Author Response to RC2

We thank the reviewer for her constructive feedback. Please find our detailed responses below.

Review of Shutkin et al. Modeling the impacts of climate trends and lake formation on the retreat of a tropical Andean glacier (1962-2020) submitted to The Cryosphere

Shutkin et al. present OGGM simulations of Queshque Glacier in the Peruvian Andes. The model is calibrated through comparison with measured mass balance from 1962-2008, with the dynamic glacier simulations run from 2008-2020. They run a range of simulations based on different temperature biases, which influence the other calibrated parameters, with model results compared against observed glacier velocities. They also show results obtained through maintaining a constant temperature or precipitation while allowing the other parameter to vary (to determine the drivers of historical changes), as well as a test comparing the influence of including frontal ablation on modelled glacier volumes.

Overall it is good to see a study such as this which is able to distinguish the long term drivers of glacier change, through these modelling scenarios. The results related to the temperature/precipitation changes are clear and interesting, as is the result on the impact of including frontal ablation. The correlations with ENSO are also quite instructive and add to the growing literature on the impact on ENSO conditions on glacier change in this region. In general the paper is clearly written and has an appropriate structure.

I do though have a couple of more major points to address:

1. Model validation/confirmation

In the paper I only see direct comparison of the models with the observed glacier velocity (Figure 5). However, this comparison is not clearly quantified aside from the comparison in the figure. As well as better quantification in terms of the glacier velocities, it would also be useful to add any other validation data that may be available. For instance, the calving front positions (which are shown in Figure 1a) could be added into Figure 6. It might also be good to compare with any mass balance measurements which are available, ideally from the field, but if not the dataset of Hugonnet et al. (2021) might also be beneficial. I think this would allow the authors to better determine which of their model runs is most likely to represent conditions correctly and give confidence that the parameters in that run are reasonable.

The reviewer's major comments categorized under "Model validation/confirmation" concern quantification of the surface velocity validation, the inclusion of additional mass balance and terminus position validation datasets, and an assessment of relative model strength across the parameterizations used throughout the paper. We address these concerns below.

a) Surface Velocity

We have quantified the error of modelled altitude band average surface velocity during 2018 against measurements (Millan et al., 2022) *derived from feature tracking on satellite imagery from the same period. The results suggest that in terms of reproducing velocity, models with temperature biases of*

-7.0 through -8.0˚C (models 2-4 in Table 2 of main text) perform with the highest accuracy. This is in visual agreement with Fig. 5 (main text). These results are presented in Table 1 below:

Model Number	*Model Temp Bias (˚C)*	*RMSE (ma^{-1})*	*MAE (ma^{-1})*
1	*-6.5*	*9.8*	*4.9*
2	*-7.0*	*3.5*	*2.5*
3	*-7.5*	*2.7*	*2.2*
4	*-8.0*	*3.8*	*3.2*
5	*-8.5*	*5.6*	*5.2*
6	*-9.0*	*7.6*	*6.9*

Table 1: RMSE and MAE values of elevation band flowline surface velocities versus observations from Millan et al. (2022).

b) Glacier Terminus Positions

Direct comparison of glacier terminus position (glacier length) to satellite imagery is complicated by the elevation band flowline method selected in this study. This method was chosen to facilitate comparison between observed and modeled surface velocities. However, the decision to use the elevation band flowline approach in OGGM limits the model's accuracy when it comes to glacier length (see OGGM documentation: https://docs.oggm.org/en/latest/flowlines.html#elevation-bands-flowlines).

Nonetheless, we have conducted the suggested analysis using Landsat 8, Sentinel 2, and historical aerial images (Fig. 1). To circumnavigate uncertainties surrounding model glacier length, we have instead opted to use the elevation of the 2008 DEM where it intersects with each terminus position shown in Fig. 1. This is a more reliable approach for aligning modeled and observed glacier positions, as the elevation band flowline is built around the 2008 DEM. We use the Zonal Statistics as Table tool in ArcGIS Pro to extract mean and standard deviation elevations from the 2008 DEM where it intersects with mapped termini. The resulting values are depicted in Fig. 2. Note, however, that these calculated elevations reflect the surface of the glacier in 2008, not the actual terminus altitude in a given year. For this reason, standard deviations for years prior to 2008 (ice-free in the 2008 DEM) are much greater than for years where ice is present in the DEM. This reflects the fact that glacier ice has lower surface roughness than its surroundings.

We then leverage the terminus surface altitudes to identify the point along the elevation band flowline corresponding to the terminus position at each year. Specifically, we query the elevation band flowline such that the glacier surface elevation in 2008 is within ±2 m of the calculated terminus elevation of a given year since 2008. This small range accounts for the fact that not every possible elevation is included as an index in the elevation band model. The results show close agreement between models and observations by 2020 (Fig. 3), though performance quality differs from year to year. In particular, it appears that our models exagerate frontal retreat rates by both 2018 and 2019 before slowing to match observations by 2020.

We hypothesize that the overestimation of retreat at certain years relates to the calving model's first order dependence on lake depth (see methods, main text), which we believe to be over-estimated in our model (see discussion, main text). This source of uncertainty may also lead to an overestimation of frontal retreat rates during certain years. It is notable that 2019-2020 marks the *transition to shallower modeled lake depths, accounting for the deceleration of retreat. We elaborate on this model limitation in our discussion section.*

Finally, although a calving rate parameter calibration could theoretically force each of our six models to match observed frontal retreat perfectly, doing so would eliminate the insight we gain through our calving experiments (as per RC2 comment on Line 412 and throughout). Rather, deploying a constant calving parameterization across models ensures that model performance is not the result of calibration and does indeed reflect the impact of the onset of frontal ablation. This is addressed more specifically in the line edits below.

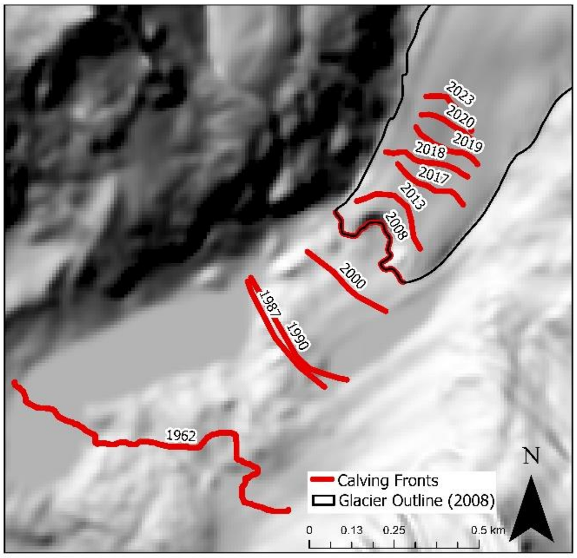

Fig. 1: The mapped glacier terminus positions overlaid on the 2008 DEM.

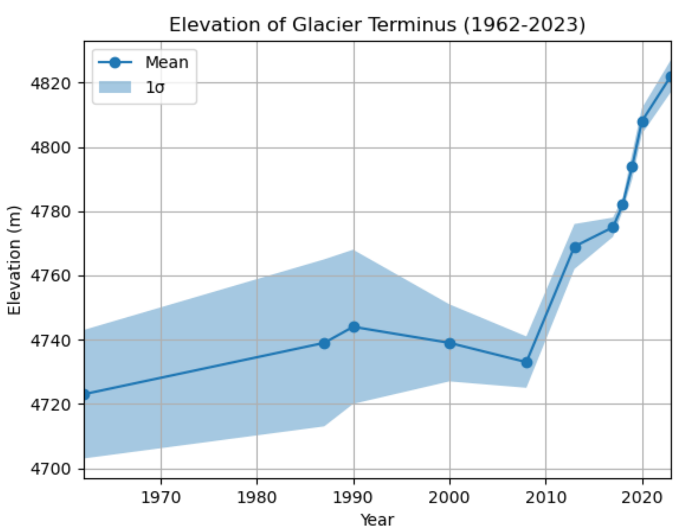

Fig 2: The glacier terminus elevations for each observation year derived from the map in Fig. 1 by averaging the DEM elevations where they overlap mapped glacier termini.

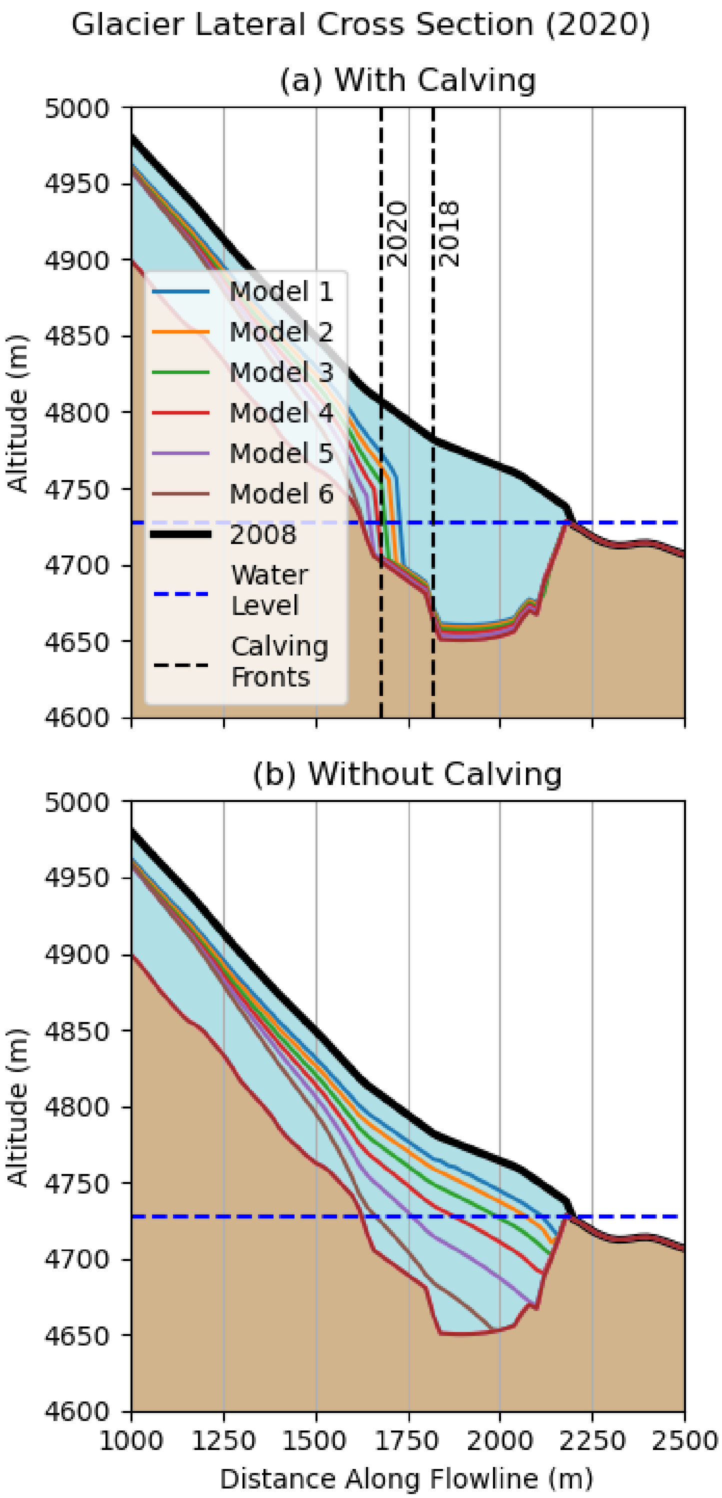

Fig 3: An update of Fig. 6 (main text) including 1) different symbology for observed (2008) vs. modeled glacier surfaces, 2) vertical lines representing mapped glacier terminus positions from 2018 and 2020, corresponding with the ice fronts mapped in Fig. 1 (main text).

c) Geodetic Mass Balance Observations

We have compared our ensemble mean specific mass balance (SMB) model outputs for each of the 42 epochs provided in the Hugonnet et al. (2021) study. We find that averaged across the years of a given epoch, our ensemble consistently overestimates mass loss as compared to the Hugonnet et al. study. In other words, our model results suggest that Queshque Glacier is retreating faster than the best estimate from the global geodetic mass balance study.

Nevertheless, our results are within the uncertainty bounds provided by Hugonnet et al. during all but 3 of the 42 epochs. All three of these epochs consider change as of 2020 including 2000-2020, 2015-2020, and 2016-2020. This suggests that a systematic bias exists either in the 2020 data used in taking the geodetic mass balance or in our simulation occurring around that time. Other epochs up to 10 years in duration (the second longest duration after the single 20-year measurement) show agreement between Hugonnet et al.'s and our own data. In summary, this comprehensive comparison indicates general agreement between the two datasets, bolstering confidence in our mass balance simulation.

d) In-Situ Ablation Measurements

We use ablation stake data measured by the National Water Authority of Peru between the years 2015 through 2019 to further evaluate our mass balance models and to compare the relative performance of individual parameterizations. This dataset was not available during our original analysis. The data comprise individual abaltion stake measurements spanning about 4700-5150 m in altitude that have been converted to water equivalence. Some measurements report altitudes occuring below the glacier terminus elevation in 2008. While lower altitudes in the stake data may be in part linked to glacier thinning, altitudes below our calving water level of 4727 m cannot be explained in this way. It appears, rather, that some level an innacuracy or negative bias exists in the altitude data. We therefore apply a uniform bias correction of 26 m across all elevations reported in the stake data such that the lowest stake measurment reaches 4727 m. We recognize that this correction is a source of considerable uncertainty, however, we determined it to be necesssary since we lack additional GPS metadata. Due to inconsistencies in the duration of the stake measurements, ablation measurments were converted to m w.e. d^{-1} then multiplied by 365 days to arrive at standard units of m w.e. a^{-1}. Having made these corrections and standardizations, we can then compare the observed ablation to our modeled mass balance profiles.

Two analyses of the abaltion measurments inform our evaluation. First, we condsider the magnitude of ablation in the lower altitudes (defined as 4800 m and below) of the ablation zone as a constriant on the realistic melt rates near the glacier terminus. We then consider the observed ablation gradient in comparison to our models.

i) Magnitude of Ablation

Observed annual melt rates below 4800 m range from about -11.4 to -3.5 m w.e. a^{-1}, averaging at -7.5 m w.e. a^{1}. Melt rate are greatest during the El Niño year of 2016 which is consistent with our simulation of overall specific mass balance. The range described above provides a limit on the magnitude of ablation we should expect to produce near the glacier terminus in our models. Average ablation rates at the lowest altitudes (4727-4800 m) during the years 2015 through 2019 are *presented in Table 2 below. We find that models 2, 3, 4, and 5 fall within the bounds of observations, with model 4 producing ablation rates closest to the observed mean. These results are generally consistent with our velocity validation, in which models 2-4 outperform the others and model 3 performs best (Table 1). We note that this is unsurprising, as the magnitude of ablation is related to overall ice flux, which controls surface velocity.*

Model Number	*Degree-Day Factor (mm w.e. d^{-1} °C^{-1})*	*Mean Terminus Ablation (m w.e. a^{-1})*
1	*1.74*	*-3.1*
2	*2.58*	*-4.2*
3	*3.87*	*-5.5*
4	*5.92*	*-7.3*
5	*9.31*	*-9.7*
6	*15.31*	*-13.0*

Table 2: Mean annual ablation produced by each model at 4727-4800 m during the years 2015-2019 which coincide with the timespan of in-situ mass balance measurments. The degree-day factors are included for reference.

ii) Ablation Gradient

While our models can reproduce the observed ablation rate at the glacier terminus, we find that their ability to reproduce the observed ablation gradient is limited. Fitting a linear trend to all negative stake observations, we calculate that on average ablation decreases (becomes less negative) by 8.5 mm w.e. d^{-1} for every 100 m in elevation. By dividing this value by the lapse rate of -0.65 ˚C per 100 m, we arrive at an estimated temperature sensitivity (positive degree-day factor) of approximately 13 mm w.e. d^{-1} ˚C^{-1}. Based on the maximum lapse rate seasonality identified by Hellström et al. (2017)*, we note that this value could in reality range between 9.3 and 14.6. However, our model mass balance calibration was performed under the assumption of seasonally consistent lapse rate and we therefore adopted the conventional value.*

In order to ensure that our model matches the observed ablation gradient, we recalibrated the model by adjusting the temperature bias to fit the observed geodetic mass balance using a fixed temperature sensitivity parameter (DDF) of 13 mm w.e. d^{-1} ˚C^{-1}. This DDF falls between models 5 and 6 of the submitted manuscript (Table 2) . We find that this calibration overestimates the ablation rate at the glacier terminus. We next introduce an additional temperature bias, cooling the model until it approximates the observed average mass balance profile in both magnitude of accumulation/ablation and gradient. We find, however, that this model vastly overestimates the specific mass balance and would indeed induce glacier growth since 1962. To further investigate the threshold between glacier growth and retreat, we conduct a sensitivty experiment wherein the fixed-gradient model is cooled until balanced conditions are achieved between 1962 and 2008. The results indicate that all stake observations except from the extreme El Niño year of 2016 show a more positive mass balance than would be required for balanced conditions (Fig. 4).

This experiment highlights a fundumental limitation of our model, which is that we cannot simultaneously fit the observed magnitude of ablation and the total observed mass change across the glacier. However, various model assumptions may be able to explain this discrepancy. For example, the assumption of perfect continuity (that all mass in the accumultation zone contributes *to the ablation zone) which is inherrent in OGGM may result in an overestimation of the true contributing accumulation area. This would in turn require more intensive ablation to compensate for the inflated accumulation, resulting in a model that reproduces the gradient but not magnitude of ablation (i.e., blue curve in Fig. 4a).*

Alternativly, one could correct for overestimated contributing area by raising the equilibrium line altitude thereby reducing the accumulation area. This forces lower ablation gradients, as relatively high ablation rates persist at higher altitudes. This strategy is represented in our models that fit the magnitude of observed ablation without matching the observed ablation gradient (i.e., models 2-5, Table 2).

In summary, we have compared our mass balance models against total mass change from 1962-2008, the magnitude of ablation during 2015-2019, and the ablation gradient during the same period. The models used in the submitted manuscript fit the first two metrics while missing the third. We find that it is impossible within our model to fit both the first and third, and therefore conclude that we have chosen adequate validation metrics. This is further supported by our surface velocity and terminus position validations. Both the in-situ ablation validation and the surface velocity validation suggest that our model numbers 2-4 produce the most reliable output. The frontal position mapping shows closest agreement with models 3 and 4.

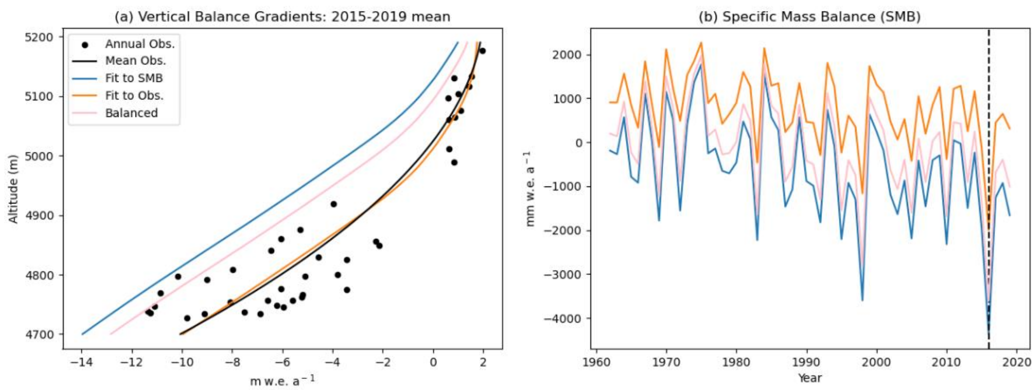

Fig. 4: (a) The observed and modeled vertical balance gradients (mass balance profiles) fit to the observed SMB (blue), the observed mass balance profile (orange), and to a glacier in long-term (1962-2008) climatic equilibrium (pink). (b) Annual SMB from the same same models, 1962-2020. The dashed line represents the strong El-Niño year of 2016, which is the only year that observed ablation rates (points left of pink curve in Fig. 4a) would produce negative mass balance across the glacier.

2. Calibrated parameters and their effect

The calibration process is quite clearly described, but it is a little odd that the temperature bias range then effects the DDF and potentially also the creep parameter (this second point should be made clearer in the paper). This effect makes sense from a calibration perspective, since the other parameters are compensating for the temperature bias to ensure the modelled mass loss matches that observed, but it has quite strong impacts on the other parameters. Note also that the precipitation lapse rate is the same in all models – if it was allowed to vary it might influence the change in the DDF, since the accumulation could change as well as ablation. There might also be impacts on the modelled frontal ablation rates, but I am not completely sure. I think the section on mass balance calibration should make clear the resulting influence of the temperature bias on all the parameters.

We agree that we should make the connections between parameters more clear in the text and that impacts of the temperature bias on other parameters are the result of calibration. In the case of flow model parameters, we calibrate against radar (thickenss) data. The A parameter therefore must be adjusted to ensure that thickness matches the observations despite different ablation rates across models. The ablation rates are themeselves a product of the DDF, which is strongly influenced by the temperature bias.

The precipitation factor is indeed a free parameter in our mass balance model. However, it is calibrated subsequently to the DDF only if the model fails to converge in the previous step. This was never the case, so our defualt value of 2.5 was used across all models. We accept that this parameter is highly uncertain as very few reliable records of accumulation exist from the tropical Andes. Our direct, though limited in duration, mass balance measurements discussed above record accumulation as high as 1.95 m w.e. a^{-1} at 5150 m on Queshque Glacier. Using the precipitation factor of 2.5, our mass balance models produce average annual accumulation rates of 1.9±0.4 m w.e. a^{-1}. This value is consistent with the limited direct accumulation measurements we have available. In contrast, the nearby Huascarán Col experiences lower annual accumumulation of about 1.4 m w.e. a^{-1}(Thompson et al., 1995; Weber et al., 2023)*. Although the defualt precipitation factor of 2.5 does seem to produce realistic accumulation values for Queshque, we note that changes to this parameter do not introduce bias into the mass balance model so long as the DDF is recalibrated, as we have done in our study* (Maussion et al., 2019).

Given this, it would be useful also if the authors could compare the calibrated DDF and creep parameters with those from the literature (ideally from this region, or similar glaciers). This would help to understand to what extent the calibrated parameters are reasonable. If DDFs themselves are not available then comparison with modelled melt rates would also be useful (which are available, as ablation is shown in Figure S5).

We will include a section elaborating on our DDFs as compared to previously published literature.

At Zongo glacier (16˚S), Fuchs et al. (2013) *use DDF values of 6.5 and 30 mm w.e. d^{-1} ˚C^{-1} to model observed glacial discharge for the dry and wet seasons, respectively. Averaged over an entire hydrological year, these values are somewhat consistent with the DDF calculated above using stake observations and support the higher sensitivty models used in our study.*

Fyffe et al. (2021) *find that on glaciers in the Cordillera Blanca, 5˚C warming induces a melt increases from 0.75-1.25 mm w.e. h^{-1} (estimated from Fyffe et al. 2021 Fig. 8). This translates to 3.6-6.0 mm w.e. d^{-1} ˚C^{-1}, matching the sensitivity of our models quite well.*

Relating to this the authors should be clear to not mistake the effects of calibration (so the differences between the models with different temperature biases) for processes related to more or less temperature sensitive glaciers (e.g. in the discussion section 5.4 and conclusion).

We understand the comment about conflation of effects of calibration for processes related to temperature sensitivity to be in reference to lines 408-414. In this paragraph, our aim is to highlight the effect of calving on variability across models. We observe that when the same calving parametirization is included across models, variability in mass loss is reduced. This is not the impact of calibration and we feel it is therefore valid to conclude that the lake calving process does indeed reduce the effects that differing temperature sensitivities have on relative mass loss rates.

We will alter the wording in the concludion (line 444) from, "...with otherwise different climatic sensitivites," to, "...with otherwise different reponses to climate warming."

One idea would be to add text in the discussion about the calibration processes, including how reasonable the calibrated parameters are, as well as how the use of these parameters (including the constant ones) may influence the results (for instance the temperature versus precipitation sensitivity of the mass balances and the influence of ENSO).

We agree that the discussion would benefit from a dedicated section comparing the calibrated parameters to data and previous literature. It would also be beneficial to discuss model sensitivities to parameterization decisions, including the constant parameters. In addition to what was elaborated upon above, we will note that although the precipitation factor does not introduce model bias, it may alter model sensitivity to temperature versus precipitaiton changes, with higher precipitation factors leading to more precipitation sensitive glaciers (Maussion et al., 2019).

I also include my minor comments below, which I hope are useful.

Yes, thank you; we appreciate the helpful edits.

With kind regards,

Catriona Fyffe

Minor comments

L21: 'precipitation amounts'

Accepted

L40 'other challenges'

Accepted

L64-73 – consider including the ETI methods you mention later also in this paragraph

Accepted

Figure 1a. Is the 2008 outline only for the Queshque Glacier, and not the other glaciers that were previously attached to it? If so maybe make this clearer in the caption.

Yes. We do so to indicate the extent of our model domain. We will clarify this in the figure caption.

L126 It might be nice to show the GPR bed topography in a little more detail. It is shown in Figure 1 but it's not so easy to see, and perhaps a specific figure in the SI would be useful.

We will include the following figure (Fig. 5) in the supplement.

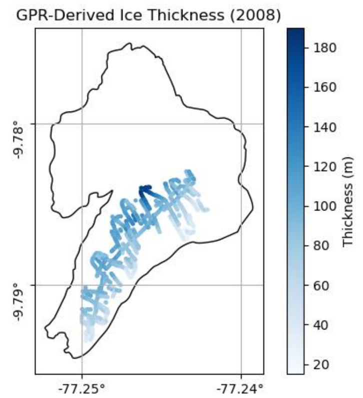

Fig. 5: Map of the GPR data used to calibrate the OGGM flow model. Though the GPR was taken in 2014, ice thicknesses in 2008 were derived by subtracting the GPR bed topography from the 2008 lidar DEM (see methods).

L131 It would be useful to add here the number of point observations you used for validation.

Only 17 georeferenced point measurements were provided by the 2009 GPR survey report. As elaborated upon in the main text, the points show general consistency with the subsequent 2014 survey (Fig. 6a). The 2009 GPR points span from the bottom of the glacier in the southwest towards the center of the glacier in the northeast of Fig. 6b. The points are located approximately along the centerline of the glacier and are each in proximity to multiple measurements from the subsequent GPR survey. There is no apparent relationship between the mean distance from the 2009 data points to their respective nearest neighbors and the resulting difference between measured and derived thickness. There does, however, appear to be a slight spatial bias, with derived thicknesses being more likely to underestimate the 2009 measurements at lower elevations. The significant outlier where the derived thickness is approximately 28 m thinner than observed occurs at a discontinuity in the 2009 survey, suggesting that the technician may have needed to navigate an obstacle which may have produced abnormalities in the ice thickness profile or potentially caused an error in measurement or radargram interpretation.

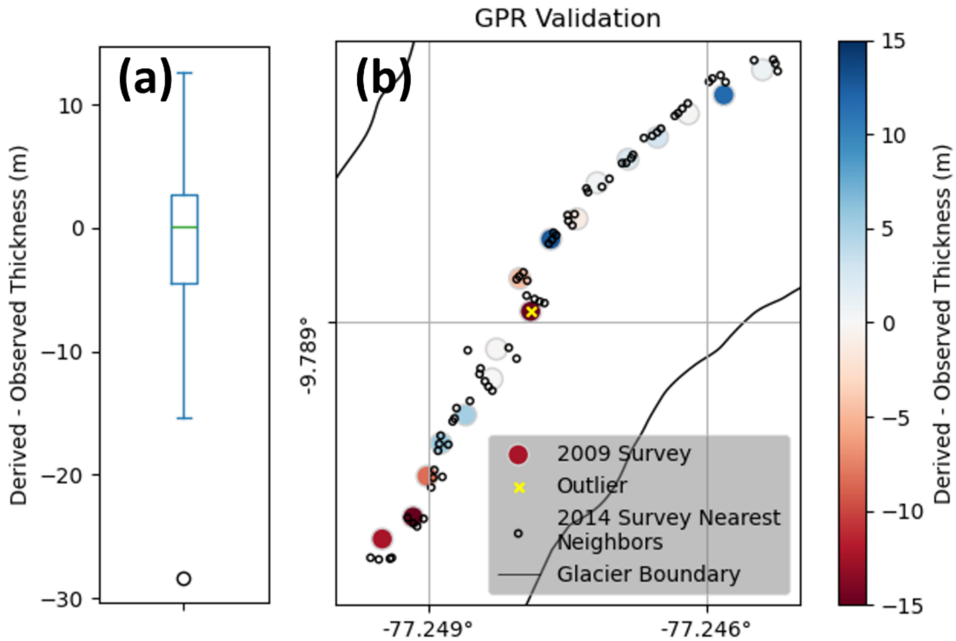

Fig. 6: To be included as an additional supplementary figure. (a) Box and whiskers plot showing the distribution of error between the derived and observed thickness values, including the position of the single negative outlier. (b) A map of the 2009 GPR survey showing the error calculated at each point.

L168 Where did you get this range of temperature biases from? It seemed that the cold bias was identified from the comparison of CRU and PISCO – was this the source?

The temperature bias range was determined by the simple requirement that temperatures needed to fall at or below freezing for a significant duration within the altitude range of the glacier. We will clarify this point in the text.

Section 3.4.2 It might be a good idea to have a sentence at the beginning of this section which summarises the approach here. This is just because the need to calculate initial ice thicknesses is unclear at the beginning since you have the measured GPR data, but then you mention later (line 190) that the GPR data is used for calibration.

We will add the following sentence at the beginning of section 3.4.2: "The GPR dataset is leveraged to calibrate the ice flow model by minimizing error between modeled and observed ice thickness."

L185 Consider adding a reference for this statement

Here we follow Pelto et al. (2020).

L189 Why did you choose a minimum slope parameter of 7.5 (and what is the unit of this parameter)? Do you have a reference for this? The sentence above about exaggerated overdeepenings would also benefit from a reference.

The correct unit for the minimum slope parameter is degrees. The text should therefore read "7.5˚". This parameter clips the glacier surface slope to ≥7.5˚ during the ice thickness inversion process, resulting in the flattest sections of ice retaining higher flow rates. This is necessary because otherwise the model will overestimate ice thickness to satisfy ice flux continuity with the steeper slopes above. The value of 7.5˚ was accepted after a series of sensitivity experiments examining the impact of this parameter on terminal ice thickness (eventual lake depth). However, as elaborated upon in the discussion section of the main text, we still suspect that terminus ice thickness (lake depth) is over estimated. This parameter is only discussed briefly in the OGGM documentation (https://github.com/OGGM/oggm/blob/master/oggm/core/inversion.py).

Table 1. The melt threshold of -1˚C may be a little cold for tropical glaciers, as there tends to be strong diurnal temperature gradient so it takes some energy in the morning to warm the ice/snow surface to melting. Pellicciotti et al. (2008) mentions this general idea, although they used a 1˚C temperature threshold and only calibrated the TF and SRF factors (but TF became negative, they say due to this effect, see Figure 12). My general point is that 'Alpine' ETI/TI parameters may not work as well in the Andes. I imagine in reality the calibrated TI factor will compensate for this, but it might be worth mentioning somewhere in the discussion.

This is an excellent point that we have elaborated upon as a limitation in the discussion section of the manuscript. We agree, ultimately, that the DDF calibration should compensate for other parameter uncertainties.

Section 3.5 Just a suggestion – but it might be useful to name each of your experiments so you can easily refer back to them later.

We appreciate this suggestion but feel it is easier for the reader to describe the experiments (e.g., "constant climatological mean temperature") rather than returning to an index with experiment names.

Lines 218-224 Although it is fairly well explained, I think you could be even more explicit in explaining how the changes in the temperature bias go through to influence the model sensitivity. For instance instead of saying 'low magnitude εT' you could write 'less negative' (and vice versa), and also say that under these conditions the calibrated DDF is lower, meaning melt is less sensitive to air temperature. It might have also helped to directly compare the PISCO data with local weather station information and use this for a bias correction. Was there a reason why this wasn't considered? You should also mention that the creep parameters are also influenced by the temperature biases – so that you have higher creep values under the most negative temperature biases.

We accept these recommendations regarding being more explicit about relations between calibrated parameters in this paragraph. We will also add discussion here regarding the precipitation factor and creep parameter, both of which are elaborated upon above in this author response.

We attempted to locate meteorological station data from the Queshque valley, but were unsuccessful, as mentioned in lines 224-225.

Table 2: You have quite a large range of calibrated DDFs due to the range in the temperature biases. Which of these is more reasonable compared to calibrated degree day models from the literature? This might help you to work out which of the models is more sensible and would give a melt response to temperature that is reasonable for the region. This might be something worth adding to the discussion.

See response to major comments.

Figure 2 I am not convinced that showing these correlations as bar graphs is necessarily the best approach. Can you make clear in the caption which model was used for the correlations? I presume the ensemble mean?

Yes, these correlations refer to the ensemble mean values. We will clarify this point in the caption. We opted for the bar graphs over other visualization methods such as correlation heatmaps because we felt this approach was the most intuitive.

Lines 257-259 The percentages here for the wet and dry season accumulation and ablation don't add up to 100%, are you in reality speaking about DJF and JJA only?

This is exactly the case. We have excluded the shoulder seasons to highlight the largest seasonal difference. We have clarified this in the text.

Figure 3 (and related Figure S5) Why not show the ensemble mean as well? Especially as it is likely more reasonable, given that Figure 5 suggests that model 3 (in the middle) is likely the most reasonable.

We feel that Figures 3 and S5 (main text & supplement) illustrate differences between models in a way that is informative to the reader. In particular, they demonstrate that all model showcase the same seasonality albeit with different magnitudes of variability. We feel that presenting mean values instead would showcase the seasonality without illustrating the later point about our model ensemble.

L277 It would be useful to show a comparison of the ice thicknesses of the different models, especially since there might be quite some differences in the accumulation zone which are not shown anywhere.

As shown in Table 3 (main text), initial thicknesses across the ablation zone where GPR data are available are stable across models. The same table shows, however, that initial ice volumes range from 7.74 x 10^7 m^3 to 8.10 x 10^7 m^3. This is primarily due to differences in the accumulation zone ice thickness.

L282 Are the real-time monthly conditions from PISCO? If so please add this for clarity.

Yes. We will rephrase to clarify that we mean the contemporaneous PISCO climate conditions as opposed to the 1962-1992 climatological means (derived from CRU and PISCO).

L285 Please give the steady state ensemble mean volume

This refers to the ensemble mean initial (2008) volume of 7.9±0.1 x10^7 m^3 (1σ uncertainty) of ice (from Table 3, main text).

L287 Give here exactly the increase in the mass loss caused by implementing the frontal ablation scheme, especially since it is a key result of the paper.

By 2020, the ensemble mean of models with calving shows additional mass loss of approximately $0.4x10^7$ m^3 as compared to models excluding the calving process.

L292-293 Can you quantify the comparison of the models with the glacier velocity measurements, even only upstream of the modelled calving front? It also looks from Figure 5 like model 3 has the best correspondence with the observations, but ideally you should quantify this and say it exactly.

See response to major comments.

Figure 4 Explain in the caption what the solid lines represent, I imagine the ensemble mean?

Yes, the solid lines are the ensemble mean values. We will clarify this in the figure caption.

Figure 6 perhaps use a different line type to differentiate the observations from the model results. Also, are there any data for 2020 which help you to determine which model is most correct? I see in Figure 1 that there are calving front elevations determined in 2018 and 2020, it would make sense to add these here to understand if any of the models are able to replicate them.

Accepted. See response to major comments.

L321 I am not sure I would use the word 'peaks' here to describe the higher wet rather than dry season, as I think you are talking about the wet season in general. The highest melt rates (for Peru at least) tend to be at the end of the dry season/beginning of the wet season anyway.

Accepted. We will change the wording from "peaks" to "increases."

L322 and L338 The exception for Shallap is likely only the case in the modelled time period in Fyffe et al. (2021) – potentially due to strong La Nina conditions. It might not generally be the case, and I would not in general say Shallap has an atypical ablation seasonality. For this reason please adjust the two sentences related to this. Gurgiser et al. (2013) found mass balances to vary between years precisely because of differences in the snow cover over the ablation zone.

Thank you for this comment. We will clarify that multiple studies of Shallap glacier have identified exceptional ablation seasonality during various years, which has been linked to the timing and phase of precipitation.

L329 'excluded by the nature'

Accepted.

Section 5.3 Just a suggestion but you could mention Lamantia et al. (2024) https://doi.org/10.5194/tc-18-4633-2024

Agreed. We will add discussion of Lamantia et al. (2024) in this section, as this paper also supports our observation of high correlation between wet season ENSO indices (ONI, SOI, and Niño-3.4) and annual specific mass balance.

L382 'the El Niño/warm'

Accepted.

L386 Be specific about which ENSO indices correlate with the SMB anomaly

Here we refer to annual or dry-season indices including ONI, Niño-3.4, and SOI. These are then contrasted with wet-season specific metrics.

L402 '2008 through to 2020' (although maybe it's ok in US English)

We believe this is standard US English but will defer to the editor's preferences.

L402 The terminology 'least sensitive calving glaciers' is a bit odd as you are really speaking about specific model runs. Consider something like 'the mass loss from the most temperature sensitive model including calving' and include the model number so it is clear which one you mean.

This statement refers to model number 1 (run with calving) surpassing the ensemble mean mass loss for all models that excluded calving. We will clarify this in the text.

L412-412 It is an odd way to talk about the model results, as it is phrased as if the glaciers themselves are accelerating frontal ablation to counteract their lower sensitivity. Instead, the calibration process is resulting in higher frontal ablation rates in order to compensate for low TF parameter values (or at least I think this is what is happening). It is very important here that you are clear what is an effect of calibration compared to what is a real glacier response.

We appreciate the reviewer bringing up this point of confusion regarding our calibration process. As frontal ablation was not a significant process throughout most of our calibration period (1962-2008), we initialize the model glacier as a land-terminating type, which is unaffected by calving. As described in section 2 (main text), the onset of frontal ablation occurred only after 2008 due to complex processes concerning the surrounding hydrology and sub-glacial topography.

Models 1-6 are therefore calibrated without respect to calving. The subsequent experiments comparing calving and non-calving glaciers (section 3.5, main text) therefore begin from model glaciers with differing temperature sensitivity parameters (DDFs). We show that in non-calving scenarios, these glaciers evolve quite distinctly, and with greater variance, than in calving scenarios. In other words, we have isolated the impact of calving on modeled glacier evolution. This is not the result of any sort of calibration procedure, as all models use the same calibration parameterization.

We will be sure to clarify this point in the text and to explicitly emphasize that the calving process is not a factor during the mass balance calibration procedure.

L420-429 – you already have a few other sources of data, but just to make you aware of the HydroLake database, I am not sure if it would be helpful or not (the depths are also estimated), but you could look at the specific depth to area ratio for lakes only in this region. As I mentioned above it might also be useful to compare with observations of the calving front, since you mention observations form 2018 and 2020, and I imagine you could find more from satellite information. https://www.hydrosheds.org/products/hydrolakes

We appreciate learning of this additional dataset, but feel that the two datasets already discussed are sufficient for this section. We will consider the HydroLake database in future research.

See response to major comments regarding calving front positions.

L438 The phrase 'out performs theoretical expectations' is a bit odd and not precise. Perhaps add here specific comparison against data (e.g. the best models ability to replicate the glacier velocity or calving front location).

We will change this phrase to refer to the specific validation metrics discussed in the main text and above in our response to major comments concerning mass balance validation.

L444 The phrase 'reduces the variability between glaciers with otherwise different climatic sensitivities' is perhaps misleading, since this variability of the calving front rates related to the models with varying TF values could be due to the calibration process rather than a 'real' glacier response.

See comments concerning this point above.

References

Hugonnet, R., McNabb, R., Berthier, E. et al. (2021) Accelerated global glacier mass loss in the early twenty-first century. *Nature* 592, 726–731, https://doi.org/10.1038/s41586-021-03436-z

Lamantia, K. A., Larocca, L. J., Thompson, L. G., and Mark, B. G. (2024) El Niño enhances snow-line rise and ice loss on the Quelccaya Ice Cap, Peru, *The Cryosphere*, 18, 4633–4644, https://doi.org/10.5194/tc-18-4633-2024

Pellicciotti, F., Helbing, J., Rivera, A., Favier, V., Corripio, J., Araos, J., Sicart, J-E., Carenzo, M. (2008) A study of the energy balance and melt regime on Juncal Norte Glacier, semi-arid Andes of central Chile, using melt models of different complexity, *Hydrological Processes*, 22 (19) 3980-3997

Additional References Cited in Author Response

Fuchs, P., Asaoka, Y., & Kazama, S. (2013). Estimation of Glacier Melt in the Tropical Zongo with an Enhanced Temperature-Index Model. *土木学会論文集b1（水工学）*, *69*(4), I_187-I_192. https://doi.org/10.2208/jscejhe.69.I_187

Fyffe, C. L., Potter, E., Fugger, S., Orr, A., Fatichi, S., Loarte, E., Medina, K., Hellström, R. Å., Bernat, M., Aubry-Wake, C., Gurgiser, W., Perry, L. B., Suarez, W., Quincey, D. J., & Pellicciotti, F. (2021). The Energy and Mass Balance of Peruvian Glaciers. *Journal of Geophysical Research: Atmospheres*, *126*(23), e2021JD034911. https://doi.org/10.1029/2021JD034911

Hellström, R. Å., Fernández, A., Mark, B. G., Michael Covert, J., Rapre, A. C., & Gomez, R. J. (2017). Incorporating Autonomous Sensors and Climate Modeling to Gain Insight into Seasonal Hydrometeorological Processes within a Tropical Glacierized Valley. *Annals of the American Association of Geographers*, *107*(2), 260–273. https://doi.org/10.1080/24694452.2016.1232615

Maussion, F., Butenko, A., Champollion, N., Dusch, M., Eis, J., Fourteau, K., Gregor, P., Jarosch, A. H., Landmann, J., Oesterle, F., Recinos, B., Rothenpieler, T., Vlug, A., Wild, C. T., & Marzeion, B. (2019). The Open Global Glacier Model (OGGM) v1.1. *Geoscientific Model Development*, *12*(3), 909–931. https://doi.org/10.5194/gmd-12-909-2019

Millan, R., Mouginot, J., Rabatel, A., & Morlighem, M. (2022). Ice velocity and thickness of the world's glaciers. *Nature Geoscience*, *15*(2), Article 2. https://doi.org/10.1038/s41561-021-00885-z

Pelto, B. M., Maussion, F., Menounos, B., Radić, V., & Zeuner, M. (2020). Bias-corrected estimates of glacier thickness in the Columbia River Basin, Canada. *Journal of Glaciology*, 1–13. https://doi.org/10.1017/jog.2020.75

Thompson, L. G., Mosleythompson, E., Davis, M. E., Lin, P. N., Henderson, K. A., Coledai, J., Bolzan, J. F., & Liu, K. B. (1995). Late-Glacial Stage and Holocene Tropical Ice Core Records From Huascaran, Peru. *Science*, *269*(5220), 46–50.

Weber, A. M., Thompson, L. G., Davis, M., Mosley-Thompson, E., Beaudon, E., Kenny, D., Lin, P.-N., & Sierra-Hernández, R. (2023). Drivers of δ18O Variability Preserved in Ice Cores From Earth's Highest Tropical Mountain. *Journal of Geophysical Research: Atmospheres*, *128*(19), e2023JD039006. https://doi.org/10.1029/2023JD039006