Referee 2

Author response

Thank you for the opportunity to review this paper.

The study by Hamed et al. investigates the effect of growing season hydro-climatic conditions, including hot-dry compound extremes, on US soybean yield variability. In a first step, the authors identify a set of most important climate and hydrological predictors that affect soybean yield variability across the US. In a second step, they fit statistical models to county-level yield time series to examine the strength and direction of the relationship between hydro-climatic predictors and yield outcomes. In particular, the study finds that the co-occurrence of hot and dry events leads to more negative yield outcomes than the effect of hot or dry conditions alone would predict. The authors finally investigate the effect of historical hydro-climatic trends on soy yields. The authors show that historically, there have been wetting and cooling trends across important production regions in the US. However, in the same regions, compound hot-dry extremes increased in frequency. These results highlight that the effect of compound events may be masked when looking at statistical relationships of individual variables alone, without considering interactions between hydro-climatic extremes.

The paper is clearly and well written. From my perspective the manuscript is largely suitable for publication as it stands. I only have a few suggestions for the authors to consider which will hopefully help improve this paper for publication.

RESPONSE: We thank the reviewer for the positive feedback on our manuscript. We are grateful for suggestions to improve our manuscript. We respond to the comments given in the text below (in bold and italics text).

General comments:

Overall, the statistical approach is robust, and limitations are clearly presented in the text. However, I would ask the authors to consider the following suggestions:

1) Predictor selection

The authors apply a strict predictor selection process, which eliminates the occurrence of highly-correlated predictors – both at the same time as well as in subsequent months.

However, I wonder whether this approach eliminates predictors that do have an important effect on soy yields. The example presented in the text is: “we excluded soil moisture in September as August soil moisture was already selected”. I understand the reasoning to avoid collinearity, but it appears a little arbitrary – likely soil moisture would be relevant in both August and September (and potentially across the whole season).

RESPONSE: We agree with this point that was also highlighted by reviewer 1. In order to avoid arbitrarily selecting from collinear predictors, we adapted the methodology to no longer intervene manually with predictor selection and only monitor multicollinearity concerns using the variance inflation factor (VIF). In the latter case, a flag is raised if the VIF exceeds a value of 3 for any variable used to fit the final model at county scale (Carter et al., 2016; James et al., 2013).

Would it be more suitable to consider three aggregations for each variable (monthly, seasonal and the whole growing season) and select only one temporal aggregation per predictor in the final model? In this configuration, a predictor of “growing season soil moisture” could have been selected by the algorithm, if it was found to have the highest correlation with yields. This would lead to more interpretable results in the context of understanding climate influences on soy yields.

RESPONSE: We understand this concern and therefore ran a test with the suggested modification.

As an initial disclaimer, the general methodology has been adapted to run the selection process and model fitting at county level as this was a particular concern for reviewer 1. This implies that different counties can now have a different set of predictors. In addition, the methodology was adapted to exclude the manual selection step as stated in the response above. To limit the number of potential predictors to select from, we reduced the initial set of considered variables to only include Minimum Temperature, Maximum Temperature, Root Zone Soil Moisture and Excessive precipitation. These predictors are supported by main findings in prior literature that highlights the damaging effects of chilling conditions, high temperature, water stress and excessive rainfall on crops grown in the US (Carter et al., 2018; Gu et al., 2008; Li et al., 2019; Mourtzinis et al., 2015, 2019; Ortiz-Bobea et al., 2019; Zipper et al., 2016). We refer the reviewer to the response to RC1 for more details on the change in the methodology. An adapted overview of the overall modelling workflow is presented in Figure R1.

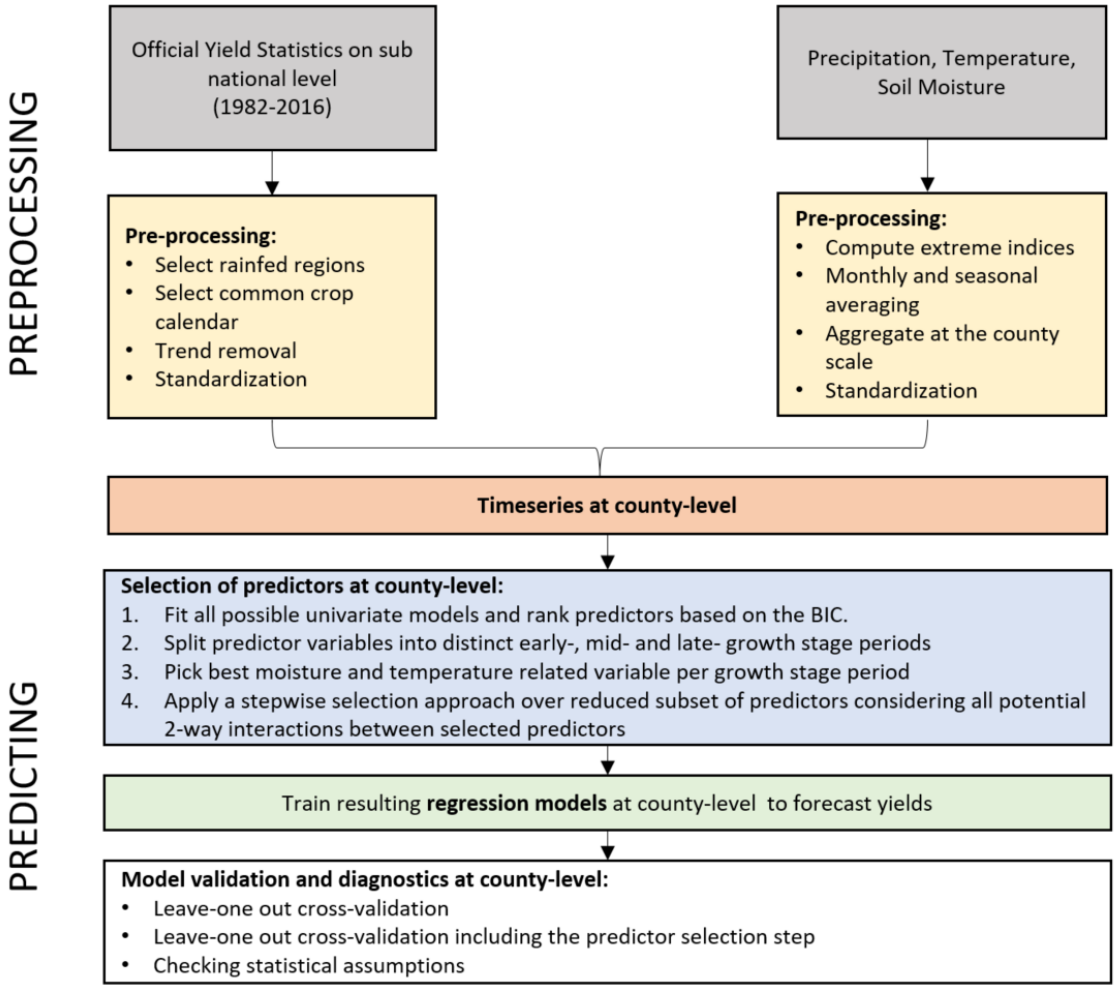

Figure R1. Overall modelling workflow applied for this study linking US yields to weather and climate variables

Back to the initial point, the reviewer suggested here a different selection approach where one temporal aggregation per predictor is selected in the final model. The main premise was to allow for growing season length predictors to be selected by the model if these were found to be most suitable to explain local soy yield variability. Consequently, for this exercise, we modified our selection approach to consider growing season predictors and to select one best temporal aggregation per predictor rather than two best moisture and temperature related variables for each distinct period of the growing season (i.e. early-, mid- and late). Results showed that the full growing season temporal aggregation was only picked up on very few occasions and only minimum temperature in northern states showed a clear signal for growing season length predictors (Figure R2). We believe that this can be explained by the changing sensitivity of soy crop yields to climatic variables across the season. For instance, warm temperature generally increases soy yields in early and late season but is associated to important reductions during the mid-season. Furthermore, short-term damaging conditions coinciding with particularly vulnerable stages of the crop growth cycle can result in important yield losses (Ben-Ari et al., 2018; Carter et al., 2018; Tack et al., 2017; Troy et al., 2015). Full season averages can mask out such details. The masking out effect can explain why full season predictors were seldomly picked-up throughout the adapted selection approach. Figure R4 displays general model performance. The model that contains season average predictors performs

qualitatively very similar to our initial setup (Figures R3 and R4). To keep the focus within this manuscript on the importance of timing with regards to soybean yield climate sensitivities, we prefer to keep this aspect of the method similar to what we initially proposed in the preprint and avoid the inclusion of seasonal averages. We will add a sentence in the revised manuscript saying that we've tested the inclusion of seasonal averages and these were not found to be critical for our setup.

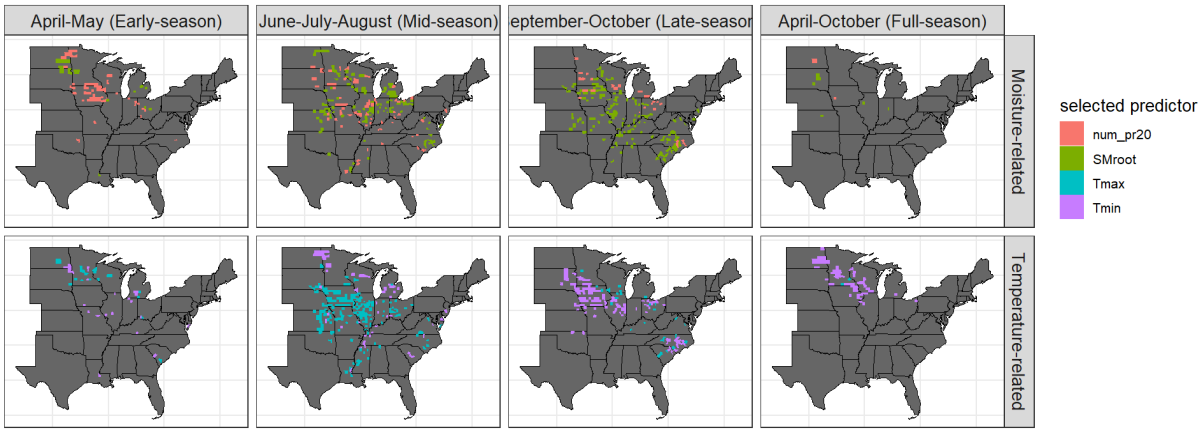

Figure R2. Region- and season- specific selected temperature and moisture related predictors when selecting for one best temporal aggregation per predictor.

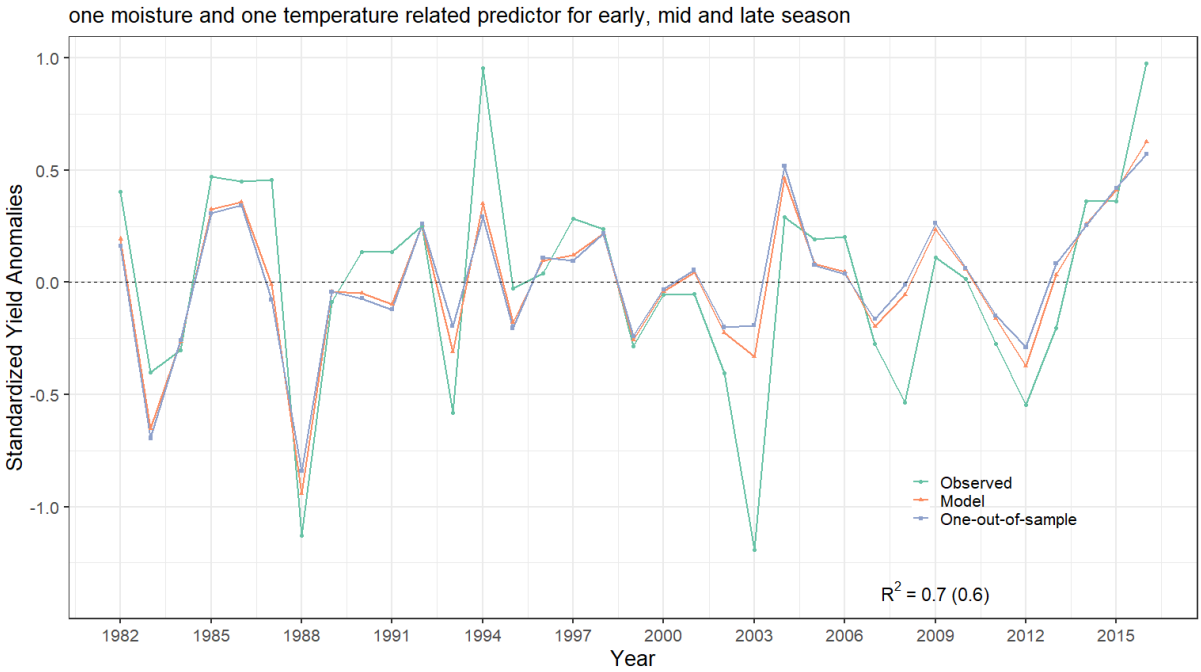

Figure R3. Average observations, model predictions, and out of sample predictions for model that selects one moisture and one temperature related predictor for early, mid and late season.

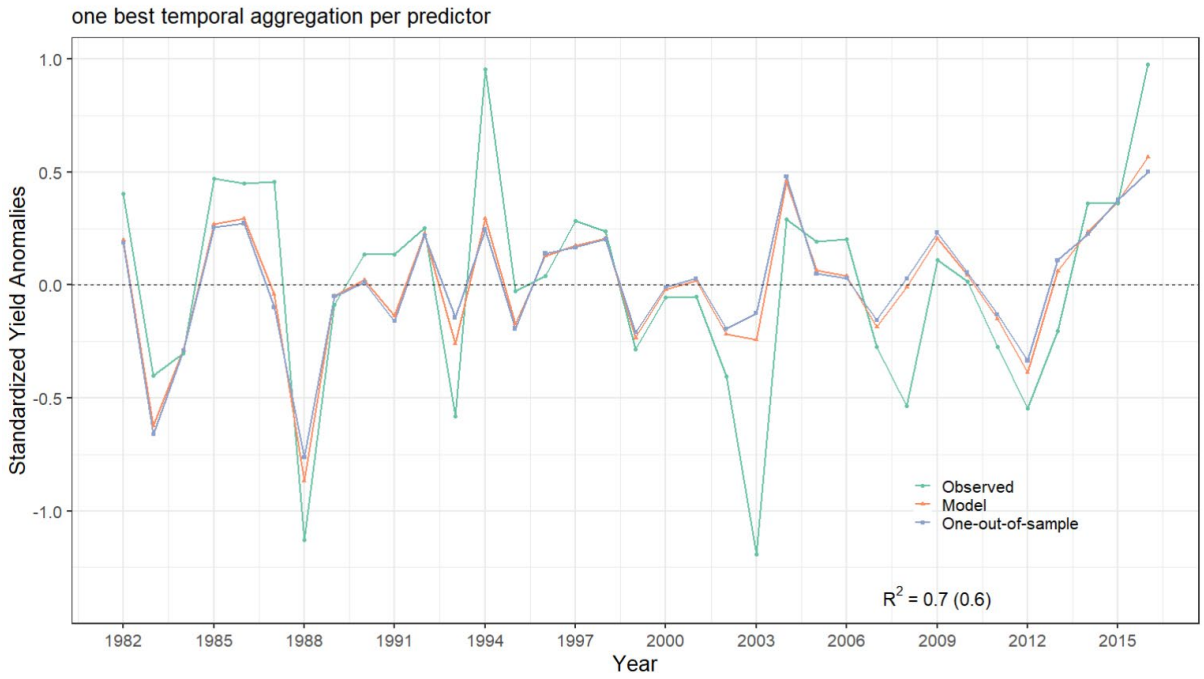

Figure R4. Average observations, model predictions, and out of sample predictions for model that selects one best temporal aggregation per predictor.

Similarly, it was not entirely clear to me why the authors selected two predictors per season (spring, summer and autumn) instead of selecting – for example – three heat and three moisture related predictors based on their individual predictive skill, irrespective of in which season they occur.

RESPONSE: The selection of two predictors per season (spring, summer and autumn) is motivated by our intent to highlight changing yield sensitivity to climatic variables across the season. This is an important element that has been discussed in recent literature (e.g. (Ortiz-Bobea et al., 2019) and references therein) and is one objective we focused on within this manuscript. Selecting three heat and three moisture related predictors irrespective of in which season they occur makes it harder to illustrate the changing sensitivity across the season. To make sure we are not compromising on model performance, we tested the proposed approach and present general model performance in Figure R5. Similar to the previous experiment, model results are not changed much and therefore we prefer to keep with the initially suggested selection approach proposed in the preprint. We will add a sentence in the revised manuscript saying that we've tested the proposed selection method and did not find important differences between the two setups.

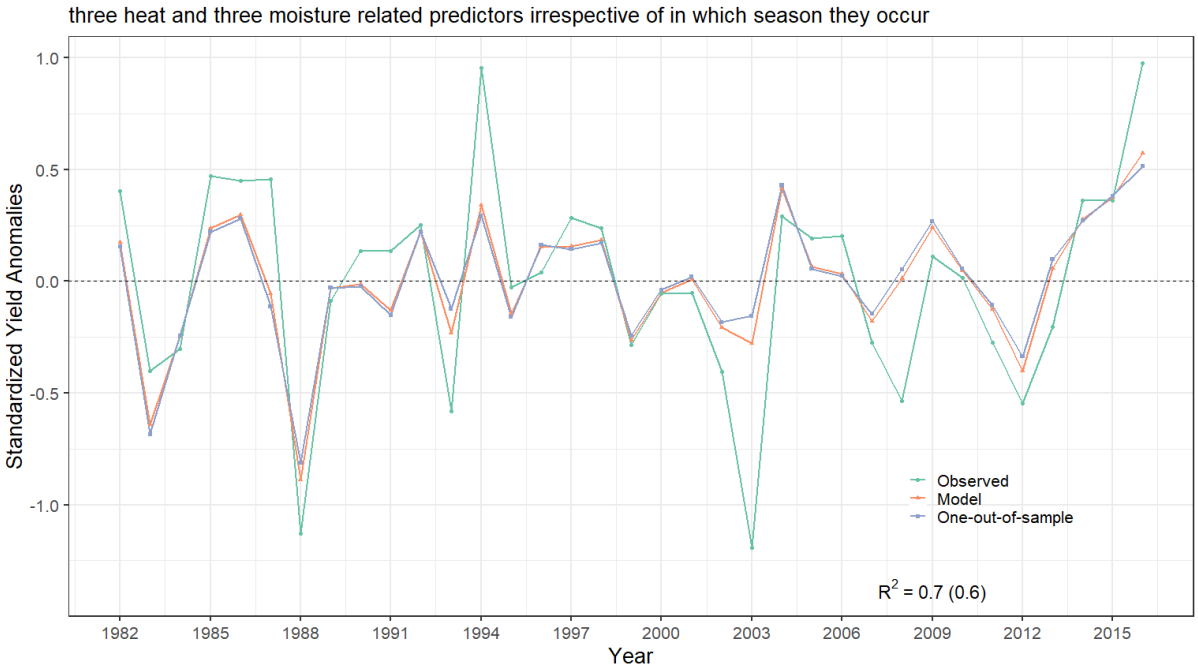

Figure R5. Average observations, model predictions, and out of sample predictions for model that selects three heat and three moisture related predictors irrespective of in which season they occur.

I have the impression, with the current way of how the predictors are selected, important predictors may be missed and less important predictors are selected. It would be great if the authors could test this or add a few clarifications in the text.

RESPONSE: We hope that tests and clarification provided above helped reduce the reviewer's concern with respect to predictor selection.

2) Cross-validation

I think it is great that the authors present the overall R2 and cross-validated (out-of-sample) R2 for their statistical models (given many studies only present an overall R2). However, the cross validation does not include the predictor selection step using the individual BICs and subsequent stepwise regression. Hence, the out-of-sample R2 will likely be over-estimated for new observations. Ideally, the cross-validation would include a predictor selection step for each iteration to obtain a true "out-of-sample" R2.

I understand the need to obtain one shared set of predictors to keep the results interpretable and to assess the influence of these predictors across all counties. I am not too concerned of overfitting, because the authors applied a very strict predictor selection process – only five predictors were selected based on all data points for the US (i.e. not selecting predictors for each county which would likely lead to overfitting) and the selected predictors are plausible. However, it should be mentioned somewhere in the paper that the cross-validation does not include the predictor selection step and may therefore lead to a potential overestimation of the out-of-sample R2.

RESPONSE: Thanks. We agree with the reviewer's concern with regards to the cross-validation method. Reviewer 1 had a particular concern with regards to predictor selection not occurring at county scale. As the reviewer thinks that the latter will be

specifically a reason of concern will respect to overfitting, we ran a cross-validation that includes a predictor selection step at the county level for each iteration to obtain a true "out-of-sample" R2. This does not only allow the calculation of a more conservative R^2 value but also allows to gain some confidence with regards to selected predictors (i.e. how frequently they are selected across iterations). We did indeed conclude a lower R^2 value although the sign of the predicted yields is still very much consistent with observations across the years (Figure R6). We also report the most frequently selected predictors (Figure R7) and associated timing within the season (Figure R8) in addition to how frequently these have been selected across the 35 iterations (Figure R9). Overall, Figures R7, R8 and R9 show high consistency with regards to selected predictors and timing within the season. We will adapt the initial preprint figures to include the true out of sample cross validation in the revised manuscript.

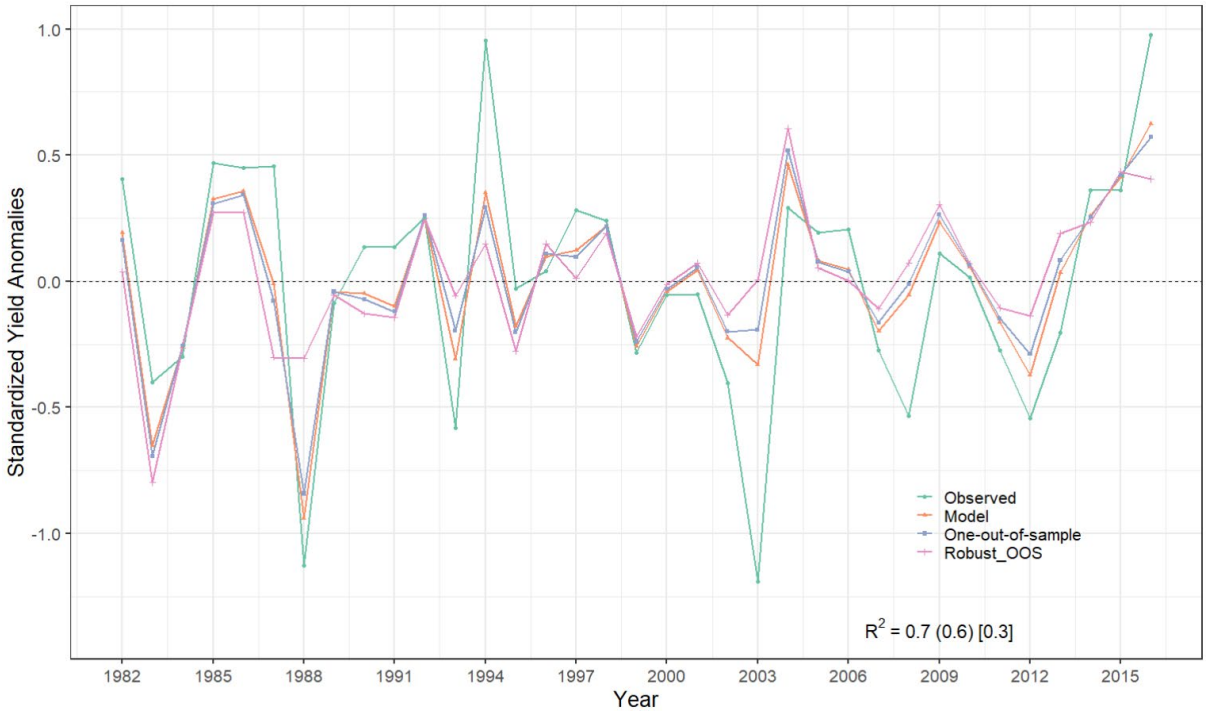

Figure R6. Average observations, model predictions, and out of sample predictions for model that selects three heat and three moisture related predictors irrespective of in which season they occur.

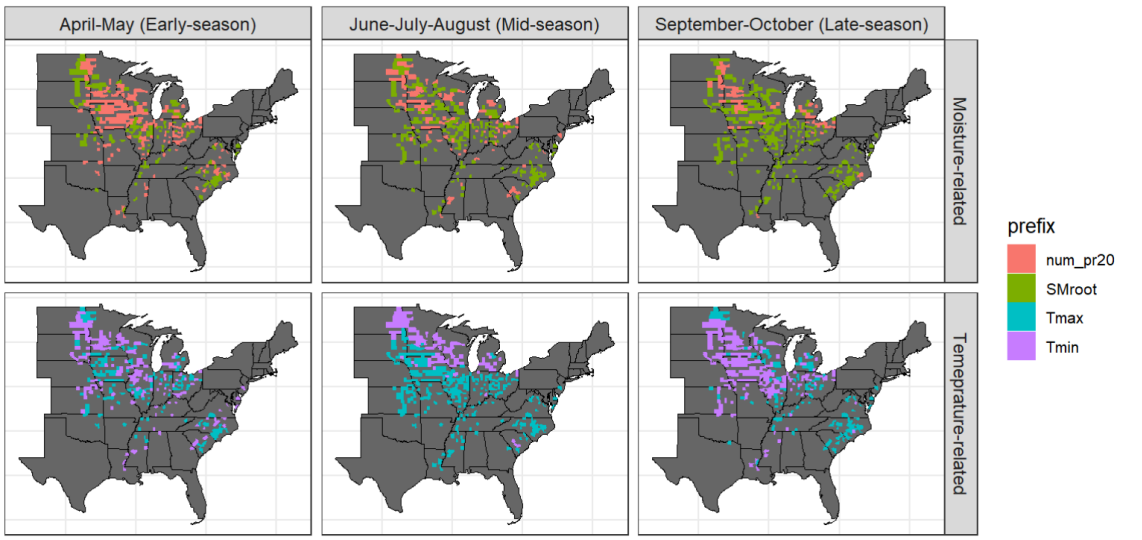

Figure R7. Most frequently selected predictors via the robust-OOS

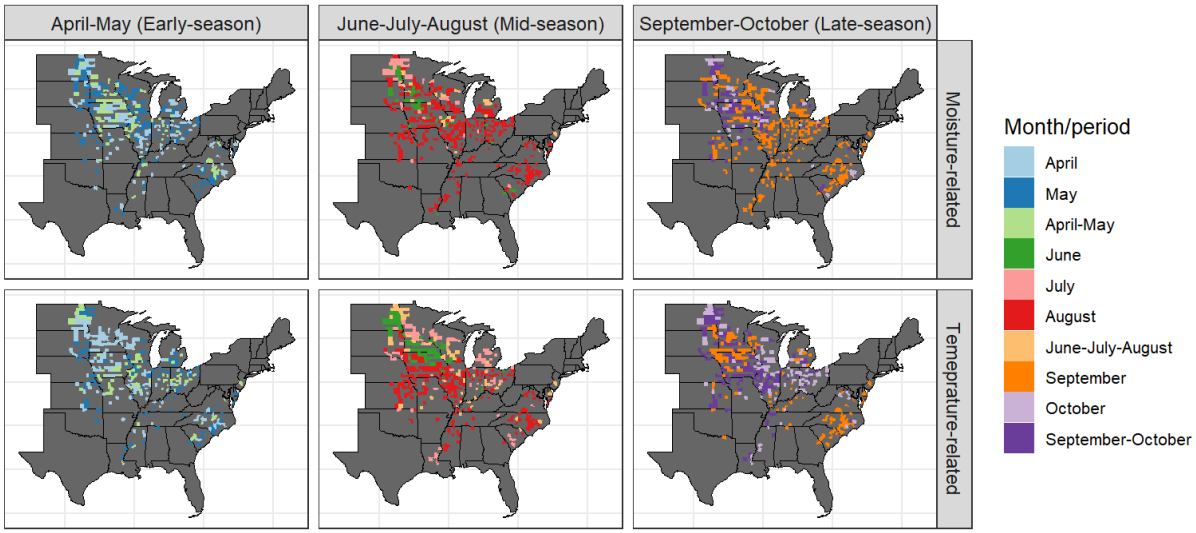

Figure R8. Most frequently selected timing via the robust-OOS

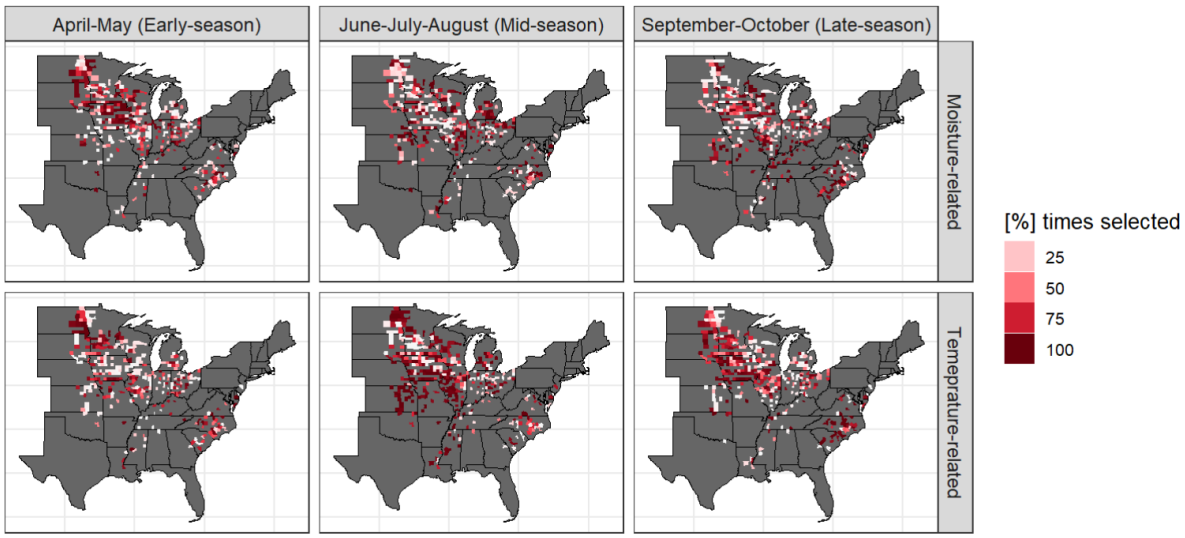

Figure R9. Frequency of the most selected predictor and timing via the robust-OOS

3) Climate and hydrological data

The authors used climate data from a global bias-corrected reanalysis dataset (WFDE5). Generally, reanalysis data can contain uncertainties and biases stemming from the earth system model used to generate the reanalysis dataset. At the same time, observation-based datasets also contain uncertainties due to the interpolation applied to the data. I would ask the authors to add a comment on why reanalysis data were used here and how the WFDE5 global reanalysis compares to observational datasets available for the US (or globally). Do the authors see any biases in the reanalysis data that could influence the results of their study?

RESPONSE: Reanalysis data was used here as these are available at daily timescales compared to the observation based CRU dataset available at monthly timescale. The daily time resolution allows for the flexible calculation of indices of interest (e.g. number of days with precipitation above a certain threshold). Furthermore, the WFDE5 dataset is specifically designed for impact studies with temperature and precipitation both bias-corrected using the CRU dataset for temperature and the CRU + GPCC datasets for precipitation (Cucchi et al., 2020). We opted to use global datasets as these make it easier to transfer similar impact assessments to other parts of the world. Nevertheless, we do see the value of leveraging as much as possible local observational data for impact assessments and we will add text on that in the revised manuscript. To monitor potential biases, a point also raised by the reviewer in the specific comments section, we calculated monthly correlations at grid-cell level between maximum temperature, minimum temperature and precipitation obtained from CRU and WFDE5 datasets (Figures R10, R11 and R12). All plots show practically perfect agreement. As CRU is not available at daily resolution, it was not possible to compare number of days with precipitation above 20mm between the two datasets. Still, we note that WFDE5 precipitation is adjusted using the CRU number of wet days variable (Cucchi et al., 2020). These very high correlations show that similar results can be expected from using the CRU dataset instead of the WFDE5 to train the statistical models. With respect to the estimation of the occurrence of joint extreme heat and drought, we checked whether similar years would have been selected if we used WFDE5 instead of CRU. For selecting hot-dry years, we calculated the percentage of grids during a given year where August maximum temperature is above the 90^{th} percentile whereas summer precipitation (JJA) is below the 10^{th} percentile. Years where the percentage of grids exceeded 15% were considered hot-dry years. We did a similar calculation for temperature above the 75^{th} percentile and summer precipitation below the 25^{th} percentile. The subset of hot-dry years is almost similar when comparing the two datasets. The only difference is that for the $90^{th}/10^{th}$ percentile pair, the WFDE5 reported 2011 as additional hot-dry year compared to the CRU subset (1983,1988). On the other hand, for the $75^{th}/25^{th}$ percentile pair, the CRU reported 2002 as additional hot-dry year compared to the WFDE5 subset (1983,1984,1988,1991,1993,1995,2003,2006,2007,2011,2012). We see this as a minor source of error that is not expected to significantly influence the results of this study.

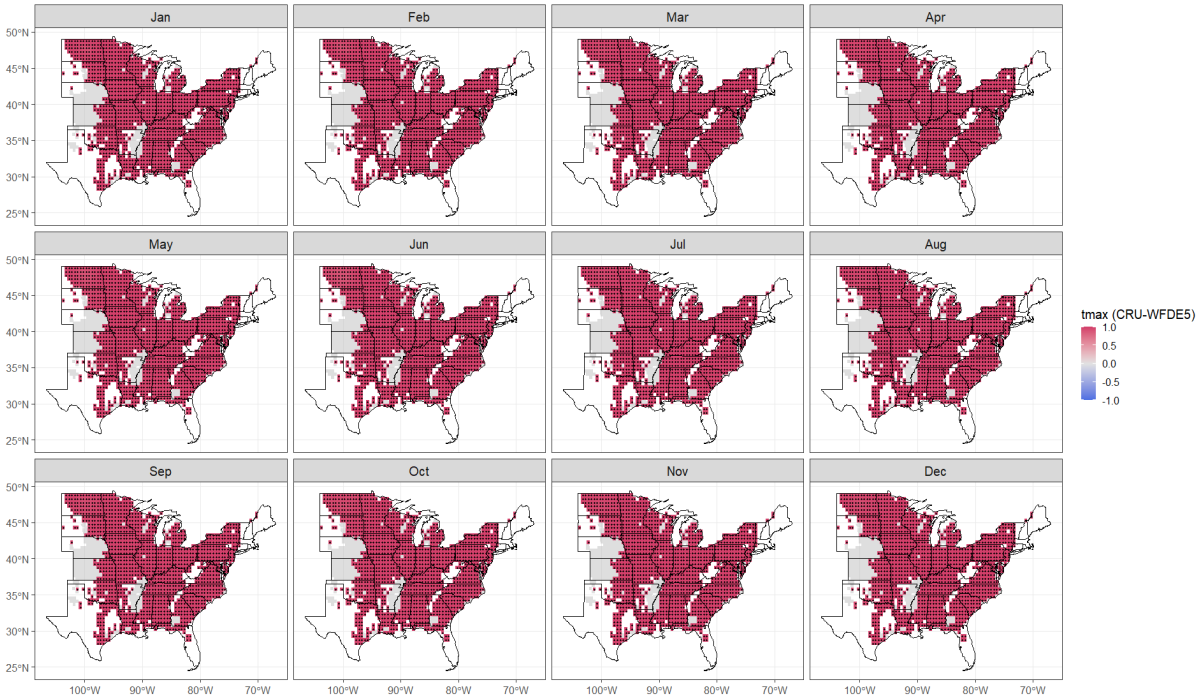

Figure R10. Grid-cell level correlation at monthly resolution for maximum temperature comparing CRU and WFDE5 datasets

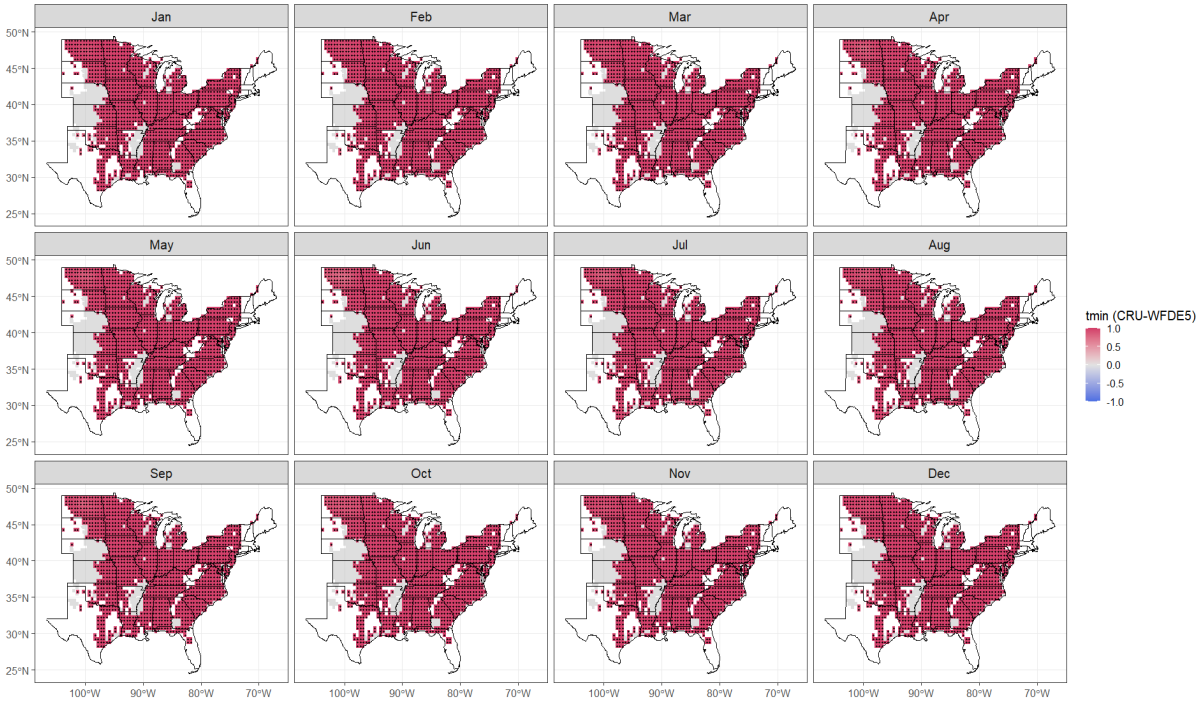

Figure R10. Grid-cell level correlation at monthly resolution for minimum temperature comparing CRU and WFDE5 datasets

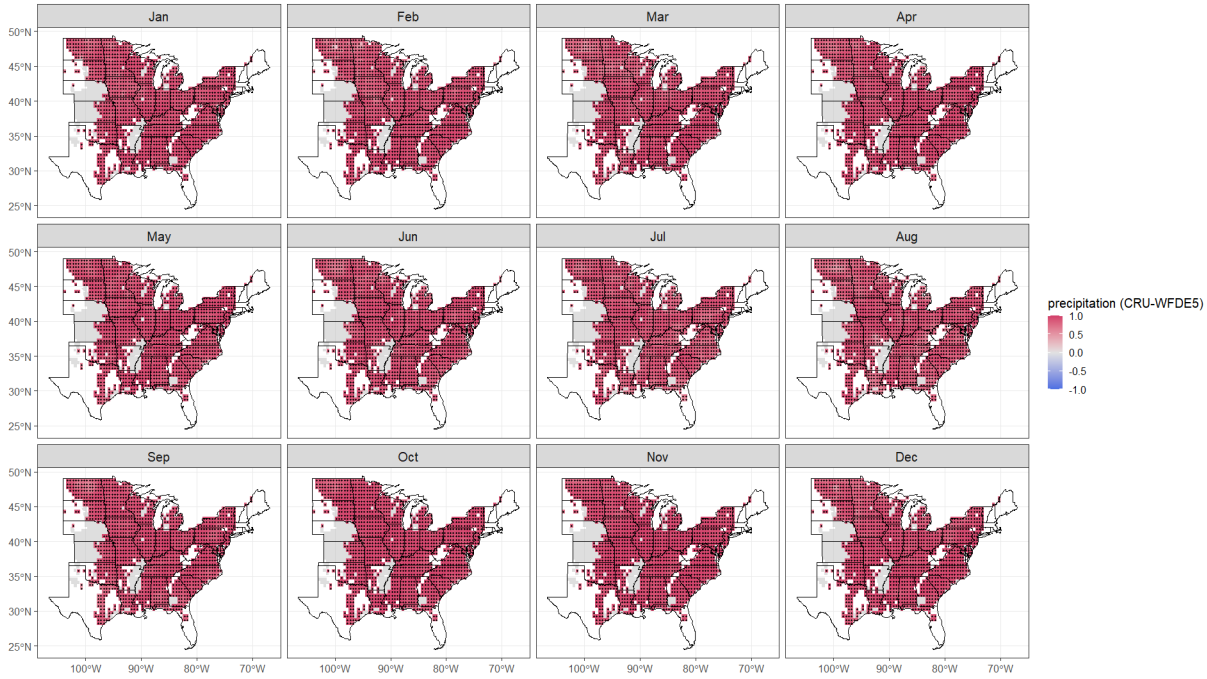

Figure R11. Grid-cell level correlation at monthly resolution for average precipitation comparing CRU and WFDE5 datasets

The hydrological indicators (actual evapotranspiration and root-zone soil moisture) are described as satellite-based, obtained from the GLEAM dataset. However, the description of the GLEAM dataset indicates that GLEAM uses a hydrological model to simulate soil moisture and actual evapotranspiration (instead of, for example, directly using satellite-based observations of soil moisture). I think a clarification in the text that soil moisture and actual evapotranspiration are not observed directly, but simulated, would help understand the data – as simulations and remote-sensed data have different uncertainties and potential sources of errors.

RESPONSE: Thank you for highlighting this. GLEAM indeed uses a hydrological model to simulate soil moisture and actual evapotranspiration. By satellite-based observations, we were referring to the assimilation of microwave satellite observations into the soil profile in addition to the use of microwave observations of the vegetation optical depth in the calculation of actual evapotranspiration (Martens et al., 2017). Nevertheless, we agree that it is misleading to call it satellite-based observations and therefore will amend the text accordingly in the revised manuscript.

Specific comments:

- Line 20-21: “Moreover, in the longer term, climate models project substantially warmer summers for the continental US which likely creates risks for soybean production.”

  Given the effect of future trends using climate model outputs was not within the scope of this study, I would suggest removing this sentence, as it is a bit vague. It might be best if the abstract includes a sentence on potential future research, e.g. future

studies are needed to understand the frequency of hot-dry compound extremes under climate change (similar to what was mentioned in the discussion).

RESPONSE: Thanks. We agree with the reviewer's comment and will adjust the revised manuscript accordingly.

- Line 74-75: "(ii) have 75 common planting dates (i.e. April-May)"

  Why was there a need for common planting dates? Would it not be better to include as many yield observations as possible, and instead subset the growing season into first, second and last third? Can you please explain this in the text?

RESPONSE: We selected for common planting dates as crops are reported to have different climate sensitivities depending on timing with respect to the crop growth stage (Carter et al., 2018). In order to not mix up the climate signal and facilitate the interpretation of results, we've selected for grid cells with planting dates starting in between the month of April and May. These are highlighted in purple color (i.e. planting month 5) in Figure R11. What is presented as 5 in the figure represents the bracket going from the 15th of April to the 15th of May.

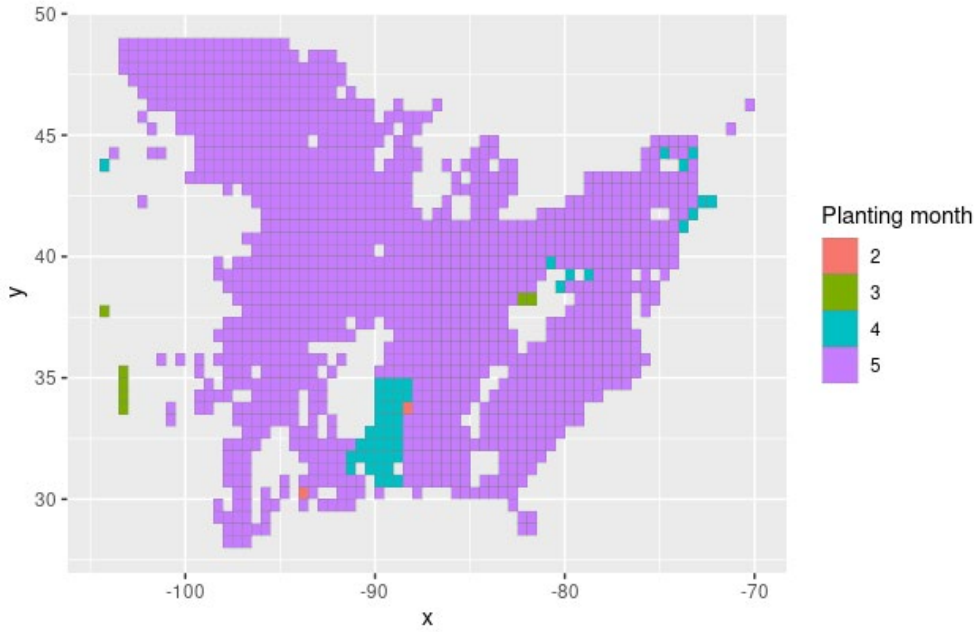

Figure R11. Planting month for rainfed soybean in the US using the MIRCA2000 dataset *(Portmann et al., 2010).*

Line 81-82: You applied a linear trend to the yield time series. Previous studies have applied cubic trends or more complex trend fitting algorithms to account for non-linear trends. Can you please confirm (possibly with a plot in the appendix) that visual examination showed that county-level yields follow a relatively linear trend?

RESPONSE: Figure R12 below shows raw averaged county-level yields (in orange) and linearly de-trended averaged county-level yields (in green). Upon visual examination, we believe yields do follow a relatively linear trend. This figure will be added to the supplementary material to be submitted along the revised manuscript.

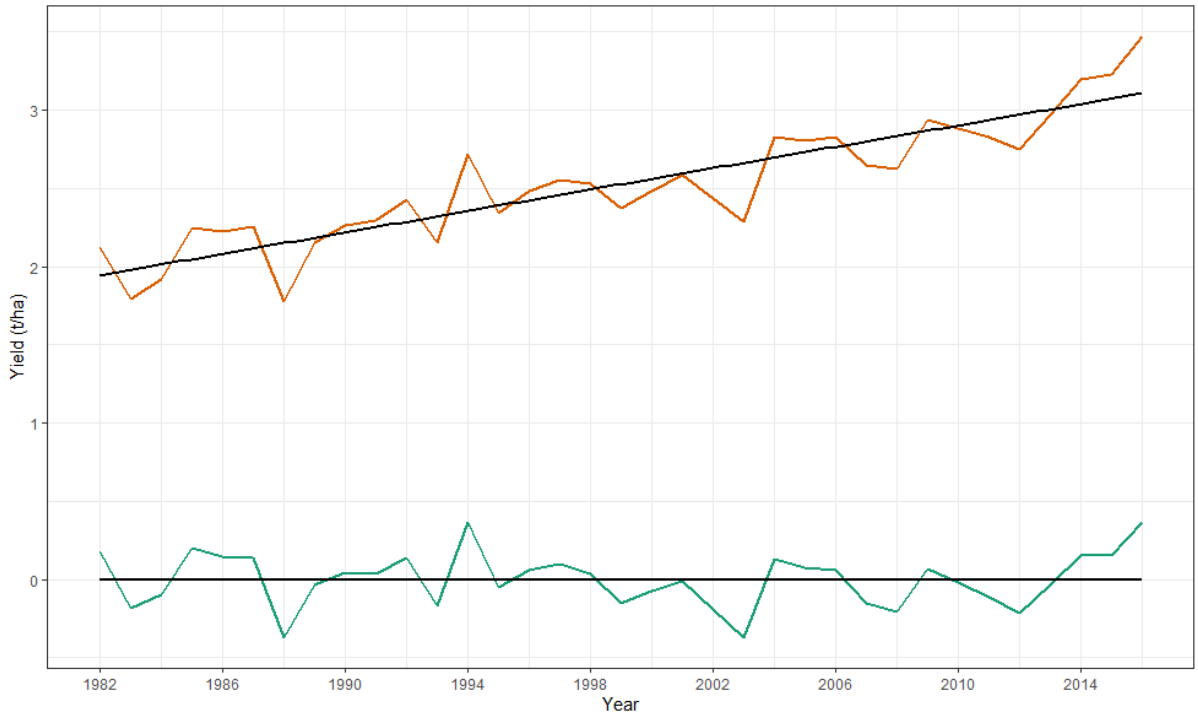

Figure R12. ***Raw averaged county-level yields (in orange) and linearly de-trended averaged county-level yields (in green).***

- Line 145-146: "To overcome this limitation, we used precipitation and temperature minimum and maximum variables from the CRU V4 global dataset (Harris et al., 2020) covering the period 1901-2019 at a spatial resolution of 0.5°."

  Why was this dataset not used to fit the statistical models, instead of the reanalysis dataset? This way you could be certain that the same data used for fitting the model is used to assess trends. Could you show the correlation between monthly Tmin, Tmax and precipitation in this observational dataset compared to the WFDE5 reanalysis (in addition to the correlations you show in Figure A.1)?

RESPONSE: We did not use this dataset to fit the statistical model as we wanted to include indices such as number of days with a precipitation above a certain threshold. These are only possible to calculate using a dataset that is available at least at daily resolution. Furthermore, we wanted to include soil moisture that is not available via the CRU dataset. We added the requested correlations under section 3 of general comments –climate data.

- Line 146-147: "Minimum temperature in the early season was used as a proxy for early season actual evapotranspiration…"

  Why did you choose minimum temperature instead of daily mean temperature (or the average of Tmin and Tmax)? Would this not capture the relationship with evapotranspiration more accurately, as it includes information on maximum daily temperatures as well?

RESPONSE: We agree with the reviewer on the possibility to use daily mean temperature as proxy for evapotranspiration. Our earlier choice was motivated by the

fact that minimum temperature was initially picked up as most relevant temperature related variable in spring in addition to literature papers that did report spring chilling conditions as a risk for soybean yields (Gu et al., 2008; Meyer and Badaruddin, 2001; Mourtzinis et al., 2019). Nevertheless, with the revised methodology, we no longer use actual evapotranspiration as a potential predictor for soy yields. It follows that setting this proxy is no longer needed.

Technical Comments:

- Line 168: I think it should by “county-level” instead of “country-level”.

RESPONSE: Thanks, we will adjust the revised manuscript accordingly.

References:

Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van Der Velde, M. and Makowski, D.: Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France, Nat. Commun., 9(1), doi:10.1038/s41467-018-04087-x, 2018.

Carter, E. K., Melkonian, J., Riha, S. J. and Shaw, S. B.: Separating heat stress from moisture stress: Analyzing yield response to high temperature in irrigated maize, Environ. Res. Lett., 11(9), doi:10.1088/1748-9326/11/9/094012, 2016.

Carter, E. K., Melkonian, J., Steinschneider, S. and Riha, S. J.: Rainfed maize yield response to management and climate covariability at large spatial scales, Agric. For. Meteorol., 256–257(March), 242–252, doi:10.1016/j.agrformet.2018.02.029, 2018.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Schmied, M., Hersbach, H. and Buontempo, C.: WFDE5: bias adjusted ERA5 reanalysis data for impact studies, Prep., (April), 1–32, doi:10.5194/essd-2020-28, 2020.

Gu, L., Hanson, P. J., Post, W. M. A. C. and Dale, P.: The 2007 Eastern US Spring Freeze : Increased Cold Damage in a Warming World ?, , 58(3), 2008.

James, G., Witten, D., Hastie, T. and Tibshirani, R.: An Introduction to Statistical Learning, 1st ed., Springer New York, New York, NY., 2013.

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E. and Peng, B.: Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States, Glob. Chang. Biol., 25(7), 2325–2337, doi:10.1111/gcb.14628, 2019.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A. and Verhoest, N. E. C.: GLEAM v3: Satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10(5), 1903–1925, doi:10.5194/gmd-10-1903-2017, 2017.

Meyer, D. W. and Badaruddin, M.: Frost tolerance of ten seedling legume species at four growth stages, Crop Sci., 41(6), 1838–1842, doi:10.2135/cropsci2001.1838, 2001.

Mourtzinis, S., Specht, J. E., Lindsey, L. E., Wiebold, W. J., Ross, J., Nafziger, E. D., Kandel, H. J., Mueller, N., Devillez, P. L., Arriaga, F. J. and Conley, S. P.: Climate-induced reduction in US-wide soybean yields underpinned by region-and in-season-specific responses, Nat. Plants, 1(February), 8–11, doi:10.1038/nplants.2014.26, 2015.

Mourtzinis, S., Specht, J. E. and Conley, S. P.: Defining Optimal Soybean Sowing Dates across the US, Sci. Rep., 9(1), 1–7, doi:10.1038/s41598-019-38971-3, 2019.

Ortiz-Bobea, A., Wang, H., Carrillo, C. M. and Ault, T. R.: Unpacking the climatic drivers of US agricultural yields, Environ. Res. Lett., 14(6), doi:10.1088/1748-

9326/ab1e75, 2019.

Portmann, F. T., Siebert, S. and Döll, P.: MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling, Global Biogeochem. Cycles, 24(1), n/a-n/a, doi:10.1029/2008gb003435, 2010.

Tack, J., Barkley, A. and Hendricks, N.: Irrigation offsets wheat yield reductions from warming temperatures, Environ. Res. Lett., 12(11), doi:10.1088/1748-9326/aa8d27, 2017.

Troy, T. J., Kipgen, C. and Pal, I.: The impact of climate extremes and irrigation on US crop yields, Environ. Res. Lett., 10(5), doi:10.1088/1748-9326/10/5/054013, 2015.

Zipper, S. C., Qiu, J. and Kucharik, C. J.: Drought effects on US maize and soybean production: Spatiotemporal patterns and historical changes, Environ. Res. Lett., 11(9), doi:10.1088/1748-9326/11/9/094021, 2016.