The authors have done a very thorough job revising the manuscript, and it is much improved. I have only minor comments before I believe the manuscript will be ready for publication.

Major comments:

I find the phrase "subgrid-scale physics schemes" a little general. For a coupled model, this could refer to the representation of hydrology in ParFlow-CLM (or other hydrologic models), but in this study it refers to physics schemes within WRF. In the abstract (or even the title), could you clarify that the physics schemes here refer to the physics within WRF?
Response: We added "in WRF" on line 19 and line 23 of the abstract.

I also have one lingering comment about the poor agreement between streamflow simulations and observations, as well as the related temperature bias. The poor streamflow fit suggests that even if regional climate models like WRF outperform observations with respect to precipitation, a temperature bias could confound this with respect to their ability to act as accurate hydrologic model forcings. I think it would be valuable to mention as early as the abstract (and possibly expand in discussion) that all the WRF configurations have a temperature bias relative to PRISM that's comparable to what we would expect due to climate change impacts. Maybe that will just tell readers that bias correction is still important, but it seems like information that should be available early.
Response: We agree with the reviewer and revised the abstract as "Despite reasonably simulating precipitation, a delay in simulated peak discharge is due to a systematic cold bias across WRF simulations, suggesting the need for bias correction."

Page 1
Line 21 – should be "uncertainty from synoptic-scale forcings…"?
Response: Revised as suggested.

Line 25 – "delayed earlier" – This is a contradiction in my view. Were flows delayed or earlier?
Response: Deleted "earlier"

Page 4
Line 101 – "However, neither… do not…" Did you intend the double-negative here?
Response: Yes, we removed the "do not"

Page 18
Table 3 – The standout conclusion from this table, in my view, is that all models have pretty terrible R_2 of precipitation relative to PRISM. Is this due to stochasticity in the daily time series generation, or something else? If it's event-scale stochasticity, would a different metric be more appropriate?
Response: We agree with the reviewers that the R2 of precipitation relative to PRISM are relatively low. The misfit is mostly due to overestimation of a few precipitation events from February through April. In the revision, we added the 95th percentile of daily precipitation in Table 3 to analyze the event-scale stochasticity of precipitation simulation.

The text on line 393 are updated as "Although NCAR-CFSR has a higher R2 than other simulations, NCAR-ERA5 has a very low R2. The BSU simulations provide a closer approximation of cumulative precipitation to PRISM. Specifically, BSU does better in simulating extreme precipitation events (i.e., 95th percentile). Therefore, we conclude that BSU WRF subgrid-scale physics schemes outperform the UCD and NCAR WRF subgrid-scale physics schemes in simulating both precipitation and temperature."

Figure 4 provides a time-averaged spatially explicit comparison between PRISM and modeled temperature; why don't we have the same thing provided for precipitation?
Response: I'm not sure if I understand the reviewer's comment or the reviewer is missing anything. The time-averaged spatially explicit comparison between PRISM and modeled precipitation is in Figure 4b).

Page 19
First line of page - Based on Table 3, NCAR-CFSR seems like a more attractive fit to PRISM than the BSU simulations. The temperature R_2 is slightly worse, but the precipitation is substantially better. Are there other factors that made the BSU simulations better?
Response: While the NCAR-CFSR has a higher R2 than others, but the NCAR-ERA5 has a very low R2 so we are concerning about its capabilities under the uncertainties of various large-scale meteorological forcings datasets. Although the BSU scenarios have relative lower R2 of precipitation, the cumulative precipitations are consistently closer to the PRISM. Also compared to BSU, the NCAR scenarios have relatively severe colder bias, which is one of the major concerns identified in this study. Overall, we decided to choose the BSU scheme.

The text on line 393 are updated as "Although NCAR-CFSR has a higher R2 than other simulations, NCAR-ERA5 has a very low R2. The BSU simulations provide a closer approximation of cumulative precipitation to PRISM. Specifically, BSU does better in simulating extreme precipitation events (i.e., 95th percentile). Therefore, we conclude that BSU WRF subgrid-scale physics schemes outperform the UCD and NCAR WRF subgrid-scale physics schemes in simulating both precipitation and temperature."