We thank Dr. Schaefli for their thoughtful review. Our responses can be found in blue throughout the following text. Please note, tables and figures specific to this response document are given with the prefix R (for example, Table R1 in the comment below). Tables and figures in the manuscript are referred to by numbers only.

This well written paper analyzes a key question for snow hydrology, which is the im- pact of precipitation phase algorithms on snow water equivalent (SWE) modelling in different climates. The paper studies four more or less different methods of precipita- tion phase computation (each with different portioning parameters) and assesses the impact of the methods on different snow accumulation and melt metrics, obtained with the model SNOWPACK at five different locations in the US. The methods are based on temperature thresholds and on bilinear regression. The analysis gives an answer to the general question of how important it is to carefully choose the precipitation phase method for different climates.

A drawback of the study is that it is purely simulation-based and does not use observed SWE data to push the study further. In fact, with the observed SWE data and SNOW- PACK, it might have been possible to estimate actual daily or hourly snow accumulation amounts and compute best parameter values for the studied precipitation phase meth- ods at the selected stations. This way, it would have been possible to judge how critical deviations from these best estimates would be at the different sites. In other words, this would allow to answer questions like "how critical is it to have a 1°C error in the air temperature threshold at a warm site as opposed to a cold site"? "How important is it to use dew point or wetbulb temperature at warm sites versus at cold sites?"

We agree this is a drawback of both this study and many other snow modeling research projects. There are, unfortunately, scant direct observations of precipitation phase in mountain regions. One of the few studies we are aware of that uses observations of precipitation phase—in this case snow board measurements—showed rain-snow partitioning errors can lead to significant biases in modeled snow accumulation at a site with a maritime climate similar to the HJ Andrews and Southern Sierra (Wayand et al., 2017). There is also evidence suggesting an optimized air temperature threshold varies throughout the snow season (Storck et al., 2002), meaning no single air temperature threshold (or range) would be applicable across sites and times.

At our study sites, there are no direct observations of precipitation phase, but we were interested in pursuing your question further. Table R1 below shows the optimized rain-snow air temperature threshold using four different data sources for each station. The second and third columns (Map and Obs.) correspond to data from earlier work that

examined the spatial variability of rain-snow partitioning across the Northern Hemisphere (Jennings et al., 2018). The methods, quoted from the paper, are as follows:

> *"To construct a spatially continuous 50% rain–snow T_s* [air temperature] *threshold product across the Northern Hemisphere, we applied the optimized bivariate model to the MERRA-2 gridded reanalysis dataset[63,64]. Hourly 2 m T_s, specific humidity (q), P_s, and precipitation data were accessed from 1980 through 2007 and summarized to a daily time step. RH was calculated from the MERRA-2 data using an empirical equation as a function of q, P_s, and T_s. Daily snowfall probability was then simulated for each grid cell using the bivariate model when precipitation was greater than 1 mm and T_s fell within the range of −8 to 8 °C. We then calculated the 50% rain–snow T_s threshold by fitting the hyperbolic tangent to binned estimates of snowfall frequency per MERRA-2 grid cell using Eq. 1."*
>
> *"We classified precipitation reports as either rain or snow using the World Meteorological Organization precipitation phase categories described in detail in Dai[40,61]. Precipitation amounts were not included in the dataset and we removed sleet as well as potential mixed-phase observations from the analysis because the relative proportions of solid and liquid precipitation during such events were not reported (i.e., it was impossible to quantify the amount of precipitation falling as snow versus rain). The classification of precipitation events was then used to quantify the rain–snow frequency per 1 °C T_s bin from –8 to 8 °C at each station. In other words, if there were 100 total precipitation observations from 1 to 2 °C, 75 of which were snow, the snowfall frequency in that bin would be 75.0%. We then calculated the 50% rain–snow T_s threshold for each station using the approach of Dai[40], where a sigmoidal curve is fit to observations of snowfall frequency per 1 °C T_s bin from –8 to 8 °C using a hyperbolic tangent function:*
>
> $$T_{50} = \frac{\tanh^{-1}\left(\frac{F}{a} + d\right)}{b} + c \quad (1)$$
>
> *where T_{50} equals the 50% rain–snow T_s threshold (°C), F equals snowfall frequency (in this case 0.5, dimensionless), and a, b, c, and d are the fitting parameters (dimensionless)."*

The fourth and fifth columns in Table R1 use changes in SWE and snow depth to estimate a rain-snow air temperature threshold. We used a modified version of the approach of Rajagopal and Harpold (2016) to predict precipitation phase by designating a

daily increase of SWE or snow depth as snowfall and a zero change or decrease as rainfall when precipitation was greater than 2.54 mm and SWE or snow depth was greater than 0 mm. As with the Map and Obs. methods detailed above, we then binned snowfall frequency per 1°C air temperature bin (Figures R1 and R2) and computed the rain-snow air temperature threshold using Eq. 1 above. The SWE approach yielded values that approximated the Map and Obs. methods, but the depth-derived values were significantly lower. We would thus argue that this method was not appropriate for our purposes, although previous work has shown it to reasonably estimate precipitation phase at subdaily time scales (e.g., Marks et al., 2013; Zhang et al., 2017).

Table R1. Optimized rain-snow air temperature thresholds for each station in the study using four different data sources: 1-Map) The spatially continuous threshold map from Jennings et al. (2018) created using reanalysis data from MERRA-2 and the bivariate binary logistic regression model; 2-Obs.) The observed threshold from the closest meteorological station (Jennings et al., 2018); 3-SWE) The threshold inferred from changes in SWE at each study station (Fig. R1); 4-Depth) The threshold inferred from changes in snow depth at each study station (Fig. R2). An NA indicates there were insufficient data to estimate the threshold from SWE and/or snow depth.

	Optimized rain-snow air temperature threshold (°C)			
Station	Map	Obs.	SWE	Depth
HJA-CEN	1.19	1.12	1.29	-0.24
HJA-VAN	1.19	1.12	0.8	-0.84
HJA-UPL	1.19	1.12	-0.4	-0.81
SSC-LWR	1.7	1	NA	0.14
SSC-UPR	1.7	1	0.87	-0.34
YOS-DAN	2.21	2.78	NA	NA
JD-125	2.25	1.25	NA	-0.97
JD-124b	2.25	1.25	NA	-1.91
JD-124	2.25	1.25	NA	0.41
NWT-C1	2.84	2.34	3.57	NA
NWT-SDL	2.84	2.34	NA	NA

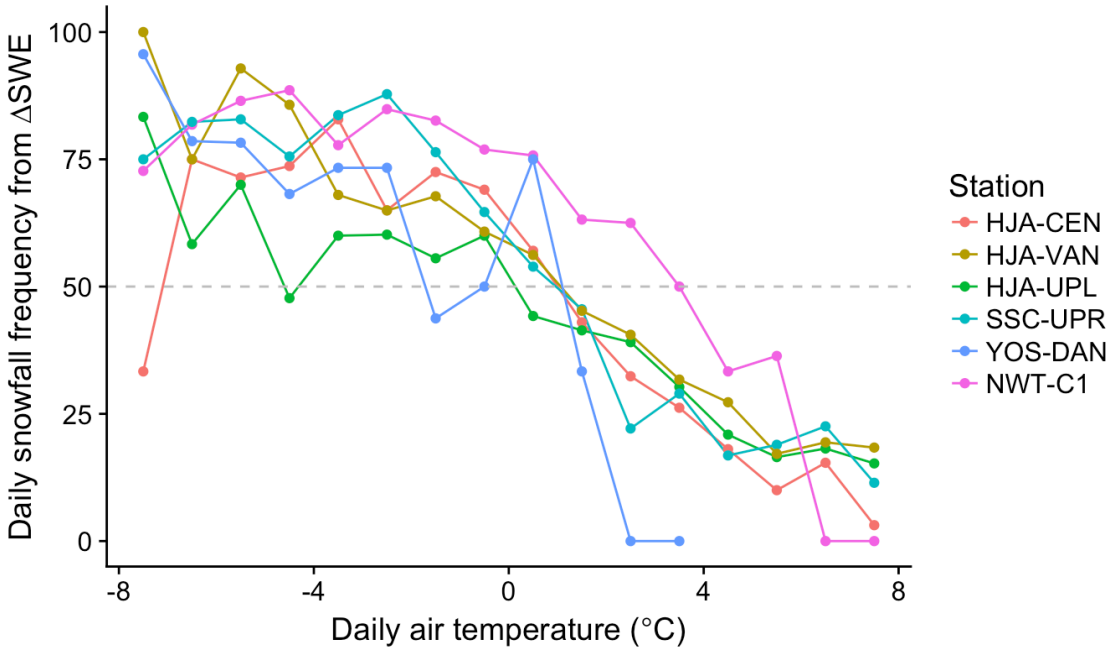

Figure R1. Snowfall frequency per 1°C air temperature bin as computed from SWE data. On days with precipitation > 2.54 mm, an increase in SWE was designated as a snowfall event, while a zero change or decrease in SWE was designated as rainfall.

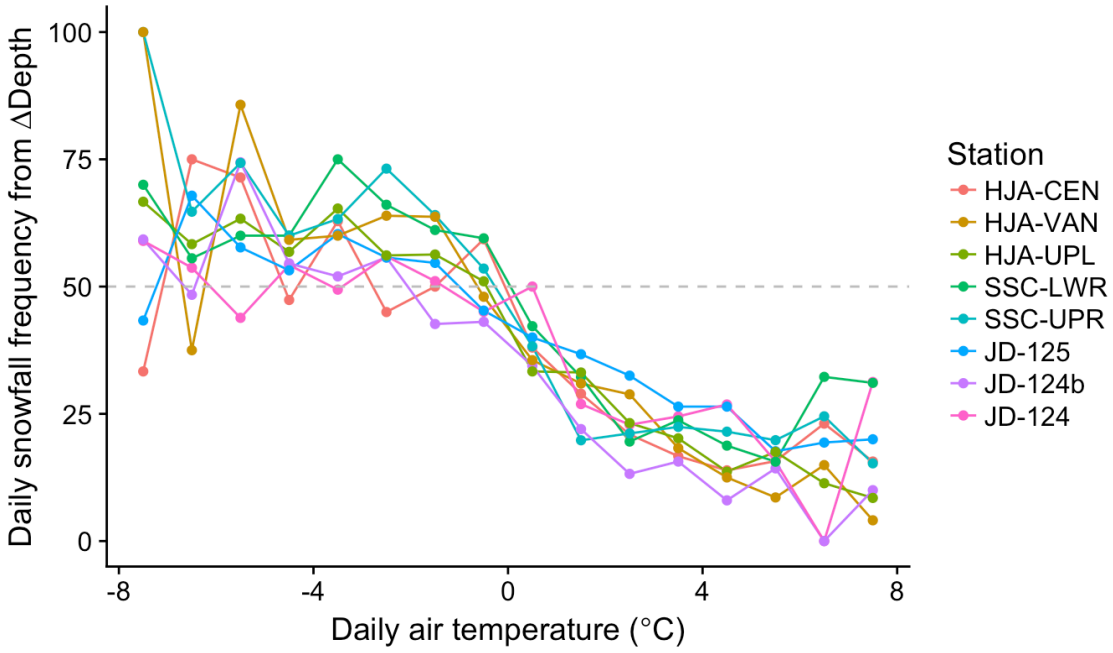

Figure R2. Snowfall frequency per 1°C air temperature bin as computed from snow depth data. On days with precipitation > 2.54 mm, an increase in snow depth was designated as a snowfall event, while a zero change or decrease in snow depth was designated as rainfall.

Returning to the question of "how critical is it to have a 1◦C error in the air temperature threshold at a warm site as opposed to a cold site," we analyzed the effect of deviating by 1°C from the mean threshold as calculated from the Map and Obs. columns in Table R1. In this context we rounded to the nearest integer degree to be consistent with our thresholds, giving the HJA stations a 1°C threshold, SSC a 1°C threshold, YOS a 2°C threshold, JD a 2°C threshold, and NWT a 3°C threshold. Because we did not include a 4°C air temperature threshold in our phase methods, we could only analyze a negative deviation at NWT. In Table R2 below, we present the mean peak SWE, peak SWE day of water year (DOWY), and snow cover duration (SCD) using the optimized air temperature threshold (center column, abbreviated Thresh.), the optimized threshold minus 1°C (left column, Thresh - 1°C), and the optimized threshold + 1°C (right column, Thresh + 1°C). Consistent with our findings in the paper, the warm maritime HJA and SSC stations are profoundly affected by deviations from the optimized threshold. Differences at these sites produced by deviating by ±1°C from the optimized thresholds range between 141 and 403 mm for peak SWE, 1 and 16 d for peak SWE DOWY, and 9 and 29 d for SCD. Compare this to 1 to 10 mm for peak SWE, 0 to 1 d for peak SWE DOWY, and 1 to 5 d for SCD at the YOS and NWT stations. The consistent story is again that threshold choice makes a much larger impact at a warm site relative to a cold one.

Table R2. Mean peak SWE, peak SWE DOWY, and SCD at the study stations using an optimized air temperature threshold as well as -1°C and +1°C deviations from the threshold.

	Mean peak SWE (mm)			Mean peak SWE DOWY (d)			Mean SCD (d)		
Station	Thresh - 1°C	Thresh.	Thresh + 1°C	Thresh - 1°C	Thresh.	Thresh + 1°C	Thresh - 1°C	Thresh.	Thresh + 1°C
HJA-CEN	414.2	528.9	611.8	128	142	144	144	159	172
HJA-VAN	564.3	645.3	726.6	132	134	145	164	173	184
HJA-UPL	984.2	1165.5	1387.3	160	166	173	190	202	210
SSC-LWR	401.5	535.6	624.6	154	161	162	137	146	151
SSC-UPR	508.4	585.4	649.7	154	155	155	142	147	151
YOS-DAN	668.8	677.8	678.8	169	169	170	206	209	209
JD-125	77.2	89.6	99.9	116	116	116	75	83	97
JD-124b	180.2	191.7	203.8	124	126	127	122	131	134
JD-124	72.3	81.5	87.3	128	115	115	78	81	92
NWT-C1	400.1	406.7	NA	204	204	NA	224	229	NA
NWT-SDL	914	914.6	NA	225	225	NA	240	241	NA

For the final question, "How important is it to use dew point or wetbulb temperature at warm sites versus at cold sites?", we would argue the best practice is to use a humidity-based temperature metric at all sites. Such methods better represent precipitation and

produce better model outcomes (e.g., Ding et al., 2014; Harder and Pomeroy, 2013, 2014; Harpold et al., 2017; Jennings et al., 2018; Marks et al., 2013). The bivariate binary logistic regression model, which performed best relative to other methods when compared to precipitation phase observations in a previous study (Jennings et al., 2018), produced snow cover metrics similar to the optimized threshold at most stations. It produced mean peak SWE, peak SWE DOWY, and SCD biases (relative to the optimized threshold) of -18.0 mm, 0.5 d, and -1.9 d, respectively.

Please note, we have not added the above material to the manuscript yet because it is consistent with the findings already presented in the submitted document. If you find this material worthy of inclusion, please let us know and we can add it as either supplementary material or as an appendix.

This having said, the study is nevertheless worth publishing and interesting for the readers of HESS. Below some general and detail comments.

General comments

I would not say that a study tests 12 different methods if only a few methods are tested with different parameter values; this oversells the study in the abstract. I would in fact say that the study tested four different methods: based on air temperature (with different 50% thresholds and different transition ranges, some of the ranges being 0), based on dew point and wet bulb temperature and based on binary regression.

Fair point. We have updated the text (see response to detailed comments below) to say we tested 5 different methods (counting the range as a different method than the threshold because the former produces mixed precipitation and the latter does not).

A key analysis of the paper is the one of "Climatic controls on precipitation phase method sensitivity".(section 4.4); it analyzes how the results vary with air temperature. Air temperature sensitivity is, however, built into each method in a different way. In the case of daily snowfall fraction: the fact that it shows the highest standard deviation for air temperatures between 0 and 4 C simply expresses the fact that several methods use thresholds in this range. The result would look different if the thresholds were between -2 and 2 C. This should be better reflected in the discussion of of the results.

Correct, the variability is tied to the methods themselves. However, we think it is important to present this information because the methods we used are based on empirical relationships (air temperature thresholds and ranges, dew point temperature thresholds), physical principles (wet bulb threshold to approximate hydrometeor

temperature (Harder and Pomeroy, 2013)), and statistical relationships (the binary logistic regression models). A threshold of -2°C would likely widen the range of variability but it would have no empirical, physical, or statistical relationship to precipitation phase partitioning except in some extremely rare, unique cases. Furthermore, this comment was similar to the feedback from Dr. Jono Conway, who noted the range in variability was likely produced by the extreme air temperature thresholds and ranges (T_{a0}, T_{ar0}, and Ta_3). To respond to his comment, we removed these methods and re-performed the analysis and the finding was the same (please see Figure R2 in our response to Dr. Conway). Even limiting the analysis to the most representative methods, the variability stays highest between 0°C and 4°C.

Additionally, it is essential to point out this air temperature range of variability for two reasons:

1. Areas most "at risk" to the snow-rain transition due to climate warming have seasonal air temperatures near and slightly above freezing (e.g., Nolin and Daly, 2006)
2. 0°C to 4°C is also the air temperature range where precipitation phase methods perform the worst (Ding et al., 2014; Jennings et al., 2018).

Thus, we have a compounded problem in that we are concerned with snow-to-rain shifts in areas with seasonal air temperatures where precipitation phase partitioning is most uncertain and our available methods exhibit downgraded performance. Given that we showed these areas (i.e., winter and spring air temperatures above freezing) also express the greatest sensitivity in terms of peak SWE magnitude and timing, plus snow cover duration, we think it is necessary to include this information.

In general, the conclusion that precipitation falling in the range 0 – 4 C explains much of the variation observed across the methods comes from the choice of the threshold values. Without actual comparison to observed data, the results are hard to generalize. Why is there no comparison to actual SWE-derived thresholds?

Please see our responses above.

Furthermore, when reading the results section where actual SWE curves are presented for the first time, it is a little disappointing to see that all studied sites show a typical seasonal snow cover with significant accumulation over many weeks. The most sen- sitive sites would typically be the ones where the snow cover might build up several times during the winter.

We should note here that the SWE curves as presented are daily averages (Fig. 4 in submitted manuscript), which has the affect of obscuring transience. As we mentioned in the Study sites and data section, the HJA and JD stations are sometimes transient (p. 4 lines 9-10 through p. 5 line 1, and p.5 lines 21-22) and they are most sensitive to phase method choice in terms of peak SWE magnitude (HJA only) as well as peak SWE timing and SCD (Fig. 5 in submitted manuscript).

Detailed comments

- The abstract does not mentioned what types of methods have been tested nor whether they have been compared to reference data or which method performed best

Yes, that is an oversight on our part. We have changed the abstract to note:

"The methods in this study included different permutations of air, wet bulb, and dew point temperature thresholds, air temperature ranges, and binary logistic regression models."

We have also added a line saying:

"Compared to observations of snow depth and SWE, the binary logistic regression models produced the lowest mean biases, while high and low air temperature thresholds tended to respectively overpredict and underpredict snow accumulation."

- Introduction: it would have been interesting to shortly discuss how /where precipitation phase is actually observed; as far as I am aware of, actual precipitation phase observations are crucially missing at most places.

Good point. We have added a line to the first paragraph of the Introduction:

"Complicating matters is the fact precipitation phase is rarely observed in mountain regions on a continuous bases over long time scales."

- Introduction: the manuscript focuses its discussion on snow-hydrological models. How do meteorological forecast models determine the limit (elevation) of snow fall? Completing the literature review with this respect would complete the picture

This is covered in discussion (p. 21 lines 4-17) and not necessary for the Introduction as we do not utilize any atmospheric model methods in this work.

- P. 2: “In general, warmer sites are more sensitive to precipitation phase method selection in terms of annual snowfall fraction variability, though it is less certain how this variability translates into divergences in simulated snow accumulation and melt. “ This statement is given without reference. In what is the apparently previously known result different from your own findings?

Text changed to: *“This previous work has shown, in general, warmer sites are more sensitive…”* in order to clearly connect the statement with the published literature in the previous line.

- Study sites: It might be useful to know the variability of the daily air temperature around the seasonal mean (ie. the anomalies, obtained e.g. by fitting a sine curve to air temperature as in the work of Woods, 2009. It is this variability that will tell something about the probability of switching from accumulation to melting conditions and about a site sensivitiy to the chosen temperature threshold.

This sounds similar to the point raised by Nayak et al. (2010), who showed the effects of switching from sub-freezing to freeze-thaw diurnal cycles on snowpacks at Reynolds Creek. It is clear fluctuations above and below freezing having important effects on snow cover energetics. However, we are unclear as to what new, relevant information such data would provide to the current study. Perhaps we are misunderstanding the comment, so please clarify if so.

- Methods: it is not clear at this stage that all stations always show a seasonal snow cover (significant accumulation over several weeks), which is important for the concept of “peak SWE” to be meaningful

It is noted in the Study sites and data section (p. 4-5) for each location whether seasonal snowpacks develop or not.

- the current definition of snowmelt rate is probably over sensitive to spurious shifts from a primary to a secondary SWE peak, which could reduce the melt dura- tion sensibly; how could this measure be made more robust? Similar comment applies to the peak SWE date that is discussed in the results section. Is this measure useful? Minor modifications of SWE accumulation can switch the SWE peak date between a spurious primary or secondary peak (Figure 4 suggest that stations with two peaks might exist, but I might be mistaken).

We noted on p. 15 (lines 3-6): *“We found the greatest differences in peak SWE dates*

were generally simulated on years with low/transient snow cover. In these cases, late-season precipitation was simulated as rain by the low T_a thresholds and snow by the high T_a thresholds, meaning an early SWE maximum was recorded as the peak in the former case and a late SWE maximum in the latter case." Given that peak SWE timing is an important measure of melt onset in the western US, we find it is necessary to highlight the variability in this metric as produced by different phase methods. Our finding indicates research on future changes to snowmelt timing is affected by modeling decisions on assigning precipitation phase.

Regarding snowmelt rate, we present the seasonal melt rate or ablation slope (e.g., Trujillo and Molotch, 2014) because of the importance of the spring snowmelt freshet to streamflow generation in many mountainous areas of the western US. However, we admit this overlooks the important winter contributions of snowmelt to groundwater and streamflow in maritime and transient snow environments. Switching the analysis to include all days when snowmelt was > 0 mm, we found marginal differences across the precipitation phase methods (mean differences were all less than 2.2 mm d^{-1}, which is less than the nominal precision of the SNOTEL snow pillows in the western US). Looking at daily average melt rate differences between the T_{a0} and T_{a3} thresholds helps illustrate why. Figure R3 below shows that generally T_{a0} produces higher melt rates than T_{a3} early in the snow cover season, while the reverse is true later in the season. Although annual average melt rates exhibit few differences, this figure shows the timing of terrestrial water inputs is important.

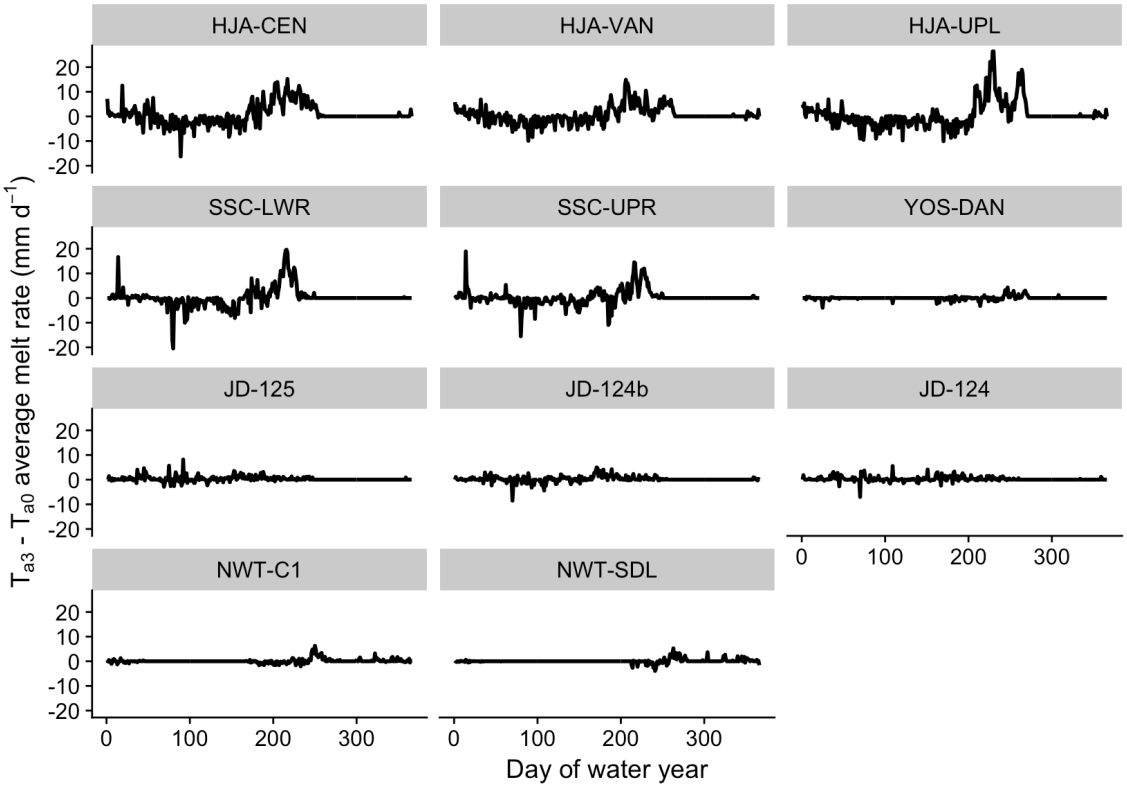

Figure R3. The difference in daily average snowmelt rate between T_{a3} and T_{a0}.

- P. 14 "meaning a significant proportion of water was simulated to have run off using one precipitation phase method versus being stored in the snowpack". This not well formulated since rainfall does not necessarily run off. It can infiltrate and recharge the groundwater.

We agree this was imprecise wording. We have changed this to, *"meaning a significant proportion of water was simulated to have infiltrated or run off using one precipitation phase method versus being stored in the snowpack…"*

- Section 4.4: Here, standard deviations are calculated across the results of all 12 computation methods. Standard deviation does not seem to be a good measure to quantify the variability of values that do not come from an actual sample of a given process but of values pertaining to different methods. (Besides: how are standard deviations obtained? First per method and then averaged over all methods?)

The standard deviation values presented in Section 4.4 and Figure 6 are computed per air temperature and RH bin across all stations and methods as noted in the text. Although standard deviation is an appropriate metric of variability in this context, we redid the

analysis using the uncertainty formulation from Harder and Pomeroy (2014). We modified it to be per RH and temperature bin. The result was the same (Figure R4):

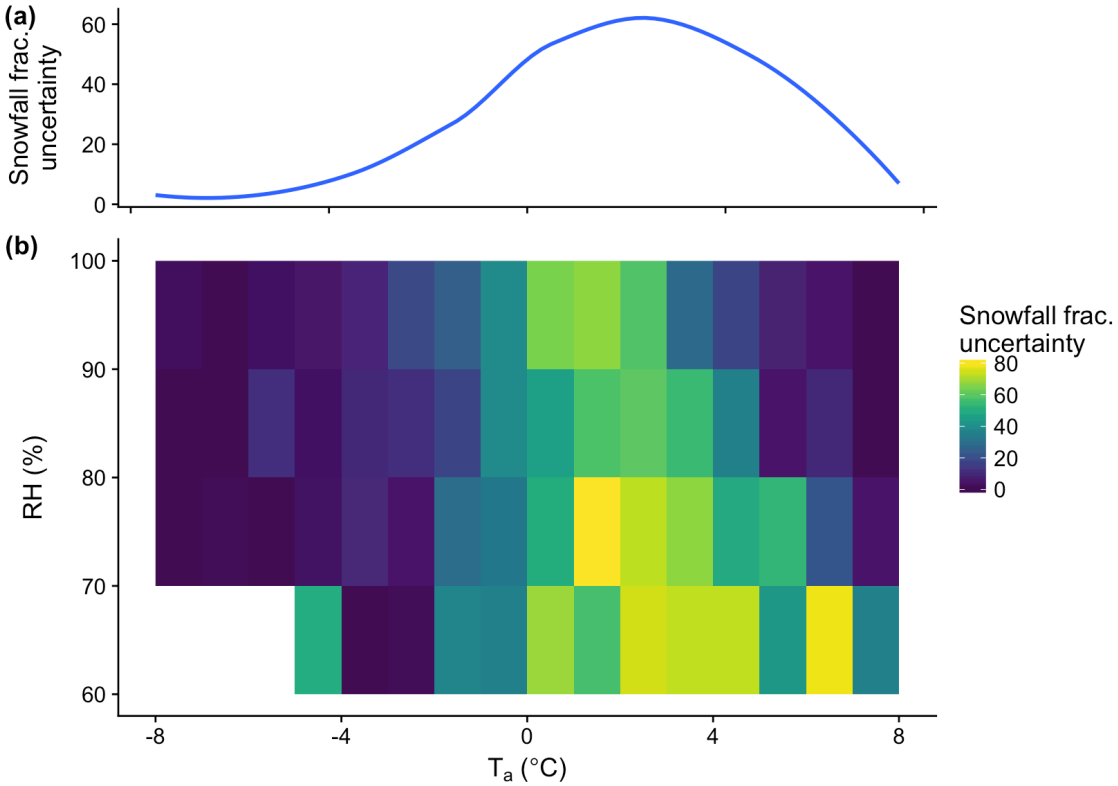

Figure R4. Same as Figure 6 in submitted manuscript, but standard deviation is replaced with the uncertainty metric from Eq. 1 in Harder and Pomeroy (2014).

Screenshot of text from Harder and Pomeroy (2014) showing the uncertainty metric equation:

$$uncertainty = \frac{\sum_{i=1}^{n}(\mathrm{Max}_i - \mathrm{Min}_i)}{n} \qquad (1)$$

where Min and Max refer to the lowest and highest values of a model output variable from the 63 model runs, i is the index (time step) of the value and n is the number of values (total time steps). The units of uncertainty are the same as the hydrological variable being considered. The uncertainty and differences between PPMs are summarized using mean values over an entire hydrological year (1 October–30 September).

References

Ding, B., Yang, K., Qin, J., Wang, L., Chen, Y. and He, X.: The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization, J. Hydrol., 513, 154–163, 2014.

Harder, P. and Pomeroy, J.: Estimating precipitation phase using a psychrometric energy balance method, Hydrol. Process., 27(13), 1901–1914, doi:10.1002/hyp.9799, 2013.

Harder, P. and Pomeroy, J. W.: Hydrological model uncertainty due to precipitation-phase partitioning methods, Hydrol. Process., 28(14), 4311–4327, 2014.

Harpold, A. A., Kaplan, M., Klos, P. Z., Link, T., McNamara, J. P., Rajagopal, S., Schumer, R. and Steele, C. M.: Rain or snow: hydrologic processes, observations, prediction, and research needs, Hydrol Earth Syst Sci, 21, 1–22, 2017.

Jennings, K. S., Winchell, T. S., Livneh, B. and Molotch, N. P.: Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere, Nat. Commun., 9, doi:10.1038/s41467-018-03629-7, 2018.

Marks, D., Winstral, A., Reba, M., Pomeroy, J. and Kumar, M.: An evaluation of methods for determining during-storm precipitation phase and the rain/snow transition elevation at the surface in a mountain basin, Adv. Water Resour., 55, 98–110, doi:10.1016/j.advwatres.2012.11.012, 2013.

Nayak, A., Marks, D., Chandler, D. G. and Seyfried, M.: Long-term snow, climate, and streamflow trends at the Reynolds Creek Experimental Watershed, Owyhee Mountains, Idaho, United States: CLIMATE TRENDS AT RCEW, Water Resour. Res., 46(6), n/a-n/a, doi:10.1029/2008WR007525, 2010.

Nolin, A. W. and Daly, C.: Mapping “at risk” snow in the Pacific Northwest, J. Hydrometeorol., 7(5), 1164–1171, 2006.

Rajagopal, S. and Harpold, A. A.: Testing and Improving Temperature Thresholds for Snow and Rain Prediction in the Western United States, JAWRA J. Am. Water Resour. Assoc. [online] Available from: http://onlinelibrary.wiley.com/doi/10.1111/1752-1688.12443/full (Accessed 23 August 2016), 2016.

Storck, P., Lettenmaier, D. P. and Bolton, S. M.: Measurement of snow interception and canopy effects on snow accumulation and melt in a mountainous maritime climate, Oregon, United States, Water Resour. Res., 38(11), 5–1, 2002.

Trujillo, E. and Molotch, N. P.: Snowpack regimes of the Western United States, Water Resour. Res., 50(7), 5611–5623, doi:10.1002/2013WR014753, 2014.

Wayand, N. E., Clark, M. P. and Lundquist, J. D.: Diagnosing snow accumulation errors in a rain-snow transitional environment with snow board observations, Hydrol. Process., 31(2), 349–363, doi:10.1002/hyp.11002, 2017.

Zhang, Z., Glaser, S., Bales, R., Conklin, M., Rice, R. and Marks, D.: Insights into mountain precipitation and snowpack from a basin-scale wireless-sensor network, Water Resour. Res., 53(8), 6626–6641, doi:10.1002/2016WR018825, 2017.