

Comment on hess-2021-9

Anonymous Referee #2

Referee comment 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" by Michael Peichl et al., Hydrol. Earth Syst. Sci. Discuss.,
<https://doi.org/10.5194/hess-2021-9-RC2>, 2021

General comments

This paper presents the modelling of non-linear effects of meteorological divers and soil moisture on winter wheat yield variability. A random forest procedure models the non-linear 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.

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 $\min T > 3$?
- 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.
- l.118/119 SMI is masked for non-irrigated agricultural land, but are these areas also discarded in yield data?
- 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?

- 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(?).
- 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?
- 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).
- l.202: What would be a solution to avoid this overfitting of the model?
- 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?
- Figure 4. Do the black and white bars in x-axis represent the distribution of the explicative variable?
- Figure 4., caption: It is not clear to me what the interval size of 100 refers to.
- l240. Would it be possible to add simple interactions in the model? (multiplication of 2 variables?)
- 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.
- 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)?

Technical corrections

- l.113 Citation in brackets
- l.154 Missing bracket
- l.202 Extra "the" in "The effects shown here are additive as the they are cleared"
- l.377 Is "(Heat8)" supposed to be in that sentence?
- Caption figure A4 "50 repetitions"