Rebuttal Letter for the Reviews on 'Machine learning methods to assess the effects of a non-linear damage spectrum taking into account soil moisture on winter wheat yields in Germany'

Michael Peichl[1*], Stephan Thober[1], Luis Samaniego[1], Bernd Hansjuergens[2], and Andreas Marx[1**]

[1]UFZ-Helmholtz Centre for Environmental Research, Department Computational Hydrosystems, Permoserstrasse 15, D-04318 Leipzig, Germany
[2]UFZ-Helmholtz Centre for Environmental Research, Department Economics, Permoserstrasse 15, D-04318 Leipzig, Germany
[*]michael.peichl@ufz.de
[**]andreas.marx@ufz.de

We thank the reviewers for their detailed comments, which helped us to improve our manuscript. Below are the responses corresponding to Reviewer 1's and Reviewer 2's comments.

Reviewer 1

Summary

The state-dependent clustering was removed in this version and all relevant figures have been pdated accordingly. I therefore recommend the article for publication after minor revision.

Dear Reviewer, thank you for your detailed suggestions for accepting the manuscript with minor revisions. We have responded to the major comments and the minor comments throughout the manuscript. We are also thankful for the comments on the ALE plots. Please find our responses below.

Major comments

Table 2 and lines 180-181: In all 12 cases the test R^2 is higher than the train R^2. In my understanding, test performance is usually below training performance. Higher test performance can occur occasionally, but it seems odd that this is the case for all your settings. Do you have an explanation for this? How did you split the train and test data set? Is there e.g. any possibility of a bias due to the splitting? Did you try various ways of splitting the data? Related to this: You mention overfitting as an explanation of underestimation of year 2019 in line 328-329. But your better test performance compared to training performance would not indicate overfitting. (However, the following argument in line 330-334 seems reasonable.) **Thank you for**

this comment. The training and testing sets are split completely randomly. For this purpose, the $slice\_sample$ command of the dplyr package is used with a seed (tested for different seeds as well as alternative random sampling commands) that randomly selects rows in the data (each row assembles the data for a specific time and space combination). The model is tuned using the training data (80% of the data). The test results shown in Table 2 were derived from predicting on the remaining 20% of the data, i.e., the test set. Since this set was randomly selected from the combined space-time space, it is reasonable to assume that the models are well suited for this realm but overfitted for either prediction in the isolated time or space dimension. As will be seen later when examining the year-specific out-of-sample predictions (lines 328-329), it appears to be the case that year-specific effects are not properly captured in the approach. This also shows potential difficulties of out-of-sample predictions of machine learning models such as random forest (l.327-328). However, we believe that such considerations should be the subject of further study. We added some information clarifying that the sampling is random based on time-space combinations (l. 149-150 and 181-182).

Minor comments

[L. 65]: What are mean average effects? Mean and average seem synonymous.

Thank you for this comment. ALE handles feature correlations by averaging and accumulating the difference in predictions across the conditional distribution (grid of realizations), thereby isolating the effects of the specific feature. The resulting ALE explanation is then centered around the feature's mean effect, so that the feature's main effect is compared to the average prediction of the data. For this reason, we used the phrase "mean average effects." However, we instead use average marginal effects here at this point.

[Figure 1] Isn't it 2003 instead of 2004?

Of course you are right, this was a typo. *[Table 2] If you explain the abbreviations T and P here, you should also explain SMI and SMIa for the sake of completeness.*

We added this information to Table 1. *L. 220-221] SMI4 (5th in non-cluster); SMI8 (6th in non-cluster)*

Thank you, we adapted this accordingly. *[Figure 4] Why is it not completely identical for "4a) no cluster" to the submitted version? SMI4 is now on fourth place instead of SMI12.*

As you can see in Figure A4 (appendix), the importance scores of SMI4 and SMI12 are very similar. To calculate the feature importance for a single feature, the model prediction loss (error) is measured before and after shuffling the feature values. The larger the increase in prediction error, the more important the feature was. The shuffling is repeated (here 50 times) to obtain more accurate results because the importance of permutation features is usually quite unstable. However, the permutation can still impact the final order. *[Figure 4] Mention in the caption that this is for PAM (4).*

Thank you, we added this information. *[Figure 4, A5-A9] You mention a grid size of 50, however in your answer to the reviews you state 100 (lines 139-148 of your rebuttal letter).*

We are sorry for the confusion. The ALE plots are based on a grid size of 50. *[L. 243] "A low drought signal of soil moisture is found for April (SMI4)." Odd phrasing, consider rephrasing.*

Thank you, we use "A low drought signal based on soil moisture is noted for April (SMI4)" now. *[L. 268] What about cluster 1 and 3? They also exhibit drought vulnerability. Maybe write "subregion such as cluster 2..."*

We now use the recommended phrasing. *[L. 270 and 338] crop yield potential*

Thank you, we now use crop yield potential. *[Figure 5] Missing data in gray should be indicated in the caption.*

We added this information in the caption. *[Fig. A4-A6] What is CR12?*

This is a carryover from the work names. For the paper we now use the name Rain12. *[L. 405] It is mean absolute error, not average. This was already mentioned in the first revision*

We apologize for repeating this error.

ALE Plots

The ALE plots are confusing because the contain a variety of small mistakes and inaccuracies. Basically I think you want to show (as indicated in lines 414-418) in these plots the 3 best cluster algorithms for cluster 2, 4 and 6 for the two soil moisture index configurations "Soil moisture for uppermost 25cm" and "Soil moisture index for both uppermost 25cm and entire soil column" (one of the plots (Fig. 4) is in the main text), right?

The wobbliness of the curves is the result of the detailed grid used to estimate the effect of the predictor variables on the target variable. In addition, it may be an artifact due to the different effect of a feature in different subregions, which is not adequately accounted for by clustering. For this reason, we added the nonlinear smoothing functions. With a less detailed grid, the ALE plots would also be smoother, but could hide relevant information. The goal here is to show how the average marginal effects depend on the underlying clustering. For example, depending on the number of clusters used and the resulting size of the clusters, information may or may not be revealed. This is noted, for example, in lines 292 - 293. *[L. 376-378] This text was not updated. It is no longer PAM (8).*

Thank you, we updated the text accordingly. *Fig. A5 is not for PAM with 2 clusters as indicated in the main text and the caption, but identical to Fig. A6.*

Thank you, something went wrong when setting the reference in latex. Also, as indicated later, Fig. A5 should represent HIERARCHICAL(2) for the configuration of the soil moisture index, considering only the top 25 cm. *Figs. A5-A9 all say "for both cluster". In three of these figures there are more than two. Also the plural-s is missing.*

Thank you, we implemented your suggestions accordingly. *First line of the caption: Instead of "SMI" call this "soil moisture index configuration" in all ALE plots as you do in Table 2.*

Thank you, this is now implemented. *Fig. A5 first line of the caption: Shouldn't this be HIERARCHICAL (2) instead of PAM (2)? According to Table 2 HIERARCHICAL (2) is the best algorithm cluster size 2 and soil moisture index configuration "soil moisture for the uppermost 25 cm." Also in Line 414-415 it is mentioned as HIERARCHICAL (2).*

Thank you very much for this comment. Indeed, the best model in terms of test R-squared is HIERARCHICAL(2) *Fig. A7 second line of the caption: 2 clusters, not 6.*

We changed this accordingly. *Figs. A5-A9 first sentence of the caption: This is a bit misleading (in all ALE plots). It should rather say something like: best combination of cluster algorithm (PAM) and the soil moisture index configuration "soil moisture*

for the uppermost 25 cm" for cluster size 2. Each plot shows a different cluster size, so it is not part of the best combination, but rather it is the best combination for a given cluster size.

Thank you. We changed the first sentence of the caption as is stated in the following for Fig.A5: "Accumulated local effects (ALE) plots for the best combination of cluster algorithm and cluster size (HIERARCHICAL with 2 clusters) for a soil moisture index configuration that only considers the uppermost 25 cm." For the other captions this sentence is adapted accordingly.

Typos

[L. 115] extra bracket

[L. 233] bracket missing

[L. 240] 4d

[L. 249] Heat8 in upper case

[L. 247] bracket missing

Thank you, we have corrected the typos mentioned above.

Reviewer 2

General comments

The authors apply a random forest procedure to explain observed yield anomaly in Germany thanks to meteorological predictors. The central result of this paper is the quantification of individual non-linear contributions of meteorological variables/indices and especially the important role of soil moisture. The ALE plots are valuable material in this study (I am not sure about the interpretation of confidence intervals though, see specific comment n.3). To my opinion, this paper deserves publication after minor revisions.

Dear Reviewer, Thank you for your detailed summary and suggestions for accepting the manuscript with minor revisions. We have responded to the specific comments and technical corrections throughout the manuscript. We also thank you for the comments on the labels of the ALE plots. Regarding the ALE plots, there is currently no standard routine for uncertainty measures of the ALE plots (as described in more detail later). Please find our answers below.

Specific comments

[Fig 1b] You mention the identification of significant trends, did you perform some trend test or shift test?

Yes, we tested for trend in both the linear as well as in the LOESS model. Both models show significant trends. *[L. 141] Is the clustering performed on raw yield directly, or anomalies? How should one interpret the clusters obtained: are the counties gathered in terms of yield magnitude or variability or occurrence of extremes?*

Thank you for this comment. As stated in the text in lines 151 - 155, we use for clustering "monthly averages and daily

observations of the meteorological data for the entire year. Soil moisture index is included for both the upper layer and the entire soil column. Average yields are also taken into account in the data for cluster formation. This is based on the intuition of taking into account time-invariant factors of each cluster that affect yields such as soil quality and average farm size. These factors are not considered in the random forest due to use of yield anomalies." Thus, the clusters aim to capture time-invariant factors, while the yield anomalies and corresponding ALE plots aim to capture the effects of extremes. *[Figure 4] the confidence intervals you obtain are related to the smoothing function, and not the robustness of the RF model itself. I am wondering to what extent it is possible to interpret it as an uncertainty of the variable effect (ex. l.238). To my understanding, this confidence interval tells us about the uncertainty of the smoothed curve, but not about the uncertainty of the local effect. (small remark: it is nice to specify the package for the ALE).*

Yes, the confidence intervals shown are based on the estimated LOESS functions and mostly refer to the number of realizations available to fit the function. Regarding the representation of uncertainty in the ALE representation, we would like to refer to an open issue in "General Pitfalls of Model-Agnostic Interpretation", an article by Molnar et al. (2020), the developer of the R package iml used for the analysis here: "While Moosbauer et al. [73] derived confidence bands for PDPs for probabilistic ML models that cover the model's uncertainty, a general model-agnostic uncertainty measure for feature effect methods such as ALE [2] and PDP [30] has (to the best of our knowledge) not been introduced yet." The package used and the corresponding command are presented in the caption of Fig. 4.

[L. 405] Do I understand correctly that "the feature is shuffled" means that the variable, for which we want to compute the importance, is shuffled?. I.e., to get the importance of e.g. SMI3, the RF is re-run on the exact same data, except for SMI3, which is shuffled in time, and then the MAE is computed?

Yes, you are right. To calculate the feature importance for a single feature, the model prediction loss (error) is measured before and after shuffling the feature values. Shuffling the feature values destroys the association between the outcome and the feature. The larger the increase in prediction error, the more important the feature was. The shuffling is repeated (here 50 times) to obtain more accurate results because the importance of permutation features is usually quite unstable.

Technical corrections

[L. 233] missing ")".

[L. 240] error in the reference to the figure

[L. 257] missing ")"

Thank you, we have corrected the technical corrections mentioned above.

It seems like there are some confusions in the figure captions:

[Caption figure 3] not "observed and predicted data for the two clusters" but "four clusters"?

Thank you, we corrected this accordingly. *[Figure 5] you chose to conduct the analysis with PAM 4 clusters, so in Fig5 it is four and not eight subregions, right?*

Yes, you are right. We corrected this. *[Figure A5 and A7] the captions don't seem to fit with the figures (clustering method and number of clusters).*

We have changed the labels. For a more detailed explanation of what has been changed, we would like to refer you to the ALE section of Reviewer 1. We thank you very much. *[L. 378] aren't the results in the appendix about PAM 4 and 6, and hierarchical 2 and 6?*

We show the ALE plots for the best combinations of cluster algorithm, cluster size, and SMI for a defined soil depth (either the top 25 cm or both the top 25cm and the total soil column) that are not presented in the main text. Those are for the uppermost 25 cm soil moisture index configuration HIERARCHICAL(2) and HIERARCHICAL(6), and for the combination of top 25cm and total soil column HIERARCHICAL(2), PAM(4), and PAM(6).

References

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B.: General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models, arXiv.org, http://arxiv.org/abs/2007.04131, 2020.