Reply to RC2

We thank the referee for their review and constructive comments. Original review comments are shown in **black** while our replies are provided in **green**.

The paper “Scale-dependency in modeling nivo-glacial hydrological systems: the case of the Arolla basin, Switzerland” investigates the transferability of parameters of a semi-lumped hydrological model within nested catchments in a high Alpine environment. It specifically explores the role of including physiographic information by implementing temperature-index models with decreasing simplicity for modelling snow and glacier melt. The authors conclude that including the effect of solar radiation in melt modelling increases the transferability of parameters, while including the effect of debris coverage of glaciers reduces parameter transferability.

The paper is well written, and presents a comprehensive modelling study and analysis that is of potential interest to the community of lumped hydrological modellers. A few concerns, however, occurred to me while reading the draft. These are outlined below and should be addressed by the authors in a revised version before the paper can be published in HESS.

The largest concern in my opinion is that the role of precipitation input is hardly discussed. The authors are making quite some efforts to explore the role of physiography for differences in production of melt water, which arguably is a major source for river discharge in their study area. At the same time, the role of physiography for differences in precipitation (snow, rain) input to the studied catchments is not analysed or discussed. The authors explain that local station data is limited and gridded data is used for precipitation, but it would be interesting to know if and how the precipitation characteristics could also explain some of the differences of the subcatchments. The focus here is on melt modelling, for which spatial differences in snow pack accumulation could be important. Similarly, also rainfall patterns could be important for shaping the discharge from different subcatchments, at least it appears that rainfall-runoff after depletion of snowpack produced some of the highest discharges in the observations (e.g., Figures 14 & E1).

I thus encourage the authors to add some analysis of the precipitation patterns, for example: Are there differences in precipitation input among the catchments? What about inter-annual variability? How do the gridded precipitation data compare to local info (at least two meteo stations are mentioned)? These issues should be discussed critically, especially if and how these relate to the parameter transferability between subcatchments.

Thank you for bringing to our attention that the current version of the paper does not pay enough attention to the role of precipitation. We currently have one graph showing the low variation between the daily precipitation in the different catchments of our study (Fig. 14a), but we will add the long-term annual precipitation trends for all catchments in the Supplement, as well as a comparison with meteorological station datasets. We will add a discussion paragraph to the article to discuss the role of precipitation.

Another doubt regards the bootstrapping approach the authors use as a benchmark. This is not critical for the evaluation of the paper, but perhaps some more explanation or even a revision would be possible. If I got it correctly, five years of observations were resampled in yearly blocks to obtain 100 benchmark series of discharge, such that each year is represented by a random choice of one of the other four years. Goodness-of-fit of observed and resampled series are calculated and averaged. The authors find that these benchmark values drop with decreasing size of subcatchments.

While I get the idea of providing a benchmark that preserves some of its characteristics like autocorrelation, it is not entirely clear to me what the assumptions behind this specific implementation of bootstrapping as a benchmark are.

The chosen bootstrapping method was retained because it is an easy metric to compute and it gives a good idea of the fit of the model in comparison with a regime simulated based on previous years. We will develop this point in the discussion.

Would it not be problematic if the precipitation dynamics were different in different years?

It would, indeed, make the bootstrapping less meaningful for years with different dynamics. However, we argue that our catchments are all dominated by the same snowmelt / icemelt dynamics that can be seen every year, even if with some temporal variability. This is why taking 100 combinations aims to compensate for outlier years. We will develop this point in the discussion.

How similar are the resampled series to each other, given that only five yearly blocks were used?

We will show in the Supplement of the revised version how different the resampled series are (by plotting them).

Why are 100 random combinations used, and not all possible combinations?

This heuristic choice was motivated by the origin of the bootstrapping method, that has a strong random component to it. This could have been done with all possible combinations.

Does it make a difference whether the series are averaged or the goodness-of-fit criteria are averaged?

Yes! Given the definition of the criteria, there is a non-linear mapping between the daily residuals (difference of the time series) and the criteria, accordingly, the criteria of the mean is not equal the mean of the criteria.

Could averaging the discharges for the same day of year provide an alternative and possibly more robust metric?

This is indeed a possible benchmark and the one introduced by Schaefli and Gupta (2007). This benchmark is particularly interesting for long time periods where it gives a robust representation of the average seasonal signal. In our case, we compute the benchmark over the period 2010 to 2014, which is short and the resulting average of the day of the year could be strongly influenced by a single year in that period, which can be avoided with the retained bootstrapping method. We will specify this in the revised version.

What insights does the drop of benchmark metrics with catchment sizes provide for similarity of the subcatchments, for example regarding their interannual dynamics?

Thanks for this comment, which relates to the drop of the benchmark values as a function of catchment size, as discussed at the start of the discussion section. Our justification based on the geomorphology (stream order) and in-stream flow paths was not clear (see a comment on the same point by reviewer 1), we will further elaborate on this. We will discuss better why the interannual streamflow variability is higher at smaller scales and also discuss the hydrological similarity between the catchments.

Furthermore, we will clarify that the bootstrapped series have each a length of 5 years, which was not clear from the current manuscript.

Further comments

Thank you for the detailed comments, which we will consider during the manuscript's revision. Below we answer those comments that go beyond simple corrections:

130-131: Does analysing the time lags between the hydrographs support your assumption?

Yes, for example, we found a time lag of about 15 minutes between the hydrograph of BI and the hydrograph of its biggest contributor, HGDA, (we took the 1rst of August 1985), which is negligible at the daily scale.

Fig. 8: It appears as if the maximum discharge of 1.0 was never captured by the model. What are the reasons for this discrepancy?

The maximum discharge of 1.0 in Figure 8 is reached by the observed dataset of BI, in June. All the discharge datasets in this plot are divided by this value, which explains why the other lines do not reach 1.0. This peak of discharge is related to a Foehn event (lines 259-261) that melted the snow but could not be captured efficiently by the model, probably due to a partial record of the temperature in the gridded input values or to the action of the wind, which is not accounted for in our model. We will explain the normalization and discuss this Foehn event further in the manuscript in link with the figure.

315-316: What about the spatial variations in precipitation? Can these also be highly variable?

Thanks, we will discuss this in the revised version.

372-375: I agree with that statement, but this directly invalidates your benchmarking approach, doesn’t it? “In contrast, even simple meteorology-based hydrological models deliver much better results” – so what would be the best option in the end?

This comment refers to the following sentences in the discussion: “Longer in-stream flow paths lead hereby to a stronger dampening effect of hillslope- and glacier-scale runoff variability. 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”

We see your point – an ideal benchmark should not depend on scale. But we do not see at this stage how to construct such a benchmark. We will make this clear in the revised version.

386: “As discussed previously” – maybe I missed it, but where was explained why simulated hydrographs should outperform NSE and match KGE?

This was not clear, we refer to the above lines “As a result, the NSE is much more sensitive to changes in bias, changes in variability or shifted yearly patterns than the KGE (see Supplementary

Material, F; Knoben et al., 2019). Thus, the benchmark KGE is a much harder criteria to meet for simulated discharges than the benchmark NSE”.

We will add the following sentence to this paragraph to make the transition smoother: “We thus expect the simulated hydrographs to outperform the benchmark NSE and match the benchmark KGE.”

399-404: If there were clear relationships of discharge and physiography - would taking them into account explicitly in your model solve a part of the transferability problem? This could hint at structural deficits of the model.

Yes we agree and will elaborate on this in the revised version.

References:

Horton, P., Schaefli, B., Hingray, B., Mezghani, A., and Musy, A.: Assessment of climate change impacts on Alpine discharge regimes with climate model uncertainty, Hydrological Processes, 20, 2091-2109, 10.1002/hyp.6197, 2006.

Schaefli, B., and Gupta, H.: Do Nash values have value?, Hydrological Processes, 21, 2075-2080, 10.1002/hyp.6825, 2007.