Dear authors, well done. The referees are pleased with the results, and I agree: job well done which helped improving the paper. There are some minor / technical revisions suggested. Please take care of these and then I will have the final look.
Kind regards
Thom

Thank you very much for your positive feedback. We respond to the two main comments of Reviewer #2 below and modified the manuscript accordingly. Reviewer #1 did not provide any additional comments (at least none were visible to us).

Best regards,

Paul Astagneau, on behalf of all authors.

Reviewer #2

I appreciate the authors' responses and revisions. I have two remaining minor comments related to my previous comment #6 and #8.
Thank you for your positive feedback. We performed the two analyses you suggested and respond to your two comments more specifically below (in blue).

> Previous comment #6: "Since the paper focuses on flow extremes, it would be good to know how the model performs in terms of reproducing the flow quantiles studied in figure 2 and later figures (i.e. the 1%, 50% and 99% annual flow quantiles). For example, this can help put the differences between bias correction methods into perspective."
>
> The authors responded that adding this is not worthwhile since "the performance of the hydrological model in simulating streamflows should not significantly impact the results".
>
> I see the point, but since the paper focuses on flow extremes, it seems natural (and relatively straightforward) to check how well the model simulates these extremes (after all, the authors do report the overall KGE values, suggesting model performance is not completely irrelevant). And it can put the results in perspective (e.g. are differences in flow between bias correction methods significant compared to errors in flow simulation).
> We checked the performance of the two hydrological models for the extrapolation period for low flow, median flow and high flow as suggested (Fig. S10). We computed (A) the KGE criterion on 3 streamflow transformations (Thirel et al., 2024) and (B) the relative bias for the 1st, 50th and 99th percentile. Please note that, due to the extremely low values associated with the first percentile of streamflow, the relative bias is unstable (Pushpalatha et al., 2012). For this reason, we also computed the low-flow bias normalized by the streamflow mean (1%*, Fig. S10B). We found that both hydrological models perform better in simulating high-flow than low-flow, which is a common limitation of hydrological models (Bruno et al., 2024). This is also likely due to the objective function used in calibration, which was computed without streamflow transformation, thus giving more weight to high-flow simulations. Nonetheless, the Cemaneige-GR5J model has better performance than the TUW model (including for low flow simulations), but the performance of the bias adjustment methods remains the same for both hydrological models (Fig. S3). This further indicates that our results do not depend on hydrological model performance because we assessed the bias adjustment methods based on the streamflow simulations rather than observations. Finally, we cannot compare the results of the 75% range criterion with the hydrological model biases because they have a different meaning (i.e. we cannot compare a bias of 10% with a fraction of control runs inside the 75%

confidence interval). We added Fig. S10 to the supplementary information and refer to this figure in the main text as follows: "*The model shows reasonable performance* ***for high-flows*** *in extrapolation between 2011 and 2019 over the catchment set, with a median KGE value of 0.81 (0.75 for the lower quartile and 0.84 for the upper quartile;* ***Fig. S10****).* ***The model shows lower performance for low than for median and high flows, which is a common limitation of hydrological models (Bruno et al., 2024). For this reason****, we check the robustness of our results with respect to the choice of the hydrological model by using a second model with a different structure (Cemaneige-GR5J;* ***Fig. S3****; Le Moine, 2008; Valéry et al., 2014; Coron et al., 2020).*" (L225-230).

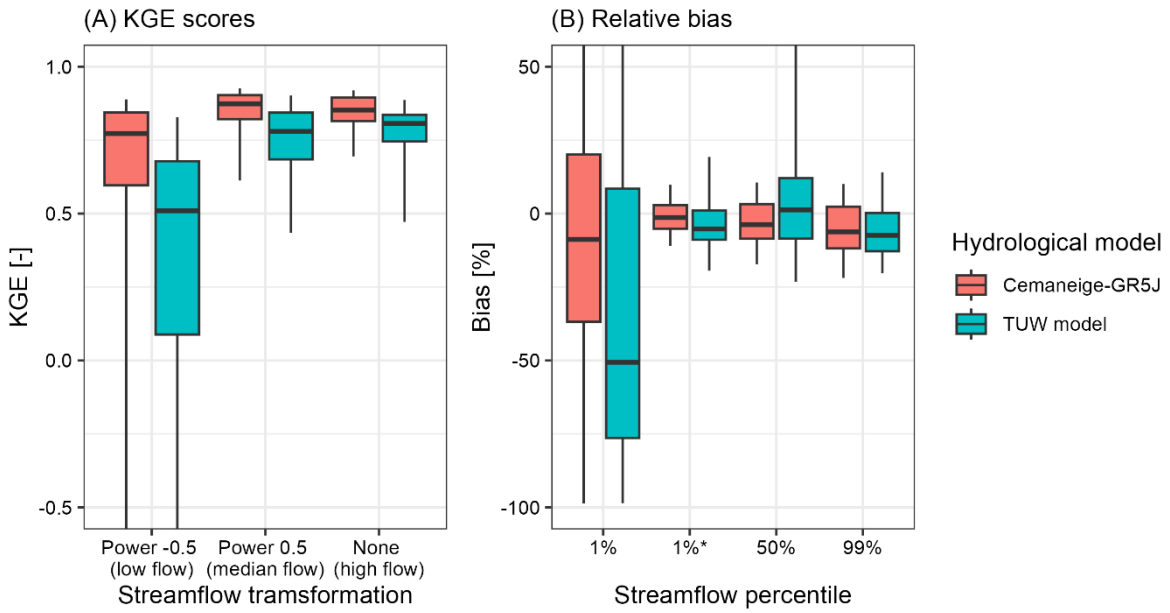

Figure S10: Hydrological model streamflow performance for the 87 catchments (extrapolation: 2011-2019). 1% refers to the relative bias calculated on the 1st streamflow percentile normalized by the streamflow mean.*

- Previous comment #8: "for evaluation of the bias correction methods, the authors adopt a method presented by Suarez-Gutierez et al. 2021; specifically they quantify the fraction of observations that fall in the 75% ensemble confidence interval. Note that those same authors also look at other aspects, e.g. they suggest making a rank histogram which should look uniform".

  The authors respond that ""We evaluated the performance of the bias adjustment methods in the historical period by looking at the 75% ensemble confidence interval introduced by Suarez-Gutierrez et al. (2021). One could investigate other confidence intervals and perform a rank analysis to explore more aspects of bias adjustment performance".

  This however leaves the concern that such a rank analysis could potentially change the conclusions of the paper (or that the conclusions change if you look at a different confidence interval). So, a useful addition could be if the authors can argue (or show) that the conclusions are robust to the choice of confidence interval.
  We reproduced Figure 2 using the same performance criterion but calculated for values outside the min–max range (see Fig. S11). The differences in streamflow performance between the bias adjustment methods are similar to those shown in Figure 2. Therefore, we argue that our conclusions are independent of the choice of confidence interval. We added this figure to

the supplementary information. We also added the following explanation to the text: *"These findings are independent of the choice of the hydrological model (see Fig. S3 in the supplementary material)* ***and of the confidence interval chosen (Fig. S11).***" (L303-304).

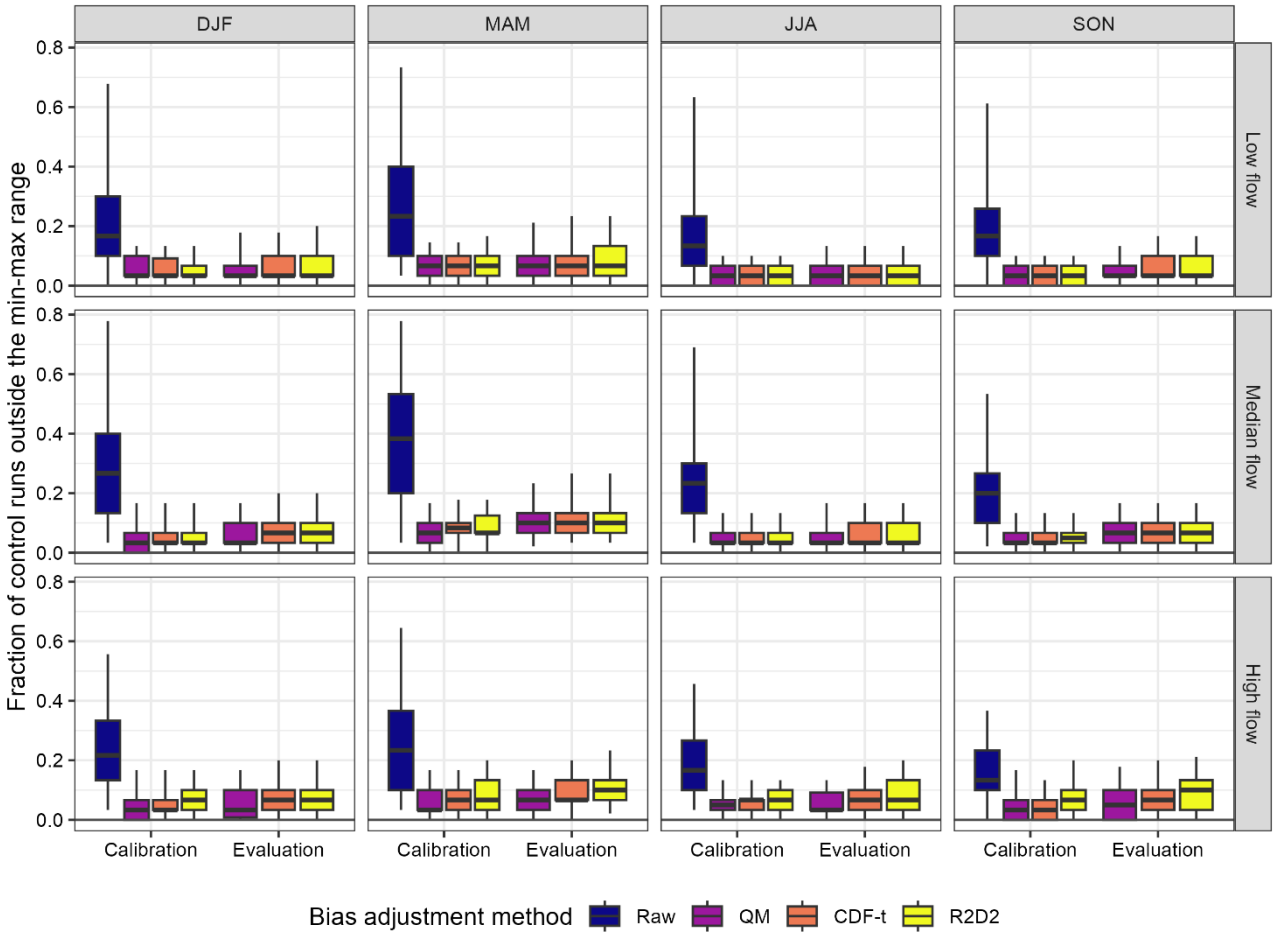

Figure S11: Ability of the three bias adjustment methods and the unadjusted ensemble (raw) to reproduce streamflow statistics of the control runs (streamflow time series simulated by the hydrological model with observed precipitation and temperature inputs) for the 87 catchments. The fraction of control runs outside the simulated min-max confidence interval was calculated for four seasons (December/January/February, March/April/May, June/July/August, September/October/November) and three streamflow percentiles (1st, 50th and 99th). The optimum value of the performance criterion is 0.

References

Bruno, G., Avanzi, F., Alfieri, L., Libertino, A., Gabellani, S., & Duethmann, D. (2024). Hydrological model skills change with drought severity; insights from multi-variable evaluation. *Journal of Hydrology*, *634*, 131023.

Pushpalatha, R., Perrin, C., Le Moine, N., & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, *420*, 171-182.

Thirel, G., Santos, L., Delaigue, O., & Perrin, C. (2024). On the use of streamflow transformations for hydrological model calibration. *Hydrology and Earth System Sciences*, *28*(21), 4837-4860.