



Supplement of

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

Eduardo Muñoz-Castro et al.

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

The copyright of individual parts of the supplement might differ from the article licence.

Contents of this file

This supplementary material file expands the results of the main manuscript to support and reinforce the findings presented there. The methodology used to generate the data and figures presented here is detailed in the main manuscript. The contents of this file are listed below:

- S1: Overall performance and hydrological consistency.
 - Figure S1: Calibration results.
 - Figure S2: Calibration results compared to the daily mean flow benchmark.
 - Figure S3: Model performance representing the seasonality of streamflow, snow water equivalent (SWE), and actual evapotranspiration (ET).
 - Figure S4: Comparison of the effect of including weights in the KGE variability term.
 - Figure S5: Comparison of the effect of streamflow transformations.
 - Figure S6: Comparison of the performance across hydrological models.
- Figure S7: Effects of the threshold used to define droughts and floods on the number of events per year.
- Figure S8: Effects of the threshold used to define droughts and floods on the models' performance in detecting streamflow extreme events.
- Figure S9: Effects of the overlapping temporal window used to identify the matches between observed and simulated droughts and floods.
- 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.
- Figure S11: CSI per type of streamflow extreme event, objective function, hydrological model and country.
- Figure S12: Results of the ANOVA applied to categorical indices.
- Figure S13: ANOVA test applied to the CSI considering different combinations of hydrological models and explanatory variables.
- Figure S14: Correlation between CSI and catchment attributes.
- Figure S15: Normalized calibrated parameters for the models tested.
- Figure S16: Forcing adjustment factors calibrated for each catchment.
- Figure S17: Parameter agreement across calibration configurations.
- Figure S18: Relative importance of parameters explaining the total variance of the CSI associated with drought, floods, and drought-to-flood transitions.
- Figure S19: Impacts of incorporating forcing adjustment parameters on parameter identifiability in the original models.
- Table S1: Parameter ranges used for calibrating the GRXJ models and the CemaNeige snow module.
- Table S2: Parameter ranges used for calibrating the TUW model.
- Table S3: Functions applied to transform the parameters of the GRXJ models.
- Table S4: Linkage between model outputs and hydrological states and fluxes.
- Table S5: Hydrological signatures computed.

S1: Overall performance and hydrological consistency

Model calibration assessment

We acknowledge that convergence of the calibration algorithm toward an optimal value of an objective function does not necessarily guarantee optimality in hydrological terms. Therefore, model performance must be evaluated to ensure that the calibrated parameters can fairly reproduce runoff generation processes in the study domain.

First, we compare the objective function values obtained from the different models (Figure S1). We found that inter-model differences are not considerable for the configurations with weights lower than two. However, when higher weights are applied, the GRXJ model's performance decreases. TUW exhibits the most “stable” performance across all configurations.

To further assess the predictive skill of our calibrated models, we compare the results with a simple benchmark based on daily mean flows (referred to as BM05 in Knoben, 2024¹). Specifically, for each day of the year, we compute the average streamflow over the calibration period and compare this reference time series with observations to calculate performance metrics Y (e.g., the KGE). Using this reference streamflow series, all objective function configurations are evaluated, yielding benchmark performance values that represent the minimum expected model accuracy.

To quantify the improvement achieved through calibration relative to the benchmark, the relative model performance is computed with the following equation:

$$Y^* = \frac{Y_{\text{Alternative}} - Y_{\text{Benchmark}}}{Y_{\text{Optimum}} - Y_{\text{Benchmark}}}$$

When Y_{Optimum} is 1 (e.g., KGE), Y^* ranges from $-\infty$ to 1, with 1 indicating optimal performance. As shown in Figure S2, all model configurations outperform the benchmark (i.e., $Y^* > 0$), indicating that our model has greater predictive skill than the long-term average streamflow series.

Overall accuracy and hydrological consistency

Our calibration approach relies on 60 objective functions derived by combining 5 KGE formulations, 3 streamflow transformations, and 4 weights. Since the objective function varies across configurations, its absolute values are not directly comparable. Here, the model's performance is assessed using biases in a set of hydrological signatures (Figure S3-S6), including seasonality, statistical properties (mean, variance), flow duration curve-derived signatures (e.g., mid-segment slope), and annual extremes (more details in Table S5).

For each hydrological signature, performance is quantified using the following efficiency metric:

$$HS_{\text{eff}} = 1 - \left| \frac{HS_{\text{sim}}}{HS_{\text{obs}}} - 1 \right|$$

Where HS represents a hydrological signature derived from the simulations (HS_{sim}) and the observations (HS_{obs}). Similar to, e.g., KGE, HS_{eff} ranges from $-\infty$ to 1, with 1 indicating optimal performance. Then, values closer to 1 indicate a better agreement between simulations and observations.

¹ Knoben, W. J. (2024). Setting expectations for hydrologic model performance with an ensemble of simple benchmarks. *Hydrological Processes*, 38(10), e15288.

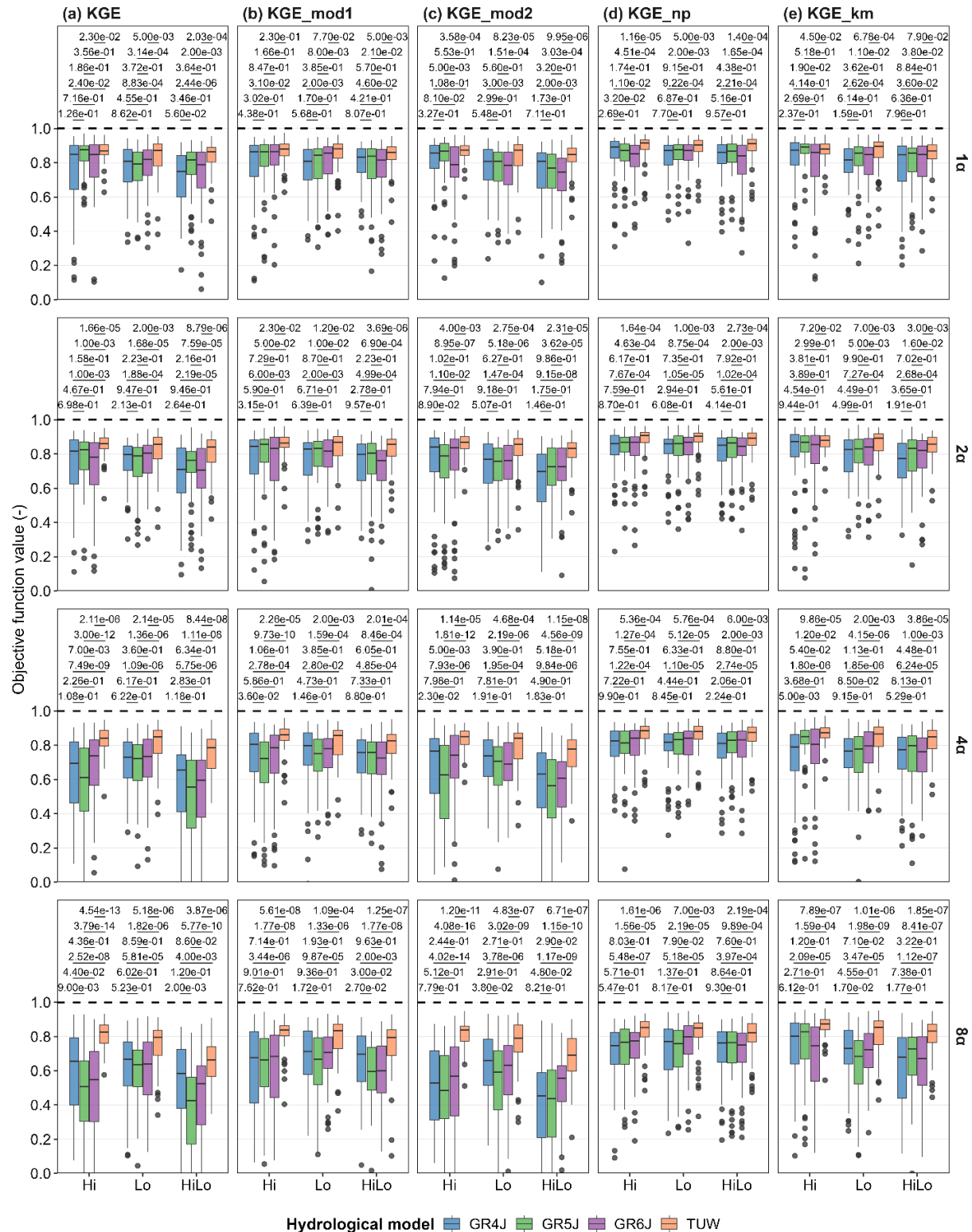


Figure S1: Calibration results for the different configurations tested. The x-axis shows the case associated with streamflow transformations, the rows show the use of weights in the KGE variability term, and the columns show the KGE formulation. The values reported on the y-axis correspond to the evaluation of the corresponding calibration configuration. Each boxplot comprises results from all 63 catchments in our study domain. The p-values corresponding to the Wilcoxon statistical significance test are included. Differences across configurations (i.e., 6 combinations of model pairs) are evaluated.

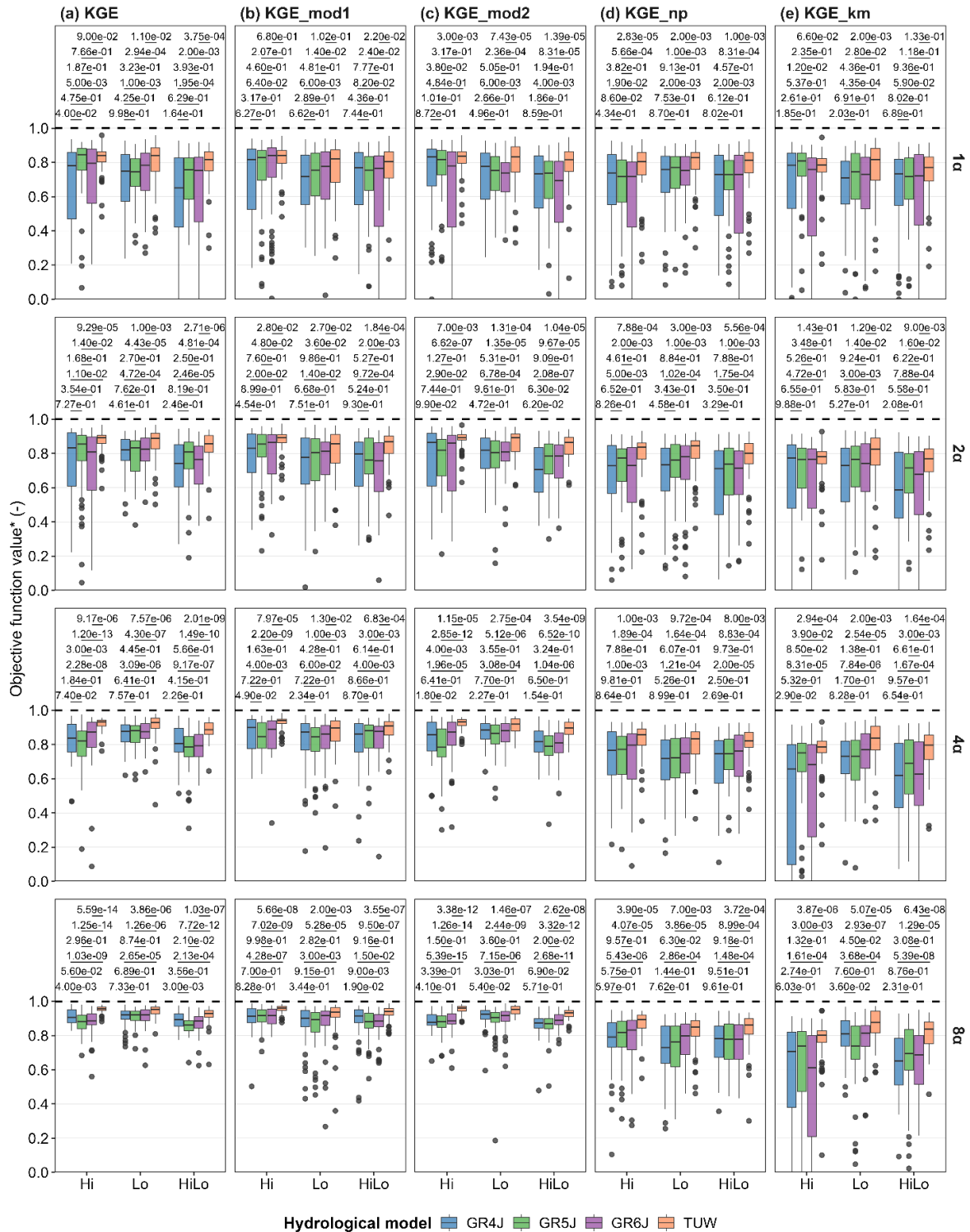


Figure S2: Calibration results relative to a daily mean flow benchmark for the different configurations tested. The x-axis shows the case associated with streamflow transformations, the rows show the use of weights in the KGE variability term, and the columns show the KGE formulation. The values reported on the y-axis correspond to the evaluation of the corresponding calibration configuration. Each boxplot comprises results from all 63 catchments in our study domain. The p-values corresponding to the Wilcoxon statistical significance test are included. Differences across configurations (i.e., 6 combinations of model pairs) are evaluated.

To assess whether the models can capture (some) key hydrological processes at the catchment scale, we analyze their ability to reproduce the seasonal timing (seasonality) of streamflow (Q), snow water equivalent (SWE), and actual evapotranspiration (ET). To minimize the influence due to, e.g., systematic biases in reference products other than streamflow (e.g., ET retrieved from GLEAM), we quantify seasonality using directional statistics (Berghuijs et al., 2025)². This approach focuses on the timing of annual cycles rather than their magnitude. Note that the hydrological year differs between the study regions: in Chile (Southern Hemisphere) it runs from April to March, whereas in Switzerland (Northern Hemisphere) it runs from October to September. Model performance is shown in Figure S3. For configurations with weights < 2, results across configurations are comparable (i.e., there are no significant differences). In contrast, higher weights lead to a systematic reduction in performance.

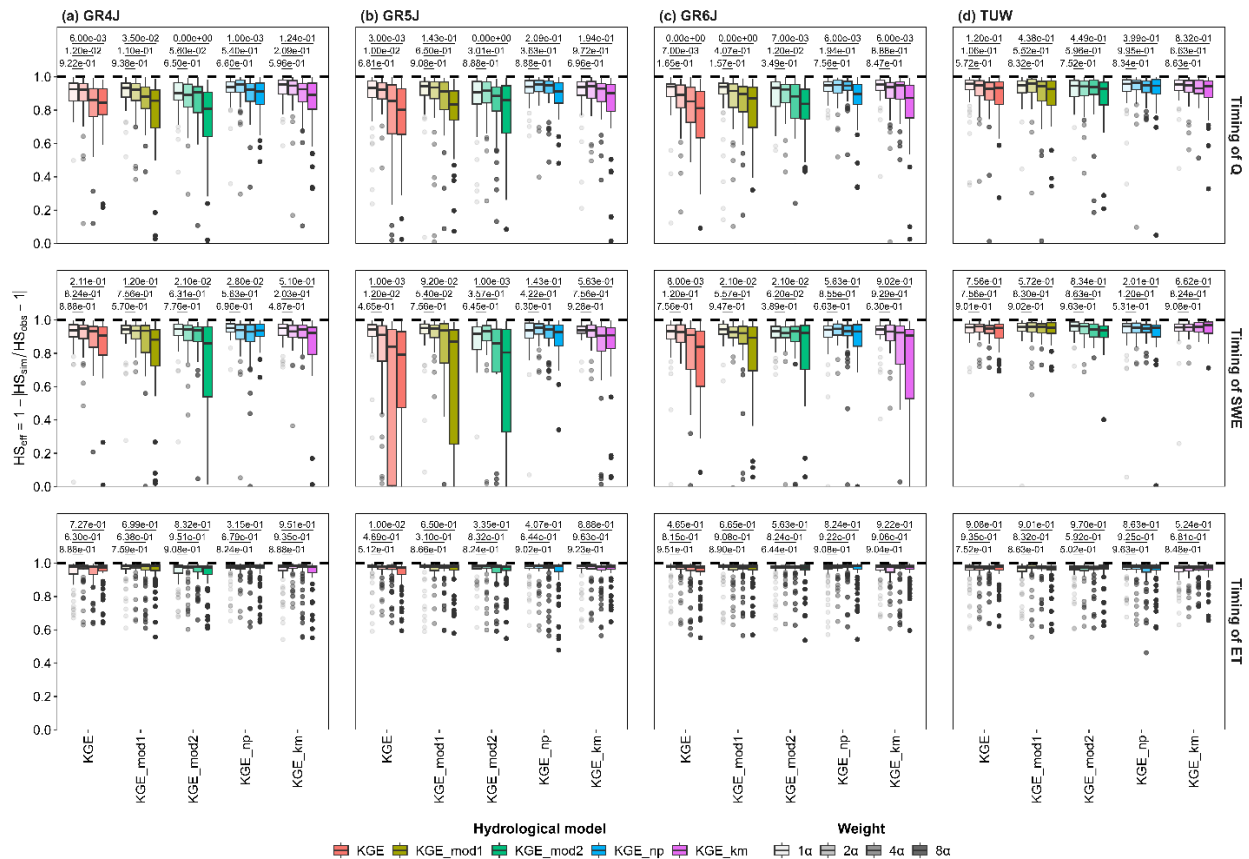


Figure S3: Model performance representing the seasonality of streamflow (Q), snow water equivalent (SWE), and actual evapotranspiration (ET), for different KGE formulations and weights, and the HiLo streamflow transformation. The p-values for the Wilcoxon significance test are provided. The differences between the unweighted case (i.e. 1 α) and the application of different weights are evaluated.

To further isolate the effect of weighting on streamflow simulations, we evaluate model performance using a set of hydrological signatures. To do so, as a sample case, we considered the original KGE formulation combined with the HiLo streamflow transformation. As shown in Figure S4, weights larger than two consistently degrade performance across indices, confirming the pattern observed in Figure S3.

We then examine the influence of streamflow transformations. Figure S5 shows that, although overall differences among transformations are modest, they are pronounced for specific indices for which such an impact could theoretically be expected. For example, the ‘Lo’ transformation yields superior performance for low-flow metrics compared to the other transformations. Finally, a comparison across models (Figure S6) indicates broadly similar performance, with no statistically significant differences among them.

² Berghuijs, W. R., Hale, K., and Beria, H.: Technical note: Streamflow seasonality using directional statistics, *Hydrol. Earth Syst. Sci.*, 29, 2851–2862, <https://doi.org/10.5194/hess-29-2851-2025>, 2025.

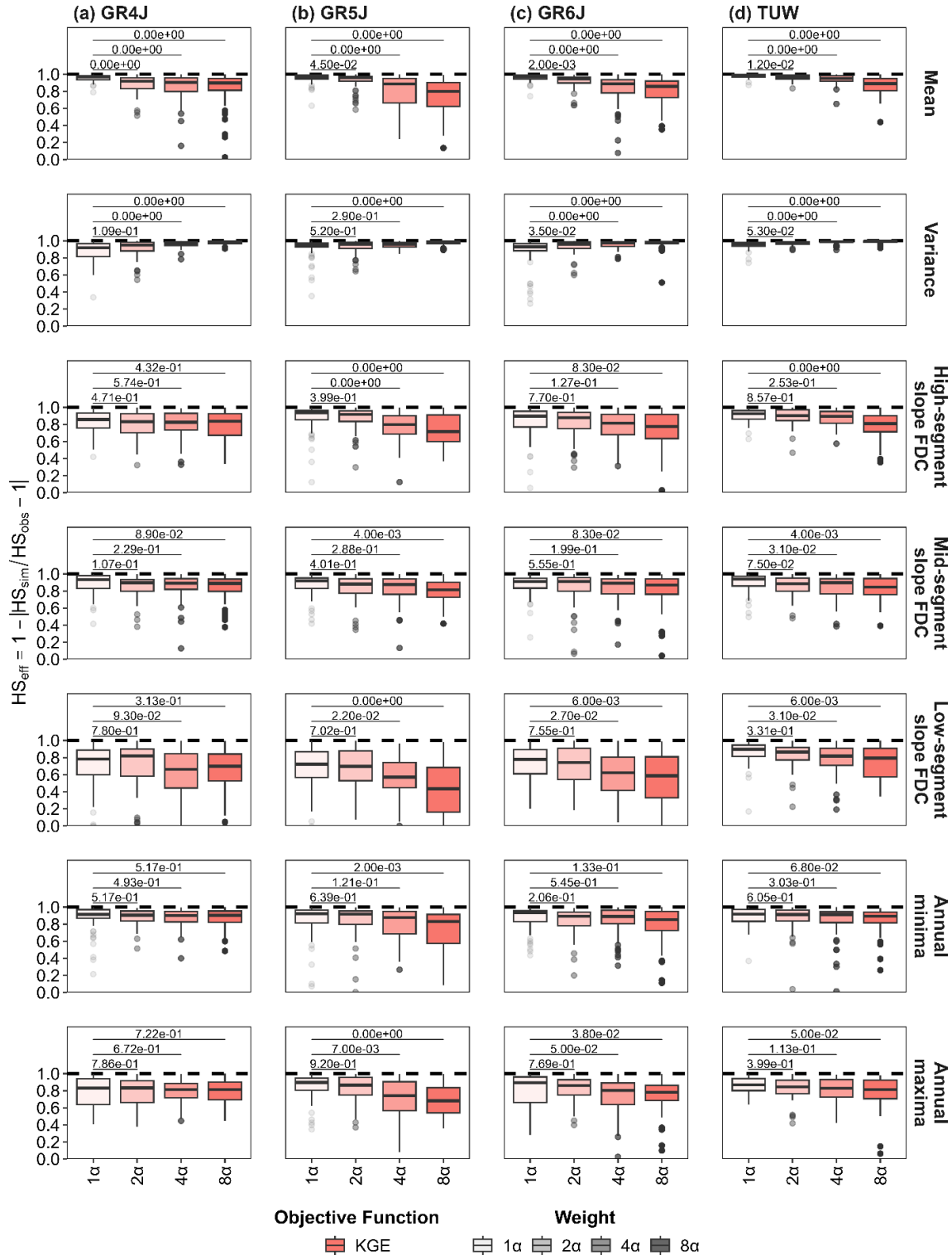


Figure S4: Comparison of the effect of including weights in the KGE variability term on different streamflow signatures shown in the different rows: mean flow, flow variance, slope of the high segment of the FDC, slope of the mid segment of the FDC, slope of the low segment of the FDC, annual minima, and annual maxima. Example for the original KGE formulation and HiLo transformation. The dashed black lines indicate the optimum values for the assessed metrics. Each boxplot contains 63 values (i.e., one per catchment). The p-values for the Wilcoxon significance test are provided. The differences between the unweighted case (i.e. 1 α) and the application of different weights are evaluated.

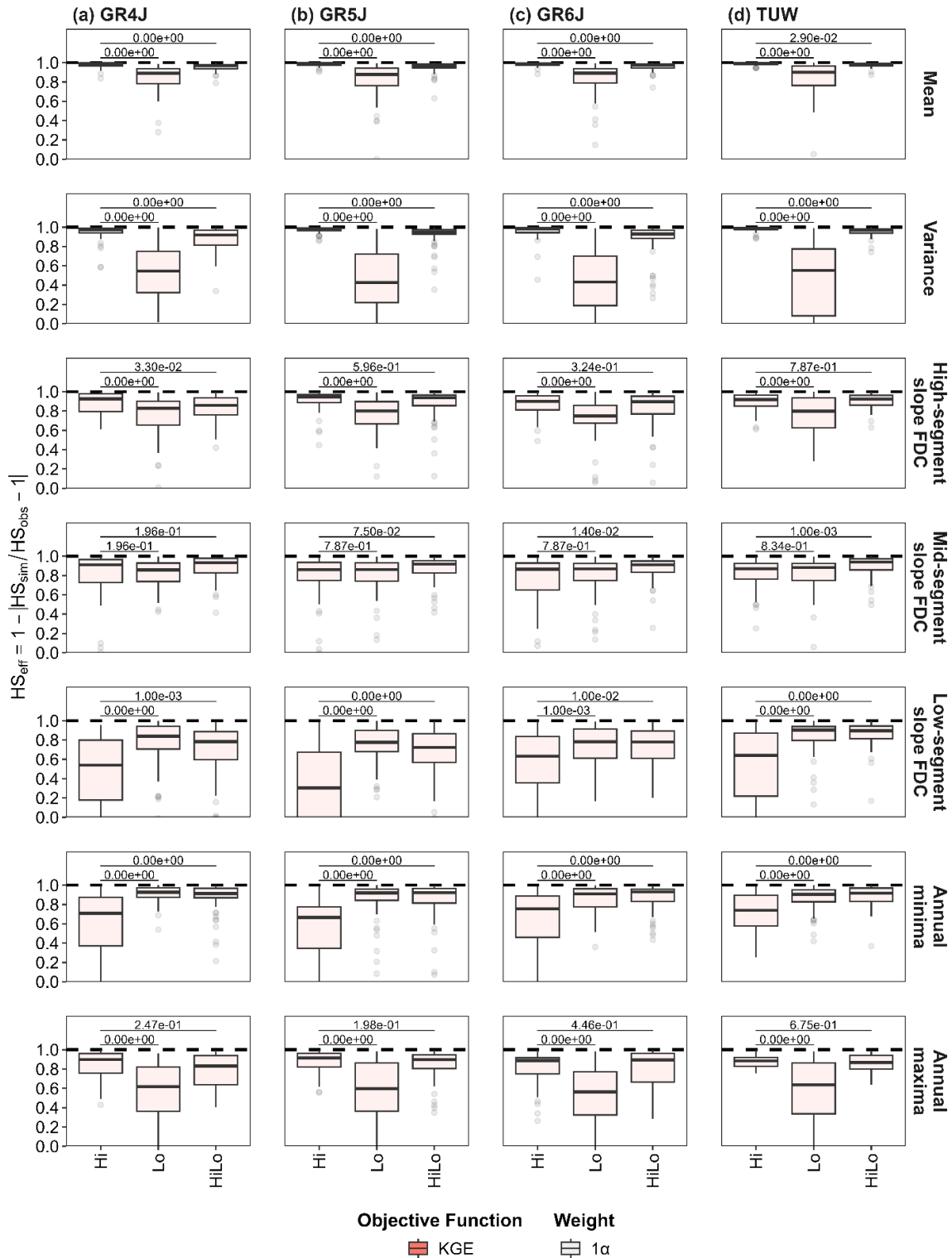


Figure S5: Comparison of the effect of streamflow transformations on different streamflow signatures shown in the different rows: mean flow, flow variance, slope of the high segment of the FDC, slope of the mid segment of the FDC, slope of the low segment of the FDC, annual minima, and annual maxima. Example for the unweighted original KGE formulation. The dashed black lines indicate the optimum values for the assessed metrics. Each boxplot contains 63 values (i.e., one per catchment). The p-values for the Wilcoxon significance test are provided. The differences between no transformations (i.e. Hi) and the application of low (Lo) and a combined HiLo transformations are evaluated.

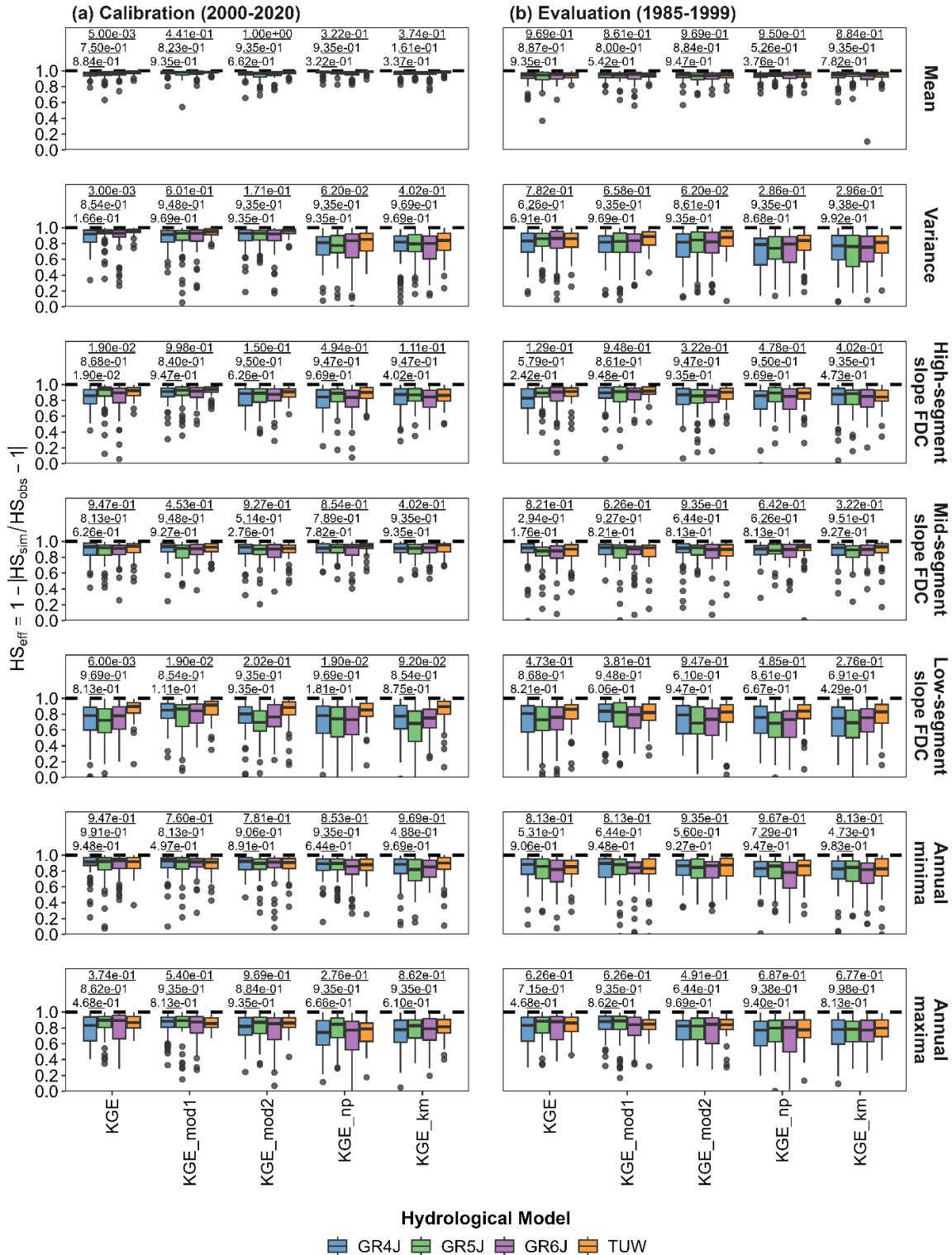


Figure S6: Comparison of the performance across hydrological models for different streamflow signatures shown in the different rows: mean flow, flow variance, slope of the high segment of the FDC, slope of the mid segment of the FDC, slope of the low segment of the FDC, annual minima, and annual maxima. Example for different unweighted and HiLo transformation KGE formulations. The dashed black lines indicate the optimum values for the assessed metrics. Each boxplot contains 63 values (i.e., one per catchment). The p-values corresponding to the Wilcoxon statistical significance test are included. The p-values for the Wilcoxon significance test are provided. In the statistical test, the results are compared with the GR4J model.

Figure S7

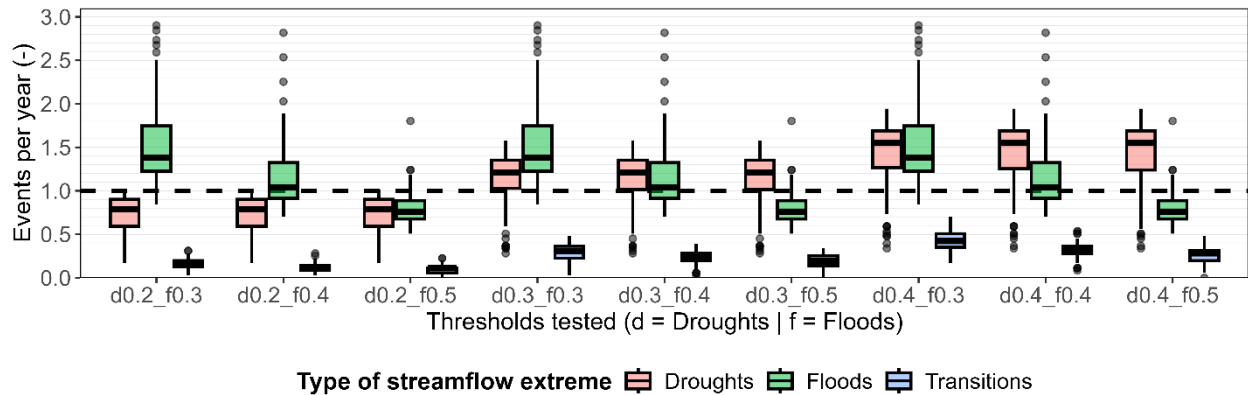


Figure S7: Effects of the threshold used to define droughts and floods on the models' performance in detecting streamflow extreme events. Number of events per year depending on the threshold used for defining droughts and floods. The notation "dX_fY" refers to the use of the Xth and Yth percentile to define the variable and fixed threshold required to identify streamflow droughts (d) and floods (f), respectively.

Figure S8

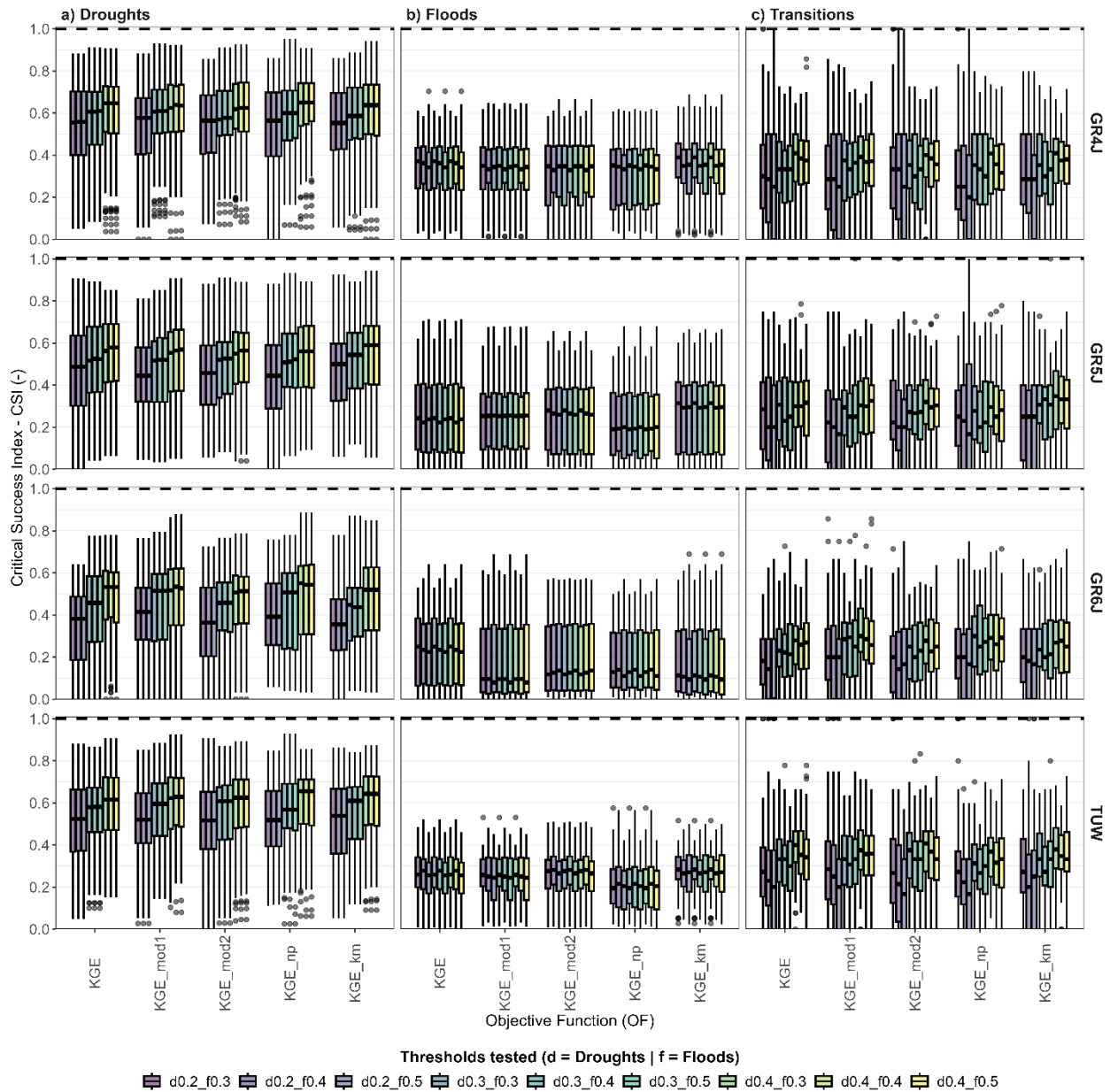


Figure S8: Effects of the threshold used to define droughts and floods on the models' performance in detecting streamflow extreme events. Performance of the GR4J, GR5J, GR6J and TUW models (in the rows) in detecting a) droughts, b) floods, and c) transitions, according to different thresholds used for the identification of streamflow extreme events. For each type of extreme event and hydrological model, the results are compared according to different formulations of the KGE (unweighted and HiLo) used as objective functions.

Figure S9

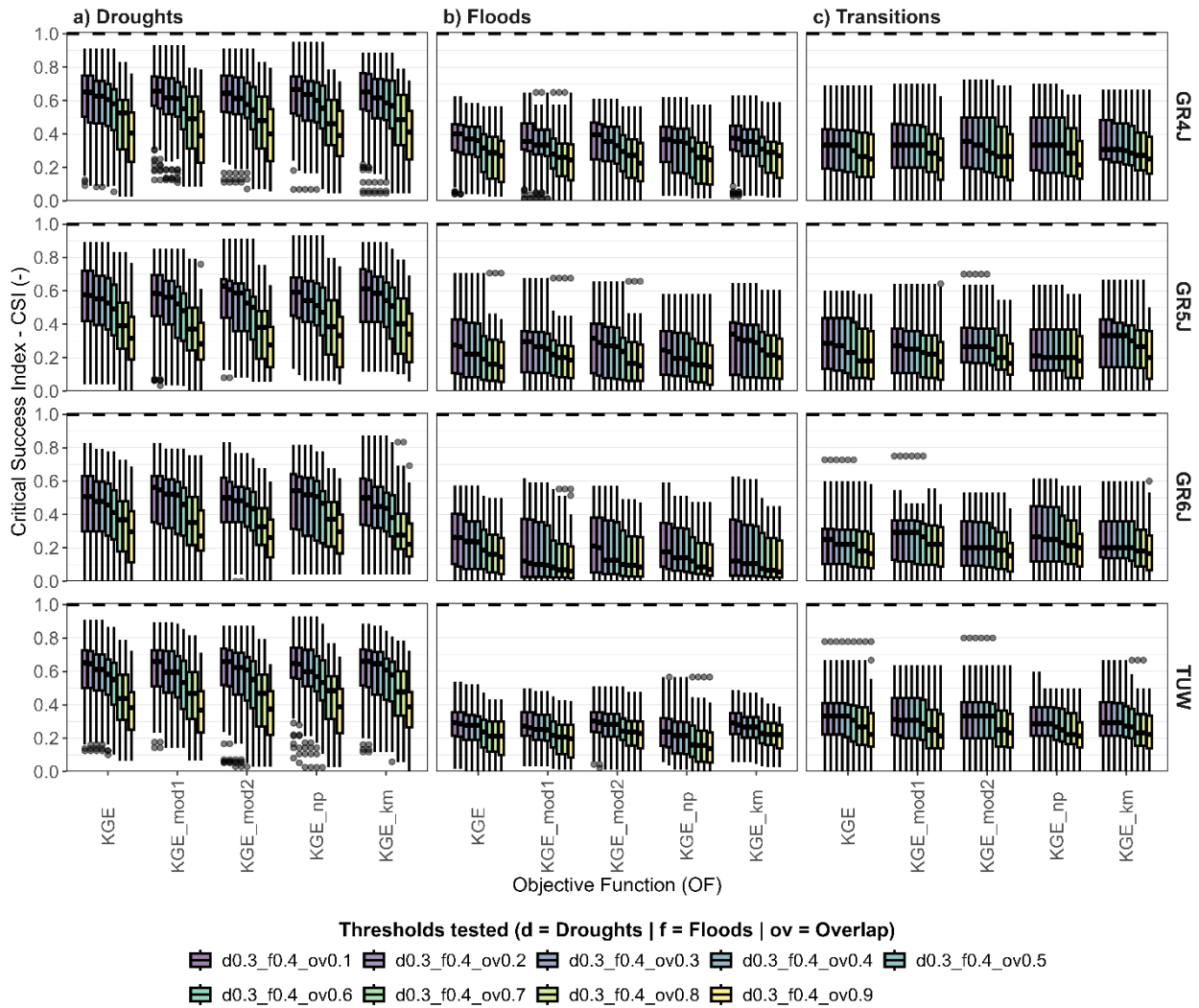


Figure S9: Effects of the overlapping temporal window used to identify observed droughts and floods. Performance of the GR4J, GR5J, GR6J and TUW models (in the rows) in detecting a) droughts, b) floods, and c) transitions, according to different overlapping temporal window thresholds (different colors) used for the identification of streamflow extreme events. For each type of extreme event and hydrological model, the results are compared according to different formulations of the KGE (unweighted and HiLo) used as objective functions (x-axis).

Figure S10

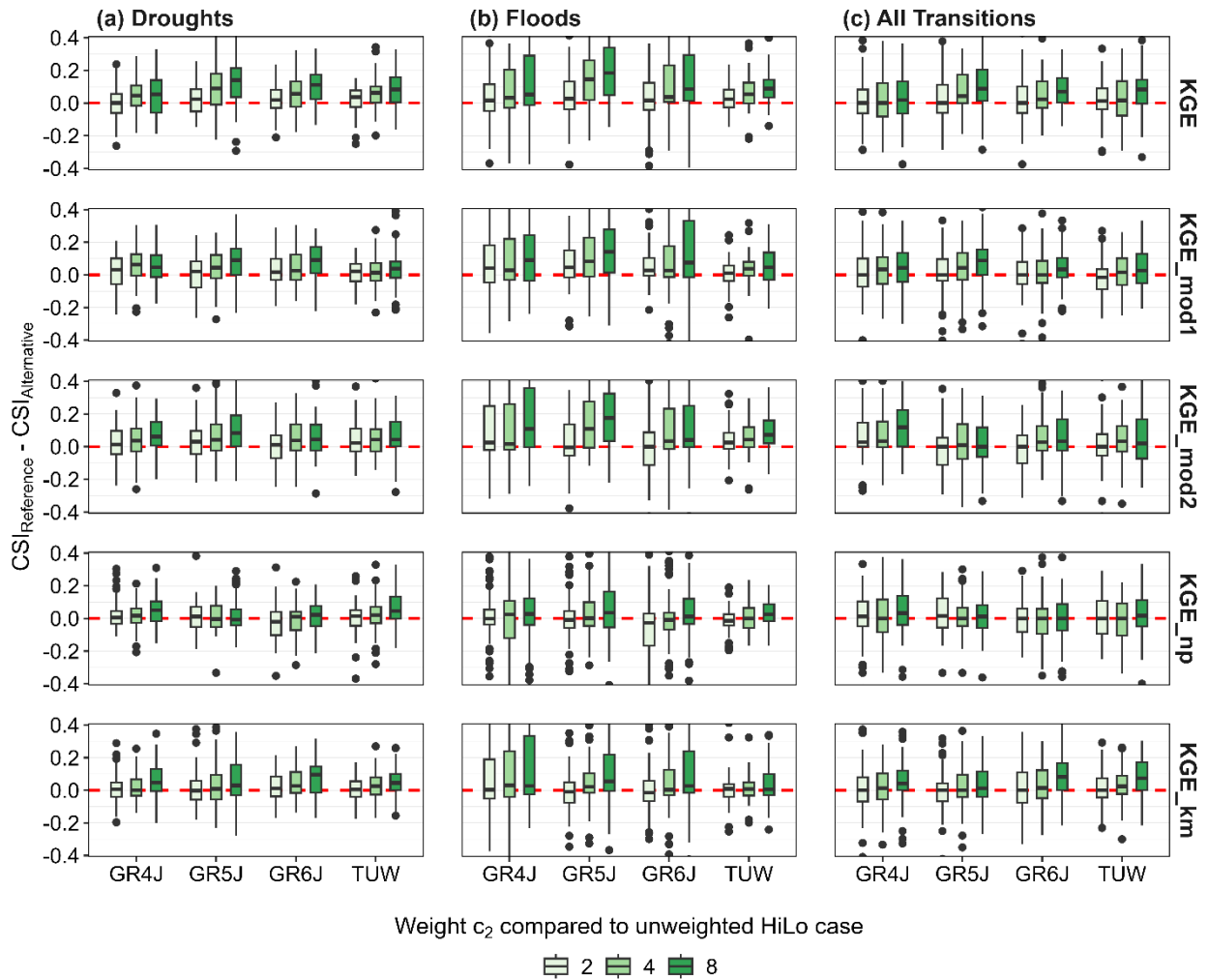


Figure S10: Difference in the CSI by using the no weights HiLo case (reference) and different weights (alternative) on the variability term of the KGE for different hydrological models. Differences are calculated as "reference - alternative" with values above (below) 0 indicating better (worse) performance 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.

Figure S11

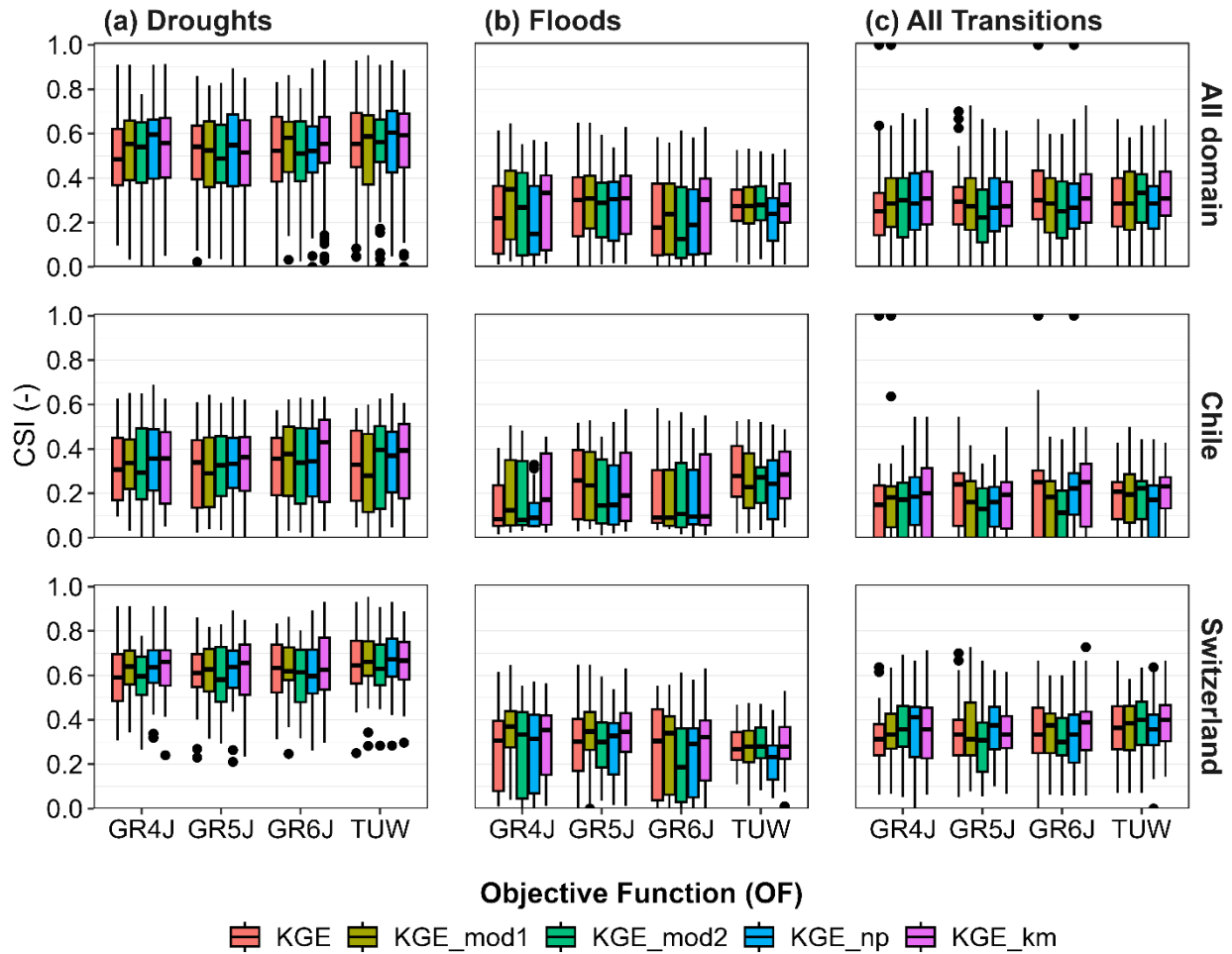


Figure S11: CSI per type of streamflow extreme event, objective function, hydrological model and country. Critical Success Index (CSI) for a) droughts, b) floods, and c) drought-to-flood transitions, based on the simulations with GR4J, GR5J, GR6J, and TUV calibrated with different unweighted HiLo KGE formulations as objective functions for All domain, Chile, and Switzerland (upper, middle, and lower panels respectively). Each boxplot contains the corresponding number of catchments (i.e., All domain = 63; Chile = 24; Switzerland = 39). Supplementary figure associated with Figure 6 in the main manuscript.

Figure S12

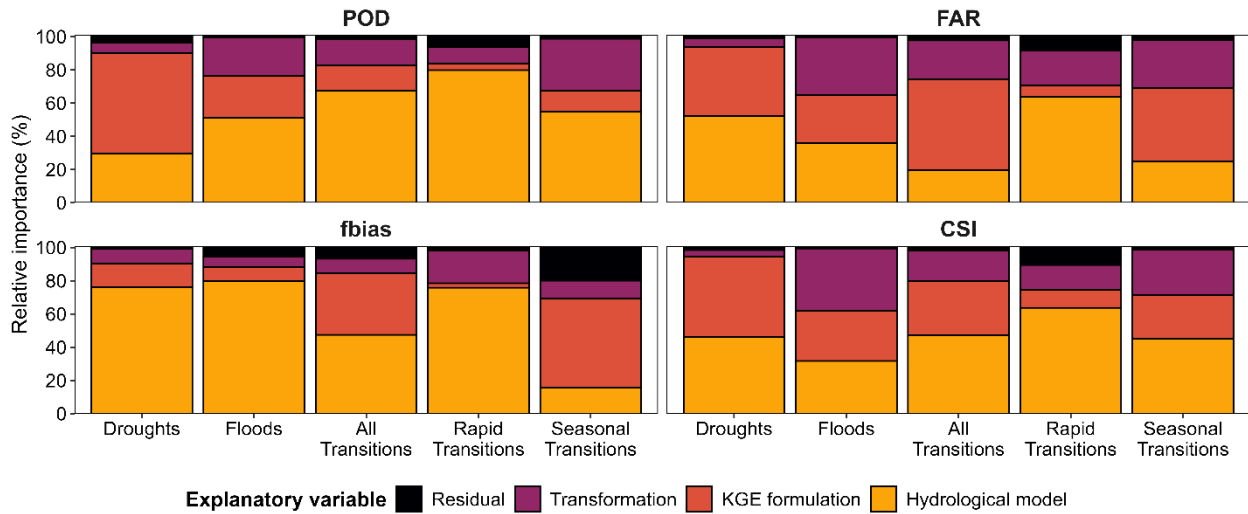


Figure S12: Results of the ANOVA applied to categorical indices. Results of the analysis of variance (ANOVA) applied to probability of detection ($POD = H/H+M$), false alarm ratio ($FAR = F/H+F$), frequency bias ($fbias = H+F/H+M$), critical success index ($CSI = H/H+M+F$) for droughts, floods, all drought-to-flood transitions (i.e., rapid and seasonal), rapid transitions, and seasonal transitions. Supplementary figure associated to Figure 8 in the main manuscript.

Figure S13

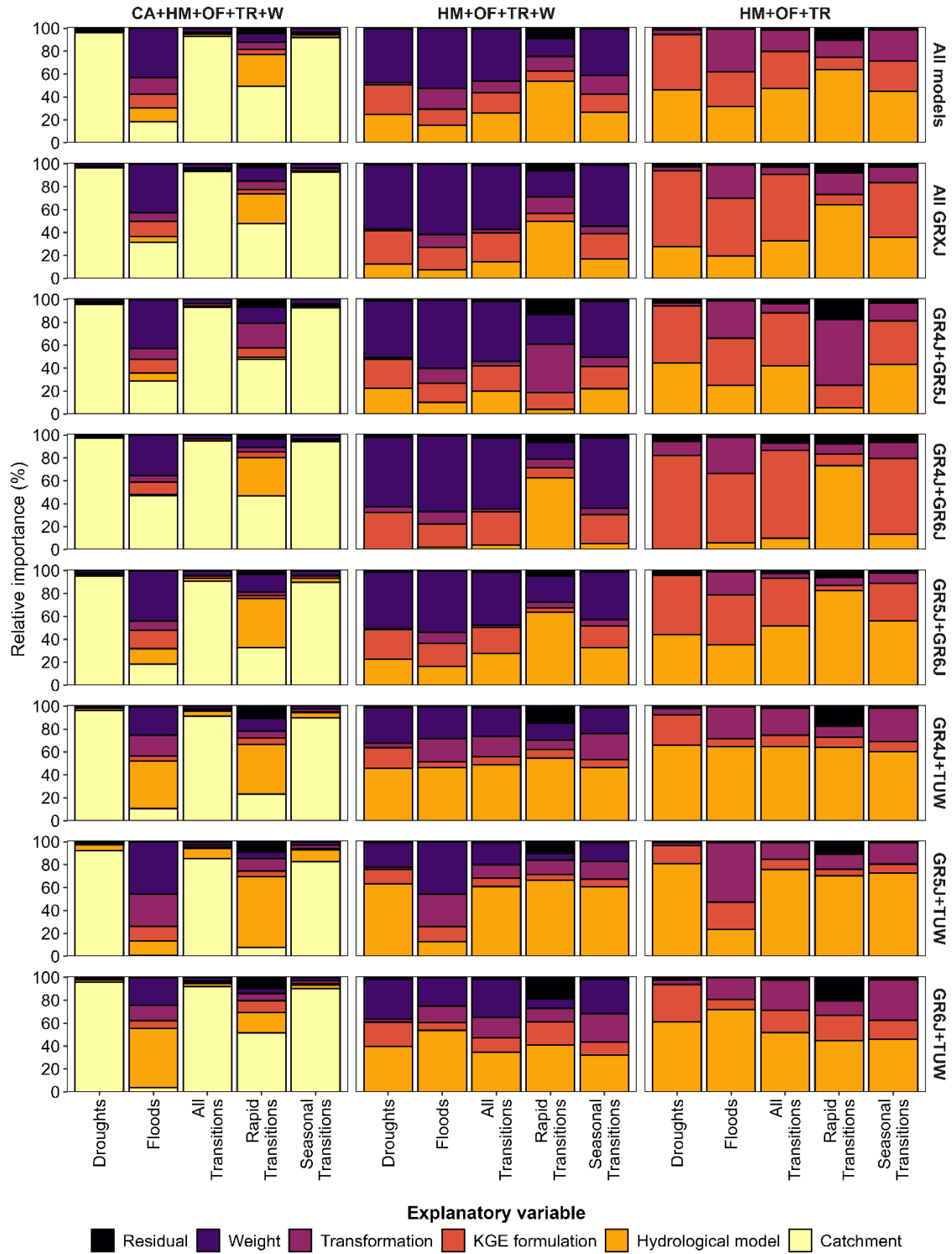


Figure S13: ANOVA test applied to the CSI considering different combinations of hydrological models and explanatory variables. CA: Catchment attributes | HM: Hydrological models | TR: streamflow transformations | W: weights.

Figure S14

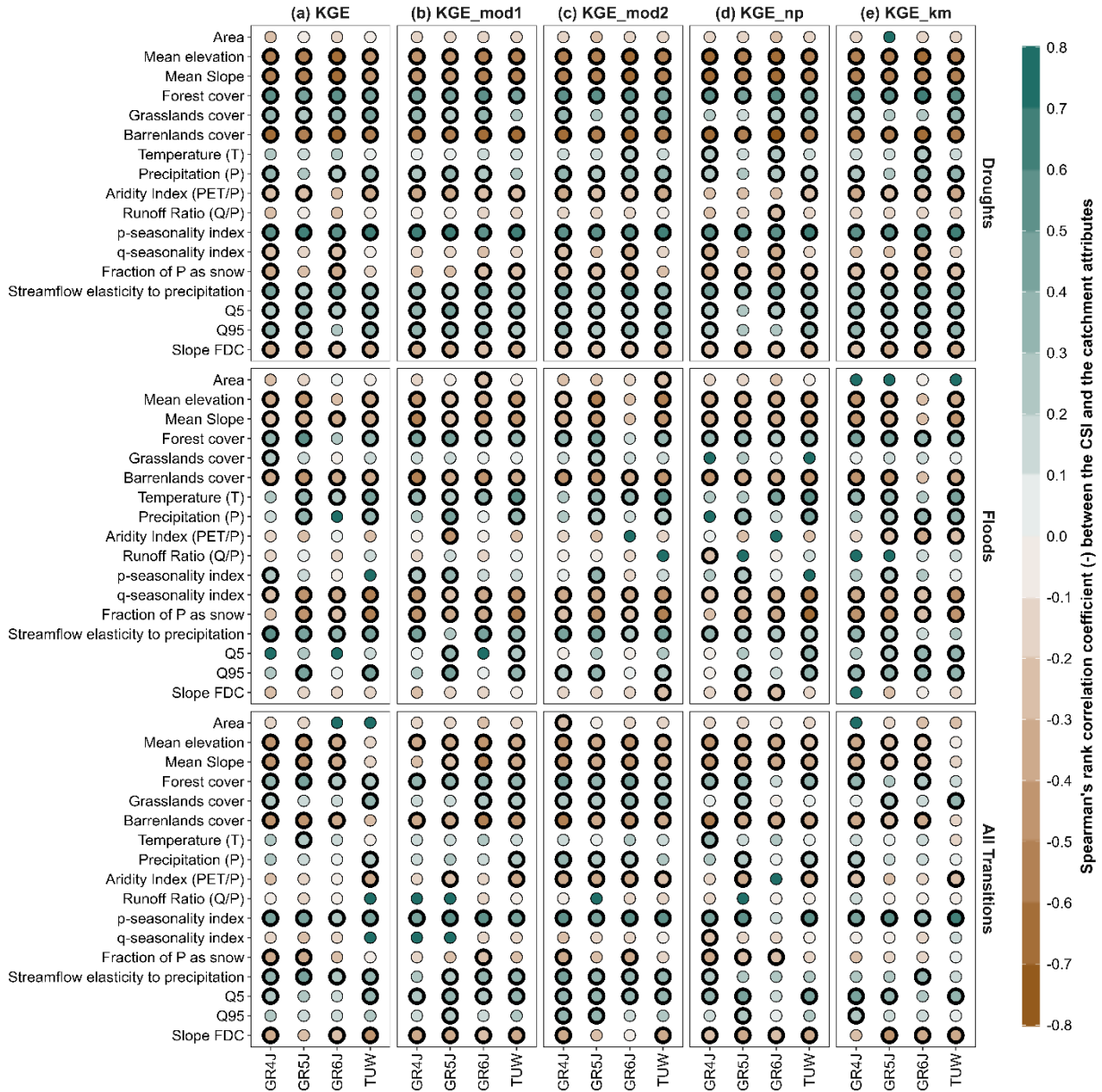


Figure S14: Correlation between CSI and catchment attributes. Correlation between CSI and catchment attributes based on results associated to different unweighted HiLo KGE formulations used as objective functions (columns) and streamflow extremes (rows). Supplementary figure associated to Figure 9 in the main manuscript.

Figure S15

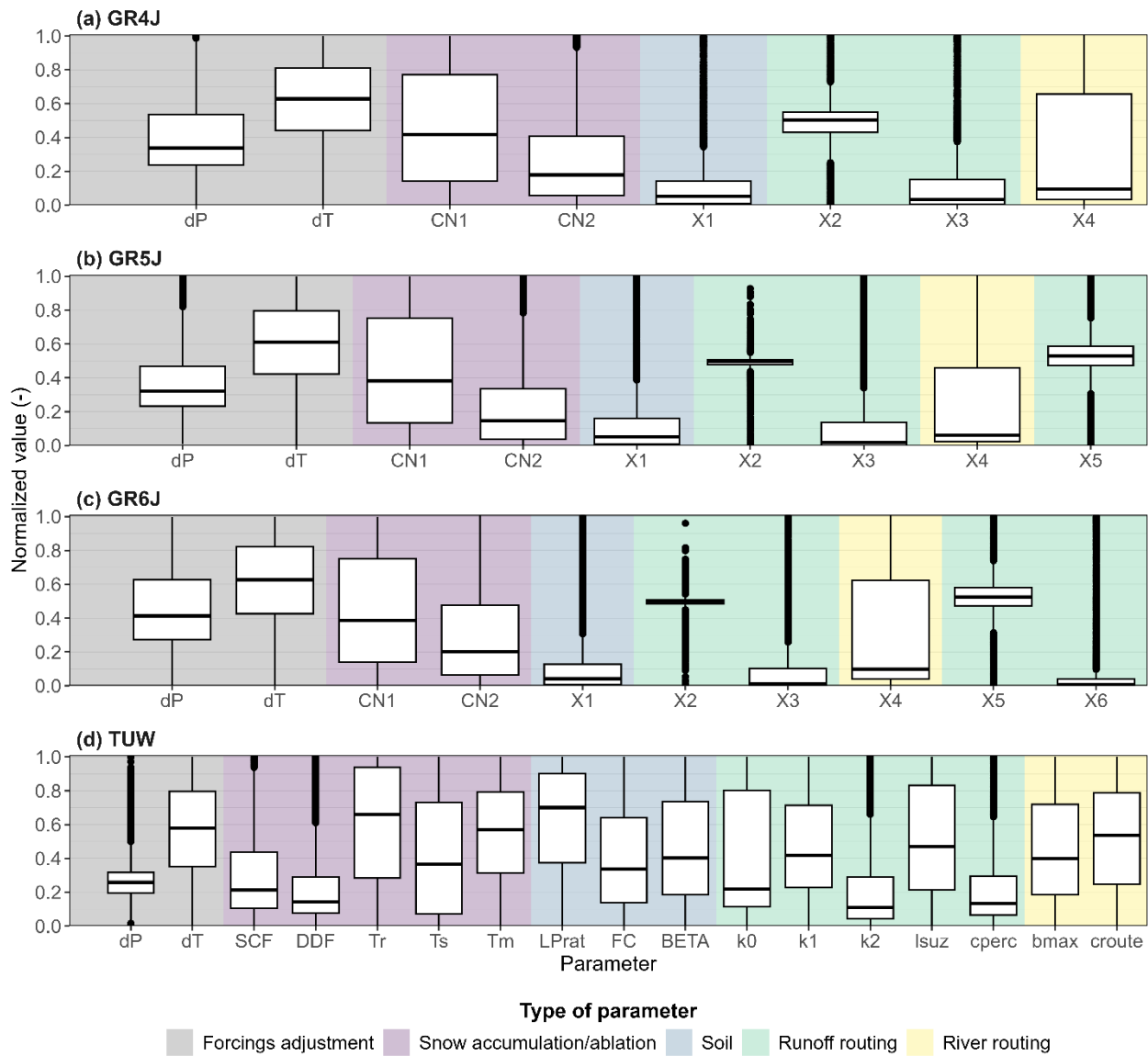


Figure S15: Normalized calibrated parameters for the models tested. Calibrated parameters normalized according to their plausible range (Tables S1 and S2). Each boxplot contains 60 sets of parameters for each of the 63 basins (i.e., 60 x 63 = 3780 values).

Figure S16

Because the forcing adjustment factors have the same meaning across all the models tested here, we present the interaction between dP and dT . Then, we can explore (i) how different they are for the same catchment, and (ii) the potential compensations for model deficiencies that these parameters could have.

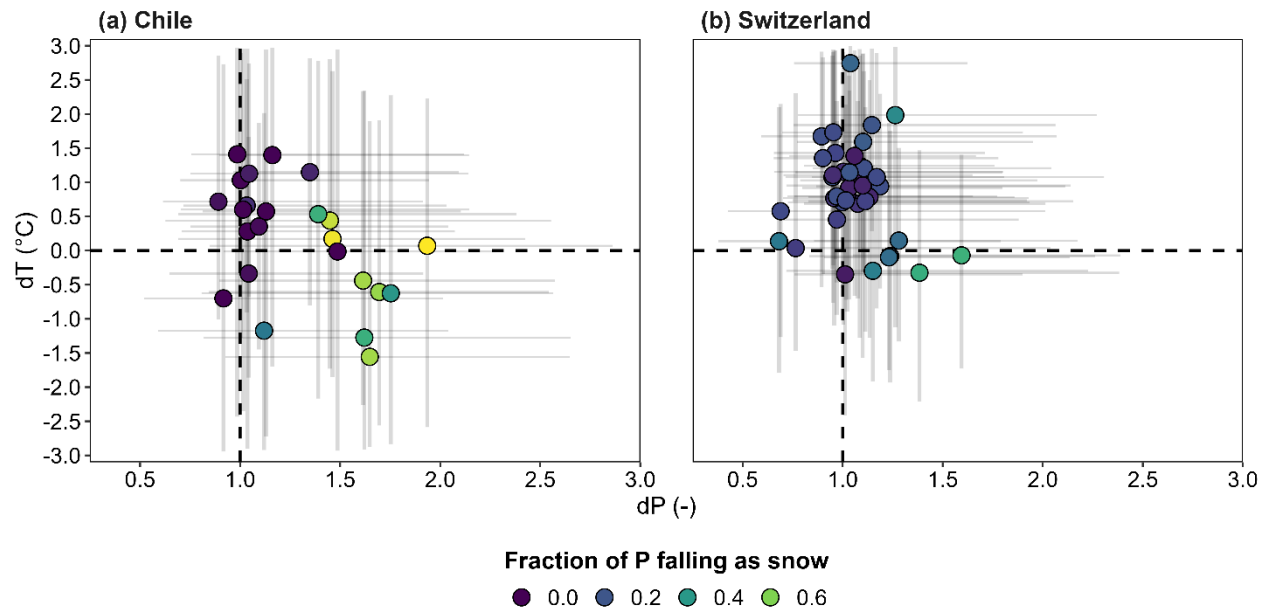


Figure S16: Forcing adjustment factors calibrated for each catchment. Forcing adjustment factors calibrated for each catchment for (a) Chile, and (b) Switzerland. The bars indicate the 10th and 90th percentiles across configurations (i.e., 60 x 4 models) per catchment, while the central shape is the 50th percentile for each catchment (i.e., 63 = 24 Chile + 39 Switzerland).

Figure S17

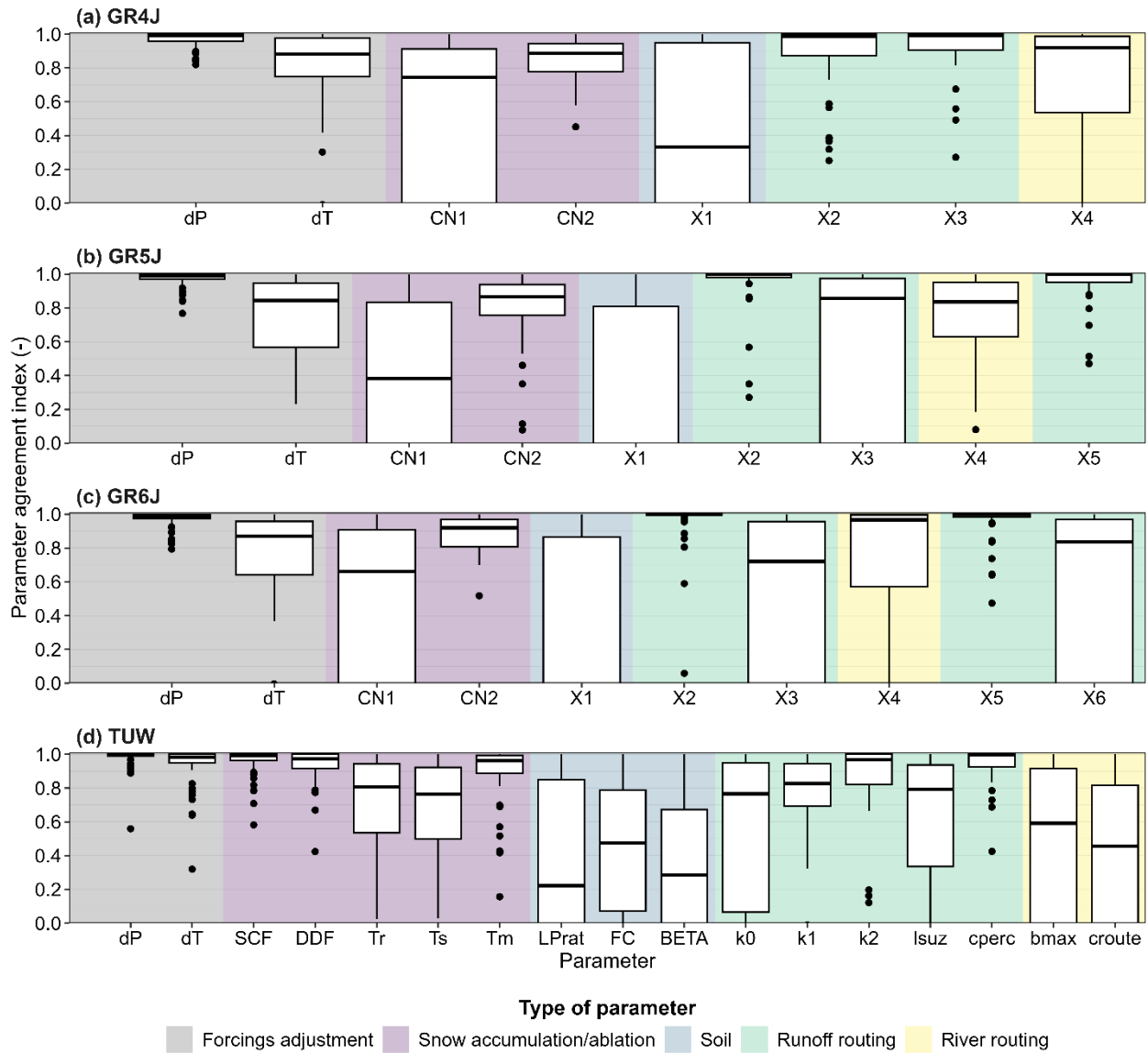


Figure S17: Parameter agreement across calibration configurations. Parameter agreement in a) GR4J, b) GR5J, c) GR6J, and d) TUV models. (Lower) Higher values in the parameter agreement index indicate (dis)agreement in the values of the parameter (i.e., more dispersion between the optimal parameter sets obtained from different calibration processes). Each boxplot comprises agreement indices from the 63 catchments included in the study domain. The parameter agreement index for each parameter and catchment – as well as the overall agreement index - has been computed using the metric proposed by Muñoz-Castro et al. (2023) .

Figure S18

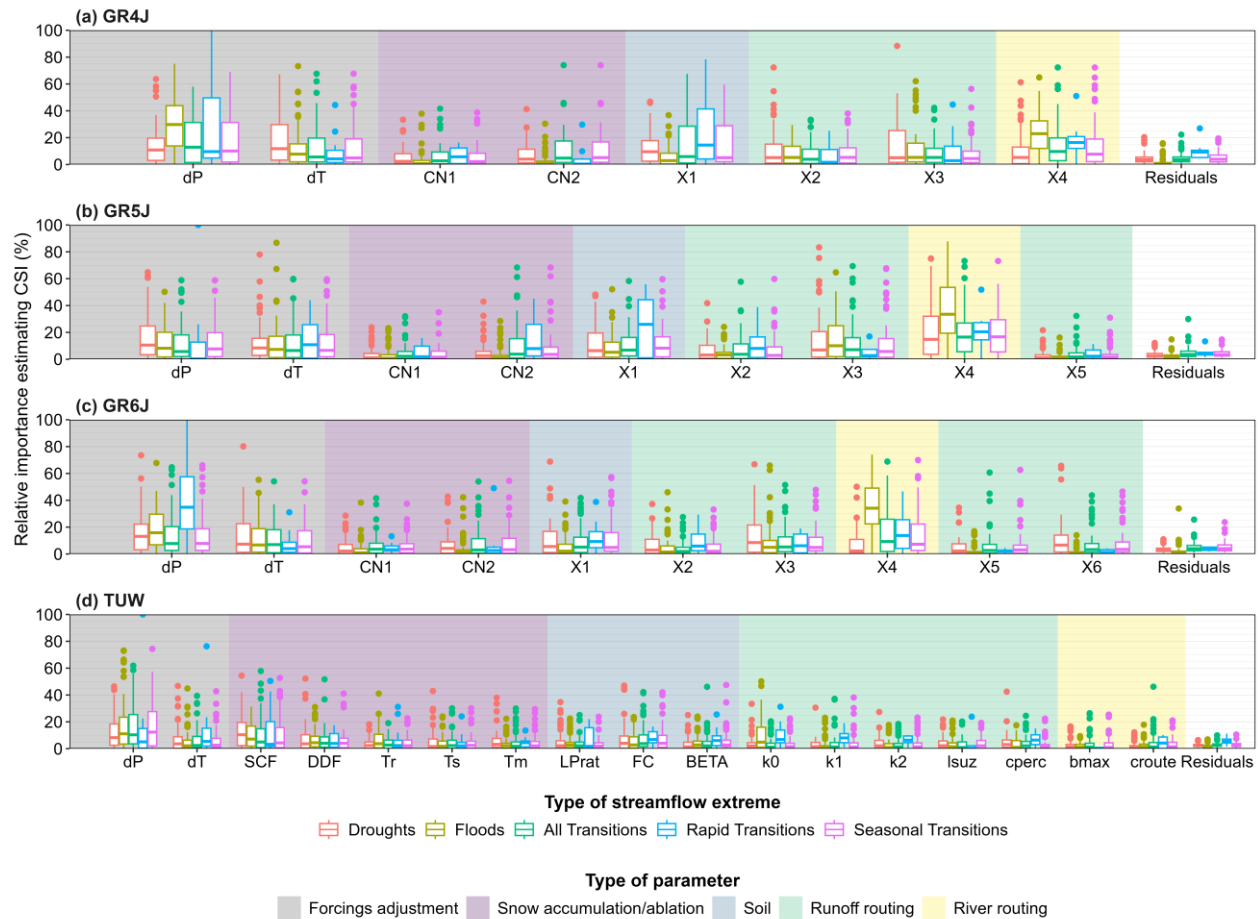


Figure S18: Relative importance of parameters explaining the total variance of the CSI associated with drought, floods, and drought-to-flood transitions. Relative importance of parameters for explaining the Critical Success Index (CSI) for models (a) GR4J, (b) GR5J, (c) GR6J, and (d) TUW based on the results of an analysis of variance (ANOVA). Each boxplot contains 63 values (i.e., one per catchment). Extended version of Figure 10 in the main manuscript.

Figure S19

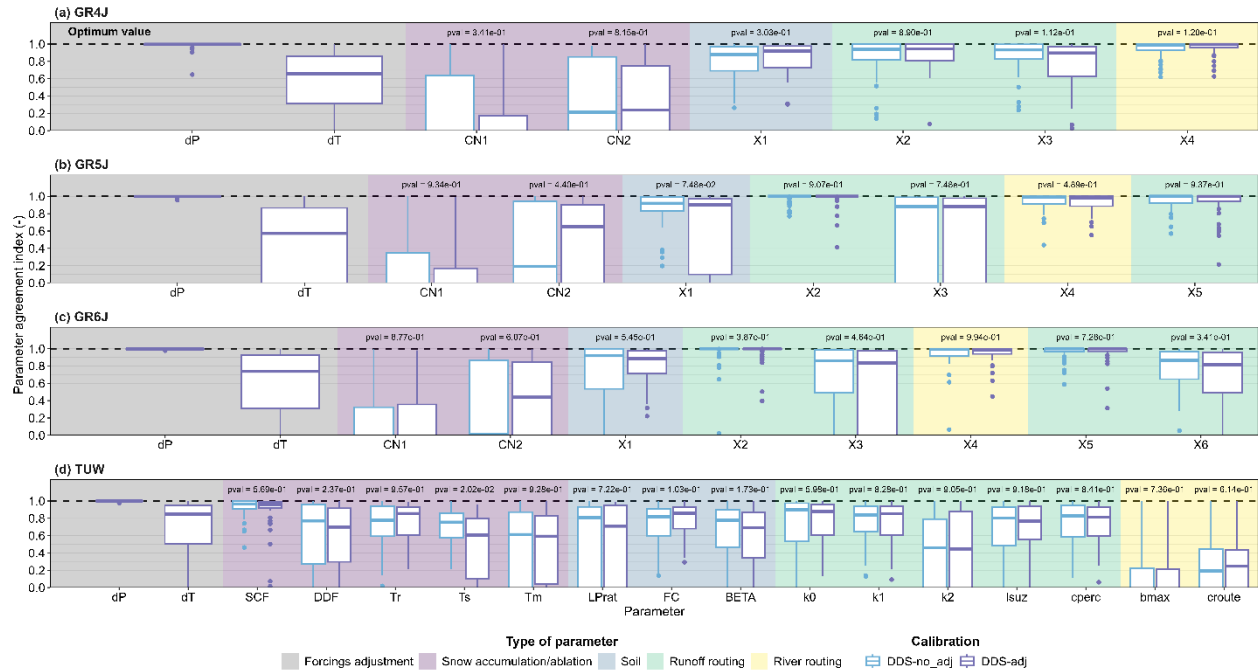


Figure S19: Impacts of incorporating forcing adjustment parameters on parameter identifiability in the original models. Comparison of parameter agreement index (R) with and without forcing adjustment parameters. Each boxplot contains 63 values, one per basin. The p-values (pval) correspond to the Wilcoxon statistical significance test. The agreement index is computed as $R = 1 - (095\% - 05\%)/|050\%|$. Then values close to (far from) 1 indicate high (low) agreement between parameters identified as equifinal.

Table S1: Parameter calibration range used for the GRXJ models and the CemaNeige snow module

Description of parameters in the GRXJ hydrological models and the CemaNeige snow module and plausible range defined for calibration. The forcing adjustment parameters introduced in the model calibration process are highlighted in gray.

Parameter	Type	Units	Description	Plausible range	
				Lower	Upper
dP	Forcing adjustment	%	Precipitation adjustment factor	0.25	3
dT	Forcing adjustment	°C	Temperature adjustment parameter	-3	3
CN1	Snow	-	Weighting coefficient for snowpack thermal state	0	1
CN2	Snow	mm/°C/day	Degree-day melt coefficient	0.01	20
X1	Soil	mm	Production store capacity	0.05	4000
X2	Runoff routing	mm/day	Intercatchment exchange coefficient	-10	10
X3	Runoff routing	mm	Routing store capacity	0.05	2000
X4	River routing	day	Unit hydrograph time constant	0.5	40
X5	Runoff routing	-	Intercatchment exchange threshold	-10	10
X6	Runoff routing	mm	Exponential store depletion coefficient	0.01	2000

Table S2: Parameter calibration range used for TUW model

Description of parameters in the TUW hydrological model and plausible range defined for calibration. The forcing adjustment parameters introduced in the model calibration process are highlighted in gray.

Parameter	Type	Units	Description	Plausible range	
				Lower	Upper
dP	Forcing adjustment	%	Precipitation adjustment factor	0.25	3
dT	Forcing adjustment	°C	Temperature adjustment parameter	-3	3
SCF	Snow	-	Snow correction factor	0.5	2
DDF	Snow	mm/°C/day	Degree-day factor	0	5
Tr	Snow	°C	Threshold temperature above which precipitation is rain	1	5
Ts	Snow	°C	Threshold temperature below which precipitation is snow	-3	1
Tm	Snow	°C	Threshold temperature above which melt starts	-2	4
LPrat	Soil	-	Parameter related to the limit for potential evaporation	0	1
FC	Soil	mm	Field capacity, i.e., max soil moisture storage	0.01	1000
BETA	Soil	-	Non-linear parameter for runoff production	0	20
k0	Runoff routing	day	Storage coefficient for very fast response	0	2
k1	Runoff routing	day	Storage coefficient for fast response	2	30
k2	Runoff routing	day	Storage coefficient for slow response	30	500
lsuz	Runoff routing	mm	Threshold storage state, i.e., the very fast response starts if exceeded	1	100
cperc	Runoff routing	mm/day	Constant percolation rate	0	10
bmax	River routing	day	Maximum base at low flows	0	30
croute	River routing	day ² /mm	Free scaling parameter for total runoff during routing	0	50

Table S3: Functions applied to transform the parameters of the GRXJ models

Transformations applied to the parameters of the GRXJ models to improve the search in the calibration process. The functions are the same as those used in airGR³ and are defined in the source code⁴. To recover the values in the original space (i.e., from transformed parameter to raw), the inverse functions should be applied. Note that this function is implemented in the airGR R-package, both for converting raw parameters to transformed ones and vice versa.

Parameter	GR4J	GR5J	GR6J
CN1	$CN1_{Raw} * 19.98 - 9.99$		
CN2	$\log(CN2_{Raw} * 200)$		
X1	$\log(X1_{Raw})$		
X2	$\text{asinh}(X2_{Raw})$		
X3	$\log(X3_{Raw})$		
X4	$9.99 + 19.98 * (X4_{Raw} - 20) / 19.5$		
X5	-	$X5_{Raw} * 19.98 - 9.99$	$X5_{Raw} * 5$
X6	-	-	$\log(X6_{Raw})$

³ 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.

⁴ <https://gitlab.irstea.fr/HYCAR-Hydro/airgr/-/tree/dev/R>

Table S4: Linkage between model outputs and hydrological states and fluxes

Table S4 presents the model outputs used to analyze some hydrological variables of interest. The notation used in the table corresponds to that used in the R packages associated with the GRXJ⁵ and TUW⁶ models.

Hydrological flux/state	Units	Hydrological model output used as proxy			
		GR4J	GR5J	GR6J	TUW
Runoff (Q)	mm/d	Qsim	Qsim	Qsim	q
Baseflow* (BF)	mm/d	QR	QR	QR + QRExp	q2
Actual evapotranspiration (ET)	mm/d	AE	AE	AE	eta
Snowmelt	mm/d	Melt	Melt	Melt	melt
Snow water equivalent (SWE)	mm	SnowPack	SnowPack	SnowPack	swe
Soil moisture (SM)	%	Prod/X1	Prod/X1	Prod/X1	moist/FC

*Here we are using the slow runoff component as a conceptual proxy for baseflow.

For GRXJ:

- Qsim : Series of simulated discharges.
- QR : Series of routing store outflow.
- QRExp : Series of exponential store outflow.
- AE : Series of actual evapotranspiration.
- Melt : Series of averaged actual snow melt per elevation bands.
- SnowPack : Series of averaged snow water equivalent per elevation band.
- Prod : Series of production store level (S; mm).
- X1 : Production store capacity (model parameter; see Table S1).

For TUW:

- q : Total runoff after routing.
- q2 : Baseflow.
- eta : Actual evapotranspiration.
- melt : Averaged snowmelt per elevation band.
- swe : Averaged snow water equivalent per elevation band.
- moist : Soil moisture (mm)
- FC : Maximum soil moisture storage (model parameter; see Table S2).

⁵ Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C. and Michel, C. (2023). airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. R package version 1.7.6, doi: 10.15454/EX11NA, URL: <https://CRAN.R-project.org/package=airGR>.

⁶ Viglione A, Parajka J (2020). TUWmodel: Lumped/Semi-Distributed Hydrological Model for Education Purposes. R package version 1.1-1, <https://CRAN.R-project.org/package=TUWmodel>.

Table S5: Hydrological signatures computed.

Hydrological signature	Abbreviation	Description
Seasonality	Timing	Center of mass timing of streamflow expressed as a fraction of a year compared to the start of a (water) year. The signature is defined using directional statistics, as proposed by Berghuijs et al. (2025) ² .
Mean of the daily series	Mean	Mean daily runoff.
Variance of the daily series	Variance	Standard deviation of the daily runoff.
Slope of the high-segment of the flow duration curve (FDC)	High-segment Slope FDC	Slope of the FDC between the log-transformed 1st and 20th streamflow percentiles.
Slope of the mid-segment of the flow duration curve (FDC)	Mid-segment Slope FDC	Slope of the FDC between the log-transformed 20th and 70th streamflow percentiles.
Slope of the low-segment of the flow duration curve (FDC)	Low-segment Slope FDC	Slope of the FDC between the log-transformed 70th and 99th streamflow percentiles.
Mean of the annual minima series	Annual Minima	Mean 7-day minimum runoff annual minima.
Mean of the annual maxima series	Annual Maxima	Mean 7-day maximum runoff annual maxima.