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. L558, 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"