In this document, we respond to the comments of reviewer 2 one by one. Whenever some entirely new text has been added to the manuscript, it has been added in italics and in red.

The proposed revised with and without track changes is added as a supplementary .pdf file.

Reviewer 2

In this manuscript, Van Tricht et al. use a 3D higher-order ice flow model to study 6 ice masses in the western part of Tien Shan in great detail. They used the 1820-1850 average climatic conditions to derive an initial state, representing the ice masses at the end of the LIA. The model was carefully calibrated based on a different set of available observations (e.g SMB measurements, moraines, ice thickness data) and various meteorological input data are used for the past. A validation based on geodetic mass balances shows a good agreement with independent observations. To provide projections over the 21th century, the model is forced with CMIP6 data (multi-model mean) considering 5 shared socioeconomic pathways and a no change scenario, representing nowadays conditions. All results show a clear reduction of ice masses, but with individual responses which are discussed in further detail. The manuscript is of great interest of the community as it gives a good picture about the detailed modelling of those 6 ice masses, especially in their calibration based on a bunch of different sources. The validations showed in the paper, give confidence that the model performs well for the future as well. The paper is well written and presented in a clear and well-structured way. However, I have the feeling that some parts of the manuscript deserve more consideration (rather equivalent to major revisions) in the study.
We would like to thank the reviewer for the detailed review. We have implemented most of the suggestions which improved the quality and clarity of the manuscript.

General comments

[RGC1.1] 'hot model' problem (selection of GCMs): A subset of the CMIP6 models are 'too hot' and simply taking an average of all available gcms leads to higher projections of warming than the IPCC's assessed-warming averages and consequently to an overestimation of glacier mass loss. Hausfather et al. (2022) suggested the following in order to avoid the overestimation: models with a TCR that lies outside the 'likely' (66% likelihood range) of 1.4-2.2 °C should be screened out. You should check out your GCM selection and consider a recalculation of multi-model means. Hint: The supplementary of Hausfather et al. (2022) contains a table with the TCR values for the different GCM's
We appreciate your suggestion, and as a response, we have made a careful selection of GCMs for our revised version. Specifically, we have chosen to include only the GCMs that have a TCR value ranging between 1.4 and 2.2°C. As a result of this selection process, we observed a slight decrease in the overall temperature increase in the multi-model mean. Subsequently, we have updated all of our results and figures based on the latest version of the temperature and precipitation forcing. We added on lines 192-193: "*Only the models with a transient climate response inside the 66% likelihood range are considered (Hausfather et al., 2022) (up to 18 GCMs)*."

[RGC1.2] Uncertainty range of projections: The flow model is only forced with a multi model mean. It is known that the spread over different GCMs is huge and thus not only the mean value, but also some kind of error range (e.g. standard deviation) should be taken into account here as well. It would give more strength to the results.

We agree with the reviewer. For the revised manuscript, we also ran simulations using temperature forcing with a gradual increase or decrease with respect to the multi-model mean to reach ± the standard deviation at the end of the century. The uncertainty ranges are clearly reflected in the results shown in figure 7.

We added at lines 223-228: "*In addition to conducting simulations using the multi-model mean of every SSP scenario, we further generate two additional scenarios for each SSP. These scenarios are designed to incorporate a gradual increase or decrease in temperature, aiming to reach approximately the standard deviation of the projected temperatures at the end of the century across the various GCMs. This approach is implemented to account for the inherent uncertainty between the different GCMs and provides a more comprehensive representation of the possible temperature variations within each SSP. For precipitation, we always consider the multi-model mean.*"

[RGC1.3] Better highlighting of the strength of the story: Especially the abstract and the introduction are at some points misleading and contradictory. On the one hand you are saying that conducting detailed studies to understand different processes is important (p.2, l.44). This is exactly the strength of your study and should be more highlighted than it is at the moment. But often you try to set the context to a regional importance. As your study is not saying anything about a regional pattern, you should try to avoid any kind of those statements and focus much more of the strength of studying those 6 ice masses in such detail. Comments about regional importance are only misleading the reader.

We agree with the reviewer, and we removed accordingly some sentences to focus more on the strength of studying 6 individual ice masses in detail. We believe this has improved the readability of the manuscript.

Specific comments

[RC1.1] p.1, l.8-10: An abstract should only concise a summary of the content of the paper. These three sentences explain more the context of the story behind and therefore belong more to the introduction than to the abstract. It raises false hopes to the reader as regional patterns are not part of the simulations nor being discussed in the paper

We agree with the reviewer and we removed these three sentences.

[RC1.2] p.1, l.17: 'since the Little Ice Age' → 'since the end of the Little Ice Age' (or even better the exact time period (years) directly)

Thank you for the suggestion. We modified this into "since the end of the Little Ice Age (1850)"

[RC1.3] p.2, l.42: But none of the 6 ice masses is debris covered, why is this relevant for the study?

We agree with the reviewer, and we removed this sentence.

[RC1.4] p.2, l. 43-44: This is true, but it calls for a regional study simulating ALL glaciers in the region individually and not for a study picking only 6 glaciers out. I'm not sure if the sentence as it is, is really a good argumentation for the relevance of the study.

We agree with the reviewer and as this was also mentioned by reviewer 3, we decided to omit this sentence and incorporate the remained of this paragraph into paragraph 1.

[RC1.5] p.2, l.53: 'today's ice masses are still reacting to past climate conditions' : Please, add a reference

We added a reference to *Christian et al. (2018).*

[RC1.6] In general, I'm missing in section 2.1 explanations why those 6 ice masses were selected. Do those glaciers represent well other glaciers in the region? Are those the glaciers with most observations? What was the reason to pick exactly those glaciers?

We selected these 6 ice masses because they are all characterised by detailed measurements needed to apply our model accurately (ice thickness and mass balance), because there are several meteorological stations on and near the ice masses, and because the ice masses are currently part of monitoring programs. Next to that, with this selection, we cover a range of different "type" of ice masses (ice cap, thin glacier, thick glacier, polythermal glacier, temperate glacier, etc.) which makes it interesting to compare the findings for the different glaciers.

Furthermore, these six glaciers were chosen in the past to initiate measurements due to their possession of several representative characteristics.

We added a paragraph in the text to elaborate on this on lines 60-65: *"The selection of the six ice masses is based on several factors, including the availability of detailed measurements necessary for model application, such as ice thickness and mass balance, the proximity of several meteorological stations to the ice masses, the distribution of the ice masses across four subregions of the Kyrgyz Tien Shan with different climates, and their inclusion in ongoing monitoring programs. Additionally, our chosen sample encompasses a diverse range of ice mass types, including an ice cap, thin and thick glaciers, polythermal, cold and temperate glaciers, thereby enabling us to make comparisons between the various findings. "*

[RC1.7] p.3, l.71: 'annually': Please, add the information since when.

Done.

[RC1.8] p.3, l.94: '500 km to the west of the previous 4': Is there an explanation why you choose one glacier so far away from all the others?

We refer to our answer in RC1.6

[RC1.9] p.5, Fig.1:
– add a legend for the altitude range of the underlying DEM in the upper panel.
– black outlines: What is the source of the 2021 outlines?
What is the date of the Sentinel-2 data?
– I would prefer lon/lat coordinates. – the ticks and labels on the axes are quite small.
– Please, order the information in the caption by panels. You first explain the background data of the 6 small figures, than the outlines of the upper panel, than the outlines of the 6 figures and again some background information about the upper panel.
A label number for the upper panel would be good as well.

We incorporated the suggestions except for the coordinates which we preferred to keep in UTM, as it is more appropriate to estimate distances on the plots and because we also illustrate the maps in UTM coordinates for clarity.

[RC1.10] 2.3 Meteorological data. You have used many different sources for the temperature data series. This will require a bias correction to combine them. How was this done?

We added info on this procedure at lines 158-159: *"All different time series used to create the meteorological dataset are debiased using overlapping periods. For a detailed description on how the data time series for these two stations have been created, we refer to Van Tricht et al. (2021b)."*

And lines 169-171:

"Like for the other two datasets, the different time series used for the Golubin AWS dataset are debiased using overlapping periods and matched for interannual variability using the standard deviation in the common period."

[RC1.11] p.8, l. 181: There are CMIP6 historical runs starting in 1850, why did you not use this one (if available).

We opted to utilize local observational data obtained from the vicinity of the glaciers, as we anticipate that these measurements would provide a more precise representation of the climatic conditions specific to the glacier's surroundings.

[RC1.12] p.9, l.188: 'are repeated until 2100': a random shuffling of the years 2001-2021 would be more realistic than just a repetition of the years.

We acknowledge the reviewer's suggestion, and following your recommendation, we have now implemented a random shuffling of the years between 2001 and 2021.

[RC1.13] p.9, l.189: Does this refer to the 'no change' scenario only?

No, it is also done for the other years. We reordered this paragraph to make this clearer.

[RC1.14] p.9, l.191: 'by using an observed data sequence': Which one exactly?

This depends on the series. For Kumtor and Chon-Kyzyl-Suu, we used 2007-2021 while for Golubin, we used 2014-2021. We added this in the text on lines 205-207: "*Finally, as was done for the historical period (Van Tricht et al., 2021b), the monthly data series are downscaled to hourly resolution by using an observed data sequence between 2007 and 2021 for the Kumtor and Chon-Kyzyl-Suu dataset 2014-2021 for the Golubin dataset*."

[RC1.15] l.193-194: Please, take my general comment concerning the 'hot-models' into account. This might largely influence your mean value!
We refer to our answer in RGC1.1

[RC1.16] p.10, Fig.2: – I'm confused a bit by the figure. You mentioned in the text, that you have downloaded the historical data for 1984-2014. The plot shows data between 1950 and 2021. Where does the data from 1950-1984 come from? – Please, rethink the choice of colors here. Unfortunately your color selection did not take colorblindness (e.g. red-green) into account. Perhaps you could use the same colors for the SSP as are used in the IPCC?
We use data from measurements in the vicinity of the glaciers for the historical period (up to 2022). Only from 2023 onwards, we use data from the GCMs. We updated the figure using the colours of the IPCC and excluded the "hot models".

[RC1.17] p.12, l.253: 'spaced more closely': Please, give an exact resolution here.
Thank you for mentioning this. With the version we use for this study, the layers are equally spaced. Therefore, we changed this sentence.

[RC1.18] 3.2 Surface mass balance model p.13, Tab.3: Please, use a consistent type of brackets. In previous figures and tables round brackets were use.
Done.

[RC1.19] p.13, l.296-297: 'typical response time of 30-60 years': Please, add a reference here
Done. We added a reference to *Johannesson et al. (1989)*.

[RC1.20] p.14, l.303: enhancement factor (m): This looks like a unit and not like a parameter name.
Done. We changed the parameter name to "E".

[RC1.21] p.14, l.305: Please, also give the step size for the basal sliding parameter
We used a predefined set of values for this range. Therefore, we added *for a predefined set of values* in the text.

[RC1.22] p.14, Tab 4: Please, use different kinds of brackets for parameter names and units.
Done

[RC1.23] p.14, l.321-322: But this additional velocity observation was not taken into account? Why?

This observation was taken into account. We reordered this section to make this clearer.

[RC1.24] p.15, l.354: 'uniform mass balance bias': Does this mean that the same bias is applied for all years?

This was indeed the case. However, following the suggestion of reviewer 1, we opted to change the approach and we are now using constant temperature and/or precipitation bias.

[RC1.25] p.15, l.354: '.. bias is searched': By which method and what is the step size used here?

We changed our method by now using temperature and/or precipitation biases. We changed this in the text at lines 385-287:

"These biases are determined by assessing combinations between ±0-0.2°C for temperature and ±0-15% for precipitation, while keeping the impact of the biases on the modelled ice thickness low."

[RC1.26] p.16, l.362: 'geometry of the LIA': The 1850 geometry would be better here.

Done

[RC1.27] p.16, l.362-363: 'larger than present-day outline': I 'can't see the present-day outline in Fig. 3. So either add them or highlight them better (with a label and description in the caption).

Added.

[RC1.28] p.16, l. 370-371: I find it very critical that the neighboring glaciers are not modeled, if they were merged in the past. This will change ice dynamics and might have a huge impact on the ice flow as it changes the boundary conditions. When does the ice cap looses the connections?

The neighbouring glaciers are not modelled separately, but the ice flow from the Grigoriev ice cap towards these glaciers is modelled. The ice is automatically removed when it flows out of the Grigoriev ice cap outline. The outlet glaciers do not have an important effect on the dynamics of the ice cap itself.

[RC1.29] p.16, Fig.3: – It would be great to add the present-day outline in the plot, as well as the in flowline profiles. – I would prefer lon/lat coordinates. – the axes ticks and the labels are too small (in all plots, the ice thickness plots and in the flowline plots) – add year of the Sentinel-2 data

Done.

[RC1.30] 4.2 Ice evolution between 1850 and 2021
p. 17, l. 384: How are the flowlines derived (in general and the 4 separated). Did you draw them manually or did you applied any kind of algorithm? You did not mentioned this before, but as you used a flowline model, this is crucial.

The flowlines were drawn manually perpendicular to the elevation contours, and optimised after modelling the flow field. The determined flowlines are only used for visualising the results. We do not use a flowline model, but a 3D ice-flow model.

[RC1.31] p.18, l. 411: 'with measured ice thickness': Which ones exactly? The ones you have used for the calibration? If so, this is not a validation. This would need a different kind of data that was not used for calibration.

We agree. We removed "validation" in these sentences and reordered this paragraph to speak about comparison rather than validation. In addition, we created a section 4.4 in which we only use validation data (geodetic mass balance and SRTM data).

[RC1.32] p.18, l. 414: 'areas with GPR measurements': Where are those regions? From where can I read this information? Is it possible to mark the areas in the profile plots?

We added references to the publications where the measurements were presented.

[RC1.33] p.19, l. 421-424: To which date/years does this refer to?

These observations refer to different years between 2012 and 2021. We added this in the text at line 456:

"for various years spanning from 2012 to 2021"

[RC1.34] p.21, Fig.7:
– Please, use for the 'no change' line the same style and color as in Fig. 2
– caption: 'Historical and future evolution': 2000-2021 isn't really historical, when you present in the same study results up to 1850 – time range: The time range is again a bit confusing. Why are you starting here from 2000? You showed before the evolution from 1850-2021. Why not combining this to one time series? Here again one wonders, if the historical line stem from CMIP6 historical data or represents a part of the previous results.

We only use CMIP6 data from 2023 onwards. Until 2022, we use local data. We modified the figure now plotting the results from 1850 onwards.

[RC1.35] p. 24, Fig. 9:
– The patch color for present day outline can only purely be distinguished from the ssp119 and ssp585 scenario.
– I would prefer lon/lat coordinates.
– The ticks and labels are too small

We modified the figure by now using the colours of the IPCC for the different SSP scenarios. Besides, we increased the size of the labels.

[RC1.36] p.25, Fig. 10 – It would be great, if you could mark the peaks of the individual ssp with a vertical colored line labeled by the year of the peak. – Why does the time series not start with the year 1850? – Why aren't all lines equal to 1 at t=2021?

We modified the figure by plotting the results starting in 1850. We also indicate the peak runoff by calculating the average peak of SSP1-2.6, SSP2-4.5, SSP3-7.0 and SP5-8.5. Now, all lines start at 1 in 2021.

[RC1.37] p.26, l. 553: One could get a better impression about the uncertainty range here, if not only the multi-model-mean, but also an error range of the GCMs would have been calculated (e.g. based on standard deviation). I highly recommend to calculate those values as well.

See our answer in RGC1.2.

[RC1.38] p.26, l. 558-569: There even exist a global study (across multiple glacier models) that gives more insights about uncertainty sources of glacier projections. You should definitely update this part based on findings from Marzeion et al. (2020)

Done. We reordered this section, made this more extensive and also referred to Marzeion et al. (2020).

[RC1.39] Comparison with geodetic mass balances -> rather see this section as part of section 4.3 (as validation with data not used for calibration). Please, consider to reorder. • title: 'Comparison' → 'Validation'
p.27, Tab. 6: A colour marking of the cells would improve the readability of the table, because one could see at a first glance which values are low/high/close to zero

Done.

[RC1.40] p.28, l.600: 'present day' : Please, use the same spelling in the hole text. Previously you wrote 'present-day'.

We removed "present-day".

[RC1.41] p.28, l. 600: 'present-day observations': To which ones you refer here, the geodetic mass balances? Please, clarify.

Done. We removed replaced "present-day observations" by "geodetic mass balances".

[RC1.42] Additional comment I would appreciate an additional section about the hardware requirements and the performance (running time). This would give the reader an impression about how easy/hard it will be to repeat those experiments for more glaciers/regions.

We do not have a good record of the spent core hours and therefore decided not to include specific numbers on the computational time. Nevertheless, the running time depends clearly on the glacier type and size.