https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

Process diagnostics of snowmelt runoff in global hydrological models: Part II - Are more complex models better?

Xiangyong Lei[1], Haomei Lin[1], Kaihao Zheng[1], and Peirong Lin[1,*]

[1]Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing, 100871, China

Correspondence: Peirong Lin (peironglinlin@pku.edu.cn)

Abstract. The added value of increased process complexity has long been a central yet unresolved question in hydrological modeling, particularly for snowmelt runoff (SMR), where multiple physical processes interact in complex ways. To address this, we develop a Tree-Based Model Complexity Scoring (TBMCS) method to systematically quantify the complexity of snow-related processes across 13 global hydrological and land surface models. Then by using SMR characteristics, i.e., total runoff (Q_{sum}), peak discharge (Q_{max}), and centroid timing (CTQ), as integrated indicators to evaluate these models, we systematically quantify the linkage between model complexity and model performance in 1,513 snow-dominated basins. Results show that (1) models differ substantially in their representation of physical processes, with the largest divergence in melting process treatments, followed by sublimation, interception and rainfall-snowfall partitioning processes. (2) While the model performance for Q_{sum} and Q_{max} shows limited sensitivity to model complexity, CTQ performance exhibits a positive correlation with model complexity ($r = 0.56$, $P < 0.05$) particularly in highly complex basins, highlighting the role of process complexity in stern conditions. (3) We also find that the model performance depends more on systematic and balanced representations of key processes than on complexity alone. High-complexity models with well-integrated processes (e.g., DBH) show high robustness, whereas models lacking critical modules exhibit poor accuracy, and even simpler models with well-designed modules (e.g., PCR-GLOBWB) can perform robustly. This study provides a quantitative framework for assessing model complexity and emphasizes that systematic process design is critical for improving SMR simulations in complex environments, offering guidance for future model development.

1 Introduction

Hydrological models play a central role in global water resources management, flood forecasting, and ecosystem protection. With advances in computational capacity and growing understanding of physical processes, these models have evolved from early conceptual models to today's complex and process-based frameworks. While more sophisticated models can represent a larger number of physical processes, numerous studies have shown that increasing complexity does not always yield statistically significant improvements in performance, in part due to over-parameterization and overfitting (Beven, 1993, 2006). At the same time, the principle of "as simple as possible, but not simpler" continues to guide model development (Valéry et al., 2014). Despite the complex basin conditions with strong spatial heterogeneity, previous studies have shown that even simple

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

models are capable of adequately describing its integrated response characteristics (Savenije, 2001; Schoups et al., 2008), as they embody fundamental physical principles (Ohmura, 2001). Therefore, a long-standing question that then arises is whether greater model complexity necessarily translates into better model performances (Schoups et al., 2008; Valéry et al., 2014; Reed et al., 2025). This remains a central challenge for model development, particularly in determining which physical processes warrant further refinement to justify added complexity and which can be simplified. As a result, a systematic, quantitative evaluation of the complexity–performance relationship is essential for the future of hydrological modeling.

This question is particularly relevant in snow-dominated basins. Compared with rainfall-driven runoff, snowmelt runoff (SMR) is a more intricate process, involving rainfall-snowfall partitioning, canopy interception, and sublimation during the accumulation period, as well as energy balance, liquid water transfer, and snow–albedo feedbacks during the melting period. The coexistence of these processes and their strong interactions creates a high degree of physical complexity, making SMR an ideal testbed to evaluate whether increased model complexity indeed translates into better performance.

However, existing models differ substantially in how these snowmelt-related processes are represented. Depending on their primary objectives, some models simplify or even ignore processes such as sublimation, relying instead on parameter calibration to approximate snow sublimation (e.g., LPJML). In contrast, others adopt a more process-based approach (e.g., CWATM), explicitly accounting for the energy requirements and resistances associated with sublimation. Similarly, while some models explicitly resolve snowmelt processes via full energy balance schemes—accounting for canopy radiative transfer (JULES-W2), snow aging (ORCHIDEE-MICT), or aerosol–albedo interactions (CLM40)—others rely on empirical degree-day formulations (e.g., LPJML, PCR-GLOBWB). Such difference makes it difficult to compare models and to understand the linkages between complexity and performance, especially given that model complexity is challenging to quantify (Best et al., 2015; Orth et al., 2015; Merz et al., 2022). Prior efforts to describe complexity have often relied on qualitative text descriptions or schematic illustrations (Telteu et al., 2021; Müller Schmied et al., 2025), which, although informative, lack a consistent quantitative framework. Therefore, there is a pressing need to develop a clear and unified framework for quantifying model complexity.

In addition, previous studies have suggested that higher complexity does not necessarily lead to superior performance (Savenije, 2001). For example, Ruelland (2023) evaluated the SIAR model in 17 basins in the French Alps and Pyrenees and found that a single-parameter model could achieve performance comparable to more complex alternatives. Similarly, Valéry et al. (2014) reported that moderately complex models outperformed both overly simple and overly complex models across 380 basins in France, Switzerland, Sweden, and Canada. Merz et al. (2022) using data from 700 basins across the continental United States, also highlighted that an intermediate-complexity model (SALTO17) performed best at the regional scale. However, several limitations remain. First, complexity is often defined merely by parameter count, overlooking the mechanistic representation of processes and thereby obscuring how process-level complexity affects performance. Second, most evaluations are based on limited model ensembles or basin samples, restricting the generalizability of their findings under large-scale and heterogeneous conditions. Third, performance assessments commonly neglect snowmelt-related metrics such as total runoff (Q_{sum}), peak discharge (Q_{max}), and centroid timing (CTQ), leading to incomplete conclusions in snowmelt-dominated regions. Collectively, these limitations hinder a robust understanding of the complexity–performance relationship and limit the transferability of insights across diverse model structures and hydrological conditions.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

To address these gaps, this study proposes a unified and quantitative framework to characterize model complexity in SMR processes. We compile 13 state-of-the-art large-scale hydrological, land surface, and dynamic vegetation models that span a broad spectrum of complexity. These include six ISIMIP2a water sector models (PCR-GLOBWB (Sutanudjaja et al., 2018), DBH (Tang et al., 2006), VIC (Liang et al., 1994), MATSIRO (Pokhrel et al., 2014), CLM40 (Oleson et al., 2010), LPJML (Schaphoff et al., 2018)) and seven ISIMIP3a models (CWATM (Burek et al., 2020), H08 (Hanasaki et al., 2008), HYDROPY (Stacke and Hagemann, 2021), JULES-W2 (Best et al., 2011), MIROC-INTEG-LAND (Yokohata et al., 2020), ORCHIDEE-MICT (Guimberteau et al., 2018), WATERGAP2-2E (Müller Schmied et al., 2024)). The ensemble is not intended to be an exhaustive set of all available models (Hou et al., 2023; Guo et al., 2024), but it aims to provide a representative and sufficiently diverse testbed for evaluating the complexity–performance relationship.

Specifically, we aim to (1) overcome the limitation of parameter-count–based definitions by developing a Tree-Based Model Complexity Scoring (TBMCS) method that quantifies process-level complexity in snow accumulation and melt; (2) address the lack of model and basin diversity by leveraging a large and representative ensemble of 13 state-of-the-art models across 1513 snowmelt-domain catchments worldwide; and (3) tackle the narrow scope of performance metrics by evaluating models against key SMR indicators—Q_{sum}, Q_{max}, and CTQ. In addition, this study seeks to provide a systematic and comprehensive assessment of whether greater model complexity leads to improved performance under large-scale and high-heterogeneity conditions, while also identifying priorities for process representation to guide future model development and selection.

2 Data and Methods

2.1 Models and study regions

This study uses 13 models from ISIMIP2a/3a, categorized into seven global hydrological models (GHMs: PCR-GLOBWB, DBH, VIC, CWATM, H08, HYDROPY, and WATERGAP2-2E), five land surface models (LSMs: MATSIRO, CLM40, JULES-W2, MIROC-INTEG-LAND, and ORCHIDEE-MICT), and one dynamic global vegetation model (DGVM: LPJML) as the testbed for examining the linkage between model complexity and performance. These categories differ in their representation of physical processes: GHMs generally provide more comprehensive descriptions of water balance processes; LSMs incorporate more detailed parameterizations of energy-related processes; and DGVMs place greater emphasis on vegetation and ecological dynamics. Analyzing differences in model performance thus offers insight into disentangling the role of process representation across model categories.

The study domain comprises 1,513 natural basins in the mid- to high-latitude Northern Hemisphere that are minimally affected by human activities, glaciers, or permafrost, and that have at least 10 years of observational records during 1979–2019 (Yin et al., 2024). The key runoff characteristics considered here are Q_{sum}, Q_{max}, and CTQ, which are crucial for water resource utilization, flood hazard prevention, and water resource management, respectively. Details of basin selection, characteristic definitions, and data quality control procedures are provided in **Section 2** of **Part 1**.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

2.2 The main snow accumulation and melt processes

Here, we examine several key cascading processes from the snow accumulation period to the snowmelt period, including rainfall–snowfall partitioning, snow interception, snow sublimation, snowmelt, canopy radiative transfer, and surface albedo changes. This design follows our experimental focus on the key characteristics of SMR (i.e., total runoff, peak runoff, and centroid timing). Among these processes, rainfall–snowfall partitioning, snow interception, and snow sublimation are related to snow accumulation and are more closely associated with Q_{sum} and Q_{max}, whereas canopy radiative transfer and surface albedo are mainly associated with snowmelt. Thus, it is necessary to score the complexity of all cascading processes. A detailed description is provided in **Section 2.2.1** of **Part 1**.

2.3 The Tree-Based Model Complexity Scoring (TBMCS) method

To quantitatively assess the relationship between model performance and model complexity, a tree-based framework is introduced to evaluate model complexity (**Fig. 1**). This framework captures structural differences—specifically tree depth, number of nodes, and number of leaves—which together characterize the overall complexity of a model. In this representation, the calculation of a variable Y at the root node (**Fig. 1**) is considered "more complex" when the model relies on a larger number of governing equations (nodes, blue boxes), input variables (leaves, grey boxes), and/or parameters (grey boxes with dashed outlines), all of which enrich the physical realism of the modeled processes. Here, "physical realism" refers to more explicit and comprehensive descriptions of land-surface heterogeneity and physical parameterizations that capture dynamic system behaviors. Such enhancements typically increase model complexity by expanding the number of nodes and leaves within the tree structure.

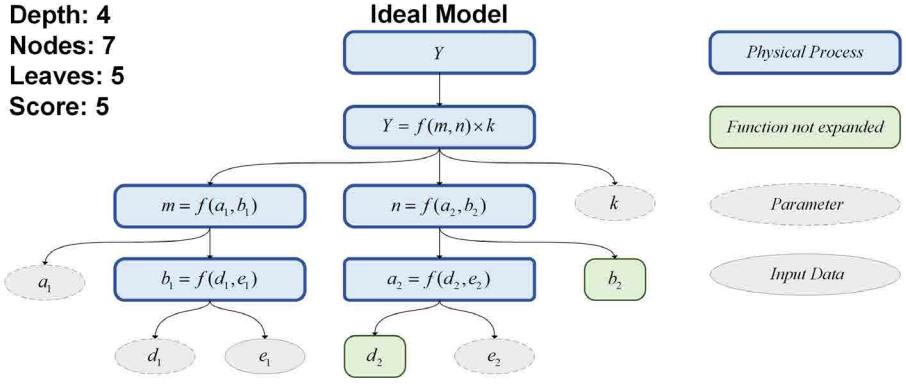

Figure 1. A conceptual diagram of the 'Tree-Based Model Complexity Scoring' (TBMCS) method. The root node is the target variable to be predicted (Y). The node will continue branching for each term in the equation if the term is further modeled using another equation or sub-model. The branching criteria for nodes follow the rules described in the text above.

To build such a tree, we start from a variable of interest and its mathematical equation in a root node. A node continues branching for each term of the equation if that term is further modeled with an additional equation (or sub-model). Branching

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

stops when one of the following criteria is met: (1) the term corresponds to an input dataset or parameter (**Fig. 1**, grey boxes); or (2) the term corresponds to a variable whose physical representation does not significantly differ among models or is already encoded in another tree (**Fig. 1**, green boxes). By constructing such trees for the same variable (e.g., Y in **Fig. 1**) and applying consistent branching criteria across different large-scale hydrological models, we establish a unified, quantitative, and measurable framework for scoring the relative (not absolute; see **Discussions**) complexity of snow-related processes.

After building the trees, the number of depths, nodes, and leaves is counted, and weights are assigned to obtain a quantitative complexity score for each model. Given the hierarchical structure that governs the model's ability to accurately simulate runoff characteristics, we assign weights of 0.6, 0.3, and 0.1 to tree depth, number of nodes, and number of leaves, respectively. These weights reflect the relative importance of each structural attribute in influencing complexity. Although the selection of weights is subjective, it does not affect the relative ranking among models, which is the primary purpose of this study. Future work may revise these weights to better represent absolute complexity.

Finally, we assess the four key physical processes described in **Section 2.2** by assigning each a complexity score that reflects the difficulty of estimating snowfall (P_{snow}), interception (P_{int}), sublimation (E_{snow}), and snowmelt (M). The overall complexity of each model is then calculated as the sum of these scores, providing a straightforward measure of both the relative emphasis on individual processes and the model's overall structural complexity.

2.4 Experiment design

To analyze the relationship between model complexity and model performance, we use the Pearson correlation coefficient (r) as the primary metric. The Pearson correlation coefficient r_{xy} between two variables x and y is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1}$$

where x_i and y_i are the values of the i-th ($i = 1, \ldots, n$) observations of variables x and y, and $\bar{x}$ and $\bar{y}$ are their respective means. In this study, x represents the model complexity score and y represents the model performance (e.g., model bias).

Basin complexity is primarily determined by two key factors—topography and vegetation. To quantify these factors, we employ four representative metrics: basin mean elevation (DEM) and its variability (DEMstd) for topography, and leaf area index (LAI) and plant functional type entropy (PFTh) for vegetation. Each metric plays a critical role in shaping the physical processes that govern snowmelt runoff. A composite basin complexity index is derived for each basin by normalizing and aggregating these four metrics, following the detailed procedure described in **Section 2.2.3** of **Part 1**.

To further disentangle the effects of individual basin complexity factors on the relationship between model complexity and model performance, we apply partial correlation analysis. This method quantifies the association between two variables while

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

controlling for the influence of a third variable, and is defined as:

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}, \tag{2}$$

where $r_{xy\cdot z}$ denotes the partial correlation coefficient between variables x and y after controlling for factor z. In this study, we define x as model complexity, y as model performance in a given basin, and z as a specific basin complexity factor (e.g., DEM, DEMstd, LAI, or PFTh).

3 Results

3.1 Quantification and Characterization of Model Complexity

We begin our analysis by quantifying the physical complexity of the rainfall–snowfall partitioning process (**Fig. 2**). Rainfall–snowfall partitioning is a key process that determines the phase of precipitation reaching the land surface, with different phases exerting substantial influence on snowpack accumulation and ablation. Air temperature is widely recognized as the primary indicator for distinguishing precipitation phase. Existing models generally implement precipitation-phase discrimination via two approaches: (1) directly using the precipitation phase from the forcing data (grey shaded area in **Fig. 2**), or (2) applying a temperature-threshold-based rainfall–snowfall partitioning scheme. Among the 13 ISIMIP models, four (MATSIRO, H08, MIROC-INTEG-LAND, ORCHIDEE-MICT) directly use precipitation phase from the forcing data. The remaining nine models employ a rainfall–snowfall partitioning scheme, which typically follows one of two formulations: a fixed single-temperature threshold (blue shaded area in **Fig. 2**) or a segmented dual-temperature threshold (green shaded area in **Fig. 2**).

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

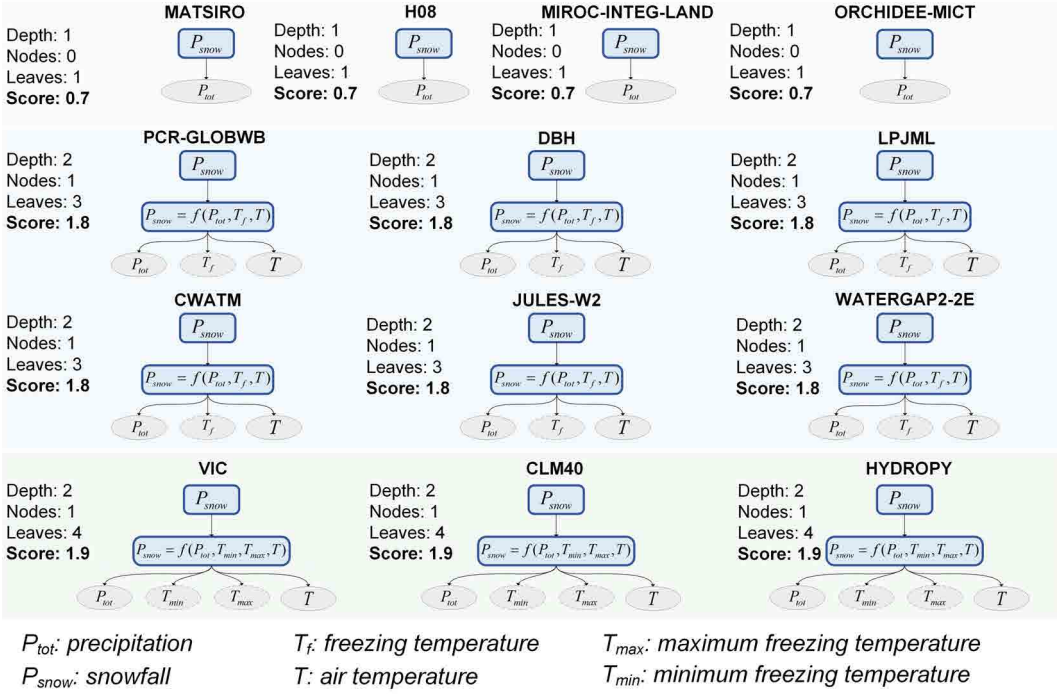

Figure 2. TBMC-based scoring for the rainfall-snowfall partitioning process. Each schematic diagram shows the tree for each model. The plotting order is arranged from the lowest to the highest score. The gray, blue, and green shaded areas represent snowfall obtained directly from the forcing data, estimated using a fixed single temperature threshold, and estimated using a segmented dual-temperature threshold, respectively.

The fixed single-temperature threshold method adopts a constant critical temperature to determine precipitation phase. When the air temperature falls below the threshold, precipitation is classified as snowfall; otherwise, it is rainfall. Models such as PCR-GLOBWB, DBH, LPJML, CWATM, JULES-W2, and WATERGAP2-2E apply this approach. In contrast, the dual-temperature threshold method employs two critical temperatures to represent a transitional range between rainfall and snowfall (e.g., VIC, CLM40, HYDROPY). Based on these structural differences, the complexity score for the rainfall–snowfall partitioning process across the 13 models ranges from 0.7 to 1.9.

Following rainfall–snowfall partitioning, snow interception is the next key process that determines the fraction of snowfall retained within the canopy versus that reaching the ground. This process also influences the exchange of energy and mass between the forest canopy and the atmosphere, thereby altering the sub-canopy snow distribution. Among the 13 ISIMIP models, all except H08 include an explicit representation of interception (**Fig. 3**). Based on the TBMCS method, the complexity scores range from 1.7 to 4.1. Most models share a similar three-layer tree depth, with differences primarily arising from the number of leaves (i.e., parameters). The magnitude of interception capacity is closely tied to vegetation representation, typically parameterized as a function of LAI (blue shading in **Fig. 3**). CWATM provides the simplest scheme, relying solely on

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

a fixed coefficient (grey shading), whereas WATERGAP2-2E offers the most detailed formulation by accounting for vegetation, temperature, and evapotranspiration effects (green shading).

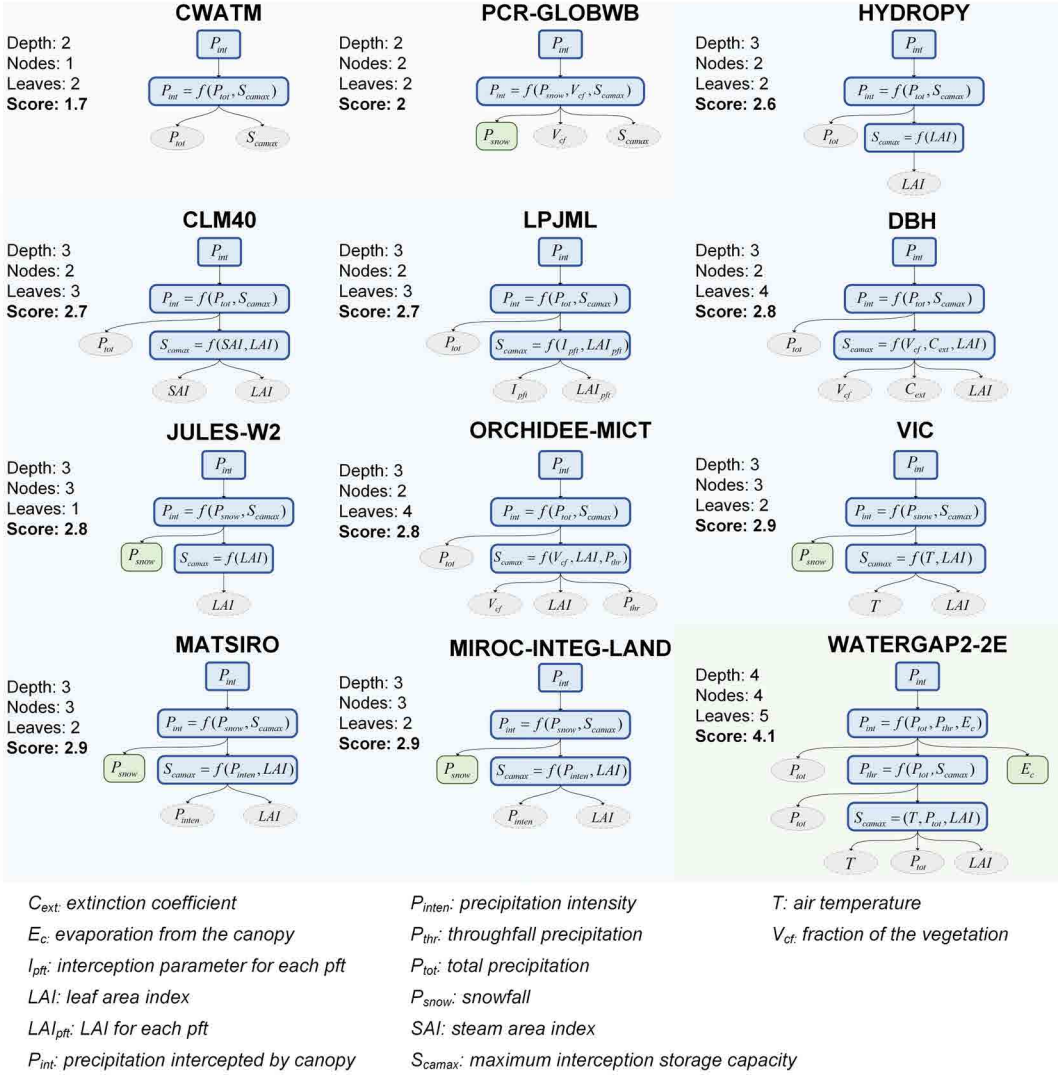

Figure 3. TBMC-based scoring for the interception process. Each schematic diagram shows the tree for each model. Since the H08 model does not include an interception module, it is therefore excluded from the scoring. The plotting order is arranged from the lowest to the highest score. Gray, blue, and green shaded areas denote interception directly linked to the parameter, with LAI effects, and with evaporation effects, respectively.

Notably, several models—including CWATM, HYDROPY, CLM40, LPJML, DBH, ORCHIDEE-MICT, and WATERGAP2-2E—do not distinguish precipitation phase in their interception parameterizations, applying a unified scheme for both rainfall and snowfall. In contrast, the remaining models employ separate parameterizations specifically for snow interception.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

Snow sublimation is a major snow-loss process during the cold season and a key component of energy and mass exchanges between the cryosphere and the atmosphere. Among the evaluated models, all except HYDROPY explicitly represent sublimation, yielding complexity scores from 1.6 to 4.7. These representations can be grouped into two main types: (1) a fixed sublimation value, as in LPJML (grey shading in **Fig. 4**); and (2) formulations that estimate sublimation based on evapotranspiration, adopted by the other models. Most models share a three- to four-layer depth and account for the effects of vapor pressure deficit (VPD), surface resistance (R_{surf}), and aerodynamic resistance (R_{air}) on sublimation (blue and green shading). More complex implementations explicitly model R_{surf} (green shading), primarily by incorporating vegetation characteristics such as LAI, as evident in JULES-W2, DBH, CLM40, PCR-GLOBWB, and CWATM. CWATM also includes a crop-specific sublimation factor when the underlying surface is cropland, resulting in the highest complexity score (4.7).

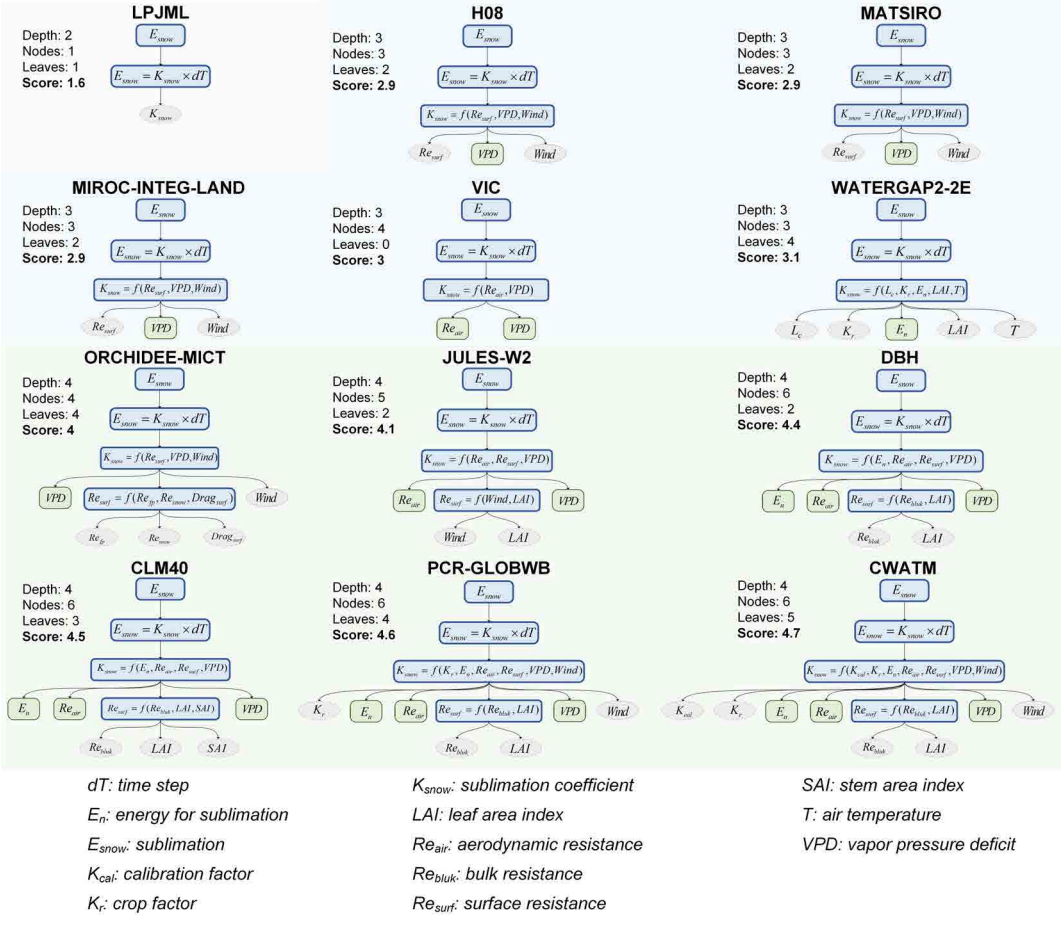

Figure 4. TBMC-based scoring for the sublimation process. Each schematic diagram shows the tree for each model. Since the HYDROPY model does not include a sublimation module, it is therefore excluded from the scoring. The plotting order is arranged from the lowest to the highest score. Gray, blue, and green shaded areas denote sublimation directly linked to a fixed parameter, accounting for a resistance parameter, and further parameterized surface resistance, respectively.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

Snowmelt, the process through which accumulated snow ablates, is the most critical determinant of runoff generation. Based on structural differences, existing snowmelt modules can be broadly classified into two categories: degree-day models (grey shading in **Fig. 5**) and energy-balance models (blue shading in **Fig. 5**). PCR-GLOBWB, LPJML, CWATM, HYDROPY, and WATERGAP2-2E employ the degree-day method, whereas H08, VIC, DBH, MATSIRO, MIROC-INTEG-LAND, ORCHIDEE-MICT, JULES-W2, and CLM40 use an energy-balance formulation.

Degree-day models estimate melt solely from a degree-day factor, producing relatively low complexity scores (1.8–2.2). Differences nevertheless exist among these models depending on whether the degree-day factor is fixed or dynamic. For example, PCR-GLOBWB, LPJML, and WATERGAP2-2E employ fixed factors, whereas CWATM uses a dynamic factor incorporating rainfall intensity and seasonality. HYDROPY also uses a dynamic factor that increases linearly with the number of snowmelt days.

In contrast, energy-balance models exhibit substantially higher complexity scores (5.2–6.1), as they compute snowmelt by explicitly resolving the snowpack energy budget, including longwave and shortwave radiation, sensible and latent heat fluxes, ground heat flux, and the latent heat of fusion. Shortwave radiation is particularly critical. H08 calculates vegetation shading directly using albedo, whereas models such as JULES-W2, DBH, and CLM40 employ the two-stream approximation (Sellers, 1985). CLM40 further enhances this representation by explicitly accounting for vegetation structure and type through parameters such as stem area index (SAI), leaf reflectance, and spectral transmittance across PFTs. Surface albedo is another major component; although snow albedo depends on snow age, melt stage, surface temperature, water content, and impurities, simpler models (e.g., VIC) represent only its decay with snow aging. In contrast, CLM40 includes solar zenith angle, impurities such as black carbon and dust, snow age, temperature, and spectral dependence, yielding a far more comprehensive treatment.

In summary, degree-day models offer simplicity and computational efficiency but rely on empirical assumptions, whereas energy-balance models provide a more physically realistic depiction of snowmelt at the cost of increased complexity and data requirements.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

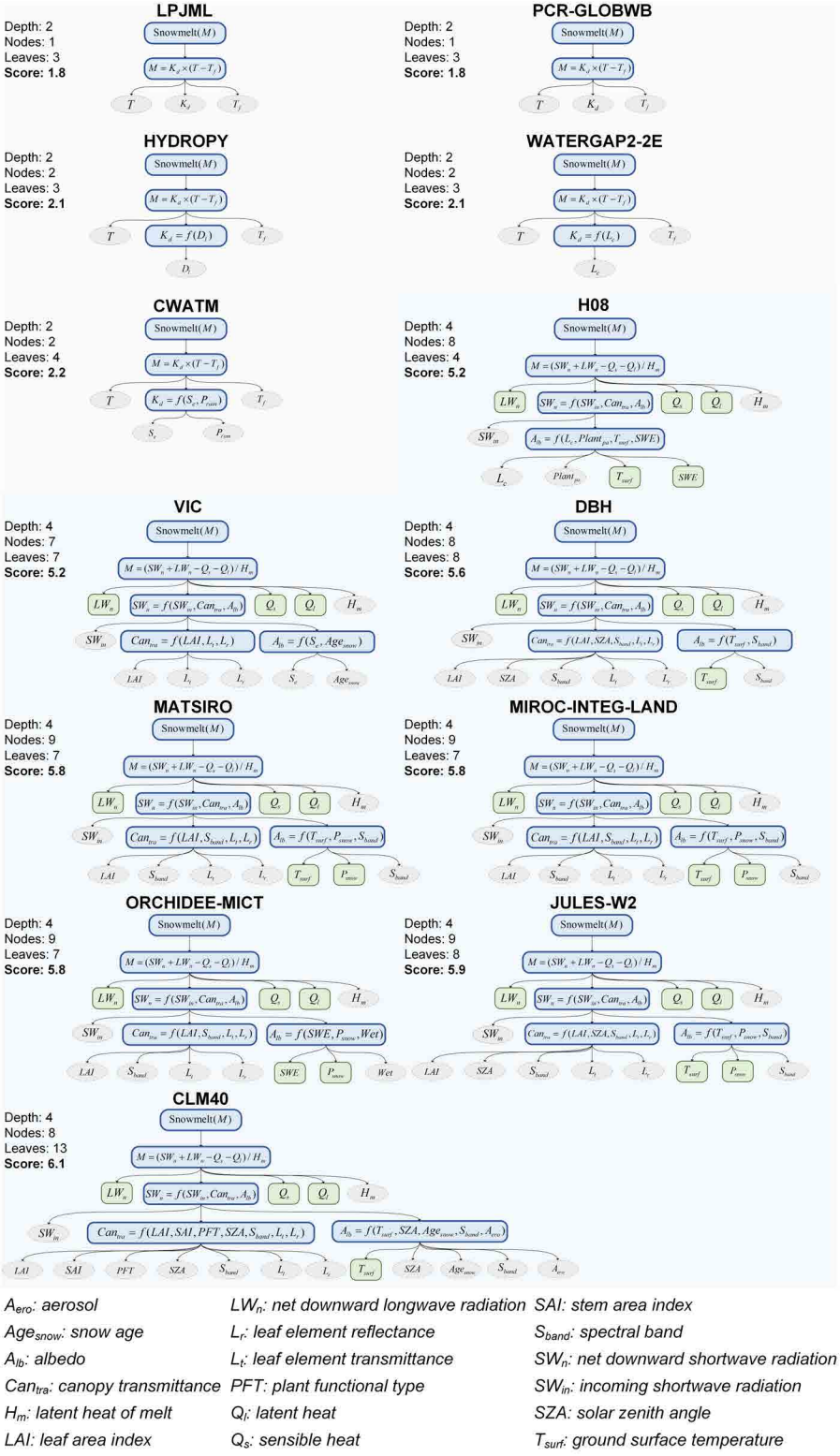

Figure 5. TBMC-based scoring for the melt process. Each schematic diagram shows the tree for each model. The plotting order is arranged from the lowest to the highest score. Gray and blue shaded areas denote the degree-day factor model and the energy balance model, respectively.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

We aggregated the complexity scores of the four key physical processes for each model to derive a total complexity score (**Table 1**). The results indicate that the three most complex models are CLM40, JULES-W2, and DBH (scores > 14), whereas the least complex are HYDROPY, LPJML, and H08 (scores < 10). Unlike earlier approaches defining complexity solely by the number of fluxes and storages (Müller Schmied et al., 2025), our framework provides a more quantitative and structurally consistent assessment by considering not only the represented variables but also the number of physical formulations, their associated parameters and input datasets.

Table 1. Model structural complexity components.

Model	Category	Rainfall–snowfall partitioning	Interception	Sublimation	Melt	Sum
HYDROPY	GHM	1.9	2.6	0	2.1	6.6
LPJML	DGVM	1.8	2.7	1.6	1.8	7.9
H08	GHM	0.7	0	2.9	5.2	8.8
PCR-GLOBWB	GHM	1.8	2	4.6	1.8	10.2
CWATM	GHM	1.8	1.7	4.7	2.2	10.4
WATERGAP2-2E	GHM	1.8	4.1	3.1	2.1	11.1
MATSIRO	LSM	0.7	2.9	2.9	5.8	12.3
MIROC-INTEG-LAND	LSM	0.7	2.9	2.9	5.8	12.3
VIC	GHM	1.9	2.9	3	5.2	13
ORCHIDEE-MICT	LSM	0.7	2.8	4	5.8	13.3
DBH	GHM	1.8	2.8	4.4	5.6	14.6
JULES-W2	LSM	1.8	2.8	4.1	5.9	14.6
CLM40	LSM	1.9	2.7	4.5	6.1	15.2

The largest contributions to total complexity and inter-model variance arise from the melt process (1.8–6.1), followed by sublimation (0–4.7), interception (0–4.1), and rainfall–snowfall partitioning (0.7–1.9). This pattern highlights that existing models place the greatest emphasis on melt processes, while snow-accumulation processes are represented less extensively. A category-level comparison further shows that LSMs generally exhibit higher complexity than GHMs and the DGVM, primarily due to their detailed energy-balance formulations within the melt modules.

3.2 Linkage between model complexity and model performance

Based on the established complexity metrics for each model, we subsequently examine the relationship between model complexity and model performance, aiming to identify the conditions under which higher complexity confers performance advantages. **Figure 6** summarizes the overall relationship between model complexity and performance. The results show that for CTQ (**Fig. 6c**), there is a positive relationship ($r = 0.45$) between complexity and model performance—defined as the propor-

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

tion of basins with acceptable bias ($\pm 20\%$ for Q_{sum} and Q_{max}, ± 5 days for CTQ across 1513 basins). For Q_{sum} and Q_{max} (**Figs. 6a–b**), model complexity exhibits little to no consistent influence on performance ($r = 0.03$, $r = -0.14$, respectively).

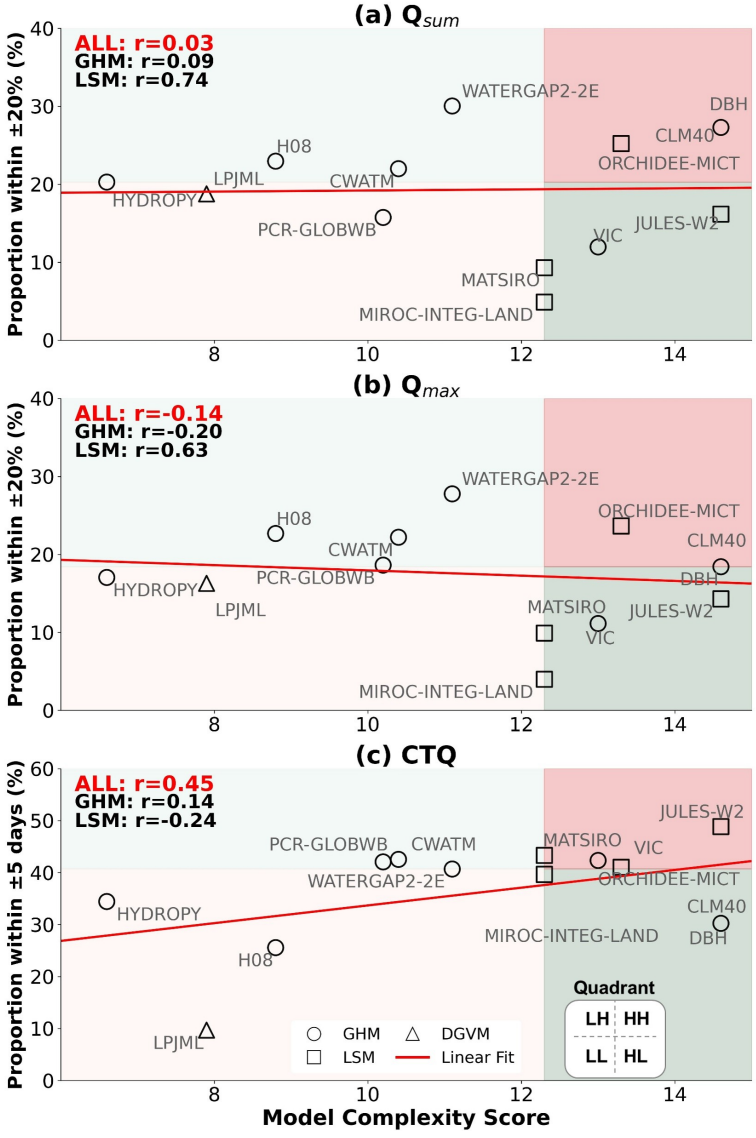

Figure 6. Relationship between model complexity and model performance. Panels (a–c) present Qsum, Qmax, and CTQ, respectively. The red line denotes the fitted regression. HH (high model complexity and high model performance), HL (high complexity and low performance), LH (low complexity and high performance), and LL (low complexity and low performance) are defined based on the median values of the corresponding x- and y-axis variables. Point shapes indicate model categories: circles for GHM, squares for LSM, and triangles for DGVM. ALL, GHM, and LSM represent the Pearson correlation coefficients calculated using all models, global hydrological models, and land surface models, respectively.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

We then divided basins into high- and low-heterogeneity groups using the median value of basin complexity as the threshold (definitions provided in **Section 2** of **Part 1**), aiming to test whether model complexity yields consistent performance gains under different heterogeneity conditions. **Figure 7** shows that, overall, the correlation between model complexity and performance is stronger under high basin complexity. Across different runoff characteristics, the correlation is not significant for Q_{sum} and Q_{max}. In contrast, for CTQ, model complexity is significantly associated with performance ($r = 0.56$, $P < 0.05$) under high basin complexity (**Fig. 7f**). Ten out of thirteen models fall within the 95% confidence interval of the fitted line.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

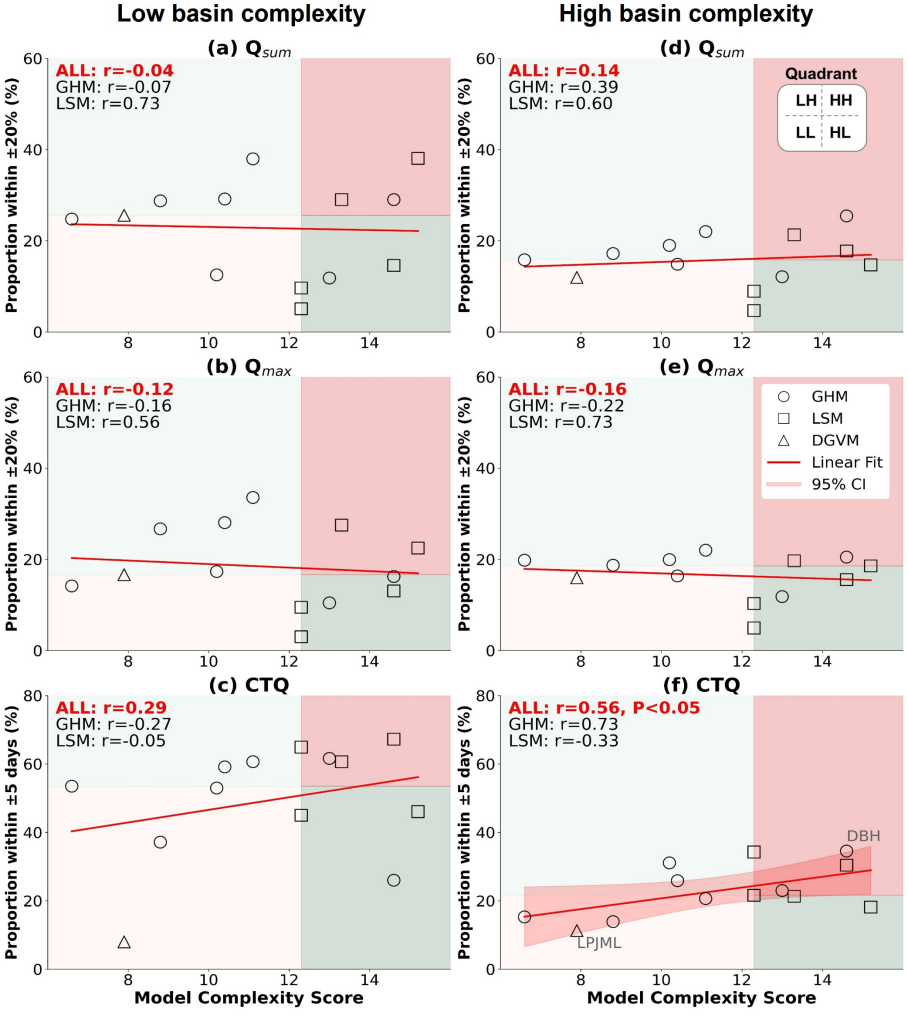

Figure 7. Relationship between model complexity and model performance under different basin complexity scenarios. (a–c) show the relationship under low basin complexity and (d–f) show the relationship under high basin complexity. The red line denotes the fitted regression. The red shading indicates the 95% confidence interval. HH (high model complexity and high model performance), HL (high complexity and low performance), LH (low complexity and high performance), and LL (low complexity and low performance) are defined based on the median values of the corresponding x- and y-axis variables. Point shapes indicate model categories: circles for GHM, squares for LSM, and triangles for DGVM. ALL, GHM, and LSM represent the Pearson correlation coefficients calculated using all models, global hydrological models, and land surface models, respectively.

The relationship between model complexity and performance differs across runoff characteristics because they rely on distinct processes. Q_{sum} is primarily constrained by water balance closure (snowfall–sublimation–snowmelt) and is therefore less sensitive to detailed process representations; even relatively simple models can yield reasonable results as long as water

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

balance is maintained. Q_{max} 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.

In contrast, CTQ is highly sensitive to energy-related processes (e.g., canopy radiative transfer, snow density, surface albedo). Simple degree-day schemes often misrepresent melt timing, whereas more complex energy-based formulations capture the onset and magnitude of snowmelt more accurately, leading to stronger improvements in CTQ performance. For example, DBH and LPJML illustrate typical cases of high-complexity–high-performance and low-complexity–low-performance, respectively: DBH employs an energy-balance approach, while LPJML relies on a fixed degree-day factor. Under high basin complexity, fixed factors fail to account for key mechanisms such as melt retardation from canopy shading or melt acceleration via snow–albedo feedbacks. This highlights that model complexity becomes particularly advantageous in heterogeneous environments, where more detailed representations are required to capture the intricate processes controlling runoff timing.

Error compensation further explains these differences. For Q_{sum} and Q_{max}, 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. In snow-dominated basins, spatial variability in energy balance and snow storage strongly controls runoff timing, giving more complex models a clear advantage that cannot be achieved through parameter tuning alone.

Building on the preceding section, **Figure 8** illustrates the mechanisms linking model complexity to performance. Partial correlation analysis shows that the impacts of individual basin complexity factors (DEM, DEMstd, LAI, and PFTh) are uneven. Model complexity exhibits limited or even negative correlations with Q_{sum} and Q_{max} and shows little sensitivity to heterogeneity, whereas CTQ is most strongly influenced, displaying a pronounced threshold effect and a positive correlation with complexity that becomes stronger under higher heterogeneity.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

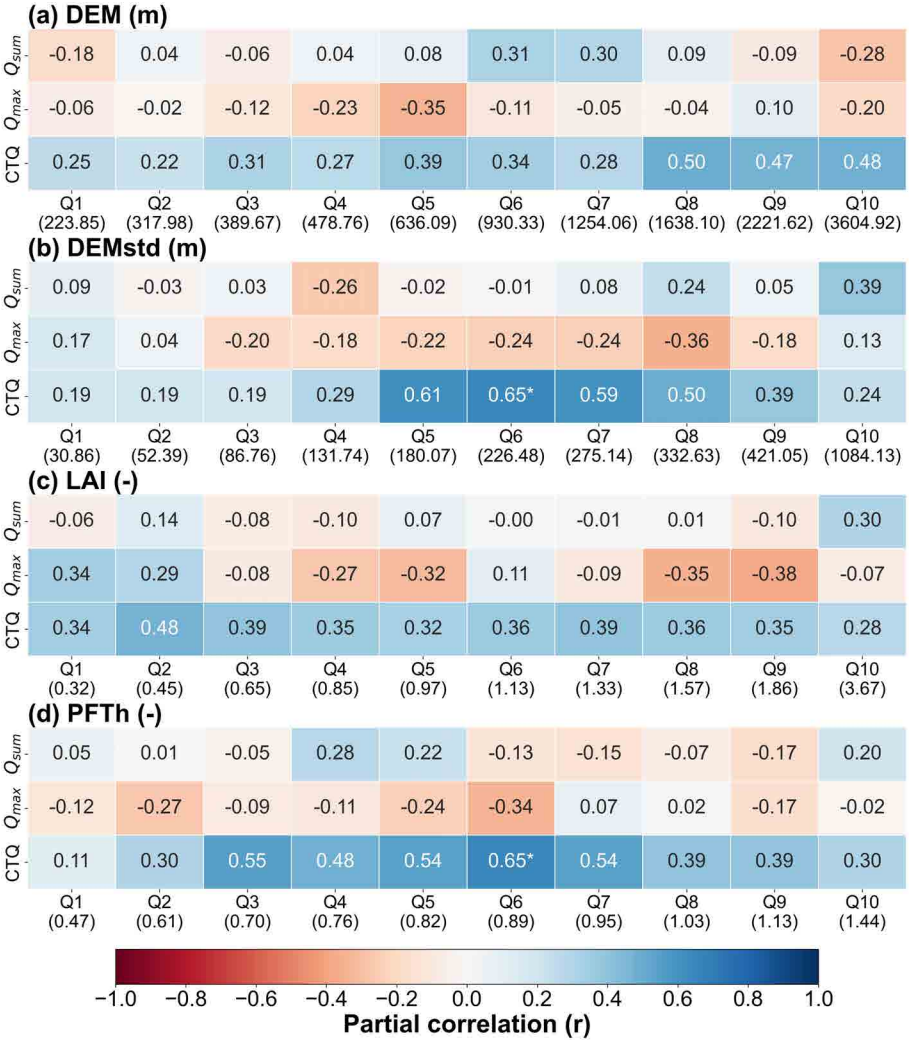

Figure 8. Influence of basin complexity factors on the correlation between model complexity and model performance. Panels (a–d) correspond to DEM, DEMstd, LAI, and PFTh, respectively. In each panel, the x-axis (Q1–Q10) represents deciles from the 10th to the 100th percentile, with values shown in parentheses. Each grid cell indicates the partial correlation, where red denotes negative correlation and blue denotes positive correlation. * indicate statistical significance at $P < 0.05$.

For the threshold effects in the relationship between model complexity and CTQ performance, the correlation with model complexity reaches its maximum at a PFTh of approximately 0.89 (**Fig. 8d**). This is likely because greater vegetation diversity enhances canopy–snow interactions (Musselman et al., 2008), but beyond this threshold, the coexistence of multiple canopy–snow processes may diminish the advantage of complex models. A similar threshold is observed for DEMstd at roughly 226.48 m (**Fig. 8b**). At this point, the gain from complexity peaks as models become fully equipped to resolve intricate snowmelt processes. However, beyond this threshold, errors in forcing data and routing, along with increased parameter uncer-

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

tainties, begin to dominate, limiting further benefits from additional complexity. For Q_{sum} and Q_{max}, threshold effects are also present, though not statistically significant. For example, when DEM reaches 930.33 m, the correlation between model complexity and Q_{sum} performance peaks at 0.31 (**Fig. 8a**). Similarly, when LAI is 0.32, the correlation between model complexity and Q_{max} performance reaches its maximum value of 0.34 (**Fig. 8c**). These findings underscore that model development and improvement should be adapted to basin-specific surface conditions. Enhancing the representation of key physical processes is crucial for improving accuracy and robustness, but efforts should be targeted to avoid excessive parameterization that may introduce additional uncertainty or overfitting.

We also find that with increasing basin complexity, the link between model complexity and performance becomes stronger, particularly under conditions of higher DEM, where more complex models are required to reliably simulate CTQ. This is physically grounded in the fact that high-elevation regions receive strong surface radiation and accumulate substantial snow, requiring energy-balance-based melt schemes to accurately capture both the onset and acceleration of snowmelt—processes that are typically represented in more complex models and are essential for reproducing CTQ. For Q_{sum} and Q_{max}, accurate simulation is largely determined by processes related to snow accumulation, whereas model complexity is primarily reflected in the representation of snowmelt. As a result, the potential benefits of added complexity are often offset by parameter uncertainties, which can even lead to a negative relationship between complexity and performance.

Finally, we extend these findings from a new perspective by examining the relationship between complexity and model robustness (**Fig. 9**). Here, robustness is redefined as a composite criterion of model performance, requiring both low bias and a strong ability to maintain accuracy under increasing basin complexity (see **Section 2** of **Part 1** for details). By incorporating this indicator, **Figure 9** complements **Figure 6** by moving beyond general accuracy and providing a more comprehensive assessment of model robustness in heterogeneous environments.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

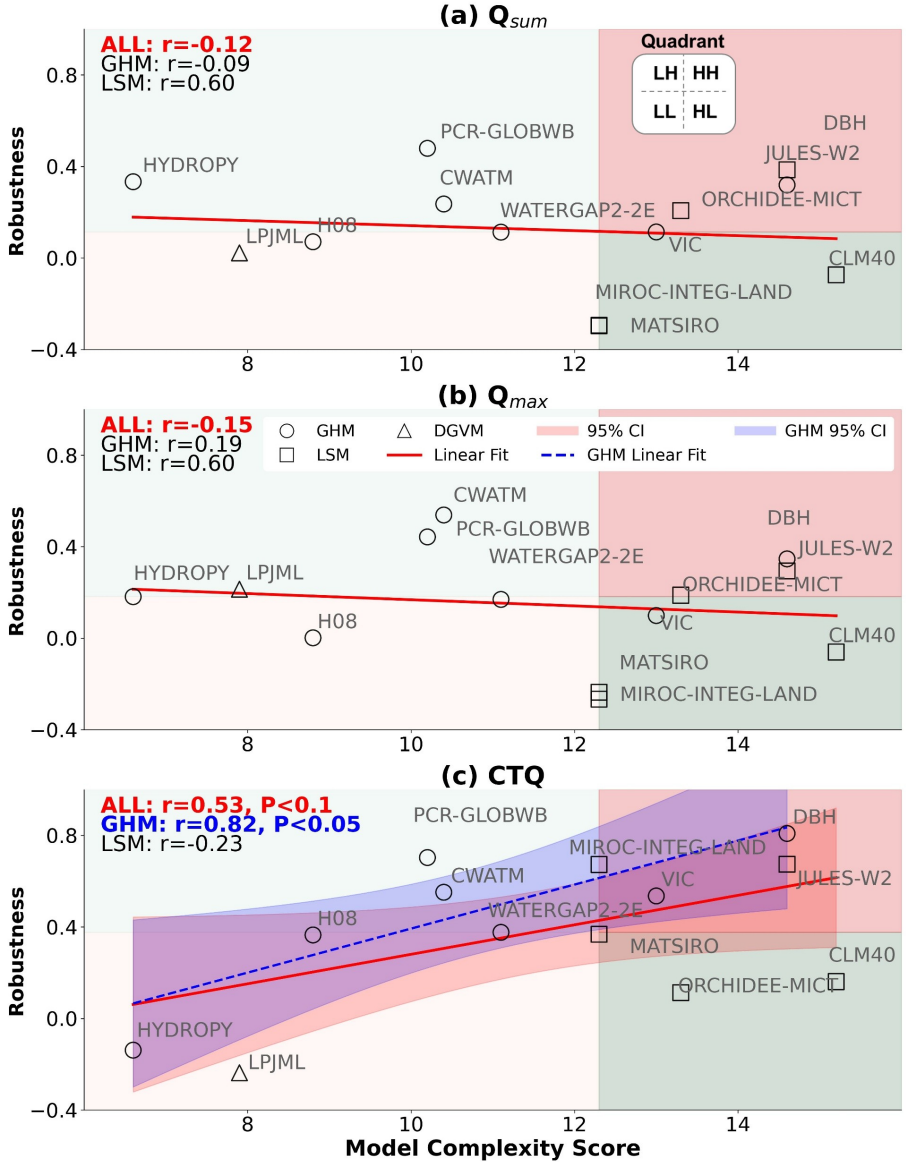

Figure 9. Relationship between model complexity and model robustness. Panels (a–c) present Q_{sum}, Q_{max}, and CTQ, respectively. The red and blue line represents the fitted regression for all models and GHMs. The red and blue shading indicates the 95% confidence interval. HH (high model complexity and high model performance), HL (high complexity and low performance), LH (low complexity and high performance), and LL (low complexity and low performance) are defined based on the median values of the corresponding x- and y-axis variables. Point shapes indicate model categories: circles for GHM, squares for LSM, and triangles for DGVM. ALL, GHM, and LSM represent the Pearson correlation coefficients calculated using all models, global hydrological models, and land surface models, respectively.

Consistent with the previous findings, complexity shows no significant correlation with robustness for Q_{sum} and Q_{max} (**Figs. 9a–b**). However, for CTQ (**Fig. 9c**), there is a significant positive correlation between complexity and model robustness

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

($r = 0.53$, $P < 0.1$). This strongly validates our previous conclusions: for simulating centroid timing, more complex models not only demonstrate a higher potential for accuracy in complex environments but also exhibit greater stability when applied across a wide range of conditions. Notably, GHMs exhibit a strong positive correlation ($r = 0.82$, $P < 0.05$).

In terms of model robustness, DBH, VIC, MATSIRO, H08, HYDROPY, JULES-W2, and WATERGAP2-2E fall within the 95% confidence interval of the regression line, indicating that increasing model complexity enhances the robustness of CTQ simulation, whereas other models show weaker associations.

3.3 Physical Process Mechanisms and Targeted Model Optimization

To further interpret the findings presented above, we conducted a process-level analysis to link model complexity with resilience and to derive implications for future model development (**Fig. 10**). Models were classified into four quadrants (HH, HL, LH, LL) according to their overall complexity and robustness (**Fig. 9**).

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

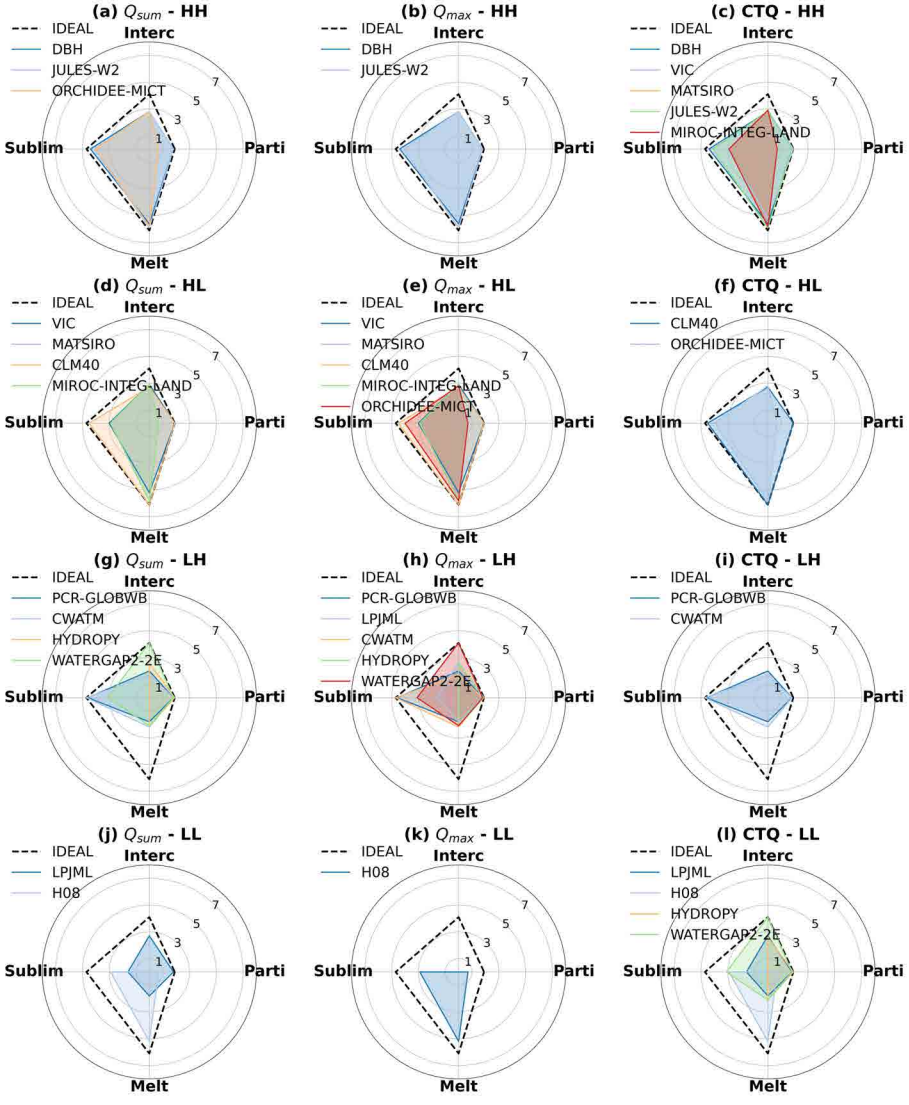

Figure 10. Cluster analysis of process complexity among different models. Columns correspond to Q_{sum}, Q_{max}, and CTQ, while rows represent the four clusters: HH (high complexity–high robustness), HL (high complexity–low robustness), LH (low complexity–high robustness), and LL (low complexity–low robustness). Each radar plot illustrates the relative scores of four key process modules (Parti: partitioning, Interc: interception, Sublim: sublimation, and Melt) for the models within the corresponding cluster. Colored polygons denote individual models, and the thick black dashed line represents the ideal model with maximum scores across all processes, serving as a benchmark for comparison.

The radar plots reveal that models in the HH quadrant (high complexity, high robustness) exhibit a more balanced distribution of complexity across key physical processes (**Figs. 10a–c**). This suggests that robust performance is not determined by the mere addition of processes, but rather by the integration of a well-structured and systematically designed process representation (e.g.,

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

DBH). In contrast, models in the LL quadrant show pronounced deficiencies in their process representations (**Figs. 10i–l**). For instance, H08 lacks a realistic treatment of canopy interception, which likely contributes to its poor performance.

Models in the HL quadrant are characterized by high structural complexity but low robustness. For example, although CLM40 incorporates a wide range of physical processes, its reliance on fixed default parameter values appears to limit performance and result in high bias (Yan et al., 2023). A comparison between HH and HL models indicates that their overall structural complexity scores are not markedly different; thus, the divergence in performance is more likely driven by parameter choices, specific process formulations, and the strategies used to couple individual components.

By contrast, models in the LH quadrant demonstrate that satisfactory resilience can be achieved even with relatively modest structural complexity. For example, PCR-GLOBWB and CWATM are consistently classified as LH across all three runoff characteristics. These models maintain a complete representation of key hydrological processes despite their lower overall complexity. Notably, CWATM enhances snowmelt simulation by introducing additional parameters to overcome limitations inherent in the degree-day factor method, illustrating that low-complexity models can still deliver robust accuracy when their process formulations are carefully designed.

To further investigate how individual physical process modules influence model performance, we selected three representative models along the linear relationship between model complexity and CTQ robustness identified in **Figure 10c**. These models illustrate a gradient of structural sophistication (**Figure 11**). The simple model H08 lacks key processes such as canopy interception, constraining its ability to realistically simulate snowmelt runoff. In contrast, the medium-complexity model CWATM incorporates all major hydrological processes but employs a simplified degree-day formulation for snowmelt estimation, which may limit its skill in capturing the timing of meltwater release. At the high-complexity end, the DBH model includes a more comprehensive suite of physical processes and represents snowmelt using an energy-balance approach, thereby improving the physical realism of melt simulations.

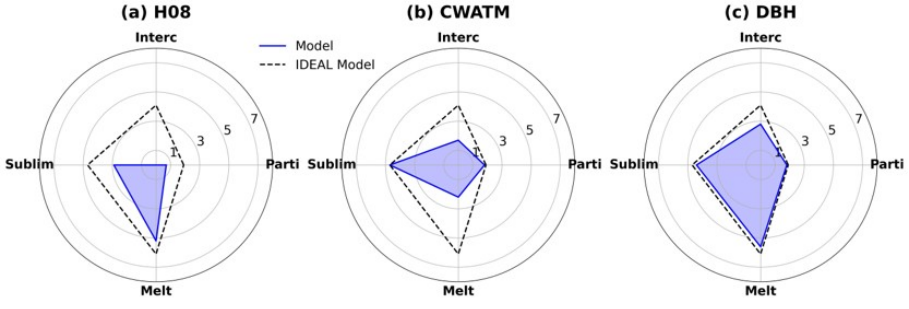

Figure 11. Complexity of physical process representation in representative models. Each radar plot illustrates the relative scores of four key process modules (Parti: partitioning, Interc: interception, Sublim: sublimation, and Melt) for the models within the corresponding cluster. Colored polygons denote individual models, and the thick black dashed line represents the ideal model with maximum scores across all processes, serving as a benchmark for comparison.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

This comparison highlights that although increasing complexity enables a more explicit representation of cryospheric processes, performance gains depend critically on whether the added complexity addresses key limitations—such as melt parameterization—rather than on structural complexity alone.

4 Conclusions and Discussions

This study develops the TBMCS method, which for the first time systematically quantifies and compares the physical complexity of snow-related processes across 13 global hydrological and land surface models. Building on this, we conducted a comprehensive assessment across global 1,513 basins to examine the relationship between model complexity and performance in simulating three key snowmelt runoff characteristics (Q_{sum}, Q_{max}, and CTQ), further revealing the pathways through which model complexity influences performance. Our findings provide new insights into the long-standing debate on where and to what extent the added complexity of models yields substantive performance gains. The primary findings are summarized as follows:

- Substantial inter-model differences exist in the representation of physical process complexity, with total scores ranging from below 10 (e.g., HYDROPY, LPJML, H08) to above 14 (e.g., CLM40, JULES-W2, DBH). The largest divergence arises from the snowmelt process (1.8–6.1), followed by sublimation (0–4.7), interception (0–4.1), and rainfall–snowfall partitioning (0.7–1.9), indicating that most of the complexity heterogeneity stems from different treatments of melt. For the different model categories, LSMs generally exhibit higher complexity than GHMs and DGVM, largely due to their inclusion of detailed energy balance schemes.

- Contrary to the conventional view that "more complex models are not necessarily better," model complexity shows a significant positive correlation with CTQ performance under high basin complexity ($r = 0.56$, $P < 0.05$), while Q_{sum} and Q_{max} remain largely insensitive to complexity gains. Basin complexity strengthens the complexity–performance linkage and introduces threshold effects, with the strongest correlations observed at PFTh = 0.89 and DEMstd = 226 m. Moreover, from a robustness perspective, CTQ exhibits a strong positive effect ($r = 0.53$, $P < 0.1$), underscoring the irreplaceable value of model complexity in highly complex environments.

- Model performance is shaped less by complexity itself than by the systematic representation of key processes. The absence of key physical processes constrains model performance more severely than process simplification. DBH (high complexity–high robustness model) achieves high robustness through balanced and well-integrated process design, whereas H08 (low complexity–low robustness model) shows poor accuracy due to missing critical modules such as canopy interception. In addition, high complexity alone does not ensure robustness due to uncertainty of parameters (e.g., CLM40), whereas well-designed snowmelt modules enable even simple models to perform resiliently (e.g., PCR-GLOBWB, CWATM).

We revealed how process complexity shapes model performance in snowmelt runoff, though several limitations remain to be refined in future work:

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

- First, as mentioned in **Section 2.3**, our proposed TBMCS method is not designed to offer the absolute score of a model's complexity in this study. Instead, the design is to allow intercomparison among different models - some processes have been simplified if their governing equations or the represented processes have no major differences from other models. Future studies could consider more specific complexity measures if the absolute score of the model is required.

- Second, snowmelt-related processes are currently represented separately, but some processes can be coupled. To maintain simplicity, we did not assign more complex scoring for potentially coupled processes. However, future research could refine this approach to provide a more realistic and accurate absolute score for each model.

- Third, our study lacks a quantitative assessment of how individual processes affect model performance, and has not yet examined the roles of parameter number, process count, or computational forms (e.g., linear vs. nonlinear). Moreover, as our focus was on snowmelt runoff, other key processes were not included. Future work should therefore quantify process-specific complexity and conduct more systematic analyses.

Appendix A

Code and data availability. All data used in this study are available from public repositories: (a) ISIMIP model outputs from https://data.isimip.org/; (b) GSHA (Yin et al., 2024) from https://zenodo.org/records/10433905. The code used in this study is available from the corresponding author upon reasonable request.

Author contributions. Conceptualization: PL, XL, HL. Investigation: XL, HL, PL. Data curation: XL, HL. Funding acquisition: PL. Investigation: XL, HL, PL. Methodology: XL, PL, HL, KZ. Visualization: XL, HL, KZ. Writing (initial): XL, HL, PL. Writing (review and editing): XL, HL, PL, KZ.

Competing interests. The authors declare no conflict of interests.

Acknowledgements. This study was supported by the National Key Research and Development Program of China (2022YFF0801303), the Beijing Nova Program (20230484302), the Beijing Nova Interdisciplinary Program (20240484647), the National Natural Science Foundation of China (42371481), and the Yunnan Provincial Science and Technology Project at Southwest United Graduate School (202302AO370012). The authors acknowledge valuable feedback from ISIMIP modelers Drs. Yusuke Satoh and Emmanouil Grillakis. We also thank Dr. Dashan Wang for insightful discussions related to this project.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

References

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), Model Description – Part 1: Energy and Water Fluxes, Geosci. Model Dev, 4, 677–699, https://doi.org/10.5194/gmd-4-677-2011, 2011.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, J. Hydrometeorol, 16, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1, 2015.

Beven, K.: Prophecy, Reality and Uncertainty in Distributed Hydrological Modelling, Adv. Water Resour, 16, 41–51, https://doi.org/10.1016/0309-1708(93)90028-E, 1993.

Beven, K.: A Manifesto for the Equifinality Thesis, J. Hydrol, 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Burek, P., Satoh, Y., Kahil, T., Tang, T., Greve, P., Smilovic, M., Guillaumot, L., Zhao, F., and Wada, Y.: Development of the Community Water Model (CWatM v1.04) – a High-Resolution Hydrological Model for Global and Regional Assessment of Integrated Water Resources Management, Geosci. Model Dev, 13, 3267–3298, https://doi.org/10.5194/gmd-13-3267-2020, 2020.

Guimberteau, M., Zhu, D., Maignan, F., Huang, Y., Yue, C., Dantec-Nédélec, S., Ottlé, C., Jornet-Puig, A., Bastos, A., Laurent, P., Goll, D., Bowring, S., Chang, J., Guenet, B., Tifafi, M., Peng, S., Krinner, G., Ducharne, A., Wang, F., Wang, T., Wang, X., Wang, Y., Yin, Z., Lauerwald, R., Joetzjer, E., Qiu, C., Kim, H., and Ciais, P.: ORCHIDEE-MICT (v8.4.1), a Land Surface Model for the High Latitudes: Model Description and Validation, Geosci. Model Dev, 11, 121–163, https://doi.org/10.5194/gmd-11-121-2018, 2018.

Guo, H., Hou, Y., Yang, Y., and Mcvicar, T. R.: Global Evaluation of Simulated High and Low Flows from 23 Macroscale Models, J. Hydrometeorol, 25, 425–443, https://doi.org/10.1175/JHM-D-23-0176.1, 2024.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An Integrated Model for the Assessment of Global Water Resources – Part 1: Model Description and Input Meteorological Forcing, Hydrol. Earth Syst. Sci, 12, 1007–1025, https://doi.org/10.5194/hess-12-1007-2008, 2008.

Hou, Y., Guo, H., Yang, Y., and Liu, W.: Global Evaluation of Runoff Simulation from Climate, Hydrological and Land Surface Models, Water Resour. Res, p. e2021WR031817, https://doi.org/10.1029/2021WR031817, 2023.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A Simple Hydrologically Based Model of Land Surface Water and Energy Fluxes for General Circulation Models, J. Geophys. Res.: Atmos, 99, 14 415–14 428, https://doi.org/10.1029/94JD00483, 1994.

Merz, R., Miniussi, A., Basso, S., Petersen, K.-J., and Tarasova, L.: More Complex Is Not Necessarily Better in Large-Scale Hydrological Modeling: A Model Complexity Experiment across the Contiguous United States, Bull. Am. Meteorol. Soc, 103, E1947–E1967, https://doi.org/10.1175/BAMS-D-21-0284.1, 2022.

Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., Gerdener, H., Kynast, E., Peiris, T. A., Schiebener, L., Schumacher, M., and Döll, P.: The Global Water Resources and Use Model WaterGAP v2.2e: Description and Evaluation of Modifications and New Features, Geosci. Model Dev, 17, 8817–8852, https://doi.org/10.5194/gmd-17-8817-2024, 2024.

Müller Schmied, H., Gosling, S. N., Garnsworthy, M., Müller, L., Telteu, C.-E., Ahmed, A. K., Andersen, L. S., Boulange, J., Burek, P., Chang, J., Chen, H., Gudmundsson, L., Grillakis, M., Guillaumot, L., Hanasaki, N., Koutroulis, A., Kumar, R., Leng, G., Liu, J., Liu, X., Menke, I., Mishra, V., Pokhrel, Y., Rakovec, O., Samaniego, L., Satoh, Y., Shah, H. L., Smilovic, M., Stacke, T., Sutanudjaja, E., Thiery,

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

W., Tsilimigkras, A., Wada, Y., Wanders, N., and Yokohata, T.: Graphical Representation of Global Water Models, Geosci. Model Dev, 18, 2409–2425, https://doi.org/10.5194/gmd-18-2409-2025, 2025.

Musselman, K. N., Molotch, N. P., and Brooks, P. D.: Effects of Vegetation on Snow Accumulation and Ablation in a Mid-latitude Sub-alpine Forest, Hydrol. Process, 22, 2767–2776, https://doi.org/10.1002/hyp.7050, 2008.

Ohmura, A.: Physical Basis for the Temperature-Based Melt-Index Method, J. Appl. Meteorol, 40, 753–761, https://doi.org/10.1175/1520-0450(2001)040<0753:PBFTTB>2.0.CO;2, 2001.

Oleson, K. W., Lawrence, D. M., Flanner, M. G., Kluzek, E., Levis, S., Swenson, S. C., Thornton, E., Dai, A., Decker, M., Dickinson, R., Feddema, J., Heald, C. L., Lamarque, J.-F., Niu, G.-Y., Qian, T., Running, S., Sakaguchi, K., Slater, A., Stöckli, R., Wang, A., Yang, L., Zeng, X., and Zeng, X.: Technical Description of Version 4.0 of the Community Land Model (CLM), 2010.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M.: Does Model Performance Improve with Complexity? A Case Study with Three Hydrological Models, J. Hydrol, 523, 147–159, https://doi.org/10.1016/j.jhydrol.2015.01.044, 2015.

Pokhrel, Y. N., Koirala, S., Kanae, S., and Oki, T.: Incorporation of Groundwater Pumping in a Global Land Surface Model with the Representation of Human Impacts, Water Resour. Res, https://doi.org/10.1002/2014WR015602, 2014.

Reed, K. A., Medeiros, B., Jablonowski, C., Simpson, I. R., Voigt, A., and Wing, A. A.: Why Idealized Models Are More Important than Ever in Earth System Science, AGU Adv, 6, e2025AV001 716, https://doi.org/10.1029/2025AV001716, 2025.

Ruelland, D.: Development of the Snow- and Ice-Accounting Routine (SIAR), J. Hydrol, 624, 129 867, https://doi.org/10.1016/j.jhydrol.2023.129867, 2023.

Savenije, H. H. G.: Equifinality, a Blessing in Disguise?, Hydrol. Processes, 15, 2835–2838, https://doi.org/10.1002/hyp.494, 2001.

Schaphoff, S., Von Bloh, W., Rammig, A., Thonicke, K., Biemans, H., Forkel, M., Gerten, D., Heinke, J., Jägermeyr, J., Knauer, J., Langerwisch, F., Lucht, W., Müller, C., Rolinski, S., and Waha, K.: LPJmL4 – a Dynamic Global Vegetation Model with Managed Land – Part 1: Model Description, Geosci. Model Dev, 11, 1343–1375, https://doi.org/10.5194/gmd-11-1343-2018, 2018.

Schoups, G., Van De Giesen, N. C., and Savenije, H. H. G.: Model Complexity Control for Hydrologic Prediction, Water Resour. Res, 44, 2008WR006 836, https://doi.org/10.1029/2008WR006836, 2008.

Sellers, P. J.: Canopy Reflectance, Photosynthesis and Transpiration, Int. J. Remote Sens, 6, 1335–1372, https://doi.org/10.1080/01431168508948283, 1985.

Stacke, T. and Hagemann, S.: HydroPy (v1.0): A New Global Hydrology Model Written in Python, Geosci. Model Dev, 14, 7795–7816, https://doi.org/10.5194/gmd-14-7795-2021, 2021.

Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., Van Der Ent, R. J., De Graaf, I. E. M., Hoch, J. M., De Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: A 5 Arcmin Global Hydrological and Water Resources Model, Geosci. Model Dev, 11, 2429–2453, https://doi.org/10.5194/gmd-11-2429-2018, 2018.

Tang, Q., Oki, T., and Kanae, S.: A Distributed Biosphere Hydrological Model (Dbhm) for Large River Basin, Proc. Hydraul. Eng, 50, 37–42, https://doi.org/10.2208/prohe.50.37, 2006.

Telteu, C.-E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., Boulange, J. E. S., Andersen, L. S., Grillakis, M., Gosling, S. N., Satoh, Y., Rakovec, O., Stacke, T., Chang, J., Wanders, N., Shah, H. L., Trautmann, T., Mao, G., Hanasaki, N., Koutroulis, A., Pokhrel, Y., Samaniego, L., Wada, Y., Mishra, V., Liu, J., Döll, P., Zhao, F., Gädeke, A., Rabin, S. S., and Herz, F.: Understanding Each Other's Models: An Introduction and a Standard Representation of 16 Global Water Models to Support Intercomparison, Improvement, and Communication, Geosci. Model Dev, 14, 3843–3878, https://doi.org/10.5194/gmd-14-3843-2021, 2021.

https://doi.org/10.5194/egusphere-2025-6073
Preprint. Discussion started: 20 January 2026

Valéry, A., Andréassian, V., and Perrin, C.: 'As Simple as Possible but Not Simpler': What Is Useful in a Temperature-Based Snow-Accounting Routine? Part 1 – Comparison of Six Snow Accounting Routines on 380 Catchments, J. Hydrol, 517, 1166–1175, https://doi.org/10.1016/j.jhydrol.2014.04.059, 2014.

Yan, H., Sun, N., Eldardiry, H., Thurber, T. B., Reed, P. M., Malek, K., Gupta, R., Kennedy, D., Swenson, S. C., Hou, Z., Cheng, Y., and Rice, J. S.: Large Ensemble Diagnostic Evaluation of Hydrologic Parameter Uncertainty in the Community Land Model Version 5 (CLM5), J Adv Model Earth Syst, 15, e2022MS003 312, https://doi.org/10.1029/2022MS003312, 2023.

Yin, Z., Lin, P., Riggs, R., Allen, G. H., Lei, X., Zheng, Z., and Cai, S.: A Synthesis of Global Streamflow Characteristics, Hydrometeorology, and Catchment Attributes (GSHA) for Large Sample River-Centric Studies, Earth Syst. Sci. Data, 16, 1559–1587, https://doi.org/10.5194/essd-16-1559-2024, 2024.

Yokohata, T., Kinoshita, T., Sakurai, G., Pokhrel, Y., Ito, A., Okada, M., Satoh, Y., Kato, E., Nitta, T., Fujimori, S., Felfelani, F., Masaki, Y., Iizumi, T., Nishimori, M., Hanasaki, N., Takahashi, K., Yamagata, Y., and Emori, S.: MIROC-INTEG-LAND Version 1: A Global Biogeochemical Land Surface Model with Human Water Management, Crop Growth, and Land-Use Change, Geosci. Model Dev, 13, 4713–4747, https://doi.org/10.5194/gmd-13-4713-2020, 2020.