Reply to editor'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)

Editor – Dr. Fabrizio Fenicia

Thank you very much for your time and for appreciating the value of our contribution. 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.

I concur with the reviewer's assessment, which acknowledges the authors' substantial revisions and improvements in clarity but also raises key methodological concerns regarding model evaluation, parameter behavior, and the robustness of the main conclusion that model structure dominates performance. In particular, the reviewer recommends an additional ANOVA analysis to verify this claim. Given these substantial and technically significant issues, the manuscript is returned to the authors for revisions before further consideration for publication.

We have addressed the reviewer's comments, and the changes implemented based on their feedback are summarized below:

1) The calibration of the hydrological models has been reviewed and updated.
2) The approach for model evaluation has been improved by considering variables such as SWE, ET, and SM, in addition to daily runoff.
3) We have conducted additional experiments to better understand the influence of forcing adjustment parameters regarding (i) the identifiability of parameters, and (ii) the hydrological behavior of the models. Although this is beyond the main scope of the manuscript, some of the results provided in the responses have been incorporated into the supplementary material, and some responses have been included in the main text.
4) The ANOVA has been run using different combinations of two models to test the influence of model subset choice on the results associated with the relative importance of the modeling decisions tested.

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

Let me start by saying that I appreciate the thorough response documents and the authors' willingness to make revisions to their paper. The paper has definitely become shorter and easier to read, but a lot of extra information is available in the SI for those readers who want it.

Thank you very much for acknowledging that. We also greatly appreciate your time and willingness to review our manuscript a second time and your valuable feedback. In the responses presented below, we seek to find a good compromise between the comments raised in the first round of review, where one of the most emphasized points was reducing the complexity of the document, and the current round.

I am mostly satisfied with the responses to my earlier comments (note: I still don't really like the choice of using NSE as you will see from my comments in the pdf, but I don't think this is a sticking point for publication), but have a few things I think are worth bringing up. This is all based on the 20 or so comments in the annotated PDF, I'm just trying to summarize them for convenience.

We very much appreciate your recommendations, which we have taken into consideration and adopted to clarify the message of our manuscript. We have compiled all the observations and comments included in the revised manuscript's PDF you have uploaded and responded to each one in this document. The changes implemented based on your feedback are summarized below:

1) We analyzed the definition of p-seasonality and improved its description in the manuscript.
2) We reviewed the influence of the adjustment parameters on the results.

[1] The methodology text in the main manuscript (Section 3.3.1) mentions the evaluation of all models on auxiliary variables (ET, SWE, SM) but directs the reader to the SI to see this.

We have updated the manuscript to emphasize the evaluation focused on event detection. To do so, we restructured Section 3.3 and placed greater emphasis on the CSI analysis and its relationship with KGE. We still included a few sentences on goodness-of-fit metrics and hydrological signatures to discuss the overall performance of the models, and the reader is referred to the supplementary material for further details.

I have a few comments about this:

- First, this sets a baseline for the reader's understanding of how well these models were calibrated at all but most readers won't see this. Including at least some text in the results section about this would be welcome, not in the least to highlight the stark differences in the performance of GR4J and TUW on the one hand, and GR5J and GR6J on the other.

In the first round of revisions, we removed some of these analyses to reduce complexity and focus the message on event detection problems, despite achieving acceptable KGE results. We agree with you that this information is important for the reader to understand what to expect from the different models. Therefore, we have added a paragraph at the beginning of the results section summarizing the model's overall performance in simulating streamflow and the other variables available for analysis (i.e., SWE and ET). We can summarize these results as follow:

1) All configurations perform better than a simple daily mean flow benchmark (i.e., the mean flow for each day across all instances over the calibration period).
2) By analyzing the differences in directional statistics, we showed that the seasonality of variables such as streamflow, snow water equivalent (SWE), actual evapotranspiration (ET) are simulated accurately, with median performance values across catchments and configurations between 0.79-0.98 (with one being the optimum).
3) Our evaluation shows that using weights greater than two can be detrimental to the overall performance of the model, both in terms of representing the seasonality of the aforementioned variables and some hydrological signatures.
4) Considering those configurations with comparable performances (i.e., removing weights greater than 2), average accuracy across configurations ranges between 0.87-0.92, 0.88-0.93, and 0.75-0.85, for, e.g., the high-, mid-, and low-segment of the slope in the FDC respectively

- Second, it seems to me that there aren't really any consequences of this extra evaluation. At least, there does not seem to have been a step that removes models from further analysis because their performance was deemed unsatisfactory. The main thing I've been able to find are two statements in Text S2 that essentially boil down to NSE is above 0 (not a particularly high bar in most cases) and if it isn't at least the correlation is higher than 0.5 (also generally not a high bar). Is this really enough to warrant keeping all models for the remainder of the analysis?

To improve the analysis and focus on the characteristics we want to evaluate, we have decided to use directional statistics (replacing NSE and r) to quantify biases in the representation of seasonality for variables such as Q, SWE, and ET. However, removing models due to poor performance in simulating variables that were not used for calibration would not fair, especially considering the end goals of the parameter estimation process (which is oriented for streamflow event detection). Nevertheless, we have improved the model performance analysis to ensure comparability between simulations. From this analysis, we have concluded that weights greater than two consistently degrade the performance of flow simulations across models. Consequently, we have completely removed such cases from the analysis.

- Last, it seems a bit strange to me to recommend to the reader in the discussion (l483-484) to use auxiliary data for various things, while this paper seems to have had this auxiliary data available but is not particularly clear if these three things were actually done and if so how.

We consider the ancillary data as part of our model evaluation to perform "sanity checks" on hydrological variables other than runoff (e.g., poorly represented ET or SWE seasonality), i.e., (i) complement model assessment, in the manuscript. Although points (ii) and (iii) have not been developed, we believe that this could be a way to improve the performance of hydrological models. Then, we have rephrased the sentence as follows:

Based on the evaluation of the models' performance, we were able to verify that, despite the dispersion of optimal parameters, the simulations are consistent with the products used to evaluate the models (Figures S3 and S6). To (i) complement model assessment, (ii) better define the parameter exploration range, and (iii) lead to parameter sets that ensure reliability and fidelity in representing hydrological processes, hydrological variables other than streamflow, such as SWE or ET, can provide useful information to improve hydrological modeling.

[2] I very much appreciate the clarification about the differences between the GRXJ versions, and being set straight about the GR4J vs GR5J comparison. However, the differences between the GRXJ versions still seem really quite large to me. The extra parameter plot (Fig S10) seems to suggest that for all calibration cases, the optimized values of X2 and X5 for both GR5J and GR6J are centered very strongly around the middle of the normalized range. Based on the table, that corresponds to X2 = 0 and X5 = 0. Flipping back to the response document, this seems to suggest that for both of these models the water import flux is essentially 0 and that the X5 threshold doesn't really do anything. For GR4J, the optimized X2 values have a very different distribution (and really, none of the other parameters have such narrow ranges at such specific 'no effect' values as these two in those two models). This just seems weird to me, and if it were my results I would really like to have some more confidence that this is actually a real outcome rather than some sort of bug or issue with calibration convergence to a local minimum.

We have conducted a comprehensive analysis of the benefits of including all parameters in the calibration process due to their interactions. To do so, we first calibrated the GR4J model by canceling the exchange parameter (X2) and transferred the calibrated parameters to the GR5J and GR6J structures. Second, we calibrated the GR4J, GR5J, and GR6J models, canceling the effect of the exchange parameters (i.e., X2 and X5) and exponential storage (i.e., X6). Third, we calibrate the GRXJ models without applying any restrictions to the parameters (i.e., allowing the calibration of X2, X5, and X6). Finally, these results are compared with each other to understand whether the GRXJ models are intrinsically different (or not) and to evaluate the benefit of calibrating the parameters X2, X5, and X6. The summarized results are as follows:

1) We have proved that despite having calibrated models with X2 and X5 parameters forced to be equal to zero, or X6 fixed at its minimum value (i.e., canceling their effects), they remain different when comparing simulation with the same models run using parameters transferred from GR4J with X2=0 (see Figure 5 and 6 in the responses document). This shows that GRXJ models are not identical, despite sharing the same core structure and development philosophy.
2) By allowing calibration of all parameters, independent of their value, which can be close to zero in the end, improvements are always obtained compared to the version with X2=X5=0 and X6 = 0.01 (i.e., cancelling their effect; see Figure 8 in the responses document). These improvements are observed for both the KGE and its components. In summary, there is a benefit to using the exchange function and the exponential storage (i.e., allow the calibration of X2, X5, and X6 depending on the model) compared to canceling them.

Additionally, we have found that the differences in performance between the GR5J and GR6J models and GR4J per catchment are close to zero when the internal calibration approach for GRXJ models is implemented. This finding has motivated us to update our calibration (and evaluation) approach to ensure comparable performance across the different models tested. We now calibrate our model with a more agile algorithm (DDS), ensuring that the performance of the calibrated models according to the different objective functions are comparable to each other (i.e., that there are no significant differences between them), and verifying that, at a minimum, the seasonality of variables other than streamflow, such as snow water equivalent (SWE) and actual evapotranspiration (ET), are fairly captured.

[3] In a practical sense, and strongly related to the evaluation comments before, I really wonder if it makes sense to retain all models in all basins for the full analysis. For example, the ANOVA analysis shows that model structure is the dominant factor (and this is echoed in the conclusions), but to what extent is that just determined by GR5J and GR6J being so much worse than the other two models? One quick way to test this might be to repeat the ANOVA but without the GR5J and GR6J results, and see to what extent the model structure factor remains important.

As we mentioned before, we have updated our calibration approach – and its assessment – to ensure comparable performance across the different models tested. Then, we carried out the experiment mentioned, which showed that the relative importance of methodological decisions is similar across combinations of hydrological models. In addition, we have included this analysis in the Supplementary Material.

Comments on the PDF

L54: The "suggests" that was here before seems more appropriate to me because it's not clear to me how either of the two examples mentioned before (peaks & volumes, spatial coherence) directly indicate that a drought-to-flood even would be difficult to model.

We have rewritten the sentence as follows:

This complexity suggests that capturing consecutive drought-to-flood events might not be trivial either. As model evaluations targeted at compound extremes have not yet been performed, it is still unclear how well hydrological models can, in fact, capture drought-to-flood transitions.

Caption Figure 1: The response document states that this approach follows Berghuijs et al (2014) which in turn follows Woods (2009) for "how in phase (or out of phase) the seasonal patterns of precipitation/streamflow and temperature are".

I interpret this as referring to the δ_p^* parameter (Eq 5 in Berghuijs; Eq 14 in Woods). E.g., Berghuijs: "The dimensionless variable [\delta_p^*] describes the seasonality of precipitation and whether or not the precipitation is in phase with the potential evaporation and temperature regimes".

However, Berghuijs (Section 3.2.1, Eq. 5) states, and from Woods the same can be derived (Section 2.6, Eq 14), that the seasonality parameter (\delta^*_p) in their equation has a range of [-1,1]. It's still not clear to me why in this figure the values go below -1.

Is this perhaps based on Berghuijs and Woods (2015) instead? This uses truncated sine curves to avoid physically unrealistic negative precip values which could result in values < -1.

Different studies have shown p-seasonality values greater (lower) than 1 (-1). For instance, Vásquez et al. (2024b, a) showed values ranging from -1.2 to +1.2, similarly to Álvarez-Garretón et al. (2018). The formula we used to calculate p-seasonality has also been used in the CAMELS dataset (Addor et al., 2017) and reads as follows:

$$\delta_P{}^* = \delta_P \cdot sgn(\Delta_T) \cdot \cos(2\pi \ (s_P - s_T)/\tau)$$

$$P(t) = \bar{P}[1 + \delta_P \sin(2\pi\ (t - s_P)/\tau_P)]$$

Without going into detail about the terms contained in each function, we can see that both the cos(), sin(), and sgn() functions are bounded below and above by -1 and 1, respectively. To obtain ${\delta_P}^*$ values greater (lower) than 1 (-1), the only possibility is for δ_P to exceed those limits. However, if $|\delta_P| > 1$, $P(t)$ could take negative values, which is physically impossible.

Looking at the R function that we used to calculate this index, we noticed that the function adjustment is not limited to values between -1 and 1 to ensure physical consistency. To fix issue, we have modified the function to restrict the adjustment associated with δ_P. Figure 1 shows how this restriction affects the p-seasonality and q-seasonality values presented in the previous version of the manuscript and the updated ones. Additionally, the fraction of precipitation falling as snow is included since, according to the formulation proposed by Woods (2009), this variable also depends on ${\delta_P}^*$ (p-seasonality). We conclude that applying the constraint to ensure physical consistency in the estimation of ${\delta_P}^*$ (i.e., $-1 \leq \delta_P \leq 1$) does not impact the signal reported by the metric, nor does it affect the estimation of the fraction of precipitation falling as snow. Still, we have updated those figures (i.e., Figure 2) where p-seasonality and q-seasonality are presented to the constrained version of the index, to be consistent with the physical definition of the variables, and we have highlighted this point in the text (see the answer to the following comment).

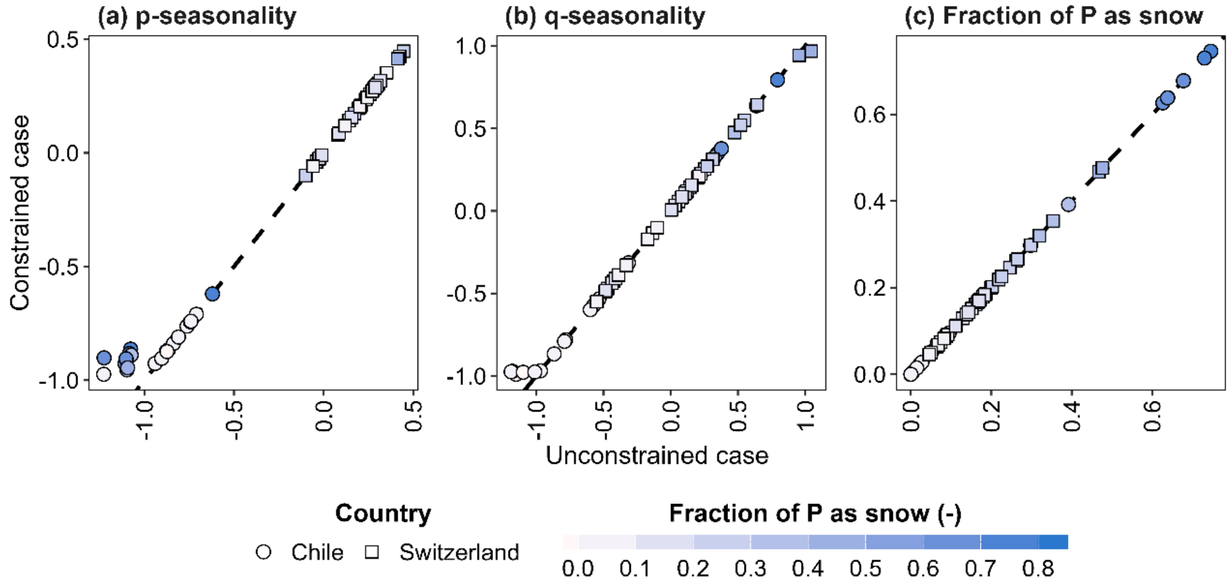

Figure 1: Comparison of estimates of (a) p-seasonality, (b) q-seasonality, and (c) fraction of precipitation falling as snow, when constraining versus not constraining the search range to ensure physical consistency in determining these indices.

In either case, I think the current description is sufficiently unclear that some extra words (as well as citations of the underlying references) are needed. Equations can go in the SI if needed but should be added.

Thank you for pointing this out. We have improved the description of the indices and included the corresponding citation. Given that the equations are well documented, we do not consider it necessary to include them in the SI. Now, the text reads as follows:

The study domain encompasses 24 and 39 near-natural catchments in Chile (CL; Figure 1a) and Switzerland (CH; Figure 1b), respectively. These catchments are selected based on the availability of complete daily streamflow records between 1981 and 2020 for at least 30 years, considering that a year is complete if all months had information for at least 90% of the days. We characterize the hydroclimatology of the catchments in our study domain by the wetness index (P/PET), the runoff coefficient (Q/P), the p-seasonality and q-seasonality indices, and fraction of precipitation falling as snow (fsnow) over the period 1985-2020. The p-seasonality index (Woods, 2009; Berghuijs et al., 2014), as well as its analogue, q-seasonality, describes the seasonality of precipitation (or streamflow) and the degree of synchronization with the seasonality of temperature. The fsnow is computed according to the formulation proposed by Woods (2009) and ranges from 0 (all precipitation falls as rain) to 1 (all precipitation falls as snow).

L122: I could imagine that processes like mountain block recharge and glacier melt play a role too. Can something be said about this?

We updated the sentence as follows to also include processes like recharge and glacier melt:

Some catchments are positioned above the water limit (i.e., Q/P = 1) or below the energy limit (i.e., Q/P = 1 - 1/(P/PET); Figure 1c), which suggests an underestimation of precipitation - which might require correcting for precipitation undercatch (e.g., Newman et al., 2015; Stisen et al., 2012; Hughes et al., 2021) - or a surplus of streamflow due to, e.g., uncertainties in stage-discharge relationships or glacier and/or groundwater contributions.

L151-153: I very much appreciate the addition of statistical tests, but feel obliged to point out that a 5% significance level suffers from much the same problems as defining, for example, NSE > 0.5 as a threshold for good models.

There is nowadays quite a bit of literature on this, for example:
- Nature: https://www.nature.com/articles/d41586-019-00857-9
- TAS: https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108
- TAS again: https://doi.org/10.1198/000313006X152649

I realize it may be a bit much to overhaul the overhauls, bit reporting p-values instead of a dichotomous significant/not-significant set of results might be better. In Figure 9 for example, marker size could represent significance (e.g., lower p-value, bigger circle). If that is deemed to cumbersome, some cautionary words about the p-values that underpin any significant/non-significant statements would be welcome.

We appreciate the recommendation, and the documents shared. While we agree with this point, adding a new layer to Figure 9 could make it difficult to read the figure and would obscure the message we are trying to convey. However, we have modified the caption to avoid defining statistical significance in binary terms, and we specify that the line thickness of the circles refers to p-values lower than 0.05. Now, the complete caption reads as follows:

Figure 9. Spearman's rank correlation coefficient between different catchment attributes and the CSI for (a) droughts, (b) floods, and (c) drought-to-flood transitions, based on the simulations with GR4J and TUW calibrated using the unweighted HiLo original KGE formulation as the objective function. The circles with thick outlines indicate correlation coefficients with p-values lower than 5%.

In the other figures (e.g., boxplots) where we seek to compare differences between configurations, we have included the p-value directly. Thus, for example, the caption of Figure 6 now reads like this:

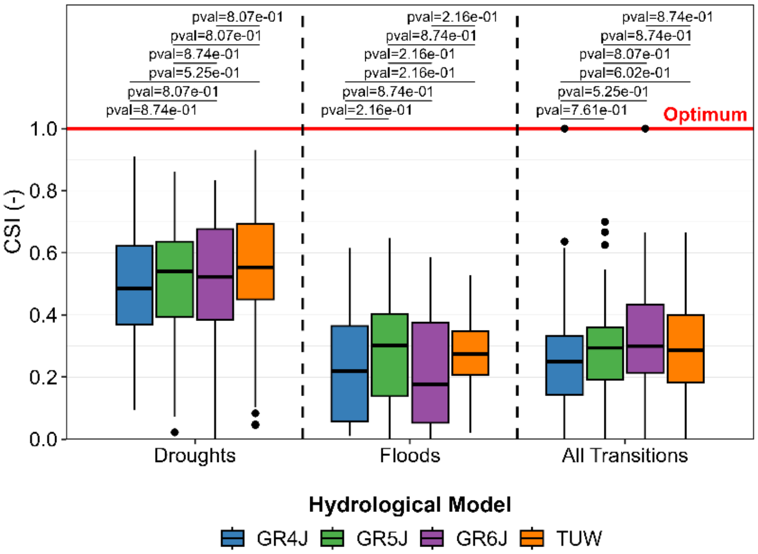

Figure 6. Critical Success Index (CSI) for (a) droughts, (b) floods, and (c) drought-to-flood transitions, based on the simulations with GR4J, GR5J, GR6J, and TUW (different colors) calibrated with the unweighted HiLo original KGE formulation as the objective function. Each boxplot displays the information of 63 values (i.e., one per catchment). The p-values (pval) correspond to the Wilcoxon statistical significance test.

L181: This phrasing makes sense to me in the context of the first set of review comments (I believe this text is meant to say that the model parameters, particularly X2, may be calibrated to values that compensate for errors in the forcing, and to reduce the chances of this happening, two dedicated forcing parameters are included in the calibration procedure), but it may be too mysterious for a reader who does not have this context. It'd be good to clarify the text a bit I think.

We have modified the text, which now reads as follows:

Bucket-style conceptual models generally include parameters and functions that allow for non-conservative adjustment of the water balance (i.e., artificially adding or leaking water). While they can help correct potential mismatches, e.g., between topographical and underground catchments, they can also compensate for biases in the forcing. To explicitly correct biases in the meteorological forcings (as illustrated in Figure 1c), two parameters were included in the calibration process in addition to the original setup for each hydrological model. Specifically, a multiplicative parameter for precipitation (dP) and an additive parameter for temperature (dT) were included to adjust systematic biases in precipitation and temperature.

L183-L185: Following up on the response to my my question about these two parameters, the values shown in Fig S16 seem really quite high to me.

The corrected precipitation in Chile on average goes up by at least 20% in all but 1 basin with various basins receiving over double the precipitation after correction. Corrections in Switzerland are lower but still, generally speaking, suggest that incoming precipitation is increased.

dT patterns should be additive and suggest that in Chile temperatures (and thus PET derived from this) are increased unless the basins have a considerable amount of snow. In Switzerland, temperatures generally seem to be increased.

Combined, I'm wondering about the equifinality in these two parameters. Precipitation generally tends to be increased but so does temperature and thus PET - to what extent is the extra P simply immediately converted into ET and thus rather meaningless? The large scatter in calibrated values for dT and dP (as the *authors correctly, I think, already mention in the updated discussion) at least suggests that the current calibration of these two parameters does not have an easily identified global optimum per basin (which is what we would need to find if these parameters truly correct forcing biases specific to the basin).*

The authors currently state as the final point in section 5.3 that having better local forcing data would be helpful. I agree of course, but I don't think that's the main point a discussion about these two parameters should focus on. What's currently missing however is a discussion about how this equifinality in calibrated dP and dT values might affect the conclusions obtained from this work. If within a given basins the models might be relatively speaking working with anything between regular and double the precipitation, there should be consequences for how well they simulate droughts and floods. The best option, I think, would be to repeat calibration without these correction factors but I realize that this may be infeasible at this point.

Thank you very much for raising this point and sharing your thoughts. We agree that this is an interesting point, which is beyond the scope of this study. However, we have conducted some calibration experiments to understand the effect of forcing adjustment parameters on (i) e.g., the partitioning of precipitation into evapotranspiration (i.e., ET/P) and runoff (Q/P), and (ii) the identifiability of hydrological model parameters. To do so, we have calibrated each model both with (adj) and without (no_adj) the forcing adjustment parameters. These experiments are performed considering the original KGE formulation, no streamflow transformations (i.e., case referred to as “Hi” in the manuscript), and no weights applied to the variability term in the KGE as the objective function.

Figure 2 shows that when we increase precipitation amounts through calibration, the mean annual runoff ratio may decrease considerably, which, in our case, translates to increases in the evaporative ratio rather than changes in storage (not shown).

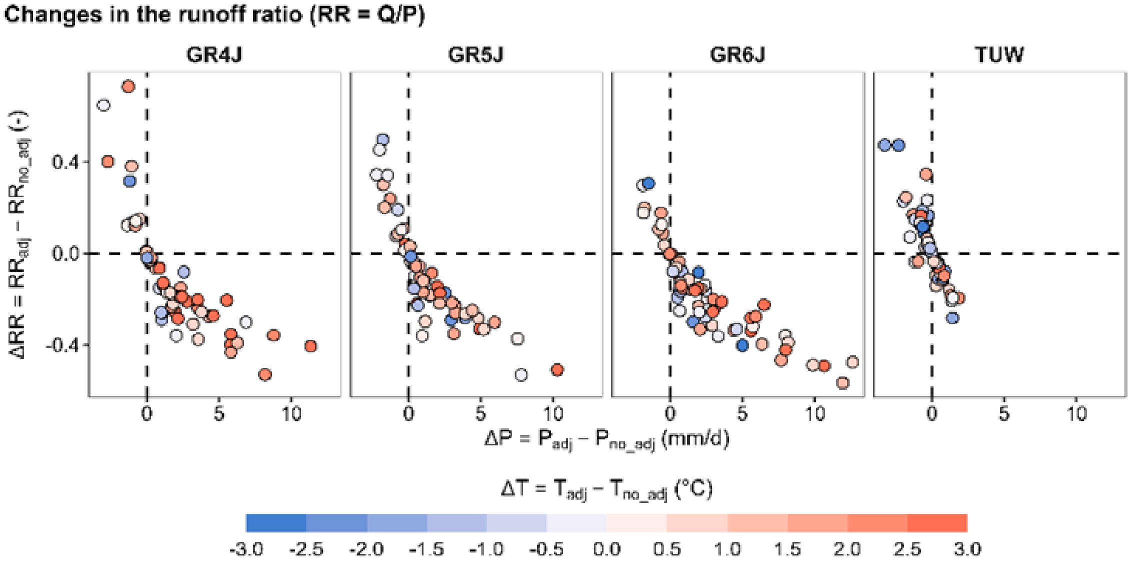

Figure 2: Changes in the partitioning of precipitation into ET and runoff in the period 1985-2020 when incorporating the forcing adjustment parameter in the calibration process. Each panel contains 63 points, one per basin.

Figure 3 shows the parameter agreement index (R) - for calibration results with the KGE – which quantifies the extent to which, in this case, a certain objective function yields unique parameter sets (Guillaume et al., 2019), with values closer to 1 indicating parameters with very similar values. To compute R, we considered the N parameter sets with KGE = max(KGE) – 0.01. Our results show that the incorporation of the forcing adjustment parameters has no significant impact on the identifiability of the hydrological models’ parameters. However, as shown in Figure S12 (and only for the adjustment parameters in Figure S11) in the previous version of the Supplementary Material, some parameters are less identifiable when different objective functions are tested. Note that we prefer to discuss identifiability or parameter agreement here, as we are not performing a "formal" equifinality analysis, which is out of scope for our study.

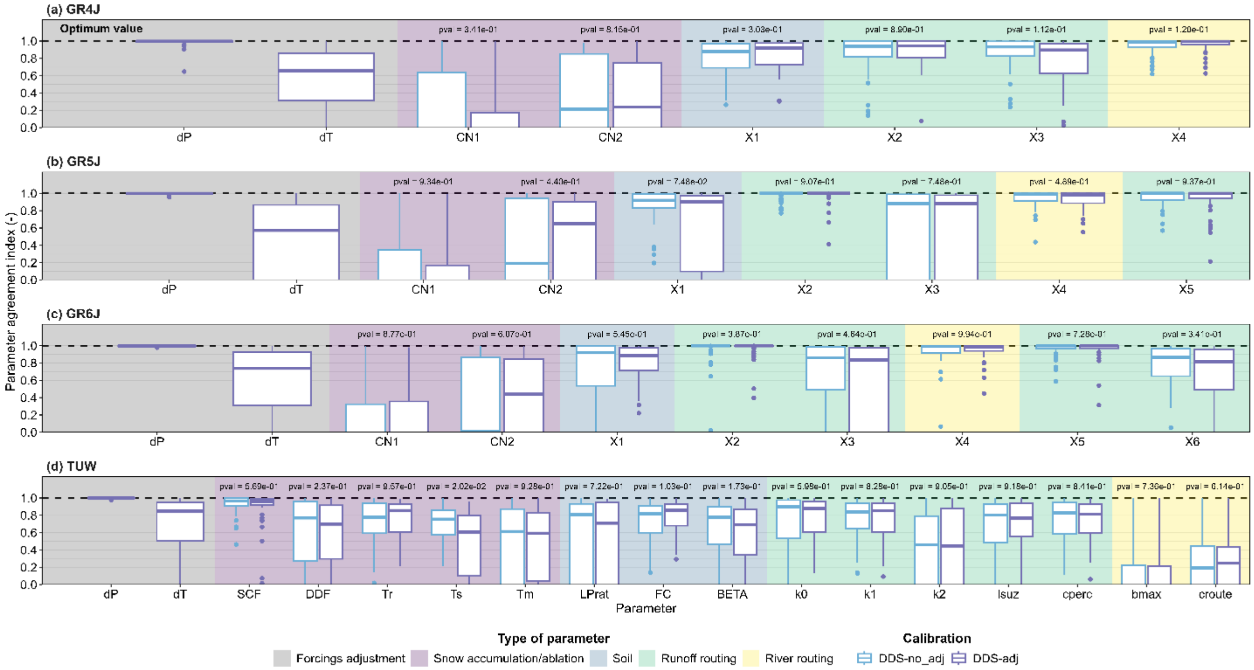

Figure 3: Comparison of parameter agreement index (R) with and without forcing adjustment parameters. Each boxplot contains 63 values, one per basin. The p-values (pval) correspond to the Wilcoxon statistical significance test. The agreement index is computed as $R = 1 - (\theta_{95\%} - \theta_{5\%})/|\theta_{50\%}|$. Then values close to (far from) 1 indicate high (low) agreement between parameters identified as equifinal.

To summarize the results obtained from this analysis, we have included the following in the discussion section of the manuscript:

We acknowledge that the incorporation of forcing adjustment parameters could have an impact on the partitioning of precipitation between ET and runoff. However, this problem also occurs when working with different parameter sets, which may come from different calibration functions. We have evaluated the impact of incorporating these additional parameters on the identifiability of the original model parameters, showing that there are no meaningful impacts (not shown).

L250: I appreciate the extensive clarification for using NSE in the response document, but I don't think I agree. We know that KGE doesn't cleanly map onto NSE but that this mapping depends on the coefficient of variation of the observations; we also know that models calibrated for one objective don't necessarily perform well on the next; and we know that NSE as a metric was not designed to be informative about the transitions the authors are interested in.

I understand the need to have a single metric to compare the 5 different KGE equations against, but it is unclear to me why NSE is the right choice. In the light of research question 2:

"Which modeling choices (e.g., [..] KGE formulation, [..]) are the most important for simulating droughts, floods and their transitions?"

How does comparing the NSE scores of the five KGE formulations help with answering this question? NSE is not a metric aimed specifically at any of these event types and the the signatures etc. shown in Text S2 seem more appropriate to me. Relying on NSE here answers the question "Do different KGE formulations differ in resulting NSE scores?" but the link with the events the authors are interested in is tenuous at best. In my opinion this section would be stronger if it shows in the main text how the different KGE formulations lead to specific changes in the ability to simulate droughts, floods and transitions, rather than relying on NSE as a proxy for this.

We agree that the model should be evaluated on the target variables, which in our case are droughts, floods, and their transitions. This is what we do and emphasize in the manuscript, for example, in Figure 4 (i.e., good KGE does not imply good detection) and Figure 5 (i.e., different KGE formulations lead to different performance in event detection). NSE was initially used to provide a general overview of model performance. We have updated the manuscript to further deemphasize evaluations on NSE and emphasize those focused on event detection. To do so, we have restructured Section 3.3 and placed greater emphasis on the CSI analysis and its relationship with KGE. We still included a few sentences on goodness-of-fit metrics and hydrological signatures to discuss the overall performance of the models, and the reader is referred to the supplementary material for further details. The text now reads as follows:

We assessed model accuracy both in terms of general model performance and the ability of the model to capture extreme events and the transitions between them. We followed a traditional split-sample test approach (Klemeš, 1986; Beven, 2025) to assess the general model accuracy over two time periods defined as (i) calibration (2000-2020) and (ii) evaluation (1985-1999). To test for general accuracy and hydrological consistency across the calibration experiments tested here, we computed several goodness-of-fit metrics (e.g., KGE) and hydrological signatures (e.g., seasonality, low- and high-flows). First, we assessed model performance across the 60 configurations by comparing the values obtained for each objective function during calibration. Second, we assesses the predictive skill of our calibrated models by comparing their performance during calibration with that of a simple daily mean flow benchmark. This benchmark is defined as the mean flow for each day, calculated from all instances over the calibration period (referred as BM05 in Knoben, 2024). Third, we assessed model performance by looking at biases in a set of hydrological signatures, including seasonality, statistical properties (mean, variance), flow duration curve-derived signatures (e.g., mid-segment slope), and annual extremes (see Table S5). We conducted this analysis in two steps: (i) we analyzed the models' ability to reproduce the seasonal timing (seasonality) of streamflow (Q), snow water equivalent (SWE), and actual evapotranspiration (ET); and (ii) we computed biases in streamflow-derived signatures. The results of this general model performance assessment are presented in Text S1 in the Supplementary Material.

Understanding the importance of explicitly showing the model's accuracy, we now begin the results section with the following paragraph:

Before looking at model performance in terms of capturing extreme events, we assessed the overall performance of the four models used. For this, we independently evaluated the calibration results for each configuration. Our results shown comparable performance across the hydrological models evaluated here (Figure S1). For instance, all configurations outperform the defined daily mean flow benchmark (see Figure S2), indicating that our models have greater predictive power with respect to the long-term observed streamflow series. Our more detailed analyses show that the seasonality of variables such as streamflow, SWE, and ET are simulated accurately, with median performance values across catchments and configurations between 0.79-0.98 (with 1.0 being the optimum). However, our evaluation shows that using weights for the variability term of KGE greater than can be detrimental to the overall performance of the model, both in terms of representing the seasonality of the aforementioned variables (Figure S3) and some hydrological signatures such as the high- and low-segments of the slope in the flow duration curve (FDC, Figure S4). In general, the use of flow transformations yields values that are consistent with what the application seeks to capture (e.g., low-flows are better simulated with "Lo" transformation and high-flows are better simulated without transformation; see Figure S5). There is little difference between different models and KGE formulations when weights and the HiLo transformation are used (Figure S6). Considering those configurations with comparable performance (i.e., removing those relying on weights greater than 2), average accuracy across configurations ranges between 0.87-0.92, 0.88-0.93, and 0.75-0.85, for the high-, mid-, and low-segment of the slope of the FDC, respectively. Further details on overall model performance are presented in Text S1 in the Supplementary Materials.

L250-L251: The phrasing here seems rather funny to me, as if the NSE is this relatively new thing that has been steadily gaining popularity instead of having been the go-to metric for calibration for most of the last half century. Perhaps consider replacing the "has become" with "is".

In line with the idea of reducing complexity as suggested in the first iteration of the review, this description has been removed from the text to focus on the analysis of the models' detection ability over "traditional" performance metrics. These analyses have been moved to the supplementary material.

L254-L255: I might have missed it earlier, but which data sources were used for this and how was the (presumably) gridded data averaged into a catchment-wide timeseries?

The description of the databases used for verification is provided in L134-L136 of the revised version of the document. However, after rereading the manuscript and in line with the idea of reducing complexity, this description has been moved to the Supplementary Material.

L261: I'm returning to this part of the manuscript after going through the paper again, because I think it might still be good to add a small section or at least mention of these outcomes somewhere in the main text. Including the current Figure ST2.2 in the main paper would be really helpful to set the reader's expectations about model performance. If space is an issue, a small new "section 4.1 Calibration and evaluation performance" bit of text that clearly points to the SI could be a good middle ground.

I think this important because the performance of the 5J and 6J models is just so different from the other two that I think the reader needs to be aware of this.

It's also not fully clear to me how this evaluation on auxiliary data is actually used. It seems to me that it's just put in the SI and then not really referenced again. As far as I can tell at least, there's no model culling step between calibration and the rest of the analysis that says "we only keep models that perform at least X on variable Y". The only current message in the SI seems to be that in most cases the models score above NSE = 0 (which tends to be a pretty low bar in any basin with some seasonality) and when they don't at least the correlation is > 0.5 (probably not super hard either given that the seasonality is already captured in the input data). Is that really good enough to consider all these models as equally plausible candidates in each basin? Is there really that much value in essentially concluding that a model that can't be calibrated well for a given basin doesn't simulate specific events very well in that basin?

We appreciate the recommendation. We have incorporated some results into the main text and improved the analysis presented in the supplementary material, as we refer the reader to this material for a detailed analysis (further details are provided in one of the responses to a similar comment above). We have decided not to include this analysis in the main text so as not to complicate the message we are trying to convey.

Figure 4: I appreciate the clarification about the different model structures (particularly that GR5J does not become GR4J for X5=0) but these differences in KGE scores (as an aside - are these calibration or validation scores?) still seem really large to me for what feels like a fairly simple change in equations.

I realize I keep piling on more work but I think it would be really good to run a handful of extra tests to figure out if these results are "real" (for lack of a better word, meant in the sense of GR5J and GR6J just being worse at fitting the data than GR4J) or if this is a calibration artifact. One possible test may be to run GR5J and GR6J with the optimized values for from GR4J (and filling in X5 and X6 either with their own calibrated values) and seeing if their performance goes up or down compared to their own calibrated outcomes. Better might be to trial a few calibrations of GR5J and GR6J with initial parameter values determined from optimized GR4J parameters for the basin.

Thank you for your suggestions. To address this point, we have used the functions available in airGR (Coron et al., 2023) to evaluate how similar the models are when calibrating GR4J considering X2 = 0, a model we have called GR3J-GR4J, and then using these parameters to create simulations with the GR5J model setting X5 = 0 (GR3J-GR5J) and GR6J with X2=X5=0 and X6 = 0.01 (minimum value that can be adopted; GR3J-GR6J). Here, we are not analyzing whether performance improves or not, but only how much it changes relative to the GR3J base case, as we seek to understand the effect that changes in the structure of the GRXJ models have on the representation of daily streamflow.

For these experiments, we used the internal calibration algorithm (Calibration Michel; Michel, 1991) and daily streamflow for the period 2000-2020 and the original KGE formulation as our objective function. Figure 5a shows the performance of the GR3J-GR4J model, while Figure 5b shows the difference in its performance when compared to the GR3J-GR5J and GR3J-GR6J versions. For the calibration period, both models show, on average across catchments, a decrease in performance compared to the GR3J-GR4J version, which is more evident in the GR3J-GR6J model. The changes in GR3J-GR6J respond to the incorporation of the exponential storage in the GR6J structure, which cannot be "deactivated". Although one might expect that with X2 = X5 = 0 (i.e., GR3J-GR5J) the model's performance should be the same (i.e., ΔKGE = 0), this is not the case because the GR5J model, in addition to having a different exchange function, has a modified routing component compared to GR4J.

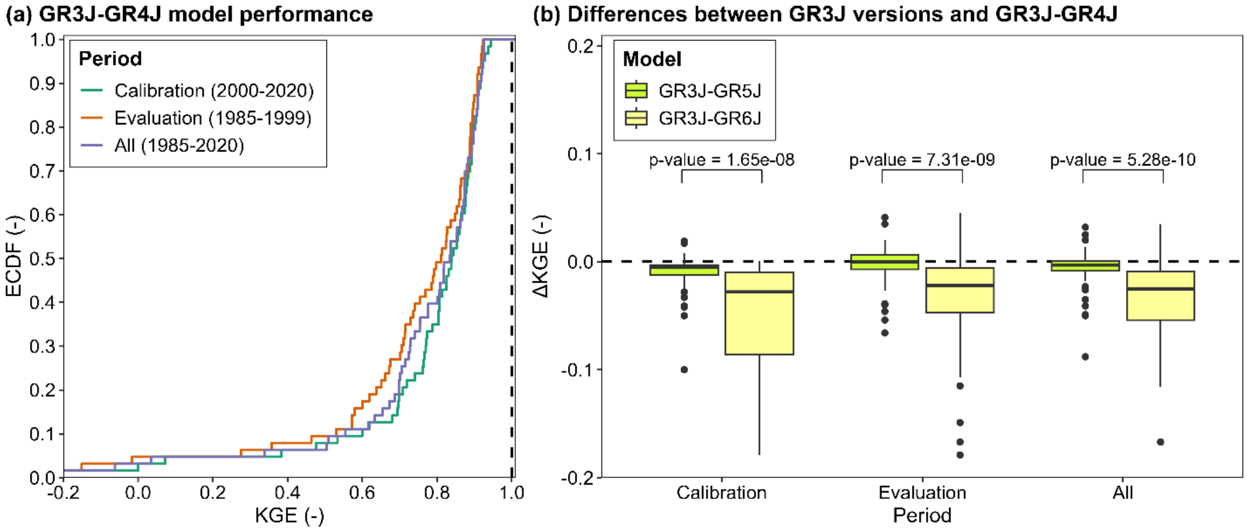

Figure 5: Performance of the GR3J-GR4J model and comparison against the GR3J-GR5J and GR3J-GR6J versions (i.e., ΔKGE = Alternative model – reference). Each line and box plot in the figure is based on 63 points (one per catchment).

To better understand the impact of these slight changes in the structure of the GR models on model performance, Figure 6 shows the changes in the KGE components. Although the bias term (Figure 6c) shows negligible differences with respect to GR3J (which is to be expected since the parameters X2 and X5 have been set to 0, and X6 to its minimum plausible value), there are changes in the representation of the dynamics of discharge (Figure 6a) and variability (Figure 6b). In general, modified GR3J models change performance in terms of dynamics.

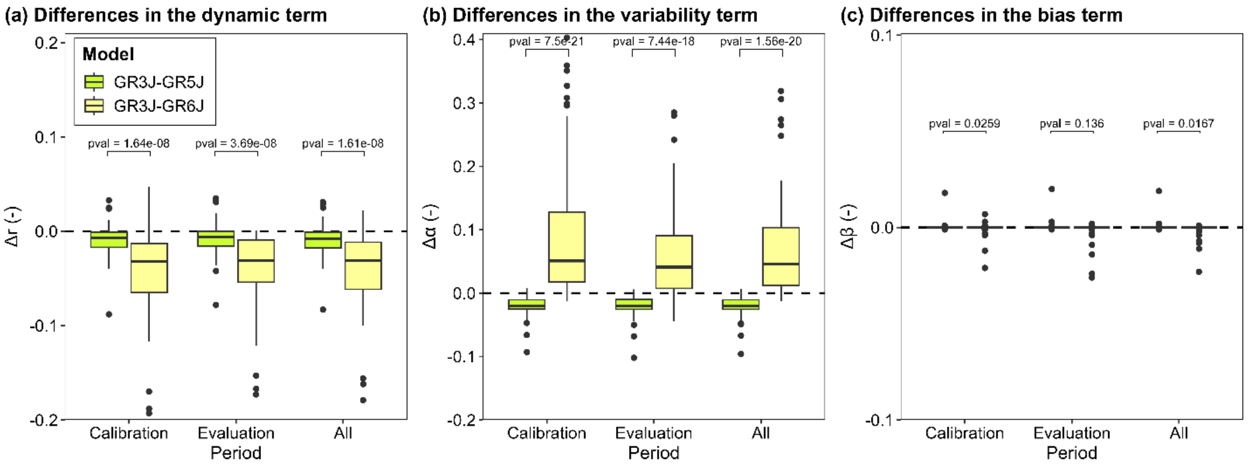

Figure 6: Changes in the KGE components associated with the modified GR3J model versions compared to the reference (i.e., Δ = Alternative model – reference). Each boxplot contains 63 points (one per catchment).

In short, we have shown that the structures of GRXJ models are inherently different regardless of the exchange function. While we would be interested in investigating these differences further, this is beyond the scope of our study. To highlight this, we have written the following in the description of the models included in the manuscript:

It is important to note that the original structure of GR4J cannot be recovered by setting the parameter X5 equal to zero in GR5J, nor can GR5J be obtained by setting parameter X6 = 0.01 (the minimum value that can be adopted) in GR6J. This is because, e.g., in GR5J the routing function differs from GR4J, whereas in GR6J the effect of the exponential storage (defined by X6) cannot be canceled. Thus, despite having the same core structure, the models are intrinsically different from each other.

Then, we calibrated the GR5J and GR6J models by setting X2=X5=0 and X6=0.01 to compare the performance of the models against GR3J_mod and GR3J_mod2, respectively. To evaluate how much is "gained" by calibrating the models, for metrics with a range of variation between minus infinity and 1, with 1 being the optimal value (e.g., KGE), we calculate the relationship between the calibrated models and the reference by applying the following equation:

$$\Delta Y^{*} = \frac{Y_{\mathrm{Alternative}} - Y_{\mathrm{Reference}}}{Y_{\mathrm{Optimum}} - Y_{\mathrm{Reference}}}$$

For metrics such as bias (β), where 1 is optimal but higher or lower values represent poorer performance, the term is transformed to express it in the form, e.g., β_mod = 1 – abs(β – 1), to have the same interpretation as in the case of, e.g., KGE. Figure 7 shows that by calibrating the models, we always get an improvement in the performance in comparison to running the GR5J and GR6J models with the calibrated parameters for GR3-GR4J. The major differences are in the variability term, while, as expected, the sweeping potential of the models is canceled by fixing X2=X5=0 and X6 to its minimum value, so changes in the bias term are close to zero. With this, we demonstrate that GRXJ models are not identical, despite sharing the same core structure and development philosophy. In other words, and returning to one of Dr. Knoben's questions, despite having models with X2 and X5 parameters very close to or equal to zero, i.e., the "GR3J" model, these remain different when comparing GR3J-GR4J, GR3J-GR5J, and GR3J-GR6J as we showed in Figures 5-7.

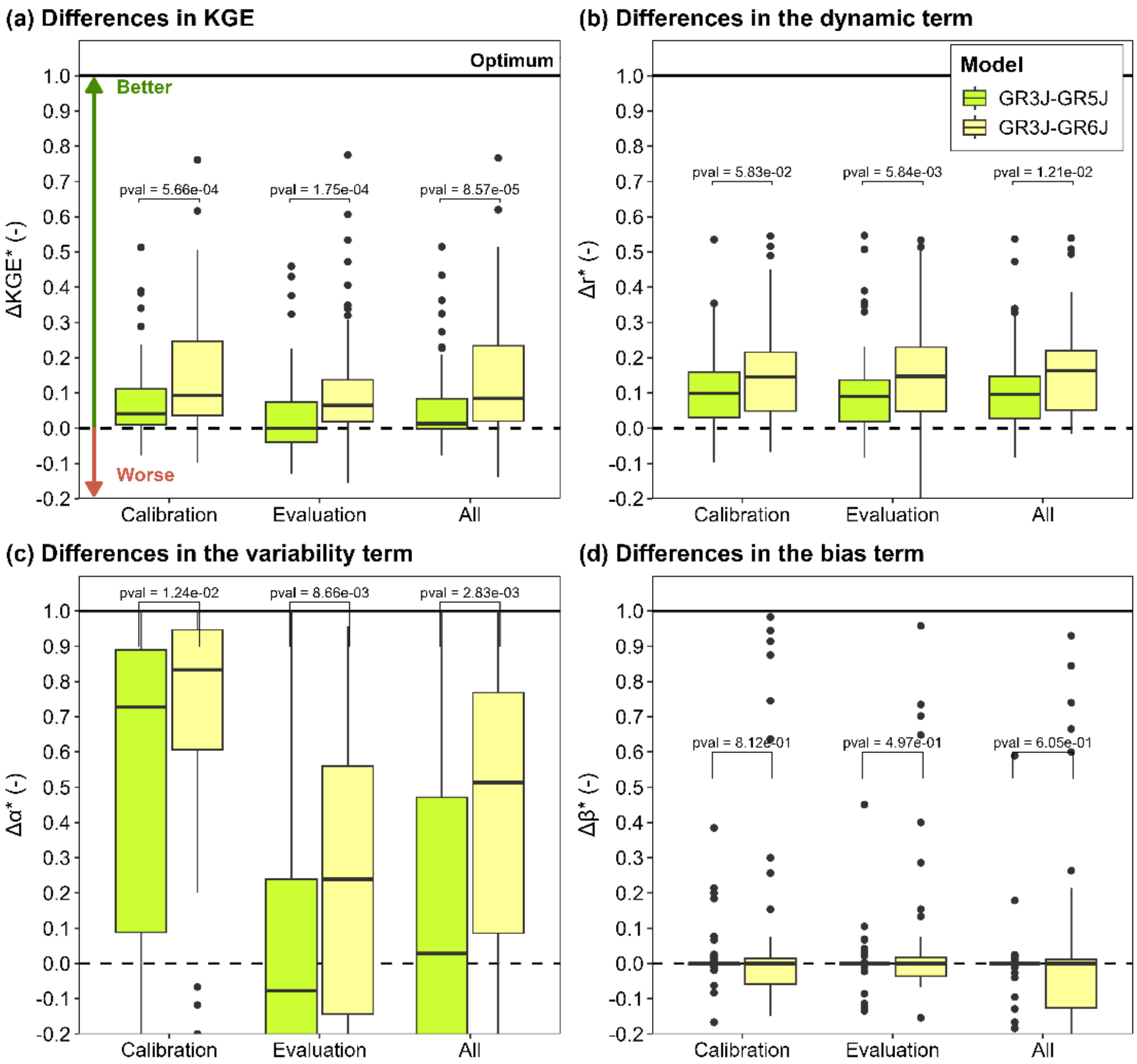

Figure 7: Added value of using calibrated GR3J-GR5J and GR3J-GR6J structures compared to their uncalibrated versions using parameters transferred from the calibration of GR3J-GR4J.

To evaluate the effect of incorporating parameters X2, X5, and X6 compared to canceling them, we recalibrated the GRXJ models without restrictions. Figure 8 shows that, by allowing calibration of all parameters, improvements are always obtained compared to the version with X2=X5=0 and X6 = 0.01 (i.e., canceling the exchange function and the exponential storage). These improvements are observed for both the KGE and its components. In summary, we prove that there is a benefit to using the exchange function and the exponential storage (i.e., allow the calibration of X2, X5, and X6 depending on the model) compared to canceling them. These findings are aligned with previous studies showing that increasing model agility improves model performance (e.g., Mendoza et al., 2015; Newman et al., 2017).

The performance difference of the GR5J and GR6J models compared to GR4J per catchment (Figure 9) is around zero. This aligns with the comment made by Dr. Knoben, who hypothesizes that the performance of these models should be similar or, in other words, the degradation/improvements in performance should not be as large as the ones shown in the previous version of the manuscript given the similarities between the models.

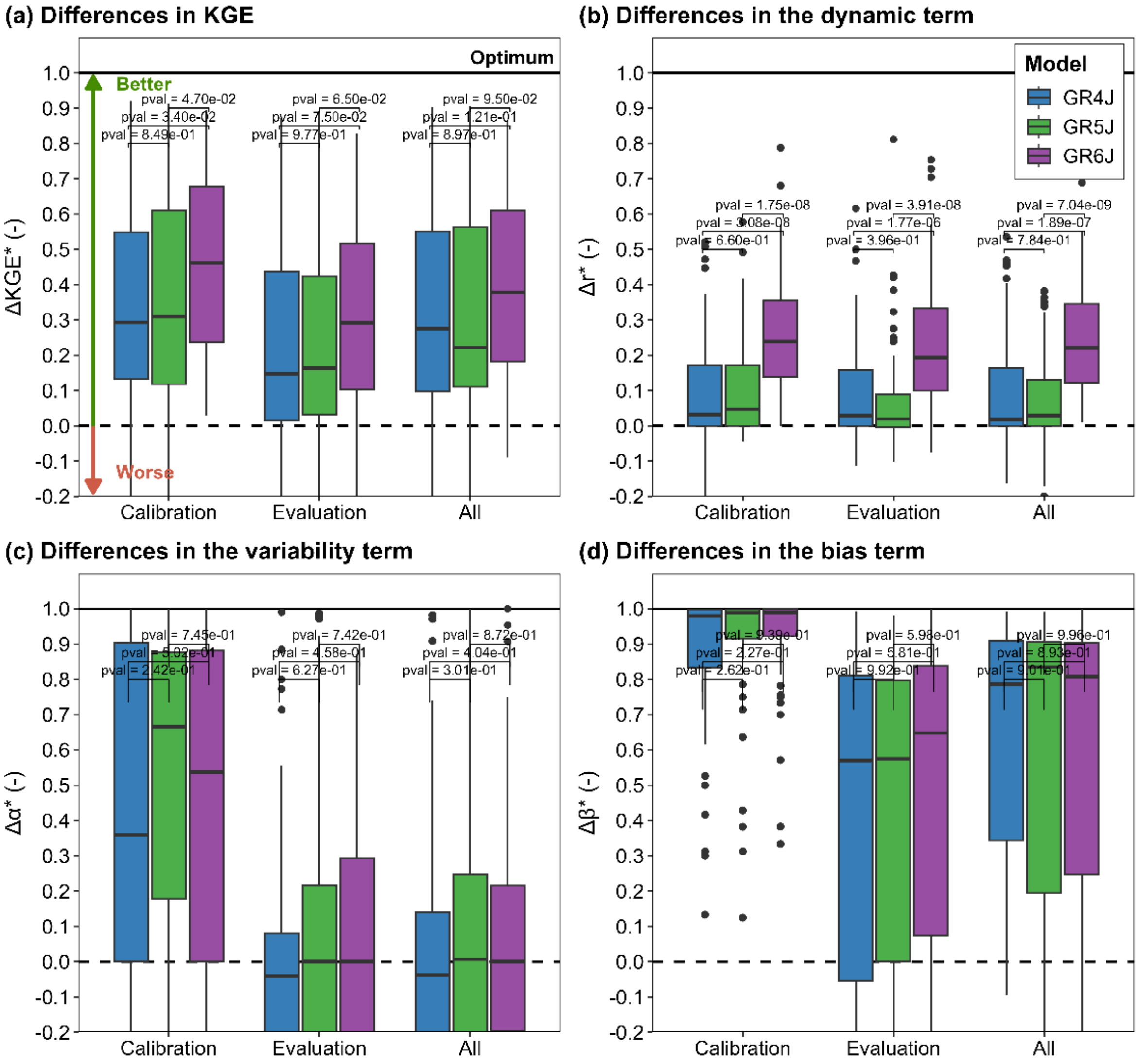

Figure 8: Added value in modeling with GRXJ models by incorporating exchange function parameters and exponential storage into model calibration (i.e., X2, X5, and X6).

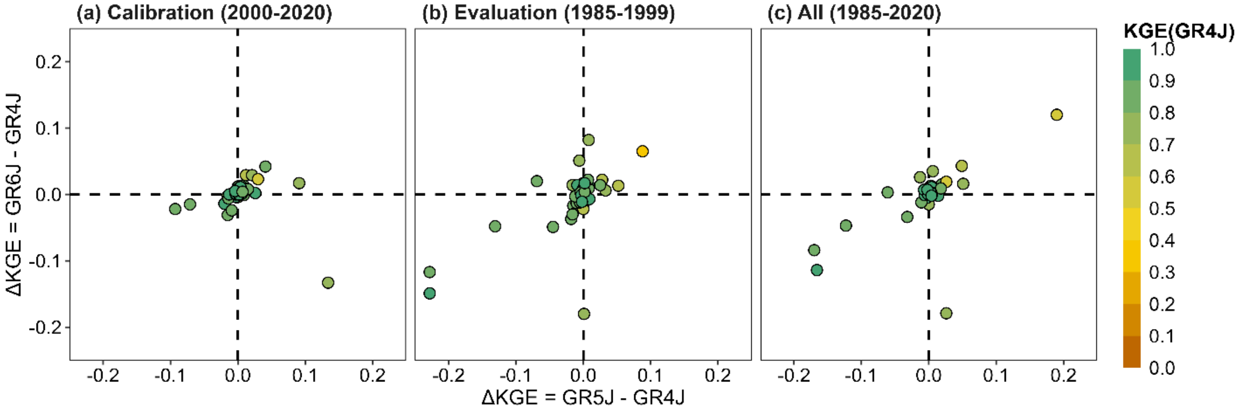

Figure 9: Differences in KGE when comparing the performance of the GR4J model with GR5J and GR6J. Values (lower) greater than 0 represent (worsening) improvements in performance compared to GR4J. Each point represents a catchment in the study domain (i.e., 63 points per panel).

This seems even more relevant when looking at the optimized parameter ranges in Fig S10. Combined with Table S1, the figure shows that for GR4J, the water exchange parameter X2 plays a large role (indicated by values being centered around normalized 0.4 of actual range [-35,35] with a wide box). In contrast, for both GR5J and GR6J X2 behaves very differently: very narrow box centered around normalized 0.5 or actual 0. X5 shows this behaviour to an even greater extent, with normalized values 0.5 and thus actual value 0.

To me this seems critical, and at the very least implies that GR5J is effectively a 3-parameter model, and GR6J a 4-parameter one (though with a different 4th than actual GR4J).

As we have discussed above and shown in Figure 8, allowing the calibration of all the parameters shows an added value in comparison to performing the calibration with fixed parameters X2=X5=0 and X6 = 0.01 (i.e., constraining the models to a "3-parameters" version). Then, independently of the value of these parameters, which can be close to 0 (i.e., cancel the effect of the exchange function and minimize the effect of exponential storage), accounting for the interaction between all the parameters in the calibration process enhances model accuracy in the calibration period.

I'm not sure if this sort of stuff should even go into the SI but if it were my results I would want some sort of clearer idea of what I'm looking at with this. It seems weird and I don't think that "if the calibration converges we consider it successful" is that strong of an argument. There's a lot of literature out there about local minima in optimization problems, as well as the impact of the numerical scheme on how jagged the response surface is and were optima may be found, that suggests that convergence does not imply optimality.

We agree with the point raised and have mentioned it in the previous manuscript. For instance, between L479 and L481 we wrote: "However, it is important to acknowledge that potential compensations for biases in meteorological forcings or model deficiencies can make the "optimal" parameter sets less identifiable (e.g., Clark and Vrugt, 2006; Vrugt et al., 2005; Beven, 2025).". Nonetheless, we agree that the statement presented in L478-L479 is a bit strong and may be misleading regarding the point we want to emphasize. Therefore, we have rewritten that point, which now reads as follows:

These models have been calibrated based on daily streamflow records, using different objective functions derived from KGE formulations, and considering the set of parameters with the best performance as the optimum. However, it is important to acknowledge that potential compensations for biases in meteorological forcings or model deficiencies can make the "optimal" parameter sets less identifiable (e.g., Clark and Vrugt, 2006; Vrugt et al., 2005; Beven, 2025).

To better understand the differences in calibration results, we compared the results presented in the manuscript (i.e., SCE) with Michel's calibration (presented in Figure 9) and two alternative calibration algorithms: DDS (Tolson and Shoemaker, 2007) and hydroPSO (Zambrano-Bigiarini and Rojas, 2013). Note that we cannot consider Michel's calibration algorithm for all our experiments because: (i) the calibration objective functions cannot be modified, (ii) the algorithm cannot be directly applied to models other than those in the GR family, and (iii) its use is not recommended in models with too many parameters. However, to provide a direct comparative case with Michel's calibration, we created a DDS calibration without adjustment parameters (DDS-no_adj). We have decided to use DDS for this experiment because it has a low computational cost. Results are presented in Figure 10. It is important to highlight that the calibration with SCE used in the manuscript does not account for parameter-space transformations in the GRXJ models, as in the other cases here tested.

From Figure 10, we can see that when forcing adjustment parameters are included, as opposed to not including them (i.e., DDS-adj vs. DDS-no_adj), the overall performance of the models decreases. While keeping the same number of iterations, this can be explained by the increase in the number of parameters to be calibrated (e.g., eight instead of six for GR4J). However, calibration with DDS and hydroPSO yields “acceptable” values for all models, while for SCE, results deteriorate as the number of parameters in the GR structure increases. This could be explained by the lack of incorporation of the parameter space transformation, aimed at enhancing the search, or by the number of algorithm iterations (the sensitivity of the method to the number of complexes and the definition of the term CCE.iter were evaluated, revealing some differences; not shown).

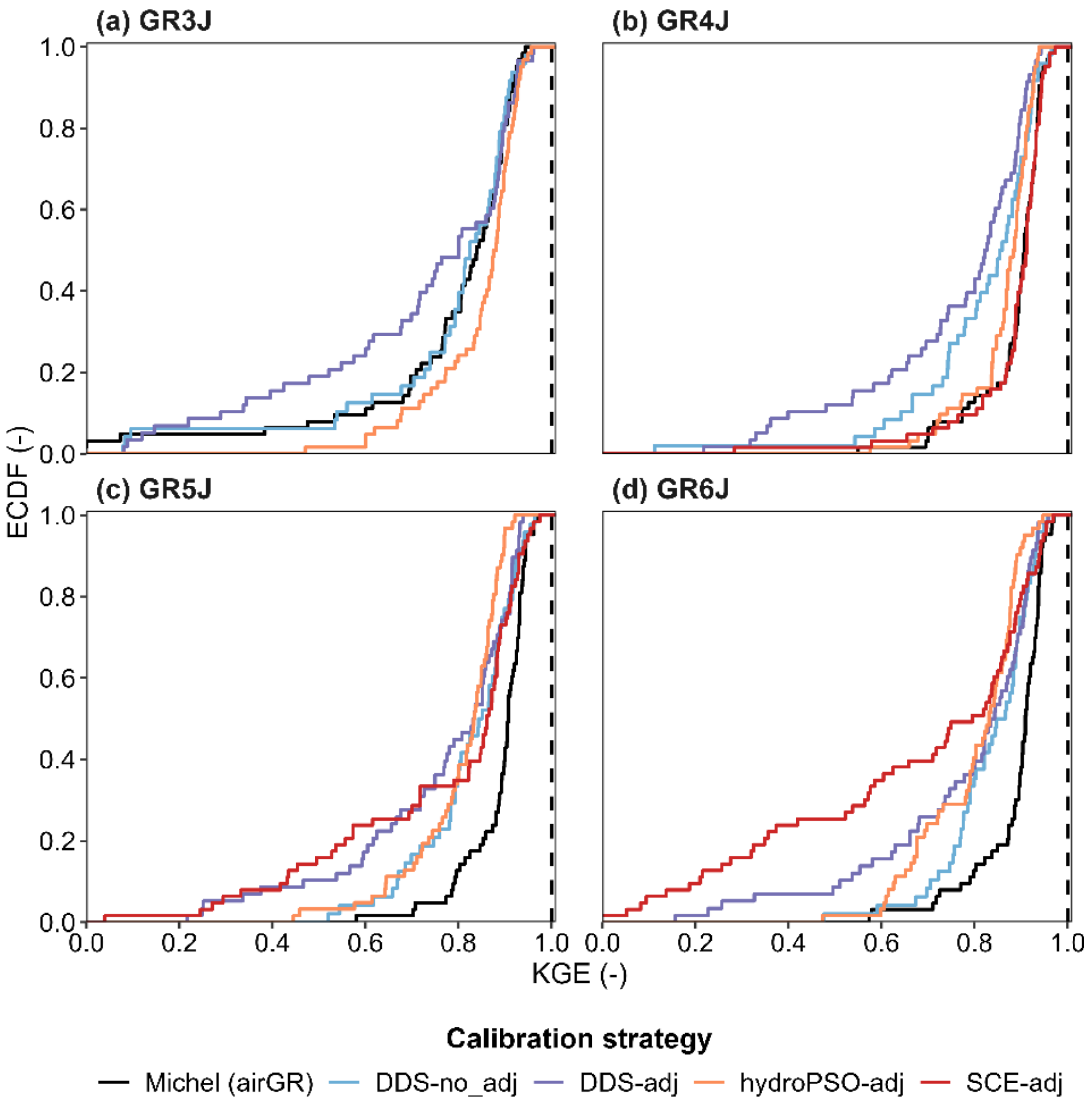

Figure 10: Comparison of the performance of GRXJ models calibrated using different calibration algorithms. Calibration is performed for the period 2000-2020 using unweighted original KGE formulation and daily streamflow without transformations.

Interestingly, for the TUW model, calibration with SCE yields results comparable to those with DDS (Figure 11), despite having more parameters than GRXJ, suggesting that the problem isn't necessarily related to the number of model parameters. Moreover, the performance of DDS-adj outperforms DDS-no_adj with same number of maximum iterations (i.e.,1000).

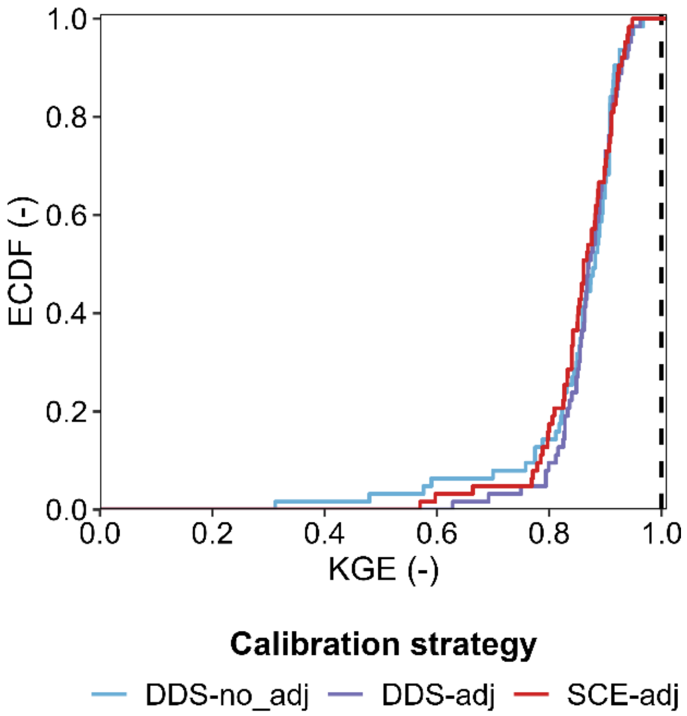

Figure 11: Comparison of the performance of TUW model calibrated using different calibration algorithms. Calibration is performed for the period 2000-2020 using unweighted original KGE formulation and daily streamflow without transformations.

To address the main comment raised during the review process (i.e., the calibration), we recalibrated the GRXJ and TUW models using the DDS algorithm and set different maximum iteration numbers based on the number of model parameters. Considering that there are no clear guidelines in the literature on how to define the number of iterations in DDS, we considered 200 iterations per parameter (i.e., Niter = 200 x Npar). This is considering that, in Figure 10, for GR3J without including forcing adjustment parameters (i.e., 5 parameters), DDS, configured with a maximum number of iterations of 1000 (i.e., 200 iterations/parameter), shows results similar to the Michel calibration. Based on several experiments, we also decided to modify the plausible range for the calibration of GRXJ models presented in the previous version. Now, for X2 and X5, we consider a range of -10 to 10 (originally -35 to 35), and for X6, a range of 0.01 to 2000 (originally 0.05 to 2000), and include a parameter-space transformation to enhance the search process during calibration.

Note that in our responses we included the “GR3J” model (i.e., GR4J with X2 = 0) because it was one of your suggestions, but in the updated manuscript we refer only to the GR4-6J models. This, considering that, in general, the GR3J model does not perform better (or comparably) than the GR4-6J versions (i.e., allowing calibration of the exchange function parameter X2-X5, and the exponential storage parameter X6, respectively; Figure 8).

Based on this analysis, we have included some of the results in Text S1 in the Supplementary Material and the following in the results section:

Before looking at model performance in terms of capturing extreme events, we assessed the overall performance of the four models used. For this, we independently evaluated the calibration results for each configuration. Our results shown comparable performance across the hydrological models evaluated here (Figure S1). For instance, all configurations outperform the defined daily mean flow benchmark (see Figure S2), indicating that our models have greater predictive power with respect to the long-term observed streamflow series. Our more detailed analyses show that the seasonality of variables such as streamflow, SWE, and ET are simulated accurately, with median performance values across catchments and configurations between 0.79-0.98 (with 1.0 being the optimum). However, our evaluation shows that using weights for the variability term of KGE greater than can be detrimental to the overall performance of the model, both in terms of representing the seasonality of the aforementioned variables (Figure S3) and some hydrological signatures such as the high- and low-segments of the slope in the flow duration curve (FDC, Figure S4). In general, the use of flow transformations yields values that are consistent with what the application seeks to capture (e.g., low-flows are better simulated with "Lo" transformation and high-flows are better simulated without transformation; see Figure S5). There is little difference between different models and KGE formulations when weights and the HiLo transformation are used (Figure S6). Considering those configurations with comparable performance (i.e., removing those relying on weights greater than 2), average accuracy across configurations ranges between 0.87-0.92, 0.88-0.93, and 0.75-0.85, for the high-, mid-, and low-segment of the slope of the FDC, respectively. Further details on overall model performance are presented in Text S1 in the Supplementary Materials.

L325: Space missing here

Thank you very much for bringing this to our attention. We have corrected it, as well as revised other spacing errors I the manuscript.

L328: Are these Figure references correct? Fig S11 for example seems to be the figure with dT and dP values. I don't think that directly contributes to the statement here?

Thank you for pointing this out. We have reviewed all references to supplementary material, which has been also updated in accordance with the recommendations provided by the reviewer.

L335: As mentioned before, I don't think "number of parameters == complexity" is entirely valid. In this case, removing the X2 component may make the model more behave linearly, and calling that "reduced complexity" (despite adding a parameter) can be argued as well. I'd suggest to rephrase this

Thank you for highlighting this point again. We have rewritten this and modified the term "complexity," which we understand is not directly related to the number of parameters.

L349-L350: There's also a very large difference in SWE which could be caused by different calibrated dP values. Is it that meaningful to compare simulated fluxes if the models have very different water budgets to work with, and there are no observations to tell us where the real values would be found? Sure, the fluxes are different, but if we put different amounts of P in that's not entirely unexpected.

We agree with your point. That is why we performed some sanity checks on hydrological variables other than runoff and, to some extent, detect potential inconsistencies in the model (e.g., poorly represented ET or SWE seasonality). Based on our updated calibrations, Figure 12 shows the model performance in simulating Q, SWE, and ET seasonality. To verify the representation of seasonality as a sanity check, and to avoid inconsistencies due to, e.g., systematic biases in reference products (e.g., ET retrieved from GLEAM), we use directional statistics (Berghuijs et al., 2025).

Based on these results, as well as the evaluation of hydrological indices derived from the streamflow series, we have included the following paragraph in the main manuscript:

Before looking at model performance in terms of capturing extreme events, we assessed the overall performance of the four models used. For this, we independently evaluated the calibration results for each configuration. Our results shown comparable performance across the hydrological models evaluated here (Figure S1). For instance, all configurations outperform the defined daily mean flow benchmark (see Figure S2), indicating that our models have greater predictive power with respect to the long-term observed streamflow series. Our more detailed analyses show that the seasonality of variables such as streamflow, SWE, and ET are simulated accurately, with median performance values across catchments and configurations between 0.79-0.98 (with 1.0 being the optimum). However, our evaluation shows that using weights for the variability term of KGE greater than can be detrimental to the overall performance of the model, both in terms of representing the seasonality of the aforementioned variables (Figure S3) and some hydrological signatures such as the high- and low-segments of the slope in the flow duration curve (FDC, Figure S4). In general, the use of flow transformations yields values that are consistent with what the application seeks to capture (e.g., low-flows are better simulated with "Lo" transformation and high-flows are better simulated without transformation; see Figure S5). There is little difference between different models and KGE formulations when weights and the HiLo transformation are used (Figure S6). Considering those configurations with comparable performance (i.e., removing those relying on weights greater than 2), average accuracy across configurations ranges between 0.87-0.92, 0.88-0.93, and 0.75-0.85, for the high-, mid-, and low-segment of the slope of the FDC, respectively. Further details on overall model performance are presented in Text S1 in the Supplementary Materials.

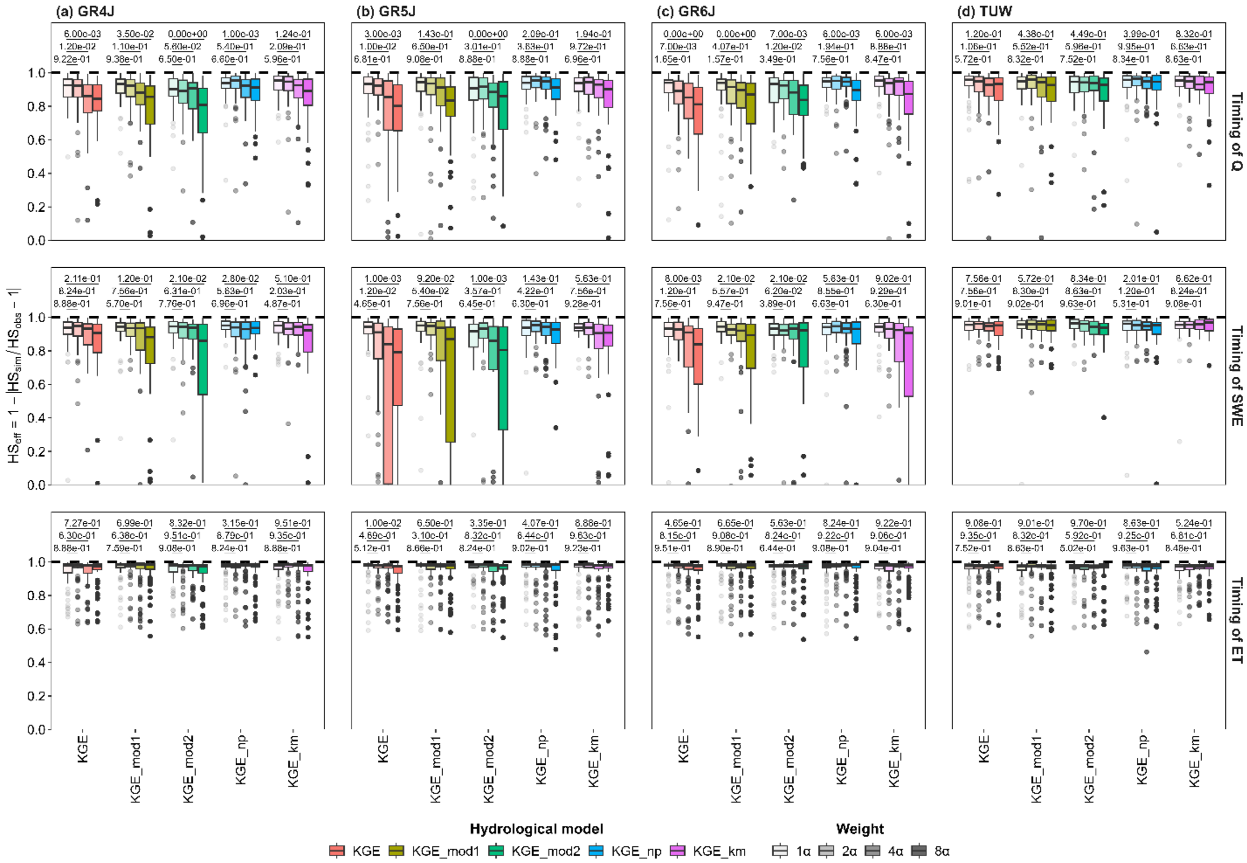

Figure 12: Model performance representing the seasonality of streamflow (Q), snow water equivalent (SWE), and actual evapotranspiration (ET), for different KGE formulations and weights, and HiLo streamflow transformation. The p-values corresponding to the Wilcoxon statistical significance test are included. Figure included in the updated Supplementary Material.

L374-L376: Regardless of whether or not the authors further investigate the GR5J and GR6J behaviour, I'm not sure this ANOVA analysis is entirely appropriate. The large influence of the model structure may be (at least partly) explained by the poor performance of the 5J and 6J models. I would recommend to add a second plot to Fig 8 where the models included are reduced to just GR4J and TUW, because it seems more reasonable to me to say that those were calibrated approximately equally well.

Thank you very much for the suggestion. First, based on the new calibration results, we have updated the ANOVA test applied to the CSI, which is presented in Figure 13. We decided to remove catchments characteristics from the explanatory variables because, as could be hypothesized, their relative importance softens the contribution of the other variables to the variability obtained for the CSI values. Then, following your recommendation, we have repeated the ANOVA test, removed two models, and tested all combinations to understand "how sensitive" our result is to this decision. Figure 14 shows that including basin characteristics (CA) accounts for much of the relative importance in detection in several of the cases evaluated. Additionally, it can be observed that, in general, the hydrological model is one of the most important decisions, even more so when comparing different models (e.g., GR6J and TUW). This decision is followed by the application of transformations, consistent previous versions of the manuscript.

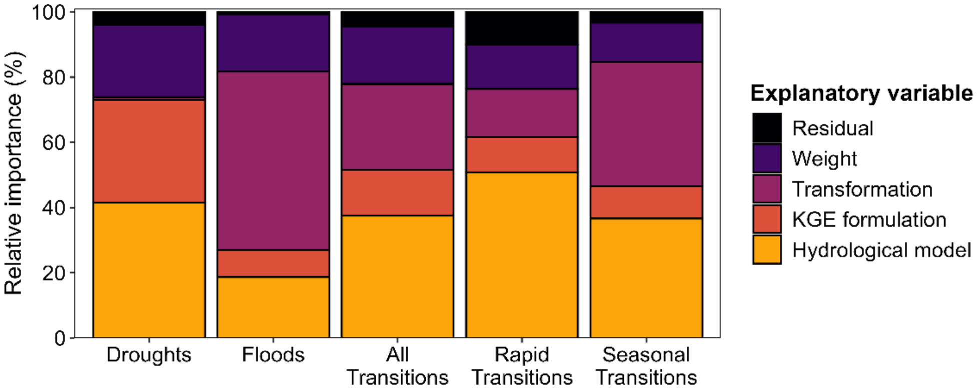

Figure 13: ANOVA test applied to the CSI. Figure updated in the main manuscript.

L382: Just highlighting another instance where I don't think "more parameters = more complex".

Thank you very much for pointing this out. As we mentioned in a similar previous comment, we have rewritten this so as not to misrefer to "more complexity = more parameters".

L412-L413: As a general repeat comment, I think the performance of GR5J and GR6J is so different from that of GR4J and TUW that calling model structure the most important choice is either incorrect (because the difference between the models with the higher performance - GR4J and TUW - seems minimal) or trivial (i.e. don't choose a model that's bad for the place you're trying to model). I think the argument about the importance of model structure made here would be much stronger if it can be based on the extra ANOVA analysis that uses just GR4J and TUW.

Thank you very much for your thoughts and suggestions. We have carried out the experiment mentioned, which has shown that the relative importance of methodological decisions is similar regardless of the combination of hydrological models used. Details of this experiment are presented in one of the previous answers (e.g., see Figure 14 and its description).

L483-L484: Maybe I'm misunderstanding something. As far as I understand, this information is available in Figure ST2.2 (model evaluation with SWE, ET, SM) but it doesn't seem to be used in the manuscript at all. Why do all the work to evaluate against these alternative data sets already but not use it, and then highlight the need to evaluate against more variables than just streamflow?

We included it in the first version of the manuscript (Figure 6 in the preprint available in EGUsphere). However, after the first round of revisions, we removed it to reduce complexity and the number of messages we were reporting as requested by the two reviewers and the editor. As we mentioned before, we consider these analyses as part of our model evaluation to perform "sanity checks" on hydrological variables other than runoff (e.g., poorly represented ET or SWE seasonality).

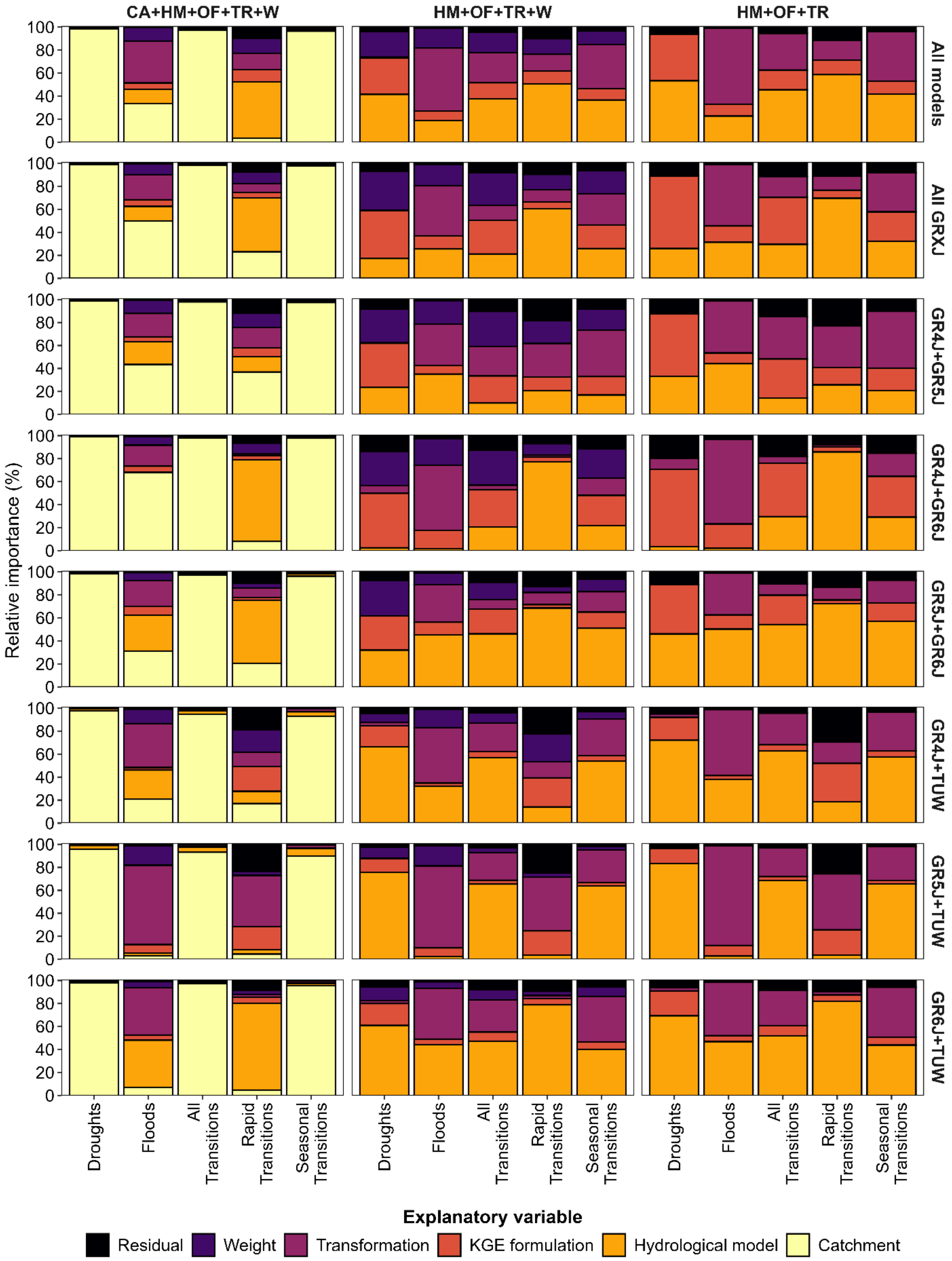

Figure 14: ANOVA test applied to the CSI considering different combinations of hydrological models and explanatory variables (CA: Catchment attributes, HM: Hydrological models, TR: streamflow transformations, W: weights). Figure included in the updated Supplementary Material.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Berghuijs, W. R., Hale, K., and Beria, H.: Technical note: Streamflow seasonality using directional statistics, Hydrology and Earth System Sciences, 29, 2851–2862, https://doi.org/10.5194/hess-29-2851-2025, 2025.

Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C., Michel, C., Andréassian, V., Bourgin, F., Brigode, P., Moine, N. L., Mathevet, T., Mouelhi, S., Oudin, L., Pushpalatha, R., and Valéry, A.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, 2023.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H.: Are we unnecessarily constraining the agility of complex process-based models?, Water Resources Research, 51, 716–728, https://doi.org/10.1002/2014WR015820, 2015.

Michel, C.: Hydrologie appliquée aux petits bassins ruraux, Cemagref, Antony, France, 1991.

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, Journal of Hydrometeorology, 18, 2215–2225, https://doi.org/10.1175/JHM-D-16-0284.1, 2017.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resources Research, 43, https://doi.org/10.1029/2005WR004723, 2007.

Vásquez, N. A., Mendoza, P. A., Lagos-Zuñiga, M., Scaff, L., Muñoz-Castro, E., and Vargas, X.: Robust spatial changes in climate classes: insights from bias-corrected CMIP6 models across Chile, Environ. Res. Lett., 20, 014061, https://doi.org/10.1088/1748-9326/ad9d5b, 2024a.

Vásquez, N. A., Mendoza, P. A., Knoben, W. J. M., Arnal, L., Lagos-Zúñiga, M., Clark, M., and Vargas, X.: The Key Role of Temporal Stratification for GCM Bias Correction in Climate Impact Assessments, Earth's Future, 12, e2023EF004242, https://doi.org/10.1029/2023EF004242, 2024b.

Woods, R. A.: Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks, Advances in Water Resources, 32, 1465–1481, https://doi.org/10.1016/j.advwatres.2009.06.011, 2009.

Zambrano-Bigiarini, M. and Rojas, R.: A model-independent Particle Swarm Optimisation software for model calibration, Environmental Modelling & Software, 43, 5–25, https://doi.org/10.1016/j.envsoft.2013.01.004, 2013.