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**.

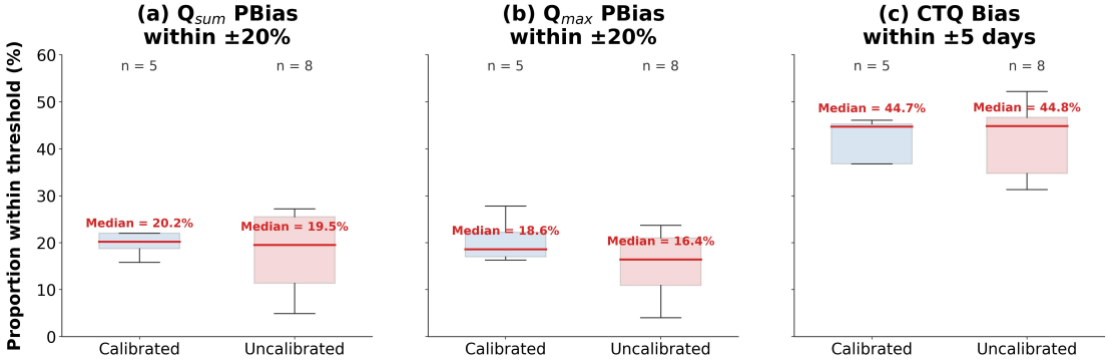

Figure 1. Influence of calibration status on model performance. *Boxplots compare the proportions of models satisfying the performance thresholds for (a) Qsum PBias within ±20%, (b) Qmax PBias within ±20%, and (c) CTQ Bias within ±5 days between calibrated and uncalibrated models.*

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 limitation. We now 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 may partly contribute 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.

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 important 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 A2 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.

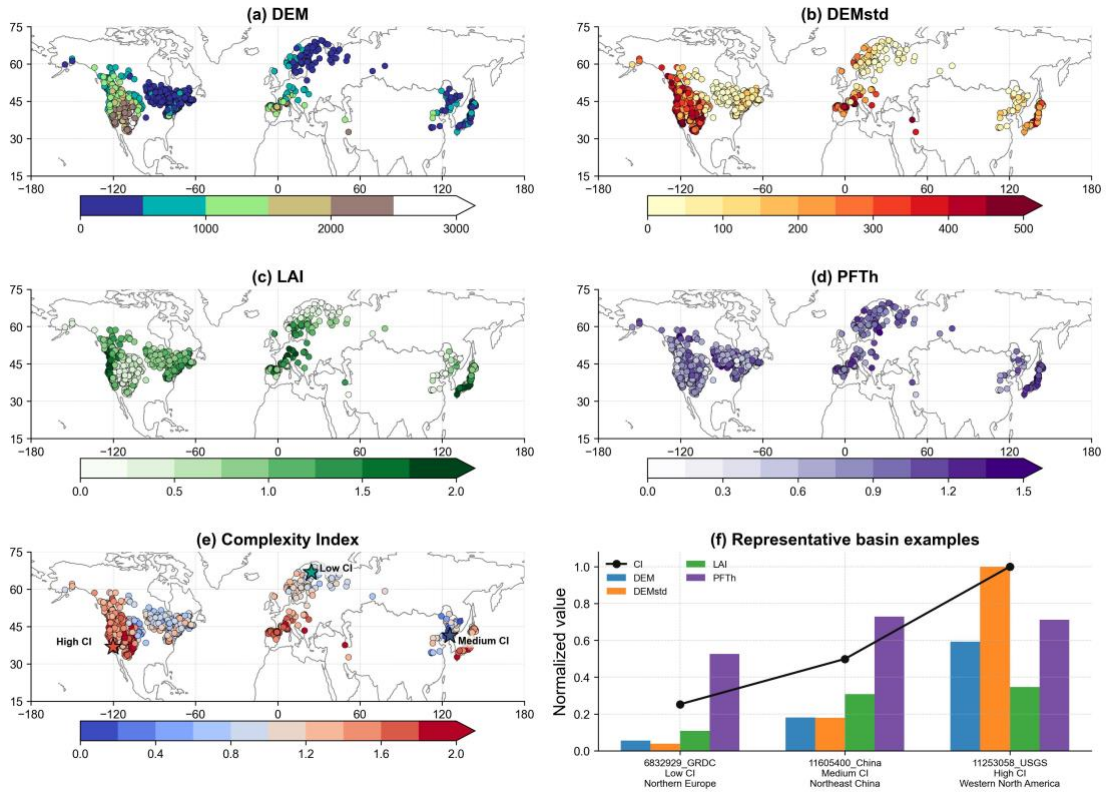

Figure 2. Spatial distribution of basin complexity components and representative examples of the integrated Complexity Index. *Panels show (a) DEM, (b) DEMstd, (c) LAI, (d) PFTh, and (e) the integrated Complexity Index across the study basins. The three starred basins in panel (e) are representative 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 A3 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.

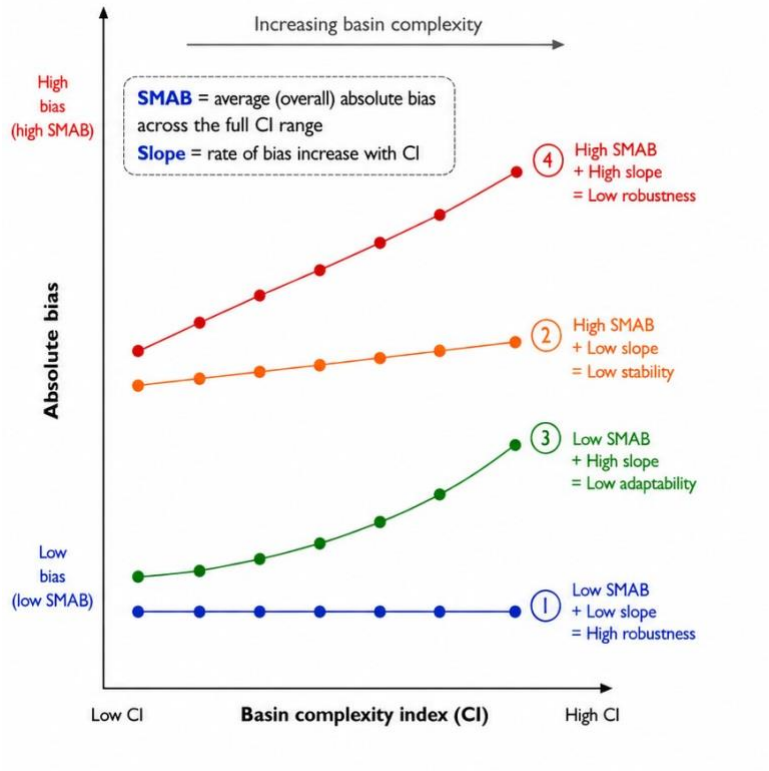

Figure 3. Conceptual illustration of the Robustness Index (RI). *The stratified mean absolute bias (SMAB) represents the overall magnitude of model bias across the full basin complexity index (CI) gradient, whereas the slope represents the rate at which model bias changes with increasing CI.*

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 commonly used 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		w = 0.5	Baseline RI formulation.
S1-1		w = 0.3	Euclidean aggregation; adaptability emphasized.
S1-2	$RI = 1 - \sqrt{w \times SMAB_{norm}{}^2 + (1-w) \times S_{norm}{}^2}$	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 resulting model rankings are highly consistent with the baseline RI formulation (Figure 4, [Figure A4 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.

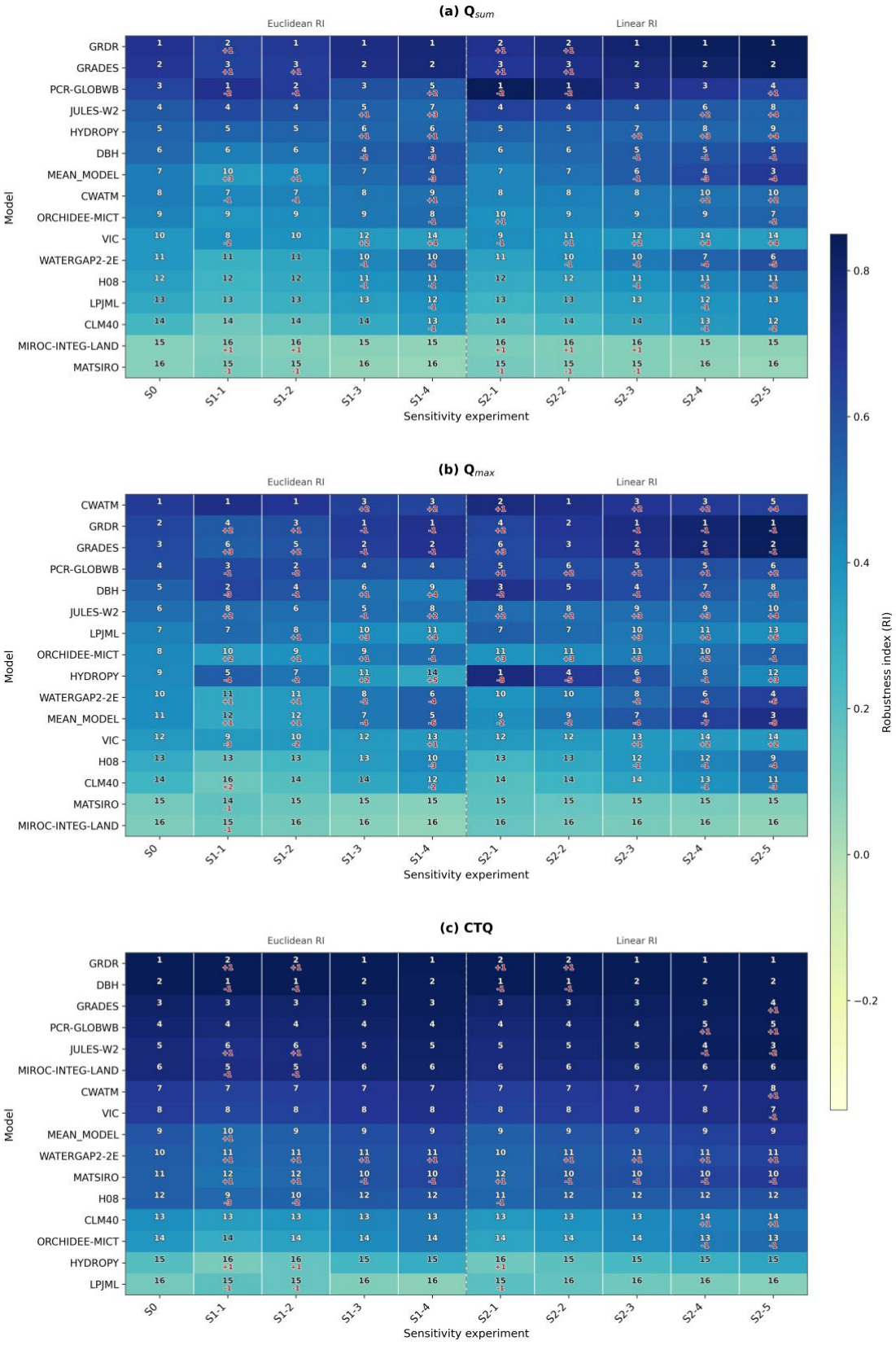

Figure 4. Sensitivity analysis of the Robustness Index (RI) formulation. *Panels (a–c) present the sensitivity of model robustness rankings for Qsum, Qmax, and CTQ, respectively. The sensitivity experiments include the baseline Euclidean-distance-based RI (S0), weighted Euclidean formulations (S1-1 to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

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 thank the reviewer for this helpful comment. 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. More specifically, 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 ample application and extension of these established diagnostic ideas to SMR model evaluation under varying basin complexity conditions.

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.

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".

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.

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.

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.

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.

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 communication with the GSHA authors and were used here for calculating the SMR characteristics during the snowmelt period. We have revised the manuscript to clarify this point.