General comments

This paper presents a threefold approach to assessing gridded precipitation datasets in mountainous regions. The approach is tested over the Upper Indus Basin, using five precipitation datasets, and observations from precipitation gauges, runoff measurements and glacier mass balance measurements. The first approach is a comparison with observations, the second a balance between precipitation, runoff and evapotranspiration, and the third uses a hydrological model (including glacier and snow components) to compare to observed runoff. Overall, the manuscript presents an interesting approach to resolving the issues of assessing precipitation datasets in data-sparse regions with complex topography. There is also an interesting analysis of the performance of these datasets over different regions of the Upper Indus Basin. In particular, the seasonal analysis over the datasets demonstrates that while most datasets represent the summer monsoon well, there is far more variation over the representation of the winter westerlies.

Major comment

My main comment concerns the second approach, referred to as the 'physical diagnosis' in the manuscript, which aims to determine whether a basin can be considered plausibly realistic using a water energy balance. Using the precipitation observations, the authors show that this catchment is not 'physically realistic' as defined using the Truc-Budyko plot and suggest that this is likely to be due to glacier melt and storage. This nicely motivates the use of SPHY, which includes a glacier component, but suggests that this second method is not suitable for assessing precipitation datasets in regions with snow and glaciers, as it is missing an essential physical component. The authors use this to assess glacier change based on each dataset, but it is not clear how this gives extra information about the reliability of the precipitation datasets themselves. In my opinion, this approach could be removed completely (or significantly reduced to just looking at observations and demonstrating why this method is not suitable). This would also help to shorten the manuscript somewhat, as it is currently quite long.

Minor suggestions

Some minor suggestions are listed below (I should note that I was not able to see any of the figure or table references in the manuscript, which may have led to some misunderstandings on my part):

Given that not all readers will necessarily be familiar with the statistical plots presented here, it would be useful to have more information in the figure captions, and to define the acronyms used in the figures. E.g. for figure 8 "the boxes represent ….." e.g. for figure 10 "each dot represents …. (an average/total for one year of data?)".

Throughout the manuscript, the authors should check that acronyms and abbreviations are defined where they first appear, and consider repeating these, or providing a nomenclature. For example on line 95, Q, P, ET_p all need defining, on line 210 ETCCDI needs defining.

Line 174: Was this adjusted using the same wind data as used in Dahri 2018? I think given the importance of this adjustment, an extra line explaining it would be useful.

Line 300: Please add references to the sentence 'The observed mass balance data were extracted from the literature'.

Line 305: Where is the data from Besham Qila held or is there a reference for this data?

Line 323/Figure 5: Could you plot the points of the stations onto the map in figure 5 so it's easier to compare the observations and colour maps? More importantly, please switch the colour scheme for either figure 3 or

5 so that they are the same, with the same scale (a divergent colour scheme for both would make it clearer which areas are high precipitation and which are low precipitation). It's tricky to compare them when blue is dry regions in one plot and wet regions in the other.

Line 326: While it's clear what you mean here, I think technically this should be 'the GDPs did not show statistically significant trends', as you cannot generally use statistical tests to prove a lack of trend.

Line 331-333: Are these discussions about bias coming from the Taylor diagram in figure 6? It might be better to talk about RMSE, as that's what you have shown here in figure 6.

Line 333: word missing '….as the better in UIB…' -> '…as the better model in the UIB…'

Line 345: It would be good to emphasize that this under/overestimation is particularly true during the winter, as it's interesting that these datasets appear to represent the summer monsoon much more effectively than the winter westerlies.

Lines 349-355: You could consider cutting figure 7 b and the accompanying text, as I think this is all shown in figure 7 a and that discussion. If figure 7 b is kept, could the seasons be put in order? I'd recommend Winter, spring, summer, autumn as this will make it easier to see the winter and spring precipitation together.

Lines 357-368: It's not quite clear how these numbers relate to figure 8, or how the values in figure 8 and the numbers in the text are calculated. Are the mean and standard deviations given in the text taken from each year? I.e. the maximum CDD taken from a year, and then averaged over all the years? Given figure 8 shows the median and 25^{th}/75^{th} percentiles, it might be more useful to discuss those? (although presumably the red dots are the means in each case).

Figure 9/lines 372-382: is the runoff value the same for each of the models? Is this a measured value? Please state in the text and the figure caption.

Lines 383-391: This section discusses correlation between runoff and precipitation. However given there is no significant correlation between the observed precipitation and runoff, except in the Karakoram, it seems that there may be other factors that need to be taken into consideration, and therefore that correlation between these two variables probably should not be used to judge the datasets?

Line 456: Are these six SPHY model runs identical, except for the precipitation dataset used? Or are there some differences, apart from the precipitation datasets?

Line 494-497: I don't quite understand this.

Line 510: Could you add some references here?

Line 589-594: you suggest undercatch in the observations here as a reason for the unbalanced water balance, but haven't you already corrected for this?

Lines 595-598: is increasing glacier mass balance also a reason for a 'leaky' catchment?

Line 613: repetition of the underestimated GPDs.

Figure 6: please keep the colours for each model matching to those in figure 5.

Figure 14: What's the difference between a and c? Does figure e include the glacier cover, or is it only snow cover? If it's only snow cover, I don't see how in figure (e) MODIS and SPHY look so similar for august, when the snow cover (green only) look quite different in b and d.