We thank Prof. Scherer for his very useful comments. We have addressed his comments below. We show how the text in the manuscript has changed, by indicating new text in boldface.

Comment: First of all, the entire study depends on the accuracy of downscaled precipitation. It would therefore be of utmost interest to better understand the uncertainties in the WRF output. As the authors correctly state, in-situ meteorological observations are scarce, and there is almost complete lack of data in the WKSK ranges, which makes it difficult to compare the WRF output with independent observations. This is especially true for high altitudes, i.e., the glacierized areas, where observational data are not available. Nevertheless, there are gridded data sets that could be used for comparison. Although they do not cover the entire study period (so far) and thus cannot substitute the ERA-Interim data used for downscaling, they could anyway be compared with the WRF results for shorter periods (as the authors have done with GLEAM data). The new ERA5 reanalysis and the High Asia Refined analysis (HAR) data set (Maussion et al., 2014) are suitable data sets in this respect. ERA5 data, especially the newest ERA5 land data set (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview), and the HAR data set do have very high spatial and temporal resolutions, such that they resolve mesoscale atmospheric processes, and thus orographically induced precipitation. HAR data are freely available at www.klima.tu-berlin.de/HAR. I would ask to authors to include a comparison of WRF output with these gridded data sets in the article. This could be put into a supplement with only a short paragraph in the main text.

Reply: We agree that such a comparison between different datasets will be a great addition to the manuscript, although we note that this does not necessarily increase the confidence in the results in WKSK, given the lack of ground truth for all these datasets.

We now added several paragraphs and two figures in the main text to deal with the comparison:

We also compare our WRF simulations with three similar data products with relatively high spatial resolutions, which have recently become available. We do note that all these datasets suffer from the lack of ground truth in WKSK, which means we cannot determine which dataset performs best in this region.

ERA5 is the follow-up of ERA-Interim (Copernicus Climate Change Service, 2017), with an improved spatial resolution of 0.25°, an improved temporal resolution, a more appropriate model input for e.g. sea surface temperatures, and more assimilated data. ERA5-Land is atmospherically forced by ERA5, and provides an even higher spatial resolution (0.1°) for land surface properties (Copernicus Climate Change Service (C3S), 2019). Finally, we include the HAR dataset with a resolution of 10 x 10 km, which uses WRF to downscale the NCEP FNR reanalysis dataset and re-initialises every day (Maussion et al., 2014). We compare temperatures between May-September, and annual precipitation, which give an indication of the parameters that are most relevant for glacier mass balance modelling. Because of the limited time overlap between the different datasets, we could only fully compare the period 2001-2010.

We binned all data to the same 0.5° x 0.5° grid to allow direct comparison. The mean values, trends, and interannual variability are compared in Figs. 3 and 4. It shows that ERA5 and ERA5-Land are nearly identical, and we only refer to ERA5 below. Our WRF model yields a warmer Karakoram than the other three datasets. Generally, the mean temperature differences are relatively minor, except for a warmer Tarim basin compared to HAR. We find very similar temperature trends as ERA5, although with smaller magnitudes. The magnitudes of the trends are also generally smaller than those in the station data (Fig. 2). The WRF interannual temperature variations correlate very well with ERA5, except two areas in the Tarim and the inner Tibetan Plateau. This is not surprising, given that our WRF model is forced by the similar ERA-Interim data. The whole western part of HMA, including WKSK, is especially well-correlated to ERA5. In that region, the correlation with HAR is weaker, but the correlation between HAR and our WRF data is very strong in East HMA. The differences with HAR might be explained by the different forcing, or by the difference in used physics modules, but this requires further study.

Differences between datasets are larger for precipitation, at least for the mean values and interannual variability. Our WRF simulations give results that are relatively wet in the Karakoram, and relatively dry in the Himalaya. However, the precipitation trends are very similar to ERA5 in both pattern and magnitude. An exception is the arid Tarim basin, which has an increasing trend in WRF, but a decreasing trend in ERA5. HAR shows a positive precipitation trend in most of HMA, with a very high trend in the Tarim basin. The correlation of the interannual variability is low in WKSK and parts of Tien Shan, which could be explained by the relatively high influence of the irrigated areas in the Tarim basin on the annual precipitation (de Kok et al., 2018, Fig. 3). Since our WRF model outcome is the only one of the four datasets that explicitly includes irrigation, this could explain the difference in annual variability.

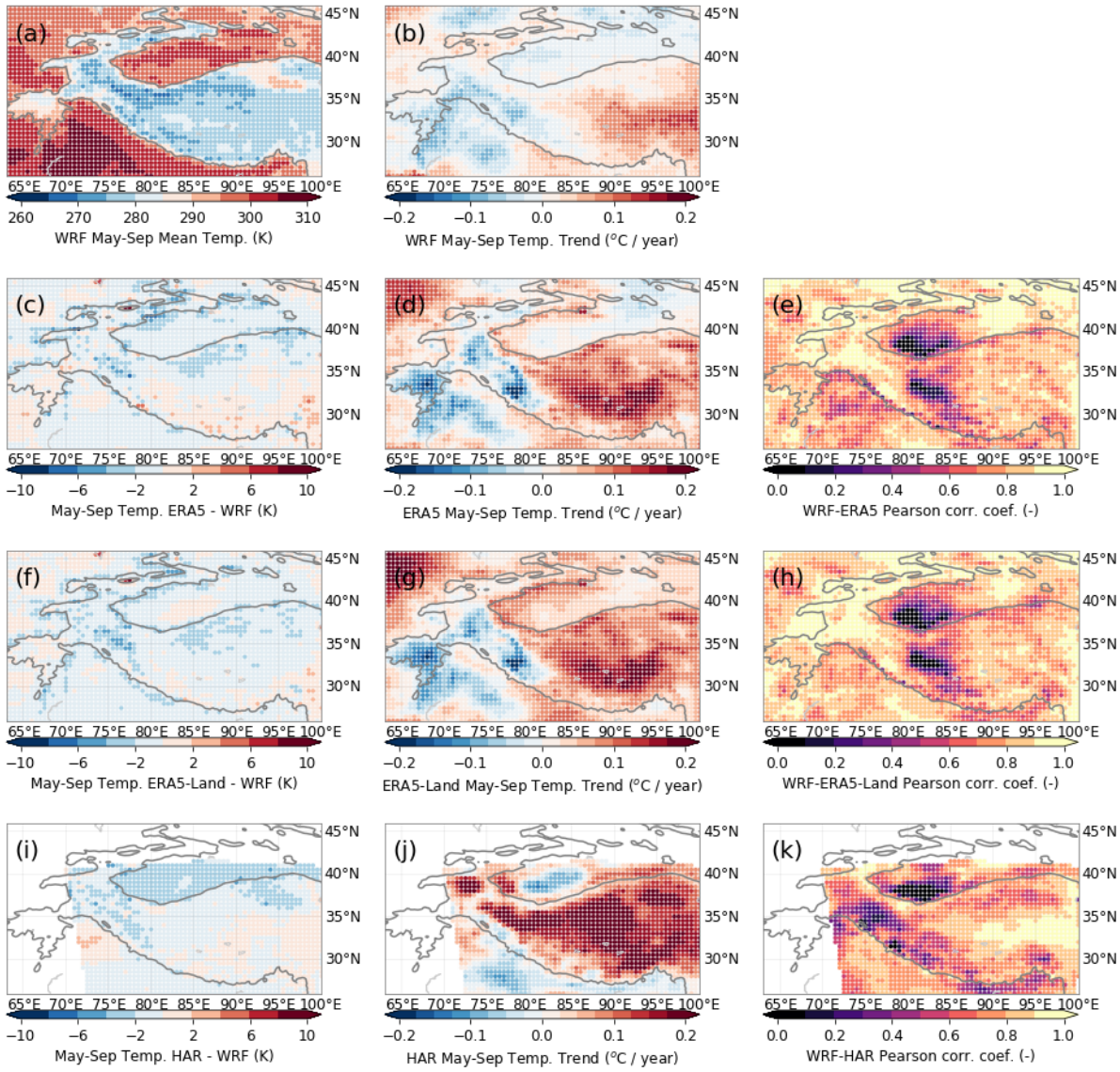

Figure 3: Comparison of WRF temperature output [a-b] with three other datasets (ERA5 [c-e], ERA5-Land [f-h], and HAR [i-k]). Columns show biases (c,f,i) with respect to the May-September mean temperature (a), May-September temperature trends (b,d,g,j), and Pearson correlation coefficients between the datasets and our WRF results (e,h,k). The 2000 m elevation contour is indicated by a solid line.

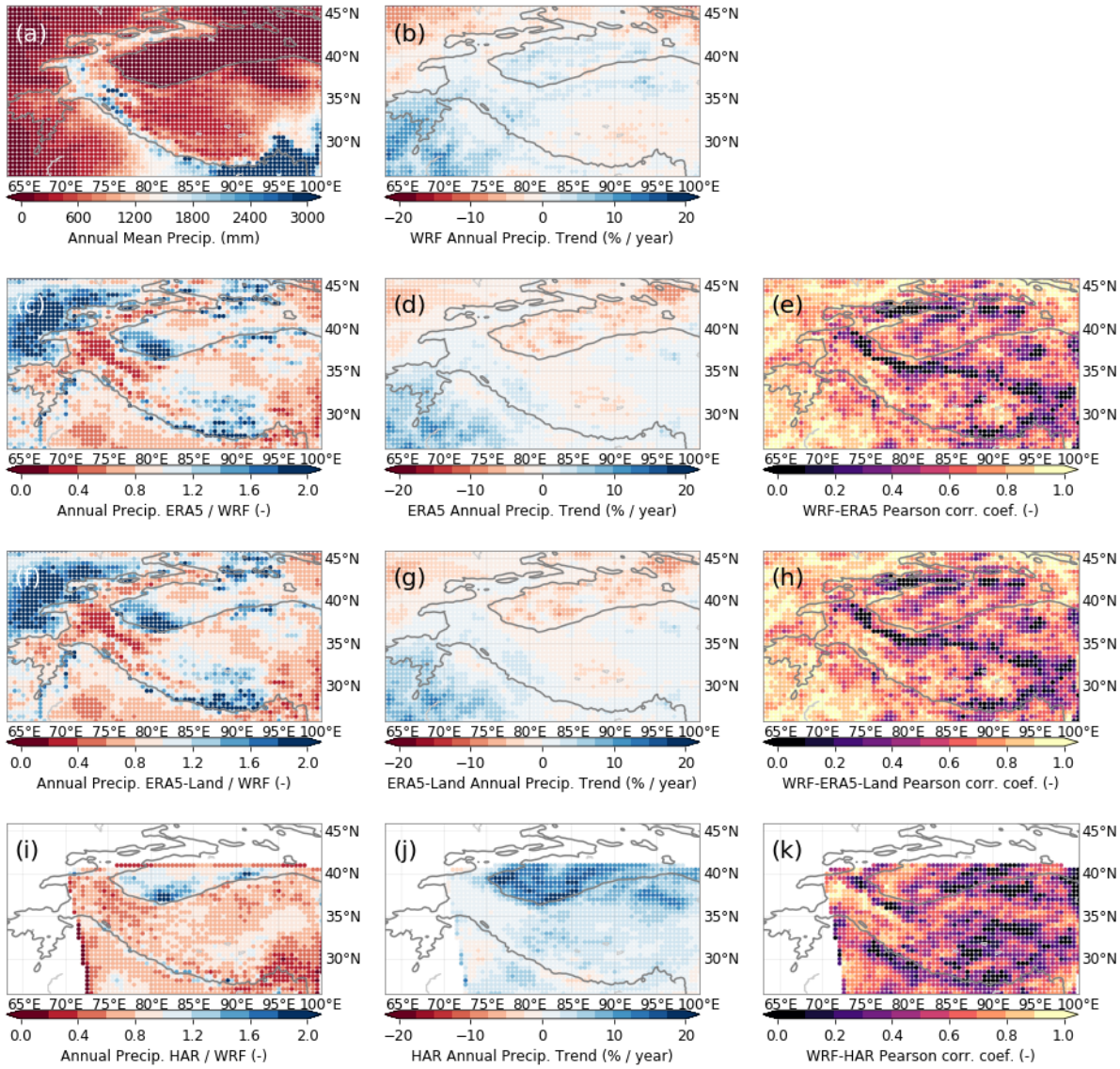

Figure 4: Comparison of WRF precipitation output [a-b] with three other datasets (ERA5 [c-e], ERA5-Land [f-h], and HAR [i-k]). Columns show precipitation multiplication factors (c,f,i) with respect to the annual mean precipitation (a), annual precipitation trends (b,d,g,j), and Pearson correlation coefficients between the datasets and our WRF results (e,h,k). The 2000 m elevation contour is indicated by a solid line.

In the discussion, we add: " **Our snowfall trends between 1980-2010 show some similarities, but also major differences with respect to a similar WRF study that did not include irrigation and used another re-analysis dataset (Norris et al., 2018). For instance, our temperature trends do not exhibit the strong summer cooling at low altitudes (e.g. the Tarim basin), and are more in line with station data (Waqas & Athar, 2018; Xu, Liu, Fu, & Chen, 2010) in that respect. However, contrasting precipitation trends in WKSK and southwestern HMA, similar to Fig. 5** *[now 6],* **are also present in ERA5 data and the Norris et al. study (see Farinotti et al. 2020). Although the interannual variability of temperature and precipitation is reasonably reproduced, and our precipitation trends are similar to those in other datasets, our model results are associated with uncertainties, which are partly irreconcilable due to a lack of *in situ* measurements in WKSK."**

Comment: The authors shall not only provide Pearson correlation coefficients but also further metrics like mean biases, r.m.s. deviations, regression slopes, etc., when comparing their WRF results with those from GHCN stations. I am not convinced that it is necessary to exclude so many GHCN stations by requesting at least 20 year of data coverage. This could be relaxed, or further comparisons may be added. I am also not convinced that it is sufficient to present results only for aggregated time periods, i.e., for annual mean air temperatures, May-September air temperatures, and July precipitation. Depending on the details of forcing the glacier model by WRF output, more detailed analyses of the WRF uncertainties are required, since snow- and ice melt can be rather variable from year to year, although years might have shown similar mean seasonal values for air temperature and precipitation.

In this respect, I would ask the authors to add more details on the WRF output and its application for forcing the glacier model simulations and the moisture tracking algorithm. In particular, I would like to know the output time step (one hour?).

Reply: It is true that the melt can be different per year. However, the glacier mass balance model does not include these subtleties. It requires a yearly input of temperature and snowfall and shifts the mass balance gradient accordingly to obtain an annual mass balance. In that sense, presentation of mean melt-season temperatures and annual mean precipitation is a reasonable representation of the data used in the glacier mass balance model. We already stated: "To modulate the mass balance gradient of the glacier over time, we applied annual precipitation changes derived from annual changes in WRF snowfall and temperature changes determined from annual changes in WRF melt season temperatures, i.e. when average daily temperature is above -5 °C. " We add a more detailed description of the glacier mass balance model as follows:

" To assess the response of the glaciers to the atmospheric forcing, we employ a glacier mass balance gradient model (Kraaijenbrink, Bierkens, Lutz, & Immerzeel, 2017). The model assumes a calibrated mass balance gradient along the glacier, and parameterises downslope mass flux in a lumped procedure that is based on vertical integration of Glen's flow law (Marshall et al., 2011). It also includes a parameterisation for the effects of supraglacial debris on surface mass balance (Kraaijenbrink et al., 2017), i.e. enhancing melt in the case of a shallow debris layer and limiting melt for thicker debris (Östrem, 1959). We modelled all individual glaciers in HMA larger than 0.4 km^2 (n=33,587) transiently for the period 1980-2010 (Kraaijenbrink et al., 2017). For ease of comparison with published observations, we select only those larger than 2 km^2 for the final analysis, which represent 95% of the glacier volume in HMA. Initial mass balance conditions in 1980 were set to be stable, while all other initial and reference conditions as described in the original study (Kraaijenbrink et al., 2017) were maintained. That is, using ERA-Interim data to locally calibrate the mass balance gradient of each glacier by constraining maximum ablation by a downscaled positive degree day climatology at the glacier terminus, and maximum accumulation by mean annual precipitation over the entire glacier area. The model simulates glacier mass change and evolution using a one-year time step, and hence requires representative annual input of temperature and precipitation. These are used to shift the mass balance curve according to sensitivity of the glacier's equilibrium line altitude to temperature changes, and adapt the maximum accumulation according to changes in precipitation (Kraaijenbrink et al., 2017)."

For the WRF output, we add: " **Results are output every 6 hours."**

Given, the annual input, we argue that it is then also reasonable to show the comparison with station data for the relevant aggregated temperature and precipitation data. Because of the mentioned difficulties with measuring snowfall accurately, we take the summer period. We now also took the period May-September, drop the 20 mm limit, and lower the number of available years to 15 to include more stations. Especially trends become very uncertain when few years are considered. The melt season used from the WRF output changes per location, but the summer months are likely to be most important. Hence, we compared these for the temperature data of the stations. Before describing the GHCN results, we add:

"**Since the glacier model requires annual input, representation of the interannual variability is especially important. Any constant biases are of less importance, since we use relative interannual variations as input for the glacier model. However, biases in temperature will have an effect on the snow-rain partition."**

Furthermore, we now add trends and biases into a new figure, which replaces Figures 2 and 3, and briefly discuss their results. We mention the median root-mean-square deviations in the text. We now write:

"We collected meteorological station data from **the Global Historical Climatology Network** (GHCN, Lawrimore et al., 2011, accessed June 2019), and selected those that have at least **15** years of full data between 1980-2010. **To be able to compare the WRF output with the station data, we apply a simple downscaling to the WRF temperatures in the grid that includes the station. We fit a linear temperature lapse rate to the temperatures and grid altitudes of a 2x2° box surrounding the station location. We then correct the WRF temperature by applying the lapse rate to the difference in altitude between the WRF grid and the station. Precipitation can also change significantly with location, but there is no clear relation between precipitation and altitude (Bonekamp et al., 2019; Collier and Immerzeel, 2015). For this simple comparison, we do not apply a downscaling of the WRF precipitation.**

Our WRF output produces May-September temperatures that are generally higher than the stations in the Tarim basin. However, biases are generally very low on the Tibetan Plateau, with values around 1°C. The median root-mean-square deviation between WRF and the stations is 1.8°C. The stations generally indicate a strong heating trend. Correlations between the annual variations in annual mean temperatures and mean temperatures between May-September are given in Fig. 2. They show generally very high correlations, with a lowest value of 0.5 (corresponding to p = 0.005). This implies that the interannual variability is very well reproduced in WRF. **This is despite the fact that many of these stations are situated in urban environments, with a potential heat island effect, a lack of evaporative cooling that is seen for irrigated agriculture, and a very difference surface energy balance than snow-covered areas. Hence, their locations might not be representative of the wider area, which might give rise to biases and trend differences when comparing the stations to the model outcome.**

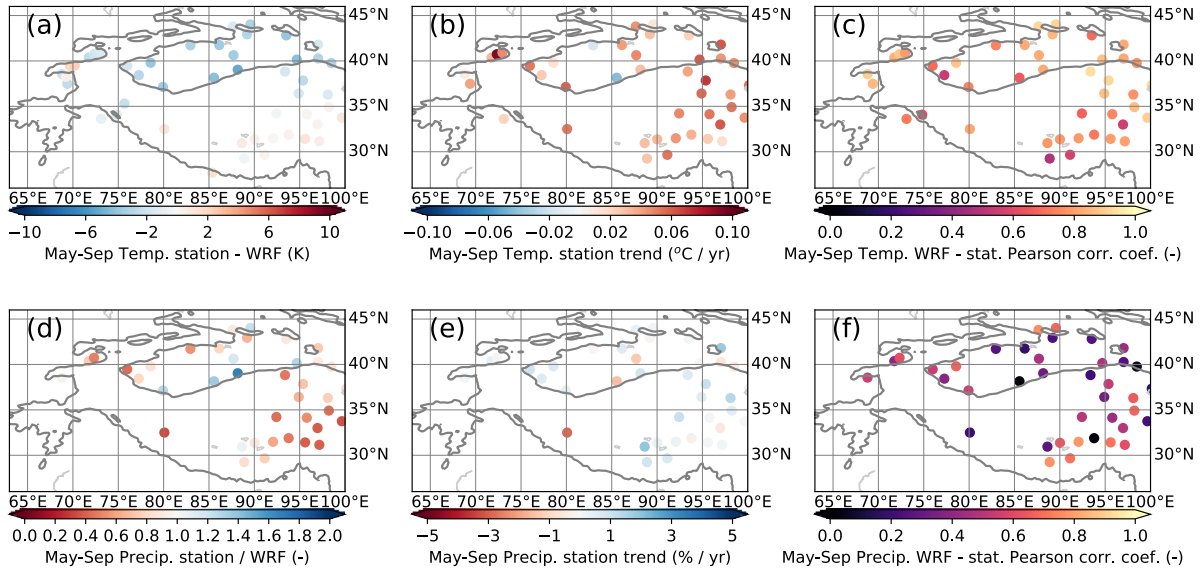

Figure 2: Comparisons between 1980-2010 time series of station data and nearest WRF grid for May-September temperatures (a-c) and May-September precipitation (d-f). Columns show temperature bias (a) and precipitation multiplication factor (d), trends (b,e) and Pearson correlation coefficients. The 2000 m-contour is indicated by a solid line

The stations in Fig. 2 closest to WKSK are almost exclusively in very arid regions, with a significant fraction of snowfall, **which is more difficult to reliably measure than rain (Archer, 1998),** making comparisons of precipitation very uncertain. Fig. 3 shows the **comparison** between time series of May-September precipitation, to limit the effect of snowfall. **Our WRF output is generally wetter than what is measured at the stations, except some locations in the Tarim basin. The median root-mean-square deviation between WRF and the stations is 11.4 mm per month. The stations show that most of the Tarim basin and Tibetan Plateau are seeing an increase in May-September precipitation.** The interannual variations are not represented by WRF as well as they are for temperature, but still show reasonable correlations for most stations, with values around 0.6. "

For the moisture tracking results, we selected the months that had the largest effect on the aggregated snowfall changes, as already stated in the text.