Response to reviewer 2 (Lindsey Nicholson and Niklas Richter)

Long-term firn and mass balance modelling for Abramov glacier, Pamir Alay

Marlene Kronenberg, Ward van Pelt, Horst Machguth, Joel Fiddes, Martin Hoelzle, Felix Pertziger

Dear Reviewers,

We would like to thank you for your attention to our manuscript. We appreciate the interest in our study and thank for your constructive review and suggestions how to improve the quality of the paper. Below, we respond point by point to all comments, and state how we plan to account for them in a revised version of the paper. The responses (normal font style) to the reviewers' comments are written directly into the reviews (displayed in italic font style). Revised/additional figures can be found at the end of this response letter and are labeled with Roman numerals to them from figures in the manuscript.

Marlene Kronenberg,

Fribourg, May 24, 2022

1 General comments

Originality: The study makes use of key datasets at Abramov glacier to perform a novel modelling study to reconstruct 52 years of mass balance with concurrent process understanding. While this in itself is novel and the glacier is unusually rich in data for the region,a more explicit statement of the scientific motivation and purpose of the paper as well as critical discussion of how it might contribute to wider glacier process understanding in this region would be welcome. The contribution of modelled subsurface firn conditions is certainly valuable for improving density assumptions for geodetic mass balance estimates and for understanding the evolving potential internal accumulation capacity of the glacier over time.The paper concludes that the effects of warming on glacier mass balance are partially mitigated by increased accumulation, which may be relevant at a regional scale, and highlights that the buffering effect of internal refreezing is diminishing over time.

We will state our scientific motivation more explicitly in the introduction and expand the discussion regarding its contribution to the regional understanding. More details are given below.

Scientific quality: The purpose of the paper could be more clearly stated with reference to what the key outputs are. For example saying that a process model allows you to build on the existing field firn study, and that the climate drivers of glacier mass change remain contested in this region, and can be investigated with a process-based model. We acknowledge that the parameter calibration is performed with rigorous manual calibration, yet by overlooking potential measurement uncertainties, as well as not using a multi-objective simulation/parameter evaluation this is not state of the art. Could you thus expand on the performance of the calibration? It would be good to justify your choice of sample number and location for the calibration datasets, and the treatment of uncertainties could benefit from recognition of a parameter equifinality issue (Rounce et al., 2020) in the optimisation. Rather than reporting only a single best-fit solution, incorporating a sensitivity analysis and its related uncertainties would be more robust. Indeed your comparison of the two sets of forcing data highlights the sensitivity of the modelled mass balance to the forcing data which demonstrates the value of a more comprehensive sensitivity study. This deserves even more emphasis given that the only difference between the two forcing data sets resides in the timespan 1980-1998, the period for which you calculated the biases and should hence be closest to the observations.

We will more clearly state the purpose (Quantification of (i) mass exchange processes in the firn and (ii) changes therin and (iii) of their contribution to the glacier wide mass balance over a long-time period) and the main outputs of our study. Our investigation on firn changes and their impacts on mass balance for Abramov glacier substantially contributes to the process-understanding for this poorly studied area located at the edge of regions with anomalous mass balance trends from a global perspective.

Regarding the calibration: A complete parameter exploration would be numerically extremely expensive and therefore not feasible. Hence, a "smart" solution as used here, selecting key parameters and choosing an appropriate order in which to calibrate them, has been adopted here. Parameters values were updated not only to reduce the bias between modelled and measured mass balances but also in order to simulate processes as realistic as possible. Doing so, the available data, which mainly consists of surface mass balance data, could optimally be used to tackle different processes relevant for a measured phenomena such as ablation (governed by snow melt, albedo degradation and ice melt).

This conceptual approach has the advantage, that we found a unique and meaningful value for each parameter with respect to simulated processes. It allows also to circumvent the issue of equifinality which would likely occur if we optimised numerous parameters with the available calibration data. The approach is based on several assumptions and related uncertainties have indeed not been quantified in the current version of the manuscript. We agree that this should be improved. We have therefore conducted model runs with perturbed parameters for selected grid points and use them to evaluate the sensitivity of the modelled surface mass balance and internal accumulation to altered parameter values (cf. Fig. I and II). We will clarify the selection of the calibration data set and its visualisation on the maps and quantify uncertainties of in situ mass balance data following Thibert et al. (2008).

We plan to include a more comprehensive sensitivity discussion including sensitivities of parameters, forcing and initial conditions.

Given the number of pre-processing steps in the forcing data and selection of model parameters, the impact of uncertainty in these choices is poorly quantified, making the model output and performance difficult to interpret, as we cannot be clear about the quality of the forcing data compared to reality. It may be helpful to show plots of the forcing data in the supplementary material as a start. This would also help addressing the difference between the two forcing data sets. Relating to the question of the overarching goal of this study: if the goal is to draw out the connection between the climate forcing and glacier response, investigations on the annual time series of mass balance alongside climatological properties - for example by adding climatological information into Table 3 or attempting to understand the causes of positive mass balance years in the timeseries - could be a starting point.The timeframe of this study allows for an in-depth investigation on the drivers of glacier response also on the decadal scale, as continuously highlighted interdecadal differences in glacier response exist. Therefore, it could be worthwhile to investigate whether there exist any breakpoints in the climatological data that could explain the simulated decadal variations in the mass balance (Table 3).

These pre-processing steps served to prepare a continuous model forcing data set. We will add plots of the model forcing data to the supplementary material (cf. Figs. III and IV). Furthermore, the presentation of a sensitivity analysis will allow to contextualize the choices and enhance the interpretation of the model output and performance. Furthermore, we will visualize the mean annual bias between the model output and the surface mass balance measurements to Figure 4 what will also contribute to the interpretation of the model performance. The visualisation of the simulated energy fluxes and the the climate will be improved to allow for a better interpretation of the decadal variation of modelling results.

Relevance: Key findings are that negative mass balance of -0.27m w.e./year persists for the period 1968/1969-2019/2020 alongside a loss of firn pore space causing a reduction of internal accumulation. Despite increasing air temperatures, no acceleration of glacier-wide mass loss was found over time as a result of increasing precipitation rates which appears to exert a strong control over annual mass balance at this glacier. This is of interest given that knowledge of precipitation in this region is quite poor and is thought to contribute to anomalous glacier behaviour in this region.

Presentation: (1) We suggest restructuring so that the model description comes before the description of the datasets - this will help guide the reader to know the constraints on the construction of the forcing data. If this move is made, we suggest headings as follows under a

'methods' section: study site, model description, forcing data, calibration data, validation data. (2) Please consider also the best way to order Sections 3.5 and 3.4 as they describe parameterisations applied before the application/analysis of the modelled results as well as bias corrections based on whole model simulations. (3) Please revisit the naming of your sites: you have sites 1 and 2 in the accumulation zone, but it would be helpful to call your ablation zone site, site 3 to be consistent in the naming convention and also to show its location on the maps. Relatedly, please can you add a statement as to how representative these chosen sites are for mean ablation zone/accumulation zone conditions. (4) Please show which stake locations/data are used for calibration cf which for validation in one of the map figures as its currently unclear if the locations for calibration are identical to the locations of validation but just a different subset of the data. (5) As currently presented the figures and tables are often far from the associated text which could be improved for the final layout, some suggested changes to figures are included in the specific comments, and some English edits suggested in an annotated pdf.

(1) We agree that restructuring the manuscript will help to guide the reader through the manuscript and thank the reviewer for their suggestion. In the revised manuscript, we will first describe the model followed by the construction of the forcing data.

(2) Sections 3.4. and 3.5. will be re-organised so that the they chronologically describe the applied steps.

(3) We will change the naming of the site in the ablation zone to 'site 3' as suggested and indicate its location on all the maps. A statement of the representativeness of the three locations will be added alongside with results for the three points.

(4) We will more clearly indicate the location of calibration/validation data in Figure 1b, 3 and 10 and clarify in the text.

(5) We will improve the figures and text following the specific comments and the annotated pdf. And the position of the figures/tables will be optimised.

---

2 Specific comments

L13: Is this heterogenous behaviour really only in the last decade?

This was a spelling error. We mean 'last decades'.

L18: Is this increase progressive, or stepwise, and what is its magnitude (to help readers not so familiar with the region)?

We will add more information here.

L25: It may be valuable to specify that the advantage of a physically based model is that it is expected to be more suitable for projecting processes into unknown climates, compared to a temperature index model calibrated for one period in time applied within a non-stationary climate projection. The additional caution of course is that many physically based models include stationary parameterizations.

We agree and will complete the statement as suggested.

L28: Please specify which processes and/or which models include/exclude them, or maybe better still just delete the sentence starting "Important ..." as it's really not needed here

We agree to delete the sentence starting "Important...".

L34: Suggest "In the data poor HMA, ..." or maybe better still just delete the sentence starting "In the ..." as it's really not needed given the following sentence

We agree to delete the sentence starting "In the ...".

L44: Suggest restructuring of this closing paragraph to more explicitly state some goals of the study. First state the problem/motivation (e.g. why do we care about this glacier/ timeseries/ region/... need for a continuous record in this poorly understood and anomalously behaving part of the mountain cryosphere?) and how doing this modelling study delivers a novel and useful solution (e.g. to evaluate decadal trends in surface energy balance fluxes, firn evolution ... and their connection to the forcing climate conditions, attempt more process understanding of previously suggested claims?)

This paragraph will be restructured as suggested.

L48: Delete "These in situ data, which are unique for the region, allow us to apply a rather complex model with relatively high data requirements."

Agreed.

L50: Delete the last sentence.

Agreed.

L64-69: Give time period of Barundun long term study; and state the long term mass balance of both studies.

Will be completed.

L69: Where is this point site on the glacier?

It is located at s2. A reference to Figure 1b will be added here.

L70: You mention a regime shift in the 1970s - please return to this point in the discussion and in the evidence from your modelled study, does your model show this increase, what are its effects on the glacier scale mass balance and firn development?

The discussion will be completed regarding this point. The modelling results revealed an increase of precipitation, but also an increase of melt energy which especially affects the lower accumulation area during the most recent decades. Whereas refreezing rates stayed similar at high elevations (also at the site presented in Kronenberg et al. (2021), they strongly decreased at lower elevations, which contributed to an overall acceleration of mass loss during the most recent years.

L76: Careful how you use this term reference glacier surface - it has a specific meaning as explained in Huss 2012, but I am not sure you mean this.

Indeed, this was confusing. We will avoid the term in the revised manuscript.

L78: The creation of the 1968 DEM needs more clarification on how you did it and how reasonable it is. Why do you use only 2 DEMs and not intervening satellite DEMs (e.g. SRMT, ASTER), to examine the trend from timestep to timestep? Would it be possible to use coarser, but more frequent dDEMs to check the reasonableness of the linear backwards extrapolation of surface height change over time, e.g. the ones mentioned from Barandun et al. (2015)?

We used the annual height change grid to calculate the DEM1968. This will be clarified. The two DEMs were used as they were readily available at a high spatial resolution and thoroughly quality checked from Denzinger et al. (2021). Glacier dynamics are not included in the EBFM which is by default run for a constant glacier grid. The application of a linearly changing elevation is a straightforward and computationally effective way to include topographical changes. A preparation and quality check of additional DEMs was beyond the scope of this study. Furthermore, we consider DEMs from SAR data as not suitable as the SAR signals are known to penetrate in the subsurface, at least in the accumulation area.

L83: Was this weather station manually monitored? Is the non-digitized data a good candidate for https://www.zooniverse.org/projects/edh/weather-rescue/

Yes, the station was manually monitored. We do not have the original handwritten records. Most of the data are available in a digital format which we plan to deposit in an online repository (we plan to provide the link/reference the link in the final paper).

L85: Suggest rephrasing to "Here, we use data from January 1968 until December 1998 for the following parameters: daily average air pressure, windspeed, relative humidity, temperature and cloud cover, as well as daily precipitation sum, daily minimum temperature and cloud cover and daily maximum temperature."

Will be rephrased.

L89: Can you say something about expected undercatch based on any other measurements in the region or on the likely mountain undercatch for the device used? To my knowledge snow undercatch can be ¿50 percent in mountain regions depending on the sensor so it would be good to mention here.

Yes, the undercatch is assumed to be substantial especially during winter months when snow is falling at low temperatures (e.g. Sevruk, 1985; Førland and Hanssen-Bauer, 2000). The comparison with monthly snow height measurements indicated a substantial undercatch during winter months, which we account for by applying precipitation bias correction factors. This issue will be mentioned here.

L91: Why 3 hour timesteps, when the ERA5 data is available hourly? If this is because the model needs this time resolution add this to the model description which should preceed this section. Did you consider better ways of reconstructing the daily cycle e.g. by using the daily cycle of the

ERA variables? For example it seems inconsistent that you introduce a daily cycle for cloud but then not for precipitation ... these choices need to be justified and their implications considered in the discussion section.

The model default resolution is 3hours. This resolution allows to represent the day-night cycle but is computationally less demanding than e.g. an hourly resolution. We will first describe the model and then the forcing to improve clarity. The choices will be assessed and discussed in a comprehensive sensitivity discussion.

Section 2.3.2: It seems that you don't actually make much use of the recent weather station - is that correct? Just used in the pressure parameterisation and the radiation parameterization? a. Consider mentioning explicitly the use of each dataset in the analysis.

We will mention the use of each data set. We used the recent AWS data for the air pressure and radiation parametrisations.

L107: Need some more explanation of the occurrence of missing data, what percentage is typically missing and what happens to the NaN data periods?

Only data from the non-NaN periods was used. Will add percentages.

L110: What do you mean by an IQ range filter? Can you specify the scale of this here please? With this and the subsequent double application of the std filter it sounds like you substantially reduce your dataset variability range in a more or less arbitrary fashion before you begin with any physically based QC? Can you explain why?

Here, we followed Stainbank (2018) who performed extensive analysis of the data set which initially contained extreme outliers. We will provide more details to the manuscript.

L116: Other publications show the way for some more rigorous quality control e.g. could correct for snow over SW sensor, using the albedo readings, and for freezing of station using non-variant wind direction, and for RH¿100 it is commonly reset to 100 according to many standard instrument manuals (e.g. Campbell scientific). RH values above 100% are usually set to 100%. Also you should report how many/what percentage of the total readings are filtered out (by each of the filtering steps)? Potentially include the whole quality controlled data series in the supplement.

We will provide detailed information about these data in the supplementary material (including quality control).

L122: Could you please just mention why you did not use the extended ERA5 1950-1979 dataset. It might be still preliminary, but did you explore using it at all?

A only very preliminary version of the extended data set was available when we compiled the model forcing data. Our interest was extending the historical *in situ* data time series (1968-1998) to recent years (rather than extending it further into the past). Therefore, we worked with the 'original' ERA5 data set.

L124: Possibly expand on what TopoSCALE does - you input upper air field and a DEM, and get out radiation modifications based on the DEM, which all other fields remain the same for the

given pressure level? For readers unfamiliar with this product it would be worth explaining also why you want to do this to your data first. As a side note, how does TopoSCALE compare to TopoCLIM now available? https://gmd.copernicus.org/articles/15/1753/2022/. Can you also show the ERA5 gridcell location on your map please?

We will explain the TopoSCALE downscaling (3D interpolation of atmospheric fields available on pressure levels and a topographic correction of radiative fluxes) in more detail. The gridcell location will be shown in Figure 1.

TopoSCALE downscales current climate reanlysis data based on a high resolution DEM, whereas TopoCLIM downscales CORDEX RCM data based on high resolution TopoSCALE data, as a best guess of current reference climate, using quantile mapping. Together these methods can produce consistent timeseries of past, current and future climate.

L128: Its not clear why you need cloud cover fraction but presumably for the SEB model, so perhaps state this at the start of the section, or better still put the model description before the data description so that its clear to the reader what is needed to force the model. Does the clear sky value come from ERA5 + TopoSCALE? Why usage of cloud cover fraction instead of readily available longwave radiation from ERA5?

We will put the model description before. The model simulates LWin using cloud cover and Tcl is furthermore used for the simulation of SWin. We followed the approach by Klok and Oerlemans (2002) to get Tcl. The ERA5 long-wave radiation cannot be corrected as no radiation data was measured at the original station. Given the large biases in various other ERA 5 parameters, we prefer not to use any parameter from ERA5 that we cannot bias correct using data from a weather station We thus used the parametrisations in the model and opted against changing them due to a lack of suitable data.

L129: You may want to add units for completeness here.

Agreed.

L135: Are the parameter values developed for a particular site, or globally valid?

The parameter values were found for Nordenskiöldbreen (Svalbard), different values were found by Greuell et al. (1997) for Pasterze glacier (Alps), where τ_{cl} depends more strongly on cloud cover. We tested both parameter sets and obtained more realistic radiation values compared to AWS data when using the values for the Arctic site by van Pelt et al. (2012). Due to a lack of data (no simultaneous measurements of cloud cover and SWin), the local values could not be calculated for Abramov glacier.

L140: "Even after downscaling (accounting for resolution diff.)..." here are you referring to what TopoSCALE does, without an explanation of what TopoSCALE is it is not clear to the uninitiated reader at this stage that it is a downscaling scheme, by which I understand it changes the resolution of the reanalysis data?.

Yes, TopoSCALE is a downscaling scheme. We will clarify.

L142: Explain why you aggregate to monthly for the bias correction when your data are available at different timescales, these steps seem confusing. Furthermore, you assume that biases 1980-1998 are the same as in recent decades, is that likely? Why didn't you for example check the other station data to correct biases. Is the height difference between the two weather station locations handled by TopoSCALE?

The two weather stations are located at different locations and elevations and different parameters are recorded with different instruments. A TopoSCALE downscaling can be performed for both locations. Unfortunately, the reproduction of the exact elevation/position of the AWS is hampered by the fact that the AWS is located on a small and exposed rock outcrop which cannot be perfectly reproduced by the DEM used. Furthermore, the AWS measurements are strongly affected by shading and warming effects by nearby peaks and rock walls. The shading effects cannot be reproduced when simulating or downscaling radiation based on DEMs, even if a relatively high resolution (e.g. 25 m) is applied. Therefore, the AWS data are of limited use for debiasing the TopoSCALE data and we opted for consistently use the weather station data available for more than three decades.

L147: After correction using monthly data, cloud cover fraction in summer showed bad fit, therefore set cloud cover to 0 for days without precipitation and to 1 if precipitation is above threshold, this seems flawed as cloudy days could occur without precipitation, would it not be better to correct this with longwave from ERA5 data? Or at least explain and justify these seemingly arbitrary choices - perhaps it is appropriate for this site, but please make a case for this here, using data or citations.

The cloud cover forcing indeed contains uncertainties which are poorly quantified. ERA5 longwave radiation can hardly be used here, as overall too humid conditions result from ERA5. We will present additional sensitivity runs using different cloud forcings and systematically analyze impacts on simulated radiation fluxes to allow for a quantification of related uncertainties.

L150: So what is the timescale of these two forcing datasets? Original ERA5 data is hourly, ERA5 TopoSCALE bias corrected data is monthly, and is this dataset 3 hourly. Here please state the purpose of this duplicate dataset ... what do you gain from running the model with two datasets. I think it tells you something about the sensitivity of your results to the forcing data, but please explain this clearly here, and what it tests and does not test..

Monthly averaged deltas were used to the correction of ERA5 TopoSCALE, but its resolution stays hourly. Yes, this model run was to analyze the sensitivity. Will be stated clearer and discussed in detail (new sensitivity discussion).

L150: Can you also show a comparison of the overlapping period of these two datasets, alongside the available weather station data so we can see how well they duplicate each other over periods of overlap - this is important to convince the reader of the homogeneity and robustness of the forcing data derived in this way. Probably fine in the supplement.

Will be added to the supplement (cf. Fig. IV).

L165: Density assumptions including a constant elevation threshold, could you please justify this, especially as your abstract implies the firn is changing density, and presumably changing in

spatial extent as well? You highlight in the text that this is a major source of error so it might be nice to consider simple uncertainty analyses based around output sensitivity to varying these assumptions. Also why is the firn line 4200m a.s.l., is this based on satellite data or the stake data, please justify this choice in the text.

This choice is done due to a lack of local density measurements at each stake. Values reported by Pertziger (1996) which were established based on average measurements from snow pit measurements. These data are only used for model validation. We will provide error bars to the validation data accounting for the uncertainty of these assumptions. The firn line was set to 4200 m a.s.l. based on field observations.

L174: Please add in the text that March is selected as the end of the winter seasons before the snowpack is modified by melt and that September is the end of the hydrological year at this site, or otherwise justify these with your measured data, or citations.

ok.

L175: What is this selection of 19 stakes for - later it becomes clear this is for calibration, but check this whole paragraph for what you really want to explain.

ok.

Section 3: Moving this section to before the data description would make it much clearer what forcing data is actually needed to drive the model and therefore make it easier to understand the data preparations steps.

Agreed.

Figure 2: It would be useful to refer to this figure more in the data description - e.g. when discussing the two forcing datasets, and also maybe include the borehole temperature data on it?

Agreed.

L197: Does the Svalbard development location mean its particularly suitable for this glacier site? Better than alternative models for example? Certainly refreezing is important in Svalbard glaciers, right?

The model is considered particularly suitable as it combines an energy balance routine with a firn model with a suitable degree of complexity regarding the availability of data and objectives of our study. The surface energy balance model explicitly accounts for local topographic effects on the radiative fluxes (e.g. shading by surrounding topography, orientation of the grid cell) which is particularly relevant in rugged mountainous terrain. Refreezing also plays an important role on Svalbard glaciers (see also Marchenko et al. (2017)). Furthermore, the surface energy balance parametrisations were originally developped and applied for Morteratsch glacier (e.g. Klok and Oerlemans, 2002), where conditions are more similar to Abramov glacier compared to other sites for which such models were developed.

L201: Could you include a comment on the model performance from these studies - did it do well in comparison to others?

Will be completed. (In Svalbard, the model has been shown to perform well compared to other models as visible from the comparison with measured data (van Pelt et al., 2019)).

L206: State here which meteorological forcing data are required.

Ok.

L212: Indication on connotation of fluxes (positive, negative) with respect to surface.

Will be completed.

L214: How does the transition occur? Binary or gradual?

It's a linear transition. Will be completed.

L216-208: This modelling of radiation seems odd as I thought this was the point of the TopoSCALE downscaling to give relevant radiative properties for a surface? Please start this section with the required forcing data for this model, and why you chose to not force it with the ERA5 incident radiative fluxes? Are cloud parameters for Svalbard (sea level and maritime) appropriate for Abramov? Can you quantify the better agreement and also state which weather station data was used here?

Our statement was indeed unclear. As stated above, the paper will be restructured and we will clearly highlight the necessary model inputs and the lacking availability of radiative fluxes for the first decades (ERA5 only from 1980 onwards, no measurements). We will also add a quantification.

L223: Longwave radiation is computed in ERA5, why not use it to begin with?

Please also refer to our answer above. This is the standard method applied by the EBFM (cf. van Pelt et al., 2012). No longwave radiation is available for the first decade as it was not measured at the station. We prefer to use a consistent method throughout the entire simulation period. Furthermore, the ERA5 longwave radiation would need to need to be pre-processed (spatial resolution, bias correction etc.).

L254: Liquid water instantly distributed following normal distribution, is this a widely used approach and do you know how it behaves in reality? Clearly this is a tricky parameter to know, but are there implications of this assumption?

The approach was extensively investigated by Marchenko et al. (2017). We also refer to Vandecrux et al. (2020) for a comprehensive review of available melt water percolation parameters.

L280: Does this initial subsurface condition differ greatly from just running the first year multiple times - would this not be a more precise way to evolve the starting conditions rather than a 20 year run? Either way its not known if this has a substantial impact on the results. Were there early borehole temperatures as well?

As e.g. visible from figure 4, the first year was particularly positive. It appears more robust, to use average conditions. The used approach represents *in situ* data well: The early borehole

temperatures indicate temperate conditions in the accumulation area as simulated by the model. The model furthermore represents the subsurface densities reasonably well.

L283: So 'adjust' here means optimisation of parameters in some way described later on .. are these point scale optimizations and how does that relate to the fact that the glacier outline is updated over time - are all optimization points within the final glacier boundary?

Yes, all optimization points are within the final boundary (big dots in figure 1b).

L303: In the comparisons does the SMB include refreezing (so the modelled climatic mass balance cf Cogley et al., 2011) or not? Looks like it from L347, but maybe clarify here?

As written, the simulated surface balance is used for comparison to stake data. We will add a reference to the description of the surface balance given earlier in the manuscript.

Table 2: Those parameters identified as optimized in this table are only those optimized through a whole model simulation. This seems a bit misleading as others are also optimized but to other data externally to a full model simulation, perhaps instead just separate literature/optimized values.

ok.

L317: Clarify pressure decay from historic average and modern average (October 2011 - 2020) as you mention the modern timeseries was only used from 2014 onwards due to data gaps (L102).

This is a typo. Will be corrected.

L336: This precipitation correction factor is performed within a SEB model simulation like the albedo optimization or is it simply applied to the precipitation timeseries compared to the glaciological data? If the latter doesn't it belong in the forcing data preparation section? Here do you compare the March accumulation in the monthly snowpits to the March accumulation in the forcing data - if so this could be stated more clearly. Maybe also justify the choice of summer bias correction (1.15) from a single different site, is this representative of a range of sites for this instrument?

Yes, the precipitation correction factor for winter months is based on a similar simulation as e.g. the albedo optimization using March snow pit data for evaluation. The value for summer is not from a single site but based on a comprehensive analysis.

L345: "Overall, the mass balance of Abramov glacier is negative for the years from 1968/1969 to 2019/2020." - what is meant here? The mean over the whole period? Give a value or consider giving the cumulative mass balance numerically (which could also be included in Figure for as a line plot on a secondary axis).

Yes, there is an overall mass loss. Will be clarified and completed with numbers.

L347: Consider how to refer to the results - e.g. rather than listing what is shown in each figure you could focus on the description of what is revealed and cite the figure: e.g. The distributed

mean annual mass balance for 1968/1969-2019/2020 is shown in Fig. 3. The mass balance of Abramov glacier is predominantly negative for the years from 1968/1969 to 2019/2020, despite part of this being taken up by glacier retreat, and shows no significant trend in annual mass balances (+0.0002mw.e. a1, p-value=0.979). The most negative modelled mass balances occur at the start of the period, while the two decades between 1978 and 1998 are characterised by almost balanced mass budget, thereafter becoming more negative again (Figure 4). The modelled mass balance gradient shows that accumulation is lowest during the first decade (1968/1969-1977/1978) and highest during the last modelled decade (2008/2009-2017/2018), and ablation is largest during the first decade, followed by the second last decade (1998/1999-2007/2008).

Good point. Thanks for the suggestion. Will be improved accordingly.

Figure 3: Please highlight the different glacier extents with dates or emphasize the fact that the average over the whole timespan with the adjustment of the glacier area leads to the distinct pattern. It might also be nice to report a long term mean ELA in this or the gradient figure (5b).

Will add the extents to Figure 3 and also visualize the ELA in the gradient figure.

Figure 4: It woulds be nice to show the cumulative modelled mass balance on a secondary axis.

The cumulative mass balance will be shown here as well.

---

*L357: Again here, avoid just listing what is in the figures, but rather tell us what the figures *show*. The title makes it seems like this is part of the model validation but in fact here you also show the model calibration results, right? Suggest moving the calibration part to the end of section 3.5. (e.g. The model calibration was performed on eight snow pits locations at the end of March and annual mass balance measurements from 19 stakes and up to eight snow pit locations for the period 1968/1969-1997/1998; the comparison of optimized model simulations to these in-situ observations used for model calibration shows a stronger RMSE for the annual mass balance, and despite some bias in the linear fit, the datasets generally approach the 1:1 linear fit line (Figure 6).) Then here in this section on validation say: e.g. For model validation we compare the simulated mass balance to a set of point measurements not used in the calibration, which can therefore provide an independent measure of model performance. This comparison is based on 146 stake readings not used in model calibration for 1968/1969-1997/1998 (Figure 7a) and all available point mass balance measurements for 2011/2012-2019/2020 (Figure 7b). In both cases the comparison scatter plots cluster around the unity fit line, but inherit the bias tendencies seen in the calibration (Figure 6)) Consider colouring the points in Figure 6 by decade or year as then you might also reveal periods of better/worse model calibration, which could be interesting.*

Will be improved as suggested.

Figure 6: The different axis labelling in (b) might create a false impression, please homogenize them

Agreed.

Figure 7: So the performance isn't any better in the measurement period (a) compared to down-scaled ERA5 period (b)? This could also be rephrased to the fact that both periods show good

agreement, but given the impact of the two different forcing data sets, I'm a bit hesitant to jump to that conclusion. Further: Could you maybe indicate the number of observations here just to be clear. The scatterplot also doesn't give us an indication of the point or location of the measurements. Generally, I would have preferred to clearly see the points on a map maybe with a highlighting of the calibration sites and also evaluate how you chose the calibration measurement. For the ablation stakes you use less than 10% to calibrate - where does this number come from?

The calibration/validation stakes are shown in the current Figure 1 using different symbols. We will use clearer symbols for distinction and refer to the figure here. We will also add the number of points used for validation in Figure 7. A very direct comparison of the model performance for both periods is hampered by the different amount of available measurements which are furthermore based to the ablation area (especially for the recent years). The bias is also larger for the recent years mainly due to an underestimation of ablation and possibly due to a lack of accumulation measurements in the scatter plot. Will plot the mean annual bias in figure 4 (along with a visualisation of uncertainties of measurements) to allow for a better interpretation.

Figure 8: Depending on the direction of the analysis, the decadal analysis could be further complemented with monthly or seasonal plots.

Agreed.

Table 4: Here annual trends, but earlier figure showed decadal MB, would be interesting to see if there are certain breakpoints in the timeseries corresponding to external forcing (disturbances in atmospheric forcing etc.)?

Will perform additional analysis to detect breakpoints and report in Table 4.

L371: Interesting that you have decreasing RH but increasing precipitation? Does need explanation?

This is likely related to changes in distinct seasons. Will report trends for seasons in the supplementary material.

L396: the correspondence in Fig 13 seems really good to me, perhaps comment in line here that these biases are accompanying a generally good agreement, as evidenced in the figure? Is this match in line with expectations from modelled densification?

Will add a comment in the text as suggested.

L403: Do you return to explain this cooling trend in the discussion?

A discussion of the cooling related to the reduced refreezing (less latent heat release) will be added to the discussion.

Figure 11 and 12 could be a bit larger.

ok.

L421: Can you add the value you find in the text here to facilitate comparison with the other studies for which you do quote the values.

ok.

Figure 15: Its cool to see these data which have not been widely included in previous Abramov studies. What is the added value of this study compared to the others, how do the different studies compare to each other? (Please check the data of the Barandun paper - its labelled 2015 in legend and 2021 in cap- tion). This should be mentioned in the discussion. For example, due to the updating of the glacier extent and glacier elevation, the results of this study are more positive compared to the assumption of static glacier extents in the other studies? You could also include your model simulation results from the recent period and the more recent mass balances studies to complete the picture (e.g. given in Table 6 Barandun et al., 2018 and Barandun et al., 2021 Zenodo data set). This would also open the possibility to compare your modelled results to geodetic mass balances for the glacier available e.g. Hugonnet et al, 2021 and/or Miles et al., 2021 which includes an ASTER based elevation change rate based on results of Brun et al., 2017 (https://doi.org/10.5281/zenodo.3843292).

The figure will be extended for recent years as suggested. Our study includes a process-based simulation of internal accumulation which is not included in most of the previous estimates (Barandun et al. (2015) use a simple parametrisation) and also a more detailed (energy-balance) approach at the surface.

L472: Move these first 2 sentences to the conclusions, as they seem a nice summary of what you have already said in this discussion section.

Agreed.

L485: This statement of performance is related to the validation against point measurements being better for the main run than the alternative run? Where do we see this in the results shown? Or is it just in the supplement. Suggest moving this statement to the part on uncertainty due to forcing data as it deserves more full discussion.

See answer to comment on line 513.

L490: Ideally this section would provide a formal uncertainty assessment. In your situation, as there are many unquantified errors it would be good to include a sensitivity analysis. If this is at all feasible, we strongly encourage you to add this, otherwise please justify why not. You can then describe your approach at end of this paragraph: " As the sources of uncertainty are diverse and sometimes not readily quantifiable, we here discuss the general sources and likely implications, rather than producing a formal uncertainty assessment for the modelled output."

We agree that this section should be improved. A comprehensive uncertainty assessment for a point location such as e.g. performed by Machguth et al. (2008) is beyond the scope of this study. Different types of errors could be accounted for by such a Monte Carly simulation. However, it gets extremely laborious and computationally heavy to reasonably well quantify the various errors and build the actual simulation, accounting for all identified sources of error. Therefore, we will rather present a relatively comprehensive sensitivity analysis. This analysis includes

tests for different parameter values following Klok and Oerlemans (2002), model forcings and initial conditions.

L512: What is the implication of these large differences? Does that mean that the more recent timeframe is wrongly estimated? How can this happen if the bias correction is also based on this period? To get an overview over this, plots of the different forcing datasets (at least in Supplement) could help.

First and foremost, these differences emphasize the high model sensitivity to the meteorological forcing and highlight that the lack of continuous weather station data is problematic. On the other hand, the comparison to *in situ* data (cf. Fig. VII) show that the bias are not systematically different for both mass balance simulations. The overall simulation for the recent years is thus not systematically wrong. As the used cloud forcing is assumed to be prone to larger uncertainties than other forcings, the sensitivity to different cloud cover forcing will be separately evaluated and discussed.

L513: Where have you shown that the standard run is 'better' than the alternative run? Can you here add some quantification on the goodness of fit of the main and alternative run? For example in Figure S5 it would be good to see a comparisons with the actual measured MB for this period and report on which forcing data version matches most closely. This is of major importance as, if the alternative forcing produces a very poor match, then it calls into question the validity of the simulated mass balance from 1998 onwards, right? Also consider if this is not better in the main paper rather than the supplement?

Measurements of surface accumulation were overall better represented by the original model run (mean bias: -0.21 m w.e., RMS: 0.44 m w.e. for 1980-1998) than by the alternative model run (mean bias: -0.35 m w.e., RMS: 0.56 m w.e. for 1980-1998), which is relevant for our study with a strong focus on accumulation. The overall bias of measured surface ablation was lower for the alternative model run (+0.18 vs. +0.39 m w.e. for 1980-1998). Whereas the original run similarly underestimated the ablation throughout the entire ablation area (linear regression for 1980-1998 between measured and modelled ablation original run: $y = 0.88 \times x + 0.16$ ($R^2 = 0.84$), too negative mass balances were simulated around the ELA which compensated for the too low ablation rates on the glacier tongue (linear regression for 1980-1998 between measured and modelled ablation alternative run: $y = 0.72 \times x - 0.37$ ($R^2 = 0.80$)). The respective scatter plots will be provided in the supplementary material (cf. Figs. V and VI)). Furthermore, the annual mean surface accumulation and ablation bias of both model runs will be visualized in Figure S5 (cf. Fig. VII) which will be moved to the main manuscript. It is visible from Figure VIIb that (i) both model runs systematically underestimate surface accumulation and ablation and that (ii) there is no systematic difference between both runs (see also answer to previous comment).

3 Figures

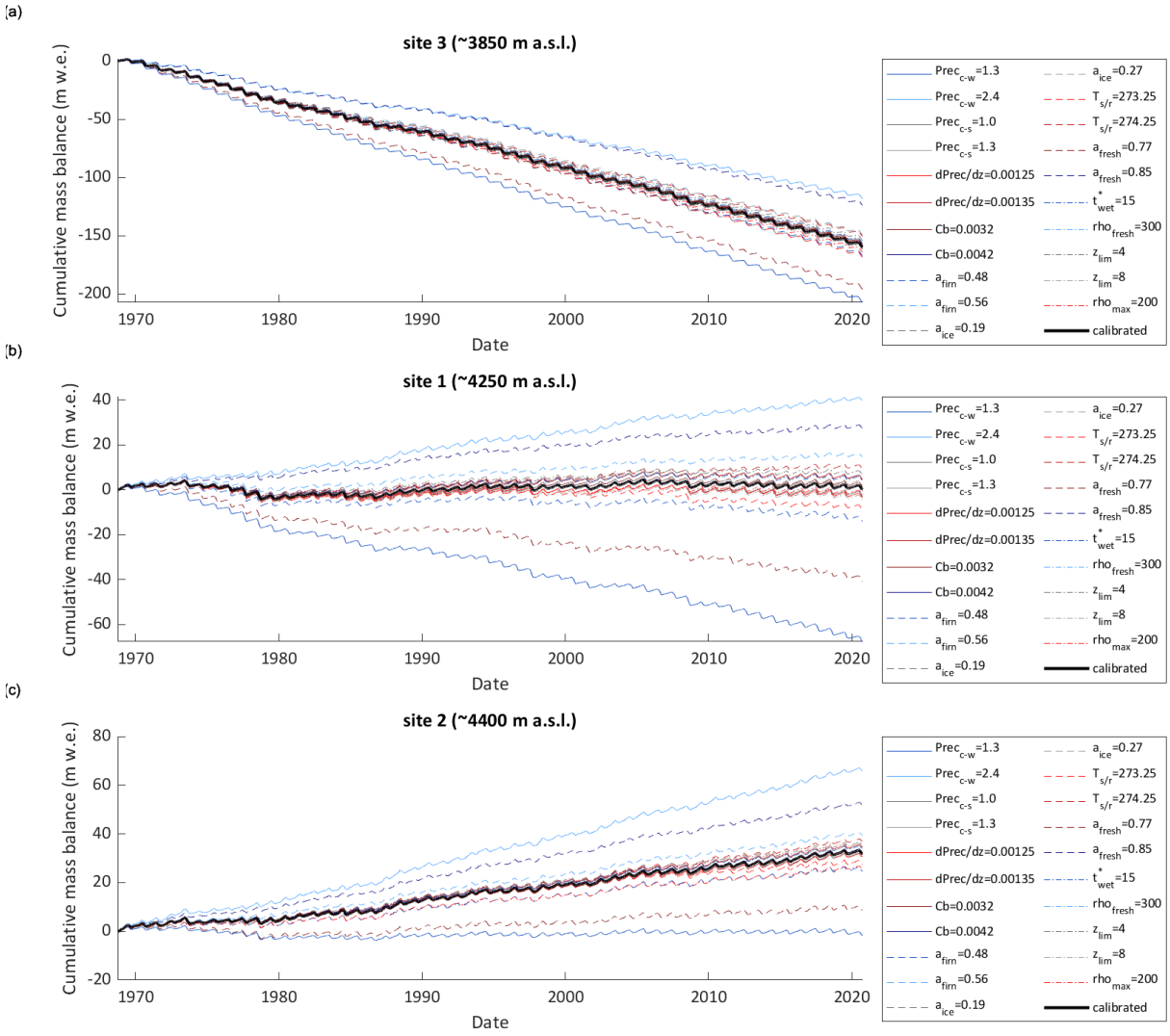


Figure I: Cumulative mass balance evolution for disturbed model parameters using ranges from literature. The cumulative mass balance is shown for three selected points (ablation area (a), lower accumulation area (b) and accumulation area (c)).

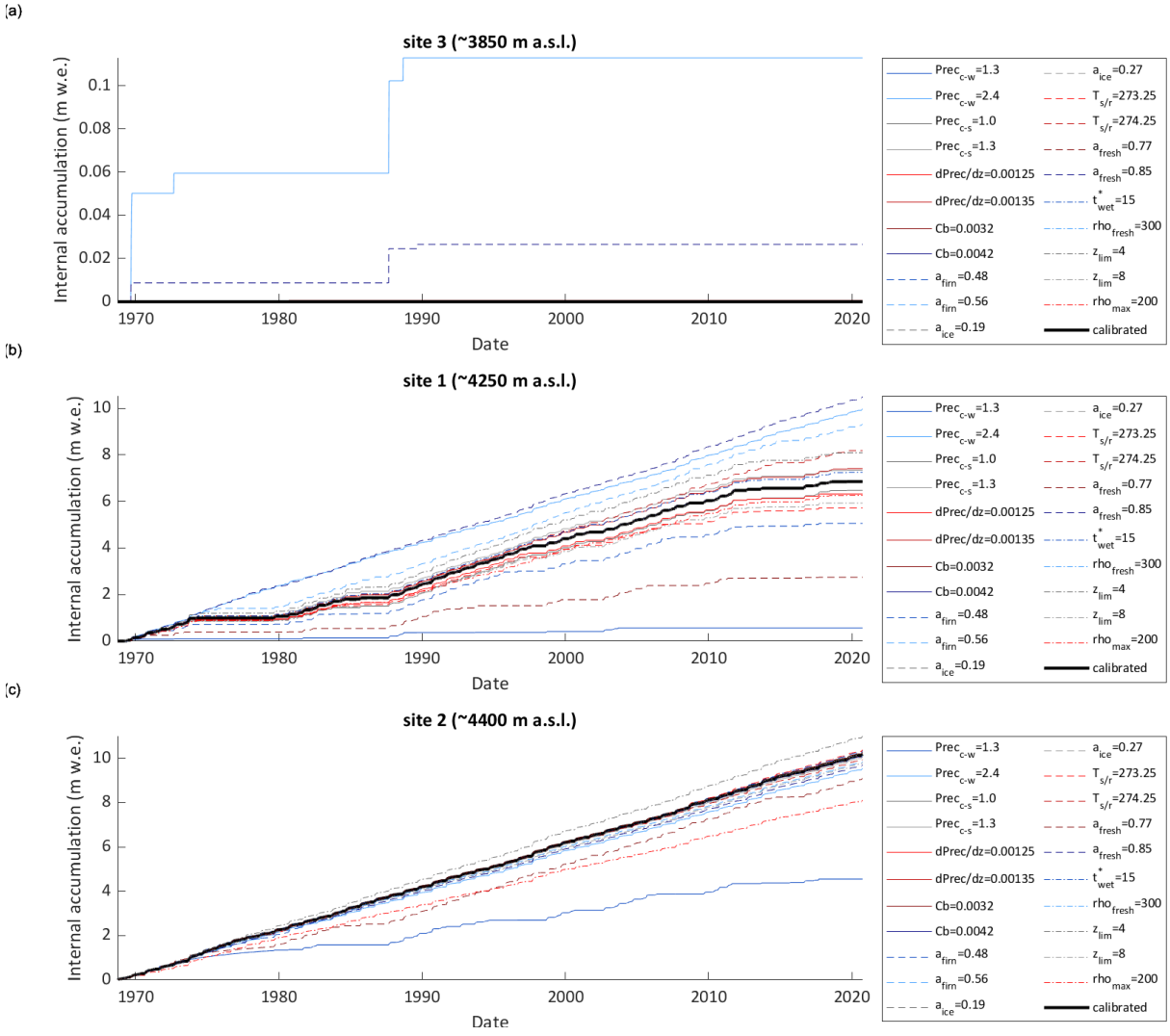


Figure II: Cumulative internal accumulation evolution for disturbed model parameters using ranges from literature. The cumulative internal accumulation is shown for three selected points (ablation area (a), lower accumulation area (b) and accumulation area (c)).

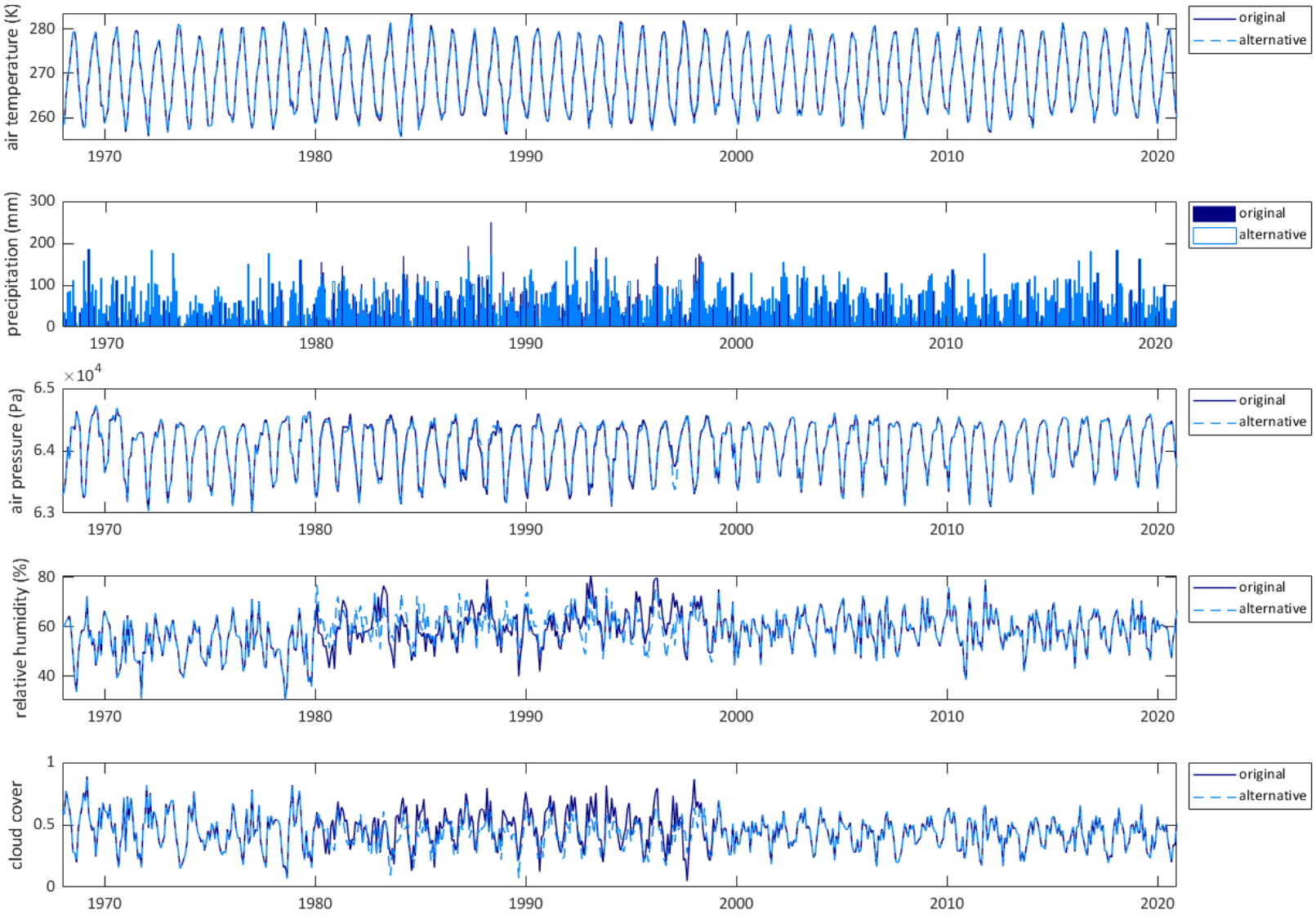


Figure III: Original and alternative model forcing for entire simulation period at monthly resolution.

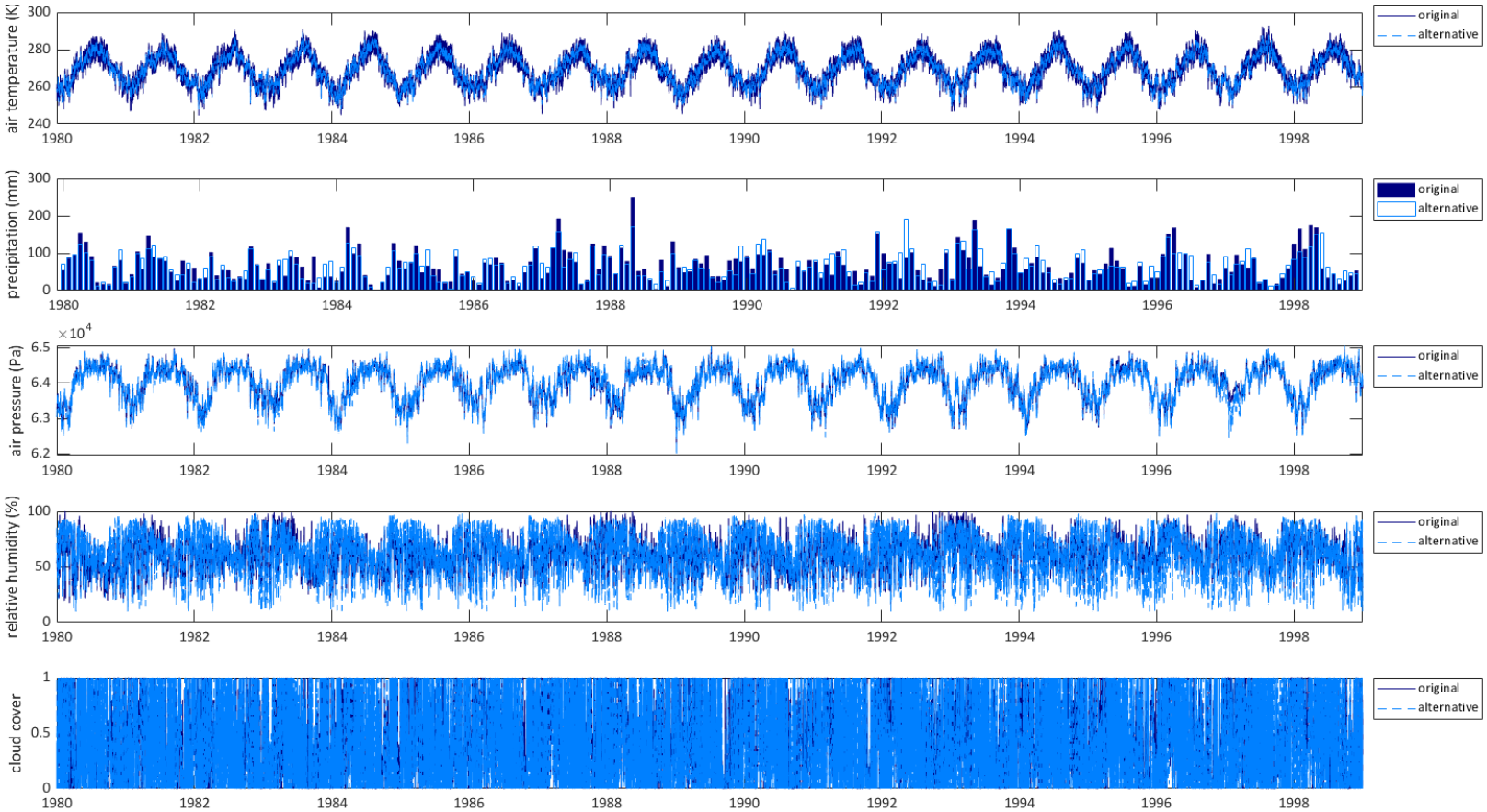
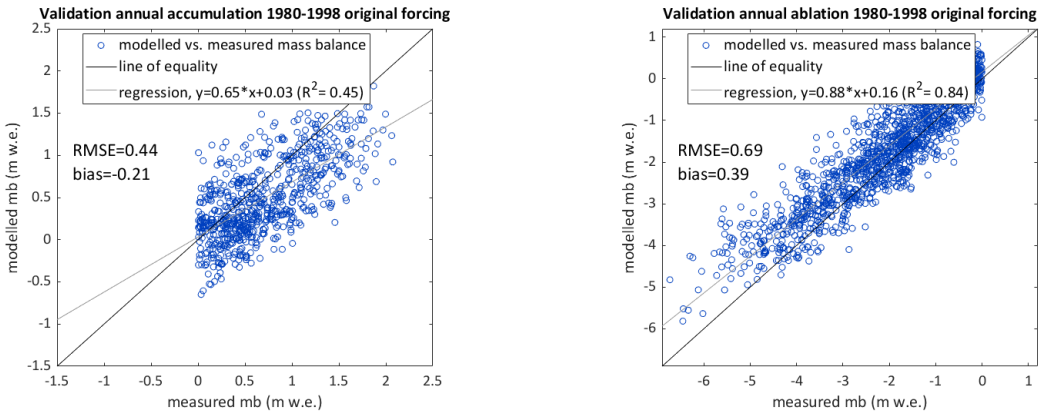
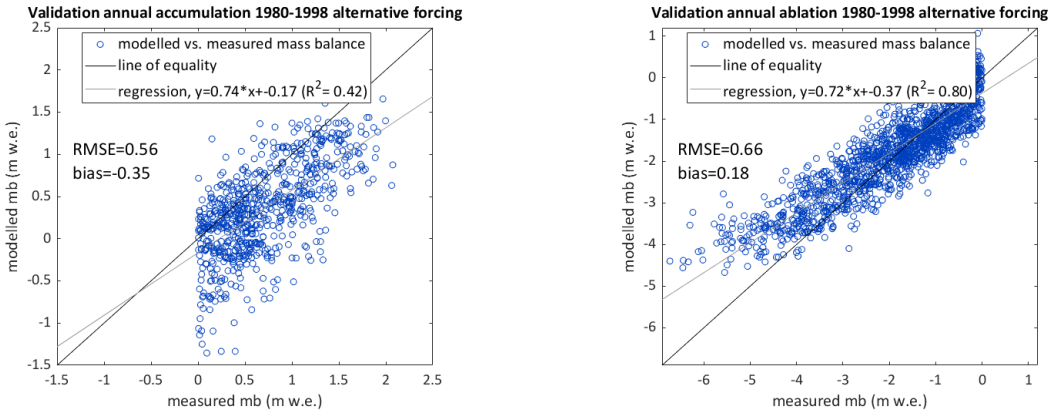


Figure IV: Original and alternative model forcing for overlapping time period (1980-1998) at three-hourly resolution (air temperature, air pressure, relative humidity and cloud cover) and monthly precipitation sums are shown.



(a) (b)

Figure V: Validation of original model run against an independent set of point surface mass balance measurements for 1979/1980-1997/1998. Measured versus modelled annual surface accumulation (a) and surface ablation (b).



(a) (b)

Figure VI: Validation of alternative model run against an independent set of point surface mass balance measurements for 1979/1980-1997/1998. Measured versus modelled annual surface accumulation (a) and surface ablation (b).

Note: Figure 4 in the main manuscript will be completed with a visualisation of the mean annual bias as it is done in Fig. VII. In addition the uncertainty of the validation measurements will be visualized (quantification following Thibert et al. (2008)).

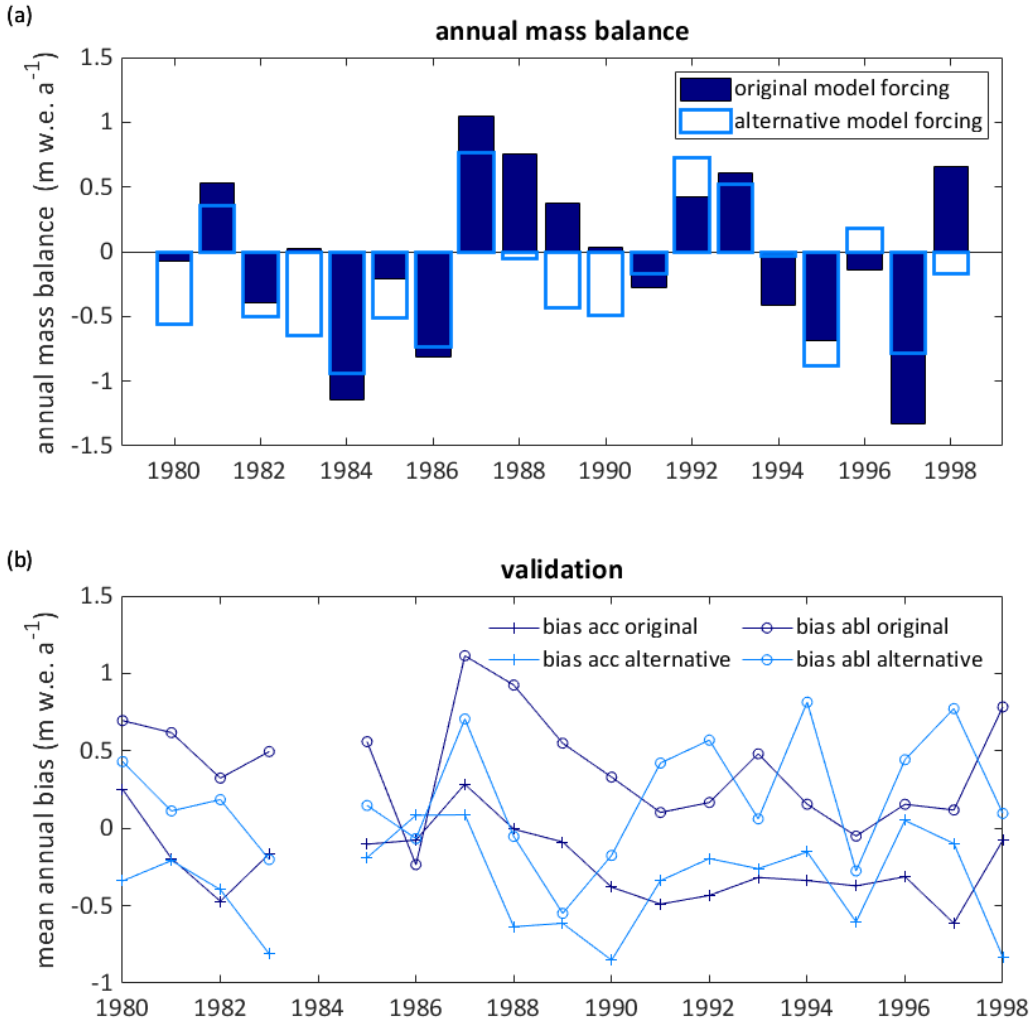


Figure VII: Comparison of mean annual mass balances of the original model run and the alternative model run (a) and mean annual bias between modelled and measured point accumulation for both model runs (b).

References

Barandun, M., Huss, M., Sold, L., Farinotti, D., Azisov, E., Salzmann, N., Usubaliev, R., Merkushkin, A., Hoelzle, M., A.Merkushkin, and Hoelzle, M.: Re-analysis of seasonal mass balance at Abramov glacier 1968-2014, Journal of Glaciology, 61, 1103–1117, https://doi.org/10.3189/2015JoG14J239, 2015.

Denzinger, F., Machguth, H., Barandun, M., Berthier, E., Girod, L., Kronenberg, M., Usubaliev, R., and Hoelzle, M.: Geodetic mass balance of Abramov Glacier from 1975 to 2015, Journal of Glaciology, 67, 331–342, https://doi.org/10.1017/jog.2020.108, 2021.

Førland, E. J. and Hanssen-Bauer, I.: Increased Precipitation in the Norwegian Arctic: True or False?, Climatic Change, 46, 485–509, https://doi.org/10.1023/A:1005613304674, 2000.

Greuell, W., Knap, W. H., and Smeets, P. C.: Elevational changes in meteorological variables along a mid-latitude glacier during summer, Journal of Geophysical Research, 102, 25 941–25 954, https://doi.org/10.1029/97JD02083, 1997.

Klok, E. J. and Oerlemans, J.: Model study of the spatial distribution of the energy and mass balance of Morteratschgletscher, Switzerland, Journal of Glaciology, 48, 505–518, https://doi.org/10.3189/172756502781831133, 2002.

Kronenberg, M., MacHguth, H., Eichler, A., Schwikowski, M., and Hoelzle, M.: Comparison of historical and recent accumulation rates on Abramov Glacier, Pamir Alay, Journal of Glaciology, 67, 253–268, https://doi.org/10.1017/jog.2020.103, 2021.

Machguth, H., Purves, R. S., Oerlemans, J., Hoelzle, M., and Paul, F.: Exploring uncertainty in glacier mass balance modelling with Monte Carlo simulation, The Cryosphere, 2, 191–204, 2008.

Marchenko, S., Pelt, W. V., Claremar, B., Pohjola, V., Pettersson, R., Machguth, H., and Reijmer, C.: Parameterizing deep water percolation improves subsurface temperature simulations by a multilayer firn model, Frontiers in Earth Science, 5, 1–20, https://doi.org/10.3389/feart.2017.00016, 2017.

Pertziger, F.: Abramov Glacier Data Reference Book: Climate, Runoff, Mass Balance, Central Asian Hydrometeorlogical Institute, Tashkent, 1996.

Sevruk, B.: Systematischer Niederschlagsmessfehler in der Schweiz, in: Der Niederschlag in der Schweiz, Beiträge zur Geologie der Schweiz - Hydrologie, edited by Sevruk, B., vol. 31, pp. 65–75, 1985.

Stainbank, W. D.: Simulation of Abramov glacier energy balance and firn properties, Msc thesis, University of Fribourg, 2018.

Thibert, E., Blanc, R., Vincent, C., and Eckert, N.: Glaciological and Volumetric Mass Balance Measurements: Error Analysis over 51 years for the Sarennes glacier, French Alps, Journal of Glaciology, 54, 522–532, 2008.

van Pelt, W., Pohjola, V., Pettersson, R., Marchenko, S., Kohler, J., Luks, B., Hagen, J. O., Schuler, T. V., Dunse, T., Noël, B., and Reijmer, C.: A long-term dataset of climatic mass balance , snow conditions , and runoff in Svalbard (1957 – 2018), The Cryosphere, 13, 2259–2280, https://doi.org/10.5194/tc-13-2259-2019, 2019.

van Pelt, W. J. J., Oerlemans, J., Reijmer, C. H., Pohjola, V. A., Pettersson, R., and Van Angelen, J. H.: Simulating melt, runoff and refreezing on Nordenskiöldbreen, Svalbard, using a coupled snow and energy balance model, The Cryosphere, 6, 641–659, https://doi.org/10.5194/tc-6-641-2012, 2012.

Vandecrux, B., Mottram, R., L. Langen, P., S. Fausto, R., Olesen, M., Max Stevens, C., Verjans, V., Leeson, A., Ligtenberg, S., Kuipers Munneke, P., Marchenko, S., van Pelt, W., R. Meyer, C., B. Simonsen, S., Heilig, A., Samimi, S., Marshall, S., MacHguth, H., MacFerrin, M., Niwano, M., Miller, O., I. Voss, C., and E. Box, J.: The firn meltwater Retention Model Intercomparison Project (RetMIP): Evaluation of nine firn models at four weather station sites on the Greenland ice sheet, Cryosphere, 14, 3785–3810, https://doi.org/10.5194/tc-14-3785-2020, 2020.