# Review on Peichl et al. 2021

## Summary

The authors present an assessment on the relationship of wheat yield anomalies to hydro-meteorological features using cluster analysis and random forest together with accumulated local effects. They achieve good model performance and additionally present an in-depth investigation of the most important driver variables, which adds additional value to the article. They highlight that their results serve well for the regional identification of harmul seasonal hydro-meteorological conditions. Furthermore, they state the potential usage of their research e.g. for the identification of harmful features and their related thresholds and also carefully underline the limitations regarding out-of-sample predictions, underestimations of extremes and missing inclusion of interaction effects in the model. I suggest that the manuscript should be accepted with moderate revision.

## Major comments

The fact that the clusters are identical with the borders of the federated states of Germany (or groups of them) without a single exception (as far as I can see) is not discussed and does not seem obvious to me. The authors should provide an explanation for this, e.g. stating which of the included variables in the cluster analysis are so strikingly different between the federated states that it leads to a perfect match of clusters and federated states (while any environmental / climatic differences do not matter in the clustering). It seems plausible that differences between eastern and western Germany due to the different political systems in place in the past can have a significant influence, but I wonder why this should hold true at the administrative level of federated states. The variables included in the cluster analysis are average yield and monthly averages and daily observations of the meteorological data for the entire year and SMI for both the upper layer and the entire soil column, right? I cannot see how these variables should primarily be superposed by the shapes of the federated states of Germany, even though I understand your argument that average yield is connected to farm size (which differs strongly between eastern and western Germany).

## Minor comments

Title: The paper is only about random forest, so you could consider using this term here instead of the more general term "machine learning".

l. 11: “R-sqare**d**” or preferably “$R^2$”. Make this consistent throughout the paper.

l. 16: I assume you mean “crop yield variations”.

l..79: Is the trend statistically significant?

Figure 1, 2, 5, A1: There are various grey counties, I assume mostly cities and/or counties without nonirrigated arable land (Peichl et al. 2018). If all these counties are the ones with 0 years of data, this should be clearly visible and stated in Fig. A1.

Figure 1 caption: Consider specifically naming the exceptional years 2003, 2014 and 2018 here.

Table 1: It should say “**max**. T>3°C” for alternating frost according to Gömann et al. 2015

Table 1: The presented variables in Table 1 seem plausible, but I wonder how excactly you came up with them? You sometimes depart from the months suggested by Gömann et al. 2015, e.g. they state the usage of Jan-Apr for Alternating Frost, while you also included May.

Table 1: It would be beneficial to have all predictor variables for the random forest in one overview table, so consider including SMI here, too.

l. 100: “each grid **point** depends”

l. 101-102: You compare your soil moisture index to simulations, but the soil moisture index used here is also simulated, so I do not understand this sentence fully. Could you elaborate please?

l. 113, 350, 368: adjust citation brackets

l. 113: I understand the line of argumentation of the authors, that soil moisture is a slow-responding variable with a long memory, but relating this to autocorrelation per se might be a bit misleading here, because temperature is of course also auto-correlated. So consider writing something like "The autocorrelation / long-term temporal persistence of soil moisture is comparatively high in comparison to temperature".

l. 116: State here also specifically that on the other hand for SMI all months of the growing season were used (only SMIa is reduced to four months).

l. 119: My understanding is that masked grid cells are excluded, but you mean the opposite that you kept only non-irrigated agricultural land, right?

l. 145: Did you run the random forest model for each cluster size from 2 to 16 for each algorithm as you did for your standard internal validation?

l. 147-49: I wonder whether the statements on the data inlcuded in the cluster analysis should be rather mentioned adjacent to the data included in the random forest model in section 2 to have the complete overview at once.

Table 2: While SMIa stands for the entire soil column, usage of the term SMI is ambigous. You use it to refer to the uppermost 25 cm (e.g. l.111), but also as a general term for both depths (e.g. l. 148, Table 2). Consider using separate terms.

Figure 2: You use dark grey for missing data and light grey for cluster 2. Consider taking another color for cluster 2 to make it better distinguishable.

Figure 2 and A1: coordinate degree sign as superscript

l. 189: You mention structural differences between western and eastern Germany. Consider mentionening briefly here the large differences in farm sizes due to different political systems in place, as an international audience might not be aware of this.

Figure 3: Indicate that the regression line is shown in bold black in the figure description. I assume the ellipses are point densities? Please specify this.

l. 227: I assume this reduction is to be compared with an average month, right?: “for each month with an SMI value lower than 0.125 **compared to an average month.**”

Figure 4: Is there a reason why there are many more red dots for the SMI compared to Heat and PS in a given subpanel?

Figure 4: You state for Fig. A4 and in the text section describing Fig. A5 that the average of 50 repititions is taken. Is this also the case here?

Figure 4 and A4: You chose an interval size of 100, whereas for Figure A4 it is 50. Why did you choose different interval sizes?

236: SMI **year-round**. “entire” could be confused with the entire soil column.

l. 239: “**led** to large losses”

l. 241: “but **need** to be”

l. 243: I assume “crop **yield** potentials” might be more precise.

l. 245: “water govern**s**”

l. 246: delete “this”

l. 266: “**eleventh** day”

l. 268: Why especially East Germany? It is valid for non-cluster and cluster 1, too, is it not?

l. 274: delete “years” in front of “in-sample”

l. 274: It did not directly become clear to me that 2019 is an out-of-sample year at the first read. I suggest explicitly naming it an out-of-sample year in this context (as you also do below).

l.286: “in **some of** the easternmost”

l. 288: “the **highest** negative”

l. 314: You use the terms East / eastern Germany and northeastern Germany. If all these mean the same region, consider using only one term or are you specifically referring only to cluster 7 with northeastern Germany?

l. 324: GCM: Abbreviations should be written out at first occurrence.

Code and data availability: The scripts seem not available online yet.

l. 355: “**a** small number” “**a** rather large number”

l. 367: Do you mean "mean **absolute** error"?

l. 376/378: Do you refer to SMI11 and SMI12 as "lagged"? Simply because they belong to another calendar year does not make them "lagged" variables in my opinion, so this term might be inappropriate here.

l. 377: Is SMI11 or Heat8 meant here?

Figure A4 caption: “50 **repetitions**”

Figure A5 caption: The caption should specify that the range of each variable stems from 50 (or 100?) repititions.