Referee Comments #3

This work presents a model study in a headwaters catchment in the Upper Colorado Watershed. In general the work is interesting and well-written but the presentation is somewhat confusing. There are some major points I think the authors need to address before the work's suitability for publication can be assessed. They are detailed below.

We thank the reviewer for his or her comments on this work being interesting and well-written. We have addressed all points the reviewer requests clarification on below.

General Comments:
-The terminology of so-called IMP's used is confusing along with the reference to these as coupled models, which I would argue they are not. It's very confusing to discuss this work in the framework of different models as opposed to just forcing used to drive the hydrologic model. The language around the different products used is very confusing and makes much of the discussion hard to follow. Some of the meterological forcing datasets appear to be used to drive the models direclty, but in the introduction it appears that only WRF simulations are used to drive models. A completely re-write of this entire section is needed to make this clear. What did the authors do with the hydrologic outputs from the WRF simulations? Why are the Noah and Noah-MP models used interchangeably but the results are not compared to ParFlow-CLM?

Response: We thank the reviewer for her/his comments noting the sophisticated tools used in this study, as well as the importance and high quality of the manuscript writing. The concern about the use of the phrase “integrated process modeling” is potentially one of semantics in this case, although we’re happy to revise to avoid any potential confusion for the reviewer and/or other readers. There is a small body of literature that does use “integrated process model” terminology (e.g., Zhang et al, 2016 doi:https://doi.org/10.5194/hess-20-529-2016; Davison et al, 2017 doi:https://doi.org/10.1002/2017MS001052) that demonstrates the utility of coupling process models built to explore discipline-specific processes as a mechanism to advance interdisciplinary research. There is also a literature comparing and contrasting one-way coupling vs two-way coupling for mountainous hydrology: e.g., Camera et al, 2020 doi:https://doi.org/10.5194/nhess-20-2791-2020 and Rudisill et al, 2022 doi:https://doi.org/10.1002/hyp.14578 where the latter paper finds that in snow-dominated watersheds such as the ERW, which found that the representation of uncertainties in the representation orographic precipitation is the single largest driver of hydrological uncertainty while the inclusion or exclusion of two-way coupling has little effect on atmosphere-through-bedrock state evolution.

We agree that the use of the word “coupled” should be used with caution, as some of these models are formally coupled within their source-code infrastructure, and others are “step-wise coupled”. To avoid any potential confusion, we have modified the manuscript to only use the word “coupled” when there is a formal two-way and self-consistent pairing of codes (for

example, ParFlow and CLM). In contrast, the associated meteorological variables used to drive ParFlow-CLM are now referred to as “forcing” only, as this is a one-way interaction between the codes/outputs. Nonetheless, to avoid confusion we would like to clarify that we are referring to a one-way coupled IPM (or WRF-Parflow-CLM) in the revised manuscript.

The name of the meteorological products are the forcing datasets for the WRF runs, not the forcing data for ParFlow-CLM. Precipitation, temperature and other hydroclimate simulations from WRF are used as the atmospheric drivers for the ParFlow-CLM simulations. We will revise the manuscript to explicitly clarify that the meteorological forcing datasets are used as the initial and boundary conditions for the WRF model, and the WRF model outputs are then used to force ParFlow-CLM.

The Noah and Noah-MP hydrologic output from the WRF simulations included ET, surface and subsurface runoff. We agree with the referee (and referee #2 who identified this too) and will compare the Noah and Noah-MP simulated ET against ParFlow-CLM. The primary reason for using ParFlow-CLM is to allow for the quantification of streamflow and groundwater storage. As the reviewer is probably aware, standalone WRF does not simulate these processes, although branches of the code (WRF-Hydro) do provide some insight into at least streamflow, although with a simplified and prescribed stream network. Groundwater in WRF-Hydro is highly simplified (shallow soil layers and a bucket model) opposed to ParFlow, which simulates the full continuum of variably saturated flow in three dimensions. Thus, while we indeed have generated multiple datasets with the different LSMs, this is not our primary objective for intercoparison and we believe showing more of these results would be distracting. We focus on the set of LSM outputs from WRF for simplicity (and to avoid confusion) and show the groundwater and streamflow outputs from ParFlow-CLM.

Response after revision: A few IPM references have now been added to the Introduction. In addition, we now explicitly define the meaning of IPM as the use of WRF via a one-way feedback to ParFlow-CLM in the Introduction (to acknowledge the reviewers point about the extensive literature using two-way feedbacks). The section in the Introduction now reads as follows,“
While discipline-specific process models, such as those used to explore and predict atmospheric or subsurface processes have advanced scientific understanding in a myriad of ways through sustained engagement with extensive user communities (Gutowski et al., 2020), Integrated Process Models (IPMs), in which these discipline-specific process models are coupled, are relatively novel and are still being vetted for various scientific applications in complex terrain. Zhang et al. (2016) and Davison et al. (2017) demonstrated the utility of coupling process models built to explore discipline-specific processes as a mechanism to advance interdisciplinary research. Furthermore, Camera et al (2020) discussed the one-way vs. two-way coupling of IPMs to understand process interactions in the mountainous hydrologic cycle. The capabilities and details of the IPM has been discussed in a series of findings….”

Except for groundwater which isn't discussed very much in the manuscript, all the same results should be in the WRF simulations. Why didn't the authors just run the WRF-ParFlow model or

even mention it's existence? They talk about everything in a coupled sense but the models are in no way formally coupled (unless something is happening that is not discussed in the manuscript); the output files from WRF simulations are saved and somehow reformatted (this is not clear) and used to drive ParFlow-CLM. They could drive any hydrologic model and it wouldn't be considered a coupled platform, likewise the standard forcing products the authors might choose to drive the simulations off the shelf are also generated with atmospheric models, yet I would never think of this as an IMP. I suggest the authours are much more transparent about this aspect and remove the terminology from a revision. They should also provide some clear language about what is actual being done here, is this a comparison between forcing generated with WRF v other approaches? Why didn't the authors just run forced by PRISM?

Response: As is hopefully now more clear based on the last answer, we use WRF outputs to force ParFlow-CLM as a mechanism to simulate the integrated hydrologic response of the watershed. We do this because WRF is not capable of simulating two key processes of interest: not just groundwater storage changes but also streamflow. We did not run the fully coupled WRF-ParFlow model platform because these simulations are extremely computationally expensive. We agree that it would be interesting to perform these 2-way coupled simulations in the future, but believe that an important first step is to determine what the hydrologic response is without the two-way interaction (i.e. WRF-forced ParFlow-CLM as we've done here). Fully coupled 2-way simulations are by no way the "standard" in hydrologic modeling, and the amount of fidelity we've included in this study stands to dwarf that of other approaches. That being said, we acknowledge that the existence of the code could be pointed to as an interesting next step. Two-way integrated meteorological and surface/subsurface hydrology simulations in complex terrain are especially limited, and a new frontier in the field.

Furthermore, we do appreciate the sentiments of the reviewer and will ensure that we will more clearly outline various methodological approaches taken in our model simulations and analyses in our revision. For example, the WRF simulations were regridded using bilinear interpolation and used as the forcing dataset to run the ParFlow-CLM model. Therefore, the simulations are, as the reviewer points out, a one-way coupled IPM. We will ensure that we mention that WRF-Parflow is also available as a two-way coupled IPM too (and cite the relevant literature).

Finally, we agree with the reviewer and can run the ParFlow-CLM simulation forced by PRISM. We plan to regrid the 800-meter resolution PRISM daily precipitation and temperature variables and temporally interpolate them using an adjustment to the hourly distributions provided by the WRF hourly output. Other than precipitation and temperature (only variables made available from PRISM), we will use the WRF hourly output to force the ParFlow-CLM simulation.

Response after revision: In the Discussion section, we have tried to be more transparent about what is meant when we use the terms ParFlow-CLM, WRF-hydro and WRF-PF: "On the other hand, ParFlow-CLM is essential in our experiment for quantifying hydrological responses, including streamflow and groundwater storage. Although WRF-Hydro provides some insights into streamflow, it still uses a simplified and prescribed stream network. Groundwater storage in WRF-Hydro is also highly simplified through its use of a bucket model, while ParFlow-CLM

simulates the full 3D continuum of variably saturated soils. We also recognize that the WRF-ParFlow model can simulate two-way coupling between the atmospheric and hydrological processes in the surface and subsurface domain, though it is computationally expensive and requires signficant efforts to set up. Importantly, in a similar fashion as the heirarchy of climate models approach oft used in the climate community (Jeevanjee et al., 2017), we would also like to assess one-way coupling performance of our IPM prior to assessing two-way coupling IPM performance."

We also performed new ParFlow-CLM experiments forced by PRISM and included these results in Figure 7, and the text has been revised as follows, "We have evaluated the aforementioned WRF subgrid-scale physical schemes and large-scale meteorological forcings in representing precipitation, temperature, snowpack and radiation fluxes, and their impacts on the integrated water budget within ParFlow-CLM. We also evaluated the simulated discharge from ParFlow-CLM forced by PRISM as a comparison with WRF forcings."

-Coupled v uncoupled processes and feedbacks. There has been a lot of work to understand the role of feedbacks between two-way coupled hydrologic models and atmospheric models. Examples include WRF-Hydro-WRF (e.g. Arnault 2016), COSMO-CLM-Parflow (e.g. Keune 2016, 2019), WRF-ParFlow (e.g. Maxwell 2012, Forrester 2020), feedbacks over complex terrain (Ban 2014), and other more conceptual approaches (e.g. Miguez-Macho 2007). This is not an exhustive list, but demonstrates that much work has been done to study these feedbacks, Some of these studies are in complex terrain and even suggest that the approach used by the authors may not be valid at high resolution without lateral flow. These studies all systematically compare different types of model physics (e.g. free drainage, the standalone atmospheric model, fully coupled system) and use varying metrics to diagnose coupling strength and changes in the atmosphere. I suggest the authors read these prior studies carefully and develop a new section that summarizes (rather than ignores the existence of) this body of work and uses this to put the current study in context. This will help frame the current work and help it look much less like a patchwork of runs that are loosely tied together. This will also help clarify my point above, to help the reader follow what is being done and what runs are conducted in the current work.

Response: We agree with reviewer that there are multiple two-way coupled IPM configurations and will ensure that our revision better contextualizes our one-way coupled IPM configuration within the broader two-way coupled literature. We can develop a new paragraph in the introduction to demonstrate the details of those models in the paper.

Response after revision: A new paragraph has been developed in the Discussion section to better contextualize one-way and two-way feedback mechanisms in the IPM community as follows, "Another methodological constraint is that our WRF and Parflow-CLM experiments were only one-way instead of two-way feedbacks, which ignores potentially important feedbacks from the subsurface hydrology to the atmosphere via ET and the radiation budget. For example, Givati et al. (2016) reported that simulated precipitation was improved with two-way coupling in WRF-Hydro compared to WRF-only and Forrester et al. (2018) showed that boundary layer

dynamics were impacted in IPM simulations in regions where shallow water tables exist. On the other hand, ParFlow-CLM is essential in our experiment for quantifying hydrological responses, including streamflow and groundwater storage. Although other fully-coupled integrated hydrology model (e.g., WRF-Hydro) provides some insights into streamflow, it still uses a simplified and prescribed stream network. Groundwater storage in WRF-Hydro is also highly simplified by using a bucket model while ParFlow-CLM simulates the full continuum of variable saturation in three dimensions. ”

-Variability in point processes compared to integrated or averaged measures. I mention this as a specific instance below, but it is also a general point, there are instances where the authors present differences locally (at a point) that do not persist synoptically. Do the different forcing products or microphysics (I think this is the point the authors make) make some difference locally for e.g. precip, radiation, but does some averaged quantity remain unaffected. It appears this is the case for much of the analysis. That is, topographic shading makes a difference locally in LH flux but the domain averaged LH flux remains unchanged between cases. The authors draw one conclusion (local differences) without acknowledging the other (same net energy flux over the domain).

Response: We agree with the reviewer and look forward to adding more discussion of locally specific versus watershed average differences in our IPM results. For example, the importance of topographic shading on the spatial distribution of radiation flux. The variances across different forcing products and subgrid-scale physical schemes are shown in our paper through both spatial distributions and average quantities. However, the 3D topographic radiation schemes only redistribute the energy flux thus affect LH flux spatial distribution, but do not show significant difference in the watershed average quantities. In the revision, we will more clearly highlight these points.

Response after revision: The text has now been revised to, “Topographic shading makes a difference locally in LH flux, by redistributing the energy flux and thus affecting LH flux spatial distribution. Nevertheless, the domain averaged LH flux remains unchanged between cases.”

-Atmospheric uncertainty. There has been much work on differences in model physics in a model such as WRF that allows different physical parameterizations to be "swapped out" easily in simulations by changing the namelist. This is an important aspect of uncertainty, but it is almost always put in the context of one of the major forms of uncertainty in the atmosphere, propogation of intial conditions. One should always determine that such a physics change is robust using (e.g.) time-shifted uncertainty in an ensemble type approach (e.g. Walser 2004). Often upon inclusion of uncertainty in the intial model state (in the atmosphere) the differences in physical paramterization no longer dominate.

Response: We acknowledge the reviewer’s well-taken point about the value of using a traditional experimental design to isolate and compare model structural uncertainty to model internal variability uncertainty to improve modeling skill and predictive power. We concur that the specific use of initial condition perturbation ensemble members enable the quantification of

model internal variability while varying subgrid-scale physics parameterizations tests model structural uncertainty. The use of different meteorological forcings do comprise different initial conditions, though this does differ from explicit initial condition perturbation analysis. We will add a few sentences in the Methods section pointing to the reviewer's important points, explicitly cite Walser and Schaer (2004) and be more explicit about how our experimental design differed from the more traditional approach. As part of what we will add to the Methods section we would like to acknowledge that although our experimental design didn't assess internal variability through the more traditionally employed initial condition perturbation approach (i.e., alter the initial conditions provided to the atmospheric model by a rounding error) we did test the WRF-ParFlowCLM configuration with a range of realistic large-scale initial boundary conditions derived from the use of several atmospheric reanalyses. This approach, while coarse, does separately assess the influence of initial and boundary conditions (e.g., integrated vapor transport) ranges from the role of atmospheric physics representations in WRF-ParFlowCLM to isolate slight differences in large-scale boundary conditions.

We believe that a major finding of our work that hasn't been explored in the literature extensively over the East River watershed is that these slight differences in large-scale boundary conditions (analogous to perturbing initial conditions by rounding errors) do not markedly change water year precipitation totals unlike changes in subgrid-scale physical parameterization choice.

Response after revision: The discussion of atmospheric uncertainty and the initial condition has now been added to, "We recognize that the output from WRF simulations may be dependent on initial conditions, which are inherently difficult to constrain (e.g., Walser and Schär, 2004), but the experimental configuration described here seeks to be insulated from that dependency by running WRF simulations with initial conditions derived from different meteorological forcings."

-Can the authors compare meterological forcings at the site? A heavily-instrumented catchment (abstract line 16-) should have observations of meterological variables and snow outside of the SNOTEL (which I don't think are used for comparison and should contain precipitation and temperature), even preciptation and temperature at gage locations would be very instrumental. It appears that the authors treat the PRISM product like observationsm, which is an unfortunate and hopefully accidental. The PRISM product is a model, even if statistical, that takes into account observations in a region. One would assume that then PRISM is ingesting precip from the SNOTEL sites in the domain but this isn't stated (are there even any obsevations that PRISM is using and is it thus totally unconstrained?).

Response: The large-scale meteorological forcings used in the WRF simulations have coarser grid resolutions than the WRF inner domain, so comparing IPM simulation results with meteorological forcing at the sites do not add scientific significance in model evaluation.

The reviewer is right about the PRISM product: it uses SNOTEL datasets and was generated using statistical methods. PRISM dataset has been widely used to evaluate climate models in the complex-terrain region, with many thoughtful uncertainty analysis (e.g., Lunquist et al., 2020). We used interpolated precipitation and temperature product PRISM to evaluate the

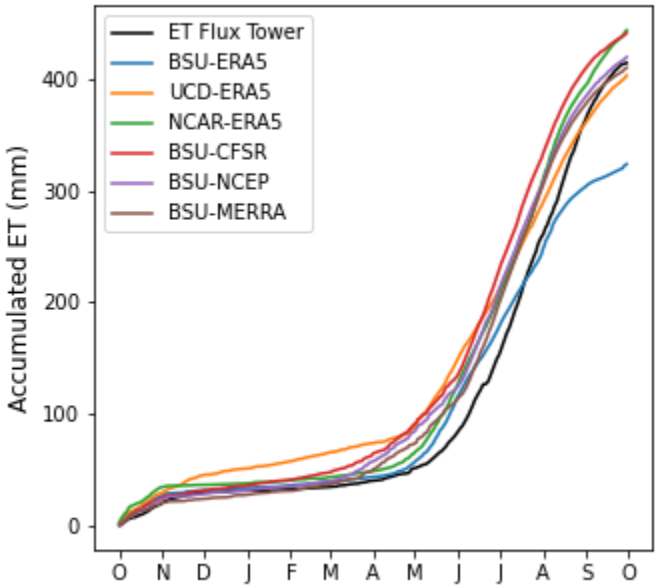
model performance over the whole domain. In the supplementary material, we presented the comparison of precipitation and temperature against snotel stations. In the paper revision, we can compare with in-situ observation data, including precipitation and temperature measurements taken at sensors installed in the East River watershed. We did not provide these in the initial manuscript to avoid distracting from the main message of the paper, but see the relevance and could add these to the supplemental in the revision.

Response after revision: The precipitation, temperature and snow water equivalent comparison among WRF simulations vs. measurements at two SNOTEL stations are now presented in Figure S3.

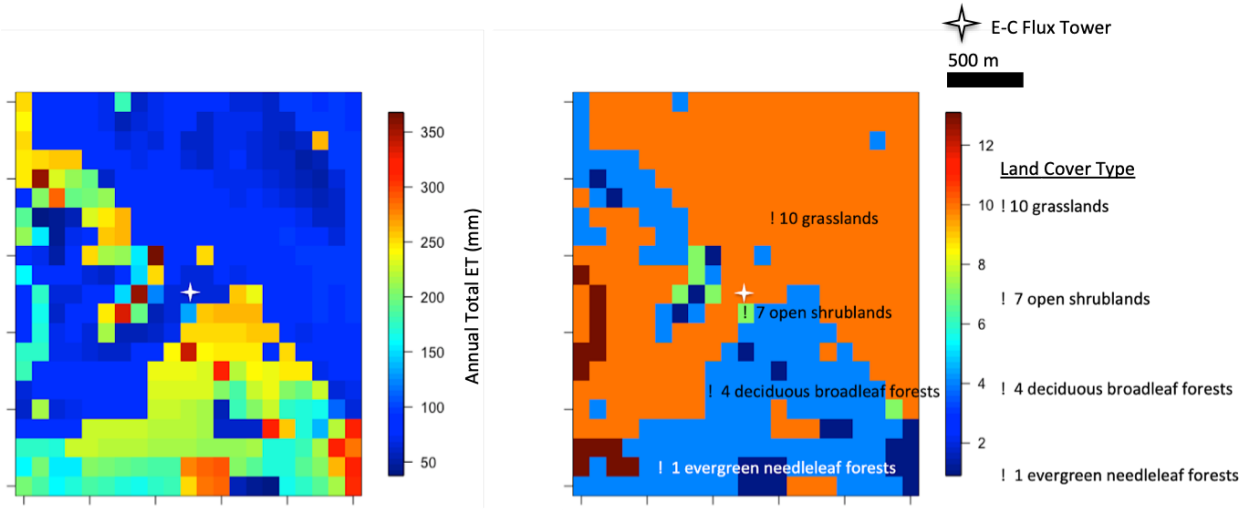
-The authors should compare to ET observations in Ryken et al 2022 to results of current work (both WRF and ParFlow-CLM). Additionally, it apears that the Ryken et al 2022 paper has meterological observations of precip, temperature and radation that might be useful to partly address my comment above.

Response: We agree with the reviewer and can add the comparison against the ET measurement at the eddy-covariance tower at the East River, but again do not want to distract from the main purpose of this study by making this an ET intercomparison so this can potentially be added to the appendix.

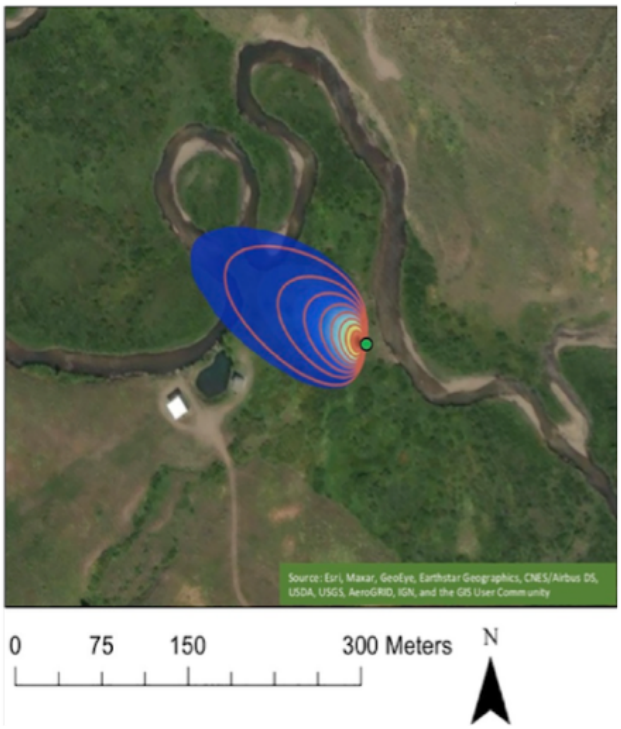
Response after revision: Per the reviewer's comment, we performed a comparison of WRF and ParFlow-CLM against the eddy-covariance tower in the East River. As shown below, the WRF simulations generally agree with the observation at the tower (black line), with the exception of the BSU-ERA5 simulation:



However, a comparison of the ParFlow-CLM ET simulations revealed lower annual ET fluxes, with annual ET on the order of 100-200 mm. This difference is largely due to a number of factors in both the ParFlow-CLM and WRF simulations. First, land cover type: WRF parameterization of land cover in this region is coarser than that of ParFlow, and represents a coniferous forest over this region, whereas ParFlow is based on NLCD land cover resolved to 30 m, which was then upscaled to 100 (m) to match the resolution of the ParFlow-CLM model. The land cover type of the cell containing the flux tower in ParFlow-CLM is show centered in the figure below with a white star, and is represented in the model as “open shrubland.” Neighboring needleleaf forests cells in ParFlow-CLM yield similar annual ET totals to that of WRF and the flux tower, which shows consistency in the predictability of the WRF simulations.



A second consideration is the flux tower fetch (see Figure 5 of Ryken et al., 2022, pasted below), which shows segments of the tower footprint over the meandering reaches of the East River that will bias the observation of ET at the tower high. Personal communication with David Gochis (who installed the tower with Ryken and Maxwell) confirmed this presumption, stating that if a land model parameterized the land cover similar to that of the surrounding land cover type (grasses and shrubs), it would likely calculate a lower flux compared to the observed ET measurement of the tower. Thus, it’s sensical that ParFlow-CLM estimates are low when only assuming the cell-based comparison of ET.



Thus, the use of the tower to benchmark the model performance at this scale is not appropriate, and we have chosen not to include the discussion in the paper.

Specific comments
line 65: is PF-CLM being cited using Maxwell et al 2015 (cited on line 617)? That paper references a simulation over large scale that as I read it is forced externally and does not use or describe the CLM model.

Response: Yes, while this is the same ParFlow, we agree a better citation can be used, the standard are: (Ashby and Falgout, 1996; Jones and Woodward, 2001; Maxwell, 2013). We will cite more appropriately in the revised version, including explicit mention of previous version of the model in the East River (Maina et al., 2022; Foster and Maxwell, 2018; Pribulick et al., 2016).

Citations:

Ashby, S.F., Falgout, R.D., 1996. A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations. Nucl. Sci. Eng. 124 (1), 145–159.

Kuffour, B. N., Engdahl, N. B., Woodward, C. S., Condon, L. E., Kollet, S., & Maxwell, R. M. (2020). Simulating coupled surface–subsurface flows with ParFlow v3. 5.0: capabilities, applications, and ongoing development of an open-source, massively parallel, integrated hydrologic model. Geoscientific Model Development, 13(3), 1373-1397.

M. Foster, L., & M. Maxwell, R. (2019). Sensitivity analysis of hydraulic conductivity and Manning's n parameters lead to new method to scale effective hydraulic conductivity across model resolutions. Hydrological Processes, 33(3), 332-349.

Jones, J.E., Woodward, C.S., 2001. Newton–Krylov-multigrid solvers for large-scale, highly heterogeneous, variably saturated flow problems. Adv. Water Resour. 24 (7), 763–774.
Maina, F. Z., Wainwright, H. M., Dennedy-Frank, P. J., and Siirila-Woodburn, E. R.: On the similarity of hillslope hydrologic function: a clustering approach based on groundwater changes, Hydrol. Earth Syst. Sci., 26, 3805–3823, https://doi.org/10.5194/hess-26-3805-2022, 2022.
Maxwell, R.M., 2013. A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling. Adv. Water Resour. 53, 109–117.
Pribulick, C. E., Foster, L. M., Bearup, L. A., Navarre-Sitchler, A. K., Williams, K. H., Carroll, R. W., & Maxwell, R. M. (2016). Contrasting the hydrologic response due to land cover and climate change in a mountain headwaters system. Ecohydrology, 9(8), 1431-1438.

Response after revision: The introduction of ParFlow-CLM is now done in the following way, “result in differences in surface and subsurface hydrologic metrics when used to force the integrated hydrologic model (ParFlow-CLM; Ashby and Falgout, 1996; Jones and Woodward, 2001; Maxwell, 2013, Maxwell et al., 2015), which has been widely applied in the UCRB (Maina et al., 2022; Foster and Maxwell, 2018; Pribulick et al., 2016). We expand upon those various sensitivity analyses in this study, including the influences of large-scale meteorological forcing and subgrid-scale physics scheme choice on the surface-through-subsurface response of the integrated hydrologic model.”

Line 240+ This section describes the PF-CLM model in general but I could not find specifics for the model domain used in this study? What is the resolution or model configuration for the PF-CLM domain? How deep is the subsurface? What is the lateral resolution? How was this matched to the forcing datasets or the WRF outputs? Was there a balance of water and fluxes between the grids? How were model parameters determined? Are there references to prior work on this model? Calibration? If not the authors might include a description of these aspects in the current manuscript and as supplemental material.

Response: The ParFlow-CLM subsurface domain is 30-meter deep, 100-meter horizontal resolution. The WRF output are re-grided using bilinear interpolation to match the ParFlow-CLM grid cells. The model parameters are based on a variety of geological and soil parameters, and calibarated using streamflow measurements. More details can be found in some previous papers (e.g., Foster et al., 2019, Pribulick et al., 2016, see the answers to the previous question). We will add more detailed descriptions of the ParFlow-CLM model in the manuscript

Response after revision: The description of ParFlow-CLM has now been revised as follows: “The ParFlow-CLM subsurface domain is 30-meter deep at 100-meter horizontal resolution. The WRF outputs are re-gridded using bilinear interpolation to match the ParFlow-CLM grid cells. The model parameters are based on a variety of geological and soil parameters, and calibrated using streamflow measurements. More details can be found in Foster et al., (2019) and Pribulick et al., (2016).”

Lines 275-281. The UCD datasets appear to have the most precip but the ear

Figure 3 caption (~line 305): a, b, c are used to identify plots in the figure but are not used in the caption. Also, it does not appear that 3c is described in the caption.

Response: We agree with the reviewer and can revise the figures accordingly.

Response after revision: Revised as suggested.

Figures 5, 6 and associated discussion. An interesting point that might be made here is that while local spatial differences are apparent in Figure 6, the domain averages (even for SWE) are the same between shaded and non-shaded formulations. This suggests that while it may be striking visually to include shading, the upscaled water balance for the catchment isn't sensitive.

Response: We agree with the reviewer that turning on and off the 3D shaded radiation scheme does not significantly affect the domain average. The reviewer is correct that the 3D shading radiation scheme does not affect the upscaling water balance, but rather redistributes the spatial distribution of radiation fluxes thus SWE and energy spatial pattern. The east side of the valley gets more radiation because it's facing west, and the western side of the valley gets less radiation as it gets less radiation in the early morning. Similar conclusions have been observed in Arthur et al. 2018.

Response after revision: The text has now been revised as follows, "The 3D radiation shading scheme does not significantly affect the total water balance, but rather the spatial distribution of radiation fluxes. Thus, despite having minimal impacts on water impacting on the water balance, the scheme does have important localized impacts on SWE and surface energy budget spatial patterns" and "In summary, the simulations show that, while local spatial differences in surface radiation with and without realistic topography are apparent in Figure 6 , the domain averages (even for SWE) are the same between shaded and non-shaded formulations. This suggests that while it may be striking visually to include shading, the impact of topographic shading on upscaled water balance for a domain like the ERW is negligible."

lines 383- I'm not sure I agree with these conclusions. While the cumulative variability in outflow resulting from the different forcing products creates different cumulative outflows, Figure 7a indicates that there is no difference in timing across all the forcing datasets. My suspicion is that the differences in outflow are due to total water quantity (Figure 5a suggests this as well) and are simply a precip bias artifact in the different WRF runs.

Response: We agree with the reviewer that Figure 7a indicates the difference in timing in all the IPM across all forcing datasets. The difference of precipitation in WRF simulations leads to the differences in streamflow simulation in ParFlow-CLM. We can revise this sentence to clarify and avoid misunderstanding.

Response after revision: We revised the sentence as follows, "Discharge at the watershed outlet (see exact location on Figure 1) shows a different timing across the various WRF subgrid-scale physics scheme configurations and large-scale meteorological forcings that leads to a temporal

shift in simulated streamflow, where the daily averaged time series (left) shows only minor differences through time."

line 443: "Here, we ... coupling WRF and ParFlow" rephrase, this sentence isn't correct, the models were not coupled

Response: We can reword this as 'one-way coupling IPM' or "WRF-ParFlow-CLM" to avoid misunderstanding.

Response after revision: We have revised this sentence in the text as follows, "here, we used an IPM with one-way feedbacks from WRF to ParFlow-CLM".

line 476+: This text appears to acknowledge the lack of coupling in the current work (as an aside, what is "one-way coupled" this is not actually coupled at all, as it appears the results from WRF were simply used as forcing for the PF-CLM model. This isn't bad, but as mentioned above should be discussed up front. The arguments here regarding computational expense as an excuse for not running coupled simualtions are incorrect, prior studies have shown with the e.g. WRF-PF model that ParFlow is approximately 1% of the total computational time compared to WRF which is 99% of the computational time. Thus if the authors ran WRF for this domain, the additional expense to run with WRF-PF is a negligible increase in cost. Also the authors might want to correclty identify that the Forrester et al study (line 483) was run with WRF-PF and the authors might want to read and cite Forrester 2020 which discusses limitations of running high resolution, uncoupled WRF simulations in mountain terriain (the CO headwaters was studied) where the lack of lateral flow caused changes in the surface energy budget and hight of the boundary layer.

Responses: The reviewer is correct that the WRF simulation used most of the computational and throughput time. However, we are unaware of any study which shows only a 1% CPU demand for ParFlow-WRF opposed to standalone WRF. Personal communication with the primary ParFlow developer confirmed this is not a strict rule of thumb, and that there are various factors which determine the additional CPU time to include ParFlow in a WRF run. This is especially true in complex terrain where there are sharp wetting fronts and higher demands on the Richards' equation solver. Thus, it's difficult to know the exact additional demands of the fully coupled code without performing the simulations.

We agree to update the statement as "one-way coupling IPM" in the revision. We also concur with the reviewer that computational expense is not the major excuse for not running coupled simulations, moreso that we wished to establish a baseline set of simulations without fully coupled feedbacks before considering the 2-way interacting model.

We agree with the reviewer to cite Forrester 2020 paper, which discusses how the subsurface groundwater flow in ParFlow-CLM (lateral groundwater flow and subsurface lower boundary conditions) affects the atmospheric model. These impacts were particularly enhanced during summertime, while we run the simulation for a whole water year. We agree to highlight the

findings from previous fully-coupled WRF-PF model, and extend the discussion of one-way and two-way coupling in the revision.

Response after revision: The text has now been revised as follows, “A limitation of our study, given the computational constraints of running IPMs, is that it was infeasible to explore the full parameter spaces of WRF and ParFlow-CLM exhaustively; thus, our conclusions are limited to the selected subgrid-scale physics schemes and meteorological forcing datasets analyzed. Additional work is needed to improve the systemic cold bias in two-meter surface air temperature throughout all experiments as this may have been the major driver in the delayed snowmelt and peak discharge simulated by the IPM. Another methodological constraint is that our WRF and Parflow-CLM experiments were only one-way instead of two-way feedbacks, which ignores potentially important feedbacks from the subsurface hydrology to the atmosphere via ET and the radiation budget. For example, Givati et al. (2016) reported that simulated precipitation was improved with two-way coupling in WRF-Hydro compared to WRF-only and Forrester et al. (2018) showed that boundary layer dynamics were impacted in IPM simulations in regions where shallow water tables exist. On the other hand, ParFlow-CLM is essential in our experiment for quantifying hydrological responses, including streamflow and groundwater storage. Although WRF-Hydro provides some insights into streamflow, it still uses a simplified and prescribed stream network. Groundwater storage in WRF-Hydro is also highly simplified by using a bucket model while ParFlow-CLM simulates the full continuum of variable saturation in three dimensions. We also recognize that a coupled version of WRF and ParFlow exists, with the capability of simulating two-way coupling between the atmospheric and hydrological processes in the surface and subsurface domain, though it was not used in this study and could be explored in future efforts. ”

line 498: Is the watershed highly instrumented now or will it be? This seems at odds with statement in the abstract (line 16)?

Response: To clarify, the East River watershed is already highly instrumented due to the presence of the long-standing Rocky Mountain Biological Laboratory (RMBL), the SNOTEL network, and DOE Watershed Science Focus Area project which has been adding instrumentation to the watershed over the last ~7 years. While these observations focus primarily on surface and subsurface processes, the East River watershed has become even more instrumented in recent years (2021-2023) through the support of the U.S. DOE (SAIL campaign) and U.S. NOAA (SPLASH campaign) deployments of a comprehensive set of atmospheric instrumentations (e.g., radar and radiation measurements).

However, these SAIL and SPLASH campaign measurements were provided after our simulations were conducted and we began to prepare this manuscript. In future work, we plan to build upon the knowledge learned from this manuscript to compare the most optimally configured IPM to SAIL and SPLASH campaign observations. We can revise this sentence to best reflect the added significance of SAIL and SPLASH campaigns in future IPM studies in the UCRB region, and look forward to using those datasets to improve mountainous hydrologic cycle process understanding and model development.

Response after revision: The last paragraph of the conclusion section has been revised to “The East River watershed is already highly instrumented due to the presence of the long-standing Rocky Mountain Biological Laboratory (RMBL), the SNOTEL network, the United States Geological Survey’s Next Generation Water Observing System (NGWOS), the National Science Foundation’s Sublimation of Snow (SOS) project, and DOE Watershed Science Focus Area project which has been adding instrumentation to the watershed over the last ~7 years. While these observations focus primarily on surface and subsurface processes, the East River watershed has become even more instrumented in recent years (2021-2023) through the support of the U.S. DOE (SAIL campaign) and U.S. NOAA (SPLASH campaign) deployments of a comprehensive set of atmospheric instrumentations (e.g., radar and radiation measurements). Future work will include integration of data, either indirectly through IPM benchmarking or directly through data assimilation into the IPM, from the SAIL campaign. . SAIL is collecting a wide-array of observations with the intent to advance understanding of precipitation, snow, aerosol, aerosol-cloud interaction, and radiation processes in complex terrain and establish the minimum-but-sufficient level of process understanding to develop a robust predictive understanding of seasonal surface water and energy budgets in the ERW (Feldman et al., 2021). SAIL aims to develop a wide range of hydrometeorological datasets to constrain atmosphere, surface, and subsurface processes simultaneously. Together, these resources are contributing to the establishment of a highly-instrumented and studied UCRB watershed. We look forward to building upon the knowledge learned from this manuscript to compare the most appropriately configured IPM to SAIL and SPLASH campaign observations. Our study highlights that the benchmarking provided by these data collections will be critical in addressing the systemic IPM cold bias by providing a more constrained estimate of radiation budgets in complex terrain that ultimately shape snowmelt and discharge.”

references

Arnault, J., Wagner, S., Rummler, T., Fersch, B., Bliefernicht, J., Andresen, S., & Kunstmann, H, 2016: Role of Runoff–Infiltration Partitioning and Resolved Overland Flow on Land–Atmosphere Feedbacks: A Case Study with the WRF-Hydro Coupled Modeling System for West Africa, Journal of Hydrometeorology

Ban, N., Schmidli, J., and Schär, C. 2014: Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations, Journal Geophyscal Research Atmosphere

Forrester, M. and Maxwell, R. 2020: Impact of lateral groundwater flow and subsurface lower boundary conditions on atmospheric boundary layer development over complex terrain, Journal of Hydrometeorology

Frei, C. and Schär, C. 1998: A precipitation climatology of the Alps from high-resolution rain-gauge observations. Inter ational Journal of Climatology,

Keune, J., Gasper, F., Goergen, K., Hense, A., Shrestha, P., Sulis, M. and Kollet, S. 2016: Studying the influence of groundwater representations on land surface-atmosphere feedbacks during the European heat wave in 2003, Journal of Geophysical Research - Atmospheres

Keune, J., Sulis, M., and Kollet, S. J. 2019: Potential added value of incorporating human water use on the simulation of evapotranspiration and precipitation in a continental scale bedrock to atmosphere modeling system–A validation study considering observational uncertainty. Journal of Advances in Modeling Earth Systems

Miguez-Macho, G., Y. Fan, C. P. Weaver, R. Walko, and A. Robock, 2007: Incorporating water table dynamics in climate modeling: 2. Formulation, validation, and soil moisture simulation. Journal of Geophysical Research

Maxwell, R., J. K. Lundquist, J. D. Mirocha, S. G. Smith, C. S. Woodward, and A. F. B. Tompson, 2011: Development of a coupled groundwater–atmosphere model, Monthly Weather Review

Ryken, A. C., Gochis, D., & Maxwell, R. 2022: Unravelling groundwater contributions to evapotranspiration and constraining water fluxes in a high-elevation catchment. Hydrological Processes

Walser, A., and C. Schaer, 2004: Convection-resolving precipitation forecasting and its predictability in Alpine river catchments, Journal of Hydrology