We'd like to sincerely thank the editor and two anonymous reviewers for taking time and offering very constructive comments. Below is our point-by-point response.

Review1

Overall, this manuscript presents a substantial amount of analysis on snowmelt runoff characteristics across a large sample of basins and multiple models/products, and the overall writing and presentation are generally clear. The topic is relevant and the study has clear value for large-scale model evaluation in cold-region hydrology. In particular, the authors made considerable efforts in constructing the intercomparison framework and diagnosing runoff volume, peak, and timing during snowmelt periods. However, several issues remain insufficiently addressed, especially regarding the parameter calibration and the formulation and interpretation of the newly proposed RI metric. My detailed comments are as below.

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. A major concern is the issue of parameter calibration. It is well accepted in hydrological modeling that calibration can strongly affect model performance, especially that hydrological models typically involve many empirical assumptions, conceptual runoff generation schemes, or regionally variable parameterizations. In practice, model performance may differ substantially before and after calibration, and parameter values can also vary greatly across climatic and physiographic conditions. However, the models included in this study do not appear to have a consistent calibration status: some are calibrated, some are uncalibrated, and some may be only partially calibrated. This compromises the comparability of the intercomparison and makes it difficult to attribute performance differences purely to model structure or process representation. This problem seems inevitable since the study uses existing model outputs rather than builds its own modeling frameworks. However, this problem should be emphasized in the manuscript (in many places) and its impact on the results should be thoroughly discussed.

Response: Thank you for raising this important concern. We fully agree that parameter calibration can have an influence on hydrological model performance, especially for runoff magnitude and peak flow simulations. We also agree that the calibration status of the models is not fully uniform, because our analysis relies on public multi-model outputs rather than a set of models calibrated under a fully controlled experimental framework. Therefore, the differences in model performance should not be attributed purely to model structure or process representation. To address this concern, we have revised the manuscript in three ways.

First, we have added a clear description of the calibration status of each model in the revised manuscript. Based on the available model documentation and calibration information, five models are classified as calibrated models, including PCR-GLOBWB, LPJML, HYDROPY, CWATM, and WATERGAP2-2E. The remaining eight models, for which no explicit calibration entry is available in the original model information, are grouped as uncalibrated models in this comparison, including MATSIRO, DBH, CLM40, MIROC-INTEG-LAND, VIC, ORCHIDEE-MICT,

JULES-W2, and H08. We have now added this information to the model description table and explicitly noted the calibration status in the manuscript.

Second, we added a new supplementary analysis to examine how calibration status may influence model performance. Specifically, we compared the proportions of basins satisfying the predefined performance thresholds between calibrated and uncalibrated models for Qsum, Qmax, and CTQ. The results are shown in the **Figure 1**.

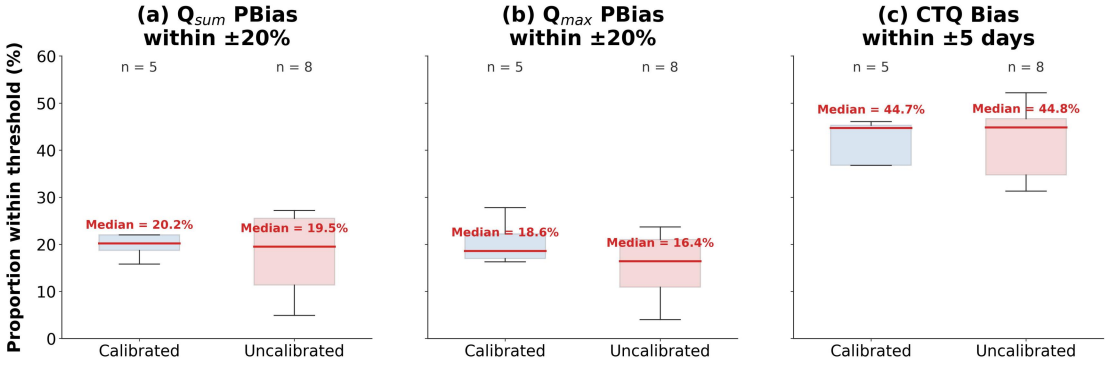

Figure 1. Influence of calibration status on model performance. *Boxplots compare the proportions of models satisfying the performance thresholds for (a) Qsum PBias within ±20%, (b) Qmax PBias within ±20%, and (c) CTQ Bias within ±5 days between calibrated and uncalibrated models.*

For Qsum, the median proportion of basins within the ±20% PBias threshold is 20.2% for calibrated models and 19.5% for uncalibrated models. For Qmax, this number drops from 18.6% to 16.4%, from calibrated to uncalibrated models. For CTQ, the median values are nearly identical between calibrated and uncalibrated models, with 44.7% and 44.8% of basins within the ±5-day threshold, respectively. These suggest that calibration status does affect model performance to some extent, particularly for Qmax, but it does not fully explain the main patterns identified in this study.

Third, we have expanded the **Discussion** to more explicitly acknowledge calibration inconsistency as an important point. We emphasize that the observed performance differences among models reflect the combined effects of model structure, process representation, calibration strategy, forcing uncertainty, and other implementation choices. In particular, we have moderated the interpretation of the differences between GHMs and LSMs. While GHMs tend to perform better for Qsum and Qmax and LSMs show relative advantages for CTQ, these differences should not be interpreted as being solely controlled by model category or process structure. **Calibration partly contributes to the stronger performance of several GHMs in runoff magnitude, whereas the similar CTQ performance between calibrated and uncalibrated models suggests that timing-related behavior is less directly improved by calibration alone and may depend more strongly on the representation of snowmelt energy-related processes.**

We have therefore revised the manuscript to present the inter-model comparison as a large-sample diagnostic evaluation, rather than a controlled attribution experiment. The new discussion clarifies that calibration inconsistency is an inherent limitation of using existing public multi-model datasets and should be considered when interpreting the results.

Corresponding manuscript lines: [Lines 90–100, 320–334, 475–478].

2. The explanation of the proposed RI in Section 2.2.4 should be strengthened. First, the manuscript does not clearly define the "complexity level/group" used in the RI calculation, including how basins were grouped along the CI gradient and how the corresponding CI values for each group were assigned. Second, the rationale for combining normalized SMAB and normalized slope into a single Euclidean-distance-based metric is not sufficiently justified. At present, it remains unclear how reliable RI is as an integrated measure of model robustness, and whether it can meaningfully balance overall bias against sensitivity to increasing basin complexity. Since RI appears to be a newly proposed metric in this study, the manuscript should provide stronger justification of its formulation, interpretation, and practical usefulness.

Response: We sincerely thank the reviewer for this comment. We agree that the definition and interpretation of the robustness index (RI) should be strengthened, because RI is central to the evaluation framework of this study. In the revised manuscript, we have substantially expanded **Section 2.2.4** to clarify three aspects: (1) how basin-complexity groups are defined, (2) why SMAB (Stratified Mean Absolute Bias) and the slope of bias against complexity index (CI) are combined to represent model robustness, and (3) why a Euclidean-distance-based formulation is used.

First, we now explicitly describe how the basin complexity levels used in RI are constructed. All basins were first ranked according to their integrated CI and then divided into ten equally-sized groups along the CI gradient. Each group therefore represents one basin-complexity level, ranging from the least complex to the most complex basins. For each group, the representative CI value was defined as the median CI of all basins within that group, and model bias was summarized using the median bias across basins in the same group. This grouping strategy reduces the influence of uneven basin distributions along the CI gradient and allows model performance to be evaluated consistently across different environmental complexity levels. To make these complexity levels more interpretable, we further added a new supplementary table (**Table 1**) summarizing the characteristics of the ten basin-complexity groups. The table reports the median values of CI, DEM, DEMstd, LAI, and PFTh for each group. The results show a clear and physically meaningful progression across the ten levels: the median CI increases from 0.73 in the very low complexity group to 1.88 in the highest complexity group; DEM increases from 371.15 m to 2351.50 m; DEMstd increases from 46.05 m to 450.74 m; and PFTh generally increases from 0.48 to 0.96. These patterns indicate that higher-complexity groups are associated with increasingly complex topographic conditions and greater vegetation-type diversity. Therefore, the CI groups are not abstract categories, but correspond to physically interpretable gradients in basin characteristics.

Table 1. Definition and characteristics of the ten basin complexity groups.

Basin complexity level	CI median	DEM median (m)	DEMstd median (m)	LAI median (-)	PFTh median (-)
very low	0.73	371.15	46.05	0.86	0.48
low	0.91	379.52	67.88	0.94	0.73
low-to-moderate	1.03	375.76	71.90	0.99	0.83
moderate-low	1.15	418.09	91.52	0.99	0.97

moderate	1.26	656.91	175.26	1.01	0.92
moderate-high	1.38	1158.79	228.02	1.05	0.79
high-moderate	1.49	1392.07	267.55	1.00	0.80
high	1.58	1765.83	311.82	0.79	0.83
very high	1.69	2049.99	348.72	0.65	0.91
highest	1.88	2351.50	450.74	0.58	0.96

We also added a new figure (**Figure 2**, [Figure 3 in the manuscript]) showing the spatial distribution of the four CI components (DEM, DEMstd, LAI, and PFTh), as well as the integrated CI. This figure provides an intuitive basis for understanding the regional differences in basin complexity. For example, northern Europe generally shows relatively low values for most complexity components, especially DEM and DEMstd, indicating comparatively simple topographic conditions. In contrast, the western United States exhibits high DEM and DEMstd values, reflecting complex mountainous terrain and strong topographic variability. Northeastern China generally lies within an intermediate range of CI, with moderate topographic and vegetation complexity. These regional differences are important for interpreting the subsequent model-performance results, because regions such as the western United States represent more challenging environments where model errors and inter-model spread may be amplified, whereas northern Europe and northeastern China provide contrasting low-to-intermediate complexity conditions for evaluating model robustness.

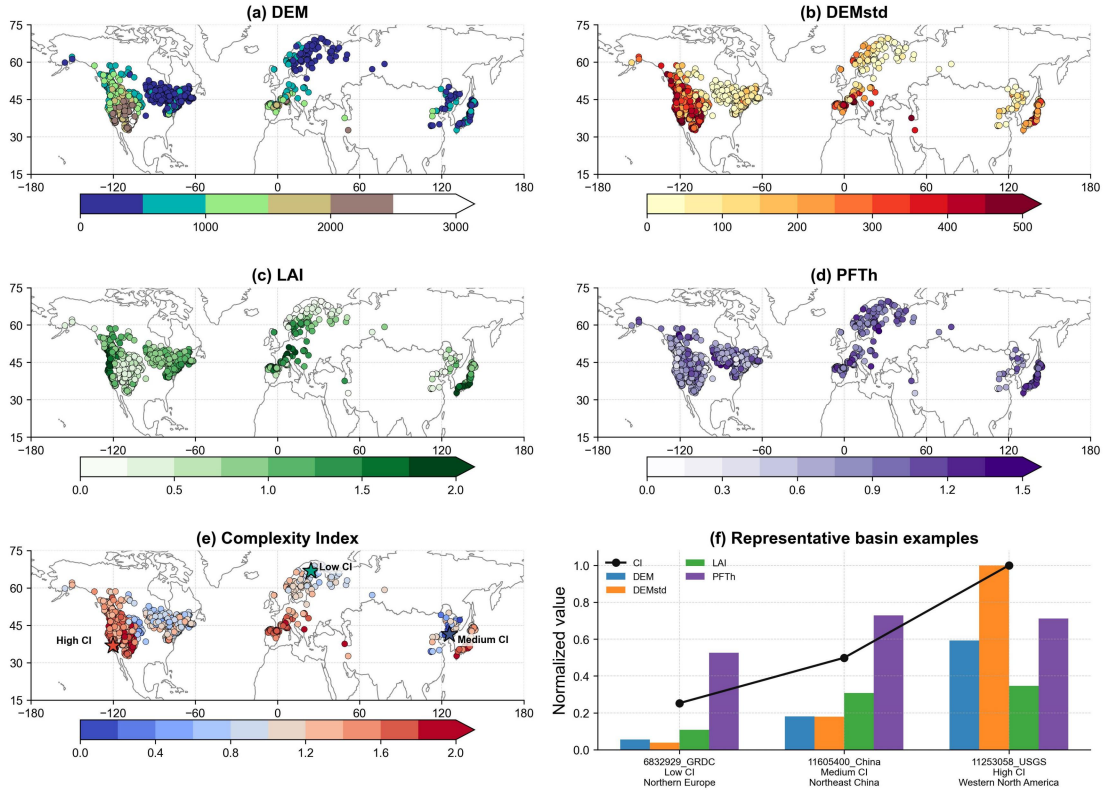

Figure 2. Spatial distribution of basin complexity components and representative examples of the integrated Complexity Index. *Panels show (a) DEM, (b) DEMstd, (c) LAI, (d) PFTh, and (e) the integrated Complexity Index across the study basins. The three starred basins in panel (e) are*

representative examples of Low CI, Medium CI, and High CI conditions. Panel (f) compares the normalized values of DEM, DEMstd, LAI, PFTh, and CI for these three representative basins.

Second, we have clarified the physical meaning of the two components of RI and added a new conceptual figure (**Figure 3**, [Figure A2 in the manuscript]) to support this explanation. In **Figure 3**, SMAB is illustrated as the overall magnitude of absolute bias across the full basin-complexity gradient. It is therefore interpreted as a measure of performance stability, because it reflects whether a model maintains a generally low bias across different complexity levels. In comparison, the regression slope of absolute bias against CI captures the rate at which model bias increases with basin complexity. We interpret this component as a measure of adaptability to complex environmental conditions, because a larger slope indicates stronger performance degradation as basin complexity increases.

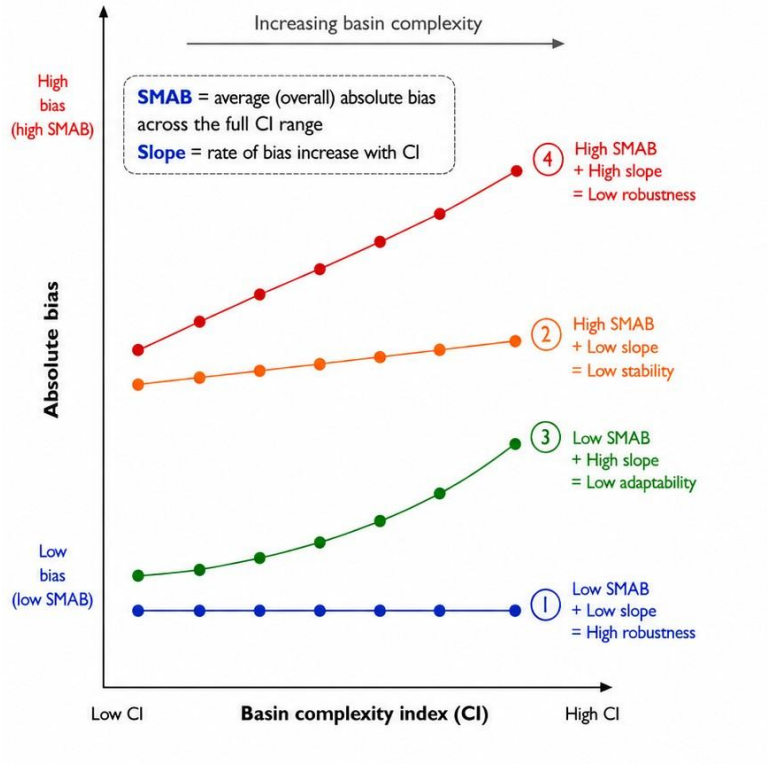

Figure 3. Conceptual illustration of the Robustness Index (RI). *The stratified mean absolute bias (SMAB) represents the overall magnitude of model bias across the full basin complexity index (CI) gradient, whereas the slope represents the rate at which model bias changes with increasing CI.*

Third, we have expanded the justification for using a Euclidean-distance-based metric. After normalization, SMAB and slope form a two-dimensional performance space in which the ideal model is located at the origin, corresponding to zero overall bias and zero degradation with increasing basin complexity. The Euclidean distance from this ideal point provides a straightforward and widely accepted way to quantify the joint deviation from the ideal condition. To further test whether the robustness ranking depends on the specific Euclidean-distance formulation, we added a sensitivity analysis in the Supplementary Information. We compared the baseline Euclidean-distance-based RI with several alternative formulations, including weighted

Euclidean indices and weighted linear aggregation indices (**Table 2**, [Table A1 in the manuscript]). These sensitivity experiments were designed to test whether the robustness ranking is sensitive to the aggregation method, the relative weighting between SMAB and slope, or a stricter criterion in which the poorer component controls the final RI.

Table 2. Sensitivity experiments used to test the robustness of the RI formulation

Experiment	Formulation	Weight setting	Purpose
S0	$RI = 1 - \sqrt{\frac{w \times SMAB_{norm}^{\ 2}}{+(1-w) \times S_{norm}^{\ 2}}}$	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	$RI = 1 - (w \times SMAB_{norm} + (1-w) \times S_{norm})$	w = 0.3	Linear aggregation; adaptability emphasized.
S2-2		w = 0.4	Linear aggregation; adaptability slightly emphasized.
S2-3		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 resulting model rankings are highly consistent with the baseline RI formulation (**Figure 4**, [Figure A3 in the manuscript]). Across the weighted Euclidean and weighted linear formulations, most models show only limited changes in both RI magnitude and relative ranking. In particular, models with high RI values and those with low RI values remain generally stable, suggesting that the identification of robust and non-robust models is not sensitive to the specific aggregation method or weighting scheme. Moderate rank changes are mainly observed for

intermediate-ranking models, especially for Qsum and Qmax, which is expected because these models may differ in the relative contributions of SMAB and slope. Importantly, the rankings for CTQ are especially stable across all sensitivity experiments. These results indicate that the main conclusions derived from RI are robust to alternative formulations and support the reliability of RI as an integrated measure of model robustness.

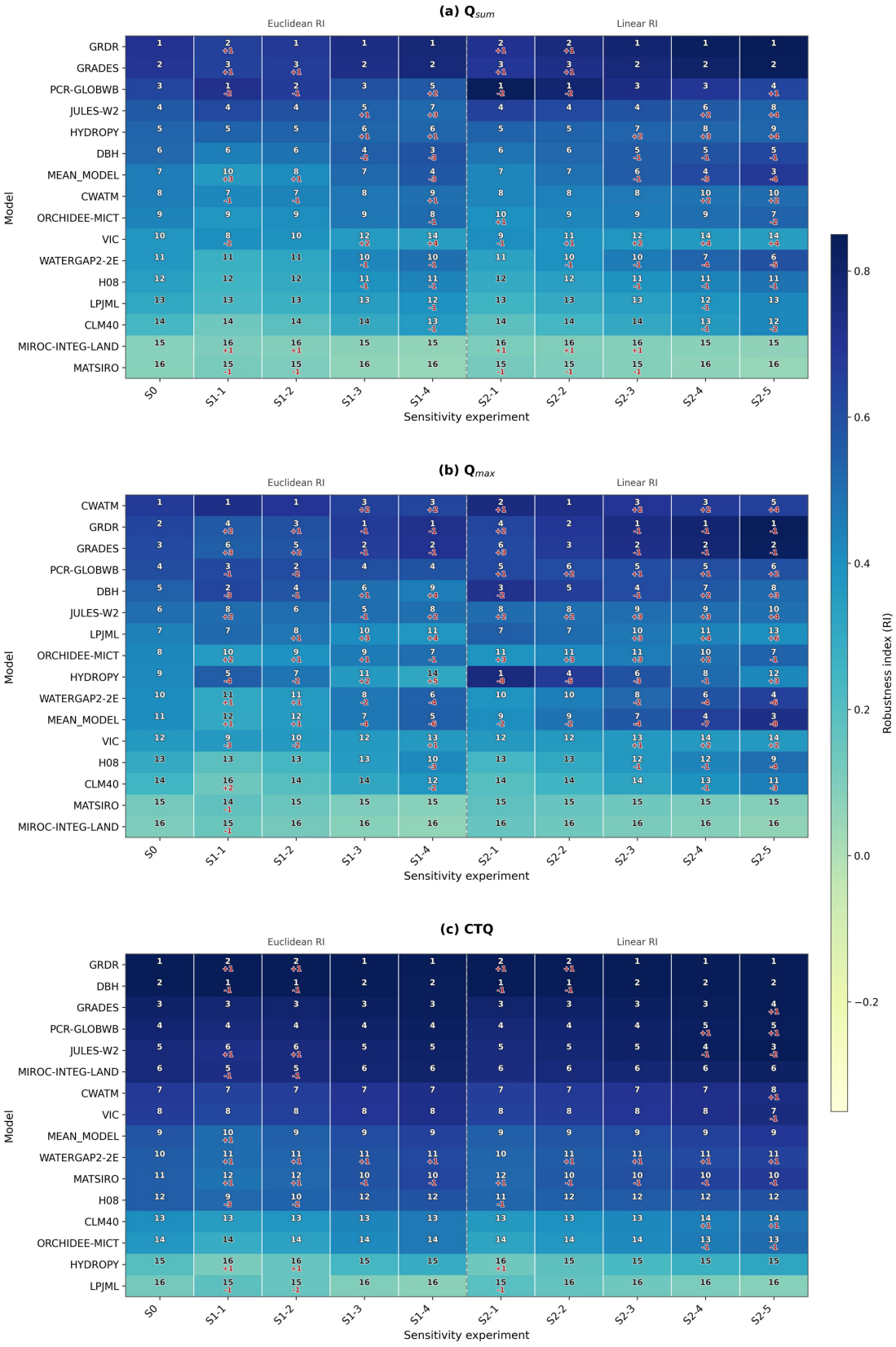

Figure 4. Sensitivity analysis of the Robustness Index (RI) formulation. *Panels (a–c) present the sensitivity of model robustness rankings for Qsum, Qmax, and CTQ, respectively. The sensitivity experiments include the baseline Euclidean-distance-based RI (S0), weighted Euclidean formulations (S1-1 to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

We have revised **Section 2.2.4** accordingly and added a supplementary table describing the physical characteristics of each CI group, as well as a supplementary sensitivity analysis comparing different RI formulations.

Corresponding manuscript lines: [Lines 210–225, 230–285].

3. I recommend that the authors moderate the novelty claims related to the evaluation framework. The volume/peak/timing analysis adopted here is useful and appropriate for process-based diagnosis, and I do not object to its use in this study. However, similar types of diagnostics have already been widely used in hydrological model evaluation, including in snow-related runoff studies. Therefore, some statements currently appear overstated. For example, the conclusion states that "This framework advances model intercomparison by moving beyond average bias evaluation," but many previous studies have gone well beyond average bias and have used multiple performance metrics and more process-oriented diagnostics. I suggest rephrasing such statements to avoid overstating methodological novelty. More generally, the manuscript would benefit from another careful round of language revision, as several statements appear stronger than the evidence supports.

Response: We agree that our original manuscript overstated the novelty of the evaluation framework in some places. We also fully acknowledge that volume-, peak-, and timing-related diagnostics have been widely used in hydrological model evaluation and snow-related runoff studies. Therefore, the novelty of this study should not be presented as the first use of such runoff characteristics or as a general move beyond traditional performance metrics.

In the revised manuscript, we have moderated the novelty claims throughout the **Abstract**, **Introduction**, and **Conclusions**. We now clarify that Qsum, Qmax, and CTQ are established and physically meaningful runoff characteristics, and that our contribution lies in applying them systematically to a large-sample, multi-model evaluation of snowmelt runoff across snow-dominated basins. Moreover, the revised framing emphasizes that this study provides a comprehensive SMR-focused diagnostic assessment by combining runoff magnitude, peak, and timing characteristics with basin-complexity gradients and the newly defined robustness analysis.

Accordingly, we have revised statements that could imply excessive methodological novelty. For example, the sentence "This framework advances model intercomparison by moving beyond average bias evaluation" has been replaced with a more moderate statement emphasizing that the framework "*This analysis complements existing model evaluation approaches by integrating established SMR characteristics with basin complexity information, thereby providing additional insights into the stability and adaptability of model performance under diverse environmental conditions.*" We have also carefully revised other strong expressions such as "novel evaluation perspective," "for the first time," and "advancing model development" to avoid overstating the contribution.

Overall, the revised manuscript now more clearly acknowledges previous work on multi-metric and process-oriented hydrological evaluation, while positioning our study as a large-sample application and extension of these established diagnostic ideas to SMR model

evaluation under varying basin complexity conditions.

Corresponding manuscript lines: [Lines 2–7, 66–70, 444–446].

4. I also suggest revising the title of Part I. At present, the emphasis on "robustness" seems somewhat overstated, as the current emphasis on robustness may not fully reflect the primary contribution of the manuscript. The paper is fundamentally a global-scale evaluation of snowmelt runoff performance across large-scale hydrological models, and the title would be more accurate if it reflected this primary focus more directly.

Response: We agree that the original title placed too much emphasis on robustness, whereas the primary contribution of Part I is a systematic global-scale evaluation of snowmelt runoff performance across large-scale hydrological models and runoff products. Although robustness remains an important component of our analysis, it is introduced as an additional perspective to examine how model performance changes along basin-complexity gradients, rather than as the sole focus of the manuscript.

Therefore, we have revised the title to better reflect the main scope and contribution of the study. The revised title emphasizes the comprehensive evaluation of snowmelt runoff performance, while leaving the robustness analysis as a key element within the manuscript.

Original title: "Process diagnostics of snowmelt runoff in global hydrological models: Part I – Model evaluation from the perspective of robustness"

Revised title: "Process diagnostics of snowmelt runoff in global hydrological and land surface models: Part I – A systematic evaluation across basins of increasing complexity"

This revised title more accurately reflects the primary focus of Part I as a large-sample model evaluation study, while maintaining consistency with Part II.

Corresponding manuscript lines: [Lines 1–3].

Other:

1. In the abstract, "little is known about their SMR performance" appears overstated. In addition, "ISIMIP3a and ISIMIP2a" are mentioned without sufficient explanation, and I suggest simplifying or removing these terms in the abstract.

Response: We have revised the abstract to use more moderate wording, emphasizing that SMR performance across diverse basin conditions remains insufficiently synthesized rather than unknown. We also agree that mentioning ISIMIP2a and ISIMIP3a in the abstract without explanation may distract from the main message. Therefore, we have simplified the abstract by referring generally to "hydrological and land surface models".

Corresponding manuscript lines: [Lines 5–15].

2. In the second paragraph of the Introduction, the statement that SMR-related simulations remain unsatisfactory should be qualified more carefully. Snow-related runoff can in some cases be easier to simulate because of its strong seasonality and regular hydrological response, whereas runoff simulation in arid regions is often more difficult.

Response: We agree that the original statement was too general. Our intention was not to suggest that SMR simulation is universally more difficult than other runoff processes, but to emphasize that accurately reproducing SMR characteristics across diverse and complex basin conditions

remains challenging. We have therefore revised the sentence to use more qualified wording and to clarify that the challenge mainly arises from coupled snow-related processes, basin heterogeneity, forcing uncertainty, and model parameterization.

Corresponding manuscript lines: [Lines 25–35].

3. Since the gridded runoff is at 0.5° resolution while the MERIT basins are much finer, the manuscript should discuss the possible impact of this scale mismatch on routing and the resulting evaluation.

Response: We agree that the scale mismatch between the 0.5° gridded runoff outputs and the finer MERIT-Basins river network may introduce uncertainties. We have added a discussion of this limitation in the revised manuscript, noting that the mismatch may affect Qmax and CTQ more strongly than Qsum because peak magnitude and timing are more sensitive to routing pathways and travel-time estimates. However, because our analysis focuses mainly on long-term mean SMR characteristics and applies the same routing framework to all models, we expect the scale mismatch to have limited influence on the relative inter-model comparison, while acknowledging that it may affect the absolute evaluation results.

Corresponding manuscript lines: [Lines 118–129].

4. The residual water-balance estimate is acceptable as a pragmatic screening proxy for identifying snowmelt-dominated basins, but it should not be interpreted as a rigorous quantification of snowmelt runoff contribution, because changes in catchment storage, delayed groundwater release, and rainfall–snowmelt interactions are not explicitly accounted for.

Response: We agree that the residual water-balance estimate should be interpreted as a pragmatic screening proxy rather than a rigorous quantification of the exact snowmelt runoff contribution. We have revised the manuscript to clarify that SMR_ratio is used only to identify basins where snowmelt signals are likely dominant during the snowmelt period, rather than to precisely partition runoff sources. We have also added a discussion of the potential influence of storage change, groundwater lag, and rainfall–snowmelt interactions on this screening procedure.

Corresponding manuscript lines: [Lines 175–185].

5. ERA5 SWE is used to define snow periods, but the manuscript should discuss the reliability and possible limitations of ERA5 SWE for this purpose, especially in complex terrain.

Response: In this study, ERA5 SWE is used mainly to define the timing of the main snowmelt period, rather than to evaluate the absolute magnitude of SWE. Nevertheless, uncertainty in ERA5 SWE may influence the identified melt start/end dates and therefore affect the calculation of Qsum, Qmax, and CTQ. We have added a discussion in the revised manuscript to clarify this limitation and to emphasize that the results should be interpreted with this uncertainty in mind, particularly for mountainous and highly heterogeneous basins.

Corresponding manuscript lines: [Lines 145–154].

6. The description of the observed discharge data source needs clarification. The manuscript states that discharge data are obtained from GSHA, but it is not clear whether GSHA directly provides the daily discharge time series used here. Please clarify this point.

Response: We thank the reviewer for pointing out this ambiguity. GSHA publicly provides annual

and monthly streamflow indices, but it does not directly release the daily discharge time series used in this study. The daily discharge data were obtained through following GSHA directed data retrieving packages, and were used here for calculating the SMR characteristics during the snowmelt period. We have revised the manuscript to clarify this point.

Corresponding manuscript lines: [Lines 164–166].

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 A2 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.

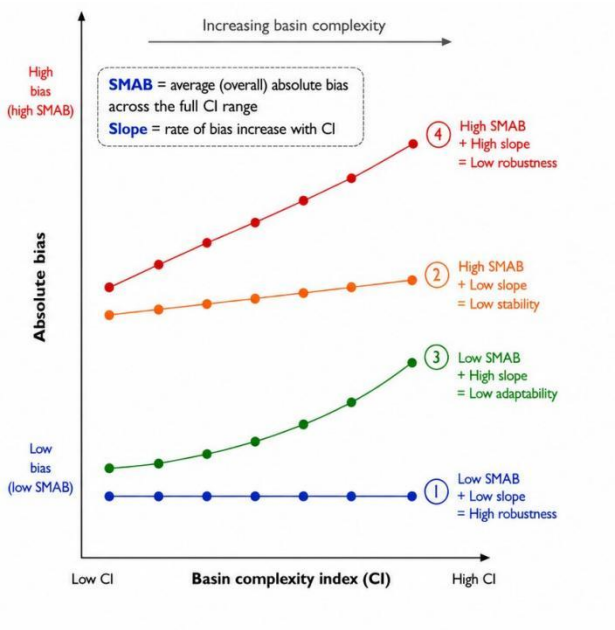

Figure 1. Conceptual illustration of the Robustness Index (RI). *The stratified mean absolute bias (SMAB) represents the overall magnitude of model bias across the full basin complexity index (CI) gradient, whereas the slope represents the rate at which model bias changes with increasing CI.*

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 A3 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.

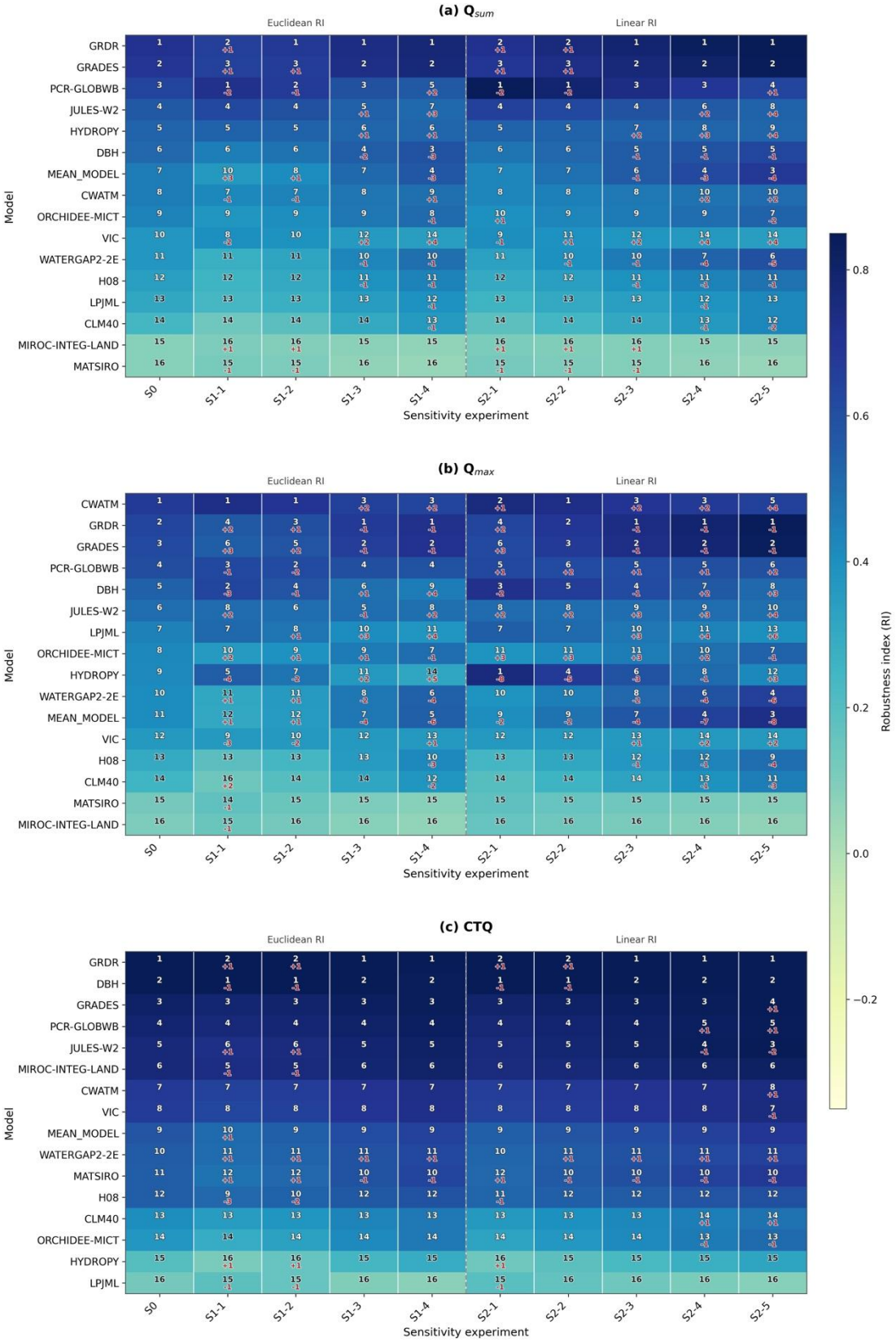

Figure 2. Sensitivity analysis of the Robustness Index (RI) formulation. *Panels (a–c) present the sensitivity of model robustness rankings for Qsum, Qmax, and CTQ, respectively. The sensitivity experiments include the baseline Euclidean-distance-based RI (S0), weighted Euclidean formulations (S1-1 to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

Corresponding manuscript lines: [Lines 230–287].

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]).

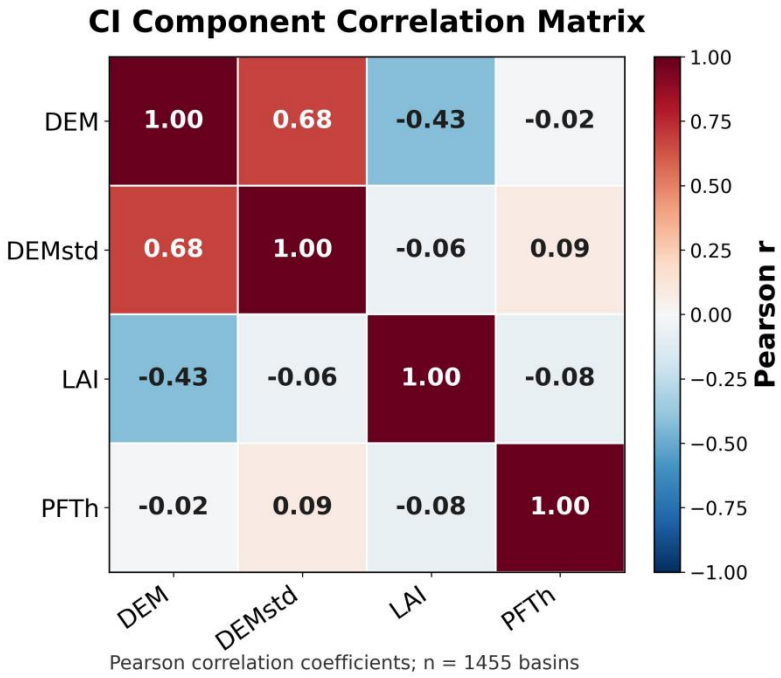

Figure 3. Pearson correlation matrix among the four components used to construct the basin complexity index (CI).

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.

Corresponding manuscript lines: [Lines 212–226].

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.

Corresponding manuscript lines: [Lines 116–128].

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**).

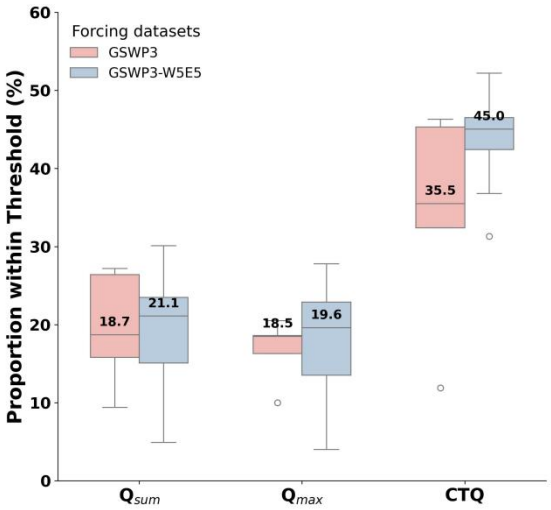

Figure 4. Effects of forcing datasets on model performance for Qsum, Qmax, and CTQ. *(a) compares the results under different forcing datasets (GSWP3 and GSWP3-W5E5). The thresholds are defined as ±20% for Qsum and Qmax, and ±5 days for CTQ. Numbers above the boxplots indicate median values.*

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 a result of both improved forcing and model-development progress, rather than forcing alone.

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

Corresponding manuscript lines: [Lines 330–334, 475–478].

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.

Corresponding manuscript lines: [Lines 315–334, 452–456].

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.

Corresponding manuscript lines: [Lines 144–147].

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.

Corresponding manuscript lines: [Lines 145–154].

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."

Corresponding manuscript lines: [Lines 9–11].

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.

Corresponding manuscript lines: [Figure 8 caption].

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