Replies (**in bold and blue**) to the comments/suggestions (*in italics*)

Dear Editor Aloïs Tilloy,

Thank you very much for considering our manuscript "A data-driven framework for assessing climatic impact-drivers in the context of food security" for publication in Natural Hazards and Earth System Sciences. We appreciate the time and effort you and the reviewers have invested in evaluating our work.

As requested, we have thoroughly revised our manuscript to address the reviewers' suggestions. We have ensured that all necessary clarifications have been made to enhance the manuscript's clarity and have referenced similar analyses to provide a more comprehensive context for our work.

We also revisited and evaluated the model codes to guarantee that the problem of spatio-temporal overfitting is clarified. All the scripts used in the manuscript are now available at the GitHub repository https://github.com/marcosrbenso/ClimateImpactML. We will explain what were the main changes in the point by point-by-point response to the reviewers comments.

Moreover, we invited prof. José Marengo, who is a well-known expert in the field of climatology to help us to validate our results.

We believe these revisions have significantly strengthened our manuscript, making our research more transparent and its contributions clearer. Please see below for our detailed responses to all the referees' comments and suggestions. We look forward to your feedback on the revised manuscript.

Best regards,

Marcos Roberto Benso

On behalf of the authors

Reviewer #1

General comments:

The study uses interpretable machine learning to identify the most important climate impact drivers for predicting maize and soybean yield variability in Brazilian states. Overall, the manuscript is quite well-written and has clear descriptions of the datasets used which are very helpful for the reader. They discuss in some detail the advantages and disadvantages of the use of different yield datasets in the region, which is crucial for the interpretation of the results of this type of analysis, and make the effort to show a comparison of the datasets and where they agree and disagree. They also use two specific examples of droughts in Brazil as case studies to examine the interpretations, which is interesting and helps to verify their approach.

The topic is very important, and novel methods such as this have a clear use-case in identifying the most relevant periods at which different CIDs impact yields. However, I *have some concerns about the methodology. The description of the methodology used is not sufficiently thorough, so these concerns may have been addressed by the authors, but this should be clarified.*

Random forests are often used for this type of study and are a good choice when working with tabular data such as this. However, care must be taken when using any machine learning method not to allow the model to overfit to dependencies or correlation between features. The training and testing method used was not explained clearly, except in Figure 1, which only states that 20% of the data was used to test the models, but not how that 20% was selected. Given that models were trained on a state level, multiple municipalities within each state would have highly correlated climate and yields. Were the datapoints split in time and/or space to account for this, or sampled randomly? If they were sampled randomly, this can lead to misleading estimations of model performance and the interpretations are less likely to represent the physical mechanisms that are intended to be studied. Particularly relevant - if soil is used as a predictive feature, which does not vary in time in the dataset used (I believe), the model can easily spatially overfit.

Overall, I find the manuscript to be quite well-written and the thorough analysis of the different datasets used and how they impact the results is interesting and excellent scientific practice. However, I think that some small changes to the methodology (most importantly, selecting a test set considering the spatiotemporal autocorrelation and estimating SHAP values using this test set, ideally using a different feature selection method such as SFS) and better explanation of the steps involved to generate the results discussed could very much improve the paper. As the paper aims to present a framework to enhance the interpretability of ML methods fo crop yield loss prediction, it is important that the framework is robust and can deal with common issues for this type of problem such as overfitting to spatiotemporal data.

Thank you for your detailed and constructive feedback on our manuscript. We have addressed your concerns in our revised manuscript as follows:

1. **To avoid problems with temporal dependencies with data, we split the data into the first 80% datapoints for training and the last 20% datapoints for testing. For the training datasets, we applied a leave-one-year-out cross-validation approach (LOYOCV) (L99-L105)**
2. **The soil dataset was disconsidered for the model since it can easily spatially overfit.**
3. **We need to clarify that the 10 most important features were selected using feature importance. That is, we applied the feature importance rank that resulted from a random forest model and selected the 10 most important features.**

Finally, given that the title of the paper and stated goal is to present a framework that can be used by other researchers, the code used should be published and made openly available, but this is not currently stated in the manuscript.

After revising the codes to make sure to address the feedback given by Reviewer #1, all the scripts used in the manuscript are now available at the GitHub repository https://github.com/marcosrbenso/ClimateImpactML.

Specific comments:

At what stage was RFE used to select features, and how was this conducted? How many features were selected? I also question the use of RFE in cases where models can overfit (e.g. when spatiotemporal data is used), as features that the model find most important are more likely to not be physically meaningful. Using, for example, sequential feature selection with a spatial or temporal cross-validation splitting method would be more likely to return relevant drivers, and I would recommend to the authors to try this if possible.

We thank for the suggestions, we believe that using sequential feature selection with a spatial or temporal cross-validation could be very interesting for future work. For this manuscript, we used, as mentioned previously, LOYOCV for selecting the best fit model and extract feature importance and selected the 10 most important features. We believe that this helped to clarify the concerns about the methodology. (L99-L105)

In Figure 4, it would be helpful to have descriptions of what features were included in the different scenarios - in particular, I could not understand what 'Complex' meant.

After careful consideration, we decided to remove this figure from the text. We opted for showing the model performance of the best-fit model.

In Figure 5 and 6, is this after RFE has been used to select only 10 features? I was confused by the fact that for maize, only February features are shown, but later in the text it states that April and May precipitation was important for some regions.

We improved the representation of the best features selection showing a new figure that represents the most important feature according to the month. This helps to improve the discussion of the physical meaning of our method of variable selection. Moreover, the complete list of most important features was added in the supplementary material.

I would strongly advice not removing correlated variables before doing the feature selection. You can expect that the highly correlated variables will not both be selected, and it is another opportunity for data leakage to enter.

Thank you for your comment. Despite the fact that we acknowledge that data leakage is an important issue, features that have a correlation higher than 0.9 considering pearson correlation coefficients tend to be too similar, therefore, redundant. We still think that is important to remove.

I think it is very useful to compare the importances between the different states and datasets, as this can help to find robust insights and identify potential problems with the datasets used. It would be useful to see uncertainty quantification here as well, as given *that similar model performance can come from many combinations of features (as shown in Figure 4), one would expect that there is significant uncertainty in the interpretations as well. I would also consider using an additional feature importance metric (permutation feature importance on held-out test set?) for comparison, but this might be out of scope.*

I also find it unusual to fit random forest models and then to use a more complex model (XGBoost) to explain them via SHAP. Normally, SHAP is used directly on the trained model to be interpreted, and if a second model was used it would normally be a simpler model. Why not use XGBoost for the initial part of the analysis instead of adding this complexity of using a second model to explain the first?

To answer the two aforementioned comments, we opted to use only a random forest model for the two modeling steps, that is, selecting the 10 most important features and then explaining the impact of the 10 most important features on crop yield prediction. For the SHAP model, we also divided into training and testing. The choice of RF models is due to the fact that this model has been widely used in the literature for many environmental applications, specially for crop yield studies. Also, XGBoost calibration can be tricky with a large number of features. That's why we used the random forest model for selecting the most important features and then applied the XGBoost. Nonetheless, to avoid confusion, in this new version of the manuscript we used only RF models.

Partial dependence plots do not need SHAP values - they can be calculated by just varying individual features and estimating the output. It might be interesting to compare this against those gained from SHAP (but again, maybe out of scope). It would at least be useful to discuss/justify in the text why the partial dependence plots gained from SHAP are more useful (which is very plausible).

We thank the reviewer for the comment. Partial dependence plots were used in other papers (e.g., Vogel et al., 2019) that were cited in our manuscript. We agree that it is an interesting idea to compare the two plots to understand what insights can be gained by the use of SHAP partial dependence plots, however, we believe that this comparison should be included in future work.

SHAP values are also sensitive to the data used to calculate them, and I would again recommend to use test sets for this that are split with consideration to the spatial and temporal correlations.

As aforementioned, we split the data into testing and training for SHAP model as well.

Interpreting the results of this type of study can be difficult, as in general, any feature used for training is one that could be a causal driver. This means that it is hard to figure out if the results are meaningful or if the model has learned some spurious correlations. The fact that only February features are shown as important for maize suggests, to me, that something strange is going on, as the authors state that this is peak planting date and in some regions, planting is not finished until the beginning of April. It seems more likely that heat, for example, would be more important during the reproductive period.

Using the different test sets as I mentioned before might help with this, as well as using permutation feature importance instead of the internal RF variable importance measure.

We agree with the reviewer that interpreting the results is difficult, especially considering the complex systems like climate and agriculture. We tried to improve the discussion and clarity of the feature selection both in the manuscript and the supplementary material by displaying the full table of variable importance tests.

Why remove heteroskedasticity? Could this be justified more in the text? As we expect more climate variability with climate change and therefore more yield variability, it isn't obvious that this should be corrected for.

After careful consideration, we decided to remove this figure from the text. We opted for showing the model performance of the best-fit model. (L87-L88)

Lines 171-172 describe a second analysis using Gaussian copulas, but I could not find this further described or any results from this in the rest of the manuscript?

We agree with the reviewer that his part of the paper was poorly explained. Therefore, we improve the description of this method by including more details of what library was used and how this analysis is performed (L125-L135)

Technical corrections:

The paragraph on interpretability (lines 53 to 56) I could not understand.

We thank the reviewer for the comment. We agree that the paragraph needs to be improved both in terms of logic and language.

Old version: The paradigm of interpretability of machine learning models is a broad topic of discussion in supervised learning (Lipton, 2018). Two essential observations related to model interpretability are: (i) and (ii) the training data can be imperfect to represent a dynamic environment that changes over time.

New version: The paradigm of interpretability of machine learning models is a broad topic of discussion in supervised learning (Lipton, 2018). The model interpretability can be achieved by means of feature engineering and using interpretable models such as linear models, that is “algorithmic transparency”. When the features, or input data, are decomposed and the number of variables make the interpretation of models difficult, post hoc interpretation can be used to extract explanations from learned models.

Please state briefly that the crop yields were detrended in the main text (the further explanation in the Supplementary is very helpful, but there is no mention of the fact that the yields are detrended in the main manuscript which is very important to interpret the results).

Thank you for your comment. The detrend of time series is now mentioned in the L185-L187

Replies (**in bold and blue**) to the comments/suggestions (*in italics*)

Reviewer #2

We thank Reviewer #2 for taking the time to read and post feedback for our manuscript. There is no substantial suggestion in the message given by Reviewer #2, therefore, we believe that, with the changes made according to Reviewer #1, we encompassed all changes necessary to improve the manuscript. However, we are looking forward to hearing further comments from reviewer #2.