In the reply, the referee's comments are in *italics*, our response is in normal text, and quotes from the manuscript are in blue.

Anonymous Referee #1

General comments

This paper presents the first analysis of the Himalayan glacier response to solar geoengineering, analyzing the response to a number of scenarios of future greenhouse gas emissions and stratospheric aerosol geoengineering. The authors employ a mixed empirical / statistical model of glacier change driven by temperature change alone and use statistical down-scaling of the GeoMIP multi-model ensemble to produce the input data for their model. They find that solar geoengineering could slow the rate of retreat of Himalayan glaciers and that this benefit would be lost were solar geoengineering to be terminated.

This study is a novel contribution to the geoengineering literature and addresses an important issue; however, there are a number of shortcomings in the study that need to be addressed.

Most importantly, the method employed by the authors excludes precipitation which is the most critical climate difference between scenarios which include solar geoengineering and those which don't. Studies have reported substantial reductions in monsoon precipitation in scenarios which include solar geoengineering and so it seems odd to exclude this from the analysis. The authors note that excluding precipitation will likely mean that their estimate of the efficacy of solar geoengineering is an overestimate but it is not clear the magnitude of this overestimate; is it of order 1%, 10%, 100%?

I'd strongly recommend that the authors either reintroduce precipitation back into their methodology (the model described can use precipitation but it was excluded) even if this is only to provide a robust sensitivity estimate that justifies its exclusion, and I'd suggest doing this even if there are serious shortcomings in the down-scaling method for precipitation or other methodological concerns. Failing that, a much more detailed justification for excluding precipitation should be made along with some robust regional estimates of the significance that the changes in precipitation should be expected to have. The authors note that Rupper and Roe (2008) calculated zonal-mean sensitivities of SMB to temperature and precipitation; I suggest using these to estimate the sensitivity. This study should make clear what the research needs are to make a better estimate of the effects of solar geoengineering on Himalayan glaciers and the significance of including precipitation is a critical question that this paper does not address well.

Reply: In the revision, we include precipitation in our methodology and consider its impact on glacier change. We used zonal-mean sensitivities of SMB to temperature and precipitation calculated by Rupper and Roe (2008). We have done the simulations with precipitation using all models under all the scenarios.

We also add a whole new section on the uncertainties in the projections (section 5) and on the method validation section 3.3.

I had a number of questions about the validity of the methodology adopted in this study

that either require the authors to change the methodology, present some additional justification for their approach, or provide some sense of the sensitivity of their approach to these issues. I list the most significant ones here but there are other minor concerns listed afterwards.

First, the authors make comparisons between the responses of a changing ensemble of models for different experiments. It is unclear how much of the difference between these experiment-ensemble means is due to the changing make-up of the ensembles and how much due to robust differences in response. This is particularly concerning in the case of G4 were the models produced very different temperature responses due to their differing treatments of the 5Tg of injected SO2.

Reply: In the revision, we have done all the simulations using every GCM ensemble member and ensemble-mean under every climate scenario. We list the result by every model under all the scenarios in Table 4 in the revision. We also test the significance of the ensemble member differences.

Volume loss using the climate projected by HadGEM2-ES under G4 is far less than that by other models (Table 4), so we exclude it when calculating the G4 model mean.

We tested the differences between RCP8.5, RCP4.5 and G4 using the 4 models in common (Table 4), and find the glacier responses are significantly different ($p<0.05$). Although there are too few models in common between G3 and G4, the dominant influence of summer melting to the mass balance (Zhao et al., 2016), and the clear difference in temperature across HMA between G3 and G4 (Figs. 2,3) suggest the glacier response in HMA is different between G3 and G4.

Second, the authors don't do enough to validate their approach. A comparison against historical data or a more explicit reference to studies which validate this approach is needed. Does this model predict stable, growing or shrinking glacier mass for pre-industrial climate conditions? Is this model in agreement with other similar models?

Reply: Our model estimates were already discussed in comparison with estimates for RCP8.5 and 4.5 by Radić et al. (2014) and Marzeion et al. (2012) in the conclusion section of previous manuscript, and we also have them in Section 5.2 of the revision. This is the first simulation of geoengineering forcing of HMA glaciers.

In our method, we need to use glacier outline and glacier surface elevation to estimate glacier mass balance. Because we do not have such input data for the pre-industrial period, we cannot do simulations of glacier change for pre-industrial climate conditions.

In the revision, we added a new Section 3.3 which is about validation of the glacier model.

3.3 Validation of the glacier model and methodology

In this section we justify the selection of various parameter values used in the method here. In section 5 we indicate how elements in the model and climate forcing affect the uncertainties of the results we produce in section 4, and how those results compare with previous estimates of glacier evolution in HMA.

A crucial parameterization concerns the SMB-altitude gradients. The field data

(Table 1) include three more glaciers than those used in Zhao et al. (2014; 2016), and include a benchmark glacier from almost every sub-region. With so few glacier observations available, there is an issue of how representative they are of the general population. For inner Tibet, there are three glaciers (Zhadang, Gurenhekou and Xiao Dongkemadi Glacier) with SMB observations, and they have almost the same SMB-altitude gradients, 0.0041 m m^{-1}, over their common elevation range (5515~5750 m, Table 1); two glaciers (Naimona'nyi and Kangwure) in central Himalaya have SMB gradients of 0.0038 m m^{-1} in their common altitude range of 5700~6100 m. These similarities suggest that the measured glaciers share some important characteristics with the vast majority which are not surveyed.

Next we consider the choices for the initial value of ELA at the start year, different V-A scaling parameters and different ELA sensitivities to summer mean temperature and annual precipitation.

In choosing the initial ELAs for each glacier, there are several reasonable alternatives (Zhao et al., 2016): i) using ELAs interpolated from the first Chinese glacier inventory, ii) median elevations from RGI dataset, iii) the elevation of the 60th percentile of the cumulative area above the glacier terminus. These three choices lead to a range of about 2.5 mm of global sea level in glacier volume loss at 2050. In this study, we use median elevations from RGI dataset, which corresponds to the median result.

Zhao et al. (2014) showed that different volume-area scaling parameterizations can lead to ±5% range of glacier volume loss. The set of parameters we use in this study corresponds to the lower bound of estimated volume loss, but one that is best matched to the observational dataset of 230 separate glaciers (Moore et al., 2013).

For the ELA sensitivity to summer mean temperature and annual precipitation, we use the zonal mean values from energy-balance modelling of glaciers in HMA by Rupper and Roe (2008). Alternatively, it can be estimated using an empirical formula for ablation and a degree-day method (Zhao et al., 2016). Zhao et al. (2016) calculated the ELA for nine glaciers in China, India and Kyrgyzstan, and compared them with the observed ELA time series by similarities of decadal trends and also annual variability. The Rupper and Roe ELA parameterization produced the best fits to observed ELA decadal trends on 9 glaciers, with a correlation coefficient of 0.6 which is significant ($p<0.05$, the values we give for p are single tailed Pearson correlation tests).

Combining the above uncertainties would require a Monte Carlo simulation since the parameters combine non-linearly to produce glacier volume and area change; this is prohibitively expensive to perform given that a single simulation of all glaciers in HMA requires about 60 cpu hours on an 8 cores computer with parallel computing in Matlab. We did estimate elevation changes for individual glaciers directly from simulated volume and area changes, then calculated the average rate of elevation change for all the glaciers in each sub-region and compared them with remote-sensing estimates from 2003 to 2009 from Gardner and others (2013), Table 2. The correlation coefficient between the Gardner et al. (2013) estimates for the 6 RGI 5.0 sub-regions with data regional and our modeled regional averages is 0.7 which is marginally significant, ($p<0.1$).

In our simulations we have used constant lapse rates for temperature (0.65℃/100

m) and precipitation (3%/100 m). To check how reliable this is we chose 5 meteorological stations close to glaciers and calculated correlation coefficients for JJA temperature and annual precipitation at the station and at the nearest downscaled grid point from 1980 to 2013 (n=34). Precipitation correlations were higher than 0.85 for all the stations (p<0.001), while temperatures correlations were 0.47-0.85 (p<0.01).

Finally we explored the sensitivity to the choice of dataset used to correct model bias in temperatures. In addition to using historical temperature from the CRU dataset, we also did the simulation using temperature from Berkeley Earth project (0.5°× 0.5° resolution; Rohde et al., 2013; http://berkeleyearth.org/data/). That simulation was done using temperature alone as the glacier driver, so precipitation for each glacier was constant over time. The simulated climate ensemble mean forced volume losses in the period 2010-2069 were +4% (G3), -9% (G4), -11% (RCP4.5) and -13% (RCP8.5) different from the results using the CRU dataset.

Third, the authors need to fix parameters for the thousands of glaciers in their study area but only have 13 glaciers to base these parameters on. As they are making a regional analysis they use only the 3 nearest glaciers to fix parameters in a regional manner. It is not immediately clear how different these 13 glaciers are from the results presented and so it is not clear how sensitive the results are to this sampling. The authors don't discuss whether one could expect the sampled glaciers to be representative of their region or why one would expect robust regional differences between the glaciers' behavior.

Reply: Although the number of sampled glaciers is only 13 and the glaciers are randomly located, we find interestingly that the SMB gradients of the few glaciers in one sub-region are similar in their common elevation range. Again in Section 3.3:

For inner Tibet, there are 3 glaciers (Zhadang, Gurenhekou and Xiao Dongkemadi Glacier) with SMB observations, and they have almost the same SMB-altitude gradients, 0.0041 m m^{-1}, over their common elevation range (5515~5750 m, Table 1); two glaciers (Naimona'nyi and Kangwure) in central Himalaya have SMB gradients of 0.0038 m m^{-1} in their common altitude range of 5700~6100 m. These similarities suggest that the measured glaciers share some important characteristics with the vast majority which are not surveyed.

In the revision, we improve Table 1 to make this clearer.

Finally, the authors downscale temperature to a high-resolution grid however they don't make any assessment of whether this downscaled temperature data matches observations. I'd expect that there could be quite large differences between the real glacier altitude and the altitude of the nearest high-resolution grid point given the complex topography. I was also wondering given that the observed altitude of the glaciers is used to calculate the ELA, why not use it directly to perform this lapse-rate adjustment?

Reply: We guess the referee did not get what we did in the data-downscaling. In fact, we did the lapse-rate adjustment for temperature data just as the referee suggests. In the revision, we used temperature data from CRU and precipitation data from GPCC, which

have good reputation of the data quality. We have done many changes in Section 3.2 and 3.3 on downscaling that hopefully help clarify it:

The temperature and precipitation on each glacier were calculated by an altitude temperature lapse rate of 0.65℃/100 m, precipitation lapse rate of 3%/100 m, and the elevation difference of the glacier surface elevation relative to the nearest fine grid point.

In our simulations we have used constant lapse rates for temperature (0.65℃/100 m) and precipitation (3%/100m). To check how reliable this is we chose 5 meteorological stations close to glaciers and calculated correlation coefficients for JJA temperature and annual precipitation at the station and at the nearest downscaled grid point from 1980 to 2013 (n=34). Precipitation correlations were higher than 0.85 for all the stations (p<0.001), while temperatures correlations were 0.47-0.85 (p<0.01).

Relatedly, I'd recommend that the authors broaden the uncertainties section into a proper discussion that covers the shortcomings of the methodology employed and discusses the sensitivity of the findings to various assumptions. A key question which should motivate the material in this section is – how would the glacier response in a model with a full surface mass balance treatment driven directly by high-resolution climate data differ?

Reply: We do broaden the uncertainties section (see Section 5 in the revision) to discuss the uncertainties caused by climate forcing and by glacier model. In the revision, we use the relatively high resolution, monthly-mean gridded 0.5°×0.5° temperature data from the CRU TS 3.24 dataset instead of the 1°×1° temperature data from Berkeley Earth project, which was used before. We also discuss possible dynamic downscaling for the glacier model in section 5.1:

The models are also relatively coarsely gridded, certainly compared with the vast majority of glaciers, and so differences may be expected between statistically downscaled forcings based on lapse rates that we use here and that produced from high resolution dynamic climate model forcing.

we note that the distribution of meteorological stations in the study region is very sparse, especially in the northwest of this region (Liu and Chen, 2000). Therefore, both the CRU gridded data and data from models projections that we used in this study may have low accuracy for specific glacier regions. This has also implications for the use of very high resolution dynamic models; one such model simulated air temperatures and down-welling radiative fluxes well, but not wind speed and precipitation, producing unstable results when used with the CLM45 land model that simulated ground temperatures and snow cover (Luo et al., 2013). Explicit glacier atmospheric mass balance modelling (Mölg et al., 2013), a technique based on very high spatial and temporal resolution climate data (hourly and 60 m) was used on Zhadang glacier (Fig. 1, Table 1) with in-situ observations available, but not across the general expanse of the glaciated region; this study also noted the importance of wind speed to glacier mass balance in the region influenced by the Indian monsoon. Maussion et al., (2013) demonstrate that 10 km resolution dynamic modelling of the region can be done

successfully, and potentially can improve the precipitation modelling over the statistical downscaling methodology we employ here, though to date this is a reanalysis dataset with no prognostic simulations. Zhao et al. (2014 and 2016) used a 25 km resolution regional climate model RegCM3 to drive their simulations of glacier response to scenario A1B. By 2050 under A1B (which is intermediate between RCP4.5 and 8.5 in temperature rises), a sea level rise equivalent to 9.2 mm was projected from HMA. In comparison, our estimates are 11.1-12.5 mm for RC4.5 and 8.5 (Fig. 4).

I had a few concerns about the structure and the focus of the paper that relate to the regional climate analysis. First, it seems strange to separate the regional climate analysis from the regional glacier analysis. Why not start with the whole-region analysis of climate and glacier response and then conduct a regional climate and glacier analysis afterwards? Second, Figures 2 and 3 seem of very little use as there are no cues on these plots to relate the results presented to the regional glacier analysis in figure 7. Given glacier response is calculated across the region or over sub-regions, I'd recommend doing a sub-regional-mean analysis as in figure 7 instead of including figures 2 and 3. This would also help to make clearer how significant excluding precipitation from the analysis was.

Reply：Agreed.

Firstly, we change the structure of the result section (Section 4) as follows:

4 Result

4.1 Climate and glacier change across HMA

  4.1.1 Temperature and precipitation over HMA

  4.1.2 Glacier change across HMA

4.2 Sub-regional climate and glacier changes

  4.2.1 Sub-regional temperature and precipitation change

  4.2.2 Sub-regional glacier changes

Secondly, we changed Fig. 2 and 3 to sub-regional analysis of temperature and precipitation (Fig. 3 in the revision), which are similar and have better connection to Fig. 7 --- the sub-regional analysis of glacier volume change.

Finally, the conclusion needs to be revised as it presents a lot of new results and material, some of which are not appropriate for a conclusion.

Reply: We rewrote the whole conclusion section 6, and moved some material to the Uncertainty section 5.

Specific comments

L18-21 – unclear sentence: 5 under G4 what?

Reply: Changed to five models under G4 and six models under RCP4.5 and RCP8.5.

L23-25 – I don't recall this point being made in the paper.

Reply: We removed this point in the revision.

L30-31 – This is a far-reaching general statement extrapolated from a specific scenario.

It holds for the specific case but not for the general case.
Reply: We remove this sentence.

L55 – delete hence
Reply: Done.

L55-57 – These are very different types of reasons and I don't see how the second matters.
Reply: We rewrite this sentence as "the glaciers are affected by both the South Asian monsoon system and the westerly cyclonic systems, depending on specific location across the region, thus the region integrates the climate response to two important global circulation systems (Mölg et al., 2013)."

L57-60 – Irvine et al. 2012 is a semi-empirical study and McCusker et al. discussed ice-sheet implications of their climate model results but did not simulate ice-sheet or sea-level response at all.
Reply: We move Irvine et al. (2012) to the correct place. We change this sentence to "glacier responses to geoengineering scenarios has been limited to studies on global responses based on semi-empirical models (Moore et al., 2010；Irvine et al. 2012) or from simplified ice sheet responses (Irvine et al. 2009; Applegate et al., 2015) or implications of climate model (McCusker et al., 2015), with nothing to date on mountain glacier impacts."

L75 – "regions of newly defined regions of " – fix
Reply: modified to "defined regions of".

L86-89 – This sentence is unclear. Why is this simple? What issues could this pose?
Reply: All the glacier outlines data are takes from RGI 5.0 dataset. The Second Chinese Glacier Inventory and the "Glacier Area Mapping for Discharge from the Asian Mountains" (GAMDAM) inventory provide data in certain regions in RGI 5.0. As the referee asked below, we improve this paragraph to make this meaning clear.
Notice the date of the data are different in different regions. And the date range for each region is short – a few years. So we take the same date for the glaciers in one region from the same data source. We do not think it would bring big issue. We add "Because the data range from one data source is only a few years" here in the revision.
The Randolph Glacier Inventory (RGI) database contains outlines of almost all glaciers and ice caps outside the two ice sheets (Arendt et al., 2015). Our study region covers HMA (26–46° N, 65–105° E), which corresponds to the defined regions of Central Asia, South Asia West and South Asia East in the RGI 5.0. According to the RGI 5.0, the study region contains a total of 94,000 glaciers and a glaciered area of about 110,000 km2. The RGI 5.0 data inside China are based on the Second Chinese Glacier Inventory (Guo et al., 2015), which provides glacier outlines from 2006–2010, except for some older outlines from the First Chinese Glacier Inventory where suitable imagery could not be found - mainly in southern and eastern Tibet (the S and E Tibet RGI 5.0 sub-

region), most of which were made in the 1970s. The RGI 5.0 data outside China are from the "Glacier Area Mapping for Discharge from the Asian Mountains" (GAMDAM) inventory (Nuimura et al., 2015) and nearly all come from 1999–2003 with images selected as close to the year 2000 as possible. Because the data range from each data source is only a few years, we take three reference years: 1980, 2009, and 2000, as start dates for our model simulations of glaciers in S and E Tibet, elsewhere in China, and outside China, respectively.

L90-95 – This approach needs some justification and citations – why is it appropriate? Is this a widely accepted approach? Etc.
Reply: Using median altitude as a proxy of ELA is a widely accepted approach (e.g. Nuimura et al., 2015).
We add this in the revision:
Following previous authors (Nuimura, 2015; Zhao et al., 2016), we use median altitude from RGI 5.0 for each glacier as a proxy for equilibrium line altitude (ELA).
The sensitivity is discussed in section 3.3
In choosing the initial ELAs for each glacier, there are several reasonable alternatives (Zhao et al., 2016): i) using ELAs interpolated from the first Chinese glacier inventory, ii) median elevations from RGI dataset, iii) the elevation of the 60th percentile of the cumulative area above the glacier terminus. These three choices lead to a range of about 2.5 mm of global sea level in glacier volume loss at 2050. In this study, we use median elevations from RGI dataset, which corresponds to the median result.

L98-99 – What precisely is done? I believe this is explained in the next section, if so make that link explicit.
Reply: The parameterizations of mass balance with altitude relative to the ELA is precisely written in Zhao et al. (2014). We add this reference here.

Fig 1 + other lat-lon plots – It would help to make sense of the lat-lon plots if there was some frame of reference beyond the lat-lon coordinates. Would it be possible to mark the numbered regions (referred to later) with boxes or outlines in this figure and then over lay this outline in the other plots. This would make these subsequent figures much more useful. Or alter natively lose the lat-lon plots entirely and focus on the area-average means for the different regions shown here and assessed in figure 7.
Reply: We focus on the area-average means for the different regions. As the referee suggested here and afterward, we change Fig. 2 and 3 (the spatial maps of temperature and precipitation) to sub-regional analysis of temperature and precipitation.

L115-145 – Is there any validation of this modeling approach? If not this is a major shortcoming. 2 previous papers are cited; if they contain a validation of this approach that should be brought up and explained.
Reply: Section 3.3 Validation of the glacier model. And Section 5.2 on uncertainties discuss these issues in detail and at length.

L116-130 – Some sense of the sensitivity of the method to this sampling of glaciers is needed. I did not find the values in the table easy to read so I couldn't tell from looking at this how large the differences were; perhaps a figure would help to illustrate this. Can you be sure that these differences are systematic regional differences rather than the quirks of individual glaciers? Why not take the average over all glaciers rather than just the 3 nearest? These questions ought to be answered here or else citations which address these questions cited. This is another source of uncertainty in the model projections but it is unclear how large it is.

Reply: We do not quite understand all the points in this question. Hopefully the changes in Section 3 make the method clearer, along with the cited reference to Zhao et al 2014. We tried to clarify our method, in section 3.3, in relation to mass balance gradients we show. The point on sensitivity to SMB, Zhao et al 2016, showed that it is important, but where the data allow the gradients to be checked within a sub-region, they agree very well between glaciers:

With so few glacier observations available, there is an issue of how representative they are of the general population. For inner Tibet, there are 3 glaciers (Zhadang, Gurenhekou and Xiao Dongkemadi Glacier) with SMB observations, and they have almost the same SMB-altitude gradients, 0.0041 m m-1, over their common elevation range (5515~5750 m, Table 1); two glaciers (Naimona'nyi and Kangwure) in central Himalaya have SMB gradients of 0.0038 m m-1 in their common altitude range of 5700~6100 m. These similarities suggest that the measured glaciers share some important characteristics with the vast majority which are not surveyed.

L140-144 – The justification for excluding precipitation here is incomplete (as it is in section 4). What matters is not the fractional precipitation change but rather the ratio of a x dT to b x dP. If you have the beta values it should be straightforward to make a rough calculation of b x dP so that a comparison can be made. This issue is critical to address properly as your model cannot distinguish GHG from solar forcing (as it uses dT alone) and solar geoengineering has a distinct effect on precipitation.

Reply: Agreed.

In the revision, we include precipitation.

L146-189 – Is there a validation for this approach, i.e. do the down-scaled, lapse-rate adjusted reanalysis data match observational data?

Reply: Yes. To validate the downscaled, lapse-rate adjusted reanalysis data, we choose 5 climate stations close to glaciers and calculated correlation coefficients for the station JJA temperature and annual precipitation and the nearest downscaled gridpoint from 1980 to 2013 (n=34). Precipitation correlation was higher than 0.85 for all the stations (p<0.001), while temperatures correlations were 0.47-0.85 (p<0.01).

In the revision, we now use the relatively high resolution, monthly-mean gridded 0.5° ×0.5° temperature data from CRU TS 3.24 dataset (Harris et al., 2014), instead of 1° ×1° temperature data from Berkeley Earth Project, because CRU has better reputation

of data quality control. We use 0.5°×0.5° monthly total gridded precipitation data from the Global Precipitation Climatology Centre (GPCC) Total Full V7 dataset (Becker et al., 2013).

L148 – "the beginning years" not clear
Reply: We run the simulations for glacier change from the relevant start years (Section 2) to the year 2089.

L150-151 – This sounds like you don't believe them, explain the limits of this approach or rewrite this sentence.
Reply: In the revision, we use temperature data from CRU instead of Berkeley Earth Project. So we rewrite this sentence as "From the start years to 2013, we use the relatively high resolution, monthly-mean gridded 0.5°×0.5° temperature data from the CRU TS 3.24 dataset (Harris et al., 2014)".

L159 – "and THE INCREASE IN greenhouse gas forcing" rather than the total forcing relative to pre-industrial.
Reply: Done.

L159-161 – Here and elsewhere need to make it clear that different models had a very different temperature, etc. response to G4's 5Tg.
Reply: OK. We add this in the revision "The across model spread of temperatures under G4 is larger than under e.g. RCP4.5, (there are too few ensemble member models under G3 to see this) because of differences in how the aerosol forcing is handled, and each model has a different temperature response to the combined long and shortwave forcing (Yu et al., 2015)."

L162 – No. Injection stops which means the aerosol forcing will decay rather than instantly disappear.
Reply: OK. We delete "driven by forcing from RCP4.5 alone".

L171 – be more specific, e.g. resolutions of ~ 2 degrees.
Reply: We show the resolution in Table 2. We add "(Table 2)" here.

L176 – Does the glacier not have an altitude of its own recorded somewhere? How large could the difference between the nearest grid-box and the glacier's actual altitude be? I suspect this could be quite large. Has this been dealt with by other papers /methods? Is there any way to get a sense of this?
Reply: We rewrite this sentence to make it clear. The individual glacier has its own altitude. There is difference between the nearest grid-box and the glacier's actual altitude. The temperature and precipitation on each glacier were calculated by an altitude temperature lapse rate of 0.65℃/100 m, precipitation lapse rate of 3%/100m, and the elevation difference of the glacier surface elevation relative to the nearest fine gridpoint.

L191-222 – Why is a regional analysis here? The regional analysis of the glacier response is made much later, it would seem sensible to present this material alongside that.

Reply: We changed the structure of the paper in the revision.

L191-222 – Some warning should be given about reading too much into the small scale "features" in these plots. These are artifacts of the ensemble average of mismatched, coarse model grids that have been downscaled – no individual model has the kind of spatial variation seen in figure 3a for example

Reply: OK. But as the referee suggested, we changed Fig. 3 to sub-regional analysis.

L193 – estimate = calculate.

Reply: We change "estimate" to "calculate".

Figure 2 and throughout – It is not clear whether the differences plotted in this figure and in all other figures are due solely to differences in the response to the experiments as is implicitly assumed or whether the different make-ups of the ensembles used in each experiment has affected the results substantially. Some measure of this effect should be made.

Reply: As the referee suggested, we changed Fig. 2 to sub-regional analysis (Fig. 3 in the revision).

There are 4 models in common under RCP8.5, RCP4.5 and G4. We tested the differences between RCP8.5, RCP4.5 and G4 using the 4 models in common (Table 4), and find the glacier responses are significantly different ($p<0.05$). There are only 2 common ensemble members (BNU-ESM, HadGEM2-ES) used in both G3 and G4, but we exclude HadGEM2-ES because this model gives a very different result from other models under G4. So there is only BNU-ESM in common under G3 and G4.

Figure 2 and other map plots – The continuous color scale is hard to read, is it possible to use a banded scale or to add contour lines?

Reply: As the referee suggested, we changed Fig. 2 from a map to a sub-regional line plot (Fig. 3 in the revision).

Figure 2 and other map plots – I'd suggest removing the stippling. First, I'm skeptical whether the measure plotted is useful, i.e. does it really give a good sense of the robustness of the regional results? Second, looking at these plots the stippling is applied at a scale much finer than the resolution of the models that generated the results. Third, it's entirely absent from some figures suggesting that the stippling code was not implemented appropriately.

Reply: As the referee suggested, we changed Fig. 2 and Fig. 3 to sub-regional line plots (Fig. 3 in the revision).

Figure 2 and 3 – Are these annual-mean or JJA-mean results?

Reply: They are JJA-mean temperature in Figure 2 and annual precipitation in Figure 3.

L197 and throughout – Are mean +/- standard deviation (which I presume these are) measures appropriate when the ensemble consists of only 4 models? Perhaps mean (min, max) may be more appropriate and informative.
Reply: For simplicity we prefer to stick with standard deviations for the results, and it also means that we can remove outliers such as the G4 result of HadGEM2-ES. We add a new table (Table 4) in the revision to show all model results, and their mean +/- standard deviation under all the scenarios.

Table 3 – I'd recommend adding much more information here and converting this into a figure which covers the full range of the model responses for each experiment across the different glacier regions. This would help to address my concern about the absence of precipitation from the simulations and the poor justification for why this was left out. Reporting the changes in precipitation in percentage change may help or else this could be combined with an estimate of the beta parameter for glacier sensitivity to precipitation (perhaps expressed as a fraction of the temperature response) to give a quantified estimate of how significant the precipitation response is. Without such information and given that precipitation is excluded from the glacier model I don't see why devoting a page to the precipitation response was necessary.
Reply: In the revision, we include precipitation in the model.

Figure 3 – These precipitation results seem to have been down-scaled, how was this done?
Reply: we downscale both the CRU gridded temperature data, the GPCC gridded precipitation data and the climate model output to a grid based on a land surface topography having resolution of $0.1126^{\mathrm{o}} \times 0.1126^{\mathrm{o}}$ using an altitude temperature lapse rate of $0.65 ℃/100$ m, an altitude precipitation lapse rate of 3%/100m, and elevation difference of the fine grid relative to the climate model grid.

L218 – "Precipitation change ratios"? not clear what this means.
Reply: We no longer use this approach in the manuscript.

L225 – "projections ensembles"?
Reply: We change it to "projections".

L233-241 – How does these results compare to previously published results? Perhaps cite the work that made the initial analyses of G3 and G4.
Reply: We cite a paper Yu et al. (2015) here, and change this sentence to "There are relative coolings of 1.05^{o}C under G3 and 0.76^{o}C under G4 compared with RCP4.5 during 2020-2069 across the whole region (Fig. 3). Yu et al. (2015) noted that G3 produced a relative cooling of 0.58^{o}C and G4 of 0.53^{o}C in globally averaged temperature over the 2030-2069 period."

L233 – Should "and the highest rate" read "at the highest rate"?
Reply: Yes. We change it.

L238 – 1.7C over what period?
Reply: The temperature rise is over the period 2070-2089 relative to the period 2050-2069. We add the period in the revision.

L239-241 – Why does this happen? Here and elsewhere more care should be taken to make clear that G4 produced very different forcing and hence temperature responses in the different models. Some evaluation of how effective G4 was at cooling the climate in different models should be made or reference should be made to a study which makes this analysis.
Reply: We add the reason in the revision. This is due to G4 having a constant stratospheric aerosol injection rate of 5 Tg SO2 per year, while G3 gradual ramps-up the aerosol so that about twice as much is needed by 2069, depending upon the sensitivity of the particular model to stratospheric sulphate aerosols. Hence, the radiative impact of terminating G3 is about twice as large as terminating G4, and the termination temperature signal is much more obvious in G3 than G4.

We made it clear that "The across model spread of temperatures under G4 is larger than under e.g. RCP4.5, (there are too few ensemble member models under G3 to see this) because of differences in how the aerosol forcing is handled, and each model has a different temperature response to the combined long and shortwave forcing (Yu et al., 2015)."

In Yu et al. (2015), they found "Over the period from 2030 to 2069, the global average SAT under rcp45 increased by 0.81 ± 0.21 °C compared with the baseline (average over 2010–2029 under rcp45); while G4, the 40 year annual global mean SAT increased by 0.28 ± 0.31 °C". So the cooling effect is about 0.54°C. Therefore, we add this paragraph in the revision

"There are relative coolings of 1.05^{o} C under G3 and 0.76^{o} C under G4 compared with RCP4.5 during 2020-2069 across the whole region (Fig. 3). Yu et al. (2015) noted that G3 produced a relative cooling of 0.58^{o} C and G4 of 0.53^{o} C in globally averaged temperature over the 2030-2069 period."

Figure 4 – There is a suspicious degree of agreement between the models in panel d. These models presumably have quite different Transient climate response (TCR)s, surely there would be greater differences than this at the regional scale. Please cite some other work or present some evidence that this is not a processing error on your behalf.
Reply: The plots in Fig. 4 (which is Fig.2 in the revision) show the average temperatures after grid-point by grid-point bias corrections. Ranges found are slightly smaller than the regional spread found by Yu et al. (2015) due to grid-point by grid-point bias correction we apply here. Under RCP8.5, temperature rises in HMA about 5 °C from the year 2020 to 2089 (panel d), which is higher than the global temperature rise of 3°C

over the same period (IPCC AR5). We do not think we made any error.

Figure 5 – Here and elsewhere the comparison is not simply between different experiments but between different ensembles of models. Some assessment of the significance of this difference ought to be made. A similar line plot like this would be a good place to do so.
Reply: We add the spread between ensemble members in this plot.

Figure 5 – Rather than going into the spatial pattern of the precip response in figure 3 (which is not used) why not add 2 panels on the precip response to this line plot? Alter natively, do the regional-average analysis I suggested earlier.
Reply: We add 2 panels on the precipitation response in the revision.

L249 – "averaged over the grids" what does this mean?
Reply: it means the averaged over the downscaled grid.

L263-264 – Explain why this happens. I presume it's because the most sensitive glaciers have already retreated / disappeared but this isn't mentioned anywhere. I'm wondering idly whether an overall sensitivity could be calculated.
Reply: Yes: The RCP4.5 and RCP8.5 scenarios produce similar continuous mass loss until approximately 2035 (Fig. 4a) mainly due to the similarity of temperatures projected by RCP4.5 and RCP8.5 in the period 2020-2035 (Fig. 2), and both show relatively slower loss rates after about the year 2050 probably because the most sensitive glaciers have already retreated before 2050.

L269 – explain why G3 > G4.
Reply: G3 reduces glacier loss more than G4, which is due to stronger temperature cooling effect under G3 (section 4.1.1). We add this in the revision.

L270-280 – There is too much reporting of values and not enough explanation in this paragraph.
Reply: We add Table 4 to show the result of all the models, and remove some values here.

L283-284 – Is this area-volume relation surprising? Aren't they explicitly linked in the model?
Reply: It is not surprising. We add "as may be expected" and delete "quite".

Figure 7 – This regional break-down is great, however it is a shame that none of the previous results are presented on the same basis. This regional (and subregional) analysis is much more useful than the spatial maps shown elsewhere, could temperature and precipitation responses be plotted for these regions?
Reply: In the revision, we add the temperature and precipitation responses in each subregion.

L288-301 – Why isn't the regional climate analysis here with the regional glacier analysis.
Reply: We have changed the structure of the result section.

L303-359 – It is unclear what this section is for. I'd suggest reframing this as a regular discussion section and broadening its scope to cover all the shortcomings of this study.
Reply: we did so in the revision.

L304-306 – This sentence is unclear.
Reply: We improve this sentence as "Firstly, only 3 ESMs participated in G3 but 5 in G4 simply because doing the G3 experiment is difficult and time-consuming to set-up."

L315-320 – why is this included? Your simulations excluded precipitation effects.
Reply: We include precipitation in the revision.

L328-333 – This paragraph is unclear, it is hard to follow the references.
Reply: We rephrase this, and split the paragraph into 4 different ones in Section 3.3. The relevant 4 lines are now: In choosing the initial ELAs for each glacier, there are several reasonable alternatives (Zhao et al., 2016): i) using ELAs interpolated from the first Chinese glacier inventory, ii) median elevations from RGI dataset, iii) the elevation of the 60th percentile of the cumulative area above the glacier terminus. These three choices lead to a range of about 2.5 mm of global sea level in glacier volume loss at 2050. In this study, we use median elevations from RGI dataset, which corresponds to the median result.

L334-338 – This is also unclear.
Reply: We delete this since we include precipitation in the revision.

L339-342 – Does it? You have not shown the relative significance of precipitation to temperature in this study and table 3 only reports the area-average results which differ from the regional responses.
Reply: We remove this sentence since we include precipitation in the revision.

L339 – if A1B shows a significant trend then surely RCP 8.5 would too given their similarity?
Reply: We remove this sentence since we include precipitation in the revision.

L344-346 – This sentence is unclear and it's not clear which observational data is referred to here.
Reply: changed to : The set of parameters we use in this study corresponds to the lower bound of estimated volume loss, but one that is best matched to the observational dataset of 230 separate glaciers (Moore et al., 2013).

L350-352 – I'm not sure how useful this measure is given the shortcomings in the approach and the systematic over-estimation due to the exclusion of precipitation. Also, extending the simulations out to 2150 could give rise to the problem that your simple approach to uncertainty bounds would include negative glacier mass.
Reply: we removed this measure in the revision.

L361-end – Much of the material presented in this conclusion is new, why does this not appear earlier?
Reply: we moved them to the uncertainty section (Section 5) before conclusion.

L372 – why is G3 extreme? There are more extreme possibilities. This is an arbitrary scenario.
Reply: we remove "extreme".

L383-385 – This is flatly wrong as written here. This result is scenario dependent. If a greater cooling were exerted by solar geoengineering more glaciers would be saved.
Reply: Perhaps we were unclear, we changed the final paragraph to:

Although G3 keeps the average temperature from increasing in the geoengineering period, G3 only slows glacier shrinkage by about 50% relative to losses from RCP8.5. Approximately 72% of glaciated area remains at 2069 under G3 compared with about 30% for RCP8.5. The reason for the G3 losses is likely to be that the glaciers in HMA are not in equilibrium with present day climate, so simply stabilizing temperatures at early 21st century levels does not preserve them. To do that would require significant cooling, perhaps back to early 20th century levels. Achieving that coling by sulphate aerosol injection may not be possible. The 5 Tg of SO2 per year specified in G4 is about the same loading as a 1991 Mount Pinatubo volcanic eruption every 4 yr (Bluth et al., 1992). G3 requires increasing rates of injection, to 9.8 Tg for the BNU-ESM at 2069. As aerosol loading increases, its efficacy decreases as particles coalesce and fall out of the stratosphere faster, while also becoming radiatively less effective (Niemeier and Timmreck, 2015). This effect is so strong that it appears unfeasible to use sulphate aerosols to completely eliminate warming from scenarios such as RCP8.5. Greenhouse gas emissions would require very drastic reduction from present levels, and net negative emissions within the next few decades, to limit global temperature rises to 1.5 or 2°C (Rogelj, et al., 2015). If such targets were met, then it is conceivable that plausible quantities of sulphate aerosol geoengineering may be able to maintain 2020 temperatures throughout the 21st century. Even if this politically very difficult combination of drastic emission cuts and quite aggressive sulphate aerosol geoengineering were done, then our simulations suggest the disappearance of about 1/3 of the glaciated area in HMA by 2069 still cannot be avoided.