Response to reviewers

We thank the three reviewers for their thoughtful and detailed comments, and acknowledge the consensus that the manuscript requires significant changes before publication. A comprehensive revision of the manuscript is underway that responds to the criticism provided, and we demonstrate considerable progress through the process in this response, including production of drafts of all of the additional figures we intend to use. The responses to all 3 reviewers begin with the same general overview and end with the set of new and revised figures which have been produced, but contain a set of specific responses to each review's points in order after the introduction.

The three key threads to the criticism in the reviews, in our eyes, are as follows: 1) the description of both the internal processes of OGGM and the way it is used in our study are ambiguous or insufficient, particularly in relation to the use of GCM data, 2) the scope of the study is not well realised, with conclusions not properly related to the stated aims, and questions raised left unanswered, and 3) the analysis of the results provided is not comprehensive enough, with insufficient quantitative measures of model performance; this also contributes to the insufficient conclusions.

What is significant is that there is no major criticism which requires additional modelling to take place; we have all the data we require, and it is produced in a clear and rigorous way, but the failure of the manuscript lies in inadequately communicating exactly what we have done and what we can determine from the results. With the review comments in mind, a rewrite of the manuscript is underway and could be completed in the classical deadline allowed by the journal for revisions. It includes a much more precise and in-depth description of OGGM and its requisite data inputs, and a considerably expanded set of figures and quantitative measures of model performance aiding a refocused narrative that we believe does a much better job of answering the interesting questions raised in the introduction.

Below we describe the proposed changes and additions to both the text of the manuscript and the data visualisations, and relate them to the specific criticisms they are intended to address. Those changes which have already been made are labelled [I] after the description of the change. Various small corrections that do not warrant special discussion (terminology, sentence structure, etc.) have also already been made, but are not mentioned for the sake of brevity, while other suggested minor text fixes will be made after the more substantial parts of the rewrite are all complete. Drafts of any new or updated figures referenced under 'changes already made' are included at the end of the document.

Review 1 itemised response

- **Climatic forcing and calibration**
  The authors explain that they use 6 general circulation models (GCM) to drive OGGM with temperature and precipitation. I suppose that they used monthly fields of surface air temperature and precipitation. Yet no details are given. Is the OGGM-inherent interpolation scheme used for these variables. I think that it is worth to specify the assumed lapse rates. However, my primary concern is the casual description of the climatic forcing. In lack of details, I assume that, after the model calibration, the GCM forcing is directly given to OGGM. Yet the GCM performance will differ around the globe making them more or less suitable for explaining length changes in various regions. In my view, the general practice is to define a common reference period with a climatologically meaningful length and apply the climatic forcing in anomaly mode. In this way, GCM biases in the recent period can be accommodated. This aspect is

even more relevant in terms of the OGGM calibration. As the authors only refer to the original OGGM publication (Maussion et al., 2019), I assume that the temperature sensitivity parameter ($\mu*$) is automatically calibrated based on the interpolated CRU data set. Correct me if I am wrong. Please specify the calibration time period in OGGM. The automated OGGM calibration implies that when you change to the GCM forcing, the model is not expected to perform well in the recent past, reverting the benefit from the CRU calibration. Changing to anomaly modes will help. The other option is to calibrate $\mu*$ to each GCM. As it stands now, I fear that it is almost impossible to interpret the results.

This is an issue of weakness in our description of the GCM data, OGGM's usage of climate variables, and the way OGGM performs calibrations. Both OGGM's behaviour and our usage of climate variables are well-defined and consistent, but not communicated well in the reviewed draft of the paper, so we address this issue by providing considerable additional description in the methods section.

We clarify the details of the scaling of climate model data to match 1900-2000 CRU data means, which is the default behaviour of OGGM [I], and describe in detail the nature of the data from the GCMs which is used [I]. We considerably enhance the description of OGGM processes, particularly focused on the way that climate variables are used in the surface mass balance calculation [I]. An explicit description of the calibration of mass balance sensitivity is also provided [I].

- **Representativeness**
  Please specify how representative the glacier sample that has length observations is for each RGI region. An idea could be to present numbers of hypsometric and glacier-area distribution of the glacier sample as compared to the entire region

  A new figure (P1) shows the distribution of Leclercq glacier lengths relative to RGI glacier lengths [I]. New figure P2 shows how well represented regions are, compared to each other and through time [I].

- **Regional length changes**
  You compute the regional relative glacier length changes as the mean of normalised length variations of all glaciers with length observations per RGI region. First, this normalisation overrates the importance of small glaciers. You can see this for Alaska in Fig.1: at several points in time, regional relative length changes exceed 0.5 in less than 10 years which should not reflect the response of the large glaciers. Second, the formation of a mean value is highly susceptible to outliers. Outliers in terms of normalised length changes are expected from the small glaciers in the region because large glacier systems will only show moderate relative length fluctuations. Is it possible to use a more robust measure. I do not think that the median will help, let alone that it will be more informative. Yet a weighting by glacier area might help.

  The new figure P1 [I] addresses concerns associated with the representativeness of the dataset by making the context of the observations in relation to the true distribution of glacier sizes clear. It is accompanied by a comment in the 'glacier data' portion of the text about the impact of glaciers of different sizes on regional mean length changes [I]. This

incorporates a defence of regional means not weighted by glacier length or area because the impact of small glaciers should be significant when we consider the real size distribution of glaciers [I].

In addition, you compute the regional mean on an annual basis only considering glaciers with measurements in that year. In many regions you see abrupt step changes in this regional value, which are likely a relict of this strategy. It is therefore very difficult to interpret the regional length change record because they can either arise from the model response or the changing number of considered glaciers from year to year. It would be highly informative to include a plot with the regional length records in addition to the sample size variations through time. An idea for removing the 'spike' behaviour is to assume a linear length change between sparse length observations. In other words, would it make sense to linearly interpolate the observed length record to a yearly timeline for glaciers with irregular sampling of length information. Admittedly, my suggestion is not ideal but it removes the sample-size dependence from the regional length values.

The suggested behaviour for linearly interpolating Leclercq data points was always used, and this is now made explicit in the text [I], and the suggested addition of a plot of the number of available Leclercq records is included in the form of new figure P2 [I]. The revised figure 2 [I] also touches on the 'spiky' behaviour that can come with the variable number of members in the Leclercq dataset by year, which is discussed in the added text related to this figure.

- **Glacier complexes**
  In the past, many of the nowadays separated glacier units in the RGI were part of large complexes comprising several glacier branches. As OGGM treats each RGI unit independently, larger glacier complexes in the past are not allowed to form. How important is this fact for the glaciers you are focussing on. The influence of tributaries might well have been important for past length changes even during the period with length observations. Is it possible to consider this effect in OGGM? If not please discuss

  This problem is not one which can be adequately tackled by OGGM, and is fundamentally an issue of any glacier modelling process which operates on a per-glacier basis. Ice masses being dynamically connected at certain times and not at others cannot be modelled by taking each dynamically separate ice mass at a reference time and modelling it through time in isolation. This is referenced in a new section on 'extensions and limitations'.

- **Calibration strategy**
  Recently, Eis et al. (2019) presented an alternative for the standard initialisation technique in OGGM. They show that it is important to not only consider present glacier length in the calibration but also the full geometry. In this way, the uncertainties in hindcast simulations can substantially be reduced. As I understand it, you use the standard OGGM spin-up which is not intended to reproduce length changes even in the last century. If there is no length change calibration in this period, why would you expect reliable performance over an entire millennium? Please justify. From my perspective, it is a prerequisite that the approach is calibrated against the length change record (Leclercq et

al., 2014) to guarantee a certain reliability over multiple centuries.

Unfortunately, the model runs were already completed before the Eis et al. (2019) work was available, with our version of OGGM kept constant at the labelled 1.1 release to ensure consistency between runs performed at different times for different regions and different GCMs. As Leclercq et al. (2014) is amongst the best sets of long-term length records for glaciers across many regions, and is still sparse both spatially and temporally, we do not feel that we have sufficient data available to both calibrate to length changes and compare against a separate set of length changes. We do, however, see benefits to considering the performance of a 'naive' model setup, whereby we do not tailor the setup of the model to reproduce one particular variable of interest; by calibrating to a certain variable, we can guarantee that the model produces results that are relatively 'correct' on the largest scale even if it is not the result of significant model skill, which is not the case if we allow the model to work in a way which is agnostic of the variables we will be examining from it. With the results we produce, we are able to find cases where the model performs well and cases where the model performs poorly, and it is the comparison with observed length changes rather than 'correct' lengths over the last 1000 years that is our primary focus.

In the new 'extensions and limitations' section of the discussion, the explanation of the sparse measurements available is reiterated (after featuring in the introduction) and it is made explicit how modelling for a period when there is essentially only one metric with sparse and inconsistent measurements imposes additional restrictions on how we can calibrate the model and what measures of model performance we can produce.

- **Manuscript structure**
  The 'Methods & Data' section and the 'Results' sections appear as single entities and they lack some structure. For the 'Methods' section, you could present OGGM with some more details on the inherent calibration. For the 'Data' part, you can specify the RGI and the length record with the pre-processing details. Moreover, the climatic forcing could be described in detail. The 'Results' section is a melange with a discussion. I would mention that in the section title. Also try to introduce some sub-divisions.

  We split the 'methods and data' section into a 'data and model description' section (with subsections for OGGM, the six GCMs used, and glacier data from Leclercq and RGIv6) and an 'experiment description' section on the way our study's specific runs were conducted [I]. We split the existing 'results' section into separate 'results' and 'discussion' sections for readability [I], and provide a heavily modified and extended discussion.

- **Objectives**
  As already mentioned above, you raise high expectations in your introduction but the conclusions appear rather vague (e.g. abstract, L35, L44, etc.). Therefore, I suggest that you better streamline the manuscript on the conclusions that you are able to draw.

  The discussion now directly responds to the stated goals and questions raised in the introduction. Two of the new figures, P3 [I] and P4 [I] provide more quantitative measures of the ability of the model to reproduce observed lengths and length changes when driven by each GCM. The new discussion frames results in response to the question of how well the model reproduces 20th century trends under each GCM input when run over longer timescales, and ties in existing data on relative roles of temperature and precipitation.

- **Open dataset**
  From your description, I appreciate the effort to link the RGI to the length change record from Leclercq et al. (2014). I therefore suggest that you provide a lookup table between the RGI-ID and the ID numbers of the length record, specifying 'positive' and 'best-effort' matches as well as not retrievable entries.

  A link to the file showing the matching between RGIv6 and Leclercq glaciers is now provided, addressing a request for this information [I].

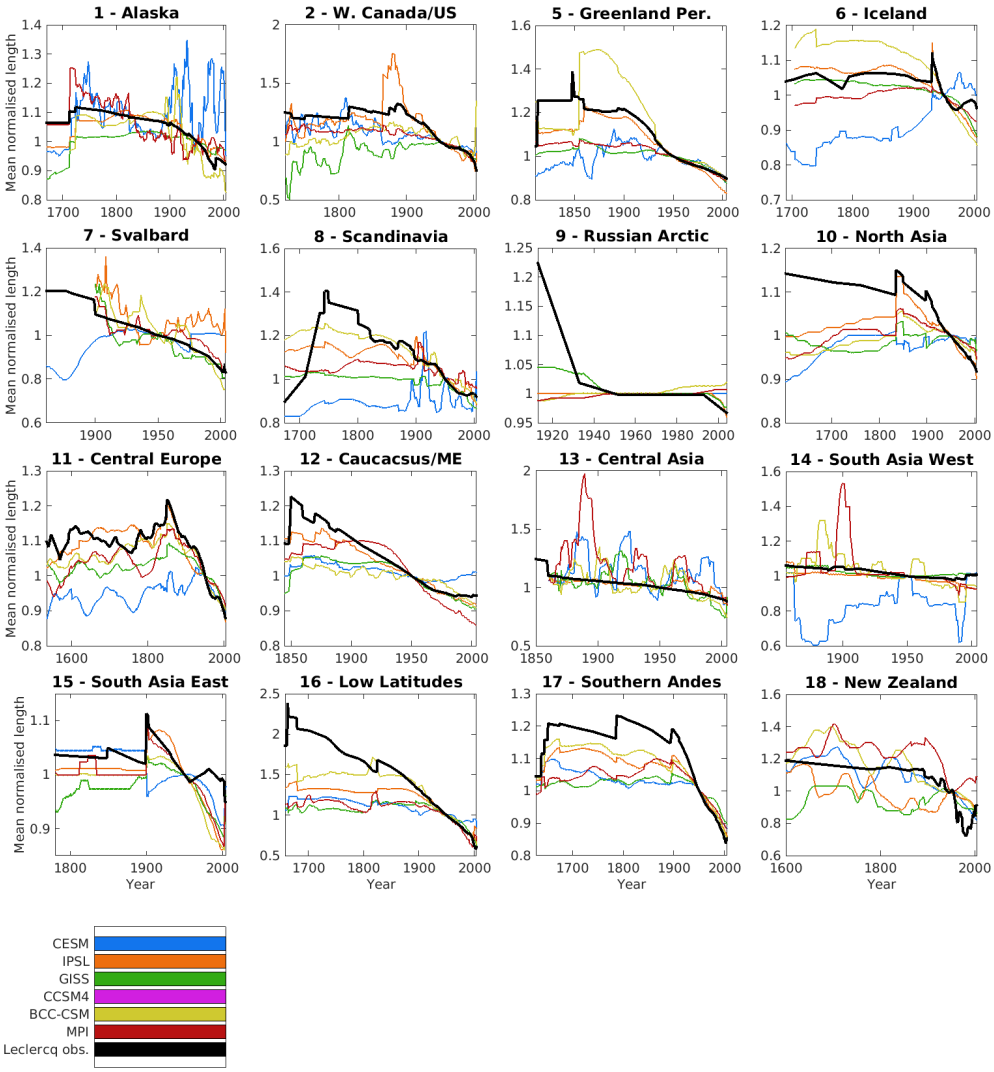
- **Detailed comments**
  L35 To what challenges do you refer here? You emphasise this point already in the introduction but you leave it vague here. Please substantiate.

  We clarify the challenge here, of needing a system that can reach and properly represent stability in a model calibrated for a time period in many cases without near-equilibrium periods for reference, and also discuss the 'theoretical equilibrium finding' behaviour of OGGM's calibration process in the expanded description of surface mass balance and calibration.

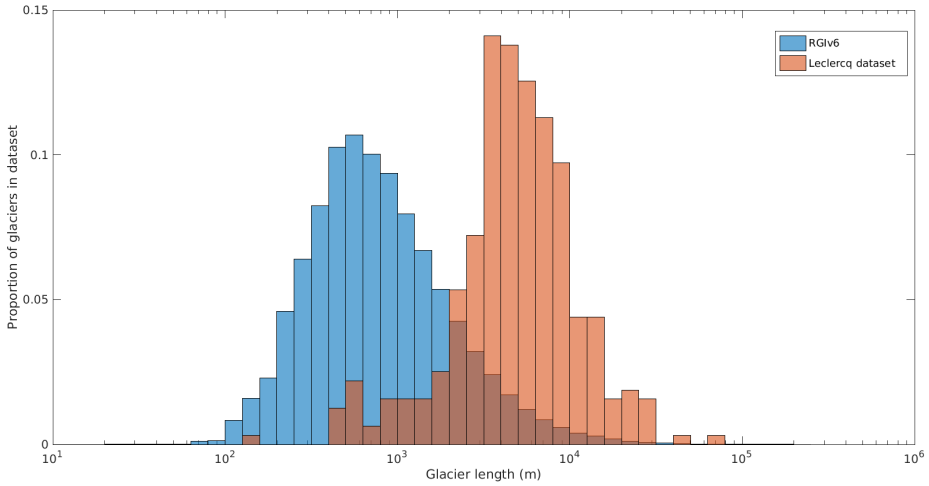
  L107-108 From this sentence, I would have expected that Table 1 comprises number of glacier units and glacierised area per RGI-region. Please add.

  Number of glaciers are added [I] but areas are not provided, as area is not featured in the Leclercq data, and the only other way to obtain areas is to take more recent area figures from the RGI (when the reference date for Leclercq is 1950) based on the matching we perform. This means that area cannot be used as a comparison between model results and observations, and therefore it is questionable what value its inclusion can bring.

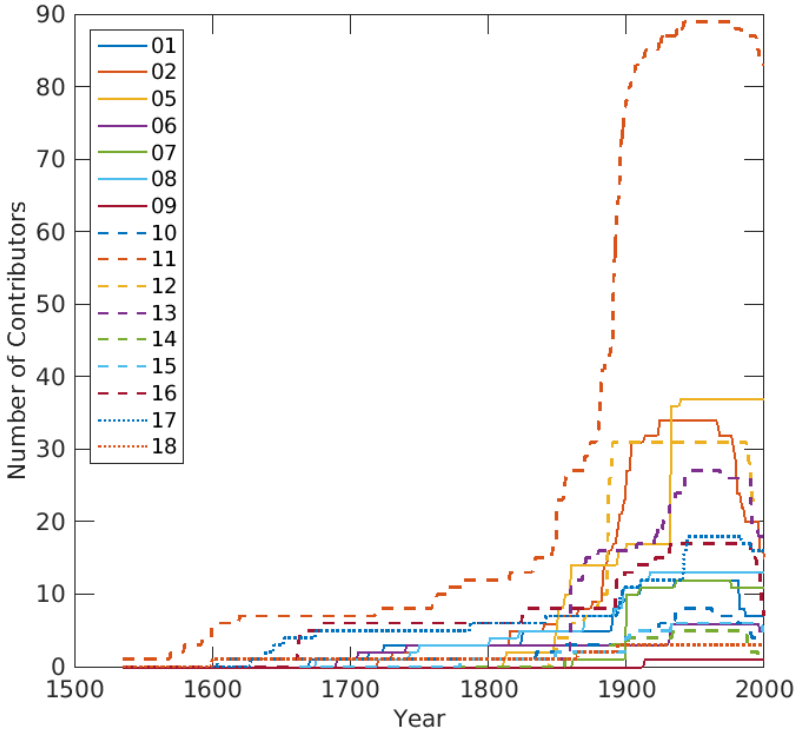
Revised figure 2: the set of modelled glaciers that contribute to the regional average over time now varies to match the set of glaciers that have Leclercq data available for any given year. The intention is to show the impact of the changing number of contributors to the regional means. Generally speaking, where spikes appear across multiple GCM runs in this new figure but are not apparent at the same time in the paper's existing figure 1, this is likely to represent an artefact of the dataset rather than an actual change in modelled glacier lengths.



Proposed new figure P1: The distribution of RGI glaciers vs the distribution of Leclercq glaciers. This is useful for general context on the datasets involved, but also illustrates the considerable bias towards larger-than-average glaciers in the Leclercq dataset, and backs up the claim that is now added; that contrary to the criticism that smaller glaciers in the Leclercq dataset disproportionately affecting normalised regional averages, the Leclercq dataset considerably overrepresents larger glaciers and the larger or more rapid normalised changes that smaller glaciers can experience are likely more representative of the bulk of glaciers.

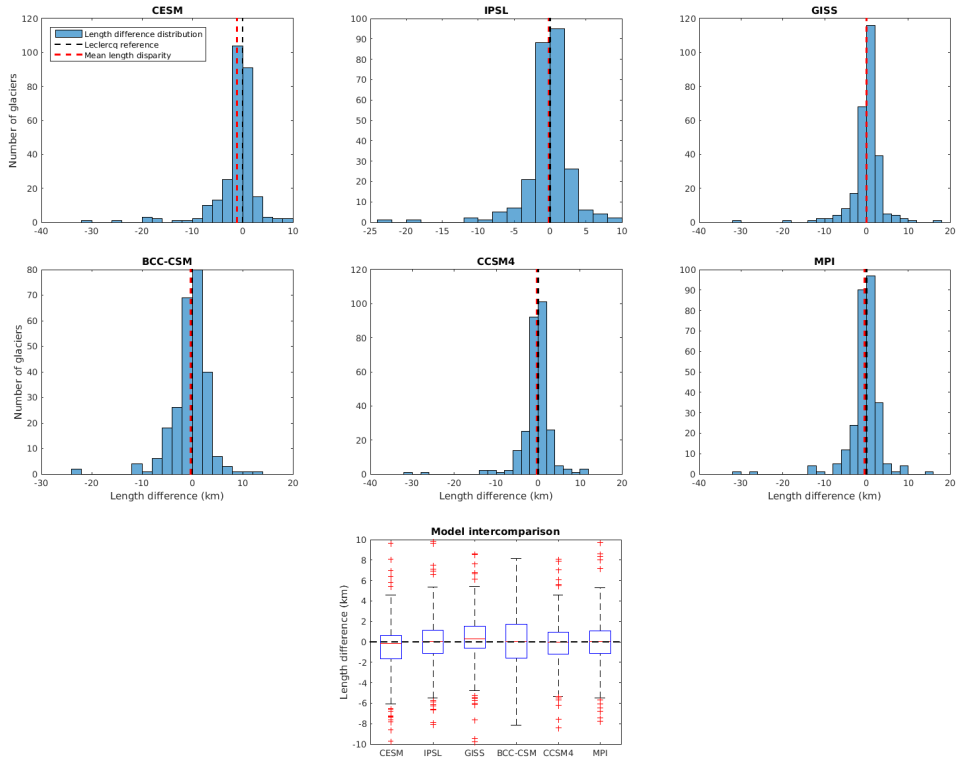


Proposed new figure P2: Changes to the number of glaciers which contribute to the Leclercq mean by year. This contextualises the potentially 'spiky' nature of the Leclercq averages; where there is a rapid jump in a particular region, it is possible that sudden changes in the mean glacier length in that year are explained as an artefact of the data, rather than representing OGGM outputting rapid changes in glacier length.



Proposed new figure P3: distribution of absolute length errors in 1950. This is part of the effort to address criticisms of the exclusive use of normalised length changes in the submitted draft. We see a moderate bias towards underestimating 1950 length from the CESM-driven runs, and towards overestimating from GISS (despite the mean not reflecting this due to the effect of outliers), and a greater range of length changes generated by the BCC-CSM-driven runs.

Distribution of per-glacier differences between modelled and Leclercq-observation length (absolute)



Proposed new figure P4: Plotting the modelled and observed per-glacier trends over the 20th century (including all glaciers which have 68 or more years in the 20th century covered by the Leclercq timeseries, which represents the point where 90% of glaciers are included). This addresses the issue raised of the glaciers being represented only through regional means. The data shows that the magnitude of observed trends on the scale of individual glaciers is not well modelled by OGGM, and that the differences in how well represented glacier changes are between models using different GCM forcings are small compared to the difference between the modelled changes and the observed changes. The less-than-parity regression line slopes for every model suggest that OGGM is likely to underestimate glacier retreat, especially for larger values of observed retreat. A similar plot for normalised trends shows an almost identical picture, but we choose the absolute trends simply because the required axis scales are less impacted by outliers.

Per-glacier 20th century trends: modelled vs observed (absolute, all glaciers)