Review2

This study presents a large-scale evaluation of snowmelt runoff (SMR) simulation across 15 global hydrological models and runoff products using 1,513 snow-dominated basins worldwide. The authors evaluate three key hydrograph characteristics: total runoff (Qsum), peak flow (Qmax), and centroid timing (CTQ), and introduces a new robustness index to quantify how model performance degrades with increasing basin complexity. The results show that most models underestimate runoff magnitude and predict earlier snowmelt timing, and that model performance degrades as basin complexity increases. Overall, the manuscript is well written and provides a valuable large sample diagnostic of model performance. The concept of evaluating model robustness across environmental complexity gradients is particularly interesting and could offer useful insights for model development. However, several issues need clarification before the manuscript is suitable for publication.

Response: We thank the reviewer for providing very useful comments for us to improve our manuscript. Below, please find our responses to address your concerns.

Major comments:

1. The robustness metric is central to the manuscript but requires further justification and interpretation. The index combines the Stratified Mean Absolute Bias (SMAB) and the slope of bias vs. complexity into a Euclidean distance metric. Why was Euclidean distance selected as the combination method? Though the authors mentioned it was done in prior studies, it would be more helpful for readers if some justification can be added here. Also, is the robustness index comparable across different runoff metrics (Qsum, Qmax, CTQ)?

Response: We thank the reviewer for this important comment. We agree that the formulation and interpretation of the Robustness Index (RI) should be more clearly justified, because it is a central metric in this study. In the revised manuscript, we have expanded **Section 2.2.4** to better explain the physical meaning of the two RI components, the rationale for using a Euclidean-distance-based formulation, and the comparability of RI across different runoff characteristics.

First, we clarified that SMAB and the slope represent two complementary aspects of model robustness (**Figure 1,** [Figure A3 in the manuscript]). SMAB measures the overall magnitude of model bias across the full basin-complexity gradient and is interpreted as a measure of performance stability. In comparison, the slope of absolute bias against CI measures how rapidly model bias changes as basin complexity increases and is interpreted as a measure of adaptability to complex basin conditions. A robust model should therefore have both low SMAB and low slope, indicating low overall bias and weak performance degradation under increasing basin complexity.

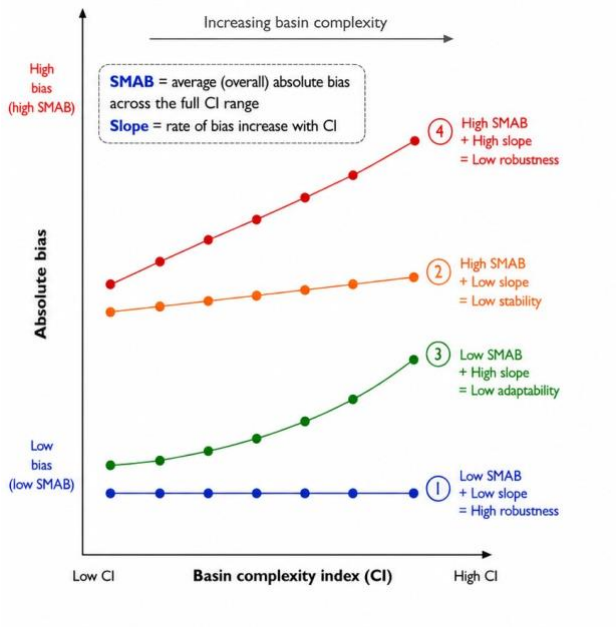

Figure 1. Conceptual illustration of the Robustness Index (RI). *The stratified mean absolute bias (SMAB) represents the overall magnitude of model bias across the full basin complexity index (CI) gradient, whereas the slope represents the rate at which model bias changes with increasing CI.*

Second, we have expanded the justification for using the Euclidean distance. After normalization, SMAB and slope define a two-dimensional metric space, where the origin represents an ideal model with zero average bias and no degradation along the CI gradient. The Euclidean distance from this ideal point provides a simple and interpretable way to quantify the joint deviation from the ideal condition. This formulation penalizes both large overall bias and strong sensitivity to basin complexity, while treating stability and adaptability as equally important. Distance-based approaches have also been widely used in previous studies to integrate multiple performance dimensions, and we now explain this more explicitly in the revised manuscript.

Third, to test whether the RI-based model ranking depends on the specific aggregation method, we added a sensitivity analysis in the Supplementary Information (**Table 1,** [Table A1 in the manuscript]). We compared the baseline equal-weight Euclidean-distance formulation with alternative formulations, including weighted Euclidean-distance formulations and weighted linear formulations.

Table 1. Sensitivity experiments used to test the robustness of the RI formulation

Experiment	Formulation	Weight setting	Purpose
S0	$RI = 1 - \sqrt{w \times SMAB_{norm}{}^2 + (1-w) \times S_{norm}{}^2}$	w = 0.5	Baseline RI formulation.
S1-1		w = 0.3	Euclidean aggregation; adaptability emphasized.
S1-2		w = 0.4	Euclidean aggregation; adaptability slightly

			emphasized.
S1-3		w = 0.6	Euclidean aggregation; stability slightly emphasized.
S1-4		w = 0.7	Euclidean aggregation; stability emphasized.
S2-1		w = 0.3	Linear aggregation; adaptability emphasized.
S2-2		w = 0.4	Linear aggregation; adaptability slightly emphasized.
S2-3	$RI = 1 - (w \times SMAB_{norm} + (1 - w) \times S_{norm})$	w = 0.5	Linear aggregation; equal weighting.
S2-4		w = 0.6	Linear aggregation; stability slightly emphasized.
S2-5		w = 0.7	Linear aggregation; stability emphasized.

The results show that most models exhibit limited changes in both RI magnitude and relative ranking across sensitivity experiments (Figure 2, [Figure A4 in the manuscript]). Models with consistently high or low robustness remain generally stable, and the rankings for CTQ are particularly insensitive to the aggregation method. Based on these results, and considering the common use and interpretability of Euclidean-distance-based metrics, we retained the equal-weight Euclidean-distance formulation as the baseline RI calculation.

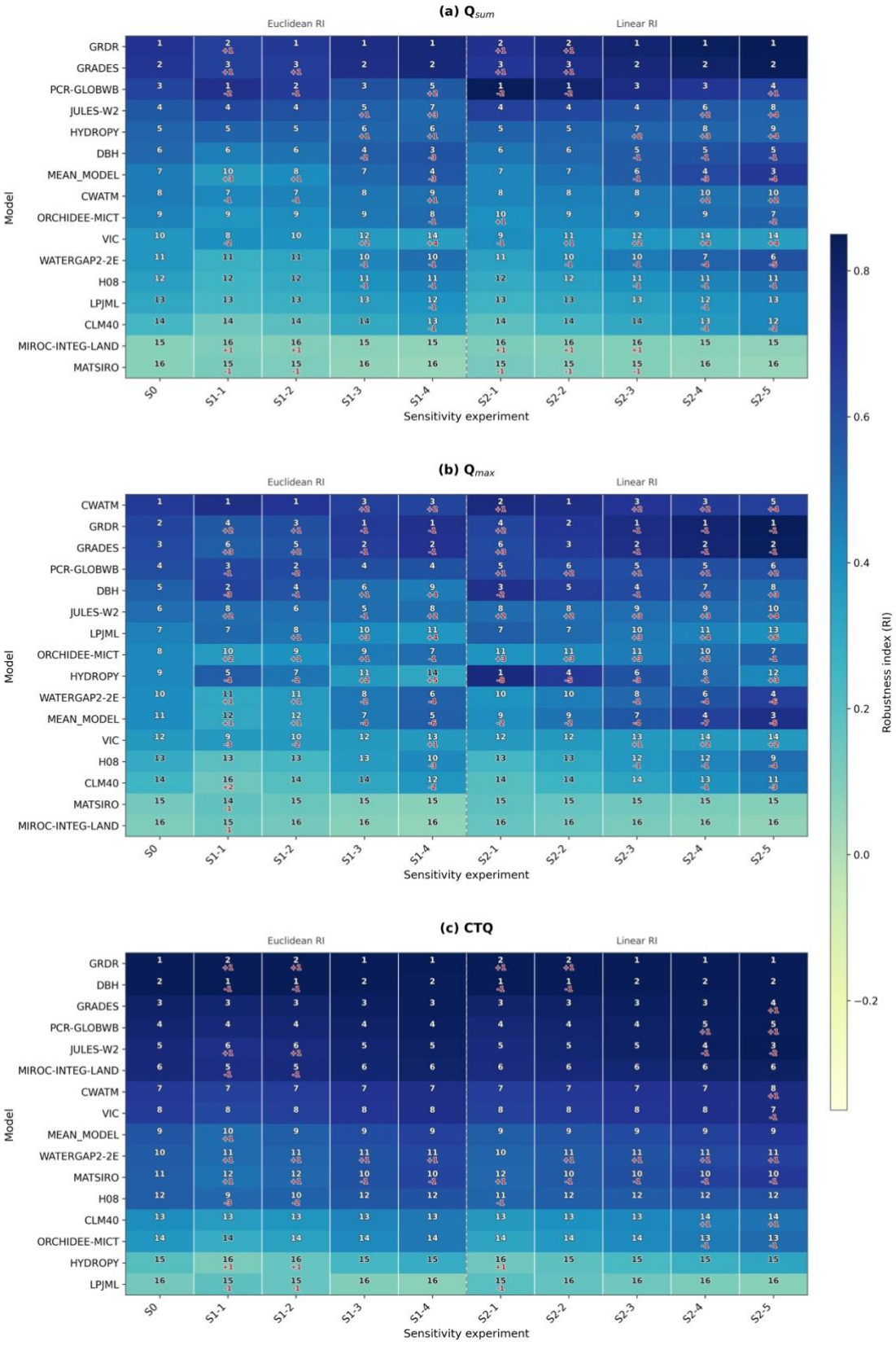

Figure 2. Sensitivity analysis of the Robustness Index (RI) formulation. *Panels (a–c) present the sensitivity of model robustness rankings for Qsum, Qmax, and CTQ, respectively. The sensitivity experiments include the baseline Euclidean-distance-based RI (S0), weighted Euclidean formulations (S1-1 to S1-4), and weighted linear formulations (S2-1 to S2-5). The color shading*

represents the RI value, and the number in each cell denotes the model rank under the corresponding experiment. Red numbers indicate rank changes relative to the baseline RI formulation.

Regarding the comparability across runoff metrics, we have clarified that RI is primarily designed for within-metric inter-model comparison, for example comparing model robustness for Qsum, Qmax, or CTQ separately. Because Qsum, Qmax, and CTQ have different physical meanings, units, bias definitions, and acceptable thresholds, their absolute RI values should not be interpreted as strictly interchangeable across metrics. However, since SMAB and slope are normalized before calculating RI, the metric allows a cautious qualitative comparison of robustness patterns among Qsum, Qmax, and CTQ, such as identifying that CTQ robustness is generally more sensitive to increasing basin complexity than runoff magnitude metrics. We have revised the manuscript to make this interpretation clearer.

2. The basin complexity index is defined as the sum of normalized DEM, DEMstd, LAI, and PFTh. While this approach is straightforward, the four variables may not contribute equally to hydrological complexity. Some factors, like elevation and topographic variability, may already be strongly correlated. It would be better if the authors can discuss why these four metrics are selected and whether dependencies among these variables influence the analysis.

Response: We agree that the selection of CI components and the potential dependence among them should be better justified. In the revised manuscript, we have expanded the explanation of why DEM, DEMstd, LAI, and PFTh were selected, and we added a correlation analysis of the four CI components to examine whether the integrated CI is dominated by strongly correlated variables.

First, we clarified that these four metrics were selected to represent two major physical dimensions of basin complexity relevant to SMR simulation: topography and vegetation. DEM and DEMstd describe topographic controls, with DEM representing the mean elevation background that affects temperature, precipitation phase, snow accumulation, sublimation, and radiation conditions, and DEMstd representing within-basin terrain variability that influences snow redistribution, surface energy balance, and runoff generation. LAI and PFTh describe vegetation controls, with LAI reflecting canopy density and its effects on snow interception, sublimation, and radiation transfer, and PFTh representing vegetation-type diversity and the heterogeneity of canopy–snow interactions.

Second, to examine potential dependencies among these variables, we added a correlation matrix in the Supplementary Information (**Figure 3,** [Figure A1 in the manuscript]).

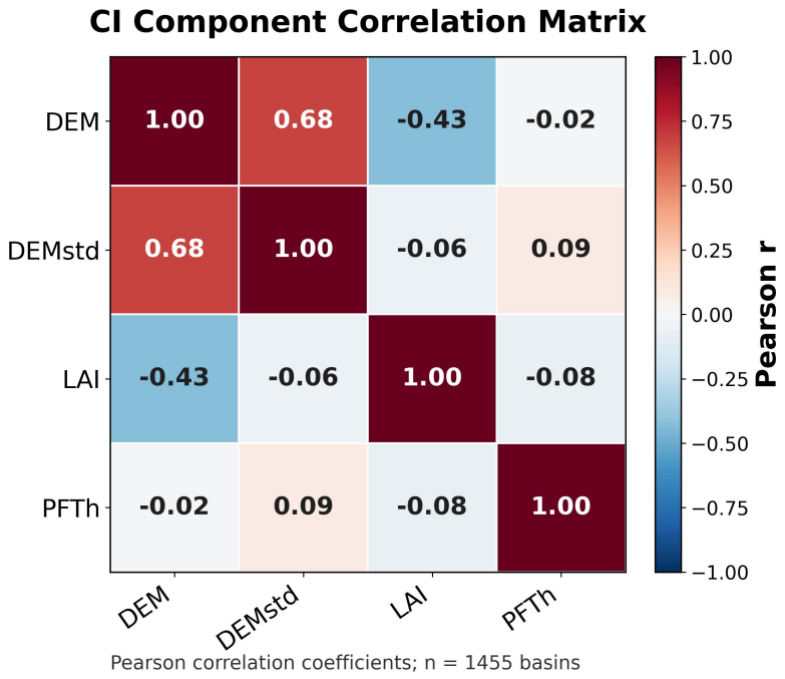

Figure 3. Pearson correlation matrix among the four components used to construct the basin complexity index (CI).

The results show that DEM and DEMstd are positively correlated (r = 0.68), indicating that high-elevation basins often also have stronger topographic relief. However, these two variables are not identical in physical meaning: DEM captures the mean elevation and associated climatic background, whereas DEMstd captures sub-basin terrain heterogeneity. LAI is moderately negatively correlated with DEM (r = −0.43), suggesting that higher-elevation basins generally have lower vegetation density, which is consistent with the transition from vegetated lowlands to colder or more mountainous environments. In comparison, PFTh shows very weak correlations with DEM, DEMstd, and LAI ($|r| \leq 0.09$), indicating that vegetation-type diversity provides largely independent information. DEMstd also shows weak correlations with LAI and PFTh (r = −0.06 and 0.09, respectively), suggesting that terrain variability and vegetation complexity are not redundant.

These results indicate that, although some dependence exists among the CI components, especially between DEM and DEMstd, the four variables capture different and complementary aspects of basin complexity. In particular, DEM and DEMstd are not redundant in their physical meanings: DEM mainly represents the background elevation control on temperature, precipitation phase, and snow accumulation conditions, whereas DEMstd reflects topographic variability and sub-basin heterogeneity, which can influence radiation, melt timing, and runoff generation. Distinguishing these two aspects is also central to our experimental design, as the study aims to examine how both mean-state environmental controls and within-basin heterogeneity affect SMR simulation. Similarly, the vegetation-related variables, LAI and PFTh, represent canopy density and vegetation-type diversity, respectively, and thus provide complementary information on canopy–snow and energy-exchange processes. 、

We have therefore retained the four-component CI as a transparent and physically interpretable composite index. At the same time, we now acknowledge in the revised manuscript that the equal-weight summation is a simplified representation. Future work could further refine this framework

by incorporating a broader set of representative complexity descriptors and by testing alternative weighting schemes to more comprehensively characterize basin complexity.

3. In this study, all runoff outputs are routed using RAPID to ensure comparability. However, the manuscript assumes that routing effects are minimal because long-term mean metrics are used. This assumption needs more justification because routing parameters can influence peak flow magnitude and CTQ. The authors should either provide a short sensitivity analysis, or cite studies showing that routing effects are negligible at the spatial scale considered.

Response: We agree that the routing scheme and routing parameters may influence simulated discharge characteristics, especially Qmax and CTQ, because these two metrics are more sensitive to flow concentration, travel time, and hydrograph attenuation than Qsum. Our original statement that routing effects are "minimal" was too strong and insufficiently justified.

In the revised manuscript, we have therefore moderated this statement and clarified that the use of RAPID is intended to ensure a consistent routing framework across all models, rather than to eliminate routing-related uncertainty. All runoff outputs were mapped to the same MERIT-Basins river network and routed using the same RAPID configuration, which improves inter-model comparability. However, we now explicitly acknowledge that routing parameters, especially the Muskingum travel-time parameter, may still affect the absolute values of Qmax and CTQ. Accordingly, we have revised the manuscript to avoid implying that routing effects are negligible. Instead, we now state that routing uncertainty should be considered when interpreting Qmax and CTQ, while emphasizing that the consistent routing framework helps reduce routing-induced differences among models.

4. Page 19, lines 340: what's the potential reason that ISIMIP 3a outperforms ISIMIP 2a in simulating Qsum and Qmax, is it because of the forcing data?

Response: We agree that the better performance of ISIMIP3a relative to ISIMIP2a may be partly related to differences in forcing data, and this point should be discussed more explicitly.

In the revised manuscript, we have added a discussion to clarify that the improved performance of ISIMIP3a likely reflects the combined effects of forcing-data improvement and model-generation advancement. On the one hand, ISIMIP3a models are driven by GSWP3-W5E5, whereas ISIMIP2a models use GSWP3. GSWP3-W5E5 includes additional bias correction and is generally expected to provide improved meteorological forcing consistency, which can affect runoff magnitude and peak flow simulation. To examine this effect, we added a supplementary comparison of model performance under the two forcing datasets (**Figure 4**).

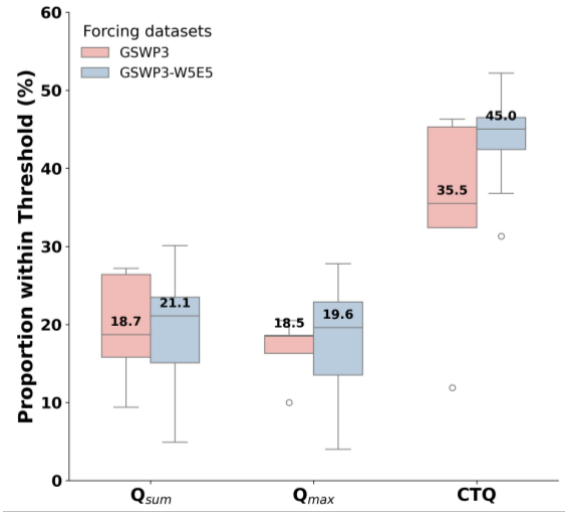

Figure 4. Effects of forcing datasets on model performance for Qsum, Qmax, and CTQ. *(a) compares the results under different forcing datasets (GSWP3 and GSWP3-W5E5). The thresholds are defined as ±20% for Qsum and Qmax, and ±5 days for CTQ. Numbers above the boxplots indicate median values.*

The results show that the median proportion of basins satisfying the predefined thresholds is consistently higher under GSWP3-W5E5 than under GSWP3, with median values increasing from 18.7% to 21.1% for Qsum, from 18.5% to 19.6% for Qmax, and from 35.5% to 45.0% for CTQ. These results support the view that forcing improvements likely contributed to the generally better performance of ISIMIP3a.

On the other hand, we do not attribute the differences solely to forcing data. Many ISIMIP3a models also include refinements in process representation compared with earlier model generations and some algorithmic updates, which may further improve their ability to simulate snowmelt runoff. Therefore, the better performance of ISIMIP3a is more appropriately interpreted as the result of both improved forcing data and model-development progress, rather than forcing alone.

We have revised the manuscript accordingly to provide a more balanced interpretation.

5. The manuscript concludes that GHMs outperform LSMs for Qsum and Qmax, while LSMs perform better for CTQ. This is an interesting finding, but the discussion remains somewhat speculative. The authors attribute the differences to energy-balance representations and runoff parameterizations, but more concrete explanations or references would strengthen this argument. In addition, some differences could result from calibration, forcing datasets, and resolution differences. These factors should be discussed more carefully.

Response: We thank the reviewer for this insightful comment. We agree that we should more sufficiently explain the mechanisms behind the contrasting performance of GHMs and LSMs. In the revised manuscript, we have substantially expanded the discussion and clarified that the observed differences likely arise from a combination of runoff-characteristic sensitivities, process representations, calibration strategies, and forcing-data differences.

First, the three runoff characteristics evaluated in this study reflect different physical aspects of the snowmelt runoff process. Qsum and Qmax are primarily related to the total water mass released during snowmelt and the resulting runoff response magnitude. These characteristics are therefore strongly influenced by runoff generation and water-balance parameterizations. In contrast,

CTQ mainly reflects the timing and rate of snowmelt release, which are more directly controlled by energy-exchange processes and snowpack evolution.

Second, the different process emphases of GHMs and LSMs likely contribute to their contrasting performance. As further quantified in Part II using the Tree-Based Model Complexity Scoring (TBMCS) framework, LSMs generally include more sophisticated energy-balance representations, canopy radiative transfer schemes, and multi-layer snowpack parameterizations. These process representations are important for capturing the timing of snowmelt and may contribute to improved CTQ performance. In contrast, many GHMs place stronger emphasis on reproducing basin-scale runoff magnitude and water balance, and several are calibrated specifically against streamflow observations. Such calibration strategies can improve Qsum and Qmax performance, even when the representation of energy-related snow processes is relatively simplified.

Third, we now explicitly acknowledge that the observed differences cannot be attributed solely to model structural characteristics. Other factors, including calibration status, forcing datasets, and spatial resolution, may also influence the comparison. For example, several GHMs in our ensemble are calibrated, whereas most LSMs are not, which may partly contribute to the superior GHMs performance for Qsum and Qmax. Similarly, ISIMIP3a models are driven by the bias-corrected GSWP3-W5E5 forcing dataset, which generally performs better than GSWP3 and may further enhance runoff simulation. We therefore revised the manuscript to emphasize that the differences between GHMs and LSMs reflect the combined influence of process representation, calibration strategy, forcing data, and model structure, rather than a single controlling factor.

Minor comments:

1. The snowmelt period is defined as the interval between maximum SWE and when SWE falls below 1 mm. This definition may not capture multi-peak melt seasons or rain-on-snow events. A brief discussion of limitations would be helpful.

Response: We agree that this definition is intended to extract the main seasonal snowmelt signal in a consistent way across all basins and all models, rather than to identify every short-term melt event. Although this choice has clear references (Trujillo & Molotch, 2014; Yang et al., 2025), and improves comparability in a global-scale large-sample assessment, but it may smooth or omit finer-scale melt processes in regions with complex seasonal transitions. In the revised manuscript, we have added a discussion of this limitation.

2. ERA5 SWE is used to define snowmelt timing. However, ERA5 has known biases in mountainous regions. The manuscript should briefly discuss how this may affect the analysis.

Response: We agree that ERA5 SWE may contain biases, particularly in mountainous regions where snow accumulation and melt are strongly affected by elevation gradients, slope, aspect, etc. Such biases may influence the identified snowmelt period and consequently affect the derived Qsum, Qmax, and CTQ. In the revised manuscript, we have added a discussion of this uncertainty. We clarify that ERA5 SWE is used mainly to identify the timing of the main seasonal snowmelt period, rather than to evaluate the absolute SWE magnitude. Therefore, although SWE magnitude biases may exist, their influence on this study is expected to be smaller than in analyses directly evaluating SWE amount. Nevertheless, uncertainties in ERA5-derived melt timing may still affect the results, especially in complex terrain, and should be considered when interpreting the evaluation.

3. “Stern conditions” is not commonly used in scientific literature, especially in hydrology or Earth system science papers. It’s a bit awkward in this context. Do you mean “challenging conditions” or “complex environmental conditions”?

Response: To improve clarity and consistency, we have replaced “stern conditions” throughout the manuscript with “complex environmental conditions.”

4. Page 18, in the title of Figure 6, change “blue circles denote GHMs, yellow squares denote LSMs, green triangles denote DGVMS, and grey diamonds denote data products.” to “circle denote GHMs, squares denote LSMs, triangles denote DGVMS, and diamonds denote data products.” Because shapes represent model types and colors shows the robustness.

Response: We thank the reviewer for pointing this out. We agree that the original figure caption was inaccurate because the model categories are represented by marker shapes, while the colors indicate robustness values. We have revised the caption accordingly.

Reference:

Trujillo, E., & Molotch, N. P. (2014). Snowpack regimes of the Western United States. Water Resources Research, 50(7), 5611–5623. https://doi.org/10.1002/2013WR014753

Yang, Y., Pan, M., Feng, D., Xiao, M., Dixon, T., Hartman, R., et al. (2025). Improving streamflow simulation through machine learning-powered data integration and its potential for forecasting in the western U.S. Hydrology and Earth System Sciences, 29(20), 5453–5476. https://doi.org/10.5194/hess-29-5453-2025