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], Pablo A. Mendoza[6,7], 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]Weather Extremes Across Scales, NSF National Center for Atmospheric Research, Boulder, CO, USA
[6]Civil Engineering Department, Universidad de Chile, Santiago, Chile
[7]Advanced Mining Technology Centre (AMTC), Universidad de Chile, Santiago, Chile

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

Referee #2 – Dr. Wouter Knoben

Thank you very much for your constructive feedback. Your comments have been very helpful and have contributed to improving the quality of our work. First, we address the comments the reviewer considers most important, included in the online discussion, and then the detailed comments provided directly in the PDF of our manuscript. For clarity, *comments are given in black italics*, and our responses are given in plain blue text. Proposed additions are highlighted in red.

Comments online

The authors have submitted a revised version of their manuscript, which, again, seems to have involved a considerable amount of work. In fact, I believe the response document undersells the changes made here.

Thank you very much for acknowledging that. We also greatly appreciate your time and willingness to review our manuscript a third time, as well as your valuable feedback.

The authors state that their main changes are (1) the addition of p-seasonality definition and (2) the addition of analysis that investigates the forcing adjustment parameters. However, unless I'm seriously misunderstanding something, the authors also re-ran the calibration of all cases with a different calibration algorithm in response to my questions/concerns about the large differences in performance between GR4J on the one hand, and GR5J and GR6J on the other. This seems to have substantially increased the performance of both models and brought GR5J/GR6J performance more in line with that of the other two, effectively addressing the main concern I still had in the previous version of this paper.

Indeed, we recalibrated the hydrological models using different calibration algorithms, and for GRXJ, we applied transformations to the parameters to improve the exploration of the parameter space. These results were used for comparison with the simulations presented in the initial versions of the manuscript. Based on these findings, we decided to use the results obtained with DDS, as they offered comparable outcomes across hydrological models.

The text in the manuscript has changed substantially compared to the last version, and I have added some additional comments in the annotated PDF. These are merely suggestions, as well as a few technical corrections, and I'm happy for the authors to decide what (if anything) to implement. I expect this would

take a few hours at most, and I don't think there is anything that would require another round of formal reviews.

We appreciate the recommendations you made in the PDF, which we have taken into consideration. While further details are provided later for each comment, the changes made to the manuscript are summarized below:

1) Following your recommendation, we have included some thoughts on the weaknesses of models, given the complexity of representing hydrological processes associated with transitions.
2) We have addressed the points regarding the low dispersion of KGE values for TUW compared to GRXJ but with similar CSI values, the loss of performance in capturing floods when weights are used in the KGE variability term, and other points highlighted in your review.
3) We have included the comparison between simulations including forcing adjustments and those without (Figure 3 in the responses during the second round of review) in the SI. Thus, we now refer to this figure instead of saying "not shown".
4) Typing errors, empty spaces, references to figures in the SI, etc., were reviewed and adjusted.

I would like to thank the authors for addressing my comments with the thoroughness shown here. I think this is a very nice piece of work, and a good example of how to deal with the complexity of contemporary modeling studies.

We are grateful for the recognition of our work, as well as for the valuable feedback you provided during the review process. The recommendations have undoubtedly allowed us to strengthen our study.

Comments on the PDF

L13: "compared to"?

Thank you for the recommendation. We have adjusted the sentence as follows:

Overall, we find that model representation of drought-to-flood transitions is generally poor, especially in semi-arid and high-mountain catchments compared to humid low-elevation catchments

L182-L183: "Bucket-type conceptual models generally include parameters and functions that allow for non-conservative adjustments to the water balance (i.e., artificially adding or leaking water)." Minor nitpick, based on my own experience I would say that most models don't include such parameters/functions but it's not exactly uncommon either. "may", "sometimes", and "occasionally" would all be more appropriate than "generally" in my opinion.

Thank you for raising this point. We agree that "generally" can be a strong term, so we have opted to use "often" instead. Now the complete sentence reads as follows:

Bucket-type conceptual models often include parameters and functions that allow for non-conservative adjustments to the water balance (i.e., artificially adding or leaking water).

L227: It might be worth explicitly mentioning that DDS was selected based on the comparison the authors ran in this second set of revisions, possibly supported by some plots in the SI taken from the current response document. That both provides some support for this choice, and also acknowledges the depth of the analysis performed (which goes beyond what is typical).

We appreciate your recommendation. However, given one of the initial suggestions during the review (i.e., reducing complexity), we believe that including this analysis would add an additional layer to the study. To limit complexity, we decided to exclude it from the study and keep it as an internal exercise that helped us better understand our results.

L290: The reasoning for removing catchments from this isn't fully clear to me. Based on the results in the response document, catchment choice is almost fully controlling in the case of detecting droughts, but hardly influential for detecting floods. If so, that might be a useful finding in its own right.

Upon further reading I notice that this will be covered later, but it might be worth noting in the text here already why catchments aren't part of the ANOVA.

We have improved the explanation regarding this decision and incorporated it into the methodological description, which now reads:

We also analyze the relative importance of the differences between catchments by including them in the ANOVA test. However, we ultimately removed this component from the explanatory variables because its influence sometimes dominated the results, thereby hiding the contribution of the intrinsic modeling decisions being tested to the variability observed in the CSI values.

We have also indicated how the results change when differences between catchments are included in the ANOVA. This is shown in the results section (see response in one of the following comments).

L338: Perhaps it's worth explicitly noting too that TUW appears to have a modestly more narrow and higher distribution of KGE scores than the GRXJ models, but that the CSI ranges seem more or less the same. Unsure if helpful, but it stood out to me so I thought I'd bring it up

Thank you for the recommendation. This is an interesting point we had not explicitly highlighted in previous versions. However, it does not provide additional information beyond what has already been presented in the manuscript to support the statement "a good KGE does not imply good performance in detecting streamflow extreme events". As the focus here is not on comparing models, we have therefore decided not to include this point directly in the manuscript.

L343: Space missing here

We have corrected this in the updated version.

L345: It might be good to add in the caption of Figure S10 the same explanation of how to interpret positive/negative values on the y-axis that is given in the caption here. It took me a little while to recognize that positive y-axis values indicate worse performance compared to reference, and that Fig S10 supports that highlighted statement even more than Fig 5 does.

Thanks for the comments. In the previous version, this information was already included in the caption of Fig. S10, but at the end of the text. To make it easier to read, we have moved this point to the beginning of the caption. It now reads as follows:

Figure S10: Difference in the CSI by using the no weights HiLo case (reference) and different weights (alternative) on the variability term of the KGE for different hydrological models. Differences are calculated as "reference - alternative" with values above (below) 0 indicating better (worse) performance of the reference (alternative). Difference in the Critical Success Index (CSI) for simulations using model calibrations with no weights and the HiLo transformation (reference) versus different weights and streamflow transformations (alternative) for a) droughts, b) floods, and c) transitions. Each alternative is

compared with its unweighted analogs and HiLo transformation. Each boxplot contains 315 values (63 catchments x 5 KGE formulations). Supplementary figure associated with Figure 5 in the main manuscript.

As an aside, perhaps it is interesting to note as well that it seems in particular for the flood case that changing the variability weights leads to substantially worse flood detection (at least for the GRXJ models). This seems unexpected, and it might be interesting follow-up work to identify why this is (does it have something to do with the models, the definition of floods as used here, etc).

Thank you for pointing this out. Yes, it is indeed something worth mentioning. To address this suggestion, we have included the following sentence:

Further, weighting the variability term can substantially worsen flood detection (e.g., GRXJ models, Figure S10).

Fig. 6. There seem to be two cases here with perfect CSI values for "all transitions", but no corresponding perfect CSI entries for droughts and floods. Based on Fig S11 this seems to be a basin in Chile and happens for two different KGE formulations. Is this expected/explainable, or are we looking at some analysis or plotting error?

This can be explained by differences in the number of events per streamflow extreme event (more events suggest more chances to capture or not capture the events). Similarly, there are catchments where no events can be captured (i.e., CSI = 0).

Caption Fig. 6. I think it would be good to add a few words to explain what is being tested here. I cannot immediately recall what the methodology section says about this

We have included more information about the test used in the methodological description. It now reads:

To assess the statistical significance of the differences between, e.g., the ability to capture streamflow extreme events across models, as well as other configurations tested in this study, we applied the Wilcoxon test (Wilcoxon, 1945) at a 5% significance level and provided p-values where possible. The Wilcoxon test is a nonparametric test used to determine whether two groups differ statistically, without making any specific assumptions about their distributions (e.g., normality).

L371: Might be worth harmonizing the text a bit here, because in lines 351-352 it is stated that there are no significant differences between models when it comes to detecting extremes/transitions.

To improve clarity, we have rewritten the text as follows:

Our previous results showed no significant differences when pooling results by model (Figure 6). However, when it comes to the relative importance in explaining the total variance of the detection skill, the results of the ANOVA show that the most important modeling decision in simulating extreme events and their transitions is the choice of a suitable model structure, followed by the choice of the streamflow transformation (Figure 8).

L379: Is this fully accurate? It definitely seems to be the case for droughts and presumably as a consequence of that it is also dominant for transitions, but for floods the impact is less extreme. I'm not fully sure what this implies but it might be worth pointing out, possibly in relation to the clear difference in CSI values between floods and droughts (Fig 6).

To improve the accuracy of the statement, we have included the following sentence:

Additionally, when catchment characteristics are included as an explanatory variable, they strongly influence drought detection, while they have little effect on flood detection (see Figure S13).

L379: "S13"

Thank you for pointing that out. We have corrected it in the updated version.

L383: Might be worth adding a few words to explain why this uses just GR4J and TUW - is this a leftover from the previous results where GR5J and GR6J performed rather poorly?

We consider only these two models as examples because the results are similar between them. We indicated this in the previous version of the manuscript in L387-L388, where we mentioned: "This result is generalizable to the other models and the different KGE formulations tested (see Figure S14 in the Supplementary Material)." To avoid misunderstandings, we have included a reference to the figure in the SI with all models. Now the sentence reads:

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 9; extended version including all models in Figure S14).

L406-L408: Is this (somehow) related to the models being run at a daily time step? Given that the observations are at a daily scale too I'm not fully sure if/how that would matter - is there any existing work out there that can provide helpful context about models' ability to predict flood peaks at the daily time scale?

While we agree that this is a very interesting topic to explore, including additional ideas about it in the manuscript, beyond those already presented, could distract us from the point we want to converge on. However, part of this is mentioned as one of the aspects to be explored in future work.

L411: This is too general. It seems difficult to capture in these four models, but that does not necessarily generalize to all conceptual models

We agree with what you have mentioned. To make the statement more accurate, we have rewritten it as follows:

Overall, our analyses highlight that these fast processes are rather difficult to capture in conceptual rainfall-runoff models like GRXJ and TUW.

L414: I feel obliged to point out that there's nothing special about KGE = 0.6, and I'm being pedantic about this because these arbitrary choices sometimes start leading lives of their own in later work. I understand that here it's simply meant as "models with relatively higher KGE scores, e.g. KGE > 0.60, still can get poor CSI scores) but it is distinctly possible that in some unrelated later paper a phrase like "Model performance of KGE > 0.60 is considered good general performance (Muñoz-Castro et al., 2026)" would appear.

Mentioning this threshold does not do much in this current sentence, so perhaps the cleaner phrasing is to simply state that there is a strong relation between low KGE and low CSI (i.e. you won't get good event detection out of an inaccurate model) but that even at the highest KGE scores seen in this work the CSI values can cover a wide range (i.e., you might also might not get good event detection out of a relatively accurate model).

We agree with the concern you raised, which is one of the key messages in our study (i.e., going beyond a good KGE, as this does not ensure that events of interest are well detected). To avoid misleading readers, we have rewritten the sentence as follows:

Even models with high accuracy, measured by traditional metrics such as KGE, struggle to capture extreme events, particularly floods and transitions from drought to flood (Figure 4)

L426: Might be good to be a bit more precise here. For floods, Fig 8 shows transformations as being the most important decision

Thank you for the recommendation. We have included the following sentence in the text:

However, choosing an appropriate transformation can be an important decision for improving the models' ability to capture flood events.

L428-L429: I think it would be really helpful to relate this finding a bit more closely to the previous sentence. It seems that on the one hand (ANOVA), model structure stands out as a key decision but on the other hand (statistical tests) there are no obvious differences between the models. How does this work?

These differences arise from the approaches used. When we state that there are no significant differences between the models, we are comparing the distributions of results aggregated by model (i.e., without distinguishing between catchments). In contrast, the variance analysis (ANOVA) incorporates, to some extent, catchment-specific information. In the revised version of the document (as noted in a previous response), we explicitly clarify these methodological differences and their impact on the results to avoid potential confusion for readers.

L432-L434: It could be helpful to add some field-/process-based literature that describes the real-world processes that are important under these conditions and that ought to be in models. E.g., changes to soil proprieties/structure due to prolonged dry conditions can substantially change infiltration rates and the amount of surface ponding/runoff, but such process are, to the best of my knowledge, hardly ever found in hydrologic models. There may be other such processes, and adding this connection with field studies would add some extra weight to these statements I think

Thank you very much for the recommendation. We have included the following in the text:

From a process perspective, hydrological model underperformance can be linked to oversimplified or poorly represented (or understood) processes (Beven, 2019; Clark et al., 2017; Hrachowitz et al., 2014; McMillan et al., 2018). For instance, in the context of drought-to-flood transitions, prolonged dry conditions can alter soil properties, such as cracking (Gimbel et al., 2016; dos Santos et al., 2016), water repellency (Doerr et al., 2007; Leighton-Boyce et al., 2007), and macropore connectivity (Or et al., 2013), changing the infiltration-runoff partitioning and potentially intensifying catchment responses to precipitation. Despite their importance for flood generation, these soil and near-surface processes remain poorly understood and, consequently, are rarely represented in conceptual or even physically-based hydrological models (Barendrecht et al., 2024; Blöschl et al., 2019; Brunner, 2023). This limits their ability to reproduce streamflow extremes and rapid shifts between opposing extremes. Therefore, it is also important to improve our understanding of the processes behind transitions and how they are represented in hydrological models.

L473: Might be good to add to the SI instead, if the results are in a publishable state

The figure we used to reach this conclusion was Fig. 3 from the second-round response document. We have decided to include this figure in the SI to refer to it directly in the text (now Figure S19).

L527-L529: Just pointing out there here too the "model structure is most important - models have similar performance" sentence order might benefit from some words explaining how both are valid at the same time

As mentioned above, these differences stem from the approaches used. When we state that there are no significant differences between the models, we are comparing the distributions of results aggregated by model (i.e., without distinguishing between catchments). In contrast, the variance analysis (ANOVA) incorporates, to some extent, catchment-specific information. To improve clarity, it now reads:

The most important modeling decision when it comes to simulating floods, droughts, and their transitions is the choice of a suitable model structure. However, in a large-sample context, we demonstrate that the four models tested here (i.e., GR4J, GR5J, GR6J, and TUW) have similar performance, showing that adding more parameters does not necessarily improve the representation of extreme events.

L531: Perhaps pointing out though that (as Fig 8 shows) it still plays a large role for floods specifically, and in general doesn't seem like an aspect that should be ignored

Thanks for the recommendation. Now it reads:

In contrast, despite it still playing a large role in floods, the choice of the objective function and its exact configuration are, overall, less important.

References

Barendrecht, M. H., Matanó, A., Mendoza, H., Weesie, R., Rohse, M., Koehler, J., de Ruiter, M., Garcia, M., Mazzoleni, M., Aerts, J. C. J. H., Ward, P. J., Di Baldassarre, G., Day, R., and Van Loon, A. F.: Exploring drought-to-flood interactions and dynamics: A global case review, WIREs Water, n/a, e1726, https://doi.org/10.1002/wat2.1726, 2024.

Beven, K.: How to make advances in hydrological modelling, Hydrology Research, 50, 1481–1494, https://doi.org/10.2166/nh.2019.134, 2019.

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K., and Živković, N.: Changing climate both increases and decreases European river floods, Nature, 573, 108–111, https://doi.org/10.1038/s41586-019-1495-6, 2019.

Brunner, M. I.: Floods and droughts: a multivariate perspective, Hydrology and Earth System Sciences, 27, 2479–2497, https://doi.org/10.5194/hess-27-2479-2023, 2023.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrology and Earth System Sciences, 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Doerr, S. H., Ritsema, C. J., Dekker, L. W., Scott, D. F., and Carter, D.: Water repellence of soils: new insights and emerging research needs, Hydrological Processes, 21, 2223–2228, https://doi.org/10.1002/hyp.6762, 2007.

Gimbel, K. F., Puhlmann, H., and Weiler, M.: Does drought alter hydrological functions in forest soils?, Hydrology and Earth System Sciences, 20, 1301–1317, https://doi.org/10.5194/hess-20-1301-2016, 2016.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resources Research, 50, 7445–7469, https://doi.org/10.1002/2014WR015484, 2014.

Leighton-Boyce, G., Doerr, S. H., Shakesby, R. A., and Walsh, R. P. D.: Quantifying the impact of soil water repellency on overland flow generation and erosion: a new approach using rainfall simulation and wetting agent on in situ soil, Hydrological Processes, 21, 2337–2345, https://doi.org/10.1002/hyp.6744, 2007.

McMillan, H. K., Westerberg, I. K., and Krueger, T.: Hydrological data uncertainty and its implications, WIREs Water, 5, e1319, https://doi.org/10.1002/wat2.1319, 2018.

Or, D., Lehmann, P., Shahraeeni, E., and Shokri, N.: Advances in Soil Evaporation Physics—A Review, Vadose Zone Journal, 12, vzj2012.0163, https://doi.org/10.2136/vzj2012.0163, 2013.

dos Santos, J. C. N., de Andrade, E. M., Guerreiro, M. J. S., Medeiros, P. H. A., de Queiroz Palácio, H. A., and de Araújo Neto, J. R.: Effect of dry spells and soil cracking on runoff generation in a semiarid micro watershed under land use change, Journal of Hydrology, 541, 1057–1066, https://doi.org/10.1016/j.jhydrol.2016.08.016, 2016.