Response to RC 2

General Remarks: This paper investigates the impacts of meteorological forcing in simulating rainfall-runoff behavior using the ATS model. Three different datasets are used to test the effects of spatial and temporal resolution of gridded meteorological forcing on surface runoff and other related hydrological processes. This is an interesting piece of work that provides certain decision support for the application of meteorological forcing.

Response: We thank the reviewer for the thoughtful and constructive feedbacks. The comments and suggestions are very helpful for improving this manuscript. We have addressed all the comments/concerns point-by-point below (highlighted in blue).

Major Comments

Reviewer Comment 2.1 — This study only tested on the Coal Creek Watershed, which has a relatively small size of about 53km2 and receives a high proportion of snowfall during cold season. Would the results be different if the gridded meteorological forcing was tested in other catchments? Is the result for individual catchment representative of general trends?

Response: We thank the reviewer for the critical comments. This question has also been raised by Reviewer 1. Please refer to the response in RC 1.10. For your convenience, we have pasted our response here:

Our study area–Coal Creek is one of the many mountain headwater catchments within the Upper Colorado River Basin, which is also one of the principal headwater basins in the U.S. and provides water for over 40 million people. The watershed is snow-dominated with strong variations in topography and land cover, which is an ideal site for testing heterogeneous spatial and temporal pattern of meteorological forcing. Our conclusions would hold for other mountainous headwater watersheds that are dominated by snow, and additional studies are needed to evaluate the GMF in other areas that are not dominated by snow. In addition, performing such studies in many watersheds is computationally expensive. We realize that this is a limitation and we have expanded the discussion on the transferability of current study in the revised manuscript.

In this study, we choose Coal Creek as an illustrative example to show the effects of meteorological forcing spatiotemporal resolution on watershed simulations. The study site has strong variations in topography and land cover, which is an ideal site for testing heterogeneous spatial and temporal pattern of meteorological forcing. Our conclusions would hold for other mountainous headwater watersheds that are dominated by snow because we did not make any site-specific assumptions. However, additional studies are needed to evaluate the GMF in other areas that are not dominated by snow.

Reviewer Comment 2.2 — The performance criterion is critical for model application. This study took modified Kling-Gupta efficiency as objective function to evaluate the model performance, which means that you specially focus on water balance. The results can be very sensitive to the selection of objective function. Have you considered some other performance criteria? Do they show different comparisons?

Response: We thank the reviewer for the suggestion of considering other metrics. In fact, we have used other metrics including the original KGE [Gupta et al., 2009], Nash-Sutcliffe Efficiency (NSE), and log(NSE) for evaluating the model performance during manuscript preparation. However, each metric focuses on a specific part of the model performance and has its own pitfalls [Clark et al., 2021]. For example, NSE is highly influenced by high flows whereas log(NSE) places a large emphasis on low flows. The modified KGE avoids the effect of input bias on the variability indicator which has an advantage over the original KGE [Kling et al., 2012]. Another advantage of using modified KGE is that it can be decomposed into three different constitutive parts (i.e., correlation coefficient, variability and bias) and thus could be used to diagnose the model performance score. In general, modified KGE follows the same trend of NSE. As a result, we choose to use a single metric that is physically explainable for better comparison of model performance. However, we can add the other metrics in the manuscript if the reviewer strongly encourage to do so.

Reviewer Comment 2.3 — The authors say that “The models were not calibrated because the focus of this study was to evaluate the effect of meteorological forcings on model simulation instead of estimating the optimal parameters used in ATS. ” The simulation results may vary for different parameters, and the parameters calibrated for different spatial and temporal resolutions may also be different. It can be seem from the hydrographs that the simulated discharge differ significantly from the observation and are underestimated for most of the time period. Does it have any effect on the results if the model parameters are calibrated in advance?

Response: We agree that the simulation results may vary if we use calibrated parameters, and the calibrated parameters may be different depending on the spatial and temporal resolution of GMF and the targeted hydrologic variables used during calibration. Therefore, we think that it is not appropriate to use the calibrated parameters from one specific spatial and temporal resolution of GMF to test the impact on model simulations forced by different spatial and temporal resolutions of GMFs. Further, the focus of the study was not on model calibration, but rather an evaluation of the effects of different GMFs on simulated watershed responses. We think that adding model calibration would detract from the current emphasis on GMF spatiotemporal effects. Although the current model is not calibrated, it still performs reasonably well (KGE=∼0.6 using Daymet) for streamflow.
In fact, we are working on a separate paper that uses novel machine learning technique for estimating model parameters for this watershed. In that paper, we are also investigating the impact of using different GMF on model calibration. We hope to cite that work when it becomes available.

Reviewer Comment 2.4 — The authors used different meteorological forcing to compare the effect of different temporal and spatial resolutions on the simulation results, which makes it difficult to state which is the greater impact to the simulation results. It is recommended that improvements in model performance at different temporal and spatial resolutions be investigated on the basis of the same data sources.

Response: We thank the reviewer for the excellent suggestion. Similar comments have also posted by Reviewer 1 (see RC1 comments 1.1 and 1.2). For your convenience, we have pasted our response here:

To be consistent with the spatial resolution comparison, we now use temporally downscaled PRISM data to study the effects of temporal resolution of meteorological forcing on watershed responses. Not surprisingly, the conclusions are still the same, although the metrics are slightly different. The simulated streamflow shows better performance using daily resolution of PRISM compared to that using sub-daily resolution of PRISM. We have updated both the comparison plots and the metric table in the revised manuscript. Also, we systematically varied each variable one at a time to isolate the effect of individual forcing. We now compared Daymet, PRISM and NLDAS by using their precipitation, or temperature or both in the model while keep the other variables the same. Not surprisingly, the resolution of precipitation plays a more important role than the resolution of temperature in streamflow simulations

Specifc Comments

Reviewer Comment 2.5 — The catchment is very small, what is the average time of concentration for the floods? Does it have an impact on the comparison of daily flood simulation results?

Response: To clarify, all the discharge plots were using hourly data. So we were comparing hourly simulated discharge with hourly observed discharge. Because this catchment is snow-dominated, most of the precipitation would first infiltrate into the subsurface as groundwater and then exfiltrate out at the stream channels. Although there is no direct way to estimate the average time of concentration, it is less likely to be a very short time (< 1 hour). In addition, ATS was taking sub-hourly timestep and the simulate discharge was integrated over hourly time window.

Reviewer Comment 2.6 — The grid size of NLDAS is larger than the actual area of the basin, it doesn't seem appropriate as a grid input.

Response: As has discussed in the manuscript, the coarse resolution of NLDAS would make the forcing highly uniform over a small catchment. As a result, NLDAS is commonly used in large basin-scale models. However, because NLDAS has the most complete dataset with hourly temporal resolution, it is still an attractive product for watershed model simulations. In this study, we were evaluating the performance of NLDAS for a small watershed to see if NLDAS could still provide any value. We found that NLDAS performed poorly for streamflow and other variables, but it preserved the dynamic watershed responses which were critical if interested in simulating flash flood or hyporheic exchange.

Reviewer Comment 2.7 — For Figure 4(also Figure 8 and 13), is it possible to show all the simulation results in one graph to have a better comparison of the different results?

Response: We have plotted all discharge time series in one graph and used flow duration curve to show better comparison.

Reviewer Comment 2.8 — It is difficult to distinguish the differences between the results in Figure 11. Is it possible to have a better comparison by using flow duration curve?

Response: We thank the reviewer for the suggestion. However, flow duration curve is commonly used for discharge and we were comparing surface ponded depth in Figure 11. We will zoom into high flow period to show more details between different depths.

References

Martyn P. Clark, Richard M. Vogel, Jonathan R. Lamontagne, Naoki Mizukami, Wouter J.M. Knoben, Guoqiang Tang, Shervan Gharari, Jim E. Freer, Paul H. Whitfield, Kevin Shook, and Simon Papalexiou. The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, aug 2021. ISSN 0043-1397. doi: 10.1029/2020WR029001. URL https://onlinelibrary.wiley.com/doi/10.1029/2020WR029001.

Hoshin V. Gupta, Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, 2009. ISSN 00221694. doi: 10.1016/j.jhydrol.2009.08.003. URL http://dx.doi.org/10.1016/j.jhydrol.2009.08.003.

Harald Kling, Martin Fuchs, and Maria Paulin. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424-425:264–277, 2012. ISSN 00221694. doi: 10.1016/j.jhydrol.2012.01.011. URL http://dx.doi.org/10.1016/j.jhydrol.2012.01.011.