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

Marcos Roberto Benso[1], Roberto Fray Silva[2], Gabriela Chiquito Gesualdo[1,5], Antonio Mauro Saraiva[2], Alexandre Cláudio Botazzo Delbem[4], Patricia Angélica Alves Marques[3], José Antonio Marengo[6,7,8], and Eduardo Mario Mendiondo[1]

[1]São Carlos School of Engineering, University of Sao Paulo, Sao Carlos-SP, 13566-590, Brazil
[2]Institute of Advanced Studies, University of São Paulo, São Paulo, SP, Brazil
[3]Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, SP, Brazil
[4] Institute of Mathematics and Computer Sciences, University of Sao Paulo, Sao Carlos-SP, 13566-590, Brazil
[5] Department of Geosciences, Pennsylvania State University, State College- PA, 16801, USA
[6]National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN), São José dos Campos, SP, Brazil
[7]Graduate Program in Natural Disasters, UNESP/CEMADEN, São José dos Campos, Brazil
[8]Graduate School of international Studies. Korea University, Seoul, South Korea

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

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 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 the mean precipitation is a highly relevant CID. However, there is a window in which crops are more vulnerable to precipitation deficit. In many regions of Brazil, for example, soybeans face an increased risk of yield losses when precipitation falls below 100 mm/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 decision-making, 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-drivers (CID) framework, introduced by Working Group 1 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, n.d.), 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 CID 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 enhance 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 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 soybean and maize. Both studies found that mean climate variables over 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).

While eliminating redundant features and variables can enhance our understanding of data structure, the challenge of interpretability remains. We propose using the model-agnostic 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 1 (IPCC WG1). 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.

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 **relevant** to explain the target variable (in this study, crop yields) based on the input variables. However, it should not be **redundant** with 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 ~~pre-processing step in machine learning~~ preprocessing step in

ML models. The filtering process removes redundant features. ~~Other relevant aspects, such as relevancy, explainability, and operationability, will be explained in the following steps~~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 ~~Random Forest~~ 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 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 ~~machine learning~~ ML algorithm. The SHAP method ~~uses a second model, most commonly the Random Forest model,~~ is used to explain how each variable was used to make each prediction.

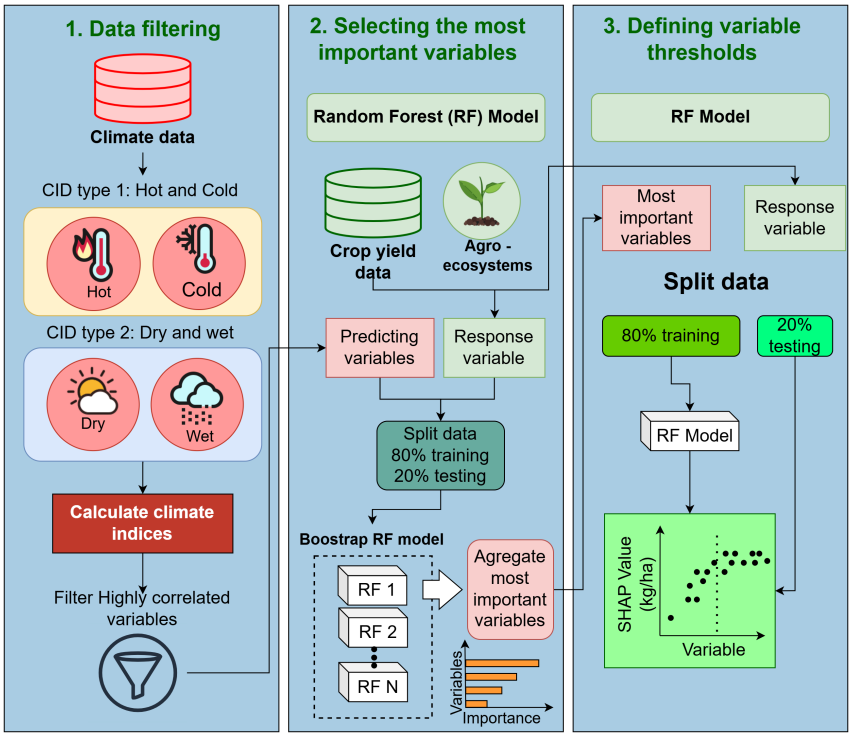

Figure 1. Flowchart illustrating the methodology proposed for analyzing the impact of climate indices on crop yield.

In the second step, we focus on identifying the most important climatic impact-drivers (CID) to assess their impact on food production. ~~To achieve this, we used the random forest model. Different models were trained considering different~~

~~combinations of input data, including precipitation means, temperature means, and combinations of means and extreme climate indices. The goal of this experiment was to identify the most importantclimate indices~~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 climatic impact-drivers (CID) on food production is customized to account for the unique characteristics, data, and conditions present in each state, rather than applying a single model uniformly across all states. The feature importance was determined based on entropy is determined by calculating the reduction in entropy (information gain) each feature provides when used to split the data at each node in the decision tree. Features that result in greater reductions in entropy across the tree are considered more important.

The creation of training, validation, and testing subsets is crucial to avoid overfitting and achieve reasonable estimates of model performance. The data set was divided ~~into the~~ according to the chronological order of the data. The first 80% ~~for training and~~ of the data, according to the timeline, was used to train the model, allowing it to learn and adjust its parameters. The remaining 20% ~~for validation data.~~ was used for validation, meaning this portion was reserved to test the model's predictions on data it hasn't 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 ~~datapoints~~ data points from neighboring municipalities from being correlated within the same year, the ~~best fit model was determined by employing~~ best-fit model was selected using a leave-one-year-out cross-validation ~~approach~~ method (LOYOCV)~~as suggested by von Bloh et al. (2023). Therefore, a fixed window of 10 years was used , and the one year~~ , 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 ~~designated~~ used for training, followed by one year as a test set. This process was repeated iteratively, leaving each year as the test set ~~and separated from the other data~~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 ~~and validation data sets, with their performance~~ 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 climatic impact-drivers (CID) to assess their impact on food production. To achieve this, we used the ~~Random Forest model. Different models were trained considering different combinations of input data~~RF model. Models trained with three different crop yield datasets, 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 the ~~Random Forest~~ RF explainability with 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 ~~Random Forest~~ RF models were implemented using the R package ranger (Wright and Ziegler, 2017), and the SHAP explanations 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 values of the adjusted and detrended crop yields $E[f(X)]$. Then, the ~~Random Forest~~ 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 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 explanations can be learned in Strumbelj and Kononenko (2010).

The ~~the~~ SHAP explanations was performed for each prediction using the R-Package treeshap (Komisarczyk et al., 2024). The package allows 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 unveil 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, considering two features i and j, their interaction is determined by subtracting the sum of their individual SHAP values from their combined contribution, i.e., Interaction $= \text{SHAP}_{i,j} - (\text{SHAP}_i + \text{SHAP}_j)$. This method captures how the joint presence of two features affects the prediction compared to their independent contributions, unveiling 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, n.d.). This 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 soybean and maize of 58 million hectares. 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), Minas Gerais (MG). The selection criteria will be explained in the sub-section Crop yield data. The map 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 that the planting dates follow a normal distribution, with the mean date being the most probable date for farmers and the maximum and minimum dates being considered twice the standard

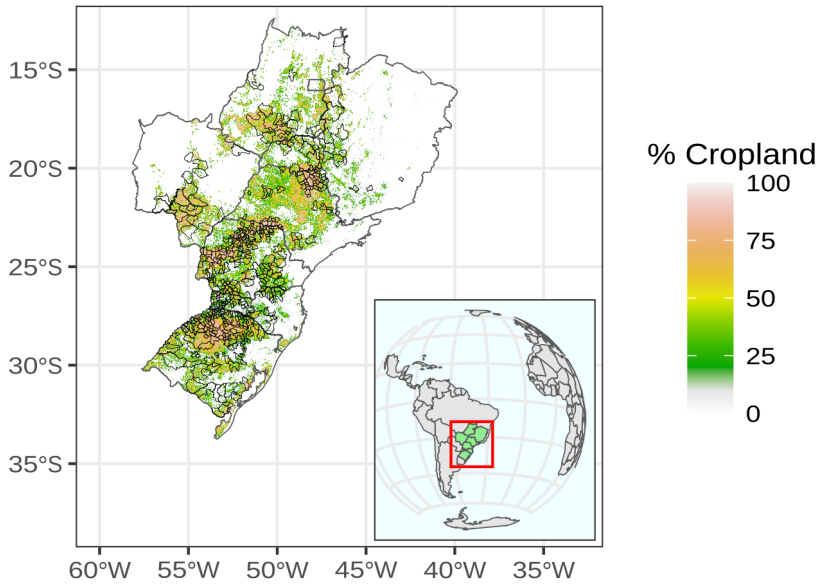

Figure 2. Location of selected study municipalities with respect with observed cropland extent between from 2003 to 2019

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 the late summer and extends to fall, from February to March. Since the second season of ~~Maize~~ maize peaks in February, we consider that the end of ~~Soybean~~ soybean is in February. For the second season of maize, planting begins at the end of January, after soybean harvest, peaks in February, and ends at 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~~ soybean and maize crops in the state of Parana. We used two criteria to select a dataset: (i) the data must comply with FAIR principles (i.e., data must have findability, accessibility, interoperability, and reusability); and (ii) climate data must be updated frequently (ideally, with minimum 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) 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 Figure 1.

Dataset	Variable	Spatial resolution	Temporal resolution	Time frame
Crop yields IBGE	Soybean	Municipality level	Annual	1974-2022
Crop yields IBGE	Maize	Municipality level	Crop season	2003-2022
Crop yields Paraná	Soybean and Maize	Municipality level	Crop season	1997-2021
Crop yields GDHY	Soybean and Maize	0.5 degree cell	Crop season	1981-2016
ERA5 Land	Precipiation, temperature	0.1 degree cell	Daily	1950-Current

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

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 municipality level. We used three data sets 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 is an annual aggregation of the total production of that crop within the municipality.

However, IBGE has started to collect 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 municipality level were filtered based on data availability. The missing years were removed from the dataset, and the municipalities with more than two 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 second season for Paraná (PR), São Paulo (SP), Mato Grosso do Sul (MS), Minas Gerais (MG) and Goiás (GO).

The Department of Rural Economy (Deral) of the Paraná state, 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 IBGE; 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 ton/ha based on Food and Agriculture Organization (FAO) country-level yield statistics and then corrected using the remote-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 on 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 municipalities 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 IBGE. 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 trend such as climate change were removed by Local Polynomial Regression Fitting (LOESS) (Cleveland et al., 2017). Moreover, systematic changes in crop yields has 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 Supplementary Material.

2.3.2 ERA5 Land reanalysis dataset

Weather data was 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 data set has a 0.1 x 0.1 lon lat-lon grid and was aggregated to the municipal area to match the spatial discretization of 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 WGI of the IPCC has presented the climatic impact-drivers (CID) 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, 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 time scales, 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 is 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 time scales representing previous meteorological conditions, typically 1 to 48 months. For agricultural applications, 3-month SPI is the most frequently used Kim et al. (2019).

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

Table 2. Description of the Climatic Impact-Drivers (CID) considered in this study and their respective Indices

CID Type	CID Category	CID Index Abbreviation	CID Index Description
Heat and Cold	Mean air temperature	Temp	Monthly temperature mean
		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
		TXN	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
Wet and dry	Mean precipitation	Prctot	Monthly precipitation sum
	Heavy precipitation	R10mm, R20mm	Monthly count of days when daily precipitation is higher than 10 and 20 mm.
		Rx1day, Rx1day	Monthly maximum 1-day and 5-day precipitation
	Agricultural and ecological drought	SPEI 3, 6	Standardised Precipitation and Evapotranspiration Index for 3 and 6 moths accumulations
		SPI 3, 6	Standardised Precipitation Index for 3 and 6 moths accumulations

3 Results and Discussion

The comparison of the datasets used in this study is important to evaluate the reliability of the data. High-quality crop yield data improves the calibration of crop growth models (Rosenzweig et al., 2014). However, they have a broader application in

geosciences. Crop yield data is 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 Parná 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, GDHY and IBGE 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 is 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 estimation of GHDY. In addition, the exact location of the planted area within each municipality can vary from year to year.

(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

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

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 ~~machine learning~~ ML technique by 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 ~~random forest~~ RF model considering different datasets. Taking the coefficient of determination, the climate variables explained the variability of soybean crop yields, on average, from 30 to 40% of 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 ~~random forest~~ 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 a ~~random forest~~ 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.

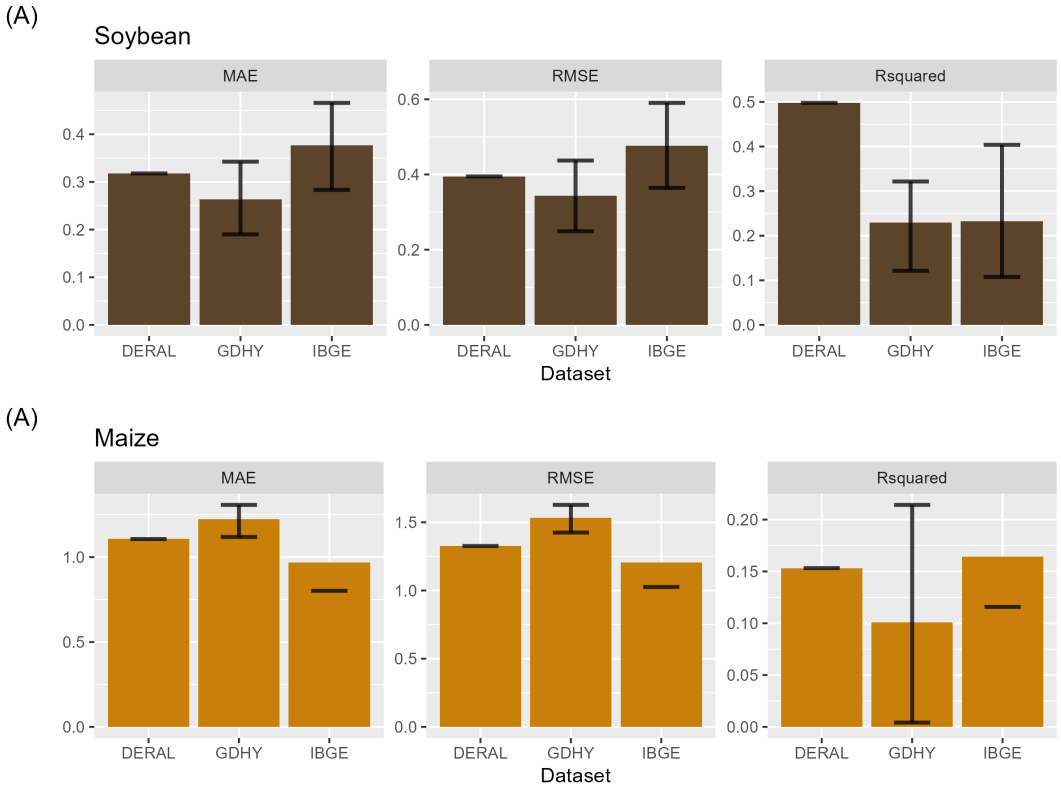

Figure 3. Performance evaluation of regression model for (a) Soybean and (b) Maize. Considering data from Department of Rural Economics (Deral) with data only for Paraná, Global Dataset on Historical Yields (GDHY) and Brazilian Institute of Geography and Statistics (IBGE) for the major historical agricultural producing municipalities

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 of crop yield more than only the temperature means, which was observed in all three datasets. For maize second season, for MS, MG, and GO, the models that combine mean temperature to mean precipitation and extremes explained more the variability of crop yields than only temperature, which was observed in the GDHY and IBGE datasets. The IBGE database showed that the maize model was much more effective in São Paulo than in any other state.

The ~~random forest~~ 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, 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 climatic

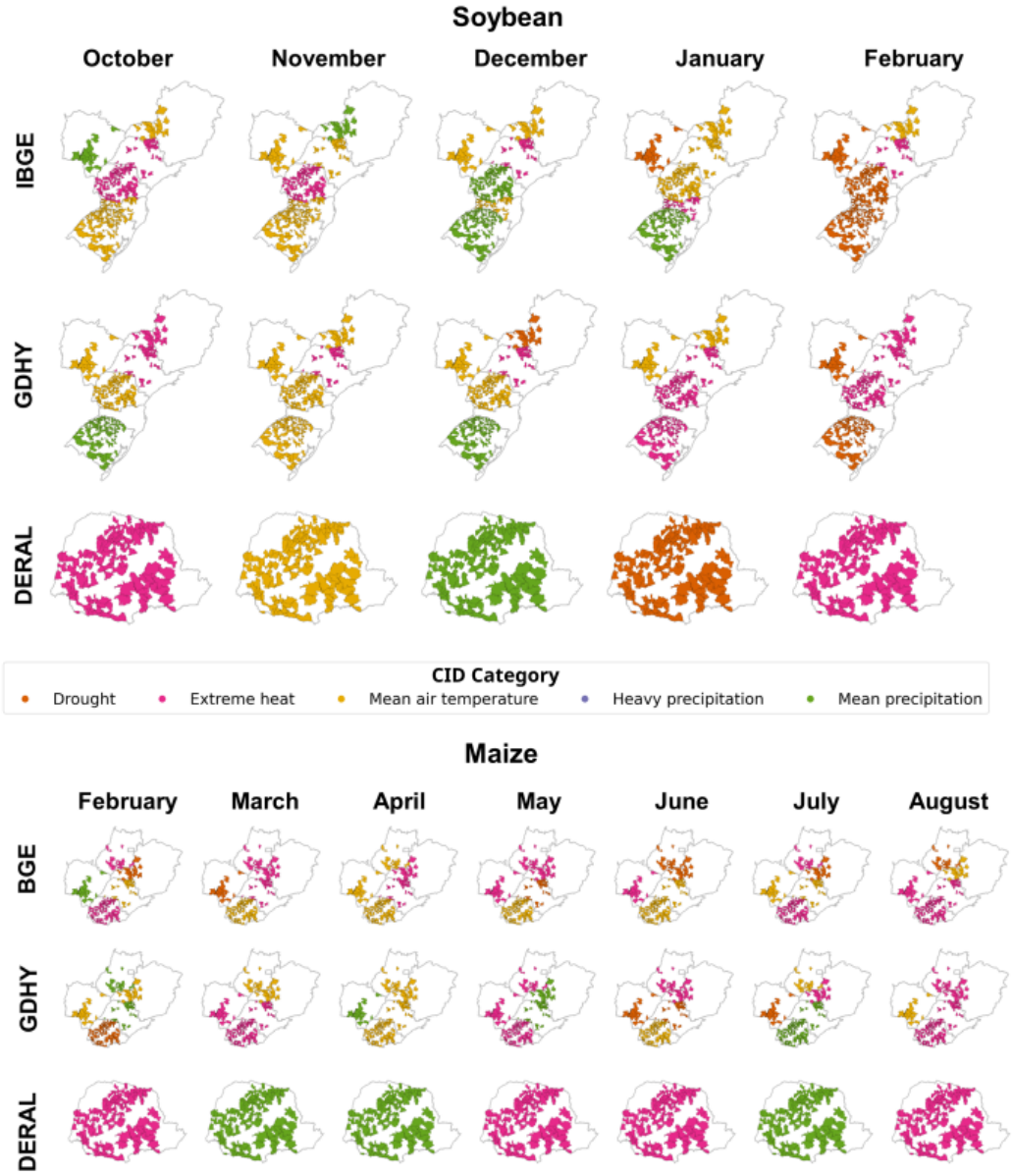

Figure 4. Importance of ~~CIDs~~ key CID categories in Predicting ~~Soybean~~ soybean and ~~Maize~~ maize crop yields. The figure displays the most significant features identified by the Random Forest model for soybean and maize.

impact-drivers considering each state and dataset. The following sections summarize the key CID categories for soybeans and maize.

Feature importance was summarized spatially and temporally. Fig. 4 highlights similatities and dissimilarities from feature importance considering the different datasets.

In the supplementary material, we show tables with the results of variable importance for all models. The analysis ~~can be~~ of variable importance for soybean datasets is shown in Table S1. The analysis identifies extreme heat (tnx), drought index (spei), and precipitation totals (prcptot) 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. ~~Overall~~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 second cycle, 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. Precipitation deficit leads to reduced available soil moisture, affecting plant development and reducing crop yields, and it is considered the most critical environmental factor that reduces crop yields (Bray, 2007). In our analysis, mean precipitation was one of the most important climatic impact-driver on soybean crop yields during January and February for RS, SC, PR, MS, and also in December in 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 impact on soybean crop yields during La Niña . The model did not indicate that mean precipitation was the most important for the state of SP. For maize production, mean precipitations during April and May were considered important for PR, MS, MG, and GO, except for SP.

Agricultural systems require minimum rainfall, or they rely on irrigation. In Brazil, SP, MS, MG, and GO states 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~~ 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 affects 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 time scales. 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 on the list of most relevant variables. This indicates that pre-sowing meteorological factors that can reduce soil moisture conditions also influence crop yields. This result corroborates Santini et al. (2022), which revealed that 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 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.

3.2 Determining Thresholds and Their Significance

With the results of the selection of critical climatic ~~impact-divers~~impact-drivers, 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 to predict, we used SHAP. This technique allowed us to extract insights ~~by coupling~~ from the results of a Random Forest model, thus providing a comprehensive perspective on 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 agro-producing municipalities, namely Marechal Cândido Rondom 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 Rondom, the Deral yields similar to IBGE, 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 driver 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, 3-month SPEI in January. The other variables represented a negligible influence on crop yield losses. For the city of Uniflor, the actual yields of Deral and IBGE were similar, 2.44 and 2.48 (Fig. 5), respectively. The main influences on crop yield losses were precipitation in December (prctot_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 two months (December and January).

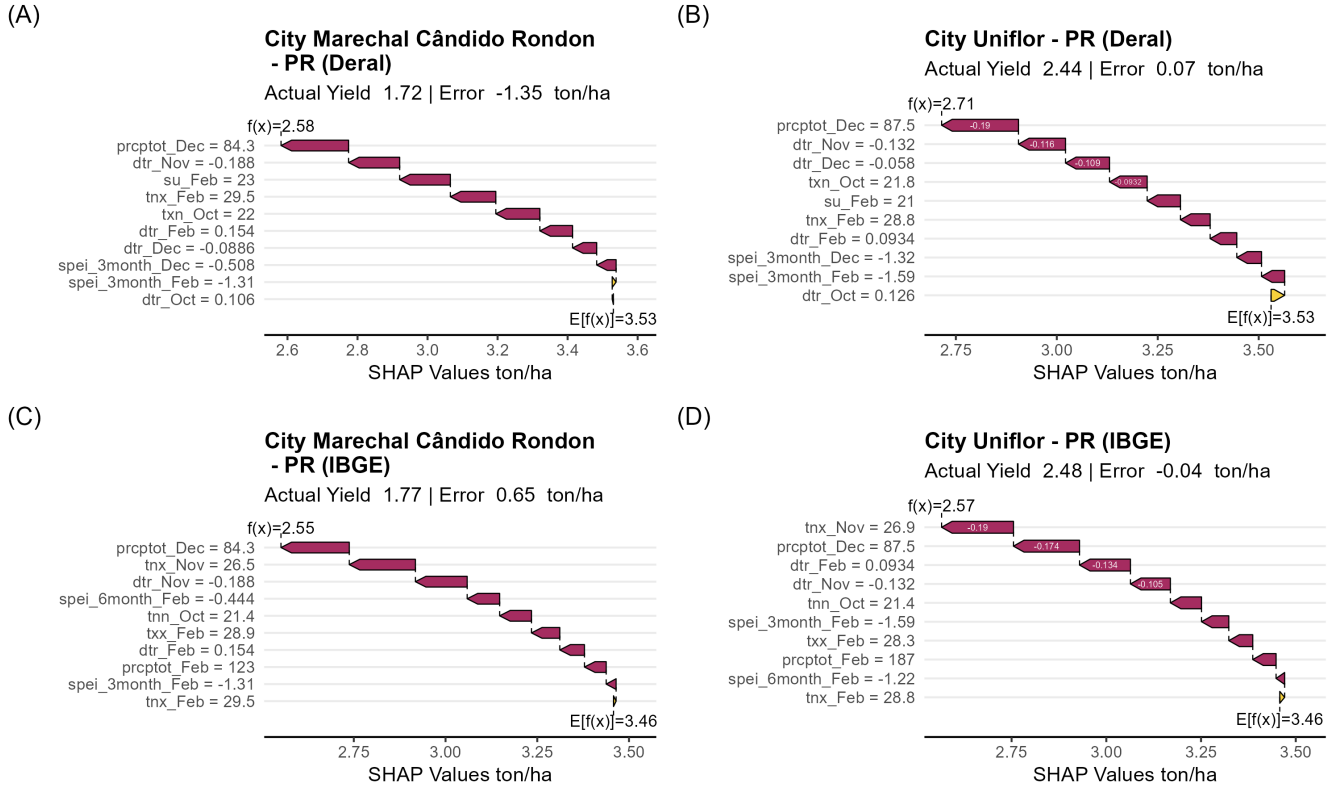

Figure 5. SHAP waterfall plot visualizing the key climatic impact-drivers contributions to crop yield losses for the state of Paraná (PR) in 2019, a drought year.Monthly Maximum Value of Daily Minimum Temperature (tnx), Minimum Value of Daily Maximum Temperature (txn), Minimum Value of Daily Minimum (tnn), Maximum Value of Daily Maximum (txx), Percentage of days when TX > 90th percentile, Standardized Precipitation Evapotranspiration Index (SPEI), Total Precipitation (prctot), Daily temperature range (DTR)

In 2014/2015, 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 results of IBGE and GDHY for two municipalities of 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 2014/2015 (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.

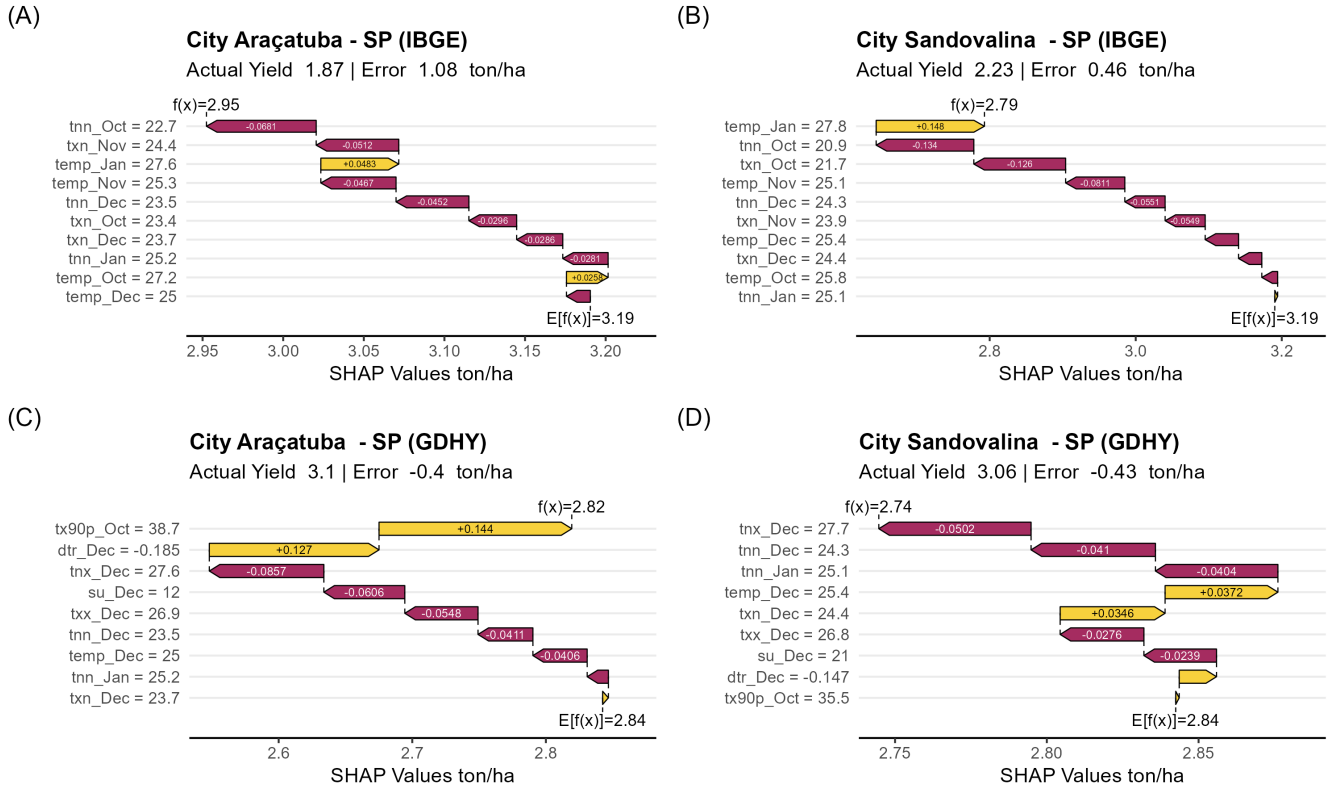

Figure 6. SHAP waterfall plot visualizing the key climatic impact-drivers contributions to crop yield losses for the state of São Paulo (SP) in 2015, a drought year. Monthly Maximum Value of Daily Minimum Temperature (tnx), Minimum Value of Daily Maximum Temperature (txn), Minimum Value of Daily Minimum (tnn), Maximum Value of Daily Maximum (txx), Percentage of days when TX > 90th percentile, Standardized Precipitation Evapotranspiration Index (SPEI), Total Precipitation (prctot), Daily temperature range (DTR)

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.

~~We compared two datasets .~~ The comparison of the two datasets can be found in Fig. 7. The IBGE data set 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 tons/ha. 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). ~~The upper limit indicates that extreme wet conditions~~Extreme wet conditions, with a threshold of 1, also affect crop yields in the RS state. ~~We find threshold 1.~~

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 ton/ha of losses. More studies are suggested on this type of hazard. The cumulative precipitation in February presented a threshold of around 100 mm and the potential to cause losses of approximately 0.4 ton/ha. 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 ton/ha; however, the threshold for IBGE is 150 mm and for GDHY, it is 120 mm.

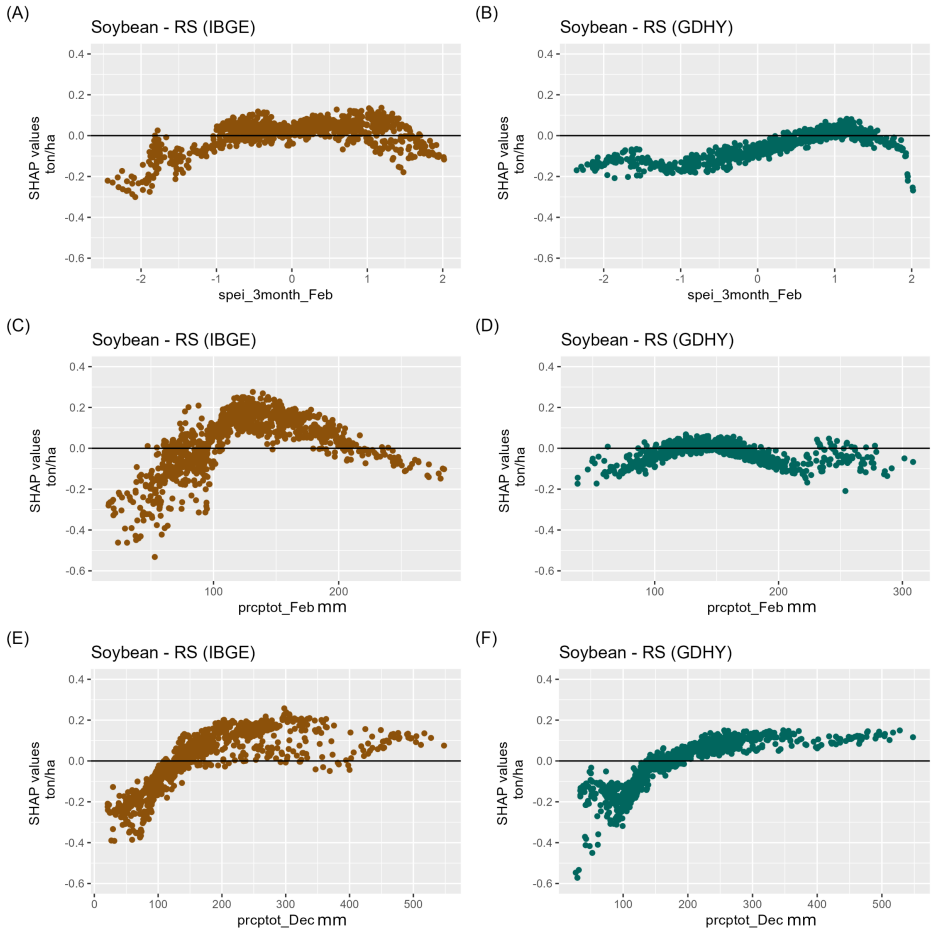

Figure 7. A comparison of the key climatic impact-drivers derived from the Brazilian Institute of Geography and Statistics (IBGE) from 2013 to 2021 and the Global Dataset on Historical Yields (GDHY) from 2009 to 2016 annual data aggregated at the municipal level was used to create a dependence plot for the soybean explanation model for validation data in RS. The data spanned from 2013 to 2021 for IBGE and from 2008 to 2016 for GDHY. Considering 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

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 data set. 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 do with 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 the model performance. However, 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 ~~exemplify~~ 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). (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 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) (see Section 3.3.2.1).

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 of 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 result corroborate that warm nights pose a great threat to crop yields Sadok and Jagadish (2020).

Our use of RF with SHAP explanation 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 Drought Index (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 ~~machine learning~~ ML algorithms. To achieve this goal, we extensively examined various datasets, focusing on soybean 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 for Historical Yields (GDHY). Through a ~~machine learning~~ 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 conducted 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

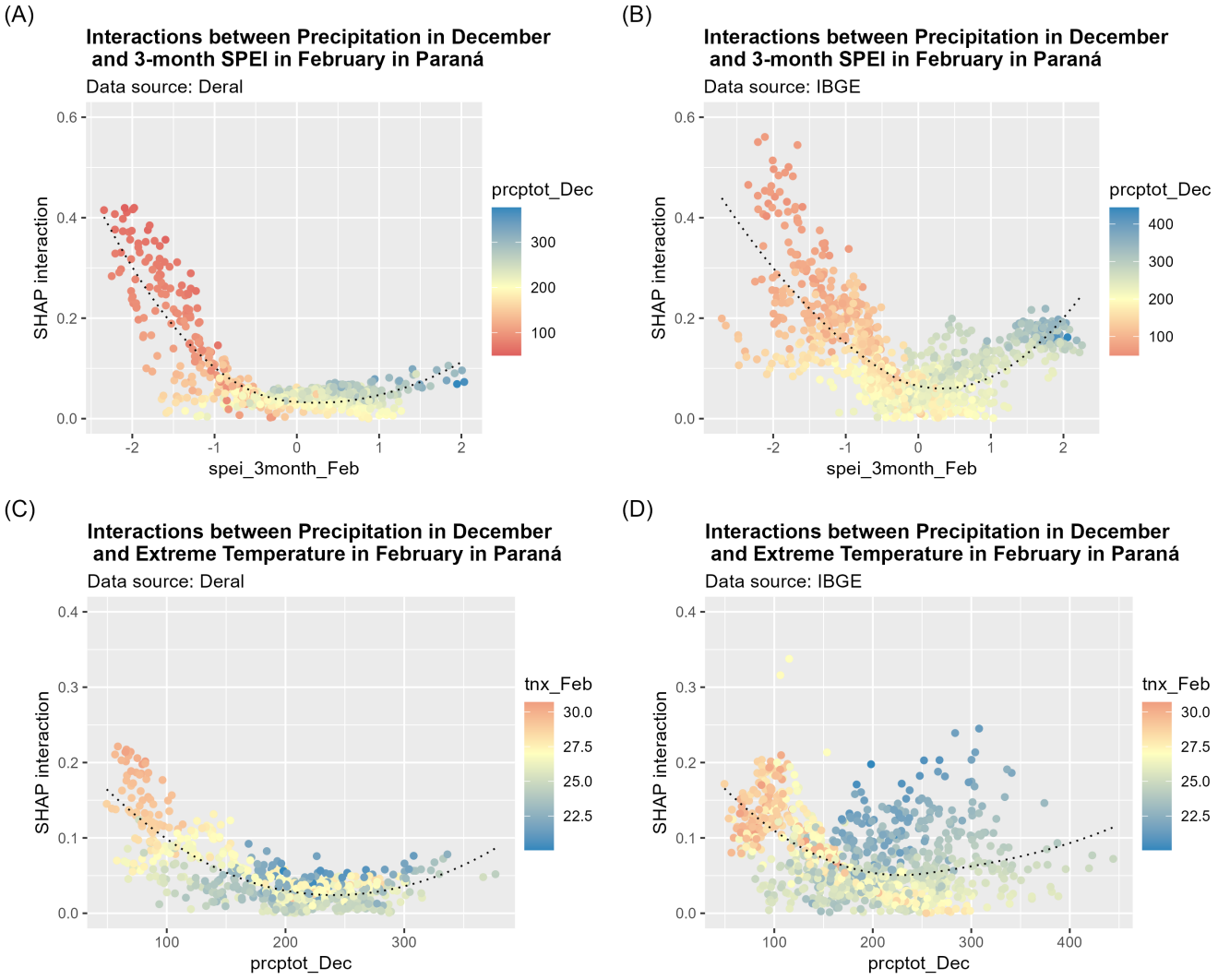

Figure 8. 2D partial dependence derived from from the Deral and IBGE of the state of Parana from 2016 to 2021 of the key climatic impact-driver total precipitation in December (prctot_Dec) with the most correlated CID and their combined impact on SHAP values (ton/ha): Considering 3-month SPEI in February (spei_3month_Feb) and Total precipitation in December (prcptot_Dec); Total precipitation in December (prcptot_Dec) and Maximum Value of Daily Minimum (tnn_Feb)

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

Code and data availability. The code to reproduce the data analysis in this manuscript can be found in https://github.com/marcosrbenso/ClimateImpactML

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

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

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 only and solely to minor text corrections for correcting grammar and improving readability.

References

Battisti, R. and Sentelhas, P. C.: New agroclimatic approach for soybean sowing dates recommendation: A case study, Revista Brasileira de Engenharia Agrícola e Ambiental, 18, 1149–1156, 2014.

Bray, E. A.: Plant response to water-deficit stress, Encyclopedia of Life Sciences, 2007.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Chiang, F., Mazdiyasni, O., and AghaKouchak, A.: Evidence of anthropogenic impacts on global drought frequency, duration, and intensity, Nature communications, 12, 2754, 2021.

Cleveland, W. S., Grosse, E., and Shyu, W. M.: Local regression models, in: Statistical models in S, pp. 309–376, Routledge, 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, 2018.

Das, S., Das, J., and Umamahesh, N.: Copula-based drought risk analysis on rainfed agriculture under stationary and non-stationary settings, Hydrological Sciences Journal, 67, 1683–1701, 2022.

de Geografia e Estatística, I. B.: Produção Agrícola Municipal 2022, http://www.sidra.ibge.gov.br/bda/pesquisas/pam, 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, Water Resour. Res, 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, (Accessed on 10/11/2023), n.d.

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, Climate research, 19, 193–212, 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, Agricultural and Forest Meteorology, 174, 110–128, 2013.

Guyon, I. and Elisseeff, A.: An introduction to variable and feature selection, Journal of machine learning research, 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 System Dynamics, 12, 1371–1391, 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, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hijmans, R. J.: terra: Spatial Data Analysis, https://CRAN.R-project.org/package=terra, r package version 1.7-29, 2023.

Iizumi, T. and Sakai, T.: The global dataset of historical yields for major crops 1981–2016, Scientific Data, 7, 97, 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., et al.: Random forests for global and regional crop yield predictions, PloS one, 11, e0156 571, 2016.

Kim, W., Iizumi, T., and Nishimori, M.: Global patterns of crop production losses associated with droughts from 1983 to 2009, Journal of Applied Meteorology and Climatology, 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, https://CRAN.R-project.org/package=treeshap, r package version 0.3.1, 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, 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, Global change biology, 25, 2325–2337, 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, 2020.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, 30, 2017.

MAPA, M. d. A. P. e. A.: Histórico de perdas na agricultura brasileira : 2000-2021, https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/publicacoes-seguro-rural/historico-de-perdas-na-agricultura-brasileira-2000-2021.pdf, [Accessed 28-10-2023], 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, Anais da Academia Brasileira de Ciências, 90, 1973–1985, 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., et al.: Extreme drought in the Brazilian Pantanal in 2019–2020: characterization, causes, and impacts, Frontiers in Water, 3, 639 204, 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), pp. 219–224, IEEE, 2022.

Mayer, M.: shapviz: SHAP Visualizations, https://CRAN.R-project.org/package=shapviz, r package version 0.9.1, 2023.

McKee, T. B.: Drought monitoring with multiple time scales, in: Proceedings of 9th Conference on Applied Climatology, Boston, 1995, 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, 2011.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., et al.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, 2021.

Ozaki, V. A., Goodwin, B. K., and Shirota, R.: Parametric and nonparametric statistical modelling of crop yield: implications for pricing crop insurance contracts, Applied Economics, 40, 1151–1164, 2008.

Parana: Levantamento da Produção Agropecuária, https://www.agricultura.pr.gov.br/deral/ProducaoAnual, https://doi.org/10.4225/13/511C71F8612C3, 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, 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, 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 Sensing, 10, 1117, 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., et al.: Chapter 12: Climate Change Information for Regional Impact and for Risk Assessment — ipcc.ch, https://www.ipcc.ch/report/ar6/wg1/chapter/chapter-12/, [Accessed 08-10-2023], 2021.

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nature communications, 6, 5989, 2015.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., et al.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proceedings of the national academy of sciences, 111, 3268–3273, 2014.

Ruane, A. C., Vautard, R., Ranasinghe, R., Sillmann, J., Coppola, E., Arnell, N., Cruz, F. A., Dessai, S., Iles, C. E., Islam, A. S., et al.: The Climatic Impact-Driver Framework for Assessment of Risk-Relevant Climate Information, Earth's Future, 10, e2022EF002 803, 2022.

Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N.: Crop planting dates: an analysis of global patterns, Global ecology and biogeography, 19, 607–620, 2010.

Sadok, W. and Jagadish, S. K.: The hidden costs of nighttime warming on yields, Trends in Plant Science, 25, 644–651, 2020.

Santini, M., Noce, S., Antonelli, M., and Caporaso, L.: Complex drought patterns robustly explain global yield loss for major crops, Scientific reports, 12, 5792, 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, 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, 2021.

Schyns, J. F., Hoekstra, A. Y., and Booij, M. J.: Review and classification of indicators of green water availability and scarcity, Hydrology and Earth System Sciences, 19, 4581–4608, 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?, Environmental Research Letters, 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 Science Informatics, 13, 345–359, 2020.

Sinnathamby, S., Douglas-Mankin, K. R., and Craige, C.: Field-scale calibration of crop-yield parameters in the Soil and Water Assessment Tool (SWAT), Agricultural water management, 180, 61–69, 2017.

Strumbelj, E. and Kononenko, I.: An efficient explanation of individual classifications using game theory, The Journal of Machine Learning Research, 11, 1–18, 2010.

Štrumbelj, E. and Kononenko, I.: Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems, 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, June 9-11, 2004. Proceedings 5, pp. 334–343, Springer, 2004.

UNDRR: What is the Sendai Framework for Disaster Risk Reduction? — undrr.org, https://www.undrr.org/implementing-sendai-framework/what-sendai-framework, [Accessed 24-10-2023], n.d.

Vapnik, V. N.: An overview of statistical learning theory, IEEE transactions on neural networks, 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, Ecological Indicators, 131, 108 200, 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, Journal of 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, Environmental Research Letters, 14, 054 010, 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, Agricultural and Forest Meteorology, 341, 109 670, 2023.

Wang, Y. and Li, Y.: Mapping the ratoon rice suitability region in China using random forest and recursive feature elimination modeling, Field Crops Research, 301, 109 016, 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, 2023.

Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of Statistical Software, 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, Journal of Agricultural and Resource Economics, pp. 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), pp. 856–863, 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, Journal of Agricultural and Resource Economics, pp. 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., et al.: Future climate risk from compound events, Nature Climate Change, 8, 469–477, 2018.