Reviewer 1 – Ben Marzeion

General comment

[RC1.01] Zekollari and colleagues present projections for glaciers in the European Alps using a glacier model that explicitly resolves ice dynamics and is forced by GCM projections dynamically downscaled in the framework of CORDEX. The manuscript presents a timely step forward for modeling of glaciers on large regional scales. It is written well and succeeds to make the assumptions and limitations of the model accessible. I am particularly impressed with high quality of the presentation of the results in the Figures. Overall, I have no doubt that this is a valuable contribution to the literature and eventually should be published. However, there are a number of minor points that should be addressed by the authors, as well as three somewhat major issues
We thank the reviewer for taking the time to read our manuscript and for his positive (and very useful!) feedback. All points raised by the reviewer have been addressed and answered, and the manuscript has been updated accordingly.

Major comments

[RC1.02] Use of CORDEX data: This is the biggest issue I have: the authors use the projections from CORDEX, present a relatively thourough validation (some comments on this below), but do not at all address a question that seems very obvious to me: from the perspective of modeling glaciers, is there added value in CORDEX? I.e., what difference does it make applying the downscaled data opposed to applying the GCM data directly? I'm aware that because of internal variability, there is not much use in using the CORDEX (or GCM) data during the 2003-2017 validation period in the same way the E-OBS data are used. But there is still some insight to be gained concerning, e.g., temporally accumulated area and volume loss. Also: through comibination with differen RCMs, the ensemble size of CORDEX is increased - however, there are GCMs not part of CORDEX. Taking the ensemble as a measure of uncertainty, how does the addition of the RCM "axis" affect the ensemble uncertainty? Is is equivalent to taking a larger GCM ensemble, or can insights be gained through the downscaling? Of course, this point is not a weakness of the manuscript. But I see the approach taken by the author as great opportunity to address some questions that are very relevant, since dynamical downscaling is very expensive, and its value is somewhat debated. I think it would be great if the authors took this opportunity.
We thank the reviewer for this suggestion, and agree that it can be instructive to compare the effect of directly forcing the glacier model with GCM outputs rather than RCM outputs (i.e. EURO-CORDEX RCMs, which are driven by a GCM). This is an analysis that we considered including when writing the original manuscript, but which we eventually decided not to include for the sake of conciseness and not to deviate too much from our main message. We have now reconsidered this, and the updated manuscript now includes new simulations in which the glacier model is directly forced with GCM data. We have however decided to not go into too many details, as this would be beyond the scope of the study and is, probably more suited for a dedicated, more in-depth analysis. Such an in-depth analysis would require a comparison between the GCM and the GCM-RCM simulations themselves, rather than only focusing on the effect this has on the glacier simulations. In the end, the focus of our study is on the glacier model, and not necessarily on the data that drive it. The focus of our new analyses with GCM forcing is on future glacier evolution, which is also the main period of interest in our paper. For the past, we have now performed an assessment of how the RCM model performs in response to a suggestion made by reviewer #2 (see our response to RC2.05). To stay in the framework of the original manuscript, we decided to only include GCM simulations that are used within EURO-CORDEX (in other

words: we do not consider GCMs that were never used to force EURO-CORDEX RCMs). For this, we considered the three GCMs with the corresponding RCM simulations in the EURO-CORDEX ensemble for a given realisation (CNRM-CM5 r1i1p1, HadGEM2-ES r1i1p1, MPI-ESM-LR r1i1p1).

The results from the GCM-forced simulations support our earlier findings, which indicated that the differences in future glacier evolution are mainly caused by differences in GCM outputs, rather than by differences in the RCMs forced by them (results from section 6.1.4.). In fact, for a given GCM, the spread between EURO-CORDEX simulations forced by the same GCM have a relatively narrow spread. As only few GCM-RCM simulations exist, it is not possible to determine whether these simulations spread evenly around the corresponding GCM simulation. The available simulations, however, do not indicate that any systematic over- or underestimation might occur. As such, we conclude that the 'addition of the RCM axis', as suggested by the reviewer, does not have a significant influence on the results. This suggests that, at least for our setting, using high-resolution RCM simulations forced with a GCM leads to relatively similar results compared to using the GCM forcing directly. As stated above, a more in-depth study would however be needed to further investigate this.

In the updated manuscript, we now introduce these new simulations and present their outcome in section 6.1.4:
The importance of the GCM-forcing also appears from additional simulations in which the original, low-resolution GCM output was used as model forcing. When comparing these results to the ones obtained by forcing the model with the corresponding EURO-CORDEX GCM-RCM combinations, a similar glacier evolution is obtained (with volume losses vs. present-day typically differing <10%, suppl. mat. Fig. S5). The limited number of GCM-RCM combinations, however, does not allow for a detailed comparison with the GCMs, but does not suggest any systematic over- or underestimation of the corresponding results.

And we now also mention this in the conclusion:
Additional simulations where the model is forced with the driving GCM only (i.e. no downscaling with an RCM) confirm the limited effect of the RCM on the modelled future evolution. More in-depth analyses on the effect of using downscaled RCM data vs. GCM data for glacier evolution modelling will be required, but our results suggest that the effect of such a downscaling on simulated glacier evolution is relatively limited – at least for the European Alps.

A figure in which these results are summarised was added to the supplementary material (suppl. mat. Fig. S5; see following page)

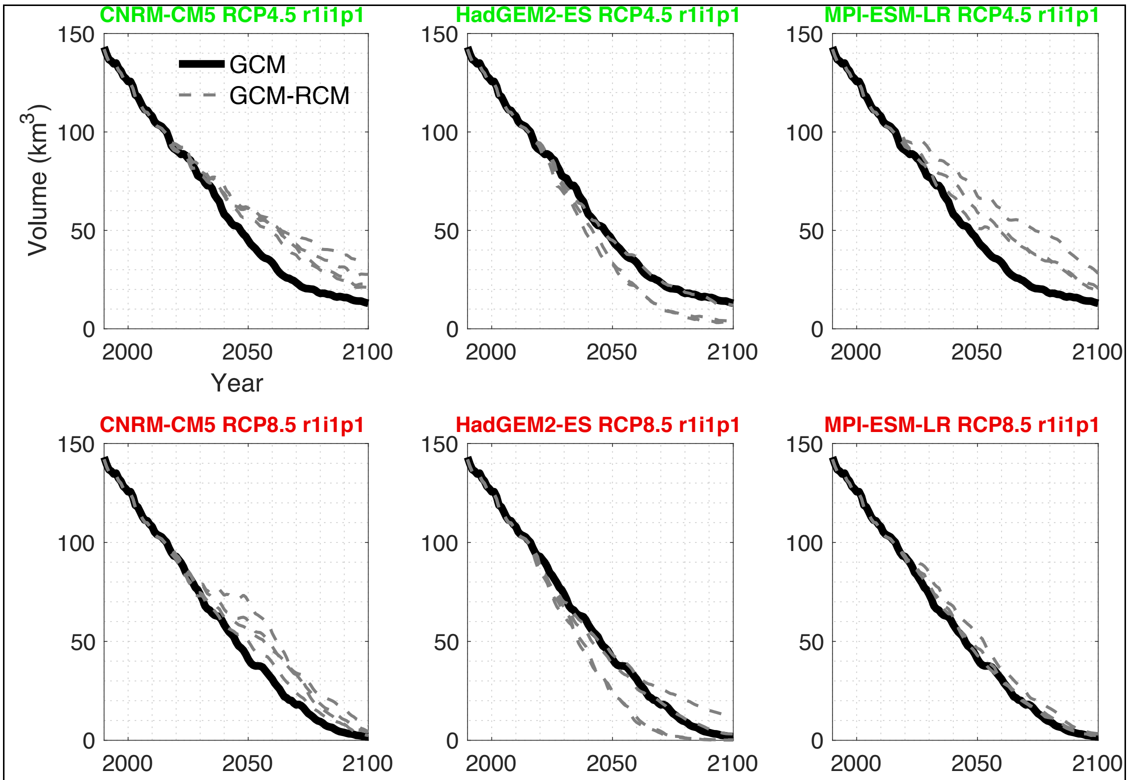

Fig. S5. Future evolution of glacier volume when the model is forced directly with GCMs vs. EURO-CORDEX GCM-RCM simulations. Results are shown for the GCMs that have the most corresponding RCM simulations in the EURO-CORDEX ensemble for a given realisation (CNRM-CM5 r1i1p1, HadGEM2-ES r1i1p1, MPI-ESM-LR r1i1p1).

[RC1.03] Initialization: Why is glacier length chosen instead of glacier area? It seems to me that area is better constrained than glacier lenght, since the length is to some degree an arbitrary choice resulting from the representation of the flow line.

Also, using the 1961-1990 climatology, was variability preserved? Where the forcing timeseries during that period detrended? As the authors say, there is no assumption about an equilibrium state of the glacier, but an assumption that the glacier woud be able to undergo the transition from an equilibrium state in 1990 to the obsserved transient state in (typically) 2003). Doesn't this correspond to an assumption that the response time of the glaciers is shorter than 13 years? You discuss this in Sect. 6.3, but it would be good to have some arguments presented already in Sect. 3.3.

We agree that the length has a dependence on the methodology used to derive it. However, for the calibration we aim at minimizing the difference between the 'reference' length and the 'calibrated' length. In this sense, considering the length or the area as a measure for this exercise is almost equivalent, as locally the area is the product of the glacier length and width. This is now explicitly mentioned in the updated manuscript and supported with quantitative information:

- We calibrate to the reference glacier length within 1%. The glacier length is thus very closely reproduced, with a standard deviation between the reference and the modelled lengths corresponding to 0.5%.
- For the glacier area, despite not being calibrated to it, the agreement between the reference and the modelled value is almost as good as for the length, with a standard deviation of 0.7%.

Although we agree that the area could have been used instead, we preferred to use the length as a criterion for our calibration, as this ensures that the position of the glacier terminus is correctly reproduced. We now explicitly mention this in the manuscript:

The glaciers are calibrated to match the length and volume at inventory date within 1% (σ *between reference and modelled volume or length of 0.6% or 0.5% of the reference value, respectively). Despite not being calibrated to it, the observed glacier area is also closely reproduced (σ of 0.7%).*

For the period 1961-1990, the variability is not preserved, nor is any detrending applied. This is not needed for our purpose, as this forcing is only used for creating the artificial steady state in 1990. In fact, the constant conditions correspond to the mean SMB over this time period (and an eventual bias), and we have now reformulated this passage to better reflect this (see updated text below)

Our setup does not assume that the glaciers have a response time of less than 13 years. In fact, we rather assume the opposite, i.e. that most glaciers have a response time of more than 13 years, so that the geometry in 2003 still depends on the geometry in 1990. This is important for our specific setup to be used. We argue that this is realistic, as most glaciers were not very far from equilibrium in the late 1980s, and since the typical response time of glaciers, in the order of a few years to decades, suggests that observed evolution in the 1990s and early 2000s was largely determined by their 1990 geometry. We have now emphasized this in section 3.3.:
The initialisation consists of closely reproducing the glacier geometry at the inventory date. At first, constant climatic conditions are imposed, until a steady state is created, which represents the glacier in 1990 (Fig. 3). These constant climate conditions correspond to the mean SMB under the 1961-1990 climate, to which a SMB perturbation is applied (detailed below). Subsequently, the glacier is forced with E-OBS data, and evolves transiently from 1990 until the glacier-specific inventory date (typically 2003). We opt for a 1990 steady-state glacier, as the glaciers in the European Alps were generally not too far off equilibrium around this period, with SMBs for many glaciers being close to zero (Huss et al., 2010a; WGMS, 2018). By imposing a steady state in 1990, the glacier length at inventory date can be influenced. Methodologically, choosing an initial steady state before 1990 would be problematic, as in this case the glacier geometry would not determine the glacier length at the inventory date anymore, as the period between the steady state and the inventory date exceeds the typical Alpine glacier response time of several years to a few decades (e.g. Oerlemans, 2007; Zekollari and Huybrechts, 2015).

[RC1.04] Validation: The geodetic mass balances used for calibration are most probably not independent of the in situ measurements used for validation. For those glaciers that have geodetic observations that do not temporally overlap with the in situ measurements, it would be good to recalibrate using these, and revalidated using the temporally independent in-situ values.

I'm also wondering about the choice of using geodetic observations for calibration, and in-situ for validation. The opposite woud seem like the more natural choice to me, since geodetic MBs can cover longer time periods, and thus allow to include effects from ice dynamics in the total uncertainty. Is there a specific reason not to make the calibration based on in-situ observations, and validate using geodetic MBs?

We agree that in-situ measurements cannot be considered to be entirely independent from the geodetic measurements for cases where both overlap in time (despite relying on different sources and techniques used to derive these values). Following the reviewer's suggestion, the SMB validation has therefore been revised, and we now only consider in-situ measurements that do not overlap in time with the geodetic mass balance measurements. Note that some in-situ measurements refer to glaciers for which no geodetic mass balance exist; these measurements were thus included in the validation as well. These latter measurements are particularly interesting, as they also serve to validate the extrapolation of the geodetic mass balance (cf. comments RC2.02 and RC2.23 by the second reviewer). The text, Figure 4 and its caption have been revised to account for this new validation procedure:
In order to ensure that the validation procedure is independent from the calibration,

validation is only performed with observations that do not temporally overlap with the geodetic mass balances used for calibration (cf. sections 2.3 and 3.1) and for glaciers without geodetic mass balance observations.
and:
When only considering SMB measurements on glaciers that have no observed geodetic mass balance (i.e. glaciers for which the geodetic mass balance used to calibrate the model was extrapolated from other, nearby glaciers), the misfit between modelled and observed values increases only little (RMSE = 0.79 m w.e. yr-1; MAD= 0.72 m w.e. yr-1; mean misfit = -0.19 w.e. yr-1), indicating that the method used to extrapolate the geodetic mass balances to unmeasured glaciers performs well.

Figure 4 was updated to account for this new validation, and so was its caption:
Fig. 4. Evaluation of modelled SMB against observations from the WGMS (2018) database. All observations are included, except those that do temporally overlap with the geodetic mass balance observations (used for calibration).

We use the geodetic mass balance for calibration because that is available for many glaciers (for ca. 1500, representing more than 60% of the total glacier area; cf. response to RC2.02). In contrast, only a few glaciers have in-situ measurements. We argue that it is more important to have a good coverage for calibration than for validation, and think that this strategy will be the only one applicable to other regions or the worldwide scale, since the availability of geodetic mass balances massively outgrows the availability of in-situ measurements by far (c.f. works by Brun et al., 2017, Nature Geoscience; Braun et al., 2019, Nature Climate Change). To clarify this, we added the following sentence in section 2.3. (that is where the mass balance data is introduced):
Note that we prefer using geodetic mass balance over SMB observations for calibration, as we argue that it is more important to have a good coverage for model calibration than for its validation. Furthermore, geodetic mass balances are becoming increasingly available at the regional scale (e.g. Brun et al., 2017; Braun et al., 2019) and outgrow the availability of in-situ measurements, making the adopted strategy applicable to other regions.

Minor comments

[RC1.05] P1 L10: ... ice flow processes, _of_ which the latter ...
This was modified.

[RC1.06] P1 L10: you could specify it is not included explicitly, since the (simple) parameterizations used are supposed to parameterize its effect.
[note that “it” in the reviewer comment refers to ice dynamics] We have now specified that this effect was not included *explicitly* in previous studies: *…, of which the latter is to date not included explicitly in regional glacier projections for the Alps*

[RC1.07] P2 L21: specify "glacier model parameters".
Two possible glacier model parameters, which are discussed in the manuscript, have now been added: *…(e.g. flow parameters and cross-section parameterization).*

[RC1.08] P2 L17: either delete "using various methods", or briefly specify.
Deleted.

[RC1.09] P2 L18 (and P15 L27): volume/length/area or volume-lenght-area scaling
This was modified as suggested by the reviewer: *volume/length/area scaling (Marzeion et al., 2012; Radić et al., 2014)*

[RC1.10] P2 L21: I suggest RCP 8.5 should not be called "extreme", since the different scenarios are not based on probabilities, nor is there a physical reason to view RCP8.5 above some limit value.
This was now modified to: *…, and an almost complete disappearance of glaciers under warmer conditions (RCP8.5).*

[RC1.11] P3 L2-4: Not sure why glaciers in the Canadian Rockies should not be controlled by topography and local effects. Either specify, or delete; I think there is value in different approaches to similar problems by itself, so I don't see a need for a very strong justification of your study here.
This passage was deleted. The previous sentence now reads: *In an RGM study for western Canada, Clarke et al. (2015) showed that relative area and volume changes are well represented by such a model, but that large, local present-day differences between observed and modelled glacier geometries can exist after a transient simulation.*

[RC1.12] P3 L9: you might also refer to Goosse et al. (2018), which has transient simulations (DOI: 10.5194/cp-14-1119-2018)
This is a good suggestion. We added: *This model was recently also used by Goosse et al. (2018) to simulate the transient evolution of 71 Alpine glaciers over the past millennium.*

[RC1.13] P4 L7: delete "in these little-known connections" (or specify why they are lesser known than the rest of the glacier).
It is not the connections that are little-known, but rather the mass transfer between these connections. We have now reformulated this: *and potential problems related to solving the little-known mass transfer in these connections.*

[RC1.14] P4 L12: I would prefer "Climate data" or "Atmospheric boundary conditions"- but that is maybe a matter of taste.
This section is now entitled: *2.2 Climate Data*

[RC1.15] P4 L19: "... has a higher resolution than the reanalysis data used in Huss & Hock (2015, ERA-INTERIM) and goes back further in time."
This was modified in response to RC2.19 and now reads: *We prefer using an observational dataset compared to a re-analysis product (e.g. ERA-INTERIM, as used in Huss and Hock, 2015), as the former has a higher resolution and goes further back in time.*

[RC1.16] P4 L21: rephrase around "several chains used for", since it is not quite clear here

what "chains" you are referring to.
We now refer to these RCM combinations as *(RCM) simulations* here and throughout the manuscript.

[RC1.17] P4 L28 (and throughout the study): perhaps it would be better to call them "model combinations" than "chains" (these "chains" have only two links anyway...).
This comment was addressed in our previous response (to RC1.16): the “chains” have been changed to “RCM simulations” throughout the manuscript.

[RC1.18] P4 L30: "... a peak-and-decline scenario ..."
Modified as suggested.

[RC1.19] P4 L32: "Note that while country-specific ..."
while was added.

[RC1.20] P4 L32: "... such as the projections recently released for the CH2018 report, ..." (or something similar - i.e., distinguish between scenarios and projections).
This was reformulated to: *Note that while country-specific projections such as the ones recently released with the CH2018 report for Switzerland (CH2018, 2018) exist,…*

[RC1.21] P5 L13: because of the difference in resolution between E-OBS and the RCMs, there is probably not a direct equivalent in the grind points. Please specify how you handle this.
The trends from the RCMs are imposed on the E-OBS grid to simulate the future SMB. For this, we rely on the nearest grid point. We now mention this in the first sentence of this paragraph: *For modelling the future SMB, debiased RCM trends from the EURO-CORDEX ensemble are imposed on the E-OBS grid based on the nearest corresponding grid cell.*

[RC1.22] P6 L13: "closest to the glacier" - there are probably glaciers that are covered by multiple grid cells - are then multiple grid cells taken into account, or are you using the glacier center point? Please specify.
In such cases, the glacier centre point is taken into account, which we now also specify: *For every glacier, the model is forced with monthly temperature and precipitation series (section 2.2) from the E-OBS (past) or RCM (future) grid cell closest to the glacier’s centre point.*

[RC1.23] P6 L15 (and L19): "temperature-index melt model"
This was modified for both occurrences in the text: *temperature-index model → temperature-index melt model*

[RC1.24] P6 L26-30: While precipitation is the least constrained boundary condition, the degree-day factors are probably the most potent paramteters for tuning. Why is the order chosen like this, and how is the defaul degree-day factor chosen? Please specify.

We keep the same setup as in GloGEM, in which the precipitation is modified first, and the degree-day factor is potentially altered after that (a third step involves a change in temperature). The reason for this is that local precipitation is often least constrained and most variable in high-mountain areas (e.g. snow redistribution). This is now also mentioned:
In the first step, overall precipitation is multiplied with a scaling factor varying between 0.8 and 2.0. This initial step focuses on the precipitation, as this is the variable that is expected to be the most poorly reproduced due to resolution issues, spatial variability and local effects (e.g. snow redistribution) (e.g. Jarosch et al., 2012; Hannesdóttir et al., 2015; Huss and Hock, 2015).

The default degree-day parameter is set to 3 mm d^{-1} K^{-1}, in line with the original GloGEM study (Huss and Hock, 2015) and is in agreement with literature values from various studies (Hock, 2003). This is now also mentioned in the text: *…(default value is 3 mm d^{-1} K^{-1}; cf. Hock (2003)) and the degree-day factor…*

[RC1.25] P8 L25: It is fine to seperate the validation of the SMB model from the ice dynamics, but there should not be much difference if the dynamics are behaving well. Also, the SMB observations are probably not taking into account the RGI outlines, such that there is some disagreement anyway. Have you tested how much the results differ if you switch on ice dynamics?

It is true that accounting for an evolving geometry while validating the SMB should work well if the glacier evolution model works well. However, we decided to not rely on a dynamically evolving glacier geometry for the SMB validation for the following reasons:

- By relying on a dynamically evolving glacier geometry, only SMB observations after 1990 (starting point of dynamic simulations) can be used for validation, This would reduce the total number of available measurements (note that the number was already reduced in the updated manuscript, to account for the interdependency issue of validation/calibration data raised by the reviewer in RC1.04)
- A related problem is that the changes modelled in glacier area shortly after 1990 are in general too small, as the glaciers needs to evolve away from the steady state (that is also the reason why our validation occurs mainly after the inventory date). Depending on the timing, additional observations may thus have to be removed from the validation dataset.

The sentence has been reformulated in order to reflect this reasoning:
As the aim is to evaluate the performance of the SMB model (rather than the coupled SMB – ice flow model) and to incorporate as many validation points as possible (which is only possible after 1990 for the dynamic simulations), these calculations are based on the glacier geometry at inventory date.

[RC1.26] P8 L29-30: I don't agree: this correlations measures the skill of the model to reproduce more positive mass balances at the upper part of the glacier than at the lower parts. A better measure for the SMB model's ability to represent the MB profile would be a comparison of the (temporally averaged) MB profiles, or correlating the deviations from the "climatological" (i.e., mean) profile.

We agree that the original formulation ('*the SMB model distributes the annual SMB relatively well over elevation*') may be perceived as misleading. However, this analysis indicates that the SMB gradient is well reproduced. This has now been reformulated:
Furthermore, the good agreement between observed and modelled balances for glacier elevation bands (r^2 = 0.60; Fig. 4b,d) suggests that, despite not being calibrated to this, the modelled and observed SMB gradient are in reasonably good agreement
Additional measures, such as the ones proposed by the reviewer, could indeed be interesting to present, but given the limited data available for such analyses (we would

need continuous time series for a meaningfully large set of glaciers), these were not included.

[RC1.27] P10 L20: delete "1x".

This was deleted.
p-value <1x10^{-3} → p-value <10^{-3}

[RC1.28] P10 L27: delete "In the literature".

By deleting *in the literature*, it gives the impression that this is something we have performed in this study. Although it becomes clearer when reading further that this is not the case, we want to avoid any possible confusion here. We have therefore rephrased this to:
Glacier area changes in the European Alps have been derived in various studies.

[RC1.29] P10 L30: "over the period 1973-1998/9".

This was modified.

[RC1.30] P11 L19: "evolution is independent of the scenario".

This has been reformulated to: …*this evolution is independent from the RCP.*

[RC1.31] P12 L6: I don't understand this sentence, please rephrase. Do you mean volume change?

We thank the reviewer for pointing this out, as this was a mistake. The sentence now reads:
…reveals that under RCP2.6, the relative volume loss has the highest correlation with the elevation range (Fig. 9 and suppl. mat. Table S3; r^2 = 0.57).

[RC1.32] P12 L11: "strong below 3200 m".

Modified.

[RC1.33] P12 L18: delete "may" (you already say it's only under certain model combinations).

may was omitted.

[RC1.34] P12 L21: "This is illustrated for Langtaler ..." Sect. 6.1.1: Just out of curiosity, it would be interesting to know about the committed mass loss past 2100.

Langtaler Ferner is projected to lose 89% of its 2017 volume by 2100 under 1988-2017 conditions. After 2100, almost no loss occurs anymore, and slightly less than 10% of its 2017 remains in the end (steady state under 1988-2017 conditions). We now also mention this in the section on the committed loss (6.1.1):
Under these conditions, the committed loss is particularly strong for small glaciers at lower elevations (e.g. Langtaler Ferner, with a committed volume loss of ca. 90% by 2100)

[RC1.35] P14 L6: correct Section number.
This was corrected: *…possible to the observed geometry (see section 4.2), which is…*

[RC1.36] Appendix: P19 L16: delete "1" at the end of the line.
This was deleted and the sentence was changed to: *…previous guesses minus 1, e.g. the third guess…*

[RC1.37] Table 1: The "committed loss" line is a bit confusing, since the committed loss should not be time-dependent. Perhaps you can call this, e.g., the "realized fraction of committed loss"?
This was indeed somewhat confusing. To clarify this, we now refer to 1988-2017 in the table and in the table caption: *The evolution under the mean SMB obtained from the 1988-2017 climatic conditions represents the committed loss.*

[RC1.38] Generally - but particularly Fig. 9: I'm not a big fan of grids in figures. If there is a need to read numbers/differences from the figures, these numbers should be mentioned on the figure or in the text, and no grid would be necessary; if there is no such need, the grid only adds clutter. Please consider removing grids.
We agree that the grid on the maps in Fig. 9 was not needed and removed this in the updated manuscript. In the other figures we prefer to leave the grids, which we think makes it easier and more convenient to read values in general (without any specific 'highlight' value).

[RC1.39] P 27 L3: "relative to 1961-1990" (instead of "with respect to").
This was modified.

[RC1.40] P27 L4: is that the mean of all "alpine" EUROCORDEX grid cells, or just the ones that contain glaciers? If the latter, I think it would make more sense to also weight them by glacier area.
We agree with the reviewer that it makes more sense to show the mean temperature and precipitation weighed by the glacier area, and have updated Fig. 2 and its legend accordingly.

[RC1.41] Fig. 2: I have a hard time seeing the transparent bands; I suggest deleting them, since that information is already shown in the distribution of the individual (thin) lines.
The transparent bands were omitted in the updated manuscript.

[RC1.42] Figs. 5, 6: are there uncertainty estimates available for the observed velocities? If so, it would be good to include them (e.g., just a single error bar so that the observation uncertainty can be compared to the differences with to modeled velocities).
We agree with the reviewer that it is interesting to incorporate error estimates on surface velocities. Unfortunately, uncertainty estimates are often absent for the measurements that we sampled from the literature. Uncertainties in observed velocities are generally <10% (e.g. Berthier and Vincent, 2012; Zekollari et al., 2013; Stocker-Waldhuber et al., 2018). We, thus, argue that the general tendency in this validation will remain unaffected.