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 (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". 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, 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.".

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 (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 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 (L517): "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 (L560): "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.