Response to Reviewer #1:

First of all, we would like to thank reviewer #1 for his/her comments on the paper. Their effort has helped us to improve the manuscript and we appreciate you agreeing to review the paper during these challenging times. Here, we provide point-by-point responses to each of reviewer 1's comments.

Reviewer 1 comments	Author response
It needs to be demonstrated how/if the results of this study depend on the single year of the simulation. Why not extend the study period to include the last 30 years or so? All of the forcings used in this study go back to early 1980s. Even MODIS goes back to the early 2000s. A longer period of analysis can also help address the question of impact of forcing uncertainties on the long-term changes in mountain system recharge, which would be of interest given the focus on global warming driven changes in precipitation and temperature. A longer analysis period could also allow for independent verification of the mountain system recharge simulations such as by using GRACE based estimates of recharge, which goes back to early 2000s. This could help identify the set of atmospheric forcings which yield the most realistic estimates of the recharge.	You are correct, thank you. The model requires hourly forcing data and the spatially downscaled, hourly forcing that we use (Princeton CONUS Forcing) is only available from 2002. During this period, California suffered from the worst drought in recorded history. Our initial thoughts were that if we used this period, it would likely bias the model results based on these extreme years. But, as you state, it would make the findings more robust. To address reviewer comment in the revised manuscript, we will add two more years of simulation in addition to the water year (WY) 2016 originally selected. Please note that (WY) 2016 approximately represents average precipitation and air temperature in the watershed. To assess the impact of hydroclimatic condition on our results, we plan to perform simulations for WY2014 and WY2011, representing extreme dry and wet years, respectively. Similar to our approach for WY2016, we will evaluate model simulations results at equilibrium to remove the impact of initial condition bias on results. Integrated hydrologic models like ParFlow.CLM model are particularly sensitive to initial conditions, which is why we chose to analyze simulations that had reached equilibrium conditions for each water year. It is not feasible to allow a 15-30 year simulation reach equilibrium because of the computational demands of the model. Therefore, we plan to perform simulations for the select dry and wet water years at equilibrium conditions to assess generality of

	our results. While using GRACE data is ideal for confirming changes in terrestrial water storages, the resolution of GRACE data is too coarse for the study basin. Furthermore, we do not include irrigation and water management options in this version of the model so it is not possible to assess the impacts of forcing uncertainty over the entire basin.
Additionally, how are the results of this study dependent on the choice of the hydrologic model? As shown by Vano et al, 2012 (cited by this manuscript too) depending on the choice of hydrologic model sensitivity of hydrologic variables (such as runoff) to changes in precipitation and temperature can vary substantially.	We agree that the choice of hydrologic model will impact our results. In the revised manuscript we will make it clear that our results are applicable to integrated surface water – groundwater models that implement 3D Richard's equation to simulate variably saturated subsurface flow across the entire subsurface, and have a fully integrated overland flow simulator. Of course, different model physics will result in different sensitivities. However, we are using the most physically-based approach for simulating surface water-groundwater processes, and have done detailed model validation to make sure major hydrologic processes are captured by the model. Ideally, one should perform such simulations using different model structures to assess the impact of all uncertainty sources on simulated hydrologic response. We will add text to discuss the role of model selection in our results.
Finally, it also should be at least discussed how the results of this study may depend on the choice of the study domain.	Yes, this is a good point. Model parameterization and geologic setting are likely play a major role in how uncertainty in meteorological forcings will propagate into groundwater. In the revised manuscript, we will expand upon this discussion. For example, our simulations are performed in a mountain region underlain by fractured, low permeability bedrock. Previous work has shown that groundwater in these regions responds quickly to changes in precipitation (Pfister et al., 2017), which would likely impact the results. In the revised manuscript we will include discussion of how the

	geologic setting of our study site, and parameterization choices that we made are likely to impact the study results. While the role of uncertainty in precipitation forcing is discussed extensively, our main goal here was to highlight the role of temperature in addition to precipitation for regions with high relief. Of course, the obtained sensitivities in different mountain settings are impacted by the quality of meteorological forcings, topography, vegetation and subsurface characteristics.
I am surprised a bit about the differences in the simulated variables generated using GridMET, NLDAS and PRISM datasets. As described in Abatzoglou 2013, GridMET is based on the NLDAS-2 and PRISM dataset. Please at least discuss why this might be the case.	We agree that it is worth further highlighting these differences in the paper. We believe that the differences among products are caused by the fact that we are using different versions of PRISM and NLDAS-2 than the version used in Abatzoglou (2013) paper to generate the Gridmet dataset. To build the Gridmet dataset, they used the 800 m resolution version of PRISM, while we used the freely available 4 km resolution of PRISM data. Additionally, we used a downscaled version of the NLDAS-2 dataset, called the Princeton CONUS Forcing dataset with ~3 km resolution. As described in the paper, the Princeton dataset is the downscaled version of the original NLDAS-2 data with ~12 km resolution and the rainfall data is updated by using the radar products. We believe that the differences in the resolution of the dataset and interpolation approach have caused the differences in precipitation forcing datasets. In the revised manuscript we will add this information to clarify.
Page 1, line 28: “high qualify” should be “high quality”	Thank you for pointing this out, it will be fixed in the revised manuscript.
Page 8, 217-219, how does this chosen threshold of 2.5 deg C for partition of precipitation into rainfall and snow, affect the results of this analysis, especially in the mid to low elevation parts of the domain?	To be clear, this threshold was not chosen by us, it is the threshold that the CLM model uses to partition precipitation into rainfall and snow. That being said, this threshold likely impacts the results. However, we did not assess its impacts.

	In the mid to low elevation portions of the domain, where precipitation can currently fall as either rain or snow, the snow melts quickly and snowpack does not accumulate due to higher temperatures in mid-elevation regions. We will add discussion to the revised manuscript, in the same location where we discuss the effect of model formulation on the results, to include the impact of temperature threshold to partition between rain and snow.
Page 22, lines 496-498, I am not sure why land surface temperature and soil moisture would not be affected by the choice of forcings, wouldn't changes in ET affect both? Please clarify.	Yes, changes in ET does affect both land surface temperature and soil moisture. At the annual scale, however, changes in soil moisture are small because changes in ET can be balanced out by changes in potential recharge and lateral soil moisture redistribution. In the revised manuscript, we will clarify the text in this section. We believe that lower sensitivity to land surface temperature is partly related to the simplification made to represent the ground heat flux calculation in CLM. Many land surface models, including CLM, only incorporate heat transport via conduction and this simplification decouples heat transport from soil moisture transport. Including heat convective transport through soil moisture distribution will increase computational time. While the ParFlowE model (Kollet et al., 2009) incorporates these processes, we did not use this version of ParFlow in our study. We will add this discussion to our revised manuscript.

References:

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, *33*(1), 121–131. https://doi.org/10.1002/joc.3413

Kollet, S.J., Cvijanovic, I., Schüttemeyer, D., Maxwell, R.M., Moene, A.F. and Bayer, P. (2009), The Influence of Rain Sensible Heat and Subsurface Energy Transport on the Energy Balance at the Land Surface. *Vadose Zone Journal*, 8: 846-857. https://doi.org/10.2136/vzj2009.0005

Pfister, L., Martínez-Carreras, N., Hissler, C., Klaus, J., Carrer, G. E., Stewart, M. K., and McDonnell, J. J. (2017). Bedrock geology controls on catchment storage, mixing, and release: A comparative analysis of 16 nested catchments. *Hydrological Processes*, 31(10), 1828–1845. https://doi.org/10.1002/hyp.11134