Referee 2

The results of this paper are significant for the hydrologic modeling and LSM communities, as they demonstrate the impacts lateral flow in LSMs and the possible implications for atmospheric fluxes. The methods of this manuscript are generally strong, and I anticipate that the results of this work could have implications for the hydrometeorological community.

However, I have some technical concerns on the methods that I ask the authors to clarify in order for this manuscript to be accepted.

We thank the referee for the encouraging comments, and for seeking clarification on some of the technical aspects. We have given a point-by-point response to all the comments, describing the changes that will be made to the revised manuscript to incorporate them. We believe that the revisions fully satisfy the referee's concerns.

Referee comments are shown in bold. Author responses are shown in plain text.

Major Comments:

1. **Section 2.1, Lines 100-110: I appreciate this discussion on the significance of horizontal routing with LSMs. I also recommend discussing the impacts of LSM depth, as the standard 2-m Noah-MP depth often does not capture groundwater processes and has resulted in dry biases with ET in some regions.**

We thank the referee for raising this and agree that the 2-m soil depth may contribute to LSM biases. In the revised manuscript, we will elaborate on this point as below.

"Here we use the standard Noah-MP LSM which has a constant soil depth of 2 m with vertically homogeneous soil parameters. This formulation can contribute to biases in runoff and evapotranspiration, which may be ameliorated by incorporating variable and higher soil depths, groundwater processes, and vertical soil heterogeneity (Gochis et al., 2010; Barlage et al., 2015; Wu et al., 2021; Yimam et al., 2025) in the modelling framework. These aspects are outside the scope of the work presented in this study and have not been explored here."

References added:

Barlage, M., Tewari, M., Chen, F. *et al.* The effect of groundwater interaction in North American regional climate simulations with WRF/Noah-MP. *Climatic Change* **129**, 485–498 (2015). https://doi.org/10.1007/s10584-014-1308-8

Gochis, D. J., Vivoni, E. R., and Watts, C. J.: The impact of soil depth on land surface energy and water fluxes in the North American Monsoon region, Journal of Arid Environments, 74, 564–571, https://doi.org/10.1016/j.jaridenv.2009.11.002, 2010.

Wu, W.-Y., Yang, Z.-L., and Barlage, M.: The Impact of Noah-MP Physical Parameterizations on Modeling Water Availability during Droughts in the Texas–Gulf Region, Journal of Hydrometeorology, 22, 1221–1233, https://doi.org/10.1175/JHM-D-20-0189.1, 2021.

Yimam, Y. T., Neely, H. L., Morgan, C. L. S., Kishné, A., Gross, J., and Gochis, D.: Evaluation of Noah-MP performance with available soil information for vertically heterogenous soils, Agrosystems, Geosciences & Environment, 8, e70048, https://doi.org/10.1002/agg2.70048, 2025.

2. **Section 2.2 and 2.3, Calibration: Please clarify why the authors use 3-day averaging for streamflow validation and calibration? This could likely underestimate major surface runoff events leading to flash flooding. Furthermore, calibration to 45-day periods may not capture the full range of processes that lead to hydrologic response. I recommend demonstrating that these parameters are consistent with a longer range calibration period. If this is not easily feasible within project constraints, I alternatively recommend discussing the assumptions and possible limitations behind this methodology decision.**

We thank the referee for the question and recommendations. Preliminary work was undertaken to arrive at the calibration settings documented in the manuscript. As computational resources preclude us from demonstrating consistency with a long-range calibration period, we will include the following information in the revised manuscript to elaborate on the calibration choices and its implications.

"Calibration of WRF-Hydro is computationally intensive and involves choices that may be aligned to the purpose of the simulations. Here we study the influence of lateral flow on seasonal timescales and hence the main purpose of calibration is to obtain better streamflow outcomes on monthly to seasonal timescales, rather that improved simulations of daily scale streamflow events. The streamflow in the domain primarily occurs in the cool season (May to October), and model simulations using default parameter values exhibit biases during these high flow months (Fig. 2). However, preliminary results showed that event-based calibration to high flow days did not translate to improved monthly flows indicating that it is necessary to use a period at least of the order of a month, which includes both high and low flow days at the four gauges for calibration. As a 45-day period is computationally feasible, and yields reasonable outcomes at monthly to seasonal timescales, this length of time is chosen for calibration. The daily streamflow data is smoothed by aggregating to 3-day flows to dampen the effect of individual high flow days."

Minor Comments:

3. **Figure 1: Consider adding the channel network (even if it is only higher order channels) to help the reader visualize the hydrologic connectivity.**

We thank the referee for the suggestion to improve the figure. We will revise Figure 1 as shown below to include the channel network.

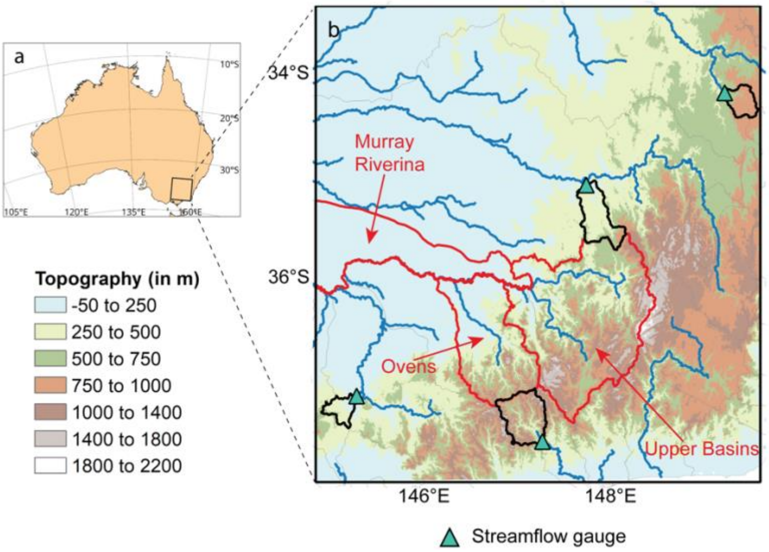

Figure 1: (a) The WRF-Hydro model domain in southeast Australia. (b) Features in the domain. Background shading indicates topography, and the surface water catchments (black outlines) that drain into the streamflow gauges (blue triangles) used for calibration are marked on the map. The basins outlined in red (Upper Basins, Ovens, and Murray Riverina) are used to analyse the influence of lateral flow in basins with varying topographic characteristics. The network of major rivers (blue lines) based on data from Geoscience Australia (https://pid.geoscience.gov.au/dataset/ga/42343) are shown in panel b.

4. **Section 2.2.1, Geographic Data: I find it surprising that the TERN dataset produced worse streamflow compared to the default soil dataset. Is this something that could eventually be improved with calibration?**

We thank the referee for the question. We agree that it is surprising that the TERN dataset produced worse streamflow, and we had briefly discussed this in the original manuscript (lines 135-145). Based on our results, rather than improving through calibration, we speculate that that it may be possible to obtained improved streamflow outcomes by utilising the TERN soil data in conjunction with regional soil parameter measurements.

We will improve upon lines 135-145 from the original manuscript as below.

"The better streamflow simulations obtained using the default soil dataset rather than the TERN dataset may be surprising, as the TERN data, which utilises regional observations (Teng et al., 2018), likely provides more accurate soil information over Australia. This is possibly because the modelled influence of soils on surface water partitioning in each land column relies on soil parameters in addition to soil type. The land model uses parameters (such as moisture at saturation, field capacity, wilting point, saturated hydraulic conductivity) for each soil type, the default values for which are defined based on scarcely available field observations in various regions (Kishné et al., 2017). Our results suggest that regional measurements that can be used to refine the default parameters values in conjunction with the regional soil datasets, such as the TERN dataset, may be necessary to obtain improved simulations."

5. **Section 3.1.2, lines 215-220: The negative ET bias might reflect the limits of the 2m LSM. I recommend connecting back to this point in the discussion.**

We thank the referee for the suggestion to include the point about soil depth in the discussion. We will include this point in the paragraph detailing avenues of future work.

"…, more accurate representation of processes including baseflow and the feedback from the channels to the soil columns may be relevant in arid and semi-arid regions, such as southeast Australia. These processes are not represented in our simulations. Our calibration results suggest that spatially varying baseflow may improve lateral flow representation in our domain. Incorporating soil layers deeper than the 2-m depth modelled in standard Noah-MP have reduced surface flux biases in some cases in other semi-arid locations (Gochis et al., 2010; Barlage et al., 2015) and may be explored in future work. Incorporating channel seepages have improved streamflow simulations in semi-arid Arizona (Lahmers et al., 2021), but the feedback of this process to the soil column has not been modelled yet."