical en-



Figure 3. Performance evaluation of regression models for (a) soybeans and (b) maize. Data from the Department of Rural Economy (Deral) with data only for Paraná, Global Dataset of Historical Yields (GDHY), and Brazilian Institute of Geography and Statistics (IBGE) for major historically agricultural municipalities. MAE: mean absolute error, RMSE: root mean square error.

vironmental factor that reduces crop yields (Bray, 2007). In our analysis, mean precipitation was one of the most important climatic impact drivers of soybean crop yields during January and February for RS, SC, PR and MS and also in December for PR.

The state of Rio Grande do Sul has historically been affected by El Niño–Southern Oscillation (ENSO), with a more substantial influence from November to May (Gelcer et al., 2013), which is responsible for droughts and the impact on soybean crop yields during La Niña. The model did not indicate that mean precipitation was the most important factor for the state of SP. For maize production, mean precipitations during April and May were considered important for PR, MS, MG and GO and not for SP.

Agricultural systems require minimum rainfall, or they rely on irrigation. In Brazil, the states of SP, MS, MG and GO have a well-defined difference between wet and dry seasons. Usually, the wet season starts in October and ends in May, and the soybean-maize double-cropping system depends on the length of the wet season in the states mentioned above.

Agricultural and ecological drought indices are directly related to a precipitation deficit and excessive temperature (Sarhadi et al., 2018; Lesk et al., 2021), which affect the ability of plants to grow and reduce plant transpiration. The duration and timing of droughts play a significant role. We observed that droughts occurring in January and February during the rainy season were the most important for soybeans. The droughts in February and March also affected the second-season crop yields of maize. Droughts that occurred at the end of the maize growing season also affected crop yields.

In this study, we considered climate extreme indices on different timescales. As we added the temporal dimension to the analysis, we revealed that a 3-month SPEI in October in the state of RS was selected for the list of the most relevant variables. This indicates that pre-sowing meteorological factors that can reduce soil moisture conditions also influence crop yields. This result corroborates the findings of Santini et al. (2022), which revealed that a drought analysis should not neglect antecedent conditions since it influences factors such as soil workability and crop development.

3.1.2 Hot and cold

The mean air temperature influences many aspects of crop cultivation. In RS and SC, the soybean growing season starts



Figure 4. Importance of key CID categories in predicting soybean and maize crop yields. The figure displays the most significant features identified by the random forest model for soybeans and maize.

when the mean temperatures exceed the minimum temperature thresholds for soybeans (Battisti and Sentelhas, 2014). As the temperature increases, the development (phenology) of the plant is affected and increased thermal stress is expected Lesk et al. (2021). Except for RS, mean temperatures during all soybean growing seasons were considered important variables. The same behavior was observed for maize; mean temperatures were considered significant during all growing seasons. Exposure to temperatures above a specific limit or threshold can lead to lower yields. The value of these thresholds depends on the crop species and farm management. For soybeans, extreme temperature indices affected crop yields throughout the growing season, especially in January and February.



Figure 5. SHAP waterfall plot visualizing the key CID contributions to crop yield losses for the state of Paraná (PR) in 2019, a drought year, with metrics for monthly maximum value of daily minimum temperature (tnx), minimum value of daily minimum temperature (tnn), maximum value of daily maximum temperature (txn), percentage of days when tx > 90th percentile, standardized precipitation evapotranspiration index (SPEI), total precipitation (prcptot) and daily temperature range (dtr).

3.2 Determining thresholds and their significance

With the results of the selection of critical CIDs, we improved the understanding of the impacts on climate variables that significantly influence crop yield losses, considering different types of indices and critical periods. The insights obtained from the combination of random forest models applied to different datasets facilitate a robust understanding of climate–crop interactions and make it possible to compare what results the datasets have in common, increasing the results' reliability. However, the random forest model did not provide information on the values of each variable that are important and can help us define the threshold values of these indicators that are associated with an increased risk of crop yield losses.

To improve our understanding of how each climate extreme indicator was used for prediction, we used SHAP. This technique allowed us to extract insights from the results of a random forest model, thus providing a comprehensive perspective of the drivers of crop yield fluctuations. The results of SHAP-derived explanations revealed a clear pattern concerning the most influential variable that affects soybean yields. We highlight critical loss events by evaluating the model prediction for a particular city in a given year.

In 2019, an important widespread drought event was observed in Brazil and was considered a mega-drought that affected many regions of Brazil, especially the Paraná River basin Marengo et al. (2021). This drought extended into 2022 and occurred during a La Niña year. We highlight the model explanation for the state of Paraná considering two important agricultural municipalities, namely Marechal Cândido Rondon and Uniflor, as shown in Fig. 5.

The two datasets presented an agreement regarding crop yields below the expected value E[X]. However, they varied in terms of the magnitude of these losses. For the municipality of Marechal Cândido Rondon, the Deral yields are similar to the IBGE yields, 1.72 and 1.77 (Fig. 5), respectively. The predictions made by the two models were similar, and the main variables also performed similarly. High temperatures, represented by the maximum value of the daily minimum temperature in February, and low precipitation in December were the main drivers of losses combined with the 3-month SPEI in February.

For Deral, accumulated precipitation in December was also a significant driver of losses, and for IBGE, the 3-month



Figure 6. SHAP waterfall plot visualizing the key CID contributions to crop yield losses for the state of São Paulo (SP) in 2015, a drought year, with metrics for monthly maximum value of daily minimum temperature (tnx), minimum value of daily minimum temperature (tnn), maximum value of daily maximum temperature (txn), percentage of days when tx > 90th percentile, standardized precipitation evapotranspiration index (SPEI), total precipitation (prcptot) and daily temperature range (dtr).

SPEI was one in January. The other variables represented a negligible influence on crop yield losses. For the city of Uniflor, the actual Deral and IBGE yields were similar, 2.44 and 2.48 (Fig. 5), respectively. The main influences on crop yield losses were precipitation in December (prcptot_Dec) and high temperatures (tnx_Feb). The 3-month SPEI values in January and February can be considered redundant. Standardized indices refer to previous conditions; therefore, the values overlap in 2 months (December and January).

In 2014/15, a severe drought occurred in southeastern Brazil, causing an unprecedented water supply shortage in the Cantareira Water Supply System and affecting many cities in São Paulo (Deusdará-Leal et al., 2019). The drought had repercussions in many regions of the state of São Paulo. Therefore, we compared the IBGE and GDHY results for two municipalities in the state of São Paulo, Araçatuba and Sandovalina. The expected values of GDHY were lower than those of IBGE, and the two datasets diverged in yields below the expected value, i.e., IBGE indicated losses, and GDHY did not. According to a report by the Brazilian Ministry of Agriculture, Livestock and Food Supply (MAPA), the state of São Paulo was one of the most affected by losses in the agricultural year 2014/15 (MAPA, 2022). This result suggests that, although it has been suggested that GDHY is recommended in data-scarce regions (Iizumi and Sakai, 2020), using this dataset requires caution.

The SHAP methodology analyzes each prediction and shows how each variable was used in the model. This helps us create a partial dependence plot, which relates the variable's value with the impact in terms of crop yield losses represented by the SHAP value. This analysis is illustrated in Fig. 7, which shows partial dependence plots for the state of Rio Grande do Sul.

The comparison of the two datasets can be found in Fig. 7. The IBGE dataset shows that the 3-month SPEI in February can influence crop yield losses and has an upper and lower threshold. The lower threshold is -1.0. Values below this can represent losses of up to 0.2 tha^{-1} . Generally, 3-month SPEI values below -1.0 are considered critical and are used as a reference for the severity of the drought (Chiang et al., 2021). Extreme wet conditions, with a threshold of 1, also affect crop yields in the RS state.

Excess rainfall can have the same impact as droughts (Li et al., 2019). However, little attention has been paid to this analysis. Our results suggest excessive precipitation can be responsible for up to 0.2 tha^{-1} of losses. More studies on



Figure 7. A comparison of the key CIDs derived from the Brazilian Institute of Geography and Statistics (IBGE) from 2013 to 2021 and the Global Dataset of Historical Yields (GDHY) from 2009 to 2016. Annual data aggregated at the municipal level were used to create a dependence plot for the soybean explanation model for validation data in RS. The data spanned 2013 to 2021 for IBGE and 2008 to 2016 for GDHY. The 3-month SPEI in February (spei_3month_Feb) for (**a**) IBGE and (**b**) GDHY, precipitation accumulated in February (prcptot_Feb) for (**c**) IBGE and (**d**) GDHY, and precipitation accumulated in December (prcptot_Dec) for (**e**) IBGE and (**f**) GDHY.

this type of hazard are recommended. The cumulative precipitation in February presented a threshold of around 100 mm and a potential to cause losses of approximately 0.4 tha^{-1} . The cumulative precipitation in December was the only indicator with a similar result in the IBGE and GDHY datasets. Regarding potential losses, both agree on a value of up to 0.6 tha^{-1} ; however, the threshold for IBGE is 150 mm, and for GDHY, it is 120 mm.

The patterns of crop yield losses observed in the region raise two main concerns. The first is that the severe crop yield losses presented in the previous examples have happened only once in the entire time series, representing an imbalance in the values of the dataset. One implication of this situation is that models may not have sufficient cases of severe failure to be trained adequately and may underestimate losses. The second concern is related to the decision to use these anomalous events. Possible solutions include using it for training, testing or removing it from the dataset. We opted to keep these events in the analysis with the warning that this might interfere with model performance. However,



Figure 8. The 2D partial dependence derived from the Deral and IBGE data for the state of Paraná from 2016 to 2021 of the key climatic impact driver total precipitation in December (prcptot_Dec) with the most correlated CIDs and their combined impact on SHAP values (tha^{-1}): considering the 3-month SPEI in February (spei_3month_Feb) and total precipitation in December (prcptot_Dec) and the maximum value of daily minimum temperature (tnn_Feb).

we wanted to evaluate the ability of the model to predict unprecedented loss events.

3.3 Evaluating combined hazards

The SHAP algorithm also allowed us to investigate the compound effect of climate indicators. In Fig. 8, we present the detection of compound event effects, considering the most important variable in the state of PR (prcptot_Dec) with four other considered important variables. We observed hot and dry compound events, characterized by high temperatures and a precipitation deficit, which have been cited as an increasing threat to food production (Zscheischler et al., 2018; Hamed et al., 2021).

This analysis also showed that precipitation in December was closely related to a 3-month SPEI in February. As discussed previously, this is expected since the SPEI considers previous conditions. However, it is essential to note that December is a critical month for droughts in PR and other states, such as Rio Grande do Sul (RS), Santa Catarina (SC) and Mato Grosso do Sul (MS).

Interestingly, indices based on the minimum daily temperature best reflected the impact of hot days. The maximum value of the daily minimum temperature in February (tnx_Feb) presented critical values of 27 °C. When minimum daily temperatures are high, it is likely that maximum temperatures are also high, and the difference between minimum and maximum daily temperatures is small; this is a possible explanation for why the daily temperature range (dtr_Oct) has a negative impact on crop yields when its values are close to zero. Since minimum daily temperature is associated with night temperature (Frich et al., 2002), our results corroborate the finding that warm nights pose a great threat to crop yields (Sadok and Jagadish, 2020).

Our use of RF with SHAP provides an advance by enabling quantification of the combined effect of multi-hazards on food production. In the realm of risk for food production, this method could be applied to explain the seasonal impact forecast made with composite indicators such as the integrated information system (IIS) (Cunha et al., 2018; Marengo et al., 2017). This approach is also readily applicable to other natural hazards, including landslides, floods and wildfires when utilizing with other datasets.

4 Conclusions

This study aimed to assess the impacts of climate extremes on food production using explainable ML algorithms. To achieve this goal, we extensively examined various datasets, focusing on soybeans and second-season maize in Brazil. Our data sources included the Department of Rural Economy, the Brazilian Institute of Geography and Statistics (IBGE), and the Global Dataset of Historical Yields (GDHY). Through an ML analysis, we examined the effects of climate extremes on crop yield production, ultimately providing critical insights for the agricultural sector. Our analysis incorporated data from several Brazilian states, including RS, SC, PR, SP, MS, MG and GO for soybeans and PR, SP, MS, MG and GO for the second season of maize.

We employed two machine models to achieve our research objectives. In the first model, we explored different combinations of input data, encompassing precipitation and temperature means, and more complex combinations, including precipitation, temperature means and extremes. This approach allowed us to determine the most relevant climate indices for the investigated regions. In particular, this experiment validated the robustness of our methodology, as it successfully identified climate indices of particular significance for regional studies.

We took the most relevant indices from the first experiment in the second model and then applied SHapley Additive exPlanations (SHAP) explanatory analysis to explore how the random forest model utilized the important indices to predict the impact of climate extremes on food production. This analysis revealed the impact of these indices and provided insights that may be crucial in establishing significant thresholds and guidelines for effective climate-driven decision-making.

In conclusion, our research exemplifies the potential of ML to understand and harness the influence of climate variables on food production. By determining the most pertinent CIDs and exploring their significance in a regional context, our findings contribute to a growing body of knowledge critical for informed decision-making, policy development and adaptive strategies in the face of climate change and its impact on agriculture. As demonstrated in our study, the combi-

nation of data-driven insights and advanced modeling techniques offers a valuable pathway toward ensuring food security under a climate change.

Code and data availability. The code to reproduce the data analysis in this paper can be found on https://doi.org/10.5281/zenodo.12612860 (Benso, 2024). References to the data used are available in Table 1.

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/nhess-25-1387-2025-supplement.

Author contributions. Conception and design of the work: MRB, RFS, GCG. Data collection and manuscript drafting: MRB. Discussion and analysis: MRB, RFS, GCG, PAAM, EMM. Critical review of the manuscript: MRB, RFS, GCG, AMS, ALCBD, PAAM, JAM, EMM. Advisor: EMM.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Special issue statement. This article is part of the special issue "Methodological innovations for the analysis and management of compound risk and multi-risk, including climate-related and geophysical hazards (NHESS/ESD/ESSD/GC/HESS inter-journal SI)". It is not associated with a conference.

Acknowledgements. We wish to express our appreciation to the reviewers and editors for their valuable feedback and contributions to this project. We would like to thank the University of São Paulo for providing a stimulating research environment and resources that made this work possible. The authors also acknowledge that the generative AI technology ChatGPT 3.5 was used solely for minor text corrections to correct grammar and improve readability of an earlier version of this paper.

Financial support. This research has been supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (grant no. 88888.057913/2013-00).

Review statement. This paper was edited by Aloïs Tilloy and reviewed by two anonymous referees.

References

- Battisti, R. and Sentelhas, P. C.: New agroclimatic approach for soybean sowing dates recommendation: A case study, Rev. Bras. Eng. Agr. Amb., 18, 1149–1156, https://doi.org/10.1590/1807-1929/agriambi.v18n11p1149-1156, 2014.
- Benso, M. R.: Climatic impact-drivers in the context of food security, Zenodo [data set], https://doi.org/10.5281/zenodo.12612860, 2024.
- Bray, E. A.: Plant response to water-deficit stress, Encyclopedia of Life Sciences, https://doi.org/10.1002/9780470015902.a0001298.pub2, 2007.
- Brazilian Institute of Geography and Statistics (IBGE): PAM Municipal Agricultural Production, _ https://www.ibge.gov.br/en/statistics/economic/agricultureforestry-and-fishing/16773-municipal-agricultural-productiontemporary-and-permanent-crops.html?edicao=31814 (last access: 3 April 2025), 2023.
- Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.
- Chiang, F., Mazdiyasni, O., and AghaKouchak, A.: Evidence of anthropogenic impacts on global drought frequency, duration, and intensity, Nat. Commun., 12, 2754, https://doi.org/10.1038/s41467-021-22314-w, 2021.
- Cleveland, W. S., Grosse, E., and Shyu, W. M.: Local regression models, in: Statistical models in S, edited by: Chambers, J. M. and Hastie, T. J., Chapman & Hall, ISBN 0-412-05291-1, 309– 376, 2017.
- Cunha, A. P. M. d. A., Marengo, J. A., Cuartas, L. A., Tomasella, J., and Leal, K. R. D.: Drought monitoring and impacts assessment in Brazil: The CEMADEN experience, ICHARM Newsletter, http://repositorio.ufc.br/handle/riufc/59873, 2018.
- Das, S., Das, J., and Umamahesh, N.: Copula-based drought risk analysis on rainfed agriculture under stationary and non-stationary settings, Hydrolog. Sci. J., 67, 1683–1701, https://doi.org/10.1080/02626667.2022.2079416, 2022.
- de Geografia e Estatística, I. B.: Produção Agrícola Municipal 2022, http://www.sidra.ibge.gov.br/bda/pesquisas/pam (last access: 15 February 2024), 2022.
- Deusdará-Leal, K., Cuartas, L., Zhang, R., Mohor, G., Carvalho, L., Nobre, C., Mendiondo, E., Broedel, E., Seluchi, M., and Alvalá, R.: Implication of the new operation rules for Cantareira System: Re-reading of the 2014/2015 water crisis, Journal of Water Resource and Protection, 12, no. 4, 261–274, https://doi.org/10.4236/jwarp.2020.124016, 2019.
- Droogers, P. and Allen, R. G.: Estimating reference evapotranspiration under inaccurate data conditions, Irrigation and Drainage Systems, 16, 33–45, 2002.
- FAO: Agricultural production statistics 2000–2021, https://www. fao.org/3/cc3751en/cc3751en.pdf (last access: 10 November 2023), 2025.
- Frich, P., Alexander, L. V., Della-Marta, P., Gleason, B., Haylock, M., Tank, A. K., and Peterson, T.: Observed coherent changes in climatic extremes during the second half of the twentieth century, Clim. Res., 19, 193–212, https://doi.org/10.3354/cr019193, 2002.
- Gelcer, E., Fraisse, C., Dzotsi, K., Hu, Z., Mendes, R., and Zotarelli, L.: Effects of El Niño Southern Oscillation on the space-time variability of Agricultural Reference Index for

Drought in midlatitudes, Agr. Forest Meteorol., 174, 110–128, https://doi.org/10.1016/j.agrformet.2013.02.006, 2013.

- Guyon, I. and Elisseeff, A.: An introduction to variable and feature selection, J. Mach. Learn. Res., 3, 1157–1182, 2003.
- Hamed, R., Van Loon, A. F., Aerts, J., and Coumou, D.: Impacts of compound hot–dry extremes on US soybean yields, Earth Syst. Dynam., 12, 1371–1391, https://doi.org/10.5194/esd-12-1371-2021, 2021.
- Han, L., Yang, G., Dai, H., Xu, B., Yang, H., Feng, H., Li, Z., and Yang, X.: Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data, Plant Methods, 15, 1–19, https://doi.org/10.1186/s13007-019-0394-z, 2019.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., F., Geer, A., Haimberger, L., Healy, S.,Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly J. Roy. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.
- Hijmans, R. J.: terra: Spatial Data Analysis, R package version 1.7-29, https://CRAN.R-project.org/package=terra (last access: 15 January 2024), 2023.
- Iizumi, T. and Sakai, T.: The global dataset of historical yields for major crops 1981–2016, Scientific Data, 7, 97, https://doi.org/10.1038/s41597-020-0433-7, 2020.
- Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., and Kim, S.-H.: Random forests for global and regional crop yield predictions, PLOS ONE, 11, e0156571, https://doi.org/10.1371/journal.pone.0156571, 2016.
- Kim, W., Iizumi, T., and Nishimori, M.: Global patterns of crop production losses associated with droughts from 1983 to 2009, J. Appl. Meteorol. Clim., 58, 1233–1244, https://doi.org/10.1175/JAMC-D-18-0174.1, 2019.
- Komisarczyk, K., Kozminski, P., Maksymiuk, S., and Biecek, P.: treeshap: Compute SHAP Values for Your Tree-Based Models Using the 'TreeSHAP' Algorithm, R package version 0.3.1, https://CRAN.R-project.org/package=treeshap (last access: 15 January 2024), 2024.
- 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.
- Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., and Peng, B.: Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States, Glob. Change Biol., 25, 2325–2337, https://doi.org/10.1111/gcb.14628, 2019.
- Liu, Y. and Ker, A. P.: When less is more: on the use of historical yield data with application to rating area crop insurance contracts, Journal of Agricultural and Applied Economics, 52, 194– 203, https://doi.org/10.1017/aae.2019.40, 2020.
- Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st In-

M. R. Benso et al.: Data-driven assessment of climatic impact drivers

ternational Conference on Neural Information Processing Systems, NIPS'17, Long Beach, California, USA, Curran Associates Inc., Red Hook, NY, USA, 4768–4777, https://doi.org/10.5555/3295222.3295230, 2017.

- MAPA: Histórico de perdas na agricultura brasileira: 2000– 2021, https://www.gov.br/agricultura/pt-br/assuntos/riscosseguro/seguro-rural/publicacoes-seguro-rural/historico-deperdas-na-agricultura-brasileira-2000-2021.pdf (last access: 28 October 2023), 2022.
- Marengo, J. A., Alves, L. M., Alvala, R., Cunha, A. P., Brito, S., and Moraes, O. L.: Climatic characteristics of the 2010-2016 drought in the semiarid Northeast Brazil region, An. Acad. Bras. Cienc., 90, 1973–1985, https://doi.org/10.1590/0001-3765201720170206, 2017.
- Marengo, J. A., Cunha, A. P., Cuartas, L. A., Deusdará Leal, K. R., Broedel, E., Seluchi, M. E., Michelin, C. M., De Praga Baião, C. F., Chuchón Angulo, E., Almeida, E. K., Kazmierczak, M. L., Mateus, N., Pedro, A., Silva, R. C., and Bender, F.: Extreme drought in the Brazilian Pantanal in 2019–2020: characterization, causes, and impacts, Frontiers in Water, 3, 639204, https://doi.org/10.3389/frwa.2021.639204, 2021.
- Mariadass, D. A., Moung, E. G., Sufian, M. M., and Farzamnia, A.: Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture, in: 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), 17–18 November 2022, USA, IEEE, 219–224, https://doi.org/10.1109/ICCKE57176.2022.9960069, 2022.
- Mayer, M.: shapviz: SHAP Visualizations, R package version 0.9.1, https://CRAN.R-project.org/package=shapviz (last access: 3 April 2025), 2023.
- McKee, T. B., Doesken, N. J., and Kleist, J.: Drought Monitoring with Multiple Time Scales. In Proceedings of the Ninth Conference on Applied Climatology, Dallas, TX, American Meteorological Society, 233–236, 1995.
- Moriondo, M., Giannakopoulos, C., and Bindi, M.: Climate change impact assessment: the role of climate extremes in crop yield simulation, Climatic Change, 104, 679–701, https://doi.org/10.1007/s10584-010-9871-0, 2011.
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349– 4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.
- Ozaki, V. A., Goodwin, B. K., and Shirota, R.: Parametric and nonparametric statistical modelling of crop yield: implications for pricing crop insurance contracts, Appl. Econ., 40, 1151–1164, https://doi.org/10.1080/00036840600749680, 2008.
- Parana: Levantamento da Produção Agropecuária, https: //www.agricultura.pr.gov.br/deral/ProducaoAnual (last access: 4 April 2025), 2021.
- Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century, Nature Food, 3, 19–28, https://doi.org/10.1038/s43016-021-00429-z, 2022.

- Proctor, J., Rigden, A., Chan, D., and Huybers, P.: More accurate specification of water supply shows its importance for global crop production, Nature Food, 3, 753–763, https://doi.org/10.1038/s43016-022-00592-x, 2022.
- Pullanagari, R. R., Kereszturi, G., and Yule, I.: Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression, Remote Sens.-Basel, 10, 1117, https://doi.org/10.3390/rs10071117, 2018.
- Ranasinghe, R., Ruane, A. C., Vautard, R., Arnell, N., Coppola, E., Cruz, F. A., Dessai, S., Saiful Islam, A., Rahimi, M., Carrascal, D. R., Sillmann, J., Sylla, M. B., Tebaldi, C., Wang, W., and Zaaboul, R.: Climate Change Information for Regional Impact and for Risk Assessment, in Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1767–1926, https://doi.org/10.1017/9781009157896.014, 2021.
- Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nat. Commun., 6, 5989, https://doi.org/10.1038/ncomms6989, 2015.
- Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. A. M., Schmid, E., Stehfest, E., Yang, H., and Jones, J. W.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, P. Natl. Acad. Sci. USA, 111, 3268– 3273, https://doi.org/10.1073/pnas.1222463110, 2014.
- Ruane, A. C., Vautard, R., Ranasinghe, R., Sillmann, J., Coppola, E., Arnell, N., Cruz, F. A., Dessai, S., Iles, C. E., Saiful Islam, A. K. M., Jones, R. G., Rahimi, M., Carrascal, D. R., Seneviratne, S. I., Servonnat, J., Sörensson, A. A., Sylla, M. B., Tebaldi, C., Wang, W., and Zaaboul, R.: The Climatic Impact-Driver Framework for Assessment of Risk-Relevant Climate Information, Earths Future, 10, e2022EF002803, https://doi.org/10.1029/2022EF002803, 2022.
- Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N.: Crop planting dates: an analysis of global patterns, Global Ecol. Biogeogr., 19, 607–620, https://doi.org/10.1111/j.1466-8238.2010.00551.x, 2010.
- Sadok, W. and Jagadish, S. K.: The hidden costs of nighttime warming on yields, Trends Plant Sci., 25, 644–651, https://doi.org/10.1016/j.tplants.2020.02.003, 2020.
- Santini, M., Noce, S., Antonelli, M., and Caporaso, L.: Complex drought patterns robustly explain global yield loss for major crops, Sci. Rep.-UK, 12, 5792, https://doi.org/10.1038/s41598-022-09611-0, 2022.
- Sarhadi, A., Ausín, M. C., Wiper, M. P., Touma, D., and Diffenbaugh, N. S.: Multidimensional risk in a nonstationary climate: Joint probability of increasingly severe warm and dry conditions, Science Advances, 4, eaau3487, https://doi.org/10.1126/sciadv.aau3487, 2018.
- Schierhorn, F., Hofmann, M., Gagalyuk, T., Ostapchuk, I., and Müller, D.: Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages, Climatic Change, 169, 39, https://doi.org/10.1007/s10584-021-03272-0, 2021.

- Schyns, J. F., Hoekstra, A. Y., and Booij, M. J.: Review and classification of indicators of green water availability and scarcity, Hydrol. Earth Syst. Sci., 19, 4581–4608, https://doi.org/10.5194/hess-19-4581-2015, 2015.
- Sidhu, B. S., Mehrabi, Z., Ramankutty, N., and Kandlikar, M.: How can machine learning help in understanding the impact of climate change on crop yields?, Environ. Res. Lett., 13, 345–359, https://doi.org/10.1088/1748-9326/acb164, 2023.
- Silva Fuzzo, D. F., Carlson, T. N., Kourgialas, N. N., and Petropoulos, G. P.: Coupling remote sensing with a water balance model for soybean yield predictions over large areas, Earth Sci. Inform., 13, 345–359, https://doi.org/10.1007/s12145-019-00424-w, 2020.
- Sinnathamby, S., Douglas-Mankin, K. R., and Craige, C.: Fieldscale calibration of crop-yield parameters in the Soil and Water Assessment Tool (SWAT), Agr. Water Manage., 180, 61–69, https://doi.org/10.1016/j.agwat.2016.10.024, 2017.
- Strumbelj, E. and Kononenko, I.: An Efficient Explanation of Individual Classifications using Game Theory, J. Mach. Learn. Res., 11, 1–18, 2010.
- Štrumbelj, E. and Kononenko, I.: Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst., 41, 647–665, 2014.
- Svetnik, V., Liaw, A., Tong, C., and Wang, T.: Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules, in: Multiple Classifier Systems: 5th International Workshop, MCS 2004, Cagliari, Italy, 9–11 June, 2004, Proceedings 5, Springer, 334–343, https://doi.org/10.1007/978-3-540-25966-4_33, 2004.
- UNDRR: What is the Sendai Framework for Disaster Risk Reduction?, undrr.org, https://www.undrr.org/ implementing-sendai-framework/what-sendai-framework (last access: 24 October 2023), 2025.
- Vapnik, V. N.: An overview of statistical learning theory, IEEE T. Neural Networ., 10, 988–999, 1999.
- Viana, C. M., Santos, M., Freire, D., Abrantes, P., and Rocha, J.: Evaluation of the factors explaining the use of agricultural land: A machine learning and model-agnostic approach, Ecol. Indic., 131, 108200, https://doi.org/10.1016/j.ecolind.2021.108200, 2021.
- Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I.: A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index, J. Climate, 23, 1696–1718, 2010.

- 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, Environ. Res. Lett., 14, 054010, https://doi.org/10.1088/1748-9326/ab154b, 2019.
- von Bloh, M., Júnior, R. d. S. N., Wangerpohl, X., Saltık, A. O., Haller, V., Kaiser, L., and Asseng, S.: Machine learning for soybean yield forecasting in Brazil, Agr. Forest Meteorol., 341, 109670, https://doi.org/10.1016/j.agrformet.2023.109670, 2023.
- Wang, Y. and Li, Y.: Mapping the ratoon rice suitability region in China using random forest and recursive feature elimination modeling, Field Crop. Res., 301, 109016, https://doi.org/10.1016/j.fcr.2023.109016, 2023.
- Wikle, C. K., Datta, A., Hari, B. V., Boone, E. L., Sahoo, I., Kavila, I., Castruccio, S., Simmons, S. J., Burr, W. S., and Chang, W.: An illustration of model agnostic explainability methods applied to environmental data, Environmetrics, 34, e2772, https://doi.org/10.1002/env.2772, 2023.
- Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, J. Stat. Softw., 77, 1–17, https://doi.org/10.18637/jss.v077.i01, 2017.
- Yang, S.-R., Koo, W. W., and Wilson, W. W.: Heteroskedasticity in crop yield models, J. Agr. Resour. Econ., 17, 103–109, 1992.
- Yu, L. and Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution, in: Proceedings of the 20th international conference on machine learning (ICML-03), Washington, DC, USA, 856–863, https://doi.org/10.5555/3041838.3041946, 2003.
- Zhu, Y., Goodwin, B. K., and Ghosh, S. K.: Modeling yield risk under technological change: Dynamic yield distributions and the US crop insurance program, J. Agr. Resour. Econ., 36, 192–210, 2011.
- Zscheischler, J., Westra, S., Van Den Hurk, B. J., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., and Zhang, X.: Future climate risk from compound events, Nat. Clim. Change, 8, 469–477, 2018.