Dear Ole Rössler,

we appreciate the positive evaluation of our manuscript and the helpful comments.

Please find below our replies as inserted blue text.

Kind regards,

Nena Griessinger, Jan Seibert, Jan Magnusson and Tobias Jonas

The article addresses a highly relevant topic that is the added value of implementing external snow models and snow data in a hydrological model. In the present study, three different snow models of increasing complexity are attached to a HBV model and tested in 20 different mesoscale catchment within Switzerland. The catchment cover all altitudes present in the Alps. For all catchments the model performance of reproducing the runoff with in the snow melt seasons was assessed and served as the basis to judge the added value of the snow models. The authors found that the implementation of a snow model that additionally assimilates observed SWE data improves the runoff considerably, especially in high altitudes and in snow-rich years. The article is very well structured and written, concise and comprehensive at the time. The article is to my knowledge of original content and suits well in the scope of the journal. I still need to point out some more general concerns and a couple of minor comments. After a revision of the manuscript that take into account this concerns and comments, I would recommend publication:

superior comments/concern:

a) The first concern addresses the interpretation of the results. What is exactly the added value of the assimilated data set. Is it a more sophisticated and correct snow melt model or is it rather the added indirect information of precipitation amounts fallen in high altitudes where the meteorological station network is not present. My interpretation would be the latter, as the differences between model M1 and M2 (e.g. assimilation) are considerable for the highest altitudes. I would appreciate a discussion on this question.

As M1 and M2 differ in the use of the data assimilation algorithm and not in the snow melt model, the added value is based on the information coming from the point snow observations. We will in the revised manuscript in more detail point out the differences between the experiments M1 and M2 to avoid any misunderstanding, and discuss the causes leading to the differences in the results between the two experiments in more depth.

b) A follow up on this issue. The SLF station data are known to overestimate the SWE amounts. How was this issue addressed in the study and if not what are the consequences for your model as you may have calibrated your model against “differently wrong” data.

Following the feedback of referee #2, we will include a discussion about the representativeness and uncertainty of the used punctual snow depth data. The stations used in this study were chosen carefully avoiding sites that are clearly influenced by strong wind drift of snow, frequent sensor failures or known under- or overestimation of snow.

c) The LOO validation produces by nature highly variable performance values. I find it difficult to estimate differences between the models based on medians of boxplot. I would rather use a significance test. I recommend to show validation boxplots side by side and add notches to them.

Thanks for this suggestion. We will consider adapting this visualization of our results.

d) I found examples on the model performance given in Figure 3 and 4 show some room for improvements. Especially in Figure 3 it seems as the threshold for snowmelt was calibrated incorrectly. Is this threshold predefined by the external snowmodel? And if so, doesn't this mean that the snow model itself needs to be updated and calibrated against discharge? And I wonder what the upper benchmark model would look like.

The threshold (horizontal line) indicates 1.5 times the mean observed runoff during the snowmelt season of each year and is not predefined by the external snow model.

special comments/questions:

Page 1 Line 1: Abstract: The first sentence is somehow isolated from the rest of the text. I recommend to delete this sentence

We will delete this sentence.

P2 L1: and the erroneous precipitation input data at higher altitudes?

Thanks, we will add this and include references accordingly.

P3 L 32 "rain input" : which precipitation data set drives the snow model? Also the RHiresD?

Yes, daily RhiresD data were used as input to our snow model. We will clarify this.

P4 L1 ff: Is it correct that all model combinations HBV+M1-M3 as well as upper and lower benchmark models are calibrated? This is somehow suggested by Figure 5. In the calibration section I understood that a calibration was done for M3, upper and lower benchmark.

Yes, all combinations were calibrated. We will clarify this.

P5 L 2-3 what do you mean by "optimal interpolation approach". What magnitude of summed corrections can be found?

Thanks for your question. We refer to the mentioned study of Magnusson et al. (2014).

P5 L12: . . ... , but the RHiresD precipitation data set. Correct?

Yes, also here RhiresD was used as input to the snow model.

P7 L12ff and Figure4: However, the differences between M1-M3 are rather small for the snowmelt season as also indicated by the differences in NSE

Yes, we agree that the differences are rather small.

P7 L27: I wonder if the differences of the LOO validation are significant given the relatively large spread. (see general comments)

Thanks for your recommendation. We will discuss this in the revised version.

P7 L31 and Figure 5: - The benchmark lines are only the median of their respective boxplots? What is the spread of benchmark models? - The only difference between the benchmark model and M3 is a predefined DDF in M3 (cp.P5, L17-18)? Or are there further differences? If not, it is unexpected to see M3 to reach higher performance values then the upper benchmark. - Why is the performance of the benchmark model so weak in comparison to the other models especially in the lowest catchment class where snow does not really play a role?

Thank you for this comment that was also raised by referee #2.

There are further differences that lead to the performances seen here:

Dealing with liquid water content, refreezing, cold content dynamics and the partitioning of rain and snow are - among others - elements that influence the performance of temperature-index models. These elements differ between [M1, M2, M3] and HBV. In [M1, M2, M3] the representation of those processes have been particularly trained for optimal performance in the Swiss Alps.

Further, calibrating HBV for the melt season only – as done in our study – could result in a DDF that is too high during the snow accumulation period. The consequence might be an unbalanced performance with good snowmelt rates during the melt season at the price of too little accumulation earlier in the year with unwanted side effects on the snow depletion dynamics. M3 features a more moderate DDF of 2.5 $mm°C^{-1}day^{-1}$ allowing for a more balanced performance over the entire snow season.

We will adapt the manuscript and include this discussion in the revised version.

P8 L17: Please specify snow-rich: extreme snow years do not necessary result in an increased flood risks. To my understanding, largest snow melt contribution to runoff is expected if snow-covered area is largest and snow depth is widely insignificant (if SWE is above a certain minimum).

Thank you, we will clarify and discuss this issue in the revised manuscript.

P8 L30: in snow rich years the extent of snow in the lowlands is presumable larger then in snow-poor years. Accordingly, I also expected an effect of snow-rich years in the lowlands? Can you comment on this?

We did not analyze the performances for single years in the lowlands. Thanks for this remark which we will consider in the revised version.

P9 Conclusion: see superior comments

Figure 1: The blue lines on black are nearly invisible. Please change colors.

Thanks, will be changed.

Figure 2: Instead of showing one specific year, I would rather see a mean snow melt sum. In addition, maps showing differences between the models would increase readability.

Thanks for your recommendation. We discussed showing either cumulative sums or differences between the models with all authors in detail and we found this visualization appropriate for this paper.

Figure 3: Please indicate which model version is represented by the red dashed line.

Figure 3 serves only as graphical explanation of how to calculate E_{PF} and therefore the model version used here is not of importance.

Figure 4: Please add upper benchmark model

We will adapt the figure 4.

Table 1: Instead of numbers I would prefer to see the names of the catchments

We agree and will change the table.

References:

Magnusson, J., Gustafsson, D., Hüsler, F., and Jonas, T.: Assimilation of point SWE data into a distributed snow cover model comparing two contrasting methods. Water Resour. Res., 50(10), 7816-7835, doi: 10.1002/2014WR015302, 2014.