Review2

Overall, this manuscript addresses an interesting question by linking snow-related model process complexity to SMR performance across a large sample of basins. The proposed TBMCS framework is potentially useful and may provide a valuable starting point for future discussion of process complexity. However, the manuscript still has several important limitations, particularly its close dependence on Part I, the interpretation of complexity effects under inconsistent calibration status, and the limited validation of the new complexity metric. In its current form, I am not yet fully convinced that Part II is sufficiently strong as a standalone paper. Addressing these concerns may require substantial revision and further clarification of the study scope and framing.

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

Major comments:

1. I have a generally positive view of Part I, which provides substantial large-sample evaluation results and offers a useful basis for understanding model performance during snowmelt periods. In contrast, I find Part II somewhat less convincing as a standalone paper. In its current form, Part II reads more as a natural extension of Part I than as a fully independent study. Its design, performance metrics, basin framework, and much of the interpretive context are inherited directly from Part I, while TBMCS-based complexity analysis is the main new element. I therefore encourage the authors to further clarify the independence and scientific completeness of Part II as a separate contribution. Given that Part I contains only a limited number of main-text figures, whereas Part II includes a fair amount of supporting or less essential material, one possible option would be to integrate the core complexity analysis into Part I. Such integration might improve the overall coherence of the study, especially if some of the less essential analyses in Part I are simplified.

Response: Thank you for this thoughtful and constructive criticism, and we appreciate the positive assessment of Part I. We agree that Part II is closely connected to Part I in terms of study domain, evaluation metrics, and basin framework. However, we believe that Part II addresses a scientifically distinct question. While Part I asks how well models perform in simulating snowmelt runoff, Part II asks how differences in snow-process complexity influence model performance and robustness. Therefore, Part II is not simply an extension of the evaluation in Part I, but develops a process-based explanatory framework for interpreting inter-model performance differences.

We also carefully considered the reviewer's suggestion of integrating the two studies. However, we believe that keeping them as two separate papers is justified because the TBMCS framework represents a distinct methodological contribution. Previous studies, such as Telteu et al. (2021) and Müller Schmied et al. (2025), have made important contributions by describing and graphically representing the physical process formulations of large-scale models. Building on this foundation, our study goes one step further by developing a quantitative tree-based scoring method and using it to examine how process complexity relates to model performance and robustness. This methodological development and its performance-based application form the core contribution of Part II and support its role as an independent study.

At the same time, we agree that Part II should be more self-contained. We have therefore

revised the manuscript in three ways. First, we revised the Introduction to more clearly distinguish the roles of the two papers, explicitly stating that Part I focuses on performance diagnosis, whereas Part II focuses on process complexity and its influence on model behavior. Second, following both this comment and a similar suggestion from Reviewer 1, we revised **the Data and Methods section** to briefly reintroduce the key metrics and concepts used in Part II, including Qsum, Qmax, CTQ, model performance, robustness, and basin complexity. Third, we reduced the dependence on Part I throughout the text by adding concise methodological context where needed, while retaining consistency between the two papers.

Overall, these revisions clarify that Part II is scientifically complete as a standalone contribution, while still building on the diagnostic foundation established in Part I.

2. A major concern is that the interpretation of model complexity may be strongly confounded by differences in calibration status across models. In Part I, this issue is less critical because the main goal is model evaluation itself. In Part II, however, the manuscript attempts to interpret SMR performance differences in terms of process complexity and to draw implications for model development. In this context, calibration becomes a more important limitation. If some models are calibrated, some are not, and others are only partially calibrated, the reported complexity–performance relationships may not be cleanly attributable to model complexity itself. For any model, performance can change substantially with parameter values while model complexity remains unchanged. Ideally, this type of analysis would be more convincing under calibrated conditions, which also suggests that relying on existing public runoff datasets may impose important constraints on the current study design.

Response: We agree that differences in calibration status across models are an important limitation when interpreting complexity–performance relationships. This issue is indeed more critical in Part II than in Part I, because Part II moves beyond model evaluation and attempts to interpret inter-model performance differences from the perspective of process complexity. We have therefore revised the manuscript to acknowledge this limitation more explicitly.

Specifically, we now clarify that, because this study is based on publicly available multi-model datasets from ISIMIP, the calibration status is not fully uniform across models. Some models are calibrated, some are not, and the degree of calibration may also differ. As the reviewer notes, this means that model performance cannot be attributed to model complexity alone. **We therefore do not interpret complexity as the sole determinant of performance, but rather as one important explanatory factor acting alongside calibration, forcing uncertainty, and other structural differences.**

To address this concern more thoroughly, we added a new comparison figure showing the effects of calibration status on model performance for Qsum, Qmax, and CTQ (**Figure 1**).

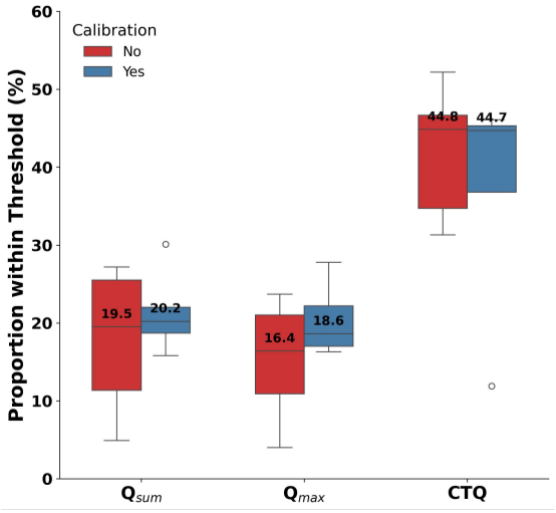

Figure 1. Effects of calibration status on model performance for Qsum, Qmax, and CTQ. *The thresholds are defined as ±20% for Qsum and Qmax, and ±5 days for CTQ. Numbers above the boxplots indicate median values.*

The results show that, overall, the effect of calibration is limited, although a somewhat clearer influence can be seen for Qmax. This suggests that calibration differences should be acknowledged as a source of uncertainty, but they do not appear to be the dominant explanation for the main complexity–performance patterns identified in this study. In contrast, our key conclusion is not based on a general complexity advantage across all metrics. Rather, the clearest complexity-related gain is found for CTQ, which is a timing-related metric. We argue that this result is less likely to be explained mainly by calibration, because CTQ depends more directly on the physical representation of melt timing and energy-related snow processes, especially under complex basin conditions.

We also note that several degree-day-based models in our ensemble are calibrated, yet this does not guarantee strong performance. For example, LPJML, despite being based on a calibrated framework, still performs relatively poorly, with only 18.7%, 16.3%, and 11.9% of basins meeting the predefined thresholds for Qsum, Qmax, and CTQ, respectively, compared with the corresponding median values across all models of 20.2%, 18.5%, and 44.7%. This further suggests that calibration alone is insufficient to overcome limitations in structurally simplified snow-process representations, especially for timing-related behavior.

Accordingly, we have revised the manuscript to (1) explicitly discuss calibration inconsistency as a study limitation, (2) add the new comparative figure and discussion, and (3) moderate our interpretation so that the complexity–performance relationship is presented as a first-order process-based diagnosis, rather than a clean causal attribution.

3. The discussion of model complexity could be further strengthened. At present, the literature review is somewhat limited and does not sufficiently engage with the broader literature on model structural complexity, structural uncertainty, and the interaction between complexity and calibration. The current framing sometimes gives the impression that the central question is whether “more complex is better,” whereas the existing literature suggests a much more nuanced picture. I

encourage the authors to broaden the review and place the present study in a more balanced methodological context.

Response: We agree that we should more sufficiently engage with the broader literature on model structural complexity, structural uncertainty, and the interaction between complexity and calibration. We have therefore expanded the Introduction to place our study in a more balanced methodological context.

Specifically, we now emphasize that the central question is not simply whether "more complex models are better," but rather when, where, and for which runoff characteristics added process complexity becomes useful. To support this revised framing, we added several references covering different aspects of the complexity debate, including studies on multi-model and multi-structure comparisons (e.g., Seiller et al., 2012; Horton et al., 2022), structural complexity and prediction uncertainty (e.g., Arkesteijn and Pande, 2013), the combined effects of model complexity and forcing uncertainty (e.g., Ludwig et al., 2009; Knoche et al., 2014), and model complexity effects in cryosphere or glacio-hydrological contexts (e.g., Muñoz et al., 2021). These additions help clarify that no single structural choice is universally optimal and that increasing complexity can also introduce structural uncertainty, parameter identifiability problems, calibration dependence, and computational cost.

We have revised the manuscript to present model complexity as a conditional and metric-dependent factor. In our results, added snow-process complexity does not universally improve model performance for all runoff characteristics. Instead, its clearest benefit appears for CTQ, especially under high basin complexity, whereas model performance for Qsum and Qmax remains less sensitive to model complexity. This revised framing is more consistent with the broader literature and avoids implying that greater complexity is always preferable.

4. Although TBMCS is an interesting idea, its validation remains somewhat limited. The manuscript does not provide enough comparison with previous or simpler ways of describing model complexity, making it difficult to assess whether TBMCS offers a clear advantage beyond being a new scoring framework. A new complexity metric would be more convincingly introduced through stricter comparisons at smaller scales, under calibrated conditions, and against existing measures. At minimum, the present limitations should be more clearly acknowledged.

Response: We thank the reviewer for this important criticism. We agree that the intended role and limitations of TBMCS should be clarified more explicitly. In the revised manuscript, we have refined the framing of TBMCS, moderated our interpretation, and added supplementary analyses to test the robustness of the main results.

We would like to clarify that TBMCS is not proposed as a fully validated or universally optimal measure of absolute model complexity. Rather, its main contribution lies in providing a process-traceable perspective for comparing model complexity across different models. Traditionally, simpler descriptions such as parameter number, process count, or qualitative model descriptions have been used. Here, TBMCS decomposes model complexity into specific snow-related process modules and represents how each process is structurally formulated through tree depth, nodes, and leaves. This allows the complexity–performance relationship to be interpreted in terms of physically meaningful process representations, rather than only through an aggregate or parameter-count-based measure.

Although stricter validation under smaller-scale, fully controlled, and uniformly calibrated conditions would be valuable, such validation is difficult at this stage because no directly comparable quantitative framework for scoring snow-process complexity across global models currently exists. Instead, we strengthened the methodological basis of TBMCS by carefully cross-checking our process interpretation and tree construction against existing model intercomparison and representation studies, particularly Telteu et al. (2021) and Müller Schmied et al. (2025), which provide detailed textual and graphical descriptions of physical process formulations in global water models. These studies offer an important reference for identifying how different models represent key hydrological and snow-related processes. Building on this foundation, TBMCS extends previous qualitative or graphical descriptions by translating process representations into a transparent and comparable scoring framework. We therefore do not present TBMCS as an absolutely validated complexity metric, but as a first-order, process-traceable diagnostic framework for comparing snow-process representations across models. To avoid overstatement, we have clarified this scope in the revised manuscript and explicitly acknowledged that future work should further test TBMCS under controlled forcing, calibration, and smaller-scale experimental settings.

At the same time, several limitations were also explicitly acknowledged. First, the present TBMCS framework focuses on four key snow-related processes and does not include all possible structural components that may influence SMR simulation, such as snow-layer design, liquid water storage, soil layers, groundwater storage, glacier storage, or routing-related schemes. Second, because the analysis relies on existing public multi-model simulations, calibration status and forcing datasets are not fully uniform across all models. Third, the weighting scheme used to combine tree depth, nodes, and leaves contains subjective choices and therefore requires sensitivity testing.

We have also revised the manuscript in four ways. First, we clarified in the Data and Methods section that TBMCS should be interpreted as a first-order, process-based diagnostic framework for snow-process complexity, rather than a complete measure of all model structural complexity. Second, we expanded the discussion of additional structural processes beyond the four selected snow-related modules and explained why they are not included in the current framework. Third, we added supplementary analyses to examine the influence of calibration status and forcing datasets, including comparisons under more controlled calibration/forcing conditions. These analyses show that calibration effects are generally limited overall, although Qmax is somewhat more affected, and that the positive relationship between model complexity and CTQ performance remains evident under fixed forcing conditions. Fourth, we added a weighting sensitivity analysis in the Supplementary Information to test whether the model-complexity ranking is robust to alternative weighting schemes.

Accordingly, we now present TBMCS as a useful and transparent diagnostic framework for linking snow-process representation to model behavior, rather than as a definitive or universally validated complexity metric. We also clearly acknowledge that future work could further test and refine the framework through controlled experiments at smaller scales, uniformly calibrated model setups, consistent forcing data, and comparison with additional existing complexity measures.

5. I also have some reservations about how rainfall–snowfall partitioning complexity is treated in TBMCS. Unlike many internal hydrological processes, precipitation phase partitioning is closely tied to forcing data and upstream processing. Some models diagnose phase internally, whereas others rely on externally provided phase information or forcing products that already contain this distinction. In the current framework, such externally handled complexity may be scored as lower simply because it is not explicitly represented in the model code, while the complexity embedded in the forcing data is not considered. For this reason, it could be better to exclude rainfall–snowfall partitioning from the complexity analysis.

Response: We thank the reviewer for this important comment. We agree that rainfall–snowfall partitioning differs from many other internal hydrological processes because it is closely linked to forcing data and upstream preprocessing. And simplifying the full complexity of the forcing-generation pipeline may generate different results.

To address this, we have revisited the scoring and found that four models in our ensemble directly use precipitation phase information from the forcing data rather than performing rainfall–snowfall partitioning internally (e.g., MATSIRO uses phase information from GSWP3, while MIROC-INTEG-LAND, ORCHIDEE-MICT, and H08 use phase information from GSWP3-W5E5). But as the reviewer pointed out, this classification does not imply that the forcing-side treatment is simple. W5E5 includes snowfall-related information over land, and GSWP3-W5E5 is generated from GSWP3 and W5E5 data. Therefore, part of the physical complexity of precipitation phase treatment may be embedded in the forcing-generation procedure, which is outside the scope of the present model-internal complexity scoring.

To test whether our main conclusions depend on this process, we added a sensitivity test in which rainfall–snowfall partitioning was excluded from the total model complexity score. The results show that the model complexity ranking remains highly stable after removing this component (Spearman rank correlation with the original ranking: $r=0.98$, $P<0.001$). More importantly, the positive relationship between model complexity and CTQ performance remains robust and even slightly increases when rainfall–snowfall partitioning is excluded ($r=0.32$ using the original TBMCS score and $r=0.33$ after excluding rainfall–snowfall partitioning). In contrast, the relationships for Qsum and Qmax remain weak in both cases.

As a result of this analysis, we have therefore determined to retain rainfall–snowfall partitioning in TBMCS because it remains a key process controlling snow accumulation and SMR generation, but does not change our main conclusions. We revised the manuscript to (1) clarify the interpretation of rainfall–snowfall partitioning as internal model-process complexity, (2) explicitly acknowledge the limitation associated with forcing-side phase preprocessing. These revisions show that the main CTQ-related conclusion is not driven by the treatment of rainfall–snowfall partitioning alone.

Other:

1. Figures 2–5 contain useful information on model structure, but some of them may be better placed in the Appendix or Supplement, since they mainly support the construction of TBMCS rather than serving as central result figures.

Response: Thank you for this helpful suggestion. We agree that Figures 2–5 contain detailed structural information and are closely related to the construction of the TBMCS framework. However, we have decided to retain these figures in the main text because they are central to

the methodological contribution of this study. Specifically, these figures show how the process complexity scores are derived for each of the four key snow-related processes—rainfall–snowfall partitioning, interception, sublimation, and melt. They therefore provide the necessary transparency for understanding and evaluating the TBMCS method, rather than serving only as supplementary supporting material. Notably, we found two EGU journal articles exclusively differentiating different models and their processes (e.g., Telteu et al. (2021) and Müller Schmied et al. (2025)). This suggests that such kind of materials are well received by the community for its contribution to understand model differences.

At the same time, we appreciate the reviewer's concern regarding readability. To address this, we have revised the corresponding text and figure captions to make the role of Figures 2–5 clearer and to emphasize that they represent the core evidence for the process-complexity assessment. We believe that keeping these figures in the main text helps readers follow how model complexity is quantified and how the subsequent complexity–performance analysis is built.

References:

Arkesteijn, L., & Pande, S. (2013). On hydrological model complexity, its geometrical interpretations and prediction uncertainty: uncertainty and ill-posedness in hydrology. Water Resources Research, 49(10), 7048–7063. https://doi.org/10.1002/wrcr.20529

Horton, P., Schaefli, B., & Kauzlaric, M. (2022). Why do we have so many different hydrological models? A review based on the case of Switzerland. WIREs Water, 9(1). https://doi.org/10.1002/wat2.1574

Knoche, M., Fischer, C., Pohl, E., Krause, P., & Merz, R. (2014). Combined uncertainty of hydrological model complexity and satellite-based forcing data evaluated in two data-scarce semi-arid catchments in Ethiopia. Journal of Hydrology, 519, 2049–2066. https://doi.org/10.1016/j.jhydrol.2014.10.003

Ludwig, R., May, I., Turcotte, R., Vescovi, L., Braun, M., Cyr, J.-F., et al. (2009). The role of hydrological model complexity and uncertainty in climate change impact assessment. Advances in Geosciences, 21, 63–71. https://doi.org/10.5194/adgeo-21-63-2009

Müller Schmied, H., Gosling, S. N., Garnsworthy, M., Müller, L., Telteu, C.-E., Ahmed, A. K., et al. (2025). Graphical representation of global water models. Geoscientific Model Development, 18(8), 2409–2425. https://doi.org/10.5194/gmd-18-2409-2025

Muñoz, R., Huggel, C., Drenkhan, F., Vis, M., & Viviroli, D. (2021). Comparing model complexity for glacio-hydrological simulation in the data-scarce peruvian andes. Journal of Hydrology: Regional Studies, 37, 100932. https://doi.org/10.1016/j.ejrh.2021.100932

Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrology and Earth System Sciences, 16(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012

Telteu, C.-E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., et al. (2021). Understanding each other's models: an introduction and a standard representation of 16 global

water models to support intercomparison, improvement, and communication. Geoscientific Model Development, 14(6), 3843–3878. https://doi.org/10.5194/gmd-14-3843-2021