Thank you very much for inviting us to revise our manuscript. We appreciate the comments and suggestions made by the three reviewers and address them in the revised manuscript. In particular, we have significantly shortened the manuscript by moving some of the analyses to the supplementary material and by reducing the complexity of some of the figures. We have strengthened the explanations in the method section, for example by incorporating a table that summarises the evaluation indicators used. Finally, we have adjusted the discussion section to emphasise the limitations and perspectives of our study.

We hope that you find the revised manuscript suitable for publication and thank you for your consideration.

Best regards,

Paul Astagneau, on behalf of all authors.

Reviewer #1 Faranak Tootoonchi

This paper is very well-written and highly relevant. The assessment of the impact of bias-adjustment techniques on SMILEs is both timely and novel. The authors have clearly put significant effort into considering important steps for bias adjustment. The results section is thorough and addresses all the proposed research questions and even goes beyond them.

Thank you for your positive and constructive feedback on our manuscript. We agree with your main remarks and answer each individual remark in blue below. Specifically, we propose several adjustments to shorten the manuscript.

I have a few **minor** remarks:

- In my view, the paper is too long and requires multiple rounds of thorough reading to absorb all the information. I understand it is not easy to cut a paper like this shorter or present it in a simpler way. Nonetheless, I encourage the authors to read the paper again and see whether more plots can be moved to the supplementary section and whether some parts of the result section can be summarized. I think certain plots from the historic analysis can be omitted. Figure 7 is particularly difficult to interpret and I am not entirely sure if I understood its key point. I could have skipped Figure 10 and limited the plots to what is shown for runoff in Figure 11. But even Figure 11 is challenging to grasp, as it represents the final output of multiple subtractions. Again, I understand that it is not too easy to cut this paper short but I think doing the laborious work of summarizing it, helps with readability.

  We agree that the paper contains too much information. We propose the following adjustments:

  - Moving Figure 3 to the supplementary information and only mention in section 3.1 that there are no obvious spatial patterns in terms of bias adjustment performance.
  - Moving Figure 7 to the supplementary information and only mention in the discussion that trends in the climate model simulations between P1 and P2 might be an explanation for the differences in performance between QM and CDF-t.
  - Showing only the period 2081-2099 for Figure 9.
  - Removing Figure 10b and the text associated.
  - Switching Figure 11 and 12 to improve the reading flow.

- The authors did not find significant benefits of the multivariate bias adjustment method compared to the univariate approaches, and I find this result reasonable. They attribute this outcome to the well-preserved correlation in this particular SMILE. In my view, the relatively low P–T correlation in the observational data (Figure 4b) also contributed to this result, as there was no strong correlation that needed to be preserved. When the correlation is weak, bias adjustment for separate months may be sufficient to maintain a reasonable dependence between precipitation and temperature. In such cases, I would argue that preserving temporal order might be more important. Ultimately, I would recommend that impact modelers evaluate whether correlation (or even chronology) is important for their specific application and choose a simple method that adjusts just enough, but not more. If the authors agree with this point, I suggest including it in the final discussion and recommendations.

  We agree that the P-T correlation is relatively low, probably because it is calculated at the annual time scale. We now stress in the discussion and the recommendations sections that impact modelers should determine which aspects are the most important for their specific application and choose a bias adjustment method accordingly. L560 : "In general, we recommend that impact modellers determine the most important aspects of their specific application and choose a bias adjustment method accordingly".

- In section 2.4 (evaluation) does it help to have a table with all the indicators you evaluated, separate for P, T and Q, present and future?

  Thank you for this great suggestion. We added a new table to section 2.4 summarizing all indicators used.

Specific comments:

L3: You can remove this from 'this internal' variability.

We removed it.

L136: Mention what the five setups are and then in table 1, in the title mention that the combinations in the last two columns encompasses five bias adjustment setups.

We added in the title of Table 1 that the combinations in the last two columns encompass five bias adjustment strategies. However, we did not modify L136 because it serves as an introduction sentence.

L171: Why not the dependence?

The R2D2 method was designed that way: R2D2 reproduces the dependence of the calibration period. For the projection period, dependence structures are the same, with preservation of "the temporal dynamic of the climate model", through the use of dimension/pivot variables. R2D2 is not designed to preserve the simulated changes in the dependence structure but is actually stationary, meaning it exhibits no change.

L185-186: The sentence here is somewhat a repetition of L180-182.

We agree and removed it. We also added the following sentence to clarify this methodological choice (as suggested by Reviewer 3). L186: "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."

L219 and then L253: Why P1 and P2 are introduced in the text but are not used in any part of the result? True that you want to cross validate but if the results are shown all together, is it really necessary to introduce an abbreviation? And then considering what mentioned in the text why Figure 3 is only for one sub period? Why not to show it for the entire historic period? And what is efficiency in this figure?

P1 and P2 are used in other parts of the methods section (L202-204, 239 and 260). For Figure 3, we found similar results for the other sub-period and decided not to show it to reduce complexity. Figure 3 will be moved to the supplementary material.

Does it make sense to already mention in L219 what is later mentioned in L253? And Did I understand correctly that you name the runoff simulation through this joint combination control run? If it is so, please already mention it in the text. I had a bit of difficulty understanding what period Figure 2 is showing.

To not confuse the calibration/evaluation periods of the hydrological model with those of the bias adjustment methods, we kept the explanations of L253 in Section 2.4.

L233: Change however to instead. And the whole L233-238 requires some rewriting. The section sounds more like an statement rather than what has been done in the paper.

We changed "however" to instead, but we think that these lines are important to understand how we evaluate the biases of an ensemble. The objective of the ensemble adjustments is to avoid removing the spread of the ensemble while still removing the biases. This is what we are exploring in our study. We rephrased "which would imply removing the random biases due to internal climate variability" by "which would imply removing the fluctuations due to internal climate variability" on L240.

L249: The term 'use' is unclear to me. It is unclear 'how' you evaluated it.

Here, we wanted to say that we use streamflow simulations from the hydrological model fed with observed meteorology (control run) as our reference to evaluate the performance of the different bias adjustment methods instead of directly using streamflow observations. We rephrased the sentences of L253-255: "To evaluate the performance of bias adjustment for streamflow simulations, we use the streamflow time series simulated by the hydrological model with observed precipitation and temperature inputs as our control run to calculate the 75 % range criterion."

L259: the term 'signal' is unclear to me. Do you mean the difference between averages?

By signal we mean the relative or absolute difference between the future period and the historical period for a given percentile of a given variable. We rephrased the sentence of L259 to clarify this point. L265: "To do this, we calculate the signal (**difference**) between the future period (e.g. 2081–2099) and the reference period (1991–2020)…".

L265: Remove second. There are two firsts in the previous paragraph. So it is unclear which first this second comes after. I would have personally rephrased the previous paragraph to avoid those firsts.

We removed the second "first" of L259 (now L265).

L271: Remove the time-of-emergence and join the two sentences.

We merged these two sentences.

L275: Until here it was not mentioned that you will look at groups of catchment with different elevation levels (or did I miss it?). Cool that you did. But does it make sense to already bring it up earlier in the text and group the catchments in Figure 1 based on the three categories of elevation, to signal this to the reader?

We changed Figure 1 to group the catchments by elevation. We also introduce this point in Section 2.1, L144.

L331-332: Doesn't this belong to any other section but not the result?

Here, we wanted to emphasize that the results so far were for interannual variability and not inter-member variability. We removed the citation here as it is already mentioned in the introduction and changed "can" to "could".

Figure 7: I unfortunately did not understand Figure 7 and its aim after many tries. If it is not only me, please consider both rewriting the section and re-visualizing it, or instead think of removing the plot and the text all together.

The idea here was to explore whether trends in the raw signal of the climate model between P1 and P2 could explain the performance differences between CDF-t and QM. The results are not straightforward, therefore we moved this figure to the supplementary materials and modified the text accordingly.

L375-376: Somewhat repeats the beginning of the section in L355.

This repetition is intentional to help the reader keeping track of the explanations in the results section, therefore we kept the sentence

L414: I think setup is better than methods. Not all mentioned in the parenthesis are methods.

We changed it to "strategies" to be consistent with the rest of the manuscript.

Figure 11 is slightly complicated. Instead of showing the subtractions can you show the actual boxplots separately for each of the pairs?

We agree that Figure 11 is complex. However, since we do not have a reference of what the time-of-emergence should be after bias adjustment for hydrological projections, we need to compare the time-of-emergence between the different bias adjustment strategies. Also, since these differences in time-of-emergence seem to be catchment-dependent, boxplots showing distributions of absolute time-of-emergence would mask the important differences. We switched Figures 11 and 12 to improve the reading flow.

L434-446: This part and Figure 12 is very interesting. However, I think some part of the text belong to discussion. I would have loved to see a plot similar to Figure 9 but for runoff just to see how the methods behave for all runoff simulated components in the catchments.

This part is just a description of the results of Figure 12 and illustrates the results of Figure 11 for 3 catchments. Furthermore, we cannot reproduce Figure 9 for runoff/streamflow because there is no reference of what the streamflow signal-to-noise ratio after bias adjustment should be.

L514: Unclear what strategies mean here.

"strategies" refer generally to the combination of a statistical method with the choice of change-preserving and ensemble adjustments . We added a sentence in the text clarifying this: L141: "Note that we use the term strategy to refer generally to the combination of a statistical method with the choice of change-preserving and ensemble adjustments."

L525: Cite the plot for precipitation.

We added a reference to Figure 7 which is now in the supplementary material (Figure S4).

L558: I agree that change preserving is inherently more in line with the aim of future impact studies. But I slightly disagree with the rest of this paragraph: Apart from having the same performance for precipitation, combination of change preserving and individual bias adjustment strategy resulted in very different signal for high flow in Saltina at Brig compared to the rest (Figure 12). One might argue that 99th percentile is too extreme, but then essentially all methods are more or less similar when it comes to moderate or moderately extreme percentiles. Based on your results, your third point sounds more concrete to me. So my suggestion is to reshuffle third and second point and use an even more cautious tone in suggesting second point.

We exchanged the third point with the second point and we now use more cautious terms to refer to the second point. L557, we changed "it is more in line with the target of climate impact studies" to "it **might be** more in line with the target of climate impact studies"

Reviewer #2 Thomas Bosshard

Summary

Astagneau et al. (2025) present a comprehensive hydrological climate impact study using single model initial-condition large ensembles (SMILEs) as input data. The focus is put on the bias-adjustment step in the modelling chain. They investigate how different choices of the bias adjustment method and its application affect the outcome on hydrometeorological extremes. The choices are: Univariate vs. bivariate bias-adjustment, trend-preserving vs non-trend-preserving bias-adjustment, and grouping all ensemble members when calibrating the bias-adjustment parameters or calibrating the parameters for each member individually.

The results show that the choice of ensemble vs individual calibration, as well as trend-preserving vs non-trend-preserving method has larger impact than univariate vs. bivariate bias-adjustment and the authors recommend to use trend-preserving bias-adjustment methods in combination with ensemble calibration, and only use multivariate bias-adjustment if correlation structures are strongly biased in the raw climate model data.

General comments

The paper is highly relevant in this field of research as it combines the rather novel SMILEs with still not fully assessed issues of bias-adjustment such as intervariable dependencies and modification of the climate change signal. It is very well written and nicely illustrated. The study is so comprehensive that the paper gets a bit overloaded. Here and there it becomes apparent that the authors had to leave out interesting information because otherwise, the article would have become even longer. In my opinion, the paper could have easily been split into two papers – for e.g. one about the meteorological analysis and one about the hydrological analysis. Both the hydrological part and the results of the bivariate bias-adjustment get too little space in the manuscript. I have one general comment about the evaluation metric 'Fraction of control runs inside the 75% range'. How did you choose the value of 75% range? Reading Suarez-Gutierrez et al. (2021) but also studies about evaluation of seasonal forecasts where ensemble forecasts have been around for a longer time (see for e.g. Crochemore et al., 2016), it looks like one recommends to look at the reliability of the projection as a whole, for e.g. by using the probability integral

transform (PIT). I would argue that looking at the whole reliability rather than at one specific percentile interval gives a more complete analysis.

Overall, given the high scientific relevance and scientifically sound study and presentation, I recommend acceptance with minor revisions.

Thank you for your positive and constructive review of our manuscript. We agree that the paper has too much content. We suggested ways to significantly shorten the manuscript in our response to Reviewer 1. In particular, we suggested to move Figures 3 and 7 and the related explanations to the supplementary materials and to simplify Figures 10 and 11. Regarding the evaluation metric for the historical period, 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 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 the 75% range. 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 a sentence in the limitations and perspectives section (L572): "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". We appreciate the detailed comments below and answer each one individually.

Detailed comments

L198: P2 is half in the historical and half in the scenario part of the climate projections. I do not think that this constitutes a big issue, but I suggest to add a note about this in the text to make it clear that you are aware of the different characteristics of the two chosen periods.

We added a sentence on L204: "Note that P2 includes both historical and scenario data but this should not affect the results of our study".

L246: I do not fully agree with the statement here that the 75%-criterion evaluates bias and interannual variability of the ensemble. The over-/underconfidence of the ensemble forecast might just as well play a role in the 75%-criterion. The authors later on state that the interannual variability and the inter-member variability are equivalent. However, they also write that it does not hold this study, if I understood it correctly. Thus, it might be good to mention both the bias, the interannual variability and the inter-member variability as factors influencing the 75%-criterion.

What we mean by this statement is that, when using absolute values and not anomalies, both the bias and the interannual variability can affect the 75% criterion values. Although the inter-member variability of a SMILE has been found to be equivalent to its interannual variability (von Trentini et al., 2020), our results showed that a reduction in the inter-member variability did not result in a degradation of the 75% criterion (Figure 5 vs. 6), which means that they are not equivalent in that case Given that L246 is part of the methods section, we replaced "interannual variability" by "variability" to avoid confusion.

L294 and Figure 3: The maps of the 75%-criterion are hard to interpret. In the text, you often talk about how close the different stations are to the optimal value of 75% present. Since the difference to 0.75 is the focus, I suggest that you plot the difference of each station's result to the optimal value of 0.75 instead.

We agree that the maps of the 75%-criterion could be improved. We adjusted the design to show the difference to the optimal value of 0.75. However, to shorten the manuscript, we decided to move them and the related explanations to the supplementary information.

Text describing the results in Fig. 5: I can see that CDF-t in ensemble-mode has some difficulties in the evaluation period. You explain it later in Fig. 7 that this potentially could be linked to weak signals in the raw data between calibration and evaluation period. If that was the case, however, we should see the same issues with CDF-t in individual-mode. I would expect even worse performance in the individual-mode due to a more pronounced tendency for overfitting and hence, potentially larger drops in performance when evaluating on independent data. I would like the authors to check their argument.

We disagree with this argument. In ensemble mode, CDF-t tries to preserve the change (evolution) from the ensemble in the calibration period to the individual member in the projection period, whereas in individual mode it tries to preserve the signal of each member. Given that each member has a different signal, it might be easier for the individual mode to adjust the biases. This remains a hypothesis without further analyses. Furthermore, to shorten the manuscript, we decided to move Figure 7 and the related explanations to the supplementary information. Finally, we also mentioned in the discussion (Section 4.2), that the weak signals might not be the only reason for the lower performance of CDF-t compared to QM in ensemble mode and that other methods could be tested to preserve the variability of the ensemble after bias adjustment. These methods might have the potential to improve performance for the tail of the distribution.

Line 3: I think you have all data at hand to be more specific about this statement as you could exactly calculate both the bias and the reduction of the interannual ensemble spread. Based on your data, which of the two is contributing more to the results you see in Fig. 5?

We are not sure to which line this comment refers to, as line 3 is in the abstract and does not seem to be related to this comment. We assume that this comment is related to L330: “which means that the interannual ensemble spread is reduced or that the simulations are biased.”. We removed this part of the sentence which might be misleading.

Fig. 6: Why are the results in Fig. 6 not reflected in Fig. 5? In Fig. 6, it is apparent that for e.g. that individual-mode leads to strongly reduced ensemble spread for both the 90th and 99th percentile for precipitation (calibration period). In my understanding, this indicates an overconfident ensemble which should show too many observed data points to fall outside the ensemble spread. This is though not at all visible in Fig. 5 (B). There, ensemble and individual-mode perform equally in the calibration period. Could be please explain this apparent inconsistency? Note that I just took the case of precipitation as an example. There are other inconsistencies of the same kind between Fig. 6 and Fig. 5.

The differences between the results of Figure 5 and the results of Figure 6 are indeed surprising. This is related to the comment you made for L246. We believe that these results show that after bias adjustment in individual mode, the interannual variability and the inter-member variability are decoupled. Furthermore, we think that the 75% range criterion only assesses interannual variability and not inter-member variability. We discuss these differences in the discussion section (4.2, L505-516). “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 weak, the change-preserving method combined with ensemble adjustments tends to have lower performance compared to when this signal is stronger. For weak 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 might be related to the precipitation signal being 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). This suggests that there may be room for improvement in adjusting the tail of an ensemble distribution, while preserving the change signal.".

Text describing Fig. 7: It was difficult for me to follow the authors argumentation here and only understood the concept after having read the conclusions. I would like to ask the authors to spend 1 or two sentences even in the result section to explain the basic hypothesis a bit more. In any case, I think it is not fully convincing that a trend-preserving method cannot handle situations of week trends in the raw signal. In case of week trends, it should behave similarly than non-trend-preserving methods rather than introducing an artificial trend. This said, I speculate that the authors might have rather meant to look at deviating trends between the observational data and the climate model data between calibration and validation period. Due to natural variability it can happen that observational data show a positive (negative) trend while climate models show negative (positive) trends. If CDF-t enforces the trend in the climate model data but the observational data show a totally different trend, the performance in the evaluation period will drop. Note that this also applies to non-trend- preserving methods. However, it could be that trend-preserving methods might be more susceptible to those kind of trend-inconsistencies between the datasets.

We agree that the results of Figure 7 are not very clear. As we decided to move this figure to the supplementary information and only discuss this aspect in the discussion section, we will not do further analyses on this aspect. We agree that the difference between the observed trend and the raw trend is a sounder hypothesis to explain the performance differences between QM and CDF-t for the tail of the distribution. We started to test this hypothesis and found no clear results (Figure S9). Furthermore, the results we obtained were quite complex and would have even more overloaded the manuscript. We believe that this aspect should be studied independently and in more detail in future work. We added a sentence in the "Limitations and perspectives" section to emphasize this aspect (L574): "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 observed and raw signal on the historical period could explain these differences."

Fig. 8 (A): The results for QM, 1st percentile stick out. Do you have an idea why it is just at the lower tail where the large modification of the climate change signal happens? Have you for e.g. looked at other percentiles close to the 1st percentile or even the whole CDF to see where the modification kicks in? Do you know other studies where similar results were seen? It would be interesting to see if this might be a more general issue seen in other similar studies or if it is particular to this study.

We checked the rest of the CDF and found that this large modification of the climate signal starts between the 1st and the 10th percentile. There is a study showing that climate models often show an elevation dependence of the bias (Matiu et al., 2024), however, none (to our knowledge) that found an elevation dependence of the modification of the climate change signal for low temperature extremes.

Fig. 9: Why did you choose to show absolute values? I would prefer to see both negative and positive values to be able to better interpret Fig. 10. If I understood correctly, the SNR and time of emergence are interlinked. Thus, if QM increases (decreases) the SNR, the time or emergence should be earlier (later) than in raw projections. Is that a correct interpretation? If so, showing only absolute values in Fig. 9 makes it hard to understand/interpret why the time of emergence is earlier or later in Fig. 10.

We agree with this comment. We now show actual values in Figure 9. Additionally, to reduce complexity and shorten the manuscript, we now only include 2081-2099 in this figure.

L435: Please add a short explanation why you picked those 3 examples. They seem to be rather extreme in how the SNR is affected.

We added the following explanation (L405): "We choose these three examples because they illustrate three different cases of time-of-emergence differences originating from differences in signal and noise."

Lines 469-472: I do not understand why the seasonal adjustments used in this study improve the correlations between precipitation and temperature. Other studies also use varying adjustments throughout the year (e.g. often using months or a monthly moving window), yet they often see a clear improvement of the correlation between precipitation and temperature when using multivariate bias-adjustments and less so for univariate bias-adjustment. I would ask the authors to clarify this statement.

Here, the seasonal adjustments do not drastically improve the correlation values and do so only in a few cells. Overall, the correlation values for the raw ensemble and after univariate adjustment are very close. In fact, at the monthly scale (scale of the adjustment), these values are even closer. Furthermore, we showed correlation values for a specific temperature range and for wet days, which could explain the differences between univariate and raw correlations. Since the correlation values of the raw ensemble are already close to those of the observations, the multivariate adjustments only marginally improve these correlation values. We are discussing these points in the discussion section (4.1; L444-466).

Line 516: replace "change-preserving methods" by "the change-preserving method used in this study", because CDF-t is just one of the available change-preserving methods available and might not be fully representative for the whole group of change-preserving methods.

We made this adjustment.

Lines 525-527: If I'm not mistaken, Vaittinada Ayar et al. (2021) investigated this with CDF-t as the only one bias-adjustment method. They state at the end that the results have to be reproduced by other methods. Please state clearer that results by Vaittinada Ayar et al. (2021) are valid for the combination of ensemble adjustment and CDF-t, and not general for all sorts of ensemble adjustment. In fact, your results seem to show that it does work better for QM.

We added the following sentence (L516): "Vaittinada Ayar et al. (2021) only tested the ensemble adjustments for CDF-t, therefore these results should be confirmed using other change-preserving methods."

Lines 530-533: Your statement sounds rather general while you actually have all data at hand. You could easily analyze the change in snowmelt and see if it corresponds with the results you see in figures 11 and 12. In fact, the differences appear to be more pronounced for median flows rather than high flow (Fig. 12). Maybe, snowmelt has an impact on the median flows? It is hard to

tell based on the results given in the paper, as for e.g., one does not know when the low, median and high flows occur throughout the year.

We agree that this would be an interesting aspect to study. However, given the current length of the paper and the need to shorten it, we would like to keep this analysis for future studies.

Lines 558-562: There might be plenty of reasons why to prefer trend-preserving methods. However, the given one here might be very specific to this study. And without this reason, it boils down to the statement that trend-preserving methods are to be preferred since they are more in line with the target of climate impact studies to use change-preserving methods. I would ask the authors to sharpen their argumentation or to clearly state that the recommendation is based on two specific methods and a specific data set in the specific region – and other conclusions might be drawn in other studies.

We added the following sentences (L559): “These recommendations are based on a specific region, dataset and a selection of bias adjustment methods. Therefore, their generalizability should be evaluated in different contexts.”

References:

Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 20, 3601–3618, https://doi.org/10.5194/hess-20-3601-2016, 2016.

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” (L585). We added the following sentence before “We conclude…” (L596): “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:

- L570: "Other regions and catchments also need to be included in future analyses to improve the generalizability of our results, such as glacierized catchments that are subject to large hydrological shifts due to climate change."
- L577: "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)."
- L574: "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.572): “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, L505-516). “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 weak, the change-preserving method combined with ensemble adjustments tends to have lower performance compared to when this signal is stronger. For weak 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 might be related to the precipitation signal being 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). This suggests 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.