Review for gmd-2020-86: “Land Surface Model influence on the simulated climatologies of temperature and precipitation extremes in the WRF v.3.9 model over North America” by García-García et al.

In their article “Land Surface Model influence on the simulated climatologies of temperature and precipitation extremes in the WRF v.3.9 model over North America”, García-García et al. use 4 different land surface model (LSM) configurations coupled to the WRF regional atmospheric model over a regional North America domain, to explore the sensitivity of precipitation and temperature extremes to the choice of land model. They find similar overall patterns in the strength of land-atmosphere coupling across their simulations but show that the strength of that coupling can differ significantly across LSMs.

This study is a clear demonstration of how the uncertainty in temperature and precipitation is highly dependent on the choice of land model – a subject too-often overlooked in the study of extremes. The authors provide a valuable contribution to the role of the choice of LSM in the analysis of land-atmosphere coupling and modeling extreme events.

Prior to publication, this study would benefit from elaboration on how the authors’ results would vary had they explored multiple instantiations of the same LSM-WRF coupling, rather than a single instance of each LSM. Portions of the text were at times dense and confusing – a reworking of these sections (pointed out below) would greatly benefit the reader. Lastly, I would like to see a deeper analysis of the mechanisms driving the differences between the LSM results. The authors may consider this beyond the scope of their study, but explaining why the LSMs generate different results would be extremely useful to the community, rather than simply pointing out that different LSMs result in different distributions of extremes in temperature and precipitation. Following revision, this manuscript would be appropriate for publication in GMD.

**Specific Comments:**

**Major:**

- The authors refer to an “ensemble of simulations”. To clarify, they did not actually run an ensemble for each model setup, correct? Rather, there are 4 simulations in total evaluated: NOAH/WRF, NOAH-MP/WRF, NOAH-MP-DV/WRF, and CLM/WRF? The authors should clarify the text on this point.
- Related to (and more important than) the previous comment, the authors do not discuss what variability within a single LSM-WRF framework is expected. Had the authors instead ran an ensemble of, say, 10 NOAH/WRF simulations with slightly different initial conditions, how large a spread in the strength of land-atmosphere coupling and the statistics associated with extreme events would they see? Is the spread we’re seeing across the 3/4 LSMs simply due to the fact that WRF was run 4 times, or is it truly a physical response to the physics and mechanisms of the particular LSMs?
- A discussion of *why* the land-atmosphere coupling strength varies between their simulations would not only help show the differences between the 4 LSM/WRF simulations are actually the result of structural differences in the land model, but would

also be greatly beneficial to the reader for understanding why they should care about the spread in LSMs. Including a discussion on this topic would also help the reader select an LSM that appropriately represents the aspect of land-atmosphere coupling they may be interested in studying. I don't mean that the authors need to completely restructure the paper to address this topic. Rather, they often make statements like (Line 203, just an example) "LSM differences in the representation of VACa and VACb probabilities suggest the LSM influence on the evolution of atmospheric conditions". Rather than simply reporting differences in VACa-d, it would be useful for the authors to elaborate, and say something like "model YY has high soil moisture and cool temperatures, falling int other VACd category of land-controlled land-atm coupling. This results in <something about surface fluxes> and <something about **why** this model setup generates VACd vs VACa coupling, or no coupling>"

**Minor:**

- The authors should make it clear throughout the paper how many LSMs are being used. I would suggest saying 4 LSMs – the authors distinguish between NOAH & NOAH-MP but sometimes lump NOAH-MP and NOAH-MP-DV together (and sometimes evaluate them separately). It would be more clear if, through the whole paper, they refer to using 4 configurations of LSM: NOAH, NOAH-MP, NOAH-MP-DV, and CLM4; the inclusion of dynamic vegetation in NOAH-MP-DV is pretty fundamentally different than how the other land models distribute vegetation, therefore sometimes lumping it in with NOAH-MP just gets confusing.
- The domain appears to include ocean. If the domain isn't square & doesn't include ocean, please clarify that. If the domain does include ocean (which I assume to be the case), please clarify what method was used for SSTs (prescribed from climatology/satellite observations/reanalysis? Computed?), and how that method may influence the authors' results.
- Line 46: it would be useful if at or before this point, the authors gave a few sentences defining and giving examples of land atmosphere interactions.
- Line 102: "4 different plant functional types" – which 4? Regular CLM4 defaults to 14-16 PFTs.
- Line 119: missing from this section – how sea surface temperatures are handled
- Line 120-123 (also mentioned above): Are the 3 simulations your "ensemble"? Or, for each LSM, is an ensemble of WRF simulations run? If the former, please clarify. If the later, please provided details (# of ensemble members, how they were initialized)
- Line 121-122: Wording. "The rationale for this decrease in resolution is that this set of simulations constitutes an ensemble of WRF sensitivity experiments". The **rational** is the computational resources. You can still get meaningful results because you're doing a sensitivity study to the LSMs, rather than trying to reproduce obs. The result is a set of 4 WRF simulations that you're calling your ensemble.
- Line 135 / equation 1: This was pretty confusing the first time I read it through, and continues to be a hurdle for the reader through the text. I clear walkthrough of the conditions supporting each VAC situation in the text (and the conditions where there is no VAC) would be super helpful here, along with a description/example of the kind of coupling expected from each category.

- Table 2: a description in the text of the extreme statistics used would be hugely useful. A few of them were talked about near the end of the manuscript, which helped, but introducing them (more than just in the table) here in the methods would make the rest of the manuscript make more sense.
- Line 177: "… using daily data from three Evaluation simulations" – what are Evaluation simulations? (Can go look at table S1, but it still doesn't tell me what an "Evaluation simulation" really is, it just tells me what models were used as "Evaluation simulations".
- Line 185: "Atmospheric forcing controls surface processes at middle and high latitudes" – it controls processes more than land does, but it still appears to have <50% control. Please clarify. This is also why a walk-through of the 4 VAC terms in the methods would be useful: what happens if you aren't in a VAC category? Then there isn't strong land-atm coupling. Make that clear, and make it clear what the % in figures 1-2 are - % of all time that each VAC dominates, or % of time when *any* VAC dominates that the specific VAC in question dominates?
- Figure 1 (related to above): These don't add to 100, so can you a add a comment to the methods where you describe vac_a - vac_d on what happens if none of the 4 are true?
- Figure 1: consider using a different significant mask, e.g. putting dots over the non-significant portions, or mask-nonsignificant values with a nan, as it the dots obscure the part of the pattern that matters (I can't tell difference between anything except darkest blue or deepest red when it is under a dot, and those are the only values I really should be looking at)
- Line 193-204; Line 205-216: I found these two paragraphs pretty hard to follow. I think it would be easier to follow if the authors included some discussion about *why* each model was under atmospheric / land control in various regions / seasons. E.g. is it the evaporation, or the temperature, or both terms? As it is, I just did a lot of "read one sentence; look at figure 1 (or 2); read next sentence" without being quite sure what was interesting/important about the patterns.
  For example:
  - Line 208: "episodes over the Mexican coast is higher in CLM4… than … NOAH-MP-DV simulations in DJF **because YYYY"**
  - Line 212: "the VACc **(ie low SM and high TAS, land control due to soil moisture limitation)"** or something like that – help the reader understand what they're seeing and why what they're seeing is cool!
  - Generally, when the authors make a statement about what VACx did, it would be helpful to accompany it with something about what that means for TAS, LH, soil moisture, land control, atm control, etc - give more help to the reader, otherwise the meaning is just lost in a bunch of acronyms, especially for those unfamiliar with VAC metrics where it wouldn't be immediately obvious/intuitive as to what the IMPLICATIONS of being in VACc or VACd are).
- Line 213: at this point I wanted to see a breakdown of VACc and VACd. It is in the supplement – maybe point the reader to it here?
- Line 218: "extremes" -> "extreme indices as described in table 2"
- Line 218: "their means" – the mean of the extremes? Or just… the means of T and P?
- Line 220-222: "the WRF ensemble mean…" This is confusing since we just went from talking about a bunch of different LSMs in WRF to now discussing a WRF ensemble mean. Are we now talking about the CORDEX? What is the WRF ensemble mean? Do

you mean the mean of all 3 LSMs? Or were ensembles run for each LSM/WRF combination? If the latter, that wasn't clear in the methods.

- Section 4.2 in general: This whole section I was pretty puzzled about what was happening. Are the different LSMs no longer being evaluated/compared? Is this section just laying the ground work for what "normal" WRF looks like, then how it deviates with each LSM will be explored later? If so, please make that clear. If not, what happened to the LSM comparison? I don't think there is anything wrong with the *content* of this section, it just needs some additional motivation/transition text to allow the reader to follow why we're no longer hearing about the LSM comparison, which up to this point was the focus.
- Line 230: "Greenland, GRL" - Is Greenland actually in your domain? It isn't shown in any of the figures up to this point. And the region highlighted in Figure 4 is not Greenland. Maybe call it "Hudson Bay" instead?
- Figure 3: D = Days? (in color bar legends?) please clarify. Also, consider moving color bar labels below color bars, or putting more horizonal space between plots, so it is clear what unit goes to which color bar
- Figure 4:
    - maybe add x labels to each black-outline-box? hard to go all the way up from the bottom.
    - Are red values in "cold events" very cold, or very not-cold?
    - necessary to add a discussion of each of these extreme metrics to the methods section, more than just the table. Eg "CSDI measures YYYY; a high value means YYYY, while a low value means YYYY", and so on for each metric. (already mentioned this above, but it would be helpful for understanding this figure)
    - See above statement/question re: Greenland
- Line 241: clarify – warm events get longer but aren't as hot?
- Line 243: "all simulations represent a similar spatial pattern of the climatology of extreme indices" – was this shown? I thought Fig 3 (the climatology) was the average of all the runs. If it wasn't shown, please make it more clear what *was* shown, and point to a figure (main text or supplement) to support this sentence.
- Line 249-252: Any insight into why this might be? (CLM4 yields the highest temperatures, NOAH gives the weakest T and P extremes)… do they have super different soil moisture, different surface energy fluxes, produce different boundary layer stabilities…?
- Line 254-256: "simulations show similar spatial patterns…" – it would be nice to include some discussion of how much the land model matters (ie are they each behaving similarly, and that is why they look the same?), vs how much the extremes are set by topography, latitude, atmosphere / distance from ocean, etc.
- Line 258: "coldest night in DJF" – this is a nice concise description of one of your extreme metrics, nice! It would be great to have something like this for each of the metrics, and have it introduced in the methods (and when you talk about them in the results, rather than just reporting the acronym – it'll help the reader understand what is happening and why it is interesting).
- Line 262-264: Again, some discussion of what is causing the spread here would be useful, though maybe beyond the scope of what you'd like to cover in this study. E.g. are the ones that are super hot the ones with low soil moisture?

- Line 283-284: add some discussion… do these places correlate with dry regions? Regions of high topography? What might be generating the spread?
- Line 298-305: I found this section really hard to follow, I think because I wasn't sure what I'm supposed to be taking away from it. Needs more "why" elaboration.
- Figure 5: would be helpful to revist the extreme indices in the caption (Txx = …, TNn = … etc.)
- Line 325: Another nice helpful interpretation of the figure/acronym with "less frequent cold nights (TX10p)" – thanks! Working more text like that in would help the reader follow!
- Line 328: re: CORDEX simulations - Was WRF in CORDEX? Were other regional atm models using CLM? How does your CLM run compare to CORDEX CLM runs? How does your WRF compare to CORDEX WRF?
- Lines 338-341: another place where it would be helpful to do more hand-holding for the reader on why what is being reported here matters
- Line 345-346: precip extremes are more uncertain across CORDEX simulations than WRF simulations -> this would be expected, would it not, as the CORDEX simulations use a variety of different atmospheric models? How does the uncertainty from the choice of atmospheric model compare to the uncertainty from the choice of land model?
- Line 347-348: "…regions with large uncertainties in the simulation of precipitation extremes among the WRF simulations are also identified as areas with large uncertainty across the CORDESX ensemble." This is interesting! Suggests there may be a robust signal.
- Line 354: "results to other model ensembles" … this seems more like a single atmospheric model study exploring the sensitivity of WRF to perturbed surface fluxes (where the surface fluxes are perturbed by using different LSM components).
- Line 368-369: "The similar uncertainties of extreme evens in the CORDEX ensemble relative to the WRF simulations suggest that the LSM component may be an important source of uncertainty in the CORDEX ensemble." I don't follow the reasoning here. The CORDEX runs use different LSMs yet show similar uncertainties in extremes, so wouldn't that suggest that the LSMs aren't the driver? Please clarify/elaborate.
- Line 370: as with the previous sentence, I'm confused if the authors are trying to talk about how the CORDEX & WRF simulations are similar, or how they're different.
- Line 373: So, the spread in uncertainty within WRF (but with different land models) is bigger than the spread in uncertainty within the CORDEX simulations? Or is the spread within the two sets of simulations being compared to the spread between the two sets of simulations? Please clarify/elaborate.
- Line 384: similarities between WRF and CMIP5 mean – elaborate on why they are the same / what is controlling the DJF coldest night / JJA hottest day?
- Line 402: "or in India" -> referring to a specific heat wave (like the 2003 Europe one), or just India in general?
- Line 407: "… depending on the employed LSM component **because YYYY**" (would be nice to have some because/why discussion here)
- Lien 409: What is the authors' recommendation for selecting an LSM? (I don't mean the authors need to pick their favorite, I just would like to see a list of considerations for selecting an appropriate LSM for one's study)
- Line 419: "land atmosphere interactions **as measured by YYYY**"

- Line 422-423: especially since this is the conclusions, would give a short word sketch on what being in the VACa-d category means.
- Line 430: Include a statement on how much you think your results are the LSM differences, vs 3 instantiations of WRF -> e.g. if you initialized a slightly perturbed CLM-WRF, how different would you expect it to be from your other CLM-WRF, vs how different the various LSM-WRF simulations are?
- Line 432: This sentence would suggest previous comment is mostly LSM dominated, but some explicit discussion of the topic would be nice.

**Typos/grammar:**

- Line 6: "four simulations performed by the WRF model using three different LSMs from 1980 to 2000" – this makes it sound like the LSMs are from 1980 to 2000, but I believe the authors mean the simulations are run from 1980-2000, using three different LSMs
- Line 47: "off-line" is "offline" everywhere else
- Line 83: authors "define" NOAH-MP-DV in brackets twice, just need it once
- Line 88-89: I think the second "as" is a typo, but I'm not sure what the authors are trying to say so I don't know how to suggest fixing it. "The NOAH LSM has been extensively used for reanalysis prodcuts, as well as for RCM simulations **as** those participating in the CORDEX project…"
- Line 92: missing citation. Perhaps the NOAH technical description? https://ral.ucar.edu/sites/default/files/public/product-tool/unified-noah-lsm/Noah_LSM_USERGUIDE_2.7.1.pdf
- Line 121: "indeed" is a typo. "… counter-intuitive for a RCM experiment; indeed. The …"
- Line 142-143: typo, I'm not sure what the authors are trying to say. "… clouds and precipitation, which leads to **low vegetation activity likely rising soil moisture**."
- Line 149: "series" -> "time series" (or if that isn't what the authors mean, what is a LH series?)
- Line 162: "techniques techniques" typo
- Line 162-163: "… for the study of future climate trends and climate variability, since **they** have been proven to modify the spatiotemporal consistence of climate models as well as internal feedback mechanisms and conservation terms." This sentence is confusing; in particular, is "they" referring to future climate, or bias removal?
- Line 234: "more frequent cold events than the rest of LSM components" -> "rest of **the** LSM components"
- Line 281: "simulations in about 35 days per year" -> "simulations **by** about" ?
- Line 314: "range among WRF simulations" -> "range among our 4 WRF simulations" (unless you just used 3 – confused if NOAH-MP-DV gets used all the time or not)
- Line 323: "WRF ensemble" – see earlier comment re: confusion about what your ensemble is
- Line 366: "Thus we compare **each** model's uncertainty…" (insert "each")
- Line 367-268: "despite they used" -> typo. Maybe "despite the fact that they used" ?
- Line 404: "point out to a future" -> "point to a future" (drop "out")

- Line 405-406: typo somewhere, but I'm not sure what the authors are going for thus not sure how to fix it. "Climate model simulations are our best source of information to inform measure against climate change impacts."
- Line 419: "WRF simulations **over North America**" (specify region is North America)