Revision

I thank the authors for their replies and revisions, which helped to clarify some of the doubts raised in the first review.

My first concern was about the heterogeneity of the precipitation input, which is important in the context of parameter transferability, because the goodness-of-fit is not only determined by the appropriateness of the calibrated parameters, but also of the input data. I found the additional info you provided on the precipitation helpful. The data of the four observation stations show a very clear trend of precipitation amounts in east-west direction, and the input data you are using seem to reflect these differences, as you rightfully point out in the discussion.

Lines 391-393: I would, however, like to question why you think that calibration on a five-years period reduces the impact of rainfall patterns on transferability to neighbouring catchments? I do not see how calibrating on the period 2010-2014 would solve this problem, especially if the difference in rainfall patterns is consistent over time as indicated by the station data shown in the Appendix. Maybe consider deleting this statement.

Would it not be better to just state that an impact is likely, and leave it like that? You are focussing on snowmelt, so it should be fine just to mention that effects from rainfall variability that are not captured in the input data are possible.

The next point was the benchmarking approach. In light of the additional explanation provided in the reply to the review and the related revision, several concerns remain, which I would like to point out in the following.

The benchmark is used to assess the quality of the parameter transfer, thus it should be unbiased to provide a fair evaluation. I think that this is not the case with the approach taken by the authors.

Excluding the data for a particular year distorts the randomness of the benchmarking approach. This design of the benchmark prevents a good fit, in the sense that a good fit cannot even be randomly achieved. This means that preventing the random draw of the original year at its original position systematically decreases the NSE and KGE values. The authors claim that “The benchmark NSE and KGE correspond to the prediction potential of the discharge dataset itself.” This is not true, because due to the exclusion it is not a randomly drawn sample.

I also do not see how the approach could possibly reduce the effect of outlier years. Either there are “outliers” present in the sample of size 5, or not. If there are outliers, the NSE and KGE will be low, because an outlier year would be represented by an average year, or the other way round. I thus agree with your statement that the different years should not be too different for the approach to be meaningful (line 409). On the other hand, if the discharge series would be very similar each year, the approach also does not make much sense, it would just deliver perfect fits. As you are showing schematically in the Annex, small deviations (small in terms of the temporal resolution) in the discharge already can let NSE or KGE drop, even if the general annual pattern is similar. The question is, how much difference in dynamics is allowed and how much is required for this metric to be useful?

What also puzzles me is the limitation to five years. Discharge data is available since 1971 (line 130), the calibration period was 2009-2014, but the benchmark was only calculated for the “evaluation period” 2010-2014, which is shorter because the first year was discarded. Why do you not use the entire dataset for bootstrapping, or averaging? Limiting to five years seems to be an unnecessary restriction. Expanding the data set would also mitigate the problem of possible outlier years.

I am not sure if a benchmark is needed at all, or if the results of the study would also be valid without this comparison. The authors state that NSE and KGE have no absolute meaning (lines 263-264). But I would claim that NSE and KGE have defined properties with regard to the statistics of the sample. Many papers have explored these metrics in the context of hydrological modelling, so they may be useful on their own because the target audience can interpret them, and their changes in the transfer experiment. The properties of the chosen benchmarking approach, however, remain unknown.

In the revision, the authors added a new section to the beginning of the discussion named "A new benchmark". This claim appears a bit bold to me, and seems to be out of scope . In my opinion, introducing a new benchmark would require a convincing approach to start with, proper testing, and critical discussion in light of the existing literature, all of which I cannot see here. From the manuscript it is not clear if the authors consulted any of the existing literature on methods for bootstrapping of dependent data and their requirements, for example in terms of sample size or the effect of excluding data while drawing.

All of the above raises a major concern if the metric is a fair benchmark, or if it is designed to give results that make comparisons against it look good. I suggest considering an unbiased benchmark, either by one of the suggestions above or something that is described in the literature, or getting rid of the benchmark entirely.

Minor:

- The Figure G is meant to show that the relative impact of precipitation is small vs. snowmelt. The time period is not indicated, so the comparison with Figure F is difficult and it is hard to tell if this is a larger event, or if a minimal rain event was chosen.
- Figure F - I suggest that you add the relative position of the observation stations to the figure caption of F to give a better idea (like: "xyz km to the NE of Arolla station").
- Line 419: "predictability of discharge from past discharge signals" – your benchmark consists not only of past years, but "other" years.
- Lines 433-437: *"Given the inherent year-to-year variability in meteorological patterns, and the close link between meteorology and discharge, it ensues that in small catchments, the discharge patterns from previous years are poor predictors of the current discharge. In contrast, even simple meteorology-based hydrological models deliver much better results. An ideal benchmark should not depend on scale; however, we do not see at this stage how to construct such a benchmark."* – Again, I would like to point out that this seems contradictory. If simple meteorology-based hydrological models give good results, why not use them as a benchmark? Maybe I am missing a point here, but perhaps the last two sentences are not really needed, and could be skipped?