Summary and general assessment

In their manuscript, the authors study how the complexity of snow processes in global hydrological and land-surface models influences their representation of snow melt runoff. They introduce a new tree-based complexity scoring method and apply this to snow process formulations of 13 global hydrological and land-surface models. The authors then compare the performance of these models for snowmelt runoff over a large set of catchments. They show that (1) there is diversity in how models represent snow processes, (2) model complexity correlates well with performance for the centroid timing of snowmelt but not with the total snowmelt and the snowmelt peak, and that (3) model complexity cannot explain performance alone. They conclude that a balanced representation of snow processes is necessary for good model performance for snow melt runoff.

While I think their study could be of value to the hydrological community, I think that major revisions are necessary to improve the clarity of the text and the interpretation of the results. In particular, methodological concepts need to be better introduced, some methodological choices surrounding the complexity scoring method require additional clarification and the results require better discussion in light of the existing literature.

Please find my detailed comments below.

Response: Thank you for providing very useful comments for us to improve our manuscript. Below, please find our responses to address your concerns.

Major

1. While I understand that this is an accompanying paper to Part 1, information that is necessary to understand this paper on its own is not properly introduced or only very late in the text. For example, the term "Robustness" is only introduced around line 272 in the Results and the reader is referred to Part 1 for details, but significant parts of the subsequent analysis depend on this concept (e.g. Figure 9). The robustness index should be introduced in the methodological section. Similarly, the term "model performance" is now introduced in the results in L219, but should be defined in the methodology (where there are already references to this). Lastly, the metrics Qsum, Qmax, and CTQ should be better introduced.

Response: We agree that the original submission relied too heavily on Part 1 for some important definitions, which hindered its readability as a standalone paper. To address this, we have restructured **Section 2** by adding two dedicated subsections:

- Introduction of Qsum, Qmax, and CTQ in the **Section 2.2**

  We added a new subsection, "Key runoff characteristics during the snowmelt period", in which Qsum, Qmax, and CTQ are explicitly defined and their hydrological meanings are clarified. We also explain why these three metrics jointly provide a diagnostic framework for evaluating snowmelt runoff behavior.

- Clearer introduction of the main snow accumulation and melt processes

  We further revised a dedicated subsection, "The main snow accumulation and melt processes", to clarify how the selected physical processes are linked to different runoff characteristics. This revision strengthens the physical basis of the later complexity analysis.

- Definition of model performance and robustness in the **Section 2.4**

  We also added a new subsection, "Model Performance and Robustness Metrics", in which both

terms are now explicitly defined before the **Results** section. Model performance is now defined as the proportion of basins with acceptable bias, using thresholds of ±20% for Qsum and Qmax, and ±5 days for CTQ. In addition, the robustness index is now introduced in the **Model Performance and Robustness Metrics** section rather than only appearing later in the **Results**. We briefly explain its two components—stability and adaptability—and clarify that robustness reflects a model's ability to maintain accuracy under increasing basin complexity. Although the detailed formulation remains consistent with that established in Part 1, sufficient methodological description is now included here to ensure that the present manuscript can be understood independently.

Overall, these revisions substantially improve the logical flow of the manuscript and reduce the need for readers to refer back to Part 1 in order to understand the main methodological framework.

2. Terminology should be precise and consistent. For example, the authors use "total complexity score" and "model complexity" and "complexity", which can be confused with basin complexity; "Qsum" is sometimes used instead of "model performance for Qsum"; etc.
See other examples below.

Response: We agree that the terminology in the original manuscript was not always sufficiently precise or consistent, which may cause confusion among several related but distinct concepts, particularly for model complexity, process complexity, and basin complexity, as well as between the runoff characteristics themselves (e.g., Qsum, Qmax, and CTQ) and the corresponding model performance. To address this issue, we have carefully revised the terminology throughout the manuscript and adopted a more consistent terminology framework, as follows:

- Model complexity

  We now use "model complexity" consistently to refer to the overall complexity value of an individual model, quantified in this study as the sum of the complexity values of the four key physical processes assessed in the Tree-Based Model Complexity Scoring (TBMCS) framework. In other words, this term corresponds to what was previously referred to in some places as the "total complexity score".

- Process complexity

  When referring to the complexity of an individual physical process, such as rainfall–snowfall partitioning, interception, sublimation, or melt, we now consistently use the term "process complexity".

- Basin complexity

  We reserve "basin complexity" exclusively for the environmental heterogeneity of basins, quantified by the composite complexity index (CI).

- Model performance for runoff characteristics

  We have revised the text throughout to distinguish the runoff characteristics themselves from the corresponding model performance for Qsum, Qmax, and CTQ, where appropriate.

3. L100. What remains unclear from the text, is what the authors base their assessment of model complexity on. Do they base the assessment on previous model intercomparison papers (e.g. Telteu et al., 2021; Müller-Schmied et al., 2025), on discussions with experts, on the model code, or on the supporting publications of the models? Please elaborate.

Response: We agree that we should more explicitly clarify that our assessment of model complexity

is not based on a single source, but on a combination of multiple lines of evidence. Specifically, the scoring framework was developed with reference to previous model intercomparison and model description studies, such as Telteu et al. (2021) and Müller Schmied et al. (2025), which provide valuable syntheses of how large-scale hydrological models represent key processes. Additionally, we conducted our own process-level assessment for each model based on three main sources:

(1) the official model documentation,
(2) the available model code or technical implementation details, and
(3) the corresponding model description papers and supporting publications.

These sources were used jointly to identify how each model represents the four key snow-related processes considered in this study. To make this clearer, we have revised the manuscript to explicitly state the basis of the complexity assessment in the **Section 2.5**.

4. Table 1. A major component missing in the complexity analysis is the representation of the snow storage itself. Different models have different numbers of snow layers, and some have a liquid snow storage in the snow and others don't (e.g. Telteu et al., 2021). Furthermore, models can have sub-grid routines to differentiate between different elevation zones within the cells which can thus melt at different times (e.g. CWatM). Such complexity - in particular elevation zones and liquid water storage where meltwater is temporarily stored - might cause a delay in the release of snowmelt runoff and thus might strongly influence the metrics considered in this study.
The authors need to explain why they did not consider such model complexity or include it in additional analyses.

Response: We thank the reviewer for this important and insightful comment. We agree that additional structural features, such as snow layers, may influence snowmelt runoff simulation, particularly runoff volume and timing. We have now clarified this point in the revised manuscript and further explained the scope of the present complexity framework.

Our study is designed to focus specifically on the key physical processes that directly control the generation and release of snowmelt runoff (SMR), rather than on all possible state variables, model structural configurations, or routing-related schemes. Accordingly, the TBMCS framework evaluates four core process modules: rainfall–snowfall partitioning, interception, sublimation, and snowmelt. These were selected because they form the main cascading process chain governing snow accumulation, ablation, and runoff generation, and because they can be identified and compared consistently across most evaluated models.

In contrast, components such as snow storage itself represent an integrated outcome of multiple upstream processes, rather than an independent first-order process module. For example, snow storage reflects the combined effects of snowfall input, sublimation loss, melt, and possibly refreezing, as also discussed by Telteu et al. (2021). Similarly, structural configurations such as the number of snow layers were not included because they represent model structural design rather than an independent physical process. Sub-grid elevation routines were also not included because they are implemented only in a subset of models and are represented in highly model-specific ways, which limits cross-model comparability within a unified scoring framework.

We nevertheless agree that these features can affect runoff characteristics. In particular, liquid water retention in the snowpack may accelerate meltwater release, and sub-grid elevation routines may alter the timing of melt across heterogeneous terrain, thereby influencing both runoff volume and timing. However, we did not include these components in our scoring system for two reasons.

First, the number of snow layers represents a model structural configuration rather than an independent physical process. Although differences in snow layer structure may influence runoff simulation and vary across models, these differences are primarily reflected in the number of layers itself (**Table 1**), which does not directly indicate how the associated physical effects differ among models in a consistent and comparable way. As a result, snow layer number is difficult to incorporate into a unified process-based scoring framework.

Second, liquid water storage representation within the snowpack is substantially more heterogeneous across models and less comparable within a unified scoring framework. As summarized in Telteu et al. (2021), only a limited subset of models explicitly distinguishes separate snow water storage compartments for frozen and liquid water. For example, among the broader set of models reviewed by Telteu et al. (2021), only four models (CLM4.5, CLM5.0, MPI-HM, and VIC) include two snow storage compartments for frozen water and liquid water content. Within the model ensemble considered in our study, this distinction is represented only in HYDROPY, CLM40, and VIC (**Table 1**). **This limited representation within our model sample further reduces cross-model comparability and makes it difficult to establish a consistent scoring rule applicable to all models.** Moreover, the influence of liquid water storage is, to some extent, already reflected in the four selected process modules. In particular, the estimation of liquid water content in snow storage depends in part on rainfall reaching the snowpack after phase partitioning, meaning that the rainfall–snowfall partitioning process considered in our framework already captures part of the relevant physical control.

Table 1. Comparison of snow layer structures and liquid water storage across models.

Model	Snow layer	Liquid snow storage
HYDROPY	1	Yes
LPJML	1	No
H08	1	No
PCR-GLOBWB	1	No
CWATM	7	No
WATERGAP2-2E	1	No
MATSIRO	3	No
MIROC-INTEG-LAND	3	No
VIC	2	Yes
ORCHIDEE-MICT	3	No
DBH	1	No
JULES-W2	1	No
CLM40	5	Yes

For these reasons, we chose to prioritize core process representations that are directly related to the three target runoff characteristics (e.g., Qsum, Qmax, and CTQ), can be systematically identified across most evaluated models, and can be evaluated within a transparent and comparable framework. We now clarify in the revised manuscript that the

present TBMCS framework should be understood as a first-order process-based complexity diagnosis rather than an exhaustive description of all snow-related structural differences among models (**Section 2.3**). We have also added discussion acknowledging the potential influence of snow storage structure and sub-grid routines, and noting that incorporating these features would be a valuable direction for future work (**Discussion**).

5. Throughout the results: Results should be discussed more thoroughly and placed in the context of the wider literature. For example, performance differences between the different models are now all attributed to model complexity. The authors should reflect more on the possibility that other factors play a role, such as differences in calibration between the models, the complexity from other processes (complexity in the representation of soil layers, groundwater or glaciers), or uncertainty in the forcing data. Also differences between the presented findings and previous studies should be better discussed. For example, what explains differences between the findings here (e.g. that CTQ performance correlates with model complexity) and other studies, such as Beck et al., 2017 who found that uncalibrated "simpler" GHMs seem to outperform more elaborate LSMs in snow-dominated basins; and Merz et al., 2022 who show that complexity does not always lead to better performance?

Response: We thank the reviewer for this constructive comment. We agree that the original manuscript placed too much emphasis on model complexity alone and did not sufficiently discuss other factors that may also influence model performance. In response, we have revised the manuscript in three main ways: (1) we added new analyses to examine the effects of calibration status and forcing datasets; (2) we expanded the discussion of additional model structural differences, such as soil, groundwater, and glacier representations; and (3) we strengthened the comparison with previous studies, including Beck et al. (2017) and Merz et al. (2022), to place our results in a broader and more balanced literature context.

(1) we have added a Supplementary Information to examine how these two factors affect model performance across the three runoff characteristics (**Figure 1**). The results show that the influence of calibration on the complexity–performance relationship is relatively limited overall. Although calibration appears to exert some effect on runoff magnitude metrics, especially Qmax, it does not systematically alter the main pattern identified in this study. In contrast, differences in forcing datasets exert a clearer influence on model performance, with the strongest impact observed for CTQ, indicating that timing-related metrics are particularly sensitive to meteorological input uncertainty (**Figure 1b**).

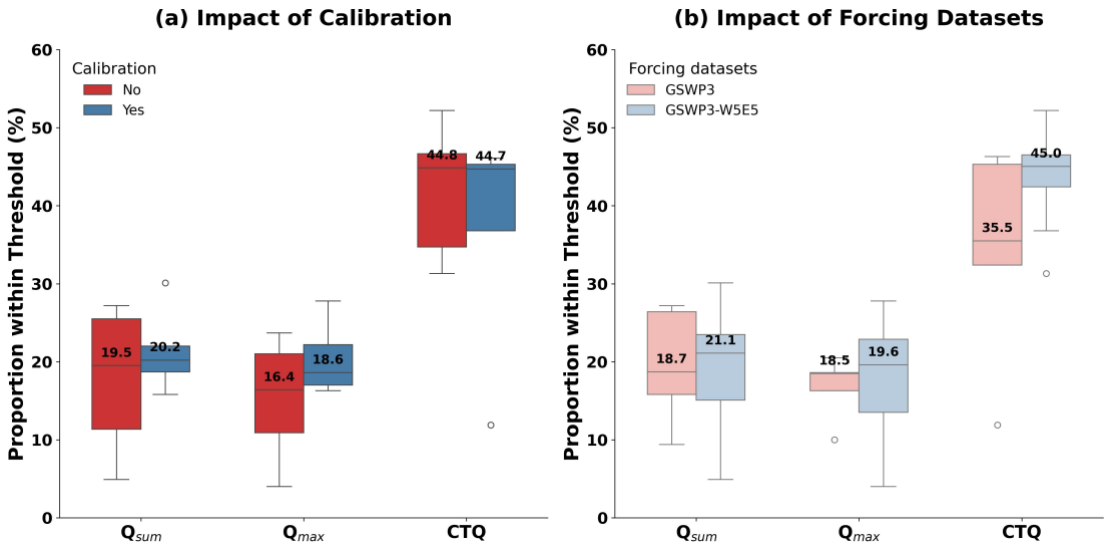

Figure 1. Effects of calibration status and forcing datasets on model performance for Qsum, Qmax, and CTQ. *(a) compares the proportion of basins within the predefined performance thresholds between calibrated and uncalibrated models, and (b) compares the results under different forcing datasets (GSWP3 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.*

Importantly, however, the forcing effect does not eliminate the complexity–performance relationship identified for CTQ regarding the role of model complexity. To test this more directly, we further examined the relationship between model complexity and model performance under controlled forcing conditions (**Table 2**).

Table 2. Pearson correlation coefficients for the relationship between model complexity and model performance for CTQ under controlled forcing conditions.

Control_variable	Control_group	Metric	Pearson_r
Forcing datasets	GSWP3	CTQ	0.44
Forcing datasets	GSWP3-W5E5	CTQ	0.82*
Calibration	Yes	CTQ	0.61
Calibration	No	CTQ	0.24

*Notes: Correlations are calculated separately for models driven by GSWP3 and GSWP3-W5E5. Model performance is quantified as the proportion of basins within the predefined thresholds (±5 days for CTQ). * denotes statistical significance at the $P < 0.05$ level.*

The results show that, even under the same forcing dataset, the positive relationship between model complexity and CTQ performance remains evident; under GSWP3-W5E5 forcing, this correlation is statistically significant ($P < 0.05$). This additional analysis suggests that although forcing uncertainty is an important co-controlling factor, especially for CTQ, it does not fully explain away the complexity–performance relationship identified in this study (i.e., for CTQ, model complexity is significantly associated with performance). We therefore revised the manuscript to

adopt a more balanced interpretation: model complexity is not the sole determinant of model performance, but its positive influence on CTQ remains detectable even after accounting for forcing conditions.

(2) we have expanded the discussion to note that process representations beyond the four snow-related modules considered here may also affect the results. In particular, differences in soil layers, groundwater storage, and glacier storage may influence runoff volume and timing, especially in basins where such processes play an important role. To make this point more explicit, we added a new comparison table in the revised manuscript (**Table 3**), which summarizes differences among models in their representations of soil, groundwater, and glacier storage.

Table 3. Comparison of soil layers, groundwater storage, and glacier storage across models.

Model	Soil layer	Groundwater storage	Glacier storage
HYDROPY	1	YES (1 layer)	No
LPJML	6	No	No
H08	1	YES (2 layers)	No
PCR-GLOBWB	2	YES (1 layer)	No
CWATM	3	YES (1 layer)	No
WATERGAP2-2E	1	YES (1 layer)	No
MATSIRO	13	YES (1 layer)	No
MIROC-INTEG-LAND	13	YES (1 layer)	No
VIC	3	No	No
ORCHIDEE-MICT	11	No	No
DBH	3	No	No
JULES-W2	4	No	No
CLM40	15	YES (1 layer)	YES (10 layers)

As shown in **Table 3**, the evaluated models exhibit substantial structural differences in these additional hydrological components. For example, the number of soil layers ranges from 1 to 15 across models, groundwater storage is represented in some models but absent in others, and glacier storage is explicitly represented only in a limited subset of models. These differences may affect the partitioning, buffering, and release of water, and therefore contribute to inter-model differences in Qsum, Qmax, and CTQ.

At the same time, we clarify why the present study prioritizes process representations over more general structural descriptors such as the number of layers. Our objective is to diagnose how the physical processes directly controlling SMR formation and release affect model performance. The four selected process modules—rainfall–snowfall partitioning, interception, sublimation, and melt—directly regulate snow accumulation, mass loss, ablation, and runoff generation, and therefore have clear physical links to Qsum, Qmax, and CTQ. By contrast, structural descriptors such as the number of soil layers, snow layers, or groundwater layers mainly describe how a model discretizes or organizes internal storage. Although such configurations can influence runoff simulation, their physical effects depend strongly on the associated governing equations,

parameterizations, and coupling schemes. For example, simply counting the number of layers does not indicate how water is transferred, stored, released, or coupled with snowmelt processes. Therefore, these structural descriptors are less directly comparable across models and less suitable as the primary basis for a unified process-based scoring framework.

We do not argue that these structural features are unimportant. Rather, we now clarify that the present TBMCS framework should be interpreted as a first-order diagnosis of snow-process complexity, rather than a complete explanation of all inter-model performance differences. To reflect this more clearly, we have revised Section 2.3 to better define the scope of TBMCS, and we have expanded the Results and Discussion sections to acknowledge that additional structural differences, including soil, groundwater, glacier, snow-layer, and routing representations, may also influence model performance. These features should be considered in future extensions of the framework.

(3) we have strengthened the comparison with previous studies, including Beck et al. (2017) and Merz et al. (2022). We now clarify that our results do not suggest that more complex models are always better. In fact, our findings are consistent with the broader literature in showing that complexity does not necessarily improve all runoff metrics: model performance for Qsum and Qmax shows limited sensitivity to model complexity, whereas the positive relationship is mainly found for CTQ, especially in basins with high basin complexity. This result is broadly consistent with Merz et al. (2022), who emphasized that added complexity does not automatically translate into better performance. Regarding the apparent difference from Beck et al. (2017), we now discuss that the two studies emphasize different aspects of model behavior. Beck et al. (2017) focused primarily on large-scale runoff performance and showed that simpler, uncalibrated GHMs can outperform more elaborate LSMs in some snow-dominated basins. In comparison, our study isolates specific snowmelt-runoff characteristics and finds that the advantage of greater model complexity emerges mainly for CTQ, a timing-related metric that is more directly linked to snow energy-balance processes and less easily compensated by runoff calibration. We therefore interpret our findings not as contradicting Beck et al. (2017), but as suggesting that the value of added complexity depends on which runoff characteristic is evaluated and under what basin conditions.

Overall, we have revised the Data and Methods, Results, and Discussion sections to make three points clearer. First, model complexity is not the only factor controlling model performance; calibration, forcing uncertainty, and other structural differences also matter. Second, TBMCS focuses on snow-process complexity because these processes are more directly linked to SMR formation and release than general structural descriptors such as layer number. Third, the main complexity-related finding is metric dependent: added process complexity does not consistently improve Qsum or Qmax, but its positive influence remains evident for CTQ, especially under complex basin conditions.

6. Finally, I would like to point out that improvements might also depend on potential changes made in the accompanying Part 1 paper.

Response: We thank the reviewer for this helpful comment. We agree that improvements to Part 2 should be coordinated with revisions to the accompanying Part 1 paper, given the close connections between the two manuscripts. Accordingly, we have revised Part 2 to be more self-contained and have also made corresponding improvements in Part 1 where needed to ensure consistency and coherence between the two papers.

Minor

Title: should it not be “global hydrological and land-surface models”?
Response: We have revised the title to “Process diagnostics of snowmelt runoff in global hydrological and land surface models: Part II - Are more complex models better?” to better reflect the model types considered in this study.

L57: Explain what centroid timing is.
Response: We have revised the text to explicitly define centroid timing (CTQ) as the timing at which cumulative runoff reaches 50% of the total runoff during the snowmelt period, thereby representing the temporal center of runoff release.

L85: While I understand that this is a paper accompanying Part 1, I would still advise to elaborate more on the set-up to make it more likely that complexity is the main explanation for the results, e.g. that model resolutions, forcing, and routing are largely held constant.
Response: We thank the reviewer for this helpful comment. We have revised the manuscript to clarify the study setup more explicitly, including that the selected models were harmonized as much as possible in terms of spatial resolution, forcing framework, and routing treatment, so that the analysis can focus more directly on differences in process representation and model complexity.

L88: “The key runoff characteristics considered here are Qsum, Qmax, and CTQ, which are crucial for water resource utilization, flood hazard prevention, and water resource management, respectively.” I would better introduce Qsum, Qmax and CTQ here and discuss how they are defined. Furthermore, references are required when stating that these are crucial for water resources and flood hazards.
Response: We have revised the manuscript to introduce Qsum, Qmax, and CTQ more clearly and then explicitly defining the three runoff characteristics based on the snowmelt period. We have also clarified their hydrological meanings and added supporting references for their relevance to water resources, flood-related applications, and runoff timing diagnostics.

L93: Canopy radiative transfer and surface albedo are later not treated as separate processes and do not have their own process tree. I would clarify which processes are the focus of the analysis and so I would not discuss these here.
Response: We have revised the text to clarify that the analysis focuses on four key process modules, and that canopy radiative transfer and surface albedo are discussed as important components within the broader snowmelt energy-balance process, rather than as separate process trees.

L116: Highlight that, after weights are assigned, everything is summed to results in a complexity score for each process.
Response: We have revised the manuscript to state more explicitly that, after weights are assigned to tree depth, number of nodes, and number of leaves, these weighted components are summed to derive the process complexity value for each process, which is then used to calculate the overall model complexity.

L118: Provide some more elaboration on why the weights were chosen this way. I do think it could influence the ranking of the models.

Response: We thank the reviewer for this important suggestion. We have revised the manuscript to better explain the rationale for assigning a larger weight to tree depth, followed by number of nodes and number of leaves. In our framework, tree depth is intended to emphasize the hierarchical structure of process representation, whereas nodes and leaves reflect the breadth of process formulation and the number of associated inputs and parameters. We therefore adopted the baseline weights of 0.6, 0.3, and 0.1 for depth, nodes, and leaves, respectively, to prioritize hierarchical process structure while still accounting for process breadth.

We also agree that the choice of weights may influence the inferred model ranking. To address this concern, we added a sensitivity analysis in the Supplementary Information (**Table 4**) using a set of alternative weighting schemes. In the main sensitivity analysis (Schemes S1–S7), we tested several alternatives while preserving the theoretical assumption that depth should receive the largest weight. In the extended stress-test analysis (Schemes S8–S12), we relaxed this assumption to examine whether the ranking remained stable under less constrained weighting choices.

Table 4. Alternative weighting schemes used in the sensitivity analysis of the TBMCS.

Scheme	Depth weight	Nodes weight	Leaves weight	Design explanation
S1	0.60	0.30	0.10	Baseline scheme
S2	0.50	0.30	0.20	Greater leaf contribution
S3	0.70	0.20	0.10	Stronger depth emphasis
S4	0.55	0.25	0.20	Moderate depth dominance
S5	0.65	0.25	0.10	Depth-dominant, close to baseline
S6	0.50	0.35	0.15	Greater node contribution
S7	0.55	0.30	0.15	Balanced depth-dominant scheme
S8	0.40	0.40	0.20	Depth and nodes equally weighted
S9	0.40	0.30	0.30	Increased leaf contribution
S10	0.33	0.33	0.33	Equal-weight scheme
S11	0.30	0.50	0.20	Node-dominant scheme
S12	0.30	0.35	0.35	Depth no longer dominant

Across all tested schemes, the model-complexity ranking remained highly stable: only DBH and JULES-W2 exchanged positions between rank 2 and rank 3, whereas the rankings of all other models remained unchanged relative to the baseline scheme (**Figure 2**, [Lines XX–XX]).

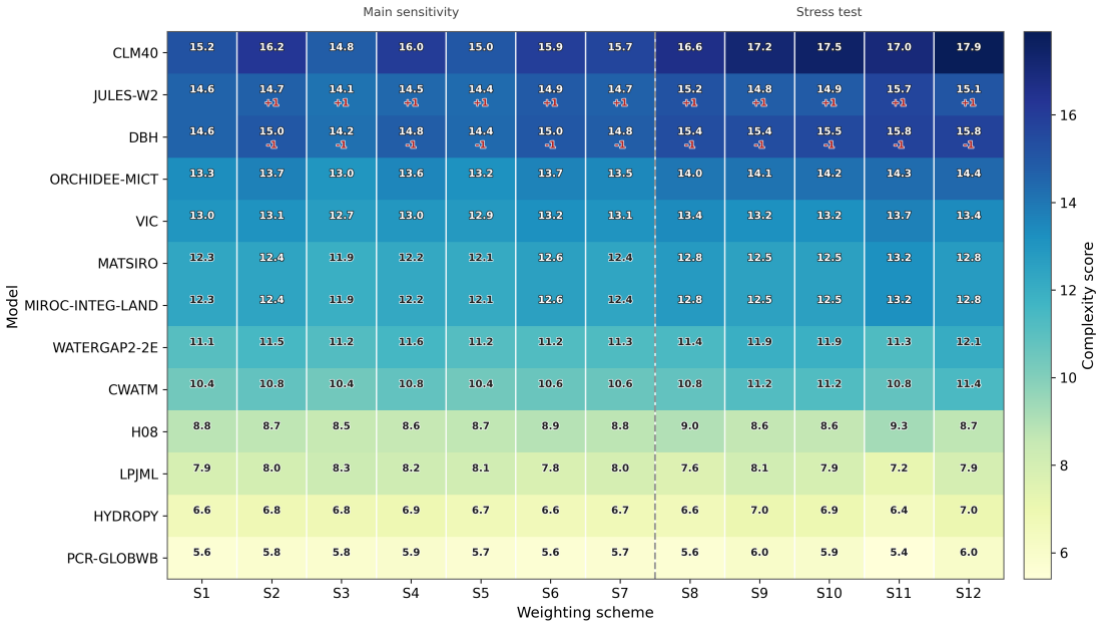

Figure 2. Sensitivity of model complexity scores to alternative weighting schemes. *Cell values indicate the complexity score of each model for a given weighting scheme. Red annotations indicate the rank change relative to the baseline scheme S1, where negative values denote an improved rank and positive values denote a lower rank. Schemes S1-S7 represent the main sensitivity experiments, whereas S8-S12 denote the stress-test schemes.*

These results indicate that the inferred ranking is robust to the weighting choice and that our main conclusions are not sensitive to the specific baseline weights adopted in this study.

L122: Highlight better in Section 2.2 what the four key processes are (see also comment on L93).
Response: We appreciate this suggestion. We have revised **Section 2.3** to explicitly list and highlight the four key processes used in this study (rainfall-snowfall partitioning, interception, sublimation, and melt).

L134: What is plant functional type entropy?
Response: In the revised text, entropy of plant functional types (PFTh) is described as an index of vegetation-type diversity within a basin, with higher values indicating greater heterogeneity in plant functional types and therefore more complex canopy–snow and energy-exchange conditions. We also briefly clarified the meanings of the other basin-complexity factors (DEM, DEMstd, and LAI) and their relevance to SMR simulation.

L150: I am wondering if taking the precipitation phase of the forcing data makes a model's process representation necessarily less complex. The phase-partitioning in the forcing data is likely very similar as the way it is handled in the model and might even be more complex than just applying a constant temperature threshold. So does it actually make sense to add this to model complexity? Please elaborate.
Response: We thank the reviewer for this comment. We agree that using precipitation phase directly from the forcing data does not necessarily mean that the overall physical treatment of phase

partitioning is simpler, because the forcing-side preprocessing may itself involve nontrivial methods. We have therefore revised the manuscript to clarify that the TBMCS framework evaluates model-internal processes complexity, rather than the total complexity of all upstream preprocessing steps.

From this perspective, directly using externally provided precipitation phase is treated as lower internal model complexity, because the phase-partitioning calculation is not explicitly represented in the model code. This is the basis on which we included rainfall–snowfall partitioning in the complexity framework. At the same time, we now explicitly acknowledge this as a limitation: forcing-side preprocessing may itself contain substantial complexity, which is not captured by the present model-structural scoring framework.

To make this clearer, we also added information on the treatment of precipitation in the forcing datasets. The ISIMIP protocols describe precipitation forcing primarily through total precipitation input, while the W5E5 forcing used in GSWP3-W5E5 includes both total precipitation (pr) and snowfall-related information (prsn) over land, with pr defined as the sum of rainfall and snowfall and prsn inherited from WFDE5-based forcing construction. GSWP3-W5E5 itself is generated by bias-adjusting and backward extending GSWP3 toward W5E5, so some models can indeed rely on forcing-side phase information rather than performing an internal rain–snow partitioning calculation.

We have therefore revised the text to emphasize that our scoring is intended to represent the complexity of model-internal process representation, not the full complexity of the forcing-generation pipeline.

L174: The authors write that almost all models explicitly represent sublimation. However, Telteu et al., 2021 and Müller-Schmied et al., 2025 classify PCR-GLOBWB as not representing snow sublimation. Please elaborate.

Response: We thank the reviewer for pointing this out. We agree that our original description of PCR-GLOBWB was not sufficiently precise. In the previous version, we interpreted the statement "*The resulting vertical fluxes for each land cover type are interception evaporation, bare soil evaporation, snow sublimation, and vegetation-specific transpiration." (Sutanudjaja et al., 2018)* as implying an explicit snow-sublimation representation. After revisiting the model description and relevant references, we agree that this process refers to evaporation from liquid water storage rather than to an explicit representation of snow sublimation. Therefore, PCR-GLOBWB should not be classified as explicitly representing snow sublimation, consistent with Telteu et al. (2021) and Müller Schmied et al. (2025).

We have accordingly revised the manuscript to correct the description of the sublimation module for PCR-GLOBWB and updated the corresponding process complexity assessment. We also revised the related table/text where the sublimation representations across models are summarized. Based on our reassessment, this correction does not materially alter the main conclusions of the study, because the overall complexity performance relationships, especially the key findings regarding the metric dependent role of model complexity, remain unchanged.

L204: Here individual complexity scores (for each process) are summed to form the total complexity score. Later on in the text, this is always referred to as model complexity. Make sure that the terminology of complexity score vs total complexity score vs model complexity vs complexity is clear and consistent (particularly because there is also basin complexity). See also comments below.

Response: As also noted in our response to the earlier comment on terminology consistency, we

have revised the manuscript to use the terminology more clearly and consistently throughout. Specifically, we now use process complexity for the complexity value of each individual process, and model complexity for the summed complexity value across the four key processes. Ambiguous expressions such as "total complexity score" or simply "complexity" have been revised where necessary to avoid confusion with basin complexity. We have also updated the relevant text around this section to make the aggregation from process complexity to model complexity more explicit.

L218: Performance needs to be introduced before in the methodology section and the text should be consistent with the terms performance vs model performance.

Response: We have revised the manuscript to introduce model performance earlier in the methodology section, specifically in **Section 2.4**.

L 224: "Figure 7 shows that, overall, the correlation between model complexity and performance is stronger under high basin complexity." This is too strong of a statement, as this is only true for CTQ and the differences are not meaningful for Qsum(-0.04-0.14) and Qmax (-0.12 to -0.16).

Response: We agree that the original statement overstated the generality of the result. In the revised manuscript, we now clarify that the correlation between model complexity and model performance becomes more evident under high basin complexity mainly for CTQ, while the corresponding differences for Qsum and Qmax are weak. We have therefore revised the text to reflect this more metric specific interpretation.

L231 "Qmax depends more on reproducing input peaks and runoff routing, yet most models remain simple in key processes such as rainfall–snowfall partitioning and interception, so added complexity offers limited benefit." I would argue that this maximum also very much depends on the rate at which the snow melts (fast melt can lead to high peaks) and thus the melt process complexity. Why would added complexity for melt parameterization not lead to better performance at Qmax? Furthermore, if the authors attribute their findings to the representation of individual snow processes, the authors should also test the correlation of model performance vs the complexity scores for individual processes (e.g. rainfall-snowfall partitioning only, interception only, snowmelt only) and see if this leads to more significant results.

Response: We thank the reviewer for this important and insightful comment. We agree that Qmax is not controlled only by input peaks and routing, but can also be influenced by snowmelt rate, such that faster melt may contribute to higher runoff peaks. We have revised the manuscript to moderate the original wording and clarify the mechanism more carefully.

In the revised text, we now explain that, at the spatial and temporal scales considered here, Qmax reflects the combined effects of precipitation inputs, snowmelt rates, and runoff concentration/routing processes. Although melt process complexity can in principle affect Qmax, its influence appears less clearly expressed than for CTQ in our large-sample, long-term mean analysis. This is likely because peak magnitude is more strongly affected by multiple interacting controls, including precipitation event intensity, rain-on-snow conditions, basin response time, and routing-related concentration effects, which can partly obscure the effect of added melt process complexity.

Following the reviewer's suggestion, we added an additional analysis examining the relationships between individual process complexity and model performance for different runoff

characteristics. Specifically, we tested the correlations between model performance and the complexity scores of rainfall–snowfall partitioning, interception, sublimation, and melt separately (**Figure 3**).

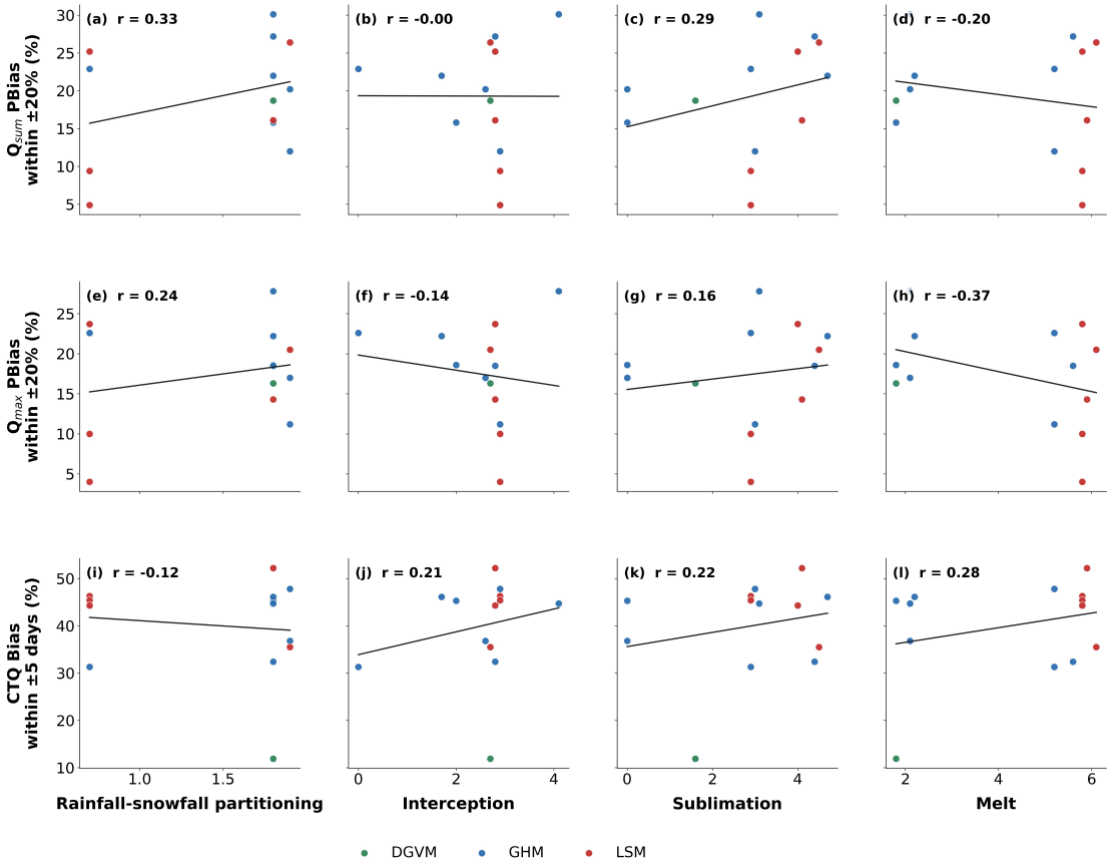

Figure 3. Relationships between individual process complexity scores and model performance metrics across the 13 models. *Columns show the complexity scores for rainfall-snowfall partitioning, interception, sublimation, and melt, respectively, while rows show the fractions of basins satisfying the performance criteria for Qsum, Qmax, and CTQ. Each panel includes a linear fit together with the Pearson correlation coefficient (r). Colors indicate different model classes.*

The results show that these relationships are indeed metric dependent. For Qmax, melt complexity shows a weak negative tendency (r=-0.37), suggesting that greater melt process complexity alone does not necessarily improve Qmax performance. In contrast, CTQ shows more consistently positive associations with process complexity, especially for melt (r=0.28), and also for interception (r=0.21) and sublimation (r=0.22). These results support our interpretation that runoff timing is more directly influenced by snow-process representation than runoff peak magnitude in the present framework. We have incorporated this additional analysis and the corresponding discussion into the revised manuscript.

L234 “Simple degree-day schemes often misrepresent melt timing, whereas more complex energy-based formulations capture the on set and magnitude of snowmelt more accurately, ...” Does this not depend on the calibration of the degree-day schemes? For example, other studies suggest that temperature index approaches and energy balance can reach similar results (e.g. Magnusson et al., 2015) and that uncalibrated simple GHMs can outperform uncalibrated complex LHMs in snow-

dominated basins (Beck et al., 2017).

Response: We thank the reviewer for this important comment. We agree that the performance difference between degree-day and energy-balance formulations does not solely depend on model structural complexity, but may also depend on calibration, model setup, and the evaluated runoff characteristic. We therefore revised the manuscript to moderate the original wording and avoid implying that energy-balance formulations are universally superior.

In particular, we now acknowledge that previous studies have shown that temperature-index/degree-day approaches and energy-balance formulations can achieve similar performance under certain conditions (e.g., Magnusson et al., 2015), and that simpler uncalibrated GHMs may outperform more complex uncalibrated LSMs in snow-dominated basins (e.g., Beck et al., 2017). These studies highlight that added process complexity does not automatically translate into better model performance.

However, our results do not suggest that calibration is sufficient to overcome the limitations of simpler melt formulations. For example, although LPJML is a calibrated degree-day-based model, its performance remains relatively poor: only 18.7%, 16.3%, and 11.9% of basins meet the predefined performance thresholds for Qsum, Qmax, and CTQ, respectively (defined as ±20% for Qsum and Qmax, and ±5 days for CTQ). In contrast, the corresponding median values across all models are 20.2%, 18.5%, and 44.7%. This comparison suggests that calibration can only partly offset the limitations of structurally simple melt formulations and is far from sufficient to guarantee strong model performance, particularly for CTQ.

Therefore, our interpretation is not that energy-balance schemes are always more accurate, but rather that, in the present large-sample framework, model complexity shows a clearer advantage for CTQ than for Qsum or Qmax, likely because timing-related process is more directly constrained by melt representation and is less easily improved through parameter adjustment alone. Accordingly, we have revised the corresponding discussion to use more cautious and literature-consistent wording.

L241: “Error compensation further explains these differences. For Qsum and Qmax, process-level errors can be offset through calibration, masking the potential advantages of higher complexity. By comparison, CTQ, as a normalized timing metric, is less amenable to such compensation because it reflects the full temporal distribution of flows.” I do not fully agree with this statement and I wonder if it is not the opposite: Qsum and Qmax depend directly on the amount of water, which is fixed (e.g. forcing input). In contrast, CTQ does not depend on the actual amount of water, but depends on the timing. Timing can be calibrated (e.g. DDF is in m/day, which is a rate). Please explain.

Response: We agree that our original wording did not sufficiently distinguish between the different ways in which calibration may affect runoff magnitude and runoff timing. We have therefore revised the manuscript to clarify this point more carefully.

We agree that Qsum and Qmax are constrained by the overall water balance and thus depend directly on the amount of water entering the system. However, the internal partitioning of that water remains uncertain. **In particular, intermediate processes such as rainfall–snowfall partitioning, interception, sublimation, and runoff generation can introduce substantial uncertainty even under the same forcing input.** In practice, these process-level errors may be partly compensated through the calibration of runoff coefficients or runoff-generation parameters, allowing the final simulated runoff magnitude to approach observations even when the internal snow-related process

representation is simplified. In this sense, the effect of higher process complexity on Qsum and Qmax may be partly masked by parameter compensation (**Figure 1a**).

By contrast, we agree that CTQ is not immune to calibration. For example, the degree-day factor (DDF) can influence melt timing because it directly controls the melt rate. We now acknowledge this more explicitly in the revised manuscript. However, most degree-day formulations used in large-scale models rely on fixed or only weakly varying DDF values. Such calibrated rate parameters may adjust the average timing of melt, but they are less able to dynamically represent the nonlinear changes in melt processes under complex environmental conditions. For example, snow–albedo feedback can accelerate melt as snow darkens, rainfall can enhance melt through additional energy input and water transfer, and forest canopy shading can delay melt by modifying radiation reaching the snow surface. These mechanisms vary across space and time and are difficult to reproduce through a single static melt-rate parameter. Therefore, although CTQ can in principle be calibrated, we argue that deficiencies in melt-process representation are generally harder to compensate using simple DDF adjustment alone, particularly in high-elevation or otherwise complex basins without additional basin-specific calibration.

We have revised the corresponding discussion to reflect this more balanced interpretation. Our intention is not to claim that CTQ cannot be influenced by calibration, but rather that, in the present framework, timing-related behavior appears to depend more directly on physically realistic melt-process representation than runoff magnitude metrics such as Qsum and Qmax.

L248-L250: These sentences are not precise and should be rephrased. Do the authors mean the following? “Model complexity exhibits ... negative correlations with model performance for Qsum and Qmax..”, “whereas the relationship between model performance for CTQ and model complexity is more strongly influenced by basin complexity factors, ...”

Response: We thank the reviewer for this helpful comment. We now clarify that model complexity shows weak or even negative correlations with model performance for Qsum and Qmax, with little sensitivity to basin-complexity factors, whereas the relationship between model performance for CTQ and model complexity is more strongly modulated by basin complexity, showing a clearer positive relationship and more evident threshold behavior.

L250: “stronger under higher heterogeneity”: What is meant by heterogeneity? And the correlations are not always stronger with higher quantiles but instead become weaker again with higher values. Why is it not “strongest under intermediate heterogeneity”?

Response: In the revised manuscript, we have replaced the ambiguous term “heterogeneity” with “basin complexity”, consistent with the terminology defined earlier in the **Section 2.4**. We also agree that the original phrase “stronger under higher basin complexity” was too general, because the correlations do not increase monotonically toward the highest quantiles. Instead, the relationship is better described as becoming more evident under intermediate-to-high basin complexity, with the strongest correlations often occurring before the highest quantile range, and in some cases weakening again at the most extreme values. We have therefore revised the corresponding text to reflect this more precise interpretation and added explanation that, under the highest basin-complexity conditions, other sources of uncertainty may increasingly obscure the effect of model complexity.

L256: “However, beyond this threshold...begin to dominate..” This statement appears very certain, but it is not tested and is a hypothesis. The statement should be supported by references or should be weakened it with words such as “likely” or “we expect”?

Response: We agree that the original wording was too certain, while this part of the discussion is intended as a hypothesis-based interpretation rather than a directly tested result. We have therefore revised the manuscript to use more cautious language and clarify that, beyond the identified threshold, the reduced benefit of additional complexity is likely related to the increasing influence of other uncertainty sources, such as forcing errors, routing effects, and parameter uncertainty, rather than stating this as a confirmed mechanism.

L278: “significant”: I would not consider a correlation with P<0.1 to be significant

Response: We agree that a correlation with $P<0.1$ should not be described as statistically significant. We have therefore revised the manuscript to use more cautious wording for the correlation between model complexity and model robustness for CTQ.

L267 “high-elevation regions receive strong surface radiation...accelaration of snowmelt” A reference should be added for this statement. Furthermore, this could also be related to the low temperatures here (which might not rise above the melt threshold of temperature index models) and the issue of snow towers (Freudiger et al., 2017).

Response: We agree that the original statement required both clearer support and a more balanced mechanism explanation. In the revised manuscript, we have therefore expanded this discussion and added the relevant references.

Specifically, we now clarify that the stronger complexity effect in high-elevation basins is not attributed to surface radiation alone. Instead, several interacting mechanisms may contribute. First, recent work has shown that radiative energy absorption can be substantial in high-elevation regions (Ma et al., 2023). Second, Ban et al. (2023) showed that runoff changes associated with the snow–albedo effect are strongest in medium-to-high elevation basins (around 2,000–3,000 m). Third, Seibert et al. (2021) pointed out that in mountainous regions with strong variability in slope, aspect, and land cover, temperature-index approaches are less able to represent spatial heterogeneity. In addition, snow redistribution processes, including snow towers and related alpine snow-transport effects, can further alter the timing and magnitude of snowmelt runoff (Freudiger et al., 2017).

Based on these considerations, we revised the manuscript to emphasize that the advantage of more complex process representation in such basins likely arises from its ability to better capture the combined effects of radiative forcing, snow–albedo feedback, terrain-related heterogeneity, and snow redistribution on snowmelt timing, rather than from a single mechanism alone.

L287: What is resilience? The term is not used before.

Response: We have revised the manuscript to define this concept explicitly in Section 2.4 before it is used later in the Results/Discussion, so that the terminology is introduced consistently and does not appear without prior explanation.

L287: HH, HL, LH, LL need to be defined in the text. Now they are only defined in the figure caption of Figure 9.

Response: We thank the reviewer for pointing this out. We have revised the manuscript to define

HH, HL, LH, and LL explicitly in the main text, rather than only in the figure caption.

L290 “This suggests that robust performance...designed process representation”. The distribution of complexity in the HH plots and HL plots look the same to me: I would not say that one has a more balanced distribution than the other. I think that this statement is not supported by this observation.
Response: In the revised manuscript, we now use a more cautious interpretation, emphasizing that high model robustness cannot be explained by overall model complexity alone and likely also depends on the specific design and interaction of individual process representations.

L302 ““CWATM enhances snowmelt simulation by introducing additional parameters”. What additional parameters?
Response: In the revised manuscript, we now clarify that, compared with other degree-day-based models using a fixed degree-day factor, CWATM introduces a more flexible formulation in which the degree-day factor is further adjusted according to rainfall intensity and the timing/seasonal progression of snowmelt. This allows the melt formulation to account for additional controls on melt variability and therefore provides a more physically comprehensive representation than a simple fixed-factor scheme. We have revised the corresponding text accordingly.

L331: Do you mean “model performance for Qsum and Qmax”?
Response: We have revised the text to explicitly use model performance for Qsum and Qmax.

L334: “CTQ exhibits a strong positive effect” Do you mean “performance for CTQ”? And a correlation with what?
Response: We have revised the text to explicitly refer to the correlation between model complexity and model robustness for CTQ, and removed the overly strong wording.

Figure 8: Caption is not completely clear to me.
“...from the 10th to 100th percentile..” Percentile of the basin complexity factor?
“...values shows..” What values, e.g. the median value in this decile?
Response: We thank the reviewer for pointing this out. We have revised the figure caption to clarify that the x-axis represents deciles of each basin complexity factor, and that the values in parentheses denote the median value of that factor within each decile.

Fig 10: What is the “ideal model”: Do you mean the most complex model? Ideal suggests the best performing model, so maybe reconsider the terminology.
Response: We have revised the terminology to avoid using “ideal model” and now refer to it more precisely as the reference model with maximum process-complexity scores.

Technical points:

Figure 6,7, 9: The color coding is confusing to me. Both the high complexity/high performance quadrant and the low complexity/low performance quadrant are colored red, which are opposite concepts. Same applies to the HL and LH, which are both green. Please reconsider the color design of these plots.

Response: We thank the reviewer for this helpful suggestion. We agree that the original color design may cause confusion because opposite quadrants shared the same color. We have therefore revised the color scheme in Figures 6, 7, and 9 so that each quadrant is more clearly distinguishable and the visual interpretation is more intuitive.

Figure 10: Often the legend is overlapping with the figures, making it difficult to read.
Response: We have revised Figure 10 by adjusting the legend position and layout so that it no longer overlaps with the figure content and is easier to read.

L306. Do you mean Figure 9c?
Response: We have corrected this to Figure 9c.

L307: I would add: "These models illustrate a gradient of structural complexity and robustness".
Response: We have added this description to improve the clarity of the text.

L356 Remove the title Appendix A
Response: We have removed the title "Appendix A" accordingly.

References:

Ban, Z., Xin, C., Fang, Y., Ma, X., Li, D., & Lettenmaier, D. P. (2023). Snowmelt-Radiation Feedback Impact on Western U.S. Streamflow. Geophysical Research Letters, 50(23), e2023GL105118. https://doi.org/10.1029/2023GL105118

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., & Schellekens, J. (2017). Global evaluation of runoff from 10 state-of-the-art hydrological models. Hydrology and Earth System Sciences, 21(6), 2881–2903. https://doi.org/10.5194/hess-21-2881-2017

Freudiger, D., Kohn, I., Seibert, J., Stahl, K., & Weiler, M. (2017). Snow redistribution for the hydrological modeling of alpine catchments. WIREs Water, 4(5), e1232. https://doi.org/10.1002/wat2.1232

Ma, Y., Yao, T., Zhong, L., Wang, B., Xu, X., Hu, Z., et al. (2023). Comprehensive study of energy and water exchange over the Tibetan Plateau: A review and perspective: From GAME/Tibet and CAMP/Tibet to TORP, TPEORP, and TPEITORP. Earth-Science Reviews, 237, 104312. https://doi.org/10.1016/j.earscirev.2023.104312

Magnusson, J., Wever, N., Essery, R., Helbig, N., Winstral, A., & Jonas, T. (2015). Evaluating snow models with varying process representations for hydrological applications. Water Resources Research, 51(4), 2707–2723. https://doi.org/10.1002/2014WR016498

Merz, R., Miniussi, A., Basso, S., Petersen, K.-J., & Tarasova, L. (2022). More Complex is Not Necessarily Better in Large-Scale Hydrological Modeling: A Model Complexity Experiment across the Contiguous United States. Bulletin of the American Meteorological Society, 103(8), E1947–E1967. https://doi.org/10.1175/BAMS-D-21-0284.1

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

Seibert, J., Jenicek, M., Huss, M., Ewen, T., & Viviroli, D. (2021). Snow and ice in the hydrosphere. In Snow and ice-related hazards, risks, and disasters (pp. 93–135). Elsevier. https://doi.org/10.1016/B978-0-12-817129-5.00010-X

Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., et al. (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. Geoscientific Model Development, 11(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

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