General response

We thank the editor Kang Yang for obtaining two valuable reviews and Sam Herreid and the anonymous reviewer for their thorough and constructive comments on our manuscript. On the following pages, we address the reviewers' comments point by point. The reviewers' comments are highlighted in grey. We hope that our responses will qualify us to submit a revised version of the manuscript.

Response to Referee Comment 1 (RC2)

The main strength of the paper is to present an open-source pipeline for the processing of the TIR imagery, which has been an ongoing concern when using the black box, proprietary software to extract temperatures from TIR images. The paper is straightforward and easy to read, and it is well-placed within the existing literature that relates debris-thickness and surface temperature. I enjoyed reading it.
Thanks for the positive feedback.
I think the manuscript would benefit from a more candid assessment of the performance of their temperature maps, which give results for the snow/ice surface temperatures that seem to have a strong spatially consistent bias, and the applications of both the empirical model and the energy-balance model given the limitations of the input data. Applying these methods is not a straightforward process, which is discussed qualitatively in the paper, but only in general terms. In my opinion, a more quantitative assessment of the model sensitivities would be beneficial.
We have taken this comment as an opportunity to perform a simple sensitivity analysis (for the results see NEW Fig. 13), to (re)calculate the surface temperatures for the snow/ice area using a proper mask (see Fig. 11), and to analyse the distribution of the snow/ice surface temperatures (see NEW Fig. 6).
Another component that is not much discussed in the paper, but I think should be added, is suggestions on how to upscale this method to a larger domain, considering the limited area tested here, and the possible complications in areas with thicker debris, as the maximum thickness here is 15 cm.
We now discuss ideas for the upscaling of this method in more details in the discussion.
L9: typo: orthophoto
corrected
L9: I suggest you mention that you calibrate the energy-balance approach “with an empirical or calibrated inverse surface energy balance”.
Revised sentence: “Finally, a high-resolution debris thickness map is derived from the corrected thermal orthophoto using an empirical or inverse surface energy balance model that relates

surface temperature to debris thickness and is calibrated against in-situ measurements.”

L84: Could you give the elevation (and elevation range) of the study area, as well as the same characteristics like slope, aspect (which influence the energy-balance application)

We added a sentence with the requested information at the end of the paragraph: “The elevation of the surveyed debris-covered area ranges from 2425 to 2480 m a.s.l. and is cut by two parallel meltwater streams running from northeast to southwest. The inclined areas (mean slope = 15°; standard deviation = 10°) face mainly towards northwest and southeast.” Maps of slope and aspect are now also included in Fig. 8.

L117: A strength here is that the flights were so short that it is unlikely that there was a significant change in surface temperature during that time, but it could still have happened. I would like to see somewhere (likely discussion) some mention of possible biases caused by changes in surface temperature during the UAV surveys, especially when it comes to aiming to do longer flights to cover larger areas. This could be made worst in partly overcast weather if cloud movement is occurring rapidly, casting changing shadows, or if flights occur late afternoon or morning. Could you mention if the temperature varied between the flight time (did it warm up or cool down, or was air temperature stable?)

No, according to our debris temperature measurements and the meteorological data from the nearby weather stations, there was no considerable change in air and surface temperature during the short survey period.

We elaborate now in more detail on the potential impact of varying meteorological conditions on the surface temperature measurements in the discussion.

Table 1: Could you change the units from ha to m2 or km2 (as in the text, L119, or at least give the conversion between ha and m2?)

We chose the unit ha for better readability. 1 ha equals 10000 m² or 0.01 km². As ha is a widely used unit for area of land (also in academia) and officially accepted for use with the SI, we think it is not necessary to update Table 1, and also do not mention the conversion factors in the text.

L146: Could you point to the figure comparing measured surface temp to UAV-corrected temperature?

We included a reference to Fig. 9.

L156: few – can you give an actual number?

The sentence was modified: “Two automatic weather station...are located 7 and 5 kilometers away from the Kanderfirn and continuously measure...”

Table 2, and elsewhere: It would be interesting to read a bit more about the uncertainties linked with using such estimates from other locations to derive the energy balance of this highly specific study site. There is a mismatch of complexity here, where you use estimates for the input to the energy-balance model, compared to the high-resolution data you use to derive the debris thickness. I think more information and discussion of the application of the energy-balance model would be interesting in the results or discussion section because it is not a trivial thing to obtain these results.

We copy and paste here our response to the other reviewer: “We fully agree that on-site measurements would be beneficial and best practice. Unfortunately, we neither had access to nor funding for the installation of an automatic weather station on the glacier. The experiment was conducted within a student’s project without any additional funding.” We discuss the potential uncertainties regarding the energy balance model in more detail in the Discussion and also provide suggestions for the further improvement of the methodology. In addition, we performed a sensitivity analysis to assess the uncertainties related to different meteorological parameters (i.e. air temperature, incoming short- and longwave radiation, wind speed) and debris properties (i.e. albedo and effective thermal conductivity).

L174: Could you have aimed for lower overlap and achieved a longer flight to cover a larger area? Could this be a suggestion for other studies? (similar to L192)

No, not really. A sufficient overlap is crucial for the generation of accurate orthophotos and DSMs. Instead of reducing the overlap, using fixed-wing or hybrid UAVs that are capable of surveying larger areas are recommended for future studies focusing on debris-thickness mapping.

L209-211: I disagree with this point. The main novel aspect of this paper is presenting an open processing for UAV-based debris thickness but then you don’t use the open access too and instead used Pix4D. I think that you should have used only WebODM for the UAV visual UAV instead of using the pix4d if you were only going to use one version. Also, I suggest moving this sentence to the end of the paragraph to avoid talking about thermal, then visual, then thermal again and avoid confusion.

We copy and paste the response to the other reviewer here: “The inconsistency originates from the history of the workflow development. We failed at the beginning to produce accurate and suitable thermal orthophotos with the open-source pipeline. Anyway, we fully agree and revised the manuscript accordingly. We now present the complete open-source pipeline and have a standalone validation section using the results from the processing with the proprietary software package.”

L218: according? Do you mean corresponding? Would radiation be radiative?

Yes, we meant corresponding. Modified.

Neither nor. “Radiation temperature” was replaced by “brightness temperature” throughout the text.

L254: This is similar to the approach discussed in Baker et al (2019)?

Baker, EA, Lautz LK, McKenzie JM and Aubry-Wake C (2019) Improving the accuracy of time-lapse thermal infrared imaging for hydrologic applications. Journal of Hydrology 571,60 – 70. doi:10.1016/j.jhydrol.2019.01.053 https://doi.org/10.1016/j.jhydrol.2019.01.053

Thanks for the hint. We included the reference.

Table 4. Any suggestion on how/why they are so variable over such a small area, and for a measurement that occurred all at the same time? What kind of bias occurred if you take the average value for reflected temperature when it is obviously very variable? Was there a spatial pattern to reflect temperature, and could you create a distributed field of reflected apparent temperature?

Good question. We do not have a definite answer, but we could imagine that the main reason for the large spread in the reflected apparent temperature at the different aluminium GCPs is the varying angle, distance and direction between the camera and GCPs during the survey. Most of the GCPs were located slightly off the flight path (see Fig. 1). Although we tried to place the GCPs in flat areas, some of them might have been slightly inclined. Mixed-pixel effects and statistical interpolations during the photogrammetric processing might also be an explanation…

Anyway, compared to other uncertainties in the methodology, the bias related to the reflected apparent temperature is negligible. Varying the reflected apparent temperature of -6.8 °C by ±3.2 °C (standard deviation) would change the surface temperature by only ±0.15 °C.

Figure 5: Could you have a different symbol for the training and validation? It's not the easiest to differentiate them at the moment with the shades of grey.

We updated Fig. 5 and used different symbols for the training and validation data.

L355: You have measured debris thickness and surface temperature from the empirical approach (and near-surface from the in-situ small sensors). Could you calculate keff instead of calibrating it? What kind of values would you get if you tried to derive them from the measurement instead?

As we did not manage to install loggers at different depths in the debris, we cannot calculate keff. We refrain from combining the mapped debris surface temperature and measured nea-surface debris temperatures to calculate keff as the small difference in depth would lead to considerable uncertainties. Moreover, it is unclear how representative k of the 2-cm-thick shale stone would be with respect to keff of the debris layer. Bisset et al. (2022) for example show that keff can vary considerably with depth.

L297: To increase the validity of your approach, you could remove these pixels that you know are not valued by creating a different mask that removed the location of the rocks an

The larger rocks that are scattered across the clean ice are numerous. Digitising and masking them manually would take ages. The only way to detect them automatically would be to apply a temperature threshold, but that's exactly what we did to remove these "outliers" from the statistical analysis.

L361: These scattered boulders – did you remove them from your analysis (removed from your temperature maps) to calculate the debris thickness field? You should probably not use your empirical equation beyond the bounds of the measurements that were used to create your empirical fit, as these modelled thicknesses above 13 cm are extrapolated and not well constrained at all. It looks like you don't have much-modelled thickness above 13cm for the empirical approach, so it might not be a big issue in this case, but something to be careful about.

No, we did not exclude them from the empirical debris thickness map because of practical reasons (see comment above). We agree that results of the empirical model beyond the bounds of the measurements should be interpreted with care, but as Fig. 6 shows, debris thicknesses above 13 cm are almost absent.

Fig 7: I think debris temperature should be before debris thickness in the results, as it is an analysis step that comes before – the measured and mapped temperature influences the modelled thickness, not the other way around. Also, instead, of having outliers in the results that are artifacts of the methods, I think these outliers should be removed from the image by designing a mask that does not include them. This figure presents the debris temperature, so it would be appropriate to remove the GCP from the results.

We agree and exchanged Fig. 6 (NEW Fig. 7) and Fig. 7 (NEW Fig. 6). We created a mask to remove the GCPs from the surface temperature map and interpolated the area of the GCPs using the surrounding pixels and inverse distance weighting (see NEW Section 3.6.4 “Snow and ice masking and GCP correction”).

Fig 8. The DSM inset could be called (e) for clarity. In the legend, you set the crevasse and ice cliff as the same feature. Is it the same feature that you refer to, or a different part of the subset image? Can you add what m stands for in the legend/caption?

DSM was called (e). Yes, it is the same feature. The crevasse is the black line and the ice cliffs are the dark greyish areas to the north. M indicates the medial moraine mentioned in the text. The caption was updated accordingly.

L376: How large is it in m x m?

ca. 170 x 390 m

L380: Could you add like Stream 1 and Stream 2 on the fig 8m, like you tag the moraine location?

We labelled Stream 1 and Stream 2 in Fig. 8. and referenced it in the text.

Section 4.4.: It would in good to see if these numbers for the difference in surface temperature are similar to those in other studies that find a bias between TIR and in-situ debris temperature in the discussion section. Is it even a useful way to assess if TIR imagery is correct as they are measuring different things?

In the study by Kraaijenbrink et al. (2018), where measured and mapped debris surface temperatures were compared, the deviation was not quantified (to our knowledge). However, the linear relationship looks similar. We think that temperature loggers installed close to the surface are useful to be sure that the thermal measurements are plausible and that the camera does not exhibit a distinct cold or warm bias. Since they measure different things, the absolute values of the two variables might differ, but they should be highly correlated in any case.

L395: The 162 image number is already mentioned in L194.

Deleted

L400: A bit contradictory to mention that it is interesting but then put it in the supplementary.

“More interesting” with respect to the small-scale deviations in the same supplemental figure, not “more interesting” than the surface temperature map. We referenced Fig. B1 earlier to make that clear.

Fig 10-11: I suggest you mask the section of the mages that should not be considered with an emissivity of 0.97. I suggest you segment the cover type (debris, ice) and only show the temperature for the section of the image where the result is valid (where it uses the proper emissivity). only sow the debris are

We followed your suggestion and created a surface type raster with the classes ice/snow and debris, and assigned the respective emissivities to create one consistent surface temperature map. In Fig. 11, we now only show the surface temperatures for the surface type classes snow and ice.

L403: Similar to the processing of the orthophoto, I think it is misleading to have this paper about an open-access pipeline but then use the result from the proprietary software in the results. I think it would be much more interesting to showcase the result with the open source, and then we could also see the processing artifacts that are mentioned above.

We followed your suggestion and revised the manuscript accordingly.

L405: I suggest removing this. This is an artifact of the processing that should not be considered a result.

Deleted as the GCPs were removed from the surface temperature map (see earlier comment).

L408 : given the actual number instead of “about 11”

Obsolete as the sentence was deleted.

Fig 11, L422-425: I find this result quite concerning. Some patterns make sense: the warmer margin for snow/ice temperature near the debris (a,b,c), but others, such as the snow patch of h-g going from ~+1 to -4 (and likely more if it wasn’t masked?), and the strong gradient in temperature between the edge and middle of the image around (c), going from -4 to +4, to -4 over ~150m. To me, these look like edge effects in the processing and suggest that only the middle third of your image is valid. If there is a reason to think that these distributed temperatures are valid, it should be explained. If these are edge effects, then the image should be segmented to only keep the middle section and remove these weird gradients. You mention these artifacts in L422, but then the next sentence states that they perform well enough, and I do not agree with that statement.

We (re)calculated the deviation from the melting point now only for the bare-ice and snow-covered area using the new ice/snow mask (see Fig. 11). We also computed the frequency distribution for the surface temperatures mapped across the bare-ice and snow-covered area (see NEW Fig. 6). The largest deviations of up to ±4 °C can be indeed observed at the edges. This is not surpring as the number of images available for the processing is smaller here and vignetting effects might therefore be more pronounced. In the central part, non-uniformities within individual images associated with external effects on the camera (ambient air, radiation etc.), are probably averaged out during the photogrammetric processing. The deviation patterns shown in Fig. 11 most likely do not represent real surface temperature variations, but rather originate from an insufficient calibration of non-uniformities in the raw thermal images. We elaborate in more detail on the challenges related to UAV-based thermal imaging in the revised discussion, summarise important operational recommendations and outline possible technical solutions. For example, we highly recommend that future studies relying on accurate absolute surface temperatures deploy a portable and light-weight calibrator (heated shutter) that has recently become available and can increase the accuracy of uncooled microbolometers considerably (see Virtue et al. 2021). For an uncooled microbolometer as the one used in this study and taking into account the complexity of thermal imaging on an alpine glacier, the accuracy seems reasonable. Most of the pixel values in the bare ice and snow-covered area (70 %) are in the range of ±2 °C. Only 7 % of the pixels deviate by more than ±3 °C. We added a section in the discussion to elaborate in more detail on the uncertainties in the modelled debris thickness caused by the inaccuracies in the surface temperature map.

L422-425: Could the relatively good average (0.4C, 0.3C) be linked to the fact that the errors are centred on 0 and so it gives a good average, when in fact it is quite spread out? Could you add information on how you define the +/- for the uncertainty of your numbers? Could you show the distribution of the ice and snow temperature in Figure 7, in addition to the map of the whole area (maybe even show both the snow/ice segmented distribution and the debris mask-only distribution?)

We added the different distributions (snow/ice, debris, total area) in the NEW Fig. 6 (previous 7). See also comment above. We used the standard deviation of the individual pixel values (mapped surface temperature minus melting point) as a measure for uncertainty. The average/median slightly above the melting point indicates that there is no consistent bias or shift in the temperature measurements. In our view, the standard devidation in the order of ±1 °C is reasonbale for an uncooled microbolometer and first attempt, and justifies the use of the mapped surface temperatures for debris thickness modelling. For comparison, Gök et al. (2023), who performed thermal imaging on a different glacier in Switzerland, state a mean ice surface temperature of 0.72-2.26 °C (and standard deviation of up to 2.87 °C).

L422-425: My understanding and experience is also that thermal infrared cameras can be quite accurate from pixel to pixel within one image, but can be quite off in terms of absolute temperature. It makes me more cautious about these spatial patterns in the temperature of the ice surface. I understand that the camera accuracy says +5 to -5, but that covers a much too large range and really limits the possibility to investigate TIR use for glacier melt, where a much smaller temperature range has large consequences for melt.

The camera accuracy given by the manufacturer is ±5 °C or 5% of the reading in the range of -25 to +135 °C. Our results indicate that the accuracy in the range of -5 to +35 °C seem to be much better (probably ±1-2 °C), depending of course on the ambient conditions during the survey.

Fig 11: Also, maybe put a dotted box around the area that is used for the quantitative assessment – is that the area near where the temperature artifact is in (g)?

We created an ice/snow mask and perform the quantitative assessment only for this area (see NEW Fig. 6 and Fig. 11).

L429: couple millimetres -> use the actual number?

Updated: “... ranges from around 1 centimeter up to 15.5 cm (Fig. 6).”

L431: relatively thick -> How thick?

Updated: “Besides this, the debris layer appears to be relatively thick (ca. 5-10~cm) in the elevated area between the parallel supraglacial meltwater streams (Fig. 12d).”

L435: Could they be linked to wetness level, which would influence the conductivity and the surface temperature, instead of thickness?

We cannot completely rule out this alternative explanation, but we think it is rather unlikely. The visual orthophoto does not indicate any spatial variations in the wettnes level. Moreover, if the wettnes level would vary in space, then the question would be wich condition could lead to the observed stripe-like pattern? Besides thickness, also the grain size, porosity and lithology of the debris layer could alter the water level/content. But then again the question would be which process leads to stripe-like variations in grain size, porosity, lithology etc.

L440: Have you tested how other parameters are sensitive in the model? Can you justify calibrating this one instead of a selection of other parameters?

Yes, we also tested other parameters (air temperature, shortwave radiation, longwave radiation, wind speed and debris albedo) in the surface energy balance model (see NEW Fig. 13). The model responds sensitively to the tested parameters, but the uncertainty introduced by these parameters concerns mainly thick debris (>10 cm). Since the (effective) thermal conductivity is the parameter that is most critical for thin debris (<10 cm), which is characteristic for the studied glacier, we chose this one for the model calibration.

L440: I find it interesting that you use a fairly complex inverse energy-balance approach, but then calibrate it with one parameter to fit the data. The other component of that model is also highly uncertain – meteorology, albedo, etc, are all very specific to the study area, and potentially even variable throughout your study area. It would be interesting to hear more about how suitable it is to use spatially homogenous input to the model when you are looking at a variable terrain. You mention this very briefly in L445, but maybe you could elaborate slightly more in the discussion.

We used the theoretical model of Evatt et al. (2015) as it is the only one that is able to reproduce the characteristic features of the empirical Østrem curve. Although it is not the main point of this comment, we would like to emphasise that the complex model used here could be easily replaced by any other model in the presented open-source pipeline.

While some of the meteorological data (e.g. air temperature or incoming shortwave radiation) extrapolated from the weather stations nearby probably describe the local conditions on the glacier fairly well, other parameters (such as incoming longwave radiation or wind speed) are subject to higher uncertainties. It is also true that some of the input parameters (for example debris albedo or air tmperature) can be expected to vary across the surveyed area. We elobrate more on the related uncertainties and the refinement of the methodology.

L470: But, if you have a fixed wing that can flight longer and further, you are likely to be able to find a patch or snow or smooth meadow where you can land adjacent to the glacier…. Another limitation to uav is that they are realy bulking to hike in to remote sites.

Theoretically yes, but

1) the legal framework in almost all countries does not support flying out of sight (or at least requires a comprehensive safety concept)

2) flying out of sight in mountainous terrain is risky

3) smooth areas are often difficult to find in glacierised and deeply-incised valleys

Yes, they are bulky, but considering the constant technological advancements in the field, smaller high-endurance UAVs that are more suitable for applications in the mountains should become available in the upcoming years...

Figure 15: I don’t think this figure is needed as you have the RMSE values on fig 13.

It is not necessarily needed, but as it is small and nicely shows the model improvement and the remaining error related to other uncertainties in the workflow, we prefer to keep it in the article.

L476: Aubry-Wake et al., 2022 might be a helpful reference about the different factors that influence TIR acquisition for debris thickness measurements because the conclusions are very different – midday is not a good time to have a strong relationship between surface temperature and debris thickness, but you focus on very thin debris overall, so a different dataset!

Aubry-Wake, C., Lamontagne-Hallé, P., Baraër, M., McKenzie, J., & Pomeroy, J. (2023). Using ground-based thermal imagery to estimate debris thickness over glacial ice: Fieldwork considerations to improve the effectiveness. Journal of Glaciology, 69(274), 353-369. doi:10.1017/jog.2022.67

Thanks for pointing out this publication. We were not aware of it before and have considered it now in the discussion.

L486: A nuance that I think needs to be clarified here is that precise surface temperatures (that are consistent together) are needed for empirical thickness calculation, but for empirical calculation, the measurements do not need to be accurate. They could be biased, but as long as they are consistent, it works. However, for energy-balance approaches, you need both accurate and precise measurements of surface temperature.

Yes, accurate surface temperature are less important for the empirical approach than for the physical approach, but temporal changes in the environmental and meteorological conditions would nevertheless also affect the empirical relationship between Ts and hd. We therefore slightly adjusted the sentence: “Since accurate and consistent surface temperatures are...”

L518: But those snow/ice temperature maps still show a large vignetting effect, so potentially these atmospheric corrections and reflected radiation are not enough to obtain precise and consistent temperature maps. These vignette effects have not been mentioned in ground-based measurements, so they might originate from the processing of the UAV images.

Vignetting effects might be more pronounced in UAV-based than in ground-based infrared thermography as the camera is more exposed to changes in the ambient conditions during the flight than at one (maybe wind-shielded) location on the ground. The larger surface temperature deviations along the margin of the orthophoto might be related to vignetting effects or in general to changes in ambient conditions (e.g. wind speed, differential heating of the camera housing related to flight direct and position of the sun). The effect is probably less pronounced in the central part of the orthophoto as the non-uniformities in individual photos should be averaged out to a certain degree during the photogrammetric processing in areas where sufficient thermal images are available. However, overall the accuracy seems reasonable for an uncooled microbolometer. The middle boxplot in NEW Fig. 6 indicates that 50 % of the surface temperature values of the ice and snow area are in the range from -0.9 to 1.8 °C. Nevertheless, we recommend that future studies that rely on accurate surface temperature measurements deploy a portable and light-weight calibrator as presented in Virtue et al. (2021) to get rid of non-uniformities in the individual images.

L530: This might be a good moment to point out that using small in-situ temperature sensors tucked in the debris, as you did, has a fairly limited use to assess the performance of UAV-based temperature.

As stated before, we think that installing temperature loggers close to the surface is justified in proof-of-concept study to be sure that the thermal measurements are plausible and that the camera does not exhibit a distinct cold or warm bias. However, we agree that the loggers are dispensable once the robustness and reliability of the camera in harsh environments has been confirmed by several studies.

We revised the sentence: “In the absence of independent debris (surface) temperature measurements, such as in the study by Gök et al. (2022), possible biases or shifts in the UAV-based debris surface temperature recordings might be overlooked”.

L558: But even applying a site-specific empirical model can lead to erroneous debris thickness if the model is based on flawed data, like a bias in the sampling of the debris thickness, or a surface temperature that does not correlate well with debris thickness due to time of day, as discussed by Herreid (2022) and Aubry-Wake et al. (2023).

Yes, the overall methodology depends in principle on a strong correlation of surface temperature and debris thickness. We discuss this aspect now in more detail.

L559: But you do not account for spatial variation right? I appreciate that you discuss how it would be nicer to have in-situ measurements of meteorology, I would like a bit more quantitative analysis of this. For example, what kind of error or uncertainty are you erasing by calibrating keff? Is the model as sensitive to air temperature as it is keff? Applying an energy balance model to a small area like this, with high spatial heterogeneity, is hard to get right even with data coming from the site, so it would be good to have a bit more information on the sensitivity of the debris thickness to the model application.

No, we have not accounted for spatial variations in this study. However, the usage of gridded data instead of single parameter values can be easily integrated in the presented open-source pipeline.

We managed to measure spatial air temperature variations in parallel with the thermal imaging in a follow-up experiment, but haven’t completed the analysis yet. Please refer to the new sections in the revised manuscript and to the comments further above regarding the model sensitivity.