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 substantially. In particular, the main comment from reviewer 1 was was very helpful because the corresponding analysis of why the clusters are identical to the state boundaries helped us identify an issue within the data provided to the clustering algorithms that was not valid. Unfortunately, this analysis is the basis of the statistical analysis and therefore has profound implications for all of the results presented in this manuscript.

Reviewer 2

General comments

This paper presents the modelling of non-linear effects of meteorological divers and soilmoisture on winter wheat yield variability. A random forest procedure models the nonlinear relationships. The model is applied on subregions of Germany, obtained with a clustering procedure. A comparison with the model trained over the whole country emphasizes the relevance of the clustering. The authors highlight the importance of soil moisture as a relevant explicative variable, more relevant than heat. The manuscript is well written. The description of the results is clear and supported by existing literature. This paper deserves publication after minor modifications/addition of complementary information.

Dear Reviewer, thank you for your detailed summary and suggestions for accepting the manuscript with minor modifications/addition of complementary information. We have responded to the specific comments and technical corrections throughout the manuscript. Please find our responses below.

Specific comments

[Table 1] The description of variables is generally self-explanatory, except maybe "alternative frost". Does it refer to the number of consecutive pairs of days with min T<-3 and then minT>3?

Thank you for drawing attention to this need for clarification. First, the minT > 3 should be declared as maxT > 3. Also, it takes into account days when both conditions apply, i.e. a day with, for example, minimum temperatures below -3 degrees at night and then during the day with degrees higher than -3 maximum temperature.

Is there a correlation between SMI and SMIa, and can this have an impact on the quality of the RF model? Same question for correlation between SMI(a) and indicators such as Heat, Heavy rain, precipitation scarcity.

There is a correlation between SMI and SMIa as well as with all meteorological variables. However, it is generally understood that this does not affect the predictive capacity of Random Forests - and this is what we assume is meant by model quality here. Nevertheless, it does affect the interpretability of the model. In our case, this is in particular the case for feature importance (if the are multiple features that contain the same information this as an impact of the order of partitioning the data). Since we are using accumulated local effects, purged of any correlation, this should be less affected by correlation issues such as multicollinearity.

[l.118/119] SMI is masked for non-irrigated agricultural land, but are these areas also discarded in yield data?

As the yield information are only available on administrative district level this masking was not possible. When it comes to the factor irrigation, we consider this neglectable as only about 5 percent of the agricultural area is possibly irrigated with an focus on crops like potatoes (https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Landwirtschaft-Forstwirtschaft Fischerei/Produktionsmethoden/Tabellen/bewaesserungsmoeglichkeiten.html)

[l.138] How can one interpret quickly subregions within Germany obtained with clustering? Are they areas where yield is of the same order of magnitude and also monthly and daily meteorological are also similar?

A more detailed description can be found in our response to the main comment of reviewer 1. Unfortunately, the cluster algorithm is provided with invalid data regarding the federal states that impact the cluster formation.

[Figure 2] Please specify in the caption or the legend that the numbers in rectangles are referring to the Rsquare obtain from the RF procedure(?).

The entire figure now has been revised. Now, we state in the caption that the number indicate the respective test R-squared.

[Figure 2a] Is it by chance that except cluster number 8, all clusters are simply connected and almost convex? What could explain this very smooth partition?

Please see reply above and to reviewer number one.

[l.186] Would it a better option to use PAM(3) (with only SMIa) than PAM(2) (with both SMI)? (To get rid of potential correlation problems between SMI and SMIa).

We apologize for the confusion caused by not further clarifying that we are relying only on the top 25 cm in this setting.

To fix this problem, we have added this sentence to the paragraph before it: "Since the data for the entire soil column do not appear to provide any additional information for the model, we rely only on the top 25 cm for further analysis."

[l.202] What would be a solution to avoid this overfitting of the model?

In this case, we compare the ALE plots of RFs fitted with either the PAM8 or the PAM2 setting. Thus, the former is clustered into eight subregions. This means that each RF model is trained on a smaller sample size such as in the PAM2 context or when no cluster is used. Consistently, this is also the model and corresponding sample in each case on which the ALE plots for that subregion are based. Because of the less smooth functional relationships shown for this configuration, we assume that the models are overfitted to this small sample size. One possible solution would be to fit to larger samples. This is done here by relying instead on PAM2, i.e., the setting that considers only two clusters and therefore has a higher sample size in each cluster.

[l.202] "The effects shown here are additive as they are cleared off the correlation to other features". I don't fully understand this sentence. Could you be more specific?

When features interact with other features in a prediction model, the prediction cannot be expressed as the sum of the feature effects, because the effect of one feature depends on the value of the other feature. Because of this compounding, the features are correlated to each other. In linear models this is commonly expressed by multiplicative expressions. As the ALE plots are purged of this correlation, we can rely on the sum of this features to derive the overall effect. This is similar to a linear model with only additive terms. We added some clarifications: "The feature effects shown here can be interpreted as additive because they are purged of correlation to other features. For example, the combined effect of soil moisture in June and July is the sum of SMI6 and SMI7."

[Figure 4] Do the black and white bars in x-axis represent the distribution of the explicative variable?

Yes, the black bars show the distribution of the respective features. White bars a results of discrete features.

[Figure 4., caption] It is not clear to me what the interval size of 100 refers to.

ALE plots predict the effect of an explanatory variable across their realisations, taking into account only a subset of the sample with observed values adjacent to the respective realisation. The size of this subset of the sample is defined by the grid size. The larger the grid size, the smaller the subset and the less smooth the visualization of the average marginal effects.

[l240] Would it be possible to add simple interactions in the model? (multiplication of 2 variables?)

In general, it would be possible to add these interactions by simply converting the features themselves into interactions, e.g. by multiplication of two variables as suggested. However, due to the strongly nonlinear structure caused by the underlying recursive partitioning, we consider such an approach not necessary for (large enough) random forests.

[l265-266] "In both clusters, heat in August, a period generally associated with ripening, has positive effects for each additional day and from day 11 onward negative effects" According to fig4, it looks like only for cluster 1, Heat8 has a negative effect from day 11 onward, not for cluster 2.

Thank you, we now distinguish between cluster 1 and cluster 2: "In both clusters, heat in August, a period generally

associated with ripening, has positive effects for each additional day and negative effects after the eleventh (cluster 1) respective sixth (cluster 2) day."

[l.291-295] Can the difficulties of out-of-sample prediction be interpreted as overfitting? Could it be Improved with longer time series (to have a larger number of configurations)?

The validation criterion the random forests are based on are out-of-bag error estimates. With tree sizes large enough the estimates converge to those found for leave-one-out cross-validation. This validation technique is more prone to overfitting compared to other methods. Potential solutions are the use of 1) an extended time-series, 2) the use of a different cross-validation technique with higher focus the variabilty of the prediction but less the bias in the training data, 3) the use of machine learning techniques that might be better suitable to extrapolated out-of-the sample compared to random forests. We clarified this in more detail in this paragraph and the conclusions.

Technical corrections

[l.113] Citation in brackets

Thank you, the citations are now corrected.

[l.154] Missing bracket

Thank you, we included the missing bracket.

[l.202] Extra "the" in "The effects shown here are additive as the they are cleared"

Thank you, we deleted the extra "the"

[l.377] Is "(Heat8)" supposed to be in that sentence?

No, it is supposed to be SMI11. Thanks for bringing this to our attention.

[Caption figure A4] "50 repetitions"

Thank you, it is of course 50 repetitions.