EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-437-RC1, 2022

Comment on egusphere-2022-437

Anonymous Referee #1

Referee comment on "Sensitivities of subgrid-scale physics schemes, meteorological forcing, and topographic radiation in atmosphere-through-bedrock integrated process models: A case study in the Upper Colorado River Basin" by Zexuan Xu et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-437-RC1, 2022

Referee comments on:

Title: Sensitivities of subgrid-scale physics schemes, meteorological forcing, and topographic radiation in atmosphere-through-bedrock integrated process models: A case study in the Upper Colorado River Basin
Author(s): Zexuan Xu et al.
MS No.: egusphere-2022-437
MS type: Research article

Summary

This study evaluates the influence of different meteorologic forcing (based on different reanalysis datasets), subgrid-scale physics schemes, and terrain shading on simulated hydrometeorology. They find that physics configurations result in more variance in simulated hydrometeorological conditions, and that meteorological forcing has a smaller impact. This type of sensitivity study is important to understanding where and how to focus further model development and observational field campaigns (as the authors note), and this particular study evaluates some sensitivities that I have not previously seen addressed. In my view, this has the potential to be a highly valuable contribution, but I believe it could use some sharpening of its framing, earlier recognition of the problems across all model configurations with respect to streamflow simulation, and more quantitative comparisons of some of the results.

Major comments

One major comment is around framing: at times, the authors imply that these results show an optimal IPM configuration but this is never clearly evaluated. At other times, the authors note that validation against observations is not a major goal of this study – in which case, it cannot indicate an optimal IPM configuration. My recommendation is to avoid implying that an optimal configuration is identified here. On a related note, I think the poor simulation of streamflow by the IPM should be mentioned earlier (perhaps in the abstract) – while it’s ok that this is the case, having this result buried in Figure 7 felt a bit deceptive.

Similarly, the authors refer in the introduction to recent arguments by Lundquist that models may be outperforming observations. In my view, they then miss a relatively easy opportunity to contribute to this debate: adding a ParFlow-CLM run forced by PRISM and reporting the results in Figure 7 would provide a case study testing whether meteorological models or observations are indeed more accurate in this case (assuming we basically believe that ParFlow-CLM is not biasing the results so much as to invert this response). I’m loathe to be the reviewer who suggests the authors do a different study than the one they have done – but in this case, the introduction led in this direction, and one additional simulation would significantly enhance the value of the present work.

Finally, it would have been useful to see more quantitative model evaluation, and some description of model evaluation in the methods. I had two specific concerns about the identification of BSU-CFSR2 as the “best” model and that used for the topographic radiation evaluation. First, I didn’t see a quantitative evaluation of models against PRISM to make this evaluation. Why not report an NSE or RMSE? Second, given the idea that PRISM is not necessarily more accurate than WRF, I’m not sure how important PRISM is as a benchmark here. Could you analyze the impact of topographic shading for two model configurations with very different results? Assuming you find similar results for a different configuration, it would just be helpful to have a sentence confirming that evaluating topographic results in a different WRF configuration had similar results.

Minor comments

Line 21 – Based on only the abstract, it's not clear to me how the "spatiotemporal variance in simulated hydrometeorological conditions" is defined. I think you mean the model response varies more across the model structure options than meteorologic – but from this sentence, another possible interpretation is that spatiotemporal variance itself (e.g., the variance of some response variable across grid cells) is greater in certain physics schemes. Is there a way to avoid this ambiguity?

Line 27 – The conclusion that these findings provide guidance on the most accurate IPM was a bit of a jump from the prior sentences, which just described model sensitivity. To justify this, it would be better to describe what analysis supports this guidance (a calibration, I presume? Against what variables?). Alternatively, your concluding point could note that these sensitivity analyses show where more effort should be focused to constrain our process-based understanding.

Line 34 – remove "that"

Line 35 – "may have"? Could you express the reason this is stated with uncertainty?

Line 44 – Is "relevant" here meaning for larger-scales? Or respective relevant scales for each process?

Line 46 - Tying the motivation for this article to recent discussions about the relative skill of process-based atmospheric models vs gridded interpolated datasets provides a great motivation for the present study.

Line 58 – “To further compound…” I think this is a good point, but could you provide an example?

Line 68 – I’m a little uncomfortable with “properly-configured” unless you feel this analysis truly fixes equifinality issues. Maybe “appropriately-configured”?

Line 120 – “We can establish” leaves the reader uncertain if you did this or not.

Line 121-126 – This motivation is very nicely stated (although I don’t think it’s a hypothesis in the context of this study) – could you state this explicitly in the abstract?

Line 127 – Is “observations” here meant to refer to gridded reanalysis products? As the Lundquist paper points out, those are also models (generally statistical interpolations), so I’d suggest another word. I also note that this section doesn’t say anything about identifying an optimal model configuration, which is an outcome highlighted in the abstract.

Line 141 – "representative" is a bit of a tough argument to make – consider "similar to many other basins in…"

Line 141 – "near" should be "nearly"

Line 153 – You noted a lack of observations earlier, which disconnects somewhat with the "heavily-instrumented" claim here. I think this could be mitigated by noting that the instrumentation is intense at this site, but it's extremely difficult to observe many processes with high accuracy at relevant scales.

Figure 1- As I read through the rest of the paper, I found I needed a more detailed study area map for the ERW specifically – with elevation and streamlines, perhaps?

Line 254 – I have trouble understanding why PRISM was used to assess model performance for meteorological fields, given the comments in the Lundquist et al. (2019) paper you cited. It seems fine to compare against PRISM, but perhaps not to "assess model performance."

Line 266 – Could you note the spatial resolution of the ASO product used here? At 50 m, point-to-grid errors could be one reason for the apparent underestimation by ASO relative to SNOTEL.

Line 272 - Results section would be easier to follow with if subheadings were included.

Figure 3 caption – would read more easily if you noted a-c in your descriptions of which variables are identified. The statistics used to evaluate these differences are essentially introduced in this caption; could you move that to the methods?

Figure 3 – I'm surprised the UCD configurations melt so much earlier when they don't appear to be warmer. Is it possible that the spatial averages here obscure spatial differences that would explain why the UCD simulations melt earlier? Figure S-4 kind of gets at this, but I think it needs more interpretation for the reader.

Line 319 – run-on sentence.

Figure 4 – Nice figure. Could you again add an introduction to these statistics in the methods so we know how you're evaluating variance earlier? Why do c and d have only two points marked on the x-axis?

Line 335 – Were there any quantitative statistics provided to determine that BSU-CFSR2 agreed best with PRISM?

Figure 6 – Some panels appear not to use their full color scale (e.g., Temperature). Is that due to outlier pixels? There’s a lot of wasted white-space in these maps – why not use the full plotting area for each map?

Line 380 – This paragraph describes Figure 7, but the next paragraph also seems to introduce Figure 7 as though it’s a new topic?

Line 401 – “The objective of this study is not to replicate the observations…” In that case, I strongly recommend changing the final sentence in the abstract, because that implies you’re identifying the best model configuration.

Line 413 – Are the differences notable or minimal? I would say minimal. Maybe better to describe quantitatively – you could note the among-model variance vs the seasonal variance?

Line 417 – “are slightly larger…” The differences are twice as big for the subgrid-scale physics schemes but are small in both cases; I would suggest rephrasing to clarify.

Line 420 – What is meant by "more muted-nature"? I think this sentence speculating about differences in groundwater signals across years would be better in the discussion.

Figure 8 – Is this color gradient perceptually uniform? It appears not to be (e.g., see Figure 1b in Cramer et al., 2020). It would be helpful to see a perceptually uniform palette here if possible.

Line 448 – "with an eye towards how to represent…" Without calibration or serious validation efforts, I don't think this study tells us about how to represent these interactions in models. I do think it tells us about where the most important uncertainties are, though (in your next sentence).

Line 454 – I don't remember a prior discussion of boundary conditions – is this referring to boundary conditions at the land surface driven by differences in the subgrid-scale physics schemes?

Line 456 – This would be more convincing if statistics on BSU-CFSR2 vs other models were presented. How does identifying this configuration allow researchers to prioritize process studies and observational constraints? What would these be, specifically?

Figure S6 – Could you use a different color scheme that doesn't have a diverging gradient? I think the diverging gradient is most appropriate for your maps showing differences (e.g., value scales that center on zero).

Line 467 – “Latent heat is posited…” by whom? Are you? I think you could state with more confidence than “posit” that other energy balance components (including but not exclusively latent) mediate the influence of shortwave spatial variability on temperature spatial variability.

Line 470 – You lost me here. This paragraph is ostensibly about how terrain shading algorithms affect radiation flux? How does this affect our ability to extrapolate findings from one mountainous watershed to another? The multiple “if” statements in here are also a little confusing – did the present study show these things or not?

References

Crameri, F., Shephard, G.E. & Heron, P.J. The misuse of colour in science communication. *Nat Commun* **11,** 5444 (2020). https://doi.org/10.1038/s41467-020-19160-7