Dear Editor Aloïs Tilloy,

Thank you very much for considering our manuscript for publication and for doing diligent work as editor. We are happy with how Reviewer #1 is engaged in improving our manuscript and for the throughout revision that was performed.

We have revised our manuscript to answer Reviewer's #1 comments. We ensured to answer all the confusion points and feedback regarding the methodology. We believe that with the revised manuscript, the reader will be much more confident in reproducing the method we proposed. We also made sure to correct all technical comments to improve readability and avoid typos and small writing mistakes.

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

First, are models trained individually for each spatial point? This is not explicitly stated anywhere, as far as I can tell.

Thank you for raising this point. To clarify, we create one model for each particular state. Specifically, for each state, we trained separate models using each of the three datasets (IBGE, Deral, and GDHY). This means that, for each state, we have three distinct models corresponding to the three datasets.

To address this more explicitly, we have added the following clarification to Section 2.1:

"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."

It is also not clear from the text how the CV and training/test evaluation is executed. For example, lines 151-152 (in the file with tracked changes): ‘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 first 80% for training and 20% for validation data.’. This does not explicitly state that the split is in time, and although a training, validation and test set is mentioned, in fact only a training and test split is done.

Thank you for your observation. We acknowledge the need for greater clarity regarding the splitting strategy. To clarify:

The training and validation datasets were split considering the chronological order of the data. Specifically, the first 80% of the time was allocated for training, and the last 20% was reserved for validation. This temporal split ensures that the model is tested on later, unseen data, in a way that happens in the real-world predictive scenarios. While we referred to the second subset as "validation data" in the manuscript, this term might have contributed to the misunderstanding, as no separate test dataset was used in this workflow.

To address this, we have revised the relevant text (lines 151–152) as follows:
"The first 80% of the data, according to the timeline, was used to train the model, allowing it to learn and adjust its parameters. The remaining 20% 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."

Additionally, it is unclear how the LOYO-CV is done - the text states a ‘fixed window’ of 10 years was used, and one year as a test set; is this repeated such that every year is a validation set and then the resulting scores are averaged, as is usually done for CV? ‘Fixed window’ reads as if only a single split is done, but I assume the authors do not mean that.

Thank you for raising this point. We recognize that our description of the LOYO-CV methodology was not sufficiently detailed, which may have led to ambiguity. To clarify:

In the LOYO-CV approach, we employed a fixed window of 10 years for training, followed by a single year as the test set. This process was repeated iteratively, leaving each year out as the test set while using the preceding 10 years for training. The evaluation metrics were computed for each iteration, and the final scores were averaged across all iterations to provide a robust estimate of model performance.

To avoid misunderstanding, we have revised the text to provide greater clarity as follows:
"A fixed 10-year window was used for training, followed by one year as a test set. This process was repeated iteratively, leaving each year as the test set while using the preceding 10 years for training. Performance metrics were calculated for each iteration and scores were averaged to obtain an overall assessment of the performance of the model."

Line 160: ‘the best fit model was determined by employing a leave-one-year-out cross-validation approach (LOYOCV)’: I do not understand this step. Did the authors train multiple models, with identical hyperparameters and features, on different subsets of years and select the best one according to its performance on each corresponding validation set? This would not be advisable, as the validation set performance would depend on the year, not just the model performance. Or were the hyperparameters selected or features selected based on the average CV performance across all folds, as would be more typical?

The hyperparameters were selected based on the average CV performance across folds. Thank you for observing this point. To improve the clarity of this point, we added the following text in the manuscript:

“the best-fit model was selected using a leave-one-year-out cross-validation method (LOYOCV), and hyperparameters were chosen according to the mean CV performance in folds”

Lines 175-177: ‘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 important climate indices.’ I would recommend describing this more explicitly. Which subsets of features were tested and how were they evaluated?

I think this paragraph does not represent what was done in the manuscript we presented. To avoid confusion, we decided to remove it from the text.

I would suggest that the authors explicitly state which years are covered in the training and test sets, on which years the LOYO-CV is done, and further on in the text, on which years/data the results are calculated on for each figure. This should also be included in Figure 1.

Thank you for your feedback. We believe the explanation regarding the training/validation split and the cross-validation (CV) procedure was clear. However, we understand that the issue arises from the datasets differing in length and the years they encompass, which may have caused some confusion.

Additionally, there are multiple ways of calculating feature importance from random forest models. I am assuming that the authors refer to the internal entropy-based feature importance. This should be stated explicitly. I would also advise that the authors consider additionally calculating the permutation-based feature importance on the validation or test years. Their agreement or disagreement with the entropy-based feature importance would aid readers in assessing the robustness of the findings.

Thank you for your comment. We believe that comparing the agreement and disagreement considering permutation-based feature importance is a great idea. Nevertheless, after careful evaluation, we decided to clarify the method we used as Reviewer #1 suggested and perhaps compare the entropy-based with permutation-based feature importance for future work.

"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."

Other points of confusion:

- Line 128, 129 introduces ‘explainable’ and ‘operational’ features, but does not define them. Also lines 137-8: ‘Other relevant aspects, such as relevancy, explainability, and operationability, will be explained in the following steps.’ As far as I can see, these terms aren’t explained later in the text.

The idea behind mentioning the terms ‘explainable’ and ‘operational’ features represent the motivation of the methodology. We believe that, by using such a method that we proposed, we can achieve that. To avoid misunderstanding we clarified this in the text:

“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.”

For simplicity and to avoid confusion, we decided to remove the passage “ Other relevant aspects, such as relevancy, explainability, and operationability, will be explained in the following steps.” from the text.

- Line 144: ‘The SHAP method uses a second model, most commonly the RF model…’ I do not think this is true. SHAP is used to explain the original model.

Thank you for pointing it out. We corrected it in the text by reformulating the sentence:

“The SHAP method is used to explain how each variable was used to make each prediction”

- Line 162: ‘The models were trained and optimized on the training and validation datasets’ This is unusual, did the authors intend to write only training datasets, not validation datasets?

Thank you for noticing, this was only applied to the training dataset. We corrected it in the text

- Outlier removal: Lines 276-277 state that the authors remove outliers using the interquartile range, but from the supplementary material it seems that only few datapoints are removed, which does not make sense. Additionally, the supplementary material discussion of the outlier removal is confusing: in lines 30-31 they state ‘Removal of outliers is a complex problem since we are working with extreme events’, but this is followed by an explanation of the trend and heteroskedasticity removal process, and not the outlier removal. Then, in line 52: ‘After obtaining a consistent time series corrected for outliers, trends, and heteroskedasticity’ - but the outlier removal occurs afterwards, as far as I can tell (line 59: ‘To eliminate potential outliers, we excluded values considering each year and state’).

I think that this issue requires clarification. We considered the immediate regions that consist of a group of municipalities that are close together. The definition of immediate region is given by IBGE. The crop yields in these regions were very similar based on the correlation index of the yields. Then, we considered the interquantile range of each year. For example, in a give year, if the yield of one municipality within the immediate region is much higher or lower than the other, this would be considered an outlier.

In the main text we added the following explanation:

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.

In the Supplementary material

“After obtaining a consistent time series corrected for trends, heteroskedasticity, and outliers”

“To eliminate potential outliers, we excluded values considering each year and immediate region within the state. This was done because he hypothesizes that within the immediate region, the crop yields should be similar”

- Lines 277-278: ‘Changes in technology in seed production, fertilizers, and land management, also known as technological trends (Liu and Ker, 2020) were removed by Local Polynomial Regression Fitting (LOESS)’ - all trends would be removed, including those due to e.g. climate change, not just technology. I would recommend mentioning this.

Thank you for pointing this out. We added it in the text:

“and other sources of trend such as climate change were removed by Local Polynomial Regression Fitting (LOESS)”

- Supplementary material line 137: the sentence ending in ‘indicating the significant role of both rainfall’ is incomplete.

We completed the sentence:

“indicating the significant role of both rainfall and temperature.”

- Figure 3: Where do the error bars come from here? Which variables were used as predictive features? Are these metrics calculated on the test set?

Since different models were trained, we used the performance of the models trained for each state to create the error bars.

- Lines 340-342: Where do these ranges come from?

These ranges come from Figure 5. To clarify, we added the reference in the text.

- Lines 349-354: Where are the results discussed here shown?

The results are the analysis of Figure 7. We added the reference and improve the writing of this passage of the manuscript.

- Figure 4: The hazard types don’t correspond, as far as I can tell, to the CID types and categories from Table 2. What do these labels mean?

Thank you for pointing this out. We actually meant CID categories. We corrected the figure.

- Line 420: ‘This technique allowed us to extract insights by coupling the results of a Random Forest model’ - I don’t think SHAP works by coupling a model to another, it is intended to explain the original model.

We agree that this sentence is a bit confusing. We corrected in the text:

“This technique allowed us to extract insights from the results of a Random Forest model”

- In lines 466-468, and in the Supplementary material (lines 95-101) the authors discuss whether or not to keep the most extreme years in the training or test set, or remove them. I understand from the text that they were kept in the dataset, but it doesn’t say if they were used in the test or training set.

All the extreme years were kept in the dataset, we mentioned it in the supplementary material.

“As our aim was to assess the effects of extreme climate events, we opted to retain all of these extreme events within the dataset”

Technical corrections:

- Line 136: I would remove the sentence ‘Feature selection is a pre-processing step in machine learning models’. It is confusing as the feature selection is described later in the text.

We agree on this correction and we removed it from the manuscript.

- ‘ML’ is introduced as an abbreviation early in the text, but the authors continue to use ‘machine learning’ afterwards. I would also advise using ‘RF’ as an abbreviation for random forest to improve readability.

In fact, this correction will improve the readability. We kept only the first time the terms “random forest” and “machine learned” appeared. After that, we only used RF and ML as suggested.

- Figure 1 is helpful, but there are multiple typos and minor formatting issues. E.g. ‘Filter Highly correlated variables’ should be ‘Filter highly correlated variables’; ‘Boostrap RF model’ should be ‘Bootstrap RF model’.

Thank you for the correction. We applied it to the figure 1

- Line 170: ‘To achieve this, we used the Random Forest model’ This is repeated multiple times in the text, and could be removed.

Thank you for the correction. We removed it from the text.

- Line 175: ‘Different models were trained considering different combinations of input data’ - I believe the authors refer to different combinations of features or variables, rather than data.

The reviewer #1 is correct. We referred to different combinations of features. However, we also need to add the we considered different crop yield datasets.

- Line 207: Typo - ‘The ~~the~~ SHAP explanations was performed’

The typo was corrected

- Lines 239, 240, 244: Maize and soybean should not be capitalised here.

This was corrected.

- Figure S1 and S2: Typo - eath should be each
This was corrected

- Supplementary line 86: The reference is erroneously capitalised: RODRIGUES et al. (2013)

This was corrected

- Supp line 130: typo - Table SS2 should be S2

This was corrected

- In the Supplementary material, Section 4 still refers to an XGBoost model.

This was corrected

- Lines 364-365: Typo - 'The analysis can be of variable importance for soybean datasets is shown in Table S1'

This was corrected

- Line 413: Typo - 'climate impact-divers' should be 'drivers'

This was corrected

- Lines 461-462: Typo - 'however, the for IBGE is 150mm'

This was corrected

- Figure 7: Please include the units for e.g. precipitation.

This unit for precipitation was included in the figure

- Line 471: 'exemplify' is the wrong word, I think - perhaps 'present'?

Reviewer #1 is correct. The word "present" is more suited. Corrected it in the text.