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 this, 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 2 itemised response

- **I find the focus of the paper a bit unclear. From the abstract, I get two objectives, which are obviously linked, but not spelled out very clearly: (i) test whether calibration of the model during the retreat phase is good enough and leads to adequate results also in times of advance or stability, and (ii) identify whether precipitation or temperature anomalies are more responsible for glacier length changes at multi-centennial time scales. I think it might help to use (ii) as the main objective, which would require (i) as an intermediate step.**

  See the response to the point two below ('In the introduction, the attribution...') for more specific discussion on the rationale for refocusing around point (i) as the primary objective and (ii) as a supporting objective. Using the constant-climate runs as a tool for examining the full GCM-driven runs rather than as an end in themselves can help us to understand why the modelled glacier lengths do well or poorly in reconstructing length changes from the observational record.

The discussion now directly responds to the stated goals and questions raised in the introduction by framing results in direct response to the question of how well the model reproduces 20th century trends under each GCM input (see also new figure P4) when run over longer timescales, and tying existing data on relative roles of temperature and precipitation. The conclusions section is rewritten with reference to the bolstered discussion section, and is able to provide more substantive points that link more directly to the introduction.

This also relates to L38-39: "it is important that these models are examined over time periods where more stable glacier geometries were expected". It could be argued that if a model is not foreseen to be applied in conditions where glaciers are stable or advance, the model does not need to be able to show such behavior (e.g., to my knowledge, the representation of ice geometry change of the model of Huss and Hock, 2015, does not provide for advancing glaciers). It would be good to give some explicit reasons why it is important.

There are two threads to the response. The first is that the calibration process for OGGM makes an assumption of the existence of equilibrium states and OGGM's ability to keep mass balance close to zero under appropriate conditions. The second is a more generalised point on the philsophy of modelling, whereby it is inherently a deficiency to have a model which cannot reach non-trivial (in this case, not zero length/volume for a glacier) stable conditions if it is intended to model a phenomenon which provably reaches non-trivial stable conditions in reality. Physically, we know that the behaviour of a glacier in response to a change in climate is - with a delay - to reach a new equilibrium, and that is true of any climate change not large enough to either have the glacier entirely melt away or grow limitlessly, and if a model cannot replicate this process of transitions between equilibria, it is somewhat assuming its own conclusions; a model that works only for periods of continual glacier retreat cannot be meaningfully used to predict or hindcast continual retreat of glaciers.

On the first point, we make a short reference to this in the introduction that directs the reader to detailed description of OGGM's mass balance calculation and sensitivity calibration in the methods section [I]. On the second point, we include a version of the argument above, but also recognise that there are cases where the model is expected to work less well (specifically glaciers with terminuses farther along the flowline than they are today, or at any point during the observational record) and later discuss this in the new 'extensions and limitations' section within the discussion.

In the introduction, the attribution of glacier length change to either precipitation or temperature anomalies seems like an afterthought. I think it would help the paper a lot to restructure with one clear objective (which I think could be this attribution – but the authors may disagree).

With the additional analysis that has been generated in the process of revising the manuscript, it has become clearer to us that it is necessary to focus on assessing the performance of the model, rather than being able to take the additional step of making the response to individual climate variable forcings the primary subject. This is because we find that despite model skill in reproducing (at least) qualitative regional trends for many regions, the reproduction of the magnitude of per-glacier trends is shown to be poorer, so the matter of model performance under 'normal' is much more complex than just a box to check on the way to assessment of single-climate-variable runs. In large part this is probably due to the 'chaining' of moving parts that are being examined; not only do we need to test the performance of OGGM running over long timescales (which is already subject to heavy limitations on data to compare against), but we are also implicitly testing the performance of each of the GCMs as a timeseries of climate variables for producing realistic glacier states. To make the constant-climate-variable runs the focus is essentially another link in this 'chain' as we assess the impact of the climate variable isolation simultaneously with OGGM's skill and the GCMs' climate datasets, all with considerable limitations on the observational dataset to compare against.

For this reason, we choose to make the constant-climate-variable runs a tool to examine the full climate runs, rather than the focus of the paper (with all the new analyses that would require), and this purpose is now made explicit, both within the introduction and within the refocused discussion and conclusions sections that have been directly tied to points made and questions raised in the introduction. Points from the above paragraph are incorporated partly into the introduction, and partly into the new 'extensions and limitations' section within the discussion.

- **It probably would be helpful if a bit more is said on the setup of climate forcing in OGGM: the authors are referring to "level 3" preprocessing of OGGM, but don't give any details on how OGGM treats climate model output before application to the mass balance model (this concerns bias correction/anomaly coupling, estimation of solid precipitation, any corrections etc). As it stands now, readers might by surprised by the apparent ability of CMIP5-type models to represent mountain climate conditions accurately enough, which is only half the truth. (see also L166, "as provided", which is not true)**

  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] and the justification for the calibration using OGGM's default method is made explicit [I]. We also clarify the details of the scaling of climate model data to 1900-2000 CRU data [I].

- **Regional averages are presented and discussed, and this comes at the relatively high cost of having to find a way how to calculate regional length changes (see discussion around L120-125). It is my impression that the calculation of regional averages is a purely graphical requirement, needed in order to avoid having to inspect 339 individual glacier time series. This points to a more significant problem, which is that the assessment of the model results depends too much on this graphic representation. I would recommend the authors to expand the analysis of results to a more quantitative evaluation, such that the assessment of regional differences depends less on visual inspection of graphs that necessarily are associated with shortcomings (such as the spikes resulting from changes in the observational ensemble).**

  There is a significant extent to which the regional grouping of data is a result of questions of how to display the data, though this is in line with many other studies in which RGI regions are considered natural partitions of glacier data. We believe that showing the data on a regional basis is appropriate, given the expectation that glaciers within a region will behave more similarly, and experience greater similarity of climate variability and trends, than glaciers in different regions. We do however recognise that the importance of the criticism that too much of the assessment of the results rests on the specific representation of regional data we have chosen.

We maintain the region-based display of data in the existing set of figures, but provide a new set of region-agnostic quantitative measures centred on new figures P3 and P4 [I] in order to remove the exclusive reliance on regionalised datasets. We also give additional context to the regional plots by providing a visualisation of the number of glaciers in the Leclercq data per region for each year in new figure P2 [I], and illustrate the impact of the changes in the observational ensemble with a modified figure 2 [I] that, in contrast to figure 1, has the set of glaciers averaged from the modelled dataset vary year-on-year with the set of glaciers available in the Leclercq dataset.

At least, it might be worth to extract data from the modeled glaciers at the same time as observations exist (adding a third version to Fig. 1 and 2), so that the modeled regional average would have the same spikes as observations if it was perfect. But I think this would not be the optimal solution. The analysis based on linear trends aims in the right direction, but doesn't really relate to the observations, so it is not helping with this issue.

The revised version of figure 2 [I] serves this function, with the set of glaciers contributing to the model average matching the set of glaciers with available observations from Leclercq, and additional context for these changes is provided by new figure P2 showing the evolution of the size of this set over time for each region.

- **The discussion of Fig. 4-6 would form a nice basis to infer something about the adequacy of the climate models if the comparison of the model results to observations was more quantitative. E.g., it would be possible to quantify how much the glacier model performance is reduced (presumably) if either temperature or precipitation information is withheld, giving further insight into their relevance (and the climate model's ability to represent precipitation and temperature evolution – after all, it is possible that a model's performance increases when one of the two variables is held constant).**

  These are great ideas for developing the analysis of constant-climate-variable runs and what they can contribute to our understanding of model performance, but in light of the point above on the paper's focus and concerns over the potential overloading of analyses after the addition of a number of other metrics (see new figures), we would prefer not to add new data analyses or figures that relate specifically to the constant-climate-variable runs unless it is considered necessary for rounding out the paper.

  I also think that the discussion of variances misses the opportunity to say more about physical, climate related causes of regional differences. This is also true for the discussion of relative variances > 1, which basically imply that there is dependency between temperature and precipitation. This is discussed on a technical level (it might be added how the way the solid fraction of precipitation is calculated in OGGM adds to this dependency), but there are also climatological reasons that should be considered here. This should include a discussion of the relevant literature on precipitation vs. temperature influence of glaciers (which is currently almost completely missing). Also, a discussion

of Marzeion et al. (2014, DOI: 10.5194/tc-8-59-2014), which includes similar experiments, may be helpful.

We give reasons above for not wanting to elevate the material on runs that keep one of the climate variables constant above the core goal of observing and assessing OGGM performance for runs forced with the full GCM data, but we nevertheless expand the discussion in the direction suggested as all of the requisite data is already available.

Within the context of the expanded discussion, the implications of the results from the constant-climate-variable runs are discussed, and they are also related directly to the full GCM climate runs and their performance in reproducing observed length changes. The suggested reference features in this discussion [1].

- **Specific/minor comments/suggestions**

Responses are given only to selected points, with corrections to sentence structure and citations straightforwardly implemented unless otherwise specified.

L3 (and throughout manuscript): delete "post-" from post-industrial (unless this is a standard term – but to me it sounds like "after the industrial age", which is not what you mean)

The logic behind the initial terminology was the idea of things happening after the onset of the industrial period, but we appreciate that this is not made sufficiently clear and recognise that the suggested use is an improvement. This is implemented throughout the paper, and any references that are related to the onset of the industrial period rather than to the whole of the subsequent period are made explicit [I].

L32: OGGM also include precipitation in the calibration

Along with the comprehensively extended section on OGGM calibration and calculation of mass balance [I], we revise all other references to OGGM's calibration and use of climate data, and where appropriate refer to the material in this new section.

L35: please spell out some of the additional challenges

We clarify the challenge here, of needing a system that can reach and properly represent stability in a model calibrated for a time period 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.

L115-117: "smaller in relative magnitude" yes, but for many applications, the relative magnitude of changes is not that important, but the absolute mass change. So this is maybe not such a big problem.

This is true, but we do feel that it is sensible to mention. In a paper that was focused on overall volume change, this would likely be a small effect, but as we focus on length changes, the effect is probably not insignificant. New figure P1 [I] provides some context, showing the real disparity in length distribution between Leclercq and the RGI.

L119-120: it is unclear to me here why a mean regional glacier length estimate is needed? Up to here it seems all comparisons to observations are done on a perglacier-basis (which seems like a better idea to me).

This point is addressed in the response to the longer 'regional averages...' comment above.

L 151: from my own experience I appreciate the difficulty of doing the matching, but 38 "not found" glaciers strikes me as a high number. Typically, glaciers with length observations tend to be more "famous" than the average glacier. Might it be worth checking in other data bases (e.g., GLIMS) based on glacier name?

The method for matching glaciers was actually chosen specifically to avoid matching glaciers which are given the same names but which do not properly match according to an objective standard for glacier location and area data. This is for two reasons: firstly, differences in the way glaciers are partitioned can make even glaciers which are genuinely the same ice mass (or parts of what was once the same ice mass) inappropriate for comparison. Where the same glacier in the RGI and in the Leclercq dataset cannot be reconciled as representing the same dynamically connected ice mass, it is not valuable to compare model results based on the RGI definition to length changes based on the Leclercq definition. Second, where glaciers can be identified across databases by name but either their location or their geometry shows a serious mismatch, it suggests that the quality of the data for that glacier in either or both inventories is not accurate enough, which is a reason not to model that particular glacier based on RGI data and expect a viable comparison with Leclercq observations.

This explanation can be added if it is deemed necessary, but out of a desire for efficiency in an already heavily expanded methods section, we choose not to include it now. The matching method is already described in detail and we also now have the files containing the matched glacier list and the method description linked as an online resource [I].

L 153: please provide a table linking RGI-ID to Leclerq's database as a supplement (as a service for similar, future studies)

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

Fig. 1 & 2: I'm surprised that also the model results are very spiky in some regions. Why is that so? It would also be helpful to use the same vertical axis range in all subplots (even if it cuts off parts of some graphs), since the normalization make the regional comparison easier (and the different axis limits make it harder).

Spiky model behaviour is typically due to glaciers which, under the model, become very small (but crucially do not disappear) and then in periods of colder/higher-accumulation conditions, have the front of the glacier considered much further along the flowline due to the way OGGM calculates the front of a growing glacier. We discuss this in a new 'extensions and limitations' section within the discussion that examines the weaknesses of OGGM as well as the potential for further use.

We agree that a fixed axis range for the subplots could make for easier comparison between models, but we would prefer not to make this change because we deem the loss of easily-readable information this would result in - cutting off parts of certain graphs, some of which are not spikes that result from modelling issues but actual large relative changes changes, and 'squashing' the output of different GCM runs together in regions which currently use smaller vertical axis ranges (e.g. Southern Andes) - to be greater than the increase in readability from a region comparison perspective.

L182-186/Fig 1&2: I don't really understand what makes regions 14, 15 and 18 stand out from the other regions? Again, I think a more quantitative comparison of the model ensemble to the observation ensemble would be very helpful.

The description of the individual features of these regions is not particularly effective, so this section is removed [I] as part of the rewritten results section with a more quantitative focus, including the additional information made available by the new graphs P3 and P4 [I].

L189: not sure what is meant by "stratified", so I also don't understand the following argument

We now use 'when the gaps between the output for different GCMs are large compared to the internal variability of a single output' instead of 'When the model results are highly stratified' [I]. Use of the term 'stratified' was overly reliant on how the lines look on the graph, while the new phrasing refers to features of the data.

L199 and following: would it make sense to do this analysis first for each glacier in the region, and then build the regional mean? This could result in an indication how robust the estimation of the "inflection" year is, and also give some more information on intra-regional variability of the glaciers' behavior.

The issues that exist with performing a split regression for each glacier and then combining them are 1) that individual glaciers are more sensitive to variability than a mean of a larger set of glaciers (see also the next reviewer point on the sensitivity of the inflection year), and 2) that there is no clear way to combine these sets of split regressions to form a useful mean. It is easily possible to determine a mean year of inflection, but then there is no intuitive idea of a mean slope before and after this point. Nevertheless, we do have an increased focus on per-glacier rather than per-region data in the form of new figure P4 [I] directly, and indirectly through quantitative measures which take per-glacier distributions rather than focusing on data from regional mean outputs (including new figure P3 [I]).

L213-125: I find this argument a bit weak, given that the "inflection" year is probably very sensitive to short-term interannual variability in the climate time series

As short-term variability is always present within the timeseries and we overwhelmingly see the 'inflection' year in a relatively short window around the onset of warming in the industrial period, it does seem that it is longer-term trends which dominate the determination of the inflection year. The sensitivity to short-term changes is also mitigated somewhat by performing the split regression on regional means rather than individual glaciers, so impacts of glaciers in situations prone to rapid advance or retreat in response to short-term changes are somewhat dampened. We do, however, agree that the idea of linking the inflection year to the year that recent retreat begins, even though we do not identify the two, may be overstating what can be determined from this data, so this point is rewritten to avoid speculation about the year retreat begins determining the overall size of retreat. Instead we comment quantitatively on the relationship (or lack thereof) between the inflection year and the slope of the 2nd part of the regression (replacing the purely quantititative expression).

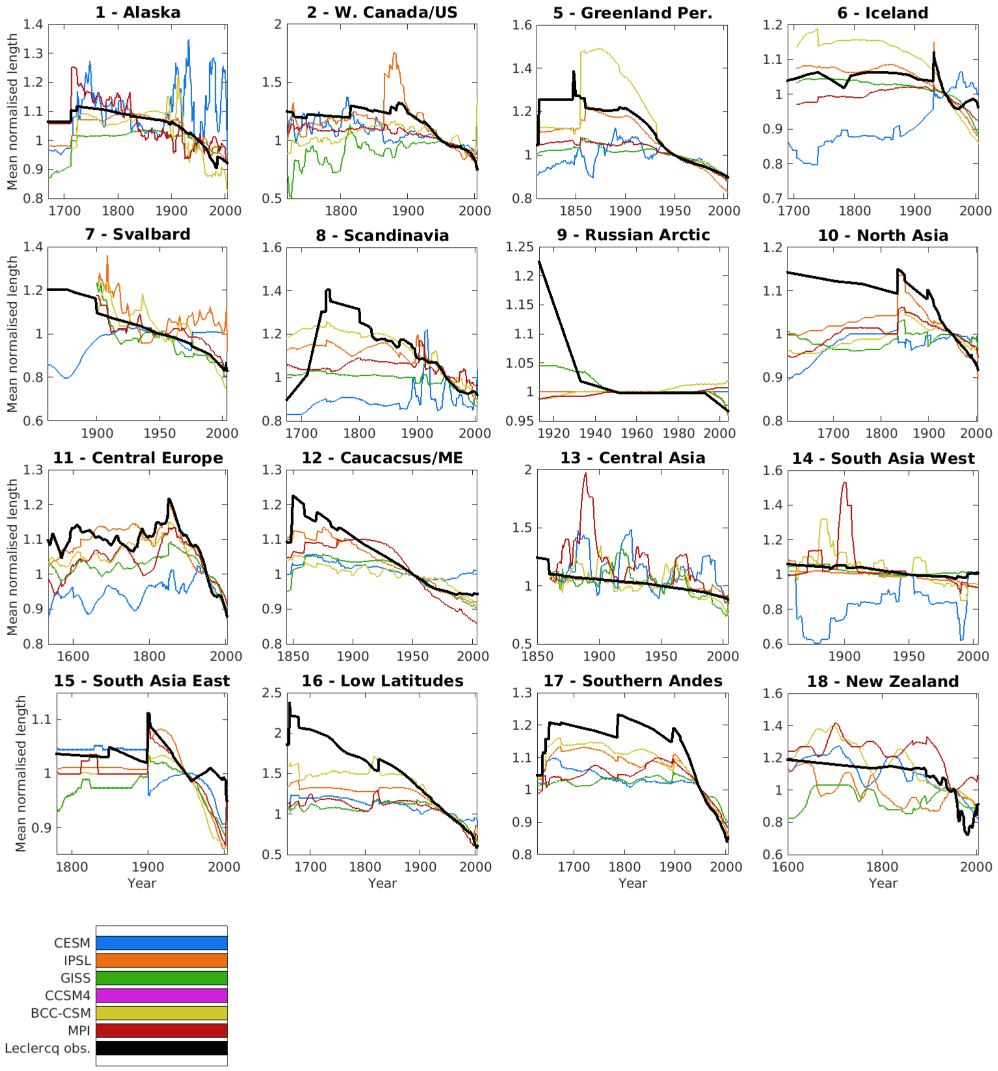
Fig. 5: was the non-zero melting threshold temperature of OGGM taken into account when calculating the temperature time series shown here (also, in L226, it says "degree-days" – is it really days, or months)?

The same melt threshold as used in OGGM forms the basis for the degree-months. With the new detail provided in the methods section on the processes in OGGM, and particular focus on surface mass balance [I], all references to OGGM method have been checked, and if needed made more specific and directly referenced the new detailed description where necessary [I].

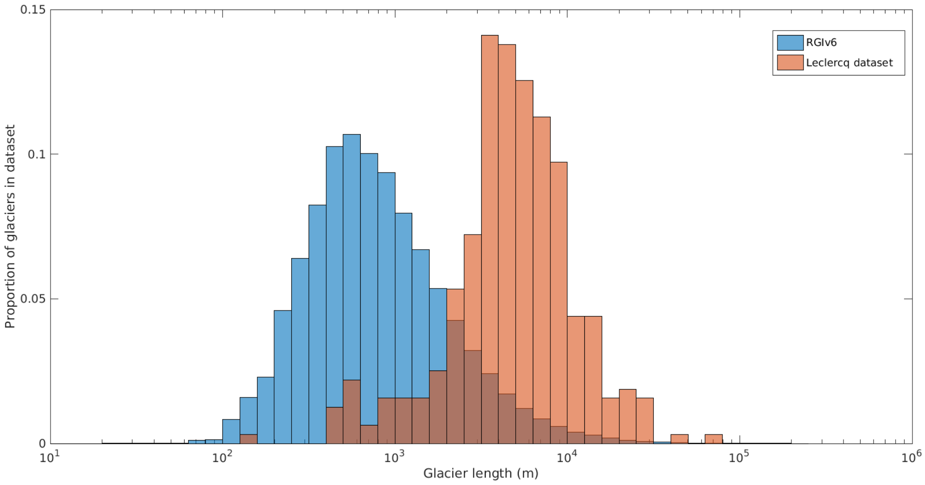
L262-263: it is not only OGGM, but (probably at least as much) the GCMs that are responsible for the level of agreement.

This specific statement is removed with the restructuring of the discussion and conclusions to directly and consistently reference the questions and aims raised in the introduction. However, we also make reference to the difficulty of conclusively determining which of OGGM and the GCM data is responsible for features we see in the outputs due to the 'chaining' of models - the GCM itself, which then feeds data into OGGM - in the new 'limitations and extensions' section.

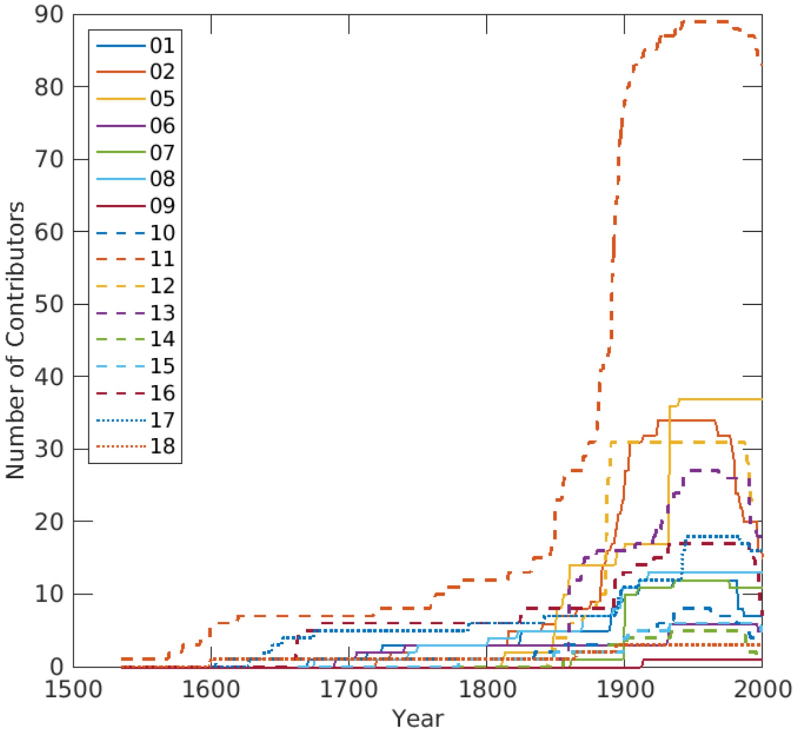
Revised figure 2: the set of modelled glaciers that contribute to the regional average over time now varies to match the set of glaciers that have Leclercq data available for any given year. The intention is to show the impact of the changing number of contributors to the regional means. Generally speaking, 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.

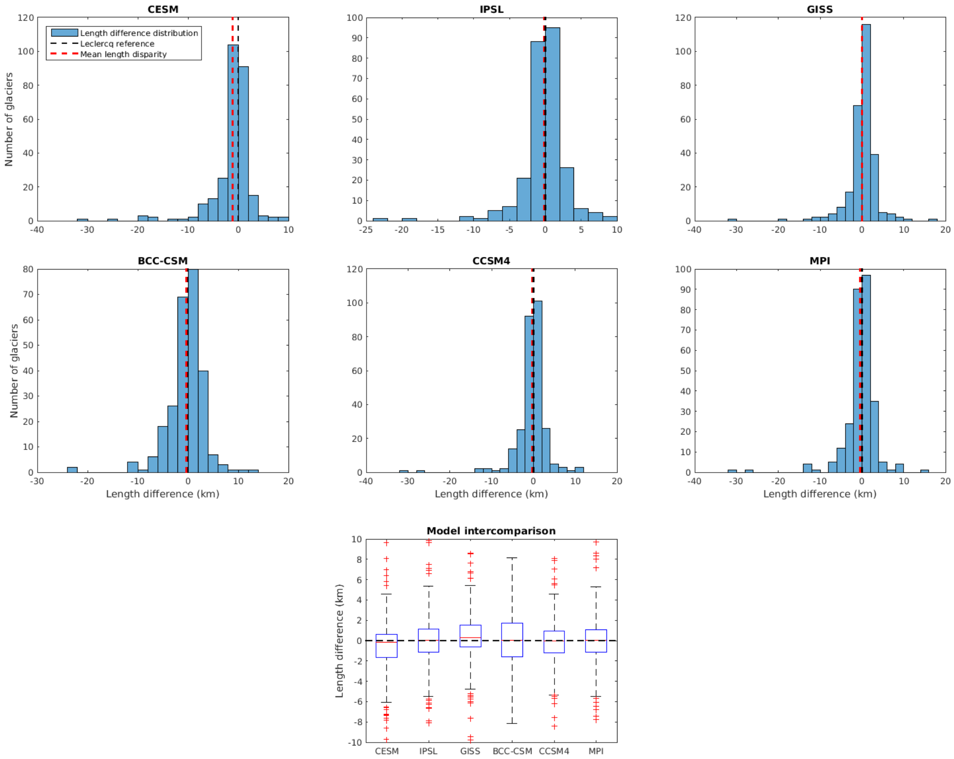


Proposed new figure P2: Changes to the number of glaciers which contribute to the Leclercq mean by year. This contextualises the potentially 'spiky' nature of the Leclercq averages; where there is a rapid jump in a particular region, it is possible that sudden changes in the mean glacier length in that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



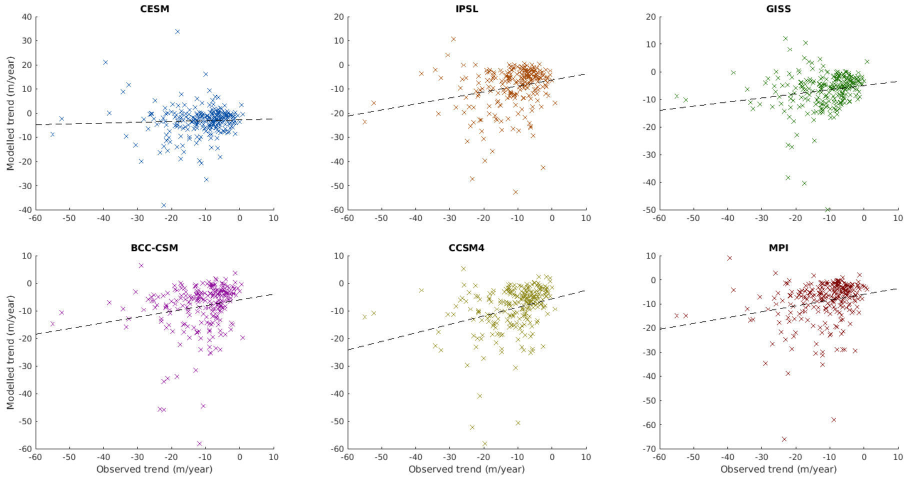
Proposed new figure P3: distribution of absolute length errors in 1950. This is part of the effort to address criticisms of the exclusive use of normalised length changes in the submitted draft. We see a moderate bias towards underestimating 1950 length from the CESM-driven runs, and towards overestimating 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)


CESM
IPSL
GISS
BCC-CSM
CCSM4
MPI
Modelled trend (m/year)
Observed trend (m/year)