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 3 itemised response

- **1. At this point, many statements a bit vague and the analysis presented are not very in-depth. Additional information (mainly quantitative information) and analyses will be required to really make point that the authors are able to use OGGM to closely reproduce past glacier changes. Two main points here: a. You chose to mainly focus on relative length changes at regional scales and explain why you present the results in such a fashion. This is definitely fine, but I do not think this should impede you from also giving some results in absolute values and for individually modelled glaciers. This is just a matter of showing the results differently and does not require any additional runs/simulations. More specifically, it would be interesting to see how the model is able to reproduce the present-day glaciers: i.e. are the glaciers that you obtain at the end of your simulation (around 2000) close to the observed ones? So far, you argue that it is important to look at the model for periods when glaciers were more stable (e.g. l.38-40), but this should not stop you from also considering**

the model performance for the recent past and its capability to closely reproduce the present-day glacier. An example: take a glacier for which the relative length change is well reproduced over the last 300 years, but where your modelled present-day length is 8 km vs. an observed length of 20 km... This means that your glacier was also 2.5 times too long 300 years ago (everywhere between 300 years ago and now): would you argue that the model does a good job at representing the changes here? Would be good to have a figure (e.g. in suppl. Mat.) with on the x-axis the observed glacier length at present-day, on the y-axis the modelled present-day glacier length (after transient run) and having every individual glacier plotted in this (for all regions together; could do this for one climate model)

Additional quantitivate measures form the bulk of the new analysis in the revised version of the paper. However, an initial point should be clarified, regarding the comparison of length changes vs the comparison of length in a given year. While the GCM data is scaled in such a way that the 1900-2000 mean precipitation and temperature are made to match the mean over this period in the CRU 1900-2000 dataset (on which the mass balance sensitivity is calibrated), we recognise that there will be overall biases in each GCM in the rest of the last millennium. This will see different sizes for modelled glaciers at the start of the 20th century when driven by different GCMs, which will result in different lengths during the 20th century regardless of matching climate variable means. Geometry responses to climate are also non-linear, so the matching climate variable means for 1900-2000 do not mean matching aggregated surface mass balance. Looking at changes (including absolute changes) allows us to focus on changes in climate over time rather than having to account for mean climate differences and changes over time simultaneously. We also note that due to the calibration of GCM climate variables to match CRU 1900-2000 means, this may have some effect of an artificial convergence of 20th century mean absolute glacier lengths (equilibria will be similar for climate datasets with the same temperature and precipitation means), but the same effect does not apply to patterns of change over the 20th century.

Three new figures are created (two of which are shown below) which deal with each of the concerns here; regional aggregation, absolute length comparison, and performance metrics for individual glaciers. New figure P3 [I] looks at the distribution of absolute differences between observed and modelled 1950 lengths - while the focus is on changes, having this context on whether models tend to over- or underestimate glacier lengths in the year other figures are normalised to is important. The new figure P4 [I] gives a scatter plot of 20th century length changes in the model vs the observations, as a quantitiative measure of the performance of the model for each glacier in a way which is not aggregated regionally. Finally, the exact figure suggested at the end of the comment is generated, but not included here; the caveat with this figure is that present-day (i.e. end-of-modelling-period) modelled lengths cannot be compared particularly well with Leclercq observations (see drop off of the number of Leclercq records available in many regions towards the end of the 20th century), so it is necessary for the observed 'present day' glaciers to instead be taken from the RGI, which also provides the geometry for initialising glaciers before the model run. This inconsistency in the nature of the length records used for comparison can reduce the clarity of the results, so we suggest this figure should appear in the supplementary material rather than the main part of the paper. A review of the justification in the text for a focus on a) changes and b) normalised lengths is conducted, incorporating the clarification above.

b. Role of the SMB. You barely mention the SMB component of the model,

which I found surprising, given that this is the main driver for the glacier behavior (the dynamics then translate your SMB forcing – with a lag due to the response time – to a length change). For instance, when considering the role of temperature vs. precipitation forcing, it would be highly relevant to describe how much these components affect the modelled SMB (with quantifiable information). Many studies have provided insights in the role of temperature vs. precipitation forcing for the SMB (e.g. Lefauconnier et al., 1999; Braithwaite & Zhang, 2000; Oerlemans & Reichert, 2000; Sicart et al., 2008; Trachsel & Nesje, 2015), and often found that the temperature is the main driver. Your finding that temperature is the main driver directly results from the calculated SMB, which is far more sensitive to temperature changes than it is to precipitation changes. Would be really nice if you could show some of the calculated SMBs and perform some basic sensitivity tests (e.g. what happens with SMB when forced with +1 degC, -1 degC, +20% precipitation,. . . etc.)

The additional detail on SMB is important to include, but we do have significant concerns about whether the suggested additional synthetic runs would actually add to the paper. Our specific intent is to examine the way OGGM reproduces glacier lengths under specific GCM climate datasets, and the runs with constant temperature or constant precipitation are ways to determine which of these explains most of the reproduced glacier changes. The suggested runs do not seem to serve this specific purpose; rather they are interesting pieces of general glacier model testing, but do not relate to the aims of this particular paper.

We do, however, include the suggested references for context on the variance explained by the constant-climate runs. In general, 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].

- **2. Title is a bit misleading: when reading 'regional' glacier length changes, I would expect that an entire region is considered. However, glaciers from various regions are selected, which in every case represent only a very small subset of all glaciers in this region. Suggest reformulating this, which could be done by simply omitting 'regional'. Or should rather mention something like: 'in various regions around the world'**

It is of course true that specific subsets of glaciers are used, but we do feel the term 'regional' is appropriate because we are clear about the number and distribution of glaciers used (see new figures P1 and P2), and because there is no cherrypicking of glaciers; we simply use all the glaciers that have available length records in one consistent format. See also the response to general comment 5 on the issues that face all modelling efforts that try to operate on a per-glacier basis; we believe that using an available dataset which is separated into regions and describing the outcomes as 'regional' is consistent with established glacier modelling literature, even when the number of glaciers is significantly larger than those used here. We are just so lacking in long-term direct observations of glaciers that our approach is as 'regional' as we can be.

A short comment on regional representativity is added, referencing the new figure P1 on

Leclercq length distribution relative to the whole of the RGI distribution, with context from the relative scarcity of longer term data and the potential incompleteness of even modern inventories [I]. The manuscript is not yet renamed, but we are open to the possibility of renaming it if it is determined the 'regional' label is considered genuinely misleading.

- **3. Almost all the 'action' occurs in the pre-frontal glacier region (compared to the present-day ice cover): how well is OGGM able to handle this? More information is needed about how the flowlines are defined here, how the cross shape is parameterized,... etc. This information is lacking at the moment.**

  It is difficult to get a hugely accurate picture of how well OGGM handles this as it's difficult to determine how much of a historical (longer than present day) glacier state generated by OGGM is due to the way that glacier geometry for flowlines extended beyond their current bounds is handled and how much is due to differences in glacier evolution between the model and reality. Naturally examples of glaciers with periods of observed advance are relatively rare, and where they exist are often the result of unique processes like surging which OGGM does not handle well. The greatest uncertainty lies in the calculation of the hypothetical flowline extending from the initial glacier geometry that is used if the glacier advances; the width and cross section of the glacier are all handled identically for glacier extent beyond modern bounds, using the total ice mass at each elevation band to determine the dimensions, but the flowline for the initialised glacier is based on an algorithm applied to observed geometry while the below-glacier-terminus flowline comes from an interation on gradient from the end of the glacier. This method can struggle to deal with cases where glacier dynamics may cause the glacier to flow in ways which are not necessarily in the direction of steepest local gradient (e.g. heading over a lip of rock that is in the direction of existing ice flow).

  The above commentary is included in the 'extensions and limitations' section [I] and referenced where there is a mention of glaciers which extend beyond their modern boundaries. The same section also lists mitigating factors - such as the feedback of tongue elevation on overall mass change (and therefore length change) - which suggest that errors in flowlines beyond the modern tongue of the glacier should not cause divergent progressions of glacier evolution between reality and the model.

- **4. I generally found the manuscript relatively easy to follow and found the figures to be simple, but clear, which is very nice. At several occasions I did however get lost in long sentences (often multiple brackets are being used...) and had to read through these several times before getting the meaning of the sentence. I therefore suggest reducing the use of brackets, and splitting up long sentences where possible. Examples are provided in the 'specific comments' section below.**

  The issue of multiple brackets was mentioned a couple of times, and the paper has gone through a pass to restructure the sentences where they appear in order to enhance readability and remove the need for excessively complex formulations. We do not list all the responses to individual sentence structure concerns below; if the problem is purely a matter of fixing the text and no meaning needs to be clarified, the specific review comment is not listed and can be assumed fixed with a simple rewrite.

- **5. You describe this study as being a kind of first attempt to reproduce past length changes with a flowline model for glaciers in many different regions and**

suggest that this would open the door to regional scale applications. I agree with the former, but have some doubts about the possibility to fully extend this to regional scales. What about glaciers that are now separate ice bodies but used to be connected? What about glaciers that disappeared by now but may have existed before (whether as separate ice bodies or tributaries to present-day glaciers) – a field in which Parkes himself authored an important study (Parkes & Marzeion, 2018). I think it would be fair to also mention these issues/challenges in your conclusion, were you provide an outlook (last sentence of the manuscript).

These concerns are entirely valid, but are fairly uniformly difficult for all large-scale glacier modelling efforts which separately model each individual ice mass. We do not expect these impacts to be greater (or indeed lesser) in the extension to modelling entire regions (which is already underway), and it therefore should not make these efforts impossible, at least by the standards of existing regional modelling. The challenges of modelling that attempts to be comprehensive, rather than restricted to a limit set of glaciers, should be discussed, but not in the sense of it being prohibitively difficult.

The added 'extensions and limitations' section is where new problems that come from modelling regions with tens of thousands of glaciers, along with problems due to uncertain on total glacier numbers or on ice masses that vary between separate and contiguous through time, are discussed.

- **Specific comments**

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

l.13-16: statement is again rather qualitative here: could you provide concrete numbers that support this statement?

In light of the enhanced quantitative measures and expanded discussion separated into specific topics - see answer to general comment 1a - the end of the abstract is entirely rewritten, rather than keeping existing points and adding data.

l.22: 'direct observations of glacier geometry': what do you consider being a direct observation?

We consider direct observations to be any observation which makes a contemporaneous measurement of the ice mass. This includes satellite mapping, for example, but does not include reconstructions of historical glacier extent from things like moraines.

l.25-26: 'though even this is likely a significant underestimate (Parkes and Marzeion, 2018)'. Well, the number of glaciers is simply subjective, as it is related to the threshold that is used to decide whether a glacier is mapped (outlined) in the Randolph Glacier Inventory (RGI Consortium, 2017). We know that the number would be higher if smaller ice bodies would also be considered, so would not refer to this as an 'underestimation' here.

I think this is a misunderstanding of what Parkes and Marzeion (2018) says. It is explicit in not lowering the threshold for the surface area of what can be considered a glacier, restricting the upscaling to glaciers above the cutoff threshold used in the RGI. That there is an underestimation of terrestrial ice masses due to the partly arbitrary nature of this cutoff has some potential impact, but that is not the underestimation being referenced here. The referenced paper suggests significant underestimations of glaciers even an order of magnitude greater area than the RGI cutoff.

l.32: 'by default calibrates the glacier sensitivity to local temperatures based on CRU data. . . ': what criterion is used for calibration? You mention that CRU data is used, but what do you (try to) match in the calibration procedure? i.e. what is the target? (e.g. measured SMB, geodetic mass balance,. . . etc.)

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].

l.39: 'we expect that. . . ': strange formulation. You expect smaller and globally less consistent temperature trends? Based on what? Or is this just what the reconstructions suggest? Would rather formulate in lines of 'Studies/Observations suggest that temperature trends were smaller and globally less consistent. . . '

This is based on the cited studues (Neukom et al. 2019; PAGES 2k Consortium 2013, 2017) on the lack of evidence for globally consistent temperature trends over the last 2 millennia outside of recent warming in the industrial period. The suggested change to refer to 'studies' is made [I].

l.45: when mentioning the glacier length changes, could make link with observed changes from Leclercq et al. (2014) (which you use later, but reader does not know at this point) and Solomina et al. (2016) in which the literature on glacier geometry changes over the last 2000 years is summarized.

The earlier reference to introduce Leclercq et al. (2014) is added, and the new reference to Solomina et al. (2016) is now included.

l.45: 'we cannot compare': you as authors? Or the literature in general?

The literature in general. We replace the statement 'Due to the limits on the observational data for comparison, we cannot compare model results with an accurate representation of pre-industrial relative glacier lengths' with the more specific 'The small number of available length records which extend back further than 150-200 years (reference to new figure P2) heavily limits any possible comparison of model results with observed pre-industrial glaciers lengths'.

l.46-47: focus is on the transition from more stable pre-LIA to retreat. This is a complex matter – and many studies have tried to shed a light on this and came up with several possible mechanisms to explain the timing of this transition

(e.g. Painter et al., 2013; Lüthi, 2014; Sigl et al., 2018). Would be surprising that your relatively simple setup (with temperature and precipitation forcing only) is able to simulate the right timing (as it is generally known that retreat starts before a real increase in temperatures is observed). Would be good if you could provide a few words of explanation on this.

We include the suggested references on how the matter has a complexity that can cause significant challenges (and which serves as a 'hard' test of the model reproducing observed glacier timeseries features).

l.48-50: observations - any reference for this?

We add a reference here to the Leclercq et al. paper once again, and to the Solomina et al. paper suggested as an addition for the first comment about line 45.

l.50-52: agree, European glaciers are indeed not representative for worldwide glacier fluctuations. This is clear from recent Glacier Model Intercomparison Project (GlacierMIP), in which a strong contrasting behavior between the evolution of glaciers in various regions is highlighted (Hock et al., 2019).

The mentioned reference is added.

l.68: '. . . comparisons of between models and differences between regions': which models are you referring to here? Glacier and/or climate models?

The statement is modified to '...comparisons between runs driven by diferent GCMs and differences between regions' to make this clear [I]. There are several instances of ambiguity in whether 'model' refers to a GCM or to OGGM, so every references to 'model', 'models', 'modelled' is checked to ensure it is explicit what is being referred to.

l.74-75: strange description for OGGM: model for glacier dynamics that accounts for geometry and ice dynamics. Ok, but is also really a model in which SMB is coupled to glacier dynamics to simulate the temporal evolution of glaciers. Would suggest already mentioning the SMB here, and giving more information about the SMB in general in the following sentences, as this is the main driver for your changes over the past centuries... (see general comment 1b)

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].

l.81-84: you use an uncalibrated version of the OGGM model. What are the implications for the modelled glacier geometries at present-day (i.e. after several centuries of transient run): do you end up having a realistic glacier shape? Would be surprising that this can be obtained without any calibration and by just taking the model as is: see also general comment 1a: would be good if

you could show the modelled present-day geometry (after transient run) vs. observed (and thus not only rely on relative changes).

OGGM is used with its default calibration, which finds sensitivity values for SMB based on CRU 20th century climate data and a set of observed glacier SMB values. It is however not calibrated in any way to the data used in this experiment - either to find sensitivity values which are specific to each of the GCM climate datasets used or to match the lengths from the Leclercq data at any point. See preceeding point for the information added on OGGM processes that makes this explicit. See end of response to general comment 1a for comparison of modern observations and end-of-run state of (a specific aspect of) glacier geometry.

Would need information about pre-frontal area and how you treat glacier changes here. See general comment 3 for more information

See general comment 3 for detail.

l.112-117: in your explanation, you link the glacier response time to its size. Is however not the case for many cases / regions (Raper & Braithwaite, 2009; Oerlemans, 2012; Zekollari et al., 2020), and quite often the main driving mechanism for the glacier response time is the surface slope. Could simply reformulate this by saying that the glaciers you consider are typically large and relatively gently sloping glaciers and that these may not be representative for all glaciers in the region when it comes to their response time, as this is driven by a combination of glacier-specific factors.

This is a good reason to diversify the explanation for response time expectations. We include the suggested references and adopt the idea of discussing the likely gentler slopes in a set of glaciers with a distribution which trends much larger than the distribution of glaciers recorded in the RGI as a whole (new figure P1 [I]).

l.118-132: OK to have regional values and relative changes, but also need to show your results in absolute values and for individual glaciers. Does not require additional simulations, just a different and elucidating way of looking at results. See comment 1a for more info.

See response to general comment 1a for list of changes made to provide additonal results that address the need for absolute values and individual glacier data.

l.139: the area given in the Leclercq dataset: area at which time period? Guess this depends on the region/glacier considered? Would be good if could give indication.

The Leclercq reference date for area varies and is not always clear. Typically we expect the year to be beyond 1950 - given that every glacier had a record of length observations which extends beyond 1950 - and to be more likely to bias later than this where the area measurement is a product of larger-scale earth observation techniques. The criteria we use are developed with the expectation that there can be 50 years difference between the Leclercq area observation date and the RGIv6 observation date.

We now make it explicit that the year for the Leclercq area measurement is variable (and not fixed to 1950 like the Leclercq reference length), and add the sentence on the expectations for the criteria [I].

l.146: 'time of the Leclercq measurement': when is this?

This refers to the same time as the preceeding point. This is now clarified as 'the year of the Leclercq area measurement' here, and specificity is added about this year as described above.

l.166-168: you explain that sometimes not fully equilibrated after 300-yr spin-up: if this is the case, why do you simply not consider a longer spinup (¿ glacier response time) of e.g. 1000 years? Seems that in any case an initial adjustment will occur, because glaciers are never entirely in steady state, and you use this as a starting point. This is OK, but could reformulate this.

This explanation was not sufficient, so we revise it to the following:

'In cases where glaciers are still undergoing significant adjustment to a new equilibrium (e.g. in region 14) after several hundred years of spin-up and the early part of the main run, this is good evidence that in a 1000 year period, responses to trends in the forcing climate variables may not actually be shown in the OGGM output. This does not invalidate the glacier model output, but the evidence of continuing adjustment leftover from the spin-up being shown in the output rather than being removed with an arbitrarily long spin-up might inform the interpretation of the rest of the timeseries. It is also the case that where continued adjustment is significant after several hundred years, the magnitude of the length changes is typically large, and in these cases adding additional spin-up centuries will not fix the fact that the modelled glacier is diverging from the size of the observed glacier. We choose to maintain the 300-year spin-up for the sake of consistency as well as these reasons.' [I]

l.180: 'models underestimate the retreat shown in the observations': how come? Maybe also role for other factors not accounted for in your model: e.g. role of aerosols (global dimming) and other mechanisms?

It is not possible to quantify the potential effect of unrepresented processes in OGGM from the information in our model runs, but we do now discuss model limitations at length in the 'extensions and limitations' discussion section, which is referenced wherever we discuss discrepancies between observed and modelled length changes which are outside the range of results provided by the 6 GCMs.

l.187: '. . . use of normalised glacier lengths removes the ability to tell which. . . ': well, you can simply also additionally give your results as non-normalised glacier lengths. Need to do this to increase insights in your results and capability of OGGM to reproduce the present-day glaciers (after multi-century transient run): see general comment 1a.

See response to general comment 1a for comprehensive comment on absolute and normalised glacier lengths.

l.213-215: 'This suggests that differences in total post-industrial retreat are more influenced by differences in when the retreat starts than by. . . ': OK, from the modelling perspective. Is this also the case when considering the observations?

The intention here was to talk about differences between GCMs (for the same region), so with one observational timeseries for the region, there isn't a comparison between observations to make. We update this to refer specifically to '...differences in total recent retreat between GCM runs for a region are...' to make sure this is clear.

l.223: '. . . we do not examine the patterns of pre-industrial length change on a per-region basis': why not? Would be interesting to examine this and this aspect would add novelty to the paper. In the end, the manuscript almost solely focuses on post-industrial time period (although not clear if modelled absolute glacier length changes are realistic, see comment 1a) and when the retreat starts here, although the simulations cover a millennial timescale.

Ultimately it is matter of a lack of data for comparison. It would be possible to compare models to each other, but there are only a handful of pre-industrial length change observations available (see new figure P2), making comparison to observations largely impossible. This prevents any discussion of how well OGGM performs for each GCM in the pre-industrial period as it restricts commentary to whether different GCMs produce similar or different results, rather than how well they are able to reproduce observations. A more robust defence of this idea along these lines is added to the text.

l.226: Results shown only for 1 climate model: definitely ok, but would be good if you could argue / give a reason why IPSL is chosen (without this explanation this seems rather arbitrary).

A line is added on IPSL being chosen due to the apparent early start of recent retreat compared to others (with the expectation that this makes differences between full GCM runs and constant-climate runs more obvious) [I]. The results for all models are shown in the supplementary material, and relatively arbitrary choice is an inevitability unless we show 12 large figures for constant-climate runs, or condense the information heavily.

l.224-256: in general a lot of rather qualitative statements are made at the end of your result section, which does not leave the reader with real take-away messages: i.e. what should one remember when reading this section? Two suggestions for topics to focus on / include in your discussion:
- I think it is important to focus more on the SMB here and link this to other studies in which the SMB and its sensitivity to temperature and precipitation changes are described + show this for OGGM's SMB component used in this study (see general comment 1b + comment on l.74-75).

See responses to general comments 1a and 1b for full story on enhanced quantitative aspects and SMB.

What about the response time? This is not really mentioned in your story, but in the end this plays an important role, as the response time will explain

the lag between the change in SMB (the effect signal of your climate input) and the change in glacier geometry (the glacier length you consider for model evaluation). Would put more emphasize on response time in this section and e.g. explain inter-region and inter-glacier differences in the timing of the post-industrial retreat and how this can be linked to response time.

This point actually gets primarily referenced in the 'extensions and limitations' section of the new discussion, given the fact we know the response time does introduce an amount of lag to the responsiveness of length changes, but we cannot directly ascribe a single response time value to a glacier that is invariate through geometry changes over time and through different types and timescales of climate variation. It is now referenced as a potential explanatory factor, but the potential for introducing uncertainty is given more focus.

l.261: 'this observed retreat is within the range of the modelled retreats': difficult to judge as only relative changes are given and over regional scales. See comment 1a and related specific comments

See response to general comment 1a for comprehensive comment on choice of comparisons.

l.262: '. . . at least qualitatively capturing major trends in glacier length in many regions': mainly for postLIA, as this is what you focus on. With some additional analyses and results, as suggested throughout this review could change the 'at least qualitatively' to 'quantitively'

See response to general comment 1 for general material on quantitative analyses.

Comments on 'qualitative' vs 'quantitative' comparisons are comprehensively reviewed in light of the new analyses discussed elsewhere. Here specifically we also change 'major trends in glacier length during the period of observational record' as this is what actually determines the time period in which we can assess the modelled trends.

l.273: 'in almost all cases temperature is the dominant forcing' and l.276: 'suggests negative feedbacks between. . . on overall glacier geometry changes': indeed, as known from many other studies in which link SMB and climatic forcing is examined (comment 1b + related specific comments). Would make story stronger if you would also explore the link between climate forcing and SMB.

See response to general comment 1b for in-depth comment.

l.277: 'using dedicated glacier models': real need to have glacier model (vs. only considering the SMB): is to have the lag between SMB forcing and geometry change due to response time. By linking your story to response time (see last suggestion on the 'results' section): reinforce your story line and gives you an additional argument to support the use of OGGM.

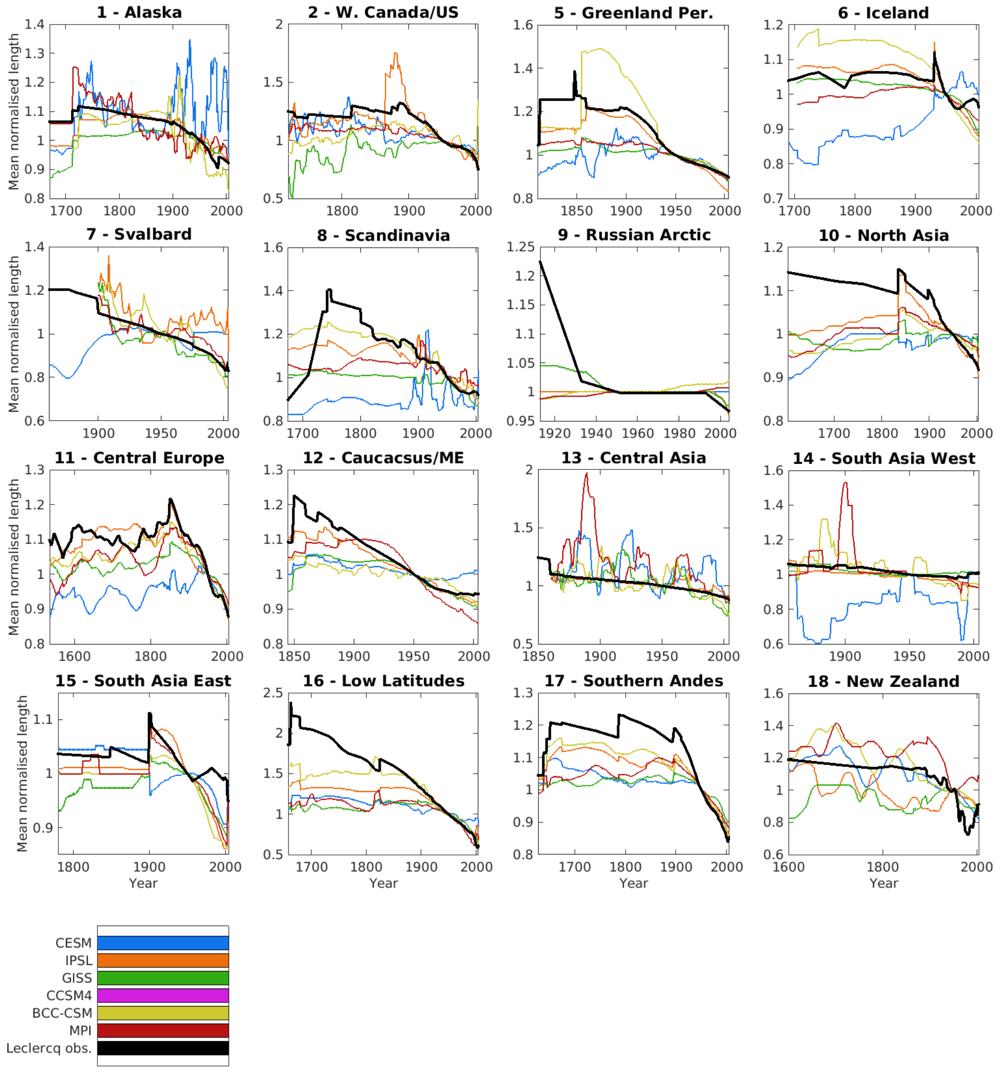
See response to 'What about response time?' above.

l.282-284: several challenges when considering all glaciers at regional scale.

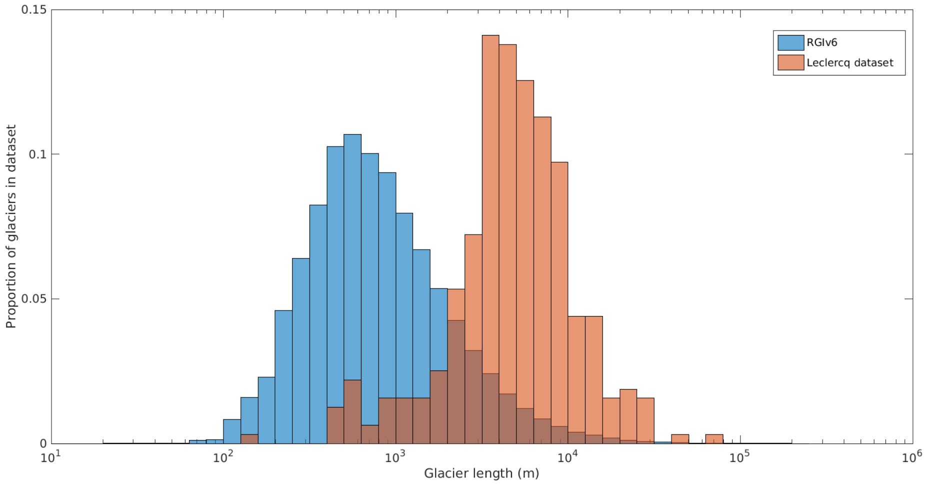
See general comment 5: would be good to at least mention some of the problems that will arise in such a case.

See response to general comment 5. The 'extensions and limitations' section goes into detail on what additional challenges there are in extending our approach to the scale of entire regions as defined by recent inventories.

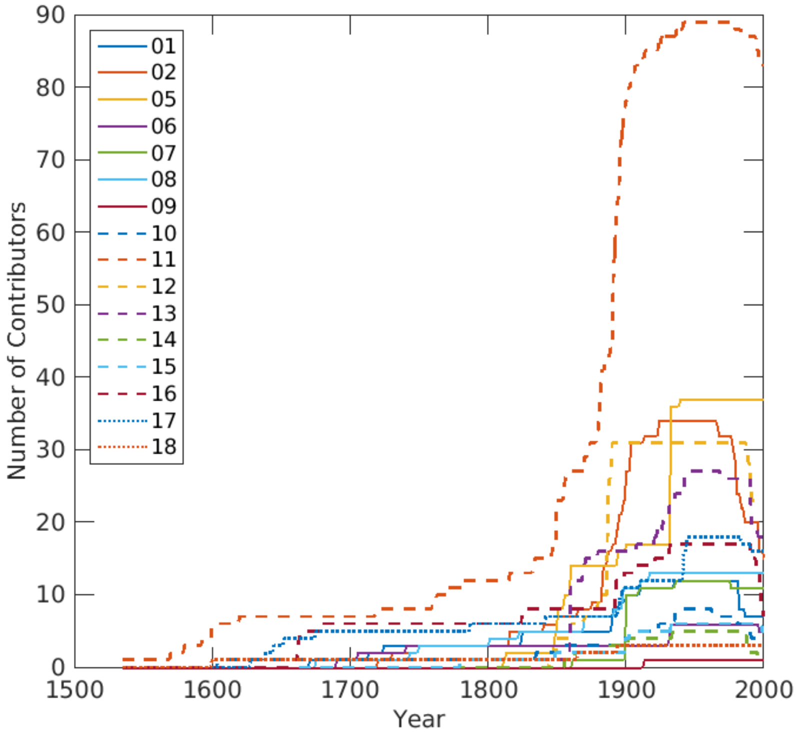
Revised figure 2: the set of modelled glaciers that contribute to the regional average over time now varies to match the set of glaciers that have Leclercq data available for any given year. The intention is to show the impact of the changing number of contributors to the regional means. Generally speaking,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

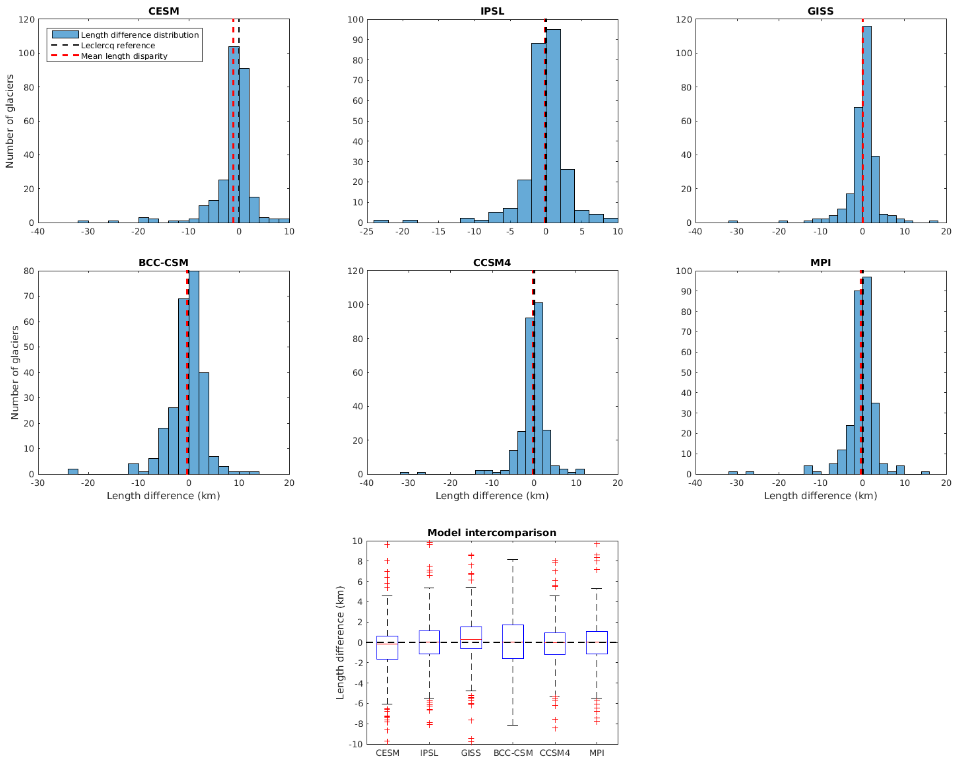


Proposed new figure P2: Changes to the number of glaciers which contribute to the Leclercq mean by year. This contextualises the potentially 'spiky' nature of the Leclercq averages; where there is a rapid jump in a particular region, it is possible that sudden changes in the mean glacier length in that 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)