International Institute for Applied Systems Analysis
IIASA

Address: IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
Email: repository@iiasa.ac.at Telephone: +43 (0)2236 807 342

YSSP Report

Young Scientist Summer Program

Future climatic impacts on soybean yields in South America and its consequences on agricultural markets

Author: Henrique Moreno Dumont Goulart
Email: henrique.goulart@deltares.nl

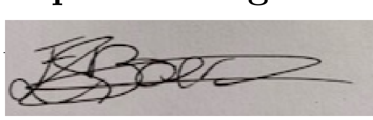
Approved by

Supervisor: Esther Boere
Co-Supervisor: Christian Folberth
Program: IBF
September 29, 2021

This report represents the work completed by the author during the IIASA Young Scientists Summer Program (YSSP) with approval from the YSSP supervisor.

It was finished by 29-09-2021 and has not been altered or revised since.

Supervisor signature:



ABSTRACT

Whilst used in all continents, the majority of soybean production takes place in specific regions. This causes the supply chain to be highly vulnerable to local climate shocks. With a changing climate, the impacts on crop productivity are expected to change as well. Crop models are commonly used to address the relation between climate and agriculture, simulating biophysical processes and estimating crop yields at a gridded location. However, they underestimate the impacts of extreme climatic conditions. In this study, a process-based crop model EPIC-IIASA is combined with a machine learning model to analyse the impact of future extreme weather events on soybean yields under different climate scenarios. The crop model outputs, such as simulated soybean yield, together with extreme climatic indices during the soybean growing season are used as input into a machine learning model that is trained on observed soybean yields. The coupling of crop model and machine learning model leads to a hybrid model that improves the crop model's ability to represent extreme weather conditions and to translate these into yield anomalies and shocks. These events are subsequently provided as input to GLOBIOM to analyse the socio-economic and market impact of extreme climate events on soybean production. We show the hybrid model can significantly improve the overall performance of the crop model, with an increased interannual variability representation, especially for extreme climatic conditions. We find low precipitation values to be responsible for low yields in the region. The global warming scenarios suggest weather shocks will likely become more frequent and intense, especially for high-emission scenarios and especially in the medium to long term. The socio-economic impacts demonstrate a non-linear relationship to the bio-physical impacts, with metrics like price per ton of soybean being significantly amplified.

ACKNOWLEDGMENTS

- Dr. Esther Boere (IIASA), Dr. Christian Folberth (IIASA), Karin van der Wiel (KNMI) and Bart van den Hurk (Deltares) for supervision

Contents

List of Figures

List of Tables

1 Introduction

Soybeans are the most common source of protein for livestock feed worldwide. They are also the second most consumed type of vegetable oil, and are widely used for human consumption. The global production of soybeans in 2019 was of 333 million tonnes, making it the 8th most produced crop in the world [FAO, 2021]. While soybeans are consumed globally, 80% of their production is concentrated in hotspot regions in the United States of America, Brazil and Argentina. These circumstances render soybeans the most traded crop in value in the world, and the soybean exports from Brazil and the US make the two countries the top exporters in value per commodity, among all commodities and countries [FAO, 2021]. As a consequence, the soybean trade entails a vast logistics network to connect production sites with consumers around the world. Population growth trends, dietary changes and an increased globalised world are all contributing factors to further extend the importance of the global soybean distribution network.

The high dependency of the global soybean market on concentrated producing regions carry a risk factor. The entire supply chain is vulnerable to disturbances in the local production sites. As for every crop, extreme weather events can lead to short-term variability and shocks to agricultural production [Zampieri et al., 2017, Zscheischler et al., 2017, Ben-Ari et al., 2018], potentially impacting the entire food supply system and posing threats to market stability and food security at a global level [Mitra and Josling, 2009, Götz et al., 2013, Shama, 2011]. In addition, with a changing climate, the conditions necessary for yield failures to happen are likely to change [IPCC, 2014], possibly intensifying the impacts on the whole supply chain of soybeans.

In order to represent the relation between climate and crops, it is common practice to use models, which can be physical or statistical. The first one, also known as crop model, uses data on weather, soil and management information as input and replicates the biophysical dynamics of real crops, ultimately generating crop-related output, such as yield and biomass. Given physical models are composed of physical equations, the internal workings and mechanisms of the physical models are explicitly modeled and, therefore, accessible. On the other hand, they are complex to build and operate, and highly time consuming in terms of computing processing. A specific case of crop models are the global gridded crop models (GGCMs), that operate at a global scale. Advantages of using GGCMs is that virtually the entire globe can be modelled for controlled and uniform time periods and forcing data sets, in contrast to observed data, where the less developed parts of the world suffer from limited data coverage or quality. The simulation periods of crop models tend to be more extensive than available observed crop data, as trustworthy and precise registries of observed yields is rather limited. Last of all, crop models are compatible with climate projections, rendering them particularly useful for climate change studies. An disadvantage of using crop models is that they underrepresent impacts from extreme climate conditions [Schewe et al., 2019]. It is argued that crop models simplify or have limited mechanisms to properly simulate the biophysical processes under extreme conditions [Feng et al., 2019].

Statistical models are another option for linking climate with agriculture. They work by establishing mathematical relationships between one or more variables and the output. They tend to be simpler to build and run when compared to the physical models, perform relatively close to process models [Ciscar et al., 2018] and are compatible with extreme climate conditions (but limited to the calibration range of values). Yet, they do not follow physics equations and are, thus, hard to understand or interpret.

The soybean shocks in yield due to weather variability are an intermediate step in the soybean chain of events. The ultimate consequences are associated with the socio-economic impacts, which includes shocks and anomalies on commodity prices, trade, exports, consumption, and food security risk around the world. The interactions between biophysical impacts and socio-economic impacts are neither linear nor uni-dimensional, with socio-economic feedbacks due to crop yields variations having a larger impact than the direct impacts of climate variability on crop yields [Calvin and Fisher-Vanden, 2017]. It is therefore necessary to explicitly represent the interactions on the socio-economic domain, which is often done through impact models. Together with the information on human and environmental processes, climate change and the crop yield shocks are studied to evaluate the potential socio-economic impacts of climate change feedback loops and uncertainties. It makes the chain of climate-crop-impacts events, attribution and integrated assessment central to the study of climate change [Moss et al., 2010].

The chain of climate-crop-impacts events has a particular drawback when dealing with climate shocks. In cases where socio-economic models rely on crop models estimations to assess impacts, the underestimation of biophysical impacts under extreme weather conditions generates a cascading effect to the socio-economic impacts [Schewe et al., 2019]. It is thus particularly relevant for socio-economic studies based on crop models to improve the representation of biophysical impacts under climate extremes. Recently, there has been a surge of studies based on hybrid models, which consists of combining physical models with statistical models for different purposes, such as downscalling [Folberth et al., 2019] and improving overall performance [Roberts et al., 2017, Ruane et al., 2017, Calvin and Fisher-Vanden, 2017]. More specifically, hybrid models have been shown to improve the representation of extreme climate, due to the addition of statistical mechanisms to crop models [Feng et al., 2019]. However, to the best of our knowledge, there has been no expansion on the use of hybrid models to integrate with socio-economic studies.

In this study, we couple a hybrid model to a socio-economic impact model to explore the socio-economic impacts that extreme climate might pose to the soybean production and distribution under global warming. More specifically, a process-based crop model EPIC-IIASA is combined with a machine learning model to analyse the impact of future extreme weather events on soybean yields under different climate scenarios and global climate models (GCMs). The crop model outputs, such as soybean biomass and yields, together with observed climatic variables during the soybean growing season, are used as input into a machine learning model that is trained on observed soybean yields, the hybrid model. The statistical model improves the crop model's ability to represent extreme weather conditions and to translate these into yield anomalies and shocks. The focus is put on low-probability high-impact extreme events and the extent to which these are expected to exacerbate in the future. These events are subsequently provided as input to a socio-economic impact model called GLOBIOM to analyse the socio-economic and market impacts of extreme climate events on soybean production. We adopt as case study the soybean production in Brazil, as it is the largest soybean producer in the world and responsible for approximately 48% of the global exports of soybeans [FAO, 2021].

2 Methods

The study structure is divided in two main blocks: The first block is the climate-crop interaction (Figure 1). We designed statistical models linking climate information from GCMs, simulated yields from GGCMs, extreme climate indices and observed soybean data. Three models were developed, all based on the random forest (RF) algorithm: a RF trained only on the crop model, a RF trained only on the extreme climate indices, and both crop model and extreme climate indices combined, the hybrid model. The second block is the socio-economic impacts estimation, which is dedicated at using the GLOBIOM model to explore the possible socio-economic consequences of extreme events. For that, we combine long-term trends of EPIC with the interannual variability of the hybrid model for better estimation of weather shocks.

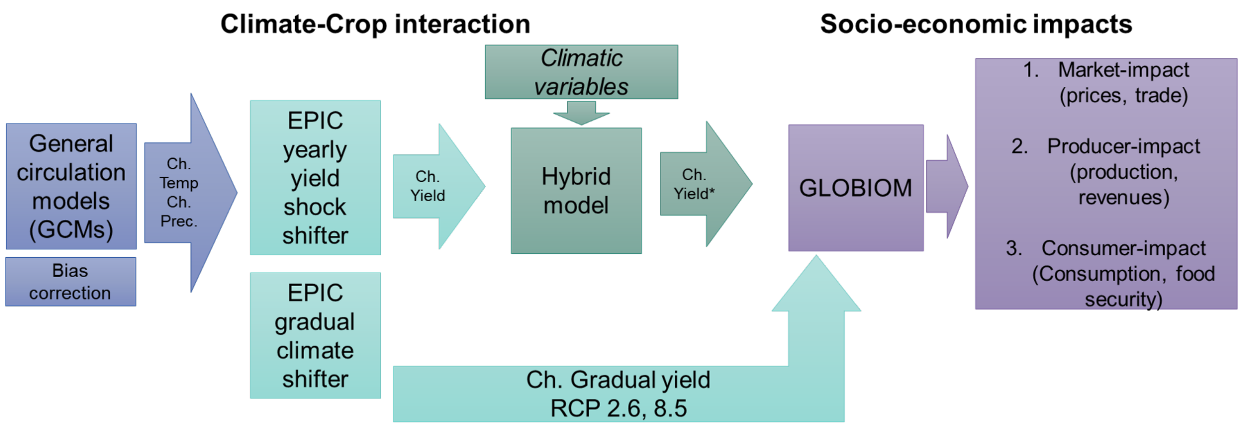


Figure 1: Study workflow from weather shocks to socio-economic impacts.

2.1 Climate-crop interaction

To represent the interactions between climate and crop, we used the physical processing of crop models, statistical relations of extreme climate indices, and the combination of them to explore the biophysical impacts of extreme climate conditions on soybeans.

Supervised machine learning models require two types of input: the independent and dependent variables. Independent variables are the ones that the model is trained on. Dependent variables are the ones that the model uses for prediction and calibration. For the independent variables, we considered the crop model outputs and the extreme climate indices, depending on the model considered. For the dependent variables, all models had as dependent variable a dataset of historical observed soybean yields.

For the creation of the climate-crop models, we followed a two-step approach based on Feng et al. [2019] to generate there climate-crop models. For the first one, we trained the statistical model only on the crop model simulated soybean yields (referred to as RF:EPIC), which worked as a bias correction mechanism. The other model was trained using the extreme climate indices alone, and no physical processing was involved (RF:CLIM). The third model, we fed the simulated yield output of the crop model (first step) into a random forest model (second step) together with the extreme climate indices, generating the hybrid model (RF:Hybrid).

The machine learning model used to train the clima-crop models is the random forest (RF) model. Random forest [Breiman, 2001] is a non-linear and non-parametric statistical model for classification and regression, being one of the most commonly used models in different fields of science. It benefits from an ensemble structure, containing multiple independent decision trees, which are each trained on random sub-samples of the data to provide different predictions. All predictions are then combined to provide a unified output. RFs do not have many parameters to be tuned, do not suffer from overfitting, and are known for high performance [Breiman, 2001]. They rank among the best classifiers for real world problems [Fernández-Delgado et al., 2014], and in the field of agriculture **?** and Wang et al. [2018] found RFs to be superior to other machine learning techniques. Finally, RFs are compatible with nonlinear relationship, which is considered relevant for extreme climate indicators that have nonlinear behaviour [Schlenker and Roberts, 2009, Feng et al., 2019].

The first step to train a random forest model was the feature selection. The feature selection consisted of two parts: First to define which individual months along the year should be selected; second which variables to use for soybean prediction. The soybean growing season encompasses multiple months, and it is necessary to select the extreme climate indices during the most relevant months of the year. The experiment is spatially gridded across large areas of Brazil, which implies a heterogeneity in seasons. Each grid cell has specific growing season properties. We adopted a planting calendar to dynamically define which months to be considered. The calendar is based on a dataset of planting windows for soybeans [Abrahão and Costa, 2018].

Among the extreme climate indices generated, some variables are more relevant than others for soybean growth, and because the indices are based on temperature and precipitation values, high correlation between some of them is expected. Irrelevant variables and multiple redundancy could add noise and decrease performance of the model. It is beneficial to the training of the model to select the most important features without high levels of correlation. We used the internal random forest feature selection module [Breiman, 2001] to select the most important features for the entire region studied. At the end of this step, the most important features for soybean growth were selected, with the temporal component dynamically defined based on each grid cell specific crop calendar.

We used a cross-validation scheme to tune the RF's hyper-parameters. We divided the dataset in 10 splits, where on is left out for validation and the rest is used to train. The process is repeated 10 times, and then 5 times more under different configurations of the split. The selected configuration for the RF is shown in Table 1.

Parameter	Description	Value
n_estimators	Number of trees in the model	200
criterion	Quality of split function	Mean Squared Error
max_features	Maximum features per split	Square root of total features
max_depth	Maximum tree depth	10

Table 1: Random forest model configuration of hyper-parameters.

We use the coefficient of determination (R^2, Equation 1), mean absolute error (MAE, Equation 2) and root-mean-square error (RMSE, Equation 3) regression loss functions to evaluate and compare the different random forest models. R2 quantifies the amount of variability explained by the model; MAE indicates the average error between the values predicted by the model and the real corresponding values, and RMSE quantifies the standard deviation of errors.

$$R^2 = \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \tag{1}$$

$$MAE = \frac{\sum |(y_i - \hat{y}|}{n} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n}} \tag{3}$$

2.2 Socio-economic exploration

Studies into the dependent demand and supply relationships in globalised systems require a framework that can take stock of both the climate-induced deviations between expected and observed prices and yields, and the impacts on the food commodity market. The GLOBIOM model [Havlík et al., 2014] is used to analyze the effects for producers, consumers and market stability of climate-induced yield shocks. We combine the yield data from the first part as input to the model.

For this study, GLOBIOM is enhanced to be able to deal with extreme fluctuations in yields. To allow for these extreme events, a “short run” response to yield shocks is implemented by limiting the possible production response to the shock. These limitations to the production responses include the restriction of land reallocation per sector for all land use sectors to reflect short-term adjustments and the reduction of possibilities for substitution between land and other inputs for crops. More specifically, GLOBIOM is adjusted to allow for a short-run response in the years 2030, 2050 and 2070, when a production shock is modelled. Up till these time-steps, GLOBIOM is run normally with the exogenous model parameters, and land conversion and conversion of land use is set to reflect 10-year periods. Subsequently, we run the year 2030, 2050 and 2070 again and adjust exogenous model parameters and land conversion and conversion of land use in such a way that they reflect a yearly response. In the case of GLOBIOM, exogenous model parameters include population, GDP, and income elasticities. Land conversion and conversion of land use include coefficients for maximum and minimum expansion and reduction of crops and livestock activities and management changes. In the year of the shock, these coefficients are set to reflect yearly adjustments. The reasoning behind these model adjustments is that the extreme events that are implemented are selected to happen only once with this order of magnitude, and not 10-years in a row, which is the normal time-step of GLOBIOM. At the same time, GLOBIOM is a linear optimization model, meaning that, without setting certain bounds, the effect of introduced shocks may lead to corner solutions that might overestimate the true magnitude of the climate-driven land abandonment. It is therefore of importance to parameterize the cost of land conversion correctly.

Because the socio-economic analysis required both long-term trends of climate change and interannual variability, we combined the climate-induced yield impact of gradual change produced by EPIC-IIASA with the yield impact induced by weather variability produce by the Hybrid model, and both are implemented in GLOBIOM. For the historical baseline (1986-2015) and each 30-year time-slice for decades between 2000 and 2070 a growth rate is computed for each crop at the half degree resolution for Brazil and the 2 degree resolution for the rest of the world. This growth rate captures the yield changes that are due to the gradual change in climate. Extreme yield losses are implemented around the year 2030, 2050 and 2070. For each of the 30-year time slices we compare the difference between the annual yield and the mean yield level of that time slice. As the focus is on the extremes and inter-annual variability, we select from every GCM-RCP-CO2 combination, and for each time slice, the year which represents the largest negative deviation between the yield weighted by area of soy bean of that year and the weighted average yield of that time slice.

Among the outputs that GLOBIOM has available, we considered four metrics to assess the global market dynamics and food security: the national production of soybeans in Brazil (Prod), the net exports (NETT), the soybean prices per ton (XPRP) and the daily calories available per capita (CALO). With these, we compared both the impacts of the weather shocks, but also the capacity of the socio-economic feedback loops to mitigate, propagate or amplify the bio-physical impacts.

3 Data

For the development of the models to represent the climate-crop interaction, three datasets were adopted: yearly values of simulated soybean yields by EPIC-IIASA (input); extreme climate indices at a monthly resolution (input), and yearly values of historical observed soybean yields (the dependent variable or output).

The adopted crop model is the global gridded crop model (GGCM) EPIC-IIASA [Balkovič et al., 2014], based on the Environmental Policy Integrated Climate [EPIC; Williams et al., 1995] field-scale crop model. It reproduces biophysical processes in the soil-plant-atmosphere system and provides crop-related outputs at a resolution of 0.5° x 0.5° based on climatic-related inputs. The simulations of EPIC-IIASA belong to the phase 3a of the Intersectoral Impact Model Intercomparison Project (ISIMIP; see https://isimip.org for details and protocols) and the Global Gridded Crop Model Intercomparison (GGCMI) initiative. To account for scenario uncertainty, EPIC-IIASA is available for multiple global climate models (GCMs), representative concentration pathways (RCPs) and different CO^2 fertilization scenarios. For the training and validation of the models, we used the reanalysis version of EPIC-IIASA, based on the GSWP3-W5E5 dataset, a combination of the GSWP3 (Global Soil Wetness Project phase 3) dataset [Dirmeyer et al., 2006] with the W5E5 dataset [Lange, 2019], which ranges from 1901 to 2016. For the global warming exploration, we used three different GCMs, UKESM1-0-LL, GFDL-ESM4 and IPSL-CM6A-LR, two RCPs, 2.6 and 8.5, and two setups for CO2 fertilization: Static CO^2 based on 2015 levels, and dynamic CO^2 levels, all of which range from 2015 to 2100. More details of the EPIC-IIASA model's performance can be seen in Müller et al. [2017].

We generated indices of climate extremes based on daily values of precipitation, maximum and minimum temperature from the same climatic datasets used for each EPIC-IIASA scenario considered in this work. The indices are monthly statistics and can be seen in table 2. The R package used to generate the indices is called Climpact [Alexander and Herold, 2015].

As dependent variable for the training of the statistical models, we used observed soybean yields for the models to be calibrated on. We used data from the Brazilian Institute of Geography and Statistics (IBGE) to build a 1 km^2 resolution soybean yield dataset from 1981 to 2016 for Brazil. It provides census data at the municipality level for the entire country, but we postprocessed the data to remove municipalities that had less than 1% of soybean harvest area.

The datasets presented incompatibilities between one another and therefore required a step of regularisation before they could be processed together. While the EPIC-IIASA outputs and the extreme climate indices have resolution of 0.5° x 0.5°, the spatial resolution of the observed soybean dataset is below 0.001°. Therefore, we upscaled the spatial resolution of the observed soybean dataset to the same as the input data, at 0.5° x 0.5°. On the other hand, the crop model and extreme climate indices cover the entire country, including regions where soybeans are not grown. We masked the growing regions based on the observed yield dataset, so all three datasets covered the same amount of territory. Figure 2 shows the resulting regions of the country considered for this study. Finally, all datasets had temporal trends. The observed yield dataset has management and technology included implicitly; the simulated crop model considers atmospheric CO2 concentration levels in the biomass growth calculation,

Index	Description	Units
FD	Days when minimum temperature is below 0°C	days
TNlt2	Days when minimum temperature is below 2°C	days
TNltm2	Days when minimum temperature is below -2°C	days
TNltm20	Days when minimum temperature is below -20°C	days
ID	Days when maximum temperature is below 0°C	days
SU	Days when maximum temperature exceeds 25°C	days
TR	Days when minimum temperature exceeds 20°C	days
TXx	Hottest day	°C
TNn	Coldest night	°C
TXgt50p	Fraction of days with above average temperature	%
TMge5	Days when average temperature is at least 5°C	days
TMlt5	Days when average temperature is below 5°C	days
TMge10	Days when average temperature is at least 10°C	days
TMlt10	Days when average temperature is below 10°C	days
R20mm	Days when rainfall is at least 20mm	days
PRCPTOT	Total wet-day rainfall	mm
DTR	Average range of maximum and minimum temperature	°C
TNx	Hottest night	°C
TXn	Coldest day	°C
TMm	Average daily temperature	°C
TXm	Average daily maximum temperature	°C
TNm	Average daily minimum temperature	°C
TX10p	Fraction of days with cool day time temperatures	%
TX90p	Fraction of days with hot day time temperatures	%
TN10p	Fraction of days with cold night time temperatures	%
TN90p	Fraction of days with warm night time temperatures	%
R10mm	Days when rainfall is at least 10mm	days
R95p	Amount of rainfall from very wet days	mm
R99p	Amount of rainfall from extremely wet days	mm
Rx1day	Maximum amount of rain that falls in one day	mm
Rx5day	Maximum amount of rain that falls in five consecutive days	mm

Table 2: Extreme climate indices at monthly resolution, their plain language descriptions and the corresponding units.

a process called CO2 fertilisation [Deryng et al., 2016, Toreti et al., 2020]; The extreme climate indicators vary according to the rising global temperature levels due to climate change. We detrended all datasets following a polynomial regression, removing influences of technology, management, global warming and CO^2 concentration levels. Therefore, only the interannual variability is present on the historical time series.

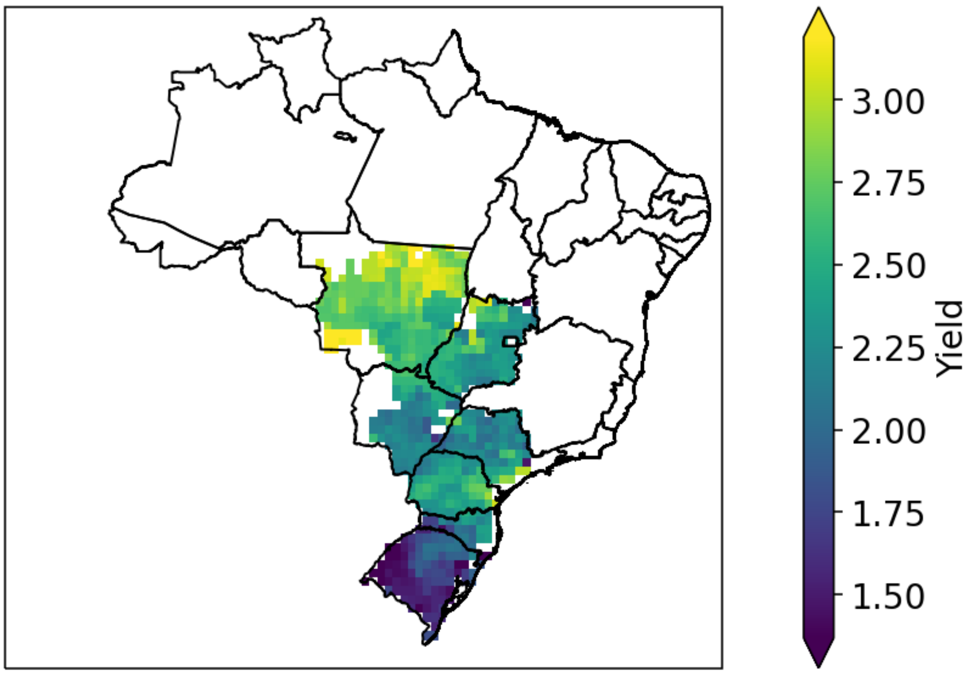


Figure 2: Mean observed soybean yields from 1980 to 2016.

For the socio-economic analysis, we adoptes the GLOBIOM [Havlík et al., 2011, 2014] model. It is a partial equilibrium model that covers the agricultural and forestry sectors, including the bioenergy sector. Commodity markets and international trade are represented at the level of up to 58 economic regions. The spatial resolution of the supply side relies on the concept of Simulation Units (SimU), which are aggregates of 5 to 30 arcmin pixels belonging to the same altitude, slope, and soil class, and also the same country (Skalský et al. 2008). For crops, grass, and forest products, Leontief production functions covering alternative production systems are parameterized using biophysical models like EPIC [Williams et al., 1995], G4M [Kindermann et al., 2008, Gusti, 2010], or RUMINANT [Herrero et al., 2013]. The biophysical models allow a precise calculation of agricultural GHG emissions (N2O and CH4). For this study, Brazil was singled-out as a separate GLOBIOM region and run at the half-degree resolution, whereas the other regions were run at a 2 degree resolution.

4 Results

In this section, we present the selection of features (Section 4.1) and the performance of the three random forest models (Section 4.2). Then we show the projections for different scenarios of global warming with and without shocks (Section 4.3). Finally, in Section 4.4 we demonstrate the socio-economic impacts.

4.1 Feature selection

Following the methods section, the first step in building the random forest models is the feature selection of the extreme climate indices. Along the soybean growing season, we dynamically select the three months following the latest planting dates at each grid cell (based on photoperiod) provided by the dataset of planting windows for soybeans by Abrahão and Costa [2018]. Each extreme climate indicator is therefore divided in three subsequent monthly values. Figure 3 shows the last planting date for each cell grid in the region studied, used as reference for the calculation of the subsequent three months. Selecting the months of extreme climate indices dynamically shows an increase of 0.1 R^2 with respect to the usual static approach.

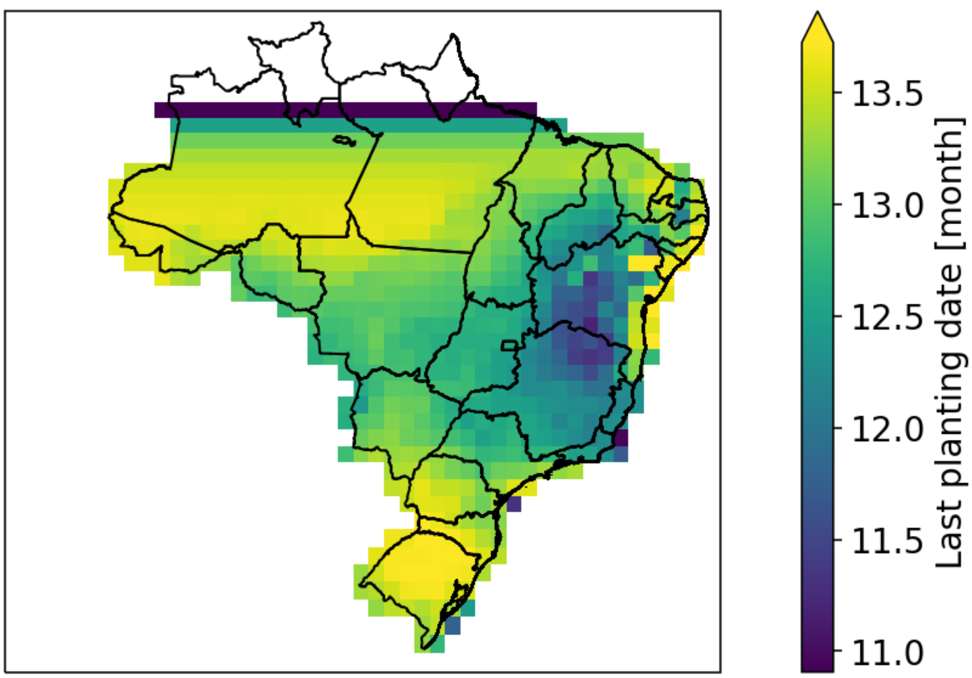


Figure 3: Last planting dates in months for the region studied. Months above 12 belong to the following year.

Using the internal feature selection module of the random forest model, the most important features according to the feature selection methods are summarised in Table 3. According to the feature selection, the soybean yields in the region studied are mostly dependent on precipitation-related indices. The only temperature related indices that obtained a high classification was diurnal temperature range.

Position	Feature
1	prcptot_2
2	prcptot_3
3	r10mm_2
4	prcptot_1
5	r10mm_3
6	dtr_1
7	r10mm_1

Table 3: Most important features in the internal feature selection module of the Random Forest model.

The use of partial dependence plots allows a examination of the interaction between each of the input variables with the output variable. In this case, as seen in Figure 4, high values of diurnal temperature range in the first month after the planting date are correlated with lower yields. Precipitation values are positively correlated with soybean yields, with a special case of less than 200 mm/month of water during the second month after planting leading to extremely low levels of soybean yield. Absolute levels of temperature seem to be less influential on yields, with maximum night temperatures being inversely proportional to the yields.

4.2 Models evaluation and comparison

The three random forest models developed in this study, RF:EPIC, RF:CLIM, RF:Hybrid, are evaluated and compared to evaluate the added value of the physical crop model information and of the extreme climate indicators. As shown in Table 4, the three models are analysed out of sample (no leakage of data) according to three different scores: the coefficient of determination (R^2), mean absolute error (MAE) and root-mean-square error (RMSE). The model based only on soybean yields of EPIC-IIASA, RF:EPIC, shows significantly lower scores than the other two models. While roughly similar, the RF:Hybrid model shows a subtle improvement with respect to the RF:CLIM, demonstrating the highest scores in all criteria. Therefore, the addition of extreme climate indicators to the crop model is shown to highly increase the the overall performance of the climate-crop interaction.

Score	RF:EPIC	RF:CLIM	RF:Hybrid
R^2	0.14	0.67	0.69
MAE	0.359	0.221	0.217
RMSE	0.470	0.296	0.291

Table 4: Out of sample performance scores on the coefficient of determination (R^2), mean absolute error (MAE) and root-mean-square error (RMSE) for the three random fores models.

The simulations of each model for the entire historical period (1980-20015) is shown in Figure 5 (not out of sample). Using the historical observed soybean series as reference, we demonstrate the increase in performance that the addition of the extreme climate indices

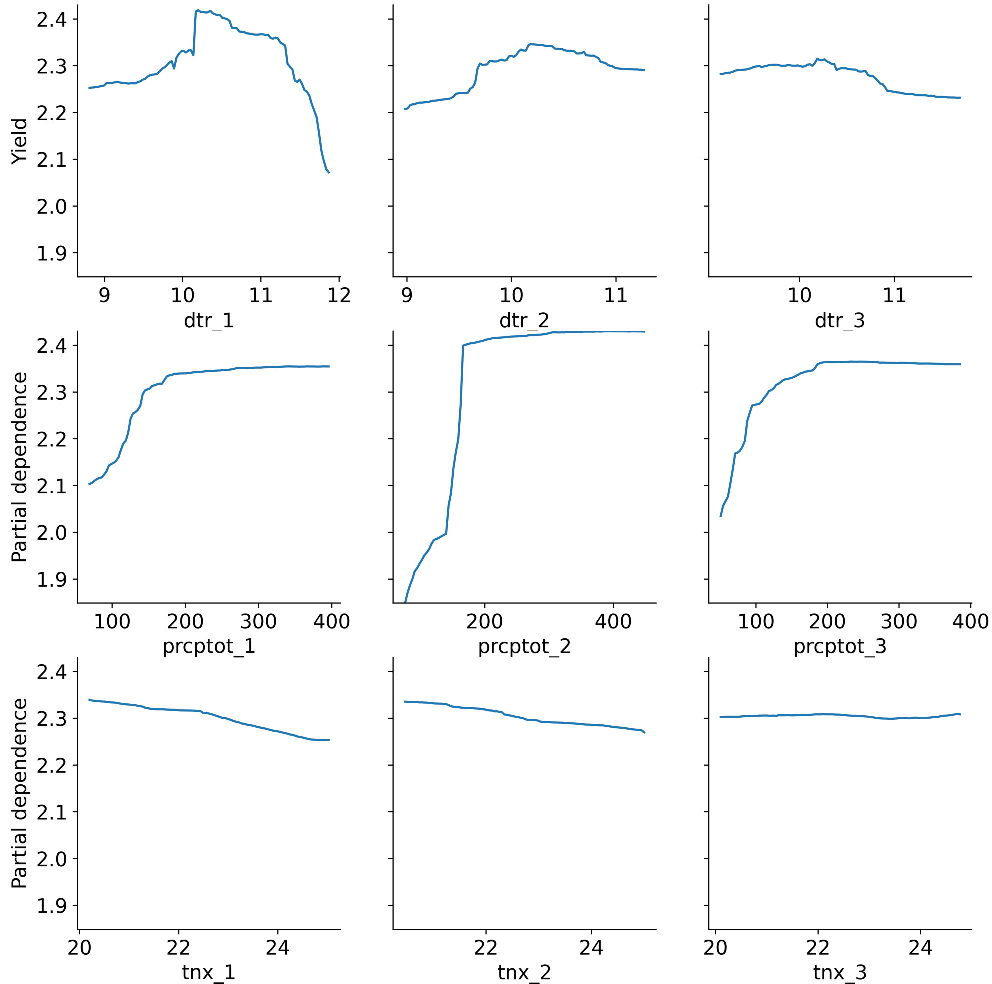


Figure 4: Partial dependence plots between individual extreme climate indices and the observed soybean yields. Numbers 1, 2, 3 indicate the relative months after the last planting month for each grid cell.

bring to the crop model. This is highlighted in years of extreme low yield, such as 1986, 1991, 2005 and 2012, in which the RF:Hybrid and RF:CLIM are capable of replicating the deep falls in soybean yields seen in the observed dataset.

In addition to the overall performance of the models, we compare their behaviour under extreme climate conditions. Based on the previous figure, two years are selected: 2005 and 2012. Figure 6 suggests that EPIC-IIASA outputs can be improved under extreme conditions when they are fed to a random forest model calibrated on observed yields, working similarly to a bias correction technique. Yet, the spatially explicit errors observed in the figure suggest an even greater improvement when EPIC-IIASA is combined with extreme climate indices, generating hybrid model.

The addition of extreme climate indices to the simulated yields of EPIC-IIASA is shown to be beneficial for the overall performance, with all score metrics improving significantly, and an increased interannual vairability. More specifically, for extreme years in which the original EPIC-IIASA model underestimates the impacts, the complement of extreme climate indices

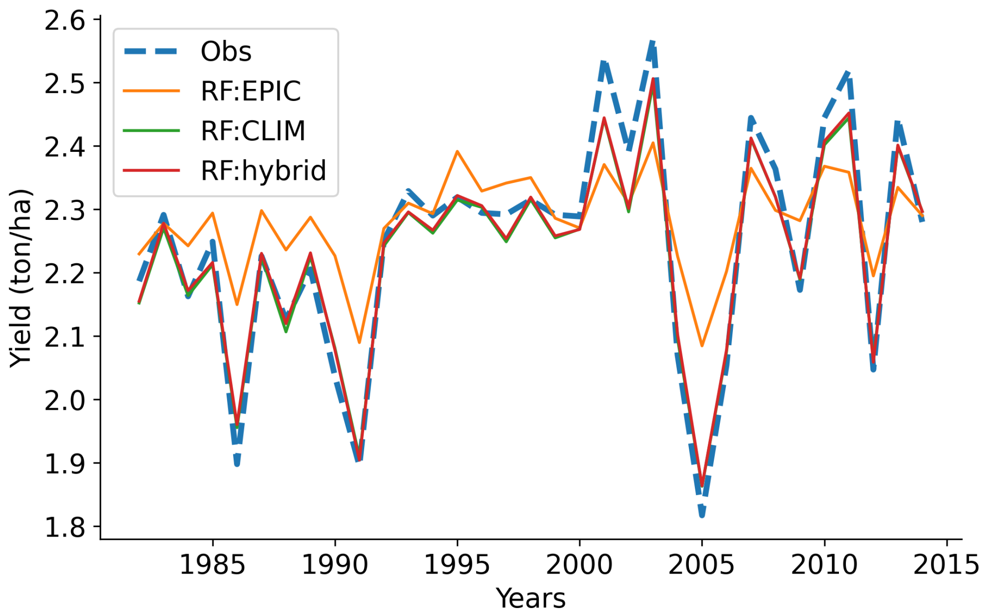


Figure 5: Weighted time series of soybean yields in the region studied for the observed soybean data as reference, and the simulated soybean yields for the RF model trained solely on the EPIC-IIASA output (RF:EPIC), the RF model trained on the extreme climate indices (RF:CLIM), and the hybrid model (RF:Hybrid).

approximates the yield losses to similar levels of the historical observed dataset. The hybrid model shows a slight improvement with respect to the model trained on extreme climate indicators, but both models have overall similar performances. Therefore, we consider the creation of the hybrid model to be a success case of improved performance, with added value for extreme conditions representation.

4.3 Future projections

For the assessment of socio-economic impacts, we combine the long-term projections of EPIC-IIASA yields at a decadal resolution with the interannual variability produced by the hybrid model. For this study, the scenarios are built on three GCM models: UKESM1-0-LL, GFDL-ESM4 and IPSL-CM6A-LR and two RCP levels: 2.6 and 8.5. These GCMs and RCPs are selected to ensure maximum spread of possible results (Figure 7). The CO2 scenario adopted is the static levels of 2015 CO^2 concentration in the atmosphere.

Long-term trends of the projections are based on the original values of EPIC-IIASA. They indicate the decadal trends of soybean yields simulated by the crop model. Among the three GCMs tested, the lower emission scenario indicates an average increase in the productivity of soybeans for Brazil until the end of the century (Figure 8 left). The higher emission scenario, however, is expected to stabilise the yields around 2030 and to slightly suffer a decrease in the second half of the century. When separating the projections by GCM to account for model uncertainty (Figure 8 right), the GFDL model shows the highest increase in soybean yields for the future. IPSL shows a stabilisation in yields from 2050. UKESM is the GCM

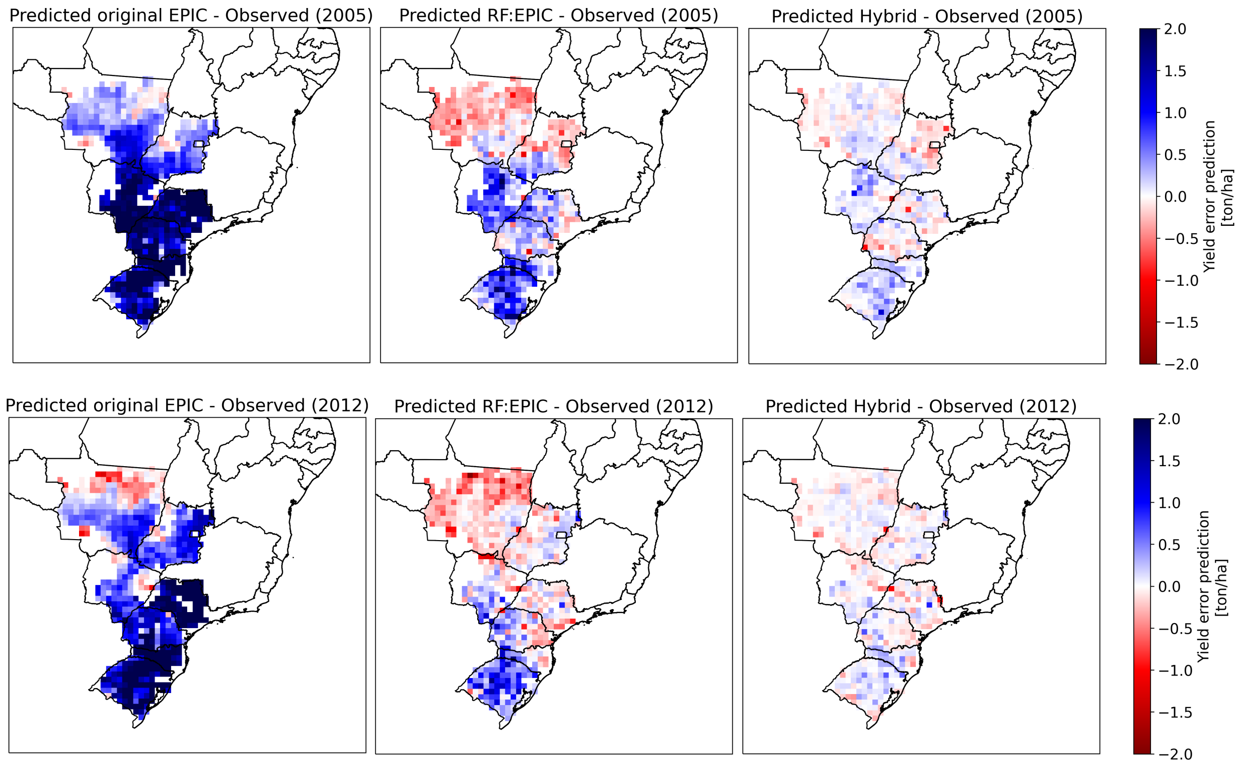


Figure 6: Comparison of residuals for two extreme years of the original EPIC-IIASA crop model (left), the RF model trained solely on the EPIC-IIASA output (RF:EPIC, center) and the hybrid model (right). Extreme years considered are 2005 (top panels) and 2012 (lower panels).

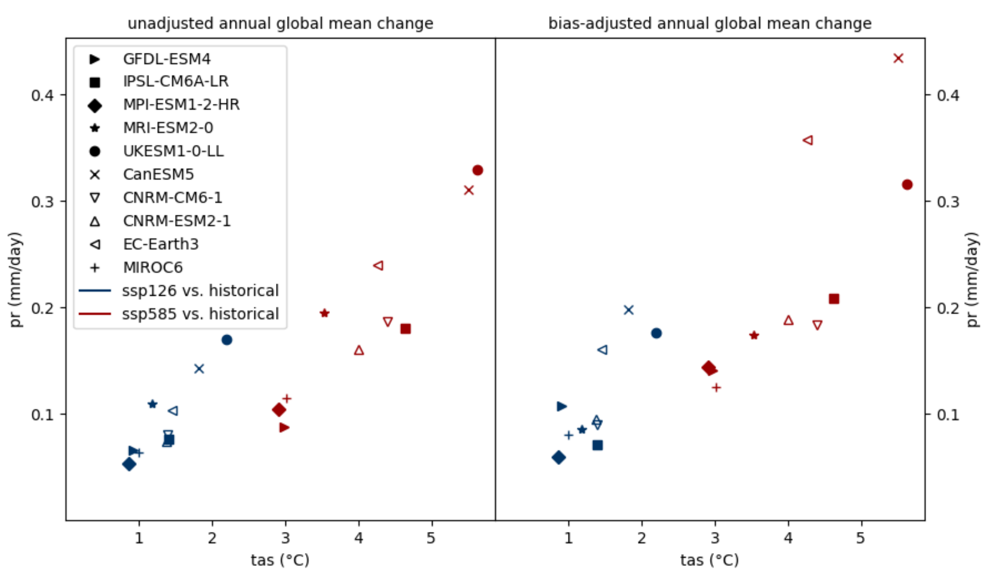


Figure 7: Spread of different bias-corrected GCMs. Source: ISIMIP.

with lowest yields values, stabilising around 3.25 ton/ha from 2040 until 2060, to slightly decay afterwards.

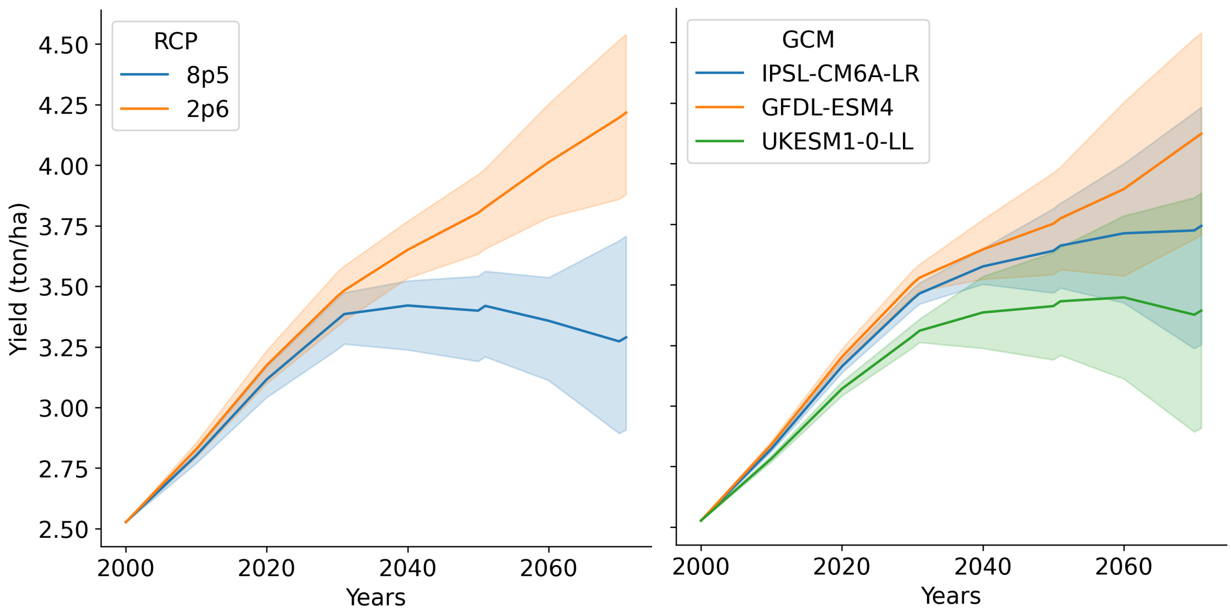


Figure 8: Long-term trends based on EPIC-IIASA. Left: projections grouped by RCPs (2.6 and 8.5); Right: Projections grouped by GCM (IPSL, UKESM, GFDL).

With the training of the random forest models on historical settings, it is possible to explore the interannual variability of future projections of soybean yields under different scenarios of climate change. A scenario with higher emissions (RCP 8.5) indicates that, across all GCMs, the mean yield outputs are expected to be lower when compared with a lower emission scenario (RCP 2.6), as shown by Figure 9. The intensity of the shocks is also expected to increase with the intensity of the emissions. Last, there is a temporal pattern only seen for RCP 8.5, where shocks become more intense and frequent closer to the end of the century (Figure 9).

The socio-economic analysis uses a combination of the long-term trends of EPIC-IIASA with the interannual variability of the hybrid model. More specifically, we select the lowest yields for periods of 30 years and add them to the long-term trends, effectively measuring the impacts of shocks on global markets and food security. The yield shocks are presented in Figure 10. As a consequence of a higher natural variability in RCP 8.5, the shocks in this scenario are more intense than the ones in RCP 2.6, reaching up to -30% of the soybean yields in Brazil (Figure 10 upper left). Among the GCMs, there are also differences, with IPSL having the least extreme shocks, UKESM the most extreme shocks, and GFDL around -15% of the yearly soybean yields (Figure 10 upper right). Finally, comparing the time blocks, there are different dynamics for each RCP. For RCP 2.6, the magnitude of the shocks decreases with time (Figure 10 lower left). On the other hand, for RCP 8.5, the shocks increase for the 2050 time block, but decrease for the last time block (Figure 10 lower right).

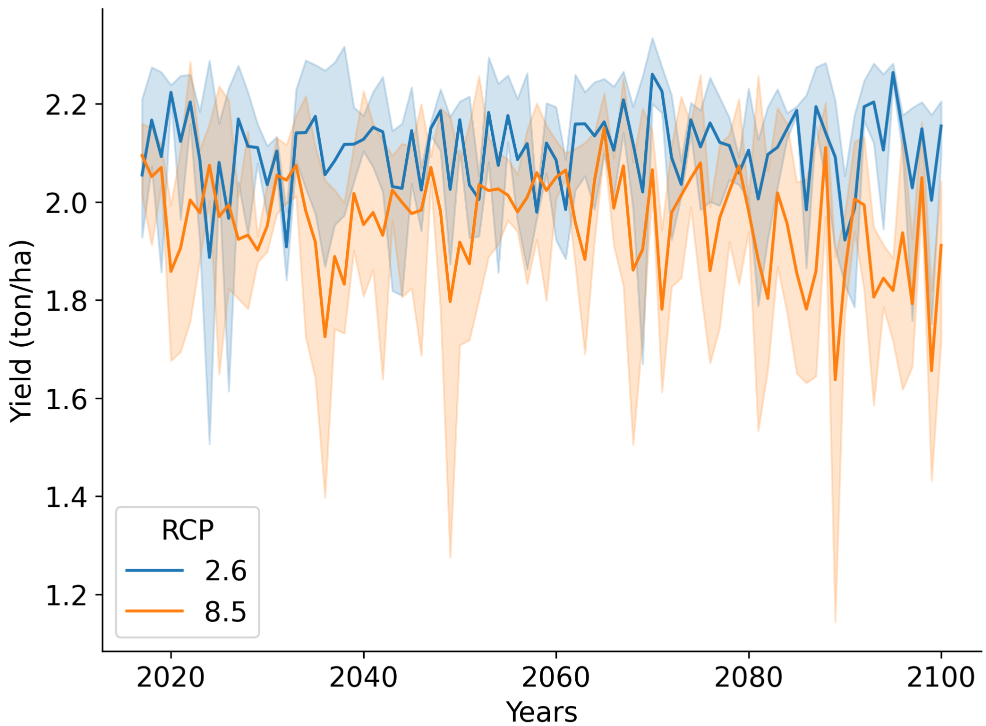


Figure 9: Hybrid model estimated projections of soybean yields for three GCMs (IPSL, UKESM, GFDL) and divided between the RCPs 2.6 and 8.5.

4.4 Socio-economic analysis

For a full analysis of the climate shocks on the soybean global market, we include a socio-economic analysis. Among the socio-economic components, we consider four metrics: the national production of soybeans in Brazil (Prod), the net exports (NETT), the soybean prices per ton (XPRP) and the daily calories available per capita (CALO).

The total production of soybeans has shown negative impacts for all RCP scenarios and blocks of time. The percentage changes for production in RCP 2.6 show a decrease in intensity, as seen for the yield metric above. For the RCP 8.5, the shocks do not seem to change in time, with percentage changes around -15%. The net exports of Brazil follow a similar behaviour, however in RCP 8.5, a slight increase in the percentage change is observed. The soybean prices follow a volatile behaviour, with prices increasing around 130% after shocks in the early century, but reaching up to 300% in RCP 8.5 late century. Calories deficits are observed for every scenario at similar levels, except for the late century, where RCP8.5 shows a much higher drop in calories per capita compared to RCP2.6.

The socio-economic metrics evaluated here indicate a non-linearity component. While the total national production of soybeans is highly correlated to the soybean yields, showing similar behaviour for the same shocks, as there are no changes in area or management in time for the shocks, other metrics show larger differences. Net exports have similar dynamics to the production, which can be partially explained to the fact that soybeans are mostly used for exports. The prices per ton of soybean show less similarities to the yield shocks. The prices have a higher sensitivity to the fluctuations in yields, and amplify the effects of the

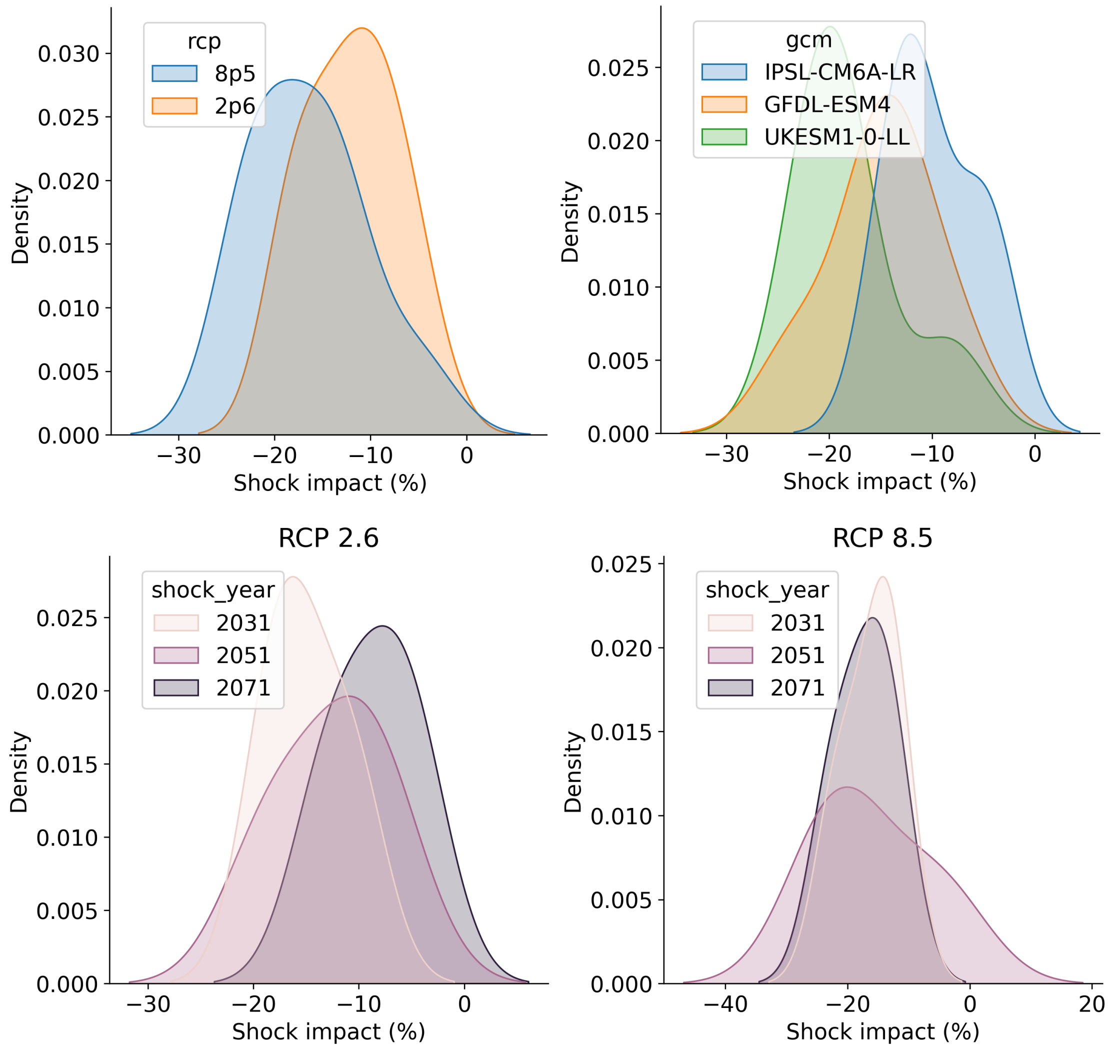


Figure 10: Shock impacts on soybean yields for different RCPs (2.6, 8.5, upper left), GCMs (IPSL, UKESM, GFDL, upper right) and time periods (2031, 2051, 2071, lower panels).

shock significantly. Calories, on the contrary, show similar variation with respect to the yield shocks. The market effects augment the shocks whereas the effects on the consumption and calories are proportional to the shock. This is probably due to the rather inelastic nature of food demand.

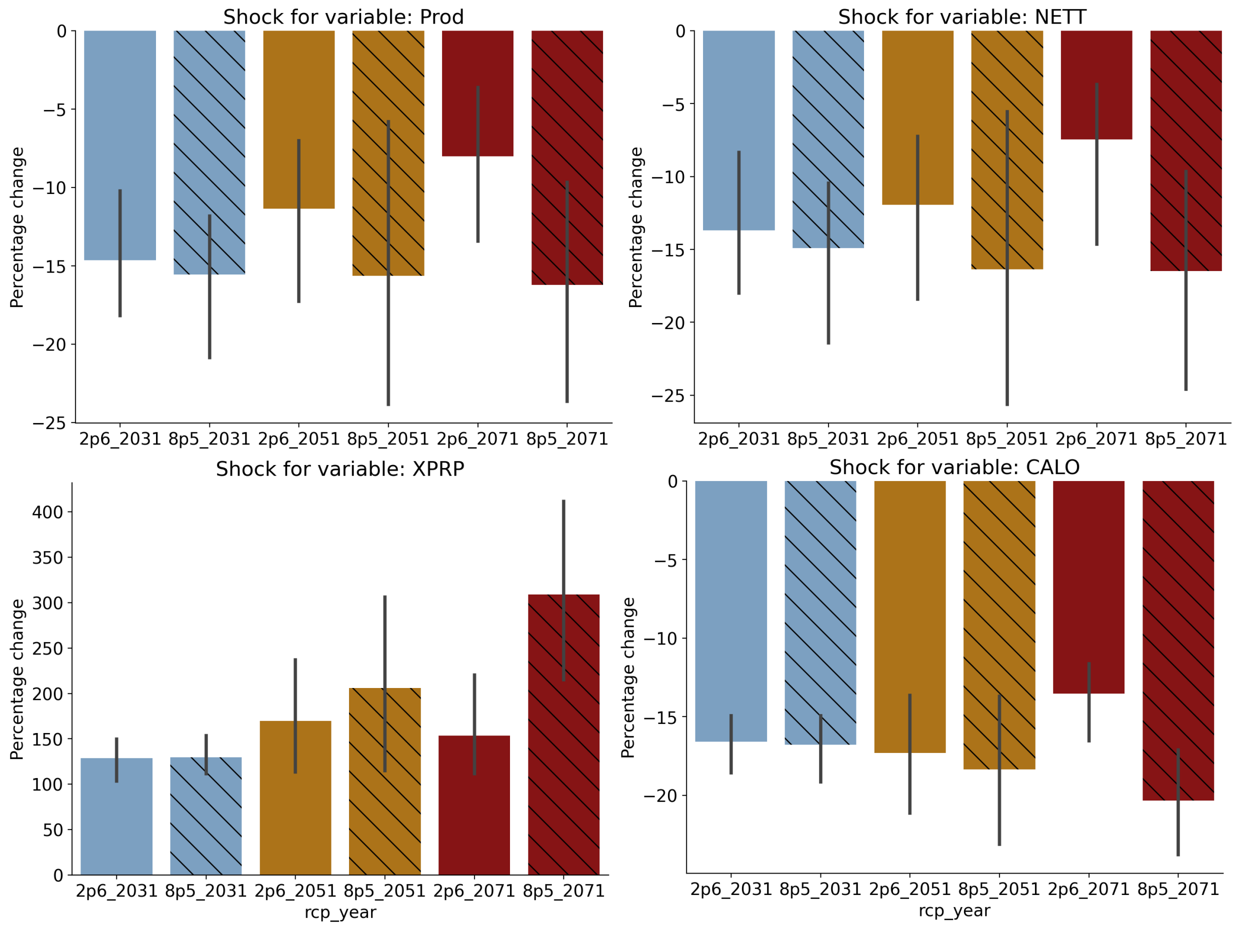


Figure 11: Socio-economic impacts of climate shocks on four different socio-economic metrics: the national production of soybeans in Brazil (Prod, upper left), the net exports (NETT, upper right), the soybean prices per ton (XPRP, lower left) and the daily calories available per capta (CALO, lower right). Hashed indicates the shock corresponds to the RCP 8.5 and confidence interval (black vertical lines) represents the difference between GCMs.

5 Discussion

While useful for the estimation of long-term trends in climate change studies, process-based crop models have shown to underperform under extreme climatic conditions [Schewe et al., 2019, Feng et al., 2019]. We argue that the creation of a hybrid model can improve the reproduction of the interannual variability of soybean yields. The addition of extreme climate indices to the simulated yields of a crop model combined in a statistical model shows an increase in overall performance and reduction of errors. The hybrid model is particularly skilled at representing extreme low yields, a known limitation of crop models [Schewe et al., 2019, Feng et al., 2019]. Our results are in line with previous works that have shown the utility of combining different models into a hybrid to improve performance [Roberts et al., 2017, Feng et al., 2019].

The hybrid model is based on the machine learning model Random Forest, as previous works have shown RFs are among the best machine learning models for agricultural studies Feng et al. [2018], Wang et al. [2018]. To determine the most relevant extreme climate indices, we use the internal feature selection of the Random Forest model. The set of selected extreme climate indices suggest precipitation-related indices are the most relevant for soybean production in the region analysed. The indices are relatively assigned based on the last planting date for each grid cell, following a dynamic calendar. We note that the inclusion of dynamic planting dates has improved the model performance with respect to static planting dates.

Among the multiple global warming scenarios tested, we find differences between the RCPs and GCMs tested. Early in the century, the differences are dominated by the GCMs. The emissions scenarios become more influential on soybean yields later in the century, but early in the century they do not seem to affect significantly. These results are in agreement with recent findings [Müller et al., 2021]. We claim interannual variability is also relevant for this type of analysis. The long-term trends of soybean yields tend to increase with time according to the projections. However, the shocks are also posed to increase with time, especially under higher emission scenarios, and they could lead to large setbacks in soybean production.

By explicitly comparing the long-term trends with individual shocks on a socio-economic impact model, we are able to measure the significance of shocks in the global supply chain of soybeans due to weather variability at the local production sites. Shocks due to weather variations can carry non-linear consequences to different socio-economic metrics. The prices of soybean are amplified by market dynamics, reaching up to 20 times the percentage change of the yield shock. On the other hand, the caloric deficit per capta shows a similar variability with respect to the yield shocks. So, following a chain of events where the production affects the prices and then the consumption, prices react in a very elastic way (high price response), but caloric deficit reacts in a very inelastic way because of the primary nature of the good.

This work is limited to the analysis of extreme climate indices at a monthly resolution, and does not consider irrigation practices, sub-surface conditions, nor CO2 fertilisation, which are relevant components for impact assessment [Schlenker and Roberts, 2009, Deryng et al., 2016, Toreti et al., 2020]. We also limit the analysis to Brazil, which is the main producer and exporter of soybeans. However, for a more thorough investigation into the global soybean markets, it would be an interesting addition to have Argentina and the United States included in tha analysis. We recommend also looking further down the supply chain into the bilateral trade of soybeans and the effects on the market for meats.

6 Conclusion

Soybeans are highly significant for the modern world, being crucial for multiple industries, such as cattle, vegetable oil and human consumption. Its global consumption requires a complex supply chain for soybean to be distributed and consumed. Yet, the majority of its production is concentrated in a few hotspots. This imbalance means disturbances in the producing sites could lead to impacts at a planetary scale, and a common source of disturbances in the agriculture field is weather variability. In this work, we explore the vulnerability of the entire supply chain of soybeans on extreme weather events at the local production sites. The chain of events linking weather events to soybean production and to global markets is complex and involves many steps. The framework adopted in this work consists of combining bio-physical and socio-economic analyses. For the bio-physical part, we generate a hybrid model, which considers process-based crop yields and extreme climate information for a better representation of interannual variability and extreme low yields. For the socio-economic analysis, we feed the hybrid model outputs under different climate change scenarios into a socio-economic impact model.

The combination of the crop model with extreme climate information is shown to significantly improve the overall performance of the model, with an increased interannual variability representation. For extreme low yield years, the hybrid models is better at capturing the impacts of extreme climatic conditions. We find that low yields in the region analysed are mainly driven by low precipitation values. The global warming projections suggest weather shocks are likely to become more frequent and intense, especially for high-emission scenarios.

The socio-economic impacts demonstrate a non-linear relationship to the bio-physical impacts. Market-related impacts, such as soybean prices, show an exacerbation of the shocks, possibly due to the market dynamics. Food security impacts, such as calorific deficit, on the other hand, show less variation, which could be a consequence of its main purpose being rather inelastic in terms of consumption.

References

FAO. Food and Agricultural Organization of the United Nations Statistical Database. Technical report, 2021.

M Zampieri, A Ceglar, F Dentener, and A Toreti. Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. *Environmental Research Letters*, 12(6):064008, jun 2017. ISSN 1748-9326. doi: 10.1088/1748-9326/aa723b. URL http://dx.doi.org/10.1038/ncomms2296{%}0Ahttp://10.0.4.14/ncomms2296{%}0Ahttps://www.nature.com/articles/ncomms2296{#}supplementary-information{%}0Ahttp://dx.doi.org/10.1016/j.jeem.2017.06.002https://linkinghub.elsevier.com/retrieve/pii/S0095069617303650https:/.

Jakob Zscheischler, Rene Orth, and Sonia Seneviratne. Bivariate return periods of temperature and precipitation explain a large fraction of European crop yields. *Biogeosciences Discussions*, (February):1–18, 2017. ISSN 1726-4170. doi: 10.5194/bg-2017-21.

Tamara Ben-Ari, Julien Boé, Philippe Ciais, Remi Lecerf, Marijn van der Velde, and David Makowski. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications*, 9(1), 2018. ISSN 20411723. doi: 10.1038/s41467-018-04087-x. URL http://dx.doi.org/10.1038/s41467-018-04087-x.

Siddhartha Mitra and Tim Josling. Agricultural export restrictions: Welfare implications and trade disciplines. *Position paper agricultural and rural development policy, IPC*, 2009.

Linde Götz, Thomas Glauben, and Bernhard Brümmer. Wheat export restrictions and domestic market effects in russia and ukraine during the food crisis. *Food Policy*, 38: 214–226, 2013. ISSN 0306-9192. doi: https://doi.org/10.1016/j.foodpol.2012.12.001. URL https://www.sciencedirect.com/science/article/pii/S0306919212001261.

Ramesh Shama. *Food export restrictions: Review of the 2007-2010 experience and considerations for disciplining restrictive measures*. Citeseer, 2011.

IPCC. *Climate Change 2014 Part A: Global and Sectoral Aspects*. 2014. ISBN 9781107641655. URL papers2://publication/uuid/B8BF5043-C873-4AFD-97F9-A630782E590D.

Jacob Schewe, Simon N. Gosling, Christopher Reyer, Fang Zhao, Philippe Ciais, Joshua Elliott, Louis Francois, Veronika Huber, Heike K. Lotze, Sonia I. Seneviratne, Michelle T.H. van Vliet, Robert Vautard, Yoshihide Wada, Lutz Breuer, Matthias Büchner, David A. Carozza, Jinfeng Chang, Marta Coll, Delphine Deryng, Allard de Wit, Tyler D. Eddy, Christian Folberth, Katja Frieler, Andrew D. Friend, Dieter Gerten, Lukas Gudmundsson, Naota Hanasaki, Akihiko Ito, Nikolay Khabarov, Hyungjun Kim, Peter Lawrence, Catherine Morfopoulos, Christoph Müller, Hannes Müller Schmied, René Orth, Sebastian Ostberg, Yadu Pokhrel, Thomas A.M. Pugh, Gen Sakurai, Yusuke Satoh, Erwin Schmid, Tobias Stacke, Jeroen Steenbeek, Jörg Steinkamp, Qiuhong Tang, Hanqin Tian, Derek P. Tittensor, Jan Volkholz, Xuhui Wang, and Lila Warszawski. State-of-the-art global models underestimate impacts from climate extremes. *Nature Communications*, 10(1):1–14, 2019. ISSN 20411723. doi: 10.1038/s41467-019-08745-6.

Puyu Feng, Bin Wang, De Li Liu, Cathy Waters, and Qiang Yu. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. *Agricultural and Forest Meteorology*, 275 (November 2018):100–113, 2019. ISSN 01681923. doi: 10.1016/j.agrformet.2019.05.018. URL https://doi.org/10.1016/j.agrformet.2019.05.018.

Juan-Carlos Ciscar, Karen Fisher-Vanden, and David B Lobell. Synthesis and review: an inter-method comparison of climate change impacts on agriculture. *Environmental Research Letters*, 13(7):070401, 2018.

Kate Calvin and Karen Fisher-Vanden. Quantifying the indirect impacts of climate on agriculture: an inter-method comparison. *Environmental Research Letters*, 12(11):115004, 2017.

Richard H Moss, Jae A Edmonds, Kathy A Hibbard, Martin R Manning, Steven K Rose, Detlef P Van Vuuren, Timothy R Carter, Seita Emori, Mikiko Kainuma, Tom Kram, et al. The next generation of scenarios for climate change research and assessment. *Nature*, 463 (7282):747–756, 2010.

Christian Folberth, Artem Baklanov, Juraj Balkovič, Rastislav Skalskỳ, Nikolay Khabarov, and Michael Obersteiner. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and forest meteorology*, 264:1–15, 2019.

Michael J. Roberts, Noah O. Braun, Thomas R. Sinclair, David B. Lobell, and Wolfram Schlenker. Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters*, 12(9), 2017. ISSN 17489326. doi: 10.1088/1748-9326/aa7f33.

Alex C Ruane, Cynthia Rosenzweig, Senthold Asseng, Kenneth J Boote, Joshua Elliott, Frank Ewert, James W Jones, Pierre Martre, Sonali P McDermid, Christoph Müller, et al. An agmip framework for improved agricultural representation in integrated assessment models. *Environmental Research Letters*, 12(12):125003, 2017.

L Breiman. Random Forests. 2001, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014. ISSN 15337928. doi: 10.1117/1.JRS.11.015020.

Bin Wang, Cathy Waters, Susan Orgill, Annette Cowie, Anthony Clark, De Li Liu, Marja Simpson, Ian McGowen, and Tim Sides. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern australia. *Ecological indicators*, 88:425–438, 2018.

Wolfram Schlenker and Michael J. Roberts. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, sep 2009. ISSN 0027-8424. doi: 10.1073/pnas.0906865106. URL http://www.pnas.org/lookup/doi/10.1073/pnas.0906865106.

Gabriel M. Abrahão and Marcos H. Costa. Evolution of rain and photoperiod limitations on the soybean growing season in brazil: The rise (and possible fall) of double-cropping systems. *Agricultural and Forest Meteorology*, 256-257:32–45, 2018. ISSN 0168-1923. doi: https://doi.org/10.1016/j.agrformet.2018.02.031. URL https://www.sciencedirect.com/science/article/pii/S0168192318300819.

Petr Havlík, Hugo Valin, Mario Herrero, Michael Obersteiner, Erwin Schmid, Mariana C Rufino, Aline Mosnier, Philip K Thornton, Hannes Böttcher, Richard T Conant, et al. Climate change mitigation through livestock system transitions. *proceedings of the National Academy of Sciences*, 111(10):3709–3714, 2014.

Juraj Balkovič, Marijn van der Velde, Rastislav Skalský, Wei Xiong, Christian Folberth, Nikolay Khabarov, Alexey Smirnov, Nathaniel D. Mueller, and Michael Obersteiner. Global wheat production potentials and management flexibility under the representative concentration pathways. *Global and Planetary Change*, 122:107–121, 2014. ISSN 09218181. doi: 10.1016/j.gloplacha.2014.08.010.

Jimmy R Williams et al. The epic model. *Computer models of watershed hydrology.*, pages 909–1000, 1995.

Paul A. Dirmeyer, Xiang Gao, Mei Zhao, Zhichang Guo, Taikan Oki, and Naota Hanasaki. Gswp-2: Multimodel analysis and implications for our perception of the land surface. *Bulletin of the American Meteorological Society*, 87(10):1381 – 1398, 2006. doi: 10.1175/BAMS-87-10-1381. URL https://journals.ametsoc.org/view/journals/bams/87/10/bams-87-10-1381.xml.

S Lange. Wfde5 over land merged with era5 over the ocean (w5e5). *GFZ Data Services*, 2019. doi: 10.5880/pik.2019.023.

Christoph Müller, Joshua Elliott, James Chryssanthacopoulos, Almut Arneth, Juraj Balkovic, Philippe Ciais, Delphine Deryng, Christian Folberth, Michael Glotter, Steven Hoek, Toshichika Iizumi, Roberto C. Izaurralde, Curtis Jones, Nikolay Khabarov, Peter Lawrence, Wenfeng Liu, Stefan Olin, Thomas A.M. Pugh, Deepak K. Ray, Ashwan Reddy, Cynthia Rosenzweig, Alex C. Ruane, Gen Sakurai, Erwin Schmid, Rastislav Skalsky, Carol X. Song, Xuhui Wang, Allard De Wit, and Hong Yang. Global gridded crop model evaluation: Benchmarking, skills, deficiencies and implications. *Geoscientific Model Development*, 10 (4):1403–1422, 2017. ISSN 19919603. doi: 10.5194/gmd-10-1403-2017.

L Alexander and N Herold. Climpactv2 indices and software. *A document prepared on behalf of the Commission for Climatology (CCl) Expert Team on Sector-Specific Climate Indices (ET-SCI), Sydney*, 2015.

Delphine Deryng, Joshua Elliott, Christian Folberth, Christoph Müller, Thomas AM Pugh, Kenneth J Boote, Declan Conway, Alex C Ruane, Dieter Gerten, James W Jones, et al. Regional disparities in the beneficial effects of rising co 2 concentrations on crop water productivity. *Nature Climate Change*, 6(8):786–790, 2016.

Andrea Toreti, Delphine Deryng, Francesco N Tubiello, Christoph Müller, Bruce A Kimball, Gerald Moser, Kenneth Boote, Senthold Asseng, Thomas AM Pugh, Eline Vanuytrecht, et al. Narrowing uncertainties in the effects of elevated co 2 on crops. *Nature Food*, 1(12): 775–782, 2020.

Petr Havlík, Uwe A Schneider, Erwin Schmid, Hannes Böttcher, Steffen Fritz, Rastislav Skalskỳ, Kentaro Aoki, Stephane De Cara, Georg Kindermann, Florian Kraxner, et al. Global land-use implications of first and second generation biofuel targets. *Energy policy*, 39(10):5690–5702, 2011.

Georg Kindermann, Michael Obersteiner, Brent Sohngen, Jayant Sathaye, Kenneth Andrasko, Ewald Rametsteiner, Bernhard Schlamadinger, Sven Wunder, and Robert Beach. Global cost estimates of reducing carbon emissions through avoided deforestation. *Proceedings of the national Academy of Sciences*, 105(30):10302–10307, 2008.

MI Gusti. An algorithm for simulation of forest management decisions in the global forest model. , 2010.

Mario Herrero, Petr Havlík, Hugo Valin, An Notenbaert, Mariana C Rufino, Philip K Thornton, Michael Blümmel, Franz Weiss, Delia Grace, and Michael Obersteiner. Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems. *Proceedings of the National Academy of Sciences*, 110(52):20888–20893, 2013.

Puyu Feng, Bin Wang, De Li Liu, Hongtao Xing, Fei Ji, Ian Macadam, Hongyan Ruan, and Qiang Yu. Impacts of rainfall extremes on wheat yield in semi-arid cropping systems in eastern australia. *Climatic change*, 147(3):555–569, 2018.

Christoph Müller, James Franke, Jonas Jägermeyr, Alex C Ruane, Joshua Elliott, Elisabeth Moyer, Jens Heinke, Pete D Falloon, Christian Folberth, Louis Francois, et al. Exploring uncertainties in global crop yield projections in a large ensemble of crop models and cmip5 and cmip6 climate scenarios. *Environmental Research Letters*, 16(3):034040, 2021.