Reviewer #3

The paper presents a performance and sensitivity analysis of different climate model bias correction methods and their effect on hydrological simulations of streamflow extremes (low and high flow). By comparing various existing methods across a range of basins in Switzerland, the paper is able to draw useful conclusions about the performance of existing bias correction methods for simulating streamflow in a historical period, and about the sensitivity of future streamflow projections to different bias correction methods.

The methodology is well designed, the paper is clearly structured, and the analysis is comprehensive, with different parts of the analysis logically connecting to one another. My comments are as follows.

Thank you for your positive and constructive review of our study. We respond to each individual comment below.

1. The abstract should be improved, both in terms of clarity and in terms of better capturing the relevant conclusions of the study. A few pointers:

-line 12: "no clear benefits from using bivariate instead of univariate bias adjustment methods when the SMILE already efficiently simulates the dependence between temperature and precipitation". I wonder how robust/general this conclusion is. Wouldn't independent (univariate) bias correction potentially alter the dependence between variables?

Some studies have shown that univariate quantile mapping methods generally preserve the inter-variable dependence simulated by the raw climate model (e.g. François et al., 2020). As this result is only valid for quantile mapping methods and not necessarily for other univariate methods, we changed “univariate adjustments” L454 to “univariate quantile mapping adjustments”. In an attempt to keep the abstract as concise as possible, we kept the related explanations in the discussion section (L452-466).

-lines 15 and 16: not clear what is meant by "precipitation and streamflow signal-to-noise ratios" and by "streamflow and precipitation time-of-emergence". This only becomes clear after reading the paper.

We clarified the sentence to “(2) that the choice … leads to large differences in the values of signal robustness indicators, including temperature, precipitation and streamflow signal-to-noise ratios and streamflow and precipitation time-of-emergence.”

-line 17: "we generally recommend to apply change-preserving and ensemble bias adjustment methods in future hydrological impact studies using SMILEs". To make the abstract more informative, it would be good to clarify in the abstract how this conclusion was reached. The abstract says that there are large differences between bias-correction methods, but does not specify why some methods are preferred over others.

In an attempt to keep the abstract as concise as possible, we kept the detailed information on how we reached this conclusion in the discussion section (4.3).

-the abstract could also mention shortcomings identified in existing methods, i.e. which improvements are necessary based on the findings in this study. The need for more research into bias correction methods is mentioned in sections 4.3 and 4.4, but without identifying which improvements are needed, even though the detailed evaluation in this study presumably provided some useful insights on this.

We added the following sentence to the abstract (L19): "Further research is needed to improve bias adjustment methods that preserve both the signal and the variability of ensemble climate projections."

2. The conclusions section (section 5) should be improved: it seems to largely focus on precipitation and temperature rather than streamflow.

We agree that the presence of streamflow could be further strengthened in the conclusions section. We changed "We found no clear advantage of using bivariate instead of univariate adjustments" to "We found no clear advantage of using bivariate instead of univariate adjustments for simulating streamflow extremes" (L583). We added the following sentence before "We conclude…" (L594): "These interactions between bias adjustment choices can result in large differences in the projection of streamflow extremes."

3. The limitations and perspectives section (section 4.4) is currently very short. Several issues identified in the comments here could potentially be addressed in this section.

4. Basin selection (section 2.1): basins with glaciers are excluded from the analysis because the hydrological model does not account for glaciers. It would be good to come back to this in the discussion, i.e. how relevant are the results and conclusions for basins with glaciers, as these regions are especially vulnerable to climate change.

5. One of the conclusions is that differences between bias correction methods are significant. One wonders whether these differences are still significant when considering all other uncertainties in the climate change modeling chain (data errors, model errors, forcing/scenario uncertainties…). Some discussion/reflection on this would be welcome.

We added the following sentences to the "Limitations and perspectives" section:

- L574: "Our analyses could be extended to other regions and catchments, for instance by including glacierized catchments that are subject to large hydrological shifts due to climate change."
- L575: "Finally, the differences in streamflow projections between bias adjustment methods should be considered in the light of other sources of uncertainties in the climate-hydrological modelling chain, such as scenario and climate model uncertainties (Clark et al. 2016)."
- L571: "The impact of the raw signal on the performance of the ensemble change-preserving method should also be further analysed by investigating whether a deviation between the observed and raw signal on the historical period could explain these differences."

6. Model errors: evaluation of the hydrological model against streamflow observations is reported in terms of KGE, which gives an indication of overall model performance (across all flow levels). 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.

Since we use the streamflow time series simulated by the hydrological model with observed precipitation and temperature inputs as our control run to calculate the streamflow performance, the performance for the different quantiles should not significantly affect the results in the historical period. Furthermore, we also present the results of our analyses with an additional hydrological model and find similar results for both models (see Fig. S1 in the supplementary

material). We rephrased “We use simulated instead of observed streamflow to reduce the dependence of our results on uncertainties in hydrological modelling.” (L256) to “We use simulated instead of observed streamflow to reduce the dependence of our results on uncertainties in hydrological modelling. This means that the performance of the hydrological model in simulating streamflows should not significantly impact the results.”.

7. Data errors: "observations" of precipitation and temperature are based on gridded (interpolated) meteorological station data, which are used as benchmark ('ground-truth') in this paper (line 155). To what extent does bias and noise in these data affect the results? E.g. typical sources of bias are under-catch of precipitation gauge measurements (especially for snow) and the absence of stations at high elevations.

While snow under-catch is a known problem at high elevations, it should not affect our results significantly because we use the control run as a reference to calculate streamflow performance. Furthermore, the hydrological model we use in our study (HBV) includes a snow-correction factor to reduce these biases.

8. Evaluation: 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 (see their figure 1). A cdf version of the same idea is in Laio et al. 2007 (figure 2 in https://hess.copernicus.org/articles/11/1267/2007/). It seems this would allow for a more complete evaluation of the ensembles. Can the authors comment on whether these methods are applicable here and why they were not considered?

We chose the 75% range to simplify the presentation of the results and because it was used as one of the main metrics in Suarez-Gutierrez et al. (2021). We did not perform the rank analysis used in Suarez-Gutierrez et al. (2021) to not overload the results even more, but we agree that it would have been an interesting aspect to look at. We added this sentence in the limitations and perspectives section (L.569): “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”.

9. Consistency in terminology: on line 136, we are introduced to "five bias adjustment setups". Later on, a distinction is made between 3 bias adjustment methods and 2 ensemble adjustment methods (e.g. figure 12), while figure 10 refers to these combinations as bias adjustment options. Would be good to be consistent and for example introduce the naming used in figure 10 from the start and use it consistently throughout the paper.

Thank you for highlighting these inconsistencies in terminology. Throughout the manuscript, we now consistently use “methods” for statistical methods (e.g. R2D2 vs. QM) and “strategy” for the combination of a statistical method with the choice of change-preserving and ensemble adjustments.

10. Line 185: "We run the adjustments at the grid scale rather than the catchment scale to avoid adding a downscaling step to the procedure, and because the catchments are of different sizes." The reasoning here is not clear to me, i.e. how does bias adjustment at the catchment scale add a downscaling step (compared to adjustment at grid scale followed by moving to catchment scale), and how does catchment size come into play?

The catchments in our dataset differ in size. Applying the bias adjustment at the catchment scale would result in mixing the bias adjustment with upscaling for large catchments and downscaling for small catchments. We wanted to separate spatial scaling from bias adjustment in this study. We replaced this sentence by “Applying the bias adjustment at the catchment scale would result in mixing the bias adjustment with upscaling for large catchments and downscaling for small catchments.” (L186).

11. Overall structure of the results section: even though this section already flows quite nicely, you could consider splitting up section 3.2 into two further sub-sections (precip/temp and streamflow), and using the same split in section 3.1 (precip/temp and streamflow). Currently, section 3.1 starts with streamflow, so opposite order of section 3.2. Not super crucial, but readability may improve by breaking up the results into smaller pieces and using consistent order in sections 3.1 and 3.2.

Thank you for these suggestions. We added subsections but kept the same order in each subsection to keep the reading flow.

12. Figure 2: clarify what is meant by "control runs" - I know it is mentioned in the methodology section, but it should be clear from the figure caption. Also, the figure axis should make clear which variable we're looking at. Suggest to rename bias adjustment method "raw" to "none" or "unadjusted". And I assume these are box plots, would be good to explicitly mention that. And which variability is captured by these box plots? Is it variability across the 87 basins?

We added the definition of “control runs” in the captions of Figure 2. We added the variable names as a subtitle to each figure (top of the figure) to not overload the y-axis titles (Figures 2, 8, 9 and 10). We kept “raw” for consistency with the text. We added the number of catchments in the caption of Figures 6, 7, 8 and 10.

13. Figure 2 and other figures focus on the 75% ensemble interval for streamflow. Why did you pick 75% and would your conclusions change if you pick another percentage? See also comment 8.

We also checked the results for outside of the 75% range and inside the maximum-minimum range and found no differences with the results found for inside the 75% range. We decided to not to present these results to reduce the complexity of the analyses (see suggestions by R1 and R2).

14. Why does figure 3 show results for one of the evaluation periods whereas figure 2 shows results for both? Also, the color bar title ("fraction of control runs") should make clear that we're looking at streamflow.

We moved this figure to the supplementary material to reduce the complexity of the manuscript but the results were similar for the other period. We also changed the color bar title to make clear that we’re looking at streamflow.

15. Figure 10: figure/axis title should make clear we're looking at precipitation. Same for figures 11 and 12, make sure the figure/axis title mentions 'streamflow'.

See previous comment.

16. Figure 12: noise is expressed as %. Is this the coefficient of variation? The axis title calls it standard deviation?

Because we express the streamflow signal in relative terms, the standard deviation between members is also expressed as %. We added this explanation in the caption.

17. Line 587: "ensemble adjustments combined with the change-preserving method are less efficient for the tails of the precipitation and temperature distributions in the historical period, probably because the raw change signals are small compared to the internal variability for many catchments". This is not clear. How does the climate change signal (second part of sentence) affect performance of the bias correction method in the historical period (first part of sentence)?

This part of the discussion is related to the results of Figure 7. The CDF-t method in the historical period has a lower performance than the QM method for the tail of the distribution for the historical period and when the ensemble strategy is used. Given that CDF-t has lower performance for the 1st percentile of temperatures when the raw signal between P1 and P2 is the weakest, we hypothesized that it might be one of the reasons explaining the performance differences. We agree that the results are not straight-forward. Therefore, to reduce the complexity of the manuscript, we moved this analysis to the supplementary material. We discuss this in the discussion section (4.2, L506-517). “However, this effect is partly due to the weak signals simulated by the raw ensemble in the historical period (see Fig. S4). More specifically, we found that when the signal of the unadjusted ensemble is low, the change-preserving method combined with ensemble adjustments tends to have lower performance compared to when this signal is larger. For low signals, the change-preserving method might try to preserve a signal which is not significant compared to internal variability. This effect is enhanced when the observations show a strong signal compared to the raw signal (Fig. S9). Therefore, the drop in performance for the tail of the distribution might be an apparent problem in the historical period but not for future projections, where the signals become larger than the internal variability. However, the relationship between the raw signal and the performance of the bias adjustment is not strong for precipitation. This is probably because the precipitation signal is weaker than the temperature signal compared to internal variability (Fig. S4). An additional explanation could be that the ensemble adjustments have a lower efficiency in preserving the variability of the distribution tail, as found by Vaittinada Ayar et al. (2021), suggesting that there may be room for improvement in adjusting the tail of an ensemble distribution while preserving the change signal.”.

18. Line 128: biased --> bias

We modified accordingly.