Reply to reviewer's comments on the preprint egusphere-2025-781:

How well do hydrological models simulate streamflow extremes and drought-to-flood transitions?

Eduardo Muñoz-Castro[1,2,3], Bailey J. Anderson[1,2,3], Paul C. Astagneau[1,2,3], Daniel L. Swain[4,5,6], Pablo A. Mendoza[7,8], Manuela I. Brunner[1,2,3]

[1]WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland
[2]Climate Change, Extremes and Natural Hazards in Alpine Regions Research Center CERC, Davos Dorf, Switzerland
[3]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[4]California Institute for Water Resources, University of California Agriculture and Natural Resources, Davis, CA, USA
[5]Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA, USA
[6]Capacity Center for Climate and Weather Extremes, National Center for Atmospheric Research, Boulder, CO, USA
[7]Civil Engineering Department, Universidad de Chile, Santiago, Chile
[8]Advanced Mining Technology Centre (AMTC), Universidad de Chile, Santiago, Chile

Correspondence to: Eduardo Muñoz-Castro (eduardo.munoz-castro@slf.ch)

Anonymous Referee #1

This manuscript presents a well-structured large-sample hydrology modeling experiment assessing the ability of four conceptual hydrological models (GR4J, GR5J, GR6J, TUW) to capture compound hydrological extremes, with a specific focus on drought-to-flood transitions. In the paper the authors examine the influence of various modeling decisions—model structure, calibration metrics, streamflow transformations, and weights—on model performance across 63 catchments in Chile and Switzerland.

The topic is relevant for the field of hydrology and fills some gaps in our understanding of model behavior under extremes events (drought-to-floods), which are of growing concern in the context of climate change.

Hence, the paper deserves to be published at HESS after some minor corrections.

Thank you for your time and constructive feedback. Your comments have been very helpful and have contributed to improving the quality of our work. We respond to each individual point below. For clarity, *comments are given in black italics*, and our responses are given in plain blue text. Proposed additions are highlighted in red.

General comments

Most of the paragraphs (e.g., L20-L38, L295-310) could benefit from some size reduction, or simply the separation of ideas. Generally speaking, one idea being introduced by paragraph would improve the readability of the text. Currently it is a bit hard to follow the paragraphs due to their size and mix of ideas together.

Thank you for this comment. We have revised each paragraph of the manuscript to shorten its length and improve its clarity.

Three of the four models come from the GRXJ family. This means that model structure diversity is somewhat limited. Could you please justify better this choice in the text? Also pointing out the reasoning of not including another conceptual model structure besides the GRXJs?

Thank you for raising this point. We decided to use the GR model family to evaluate how slight differences in structure can affect the performance of models in detecting extreme events. In L173:L177 in the preprint we motivated this choice by highlighting that "GR4J, GR5J, and GR6J (with 6, 7, and 8 parameters coupled with CemaNeige, respectively) were chosen to explore how slight changes in model structure affect simulated streamflow extremes and the TUW model (with 15 parameters) was chosen to explore how more complex models, particularly with respect to the snow routine and the representation of the processes occurring in the production storage, simulate these phenomena.". To reinforce this point and discuss its limitations, we will bring this up again in Section 5.4 of the revised manuscript, where this choice is discussed in detail as follows:

To support our analysis, we tested four bucket-type hydrological models used within the hydrological modeling community (Addor and Melsen, 2019). Even though these models are at the lower end in terms of model complexity (Hrachowitz and Clark, 2017), and three of them share the same core structure, they allowed us to perform a comprehensive analysis of different model structures at a lower computational cost than when using models with more complex structures (e.g., Clark et al., 2017; Orth et al., 2015; Poncelet et al., 2017). Furthermore, previous studies have also shown that more complexity does not necessarily imply better performance (Figure 7; e.g., Li et al., 2015; Merz et al., 2022).

The paper is dense, but could you somehow summarize better your conclusions in a maximum of three/four bullet points? I see that much can be concluded from your study, but I also think that you could benefit the readers by summarizing the main conclusions in this part rather than everything. Think about what were your hypothesis, and try to come back to them here, for example.

We agree with the sentiment on the manuscript's density. To address this comment, we have re-structured (e.g., subsubsections in subsection 4.2 have been removed) and revised the paragraphs and sentences to limit their length. Based on this, we have revised the conclusions section to emphasize the key messages of our work, which now read as follows:

1. A satisfactory general model performance, as expressed by the KGE, does not guarantee a good performance in terms of detecting streamflow extremes and their transitions. While KGE can serve as a rough proxy for low-flow performance, it cannot for high-flows and drought-to-flood transitions. Consequently, assessments of the suitability of hydrological models for simulating extreme events and their transitions should be complemented with metrics describing extreme event detection performance such as the critical success index (CSI).
2. The most important modeling decision when it comes to simulating floods, droughts, and their transitions is the choice of a suitable model structure. Here, we demonstrate that the GR4J and TUW models have similar performance – with GR4J being slightly better at detecting floods and transitions - and adding model complexity by increasing the number of parameters does not necessarily improve the representation of extreme events.
3. The choice of the exact KGE formulation and the use of weights for their variability term to define the calibration objective function do not substantially affect the simulation of extreme events. However, a joint focus on high and low flows by equally weighting them in the objective function (referred to as HiLo in our analysis) can improve model performance without compromising its ability to capture streamflow extremes.
4. A model's performance in simulating streamflow extremes and transitions primarily depends on how well it captures streamflow timing rather than other hydrological signatures or variables such as evapotranspiration or snow water equivalent.
5. Drought-to-flood transitions are more difficult to capture in semi-arid, high-mountain, and flashy catchments than in humid low-elevation catchments.

Specific comments

Figure 1: It is difficult to distinguish the basin boundaries in both subplots (A and B) of the figure. Maybe if you could reduce the line weight of the country boundaries in A and B, use another color for the basins and increase the figure size of subplots B, C and D.

We agree that the catchments are not clearly distinguishable. Hence, we have changed the color of the catchment's outlet point to red and reduced the line thickness of the international borders, as well as the catchment's boundaries to improve clarity.

L128: I think this section would benefit from this reference:

Clerc-Schwarzenbach, F. M., Selleri, G., Neri, M., Toth, E., van Meerveld, I., and Seibert, J.: HESS Opinions: A few camels or a whole caravan?, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-864, 2024.

In their study they show that most of the time using local information (as you did) can be beneficial for model simulations. If you feel that fits, please consider inserting it.

We appreciate the reviewer's observation, as one of our motivations for using "local" CAMELS databases was indeed based on the evidence presented in the work of Clerc-Schwarzenbach et al. (2024). We have included this reference in a discussion of forcing factors in Section 5.3, which reads as follows:

Here, we attempt to reduce this effect by (1) utilizing local meteorological products over global ones, based on the evidence that these may enhance hydrological modeling (e.g., Clerc-Schwarzenbach et al., 2024), and (2) incorporating adjustment factors to account for potential systematic biases associated with them (e.g., Hughes, 2024; Probst and Mauser, 2022). However, introducing forcing adjustment factors could compensate for some model deficiencies by modifying the inputs (e.g., Tang et al., 2023, 2025). This is somehow reflected by the high dispersion of forcing adjustment factors within each configuration (Figure S16 in the Supplementary Material). Therefore, an improvement in the spatiotemporal representation of precipitation and temperature, as well as of the potential interactions between these variables, could contribute to improved representations of compound streamflow extreme events in hydrological models.

L329-333: I feel that this part should rather be placed in the discussion section.

We agree that this point is interesting for discussion. However, we believe it is important to include it in the results section, as the statement directly results from that analysis. Nevertheless, following the reviewer's recommendation, this message has been reinforced in the Discussion section where it now reads:

Our comparison also highlights that the potential benefit from adjusting these choices (e.g., using other weights or other transformations) varies widely between catchments (Figure 5). This is in line with the findings of Mizukami et al. (2019), who found that the influence of weights on model performance depends on model structure and catchment characteristics. While none of the tested modifications in the objective function consistently improve the simulation of streamflow extremes across all catchments in the study domain, some of the alternative KGE formulations could improve the simulation of certain variables.

Figure 8: The current choice of line colors and types makes it hard to distinguish among the different models. Please consider restructuring it to make it easier for readers.

The colors have been modified to enhance the readability of the figure, and the caption for the figure has been rewritten. Thank you for this suggestion!

Now the figure (and the caption) is as below:

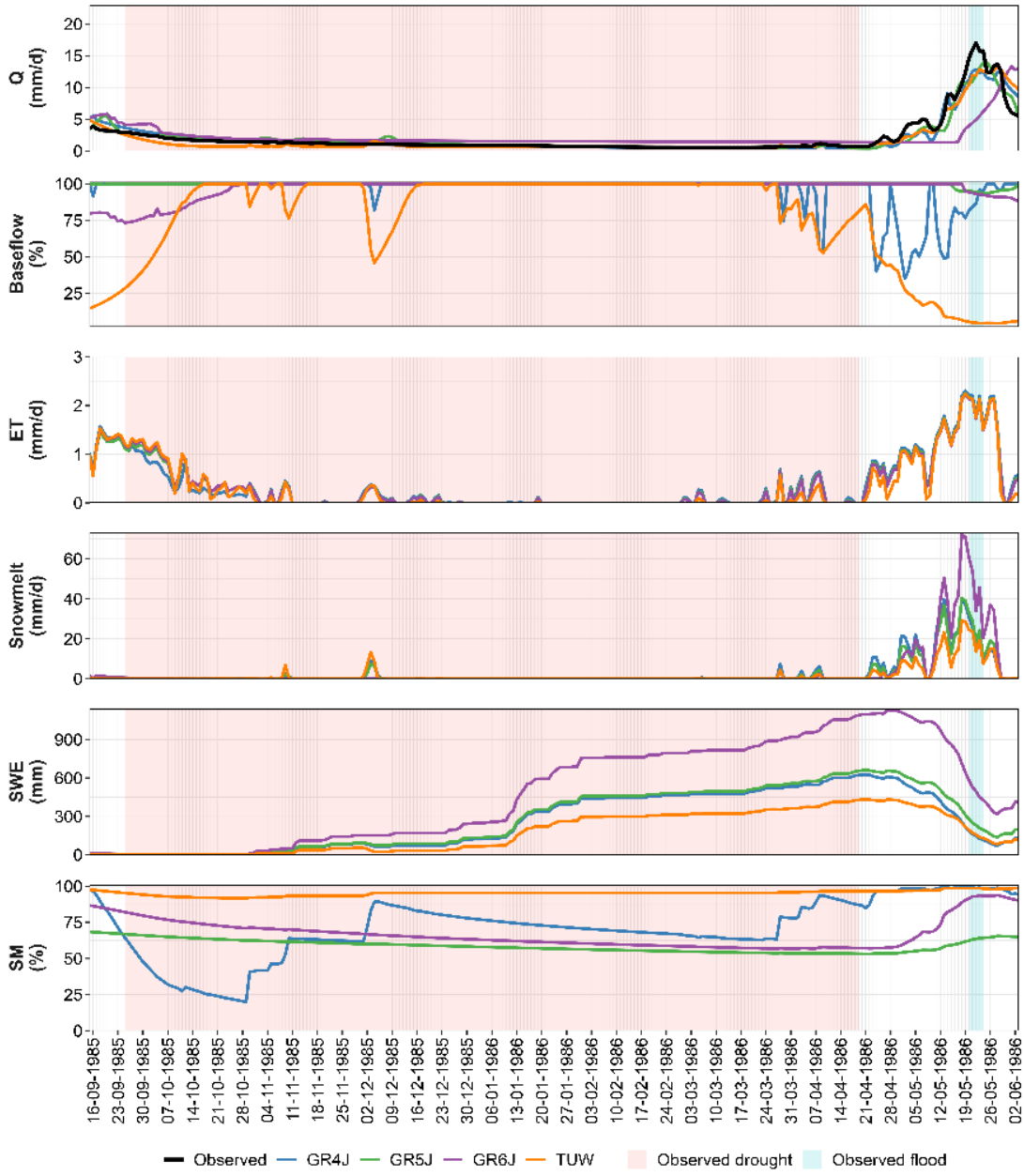

Figure 8. Example of how different hydrological fluxes and states (rows) are simulated for an observed drought-to-flood transition in the Dischma river (Switzerland) with the GR4J, GR5J, GR6J, and TUW hydrological models (colored lines) calibrated with the unweighted HiLo original KGE formulation. The shaded red and blue areas indicate periods of observed streamflow drought and flood conditions, respectively.

Section 4.5: Start by introducing the figure, then you can make your statements. Currently it is a bit confusing the way the section is structured. Also, I see the possibility of having two paragraphs here rather than just one.

The paragraph has been restructured and divided into two, following your recommendation. It now reads as follows:

We explore the relationship between model performance and catchment characteristics using the Spearman's rank correlation coefficient. To this end, we focus on the CSI obtained for the different types of extreme events of interest (droughts, floods, and transitions) generated with the GR4J and TUW models calibrated with the unweighted HiLo original KGE formulation (Figure 10).
Our results show that the model's capability for simulating extreme hydrological events and their transitions depends on catchment characteristics (Figure 10). Drought-to-flood transitions are more difficult to capture in semi-arid (negative correlation between aridity index and CSI), high-mountain (negative correlation between mean elevation and CSI), and flashy (negative correlation between the slope of the flow duration curve and CSI) catchments than in humid low-elevation catchments with high streamflow elasticity to precipitation. This result is generalizable to the other models and different KGE formulations tested (see Figure S14 in the Supplementary Material).

Section 4.6: Again, please start by introducing the figure, then you can make your statements.

To improve the clarity of the sentence and following the reviewer's recommendation, we have rewritten it as follows:

We conduct an ANOVA test to analyze the relative importance of different model parameters in detecting streamflow extremes and their transitions, whose results are presented in Figure 11 (for the extended version with rapid and seasonal transitions, see Figure S15 in the Supplementary Material).

L472-L473: Statement repetition. This idea has already been presented.

Thank you for bringing this to our attention. This statement has been merged with a similar one presented in L459:460 in the pre-print to avoid repetition. The new statement at the beginning of Section 5.3 reads as follows:

Our results show that model structure is the most important modeling decision for capturing extreme events and their transitions (Figure 9), which is consistent with previous studies focused on the independent analysis of extreme events (e.g., Alexander et al., 2023; Melsen and Guse, 2019; van Kempen et al., 2021).

L501: Not Figure 10?

In figure 10, we aim to identify, in which types of catchments the detection of extreme events works best, based on their attributes (comparison between catchment attributes and absolute CSI values). In the context of the paragraph, the idea is to highlight the role of forcing adjustment factors in explaining the variability of the CSI (i.e., the relative importance of the parameter), which may not be explicitly presented in the text. To address this, the sentence is modified as follows:

Given the relative importance shown by the forcing adjustment parameters (Figure 11), the meteorological forcings used to simulate streamflow extreme events can also have a major impact on the model's performance.

L523-L528: This idea has already been presented in the study area. Please consider keeping it just here in the discussion.

Thank you for this suggestion, which we will adopt. This idea will be removed from Section 3.2.1 (L177-L179 in the preprint).

References

Clerc-Schwarzenbach, F., Selleri, G., Neri, M., Toth, E., van Meerveld, I., and Seibert, J.: Large-sample hydrology – a few camels or a whole caravan?, Hydrology and Earth System Sciences, 28, 4219–4237, https://doi.org/10.5194/hess-28-4219-2024, publisher: Copernicus GmbH, 2024.