Comment on esd-2021-40
Anonymous Referee #1

Referee comment on "Weather-induced crop failure events under climate change: a storyline approach" by Henrique M. D. Goulart et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2021-40-RC1, 2021

Goulart et al. provide an interesting analysis of soybean failures in the US as modeled using Random Forest in the present climate and in future climates. The research is well conducted and uses appropriate methods. My main concern is not with the methods or results, but with the interpretation of those results. The manuscript is well written and the figures describe the data well. I thank the authors for presenting the research in such a complete and coherent manner.

Major comments

I rather like your use of Random Forest and you apply the model rigorously. Your use of shuffling data is also clever and provides a nice analysis structure, but your interpretation of the results could be improved. See my two comments below

- The interpretation of the compound factor becoming less important needs contextualization. I don’t think it’s quite correct to say that it becomes less important, but only that it becomes more dependent on precipitation. I think this relates to your test of importance being to shuffle the data in an attempt to break the correlation structure as compared to the ordered data. But in a warming climate, when all years are hot years, the dependence of joint hot-dry extremes becomes dependent mostly on rainfall, making the joint failure and shuffled failure years similar to one another at high return periods. I think it is misleading then to say that the compound nature of these events becomes less important, only that their occurrence depends primarily on precipitation anomalies.

- L 429-438 - There is no real evidence (from figure 7) that precipitation contributes to the increase in soybean failures overall. There is only evidence that it contributes in a secondary way to truly extreme events, such as 2012 (Figure 10), although its influence is much smaller than that of temperature
- There is an unintended consequence of using a threshold to define failures that leads to the conclusion that joint dependence will become less important in the future. When you set a threshold to identify “failure” years arising from hot years, you lose the sensitivity of yields to warming above that threshold. For example, in the future even if all years meet the criteria for hot years, drought years may be exceptionally hot. This joint temperature-precipitation correlation structure would show up if you were modeling yields as a continuous variable but is lost once you have converted your dependent variable to a binary (e.g. no discrimination between “failure” and “exceptional failure”).
  - In particular, your analysis cannot say that “the correlation between contributing meteorological variables becomes less relevant” (L392-393). You can only say that it becomes less relevant for your chosen yield threshold. If you were to up your threshold definition of “failure”, it would once again become important.
  - It is not true that “non-critical values of precipitation [...] can lead to failures” (L394). We can see in Figure 7 that in fact all of the 3C failures still occur at very low levels of precipitation, but it is true that now low levels of precipitation lead to crop failures at higher rates. My qualm here is really with the interpretation of the precipitation not DTR as you already discuss the limitations of the model with respect to DTR well in the discussion section.

**Minor comments**

Does EPIC-IIASA include changing management in the historical run or is it static?

Why do there only appear to be nine values in your partial dependence plots (Figure 4)? Or am I misinterpreting the ticks at the bottom of the plots on the x-axis?

Figure C7 - not required, but I’d suggest using different colors. It’s nearly impossible to distinguish the differences in the PDFs at the top of the graph using orange/red and yellow.

L 355 - I’d qualify that the change in temperature is much greater than the change in precipitation

L288-289 - It is not true that you see an increase in extreme dry years. The two blue PDFs are nearly identical, and the difference seems marginal. You certainly see an
increase in the failure rate at the low-end of the rainfall distribution though (e.g. “Failure 3C” PDF larger than “Failure PD” PDF). And because heat is increasing you will still see increasingly frequent joint warm and dry conditions.

Figure 7: change the axis labels to be interpretable to the reader instead of being the variable names used in the code

L 282 -283 (and relevant to 390-392) - The interpretation that compound weather extremes become less important in a warming world because extreme temperature alone drives failures is not necessarily true. It may be (and probably is) rather that compound weather extremes become more likely in your shuffled data when you uniformly warm all years. For example, if every year is an exceptionally hot year then your crop failures depend almost exclusively on rainfall, making the joint failure and shuffled failure years similar to one another at high return periods. In fact, we can see that all of the 3C failures still occur at very low levels of precipitation in Figure 7, but it is true that now low levels of precipitation lead to crop failures at higher rates.

L 255 - I don’t think the data support a robust increase of failure probability at higher precipitation levels unless you are willing to put equal weight on the corresponding dips in higher maximum temperatures leading to lower rates of failure and higher diurnal temperature ranges leading to lower rates of failure. I’d at least caution the reader that this may be noise rather than signal in the data.

L243 - define DTR at first use in the text. While it’s defined in the table it is not defined before first use in the text

L159-161 - what does assigning weights mean in this case? Was the dependent variable weighted? Are you running RF on binary data instead of crop yields generally? If so, justify this decision.

- Also, why does it matter if the failure observations are less frequent?

L158 - need to better define for the reader what a “data split” and a “shuffle” is in your cross validation procedure

- Described on L190

L100 - note that soybean yield data is available for this entire period going back to 1900 from USDA, but it is difficult to remove the management and technology changes trend. So there is still reason to use crop models and the work provides a useful complement to observation-based analyses.
L 37: “the majority of climatic shocks are compound events” - is this true?