Referee Comments #1

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.

Response: Thank you for acknowledging the contribution and novelty of this paper. We appreciate the subsequent suggestions to sharpen the framing, recognize issues across all model configurations with respect to streamflow simulation, and look forward to developing 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.

Response: We agree that framing is important and that the phrasing “optimal configuration” needs to be avoided. The manuscript can be reframed to avoid implying WRF configuration optimization by making sure that the reader is aware that the space of WRF configurations is not exhaustively or even parsimoniously sampled, given the infeasibility of evaluating all the WRF configurations. Of the configurations we investigated, we found that the BSU_CFSR exhibits the most skill with respect to the observations and observational products we sampled – this wording was modified accordingly. Furthermore, per this reviewer's suggestion, we can present information about the streamflow simulation by the IPM earlier in this paper. We can also revise the paper to discuss the bias in streamflow simulation and report the quantitative measure of the biases in simulating streamflow.

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.

Response: We thank the reviewer for the insight regarding the opportunity to contribute to the debate highlighted in Lunquist et al, 2019, BAMS. We appreciate the suggestion and agree that ParFlow-CLM run forced by PRISM will enable us to make statements about whether, in this particular domain over this particular water year, using a state-of-the-science surface/subsurface hydrology model forced with observational-based products exhibits superior or inferior performance with respect to streamflow observations, as compared to that same hydrological model when it is forced with atmospheric process model simulations.

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.

Response: We agree with the reviewer and can report the quantitative measurements (e.g., RMSE against PRISM) in the revision. While uncertainty exists in the PRISM product, it has been evaluated against SNOTEL measurements and chosen as the benchmark for the assessments in this paper. Furthermore, there are a few SNOTEL sites that provide observations in the vicinity of the East River Watershed: the Butte station is within the Watershed, the Schofield Pass station is only a few km north of the East River Watershed northern boundary and captures some of the north-south gradient in snowfall while the Taylor Park Reservoir station is only a few km east of the East River Watershed and captures some of the east-west variability in snowfall.

As stated in the original version of the paper, the impacts of topographic shading had a minimal effect within the context of spatial-aggregated hydroclimate variables but had a significant effect for spatially resolved radiation fluxes which nonlinearly affect temperature and snowmelt. The topo_shading and slope_rad only redistribute the radiation flux on the topographic edges, but maintain the total surface energy budget at the watershed-scale. We can clarify and expand on these points in a revised manuscript.

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?

Response: We are referring to the variances of simulated hydrogemeterological outputs over multiple sub-grid physical schemes or meteorological forcings in the WRF model. We can modify this sentence in the revised manuscript by clearly explaining the hydrometeorological variables that are used to analyze the numerical experiments in this study.

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.

Response: We concur with the Reviewer that the language regarding IPM accuracy requires modification in the revised manuscript. The Reviewer's suggestion to reframe the concluding point to highlight the fact that these sensitivity analyses do help guide the scientific community as it develops observational constraints on process-understanding is very well-taken and we would like to include it in the revised manuscript. To that point, our finding that the atmospheric drivers of uncertainty in discharge, ET and other hydrologic variables are associated with processes occurring within the study domain and not external to it (i.e., uncertainty in physical processes in the Upper Colorado River at the watershed scale dominates surface hydrological variable uncertainty over the uncertainty associated with large-scale dynamics that set the initial and boundary conditions of these watersheds) provides support for future research directions.

We view our study as a first exploration of these topics and we look forward to working more in the coming years with hydroclimatic scientists interested in advancing the predictive understanding of Upper Colorado River water balance to improve regional, continental, or even Earth System scale dynamics modeling. The sensitivity analyses that we performed here do show the value of observational constraints on process-based understanding and where the scientific community should be focusing its efforts moving forward. We found definitively that the scientific community should be focusing on what is going on in the watersheds specifically, rather than focusing on improving large-scale meteorological prediction. It is a significant finding and can be emphasized in the revised manuscript in the context of discussing the development of observational constraints on process-models.

Future work (such as the use of more configurations, forcings, water years, and watershed locations) could help to support this hypothesis. Future works are going to be focusing on using a baseline configuration for future process-based research concurrent to SAIL, rather than a suite of configurations that we explored here. The reason for this is that we need to be sure to reference this manuscript for the WRF runs you are doing in support of SAIL.

Line 34 – remove “that”
Response: We agree with the suggestion and look forward to revising the manuscript per this suggestion.

Line 35 – “may have”? Could you express the reason this is stated with uncertainty?
Response: This is an estimation from multiple sources mentioned in (Milly and Dunne, 2020). We look forward to revising the manuscript to clarify this point.

Line 44 – Is “relevant” here meaning for larger-scales? Or respective relevant scales for each process?
Response: “relevant” means the observational datasets can be only used to improve the understanding of physical processes at their respective scale.

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.
Response: We agree with the reviewer and can add the review of climate model assessments against a few reanalysis datasets, specifically in the complex-topography Rocky Mountains/UCRB regions.

For examples, those references can be included in the revision
Alder, J. R., & Hostetler, S. W. (2019). The dependence of hydroclimate projections in snow‐dominated regions of the western United States on the choice of statistically downscaled climate data. Water Resources Research, 55(3), 2279-2300.
Buban, M. S., Lee, T. R., & Baker, C. B. (2020). A comparison of the US climate reference network precipitation data to the parameter-elevation regressions on independent slopes model (PRISM). Journal of Hydrometeorology, 21(10), 2391-2400.
Rahimi, S., Krantz, W., Lin, Y. H., Bass, B., Goldenson, N., Hall, A., ... & Norris, J. (2022). Evaluation of a Reanalysis‐Driven Configuration of WRF4 Over the Western United States From 1980 to 2020. Journal of Geophysical Research: Atmospheres, 127(4), e2021JD035699.

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

Response: First, the snow processes cross-scale interactions are hard to manage and often necessitate downscaling of WRF to force snow process models at the scale they need to be run. One reference for this is Winstral and Marks, 2014 (Winstral, A., and Marks, D. (2014). Long-term snow distribution observations in a mountain catchment: assessing variability, time stability, and the representativeness of an index site. Water Resour. Res. 50, 293–305. doi: 10.1002/2012WR013038). Also, Siirila-Woodburn et al. (2021) has provided detailed reviews of the challenges of managing the scales of subsurface process modeling with the scales of atmospheric process modeling.

Citations:
Winstral, A., & Marks, D. (2014). Long-term snow distribution observations in a mountain catchment: Assessing variability, time stability, and the representativeness of an index site. Water Resources Research, 50(1), 293-305.
Siirila-Woodburn, E. R., Rhoades, A. M., Hatchett, B. J., Huning, L. S., Szinai, J., Tague, C., ... & Kaatz, L. (2021). A low-to-no snow future and its impacts on water resources in the western United States. Nature Reviews Earth & Environment, 2(11), 800-819.

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

Response: We will reword "properly-configured" to "appropriately-configured".

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

Response: We will change "We established" to "We can establish".

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?

Response: We agree with the reviewer on this point. We look forward to revising the manuscript to state the hypothesis of this paper by stating "Our hypothesis is that synoptic-scale forcings produce a much larger spread in surface-through-subsurface hydrology fields than subgrid-scale physics scheme choice." We will then clarify the text to walk the reader through the implications if the hypothesis is confirmed or falsified by stating: "If our hypothesis is confirmed, then scientific efforts to advance the predictive hydrology, through modeling, of the UCRB should prioritize improving large-scale weather products and analyses. Conversely, if the hypothesis is falsified, model subgrid-scale physics scheme choice produces more variability in hydrologic response, so scientific efforts should prioritize development of smaller scale atmospheric and hydrological processes affected by surface heterogeneity in the ERW."

We also look forward to explicitly stating this hypothesis in the abstract of the revised manuscript.

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.

Response: We agree with the reviewer that the "observations" here can be misleading, and can reword the revised manuscript with the following language: "regridded reanalysis products and in-situ sensor measurements”. We also look forward to revising the manuscript to briefly summarize the objectives and outcomes of this paper at the end of the introduction section.

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

Response: We agree with the reviewer and look forward to revising the manuscript with the suggested language.

Line 141 – “near” should be “nearly”

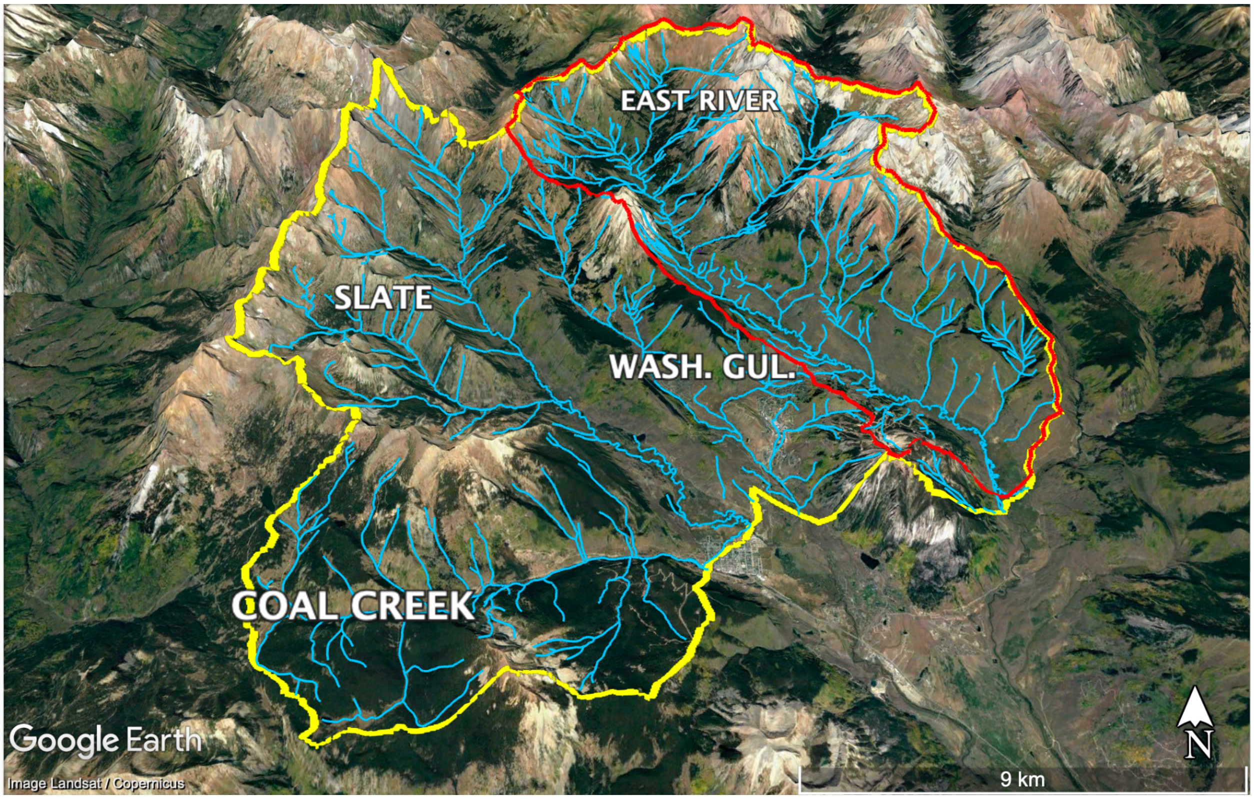
Response: We agree with the reviewer and look forward to revising the manuscript with the suggested language.

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.

Response: We agree with the reviewer and look forward to comparing our simulations with the precipitation, temperature and ET observations measured by the in-situ instrumentation in the East River watershed. We will add this tothe manuscript with the suggested language by noting the mismatch in scales directly measured by instrumentation.

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?

Response: We appreciate the sentiments of the reviewer and will now include a Google Earth overlay of watershed boundaries and streams in the revised manuscript as an additional Supplementary Material figure.



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."

Response: We included PRISM here as a reference dataset because it is one of the most widely-used gridded observationally-based datasets at sufficient resolution (800 meters) to evaluate the heteorgeneity of the UCRB. At the same time, we recognize the very issues that Lundquist et al. (2019) raised about this dataset, since those issues were strong factors in motivating the research described in this manuscript. We look forward to revising the manuscript to modify the language from "... was used as the reference dataset to assess model performance of precipitation and temperature in this study" to "... was used as a point of comparison in evaluating model uncertainty across sub-grid physical schemes and meteorological forcing datasets for precipitation and temperature in this study."

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.

Response: The raw ASO product has 50 meters spatial resolution. The ASO product is regridded to the same grid resolution as WRF outputs (500 meters) for comparison purposes. We acknowledge that the underestimation by ASO could be due to the point-to-grid errors and look forward to addressing this issue in the revision. We recognize the research on gridding SWE data (e.g., Fassnacht et al, 2003, doi: https://doi.org/10.1029/2002WR001512 and Dozier, 2011, doi:https://doi.org/10.1029/2011EO430001) and have followed the approach of the linear interpolation of the ASO data as documented in Oaida et al, 2019, doi:https://doi.org/10.1175/JHM-D-18-0009.1 and look forward to including this detail in the revised manuscript.

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

Response: We can add subtitles as “4.1. Sub-grid physical schemes vs meteorological forcings”, and “4.2. 3D topographic radiation effects”.

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?

Response: We agree with the reviewer and can add labels in the captions, and also to describe the statistical methods in the methods section.

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.

Response: We agree with the reviewer that the spatial average visualization does not show the importance of locally specific spatial differences and, therefore, is unable to explain the physical reasoning of earlier snowmelt. We can create another supplement figure of the locally specific spatial differences in the UCD configuration simulation, and add a brief discussion of the physical reasons that may have given rise to an earlier snowmelt.

Line 319 – run-on sentence.

Response: We can modify this sentence for clarity.

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?

Response: Thank you for the kudos about this figure! We look forward to revising the manuscript accordingly. The x-extents of a-d are identical; we can add the additional tick-marks to avoid confusion.

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

Response: We look forward to adding to the revised manuscript the quantitative statistic of RMSE (Root Mean Squared Error), as suggested earlier by the reviewer, for precipitation, temperature and SWE across all experiments to quantify how the BSU-CFSR2 configuration compared to the other configurations in terms of its agreement 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?

Response: We can adjust the extent of the plotting area according to this comment.

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

Response: This paragraph describes the variance of simulated streamflow across experiments, and the next paragraph introduces the comparison against in-situ streamflow observations. We look forward to revising the manuscript to add a better transition sentence to aid the reader in separating these paragraphs.

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.

Response: We agree with the reviewer and can revise the abstract to not explicitly state that the objective of this paper is to identify the optimal model configurations but rather an exercise of sensitivity analysis where one configuration will perform the best.

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?

Response: The reviewer is correct that the differences are minimal. Since this is a snow-dominated watershed and streamflow is predominantly controlled by snowmelt, the seasonal variance is not comparable with the among-model variances. We can revise the manuscript to describe the intra-model configuration variance in different seasons.

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.

Response: We agree with the reviewer and look forward to removing the word "slightly" in the revised manuscript to avoid any misunderstanding.

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.

Response: We can re-word this sentence and perhaps use the wording “less noisy” opposed to “more muted” to avoid any confusion. What’s meant here is that streamflow signals are very reactive, noisy, and change quickly, whereas groundwater signals are the product of slower processes via infiltration and vadose zone dynamics, often at longer timescales, which result in very different temporal signals as compared to streamflow.

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.

Response: We can replace with the perceptually uniform color bar for the plots, based on the suggestions in Cramer et al. (2020).

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).

Response: Here we mean that by evaluating the model uncertainties for simulating precipitation, temperature, and streamflow, we are able to identify the which process within the model has the most important uncertainties. We can revise this sentence accordingly.

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?

Response: Here we mean the large-scale forcing dataset used as the initial and boundary conditions in the WRF model. We can revise this sentence and explicitly mention that in the revised manuscript.

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?

Response: We agree with the reviewer and look forward to adding the quantitative statistics RMSE for the experiments against the PRISM observationally-based dataset.

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).

Response: We agree with the reviewer and can revise the color scale.

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.

Response: We agree with the reviewer and can revise this sentence as “Latent heat buffers differences in the shortwave radiation contribution to the radiation budget.”

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?

Response: This paragraph discusses the systemic cold bias in our current IPM configuration, and the limitations of one-way vs. two-way coupling between WRF and ParFlow-CLM. We will revise this paragraph to be more clear and to the point.

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