Responses to Reviewer 1, Sam Herreid

TC-2022-113: "*High-resolution debris cover mapping using UAV-derived thermal imagery: limits and opportunities*" by Gök, et al.

Thank you kindly for taking the time to carefully read our manuscript and for the constructive comments which helped us to improve it. In the following, we will address all comments point by point.

"High-resolution debris cover mapping using UAV-derived thermal imagery: limits and opportunities" by Gök et al. uses thermal data collected during eight repeat UAV flights throughout one day of a portion of Tsijiore-Nouve Glacier in Switzerland to evaluate estimates of debris cover thickness. The authors use two established methods to estimate debris cover, but move research on this subject forward by their thorough consideration of thermal data drift and offset, the continued advancement of using thermal cameras on UAVs, and the repetition of flights over the course of one diurnal cycle. The authors do a nice job evaluating the debris thickness estimates against field measurements, however, I think a key deficiency in the results are debris thickness difference maps showing the variability of a quantity, which, if estimated perfectly, should be a grid of zeros. Even integrating to show the change (i.e. error) in total debris volume for this study domain might, for example, help future erosion rate studies looking to use debris thickness changes as a way to quantify bedrock erosion to understand the magnitude of error in these methods. Below are mostly minor, but some more major in line comments.

ABSTRACT

L10: be explicit: e.g. supraglacial debris surface temperature or LST over a debris cover

> *Changed to: "… we present land surface temperatures (LST) of supraglacial debris cover …"*

L11: can you say at x cm resolution rather than high?

> *Changed to: "… measured from an unpiloted aerial vehicle (UAV) at high (15cm) spatial resolution."*

L16: Can you make here in the abstract a statement on which method you found to be most efficient successful/efficient if the RMSE is essentially the same for both? This could be formulated as a recommendation for future studies.

> *Added sentence: "Although the rational curve approach requires in-situ field measurements, the approach is less sensitive to uncertainties LST measurements*

compared to the SEBM approach. However, the requirement of debris thickness measurements can be an inhibiting factor that supports the SEBM."

L16: This isn't exactly true that diurnal variability in LST controls the relationship between LST and debris thickness. Maybe you mean to say the success of a rational function to express the relationship varies predictably with time of day? Maybe also conditionally with clouds cover and precipitation?

Good point. We changed the sentence to: "Because LST varies throughout the day. the success of a rational function to express the relationship between LST and debris thickness also varies predictably with the time of day."

L20 This last sentence is unclear, do you mean independent measurements of LST or a better camera on your UAV? And also a little bit of a downer. Glaciologists can't practically afford a cooled microbolometer and they certainly aren't going on UAVs anytime soon. I would either make a more achievable suggestion or try a more positive closing sentence.

Uncooled microbolometers are generally able to perform temperature measurements with sufficient accuracy. However, insulating the device from influential factors in high mountain regions (e.g. sensor cools down after UAV take-off, wind, or direct shortwave radiation) is challenging. But we think there is room for improvement to perform more accurate measurements using this sensor type. We changed the sentence to: "… which are challenging to achieve with uncooled sensors in high mountain landscapes."

INTRODUCTION

L23: Add citation(s) to first sentence, Scherler et al, 2018: Herreid and Pellicciotti, 2020

Done.

L23: Add citation supporting "debris cover is generally rather thin, usually less than a meter"

We added the reference to Rounce et al., 2021

L24: I would change "profound" to something like "exponentially compounding influence …where melt rates are accelerated…" or at least "it can have a profound"

Changed to: … "it can have a profound" …

L26: by "heavily" do you mean spatially or in thickness?

Both. To be clearer we changed the sentence to: "Consequently, glaciers with widespread and thick debris cover can persist longer at lower…"

L29: I think it is established that the advancing you cite here is occurring more because of elevation and precipitation rates, not as a function of debris cover.

We agree and change the sentence to: "…with some glaciers being stationary and some retreating."

L33: inappropriate location for citations

We agree and moved the citation to the end of the sentence.

L39: Cite the recent studies

Added: Kaushik et al. 2022

L39: There are practical reasons why there are more studies on extent change rather than thickness change, it's an easier problem with a better control on errors, it's intermediate studies like this one that may slowly help add the z component to change analyses.

Yes, we agree with this view. Yet, we don't feel this needs to be added to the text.

L43: The list followed by "vary rapidly" reads poorly, maybe you mean the abundance of these features can increase or decrease rapidly or the abundance can vary dramatically between different glaciers or different parts of one glacier?

Good point. Changed to: "In particular, the abundance of supraglacial streams, ponds and ice cliffs can increase or decrease rapidly across the glacier surface (Anderson et al., 2021)."

L45-46: what does "distribution" constitute beyond knowledge of "extent and thickness"? Also, a continuous model of debris thickness is more likely, not comprehensive observations.

Indeed, "distribution" is redundant and was removed. Although we agree that model-derived debris thickness maps are more likely than observation-based, we think the development of such models will always require comprehensive observations for evaluation.

L49: Incorrect citation to Ostrem 1959, this paper established the relation between debris thickness and melt rate.

Citation removed.

L49: In (3) explain where the debris thickness comes from, a residual from the methods you describe.

We changed the sentence to: "…the estimation of sub-debris melt by DEM differencing and converting melt rate to debris thickness based on the Østrem-curve, …"

L52: I think you might mean "fitting … to" rather than "extrapolating … using"

Changed accordingly

L57: What do you mean by recent technological advances? Uncooled microbolometers have been around for quite a while. One limitation may be that it's a single sensor imaging a single wavelength window rather than a split window approach some thermal satellite sensors use to solve for and remove atmospheric attenuation.

Good point, we changed the sentence to: "Most LST-based approaches to estimate debris-cover thickness have focussed on satellite imagery, whereas studies employing near-ground image acquisition in high resolution are less frequent."

L60: What are the opportunities and limits? What is the research question? And what is the desired resolution, acquisition frequency and look angle? Some ideas of these are needed to motivate this statement.

We removed this sentence at this point and focus on the discussion on opportunities and limits.

L60: I think it's more "approaches of acquiring a thermal image" rather than "applications"

Changed accordingly

L61-62: the physics and methodologies seem the same to me whether the thermal imagery is acquired via oblique imagery or an airborne sensor. I would say what remains elusive is mapping high-resolution thermal for entire glaciers or several glaciers. A UAV is a good candidate for achieving this, although still a step beyond the scope of this paper.

We agree and change the sentence to: "Debris thickness was recently mapped using oblique LST (Herreid, 2021), but the quantification of debris thickness from UAV thermal imagery has remained elusive."

L64 and throughout LST not "LST's"

Changed accordingly

STUDY AREA

L70: coordinate units

Changed accordingly

MATERIALS AND METHODS

L102: I understand this, but maybe say point measurements are ambiguous or difficult to make in an unbiased/representative way.

> Changed to: "The debris cover close to lateral moraines consists of very large boulders (>0.5 m) that rendered measurements impractical and thus introduced a bias on the point measurements."

L103: how does this boundary look in your debris thickness estimates? Does this mean there were areas of >30 cm debris, just not sampled?

> Unfortunately, the mentioned large boulders made measurements impossible. Based on Figure 10 f, we observed the thickest DC along the north-western margin of the ice. However, as written, it is not entirely clear where the ice stops and the moraine begins. In terms of sampling: yes, especially where large blocks and boulders impeded digging into the debris, the debris thickness is likely larger than 30 cm and we did not sample this.

Section 3.2: I think the authors did a really nice job here illustrating a problem that future glaciologists might encounter and giving a solution to it. Is the spline, rather than a median value shift, capturing spatial variability in drift?

> Yes, the measurement bias at the beginning of a flight can be different that at the end of a flight. If there are well-distributed reference temperatures (in our study ice cliffs), then the spline seems to do a good job. However, at the beginning and the end of the flight, where no ice cliffs were in the field of view, we added a constant value like the median value approach.

I don't think you said yet in the paper how long a flight path took to fly, do we assume Fig. 2 is roughly one snapshot in time? Was this approach applied to each of the flights?

> We added a sentence in section 3.1: "Each flight took between 12 and 15 minutes and captured around 600 thermal images."

Was drift and offset roughly similar for each flight?

> Yes, therefore we decided to show only one example. The evening flights showed less variations of ice surface temperatures. We also added this information to the text of section 3.2: "The drift and offset were similar for most flights, but the evening flights showed less variation in the ice temperatures."

It seems like the bare ice temperature in (c) is still not 0C? . I trust that the air temperature during the survey was well above freezing and indicative of bare ice being at the pressure melting point, but it might be worth noting that there's nothing initially alarming about ice being less than 0C. Consider changing the section title to something like thermal image drift and offset correction.

> Yes, the variation of ice surface temperatures within one frame is quite large. This in-frame variation is not captured in the calibration procedure. The title changed to: "Thermal drift and offset correction"

Section 3.3: The thermal images were stitched and then orthorectified or orthorectified and then stitched? From Fig. 1 I'm guessing the number of frames is around 600 for each flight, can you give the number here. You say that you identified the GCPs in the thermal images but there are only 6 GCPs so most thermal images have no reference, this makes me guess that you stitched the thermal images first. How did you manage overlap? Did you take a mean or trim one image? It's clear from the section above you applied some corrections for drift and offset, but still I'm guessing there are residuals and the temperature may be slightly different for some of the overlap just due to temperature change during the gap time between image acquisition paths.

> We understand that more information on the image acquisition and mosaic generation is welcome and changed the text to:
>
> "Each flight yielded around 600 thermal infrared frames (Fig. 2), of which around 400 have been used to generate orthomosaic maps and 200 were omitted as they recorded the take-off and landing of the UAV. The diurnal variation of the surface temperature and relatively low contrast of thermal images led to spatiotemporal variations in the reconstruction of the 3D point clouds. Instead of additional point cloud alignment (Rusinkiewicz and Levoy, 2001), we orthorectified the thermal images using the same digital surface model (DSM) obtained from simultaneously recorded optical images. Therefore, we identified and marked all GCPs in both the optical and thermal images prior to the photogrammetric processing to improve the image alignment and improving the calculation of the camera calibration parameters (Cook, 2017). As the footprint of the images is relatively large with respect to our area of interest, the 6 GCPs were visible in almost all thermal images. The generated DSM from the optical images was then used as the basis for the thermal image orthorectification. The overlapping parts were reduced by a weighted average during the orthomosaic generation. Agisoft Metashape software offers several options on how to handle overlap areas and we found the default setting to produce the most reasonable results. "

L150: watch section numbering

> Changed accordingly

L166: you are making the assumption that atmospheric attenuation is linear, but still I think I agree that this is a negligible term in this study.

> Yes, we agree.

L171: Convective heat transfer was more 'deemed not present' than included

> Good point! We changed the start of the sentence to: "Thermal energy fluxes…"

L173, Eq. 2: All fluxes have units of Wm^-2 so delta-S must as well, how is this a rate of change of heat, where is the time derivative coming from?

It's the rate of change in heat storage within the debris (e.g., Brock et al, 2010). To balance surface energy fluxes at sub-daily time intervals, the debris layer must be treated as a volume with variable heat storage. It requires knowledge of the average rate of mean debris temperature change, which we obtain by fitting a sine function to the diurnal LST variation in each pixel of the map and forming the first derivative for each time of flight (described in L220-L225 and examples shown in Fig. 5). We explain the delta-S in more detail later in this section with Eq. 6 and think this is sufficient. Please, let us know if there is still something unclear.

L175: What exactly is the ground in this context?

The base of the debris layer. We expanded the sentence by: "… from the debris into the underlying ice." The expression ground heat flux is often used for the conductive heat flux through the layer of debris. But to be more precise we changed ground heat flux to conductive heat flux throughout the manuscript.

L185: with a larger

Done.

L219: stored heat flux

Done.

L221: where did d in the second term come from?

Following Brock et al. (2010) we describe the rate of change of heat storage in Equation 6, where the debris thickness is included. Consequently, the SEBM now contains G(d) and ΔS(d) which then results in a quadratic equation when solving for debris thickness. Previous energy balance approaches to solve for debris thickness did not face this issue as ΔS was either not considered or roughly approximated.

L229: Fig. 7 suggests this null area of unphysical SEBM results is nontrivial, this seems like a very notable detraction, a lot of future analyses that will use debris thickness maps will need to be continuous. The lack of a meaningful solution in areas with debris also raises concern about the area where there is a solution.

Yes, we agree with the reviewer that these unphysical results are relevant, especially in the morning flights. We also agree that these uncertainties may raise concern for areas with existing solutions. We devoted much of chapter 5.1 to discussing and explaining the cause of this issue. In short, such issues arise specifically when the debris is heating up during the morning to midday when uncertainties in LST or reanalysis-derived quantities can cause no solution to the equation. In addition, we address the impact of several parameters on the model and the unphysical results (sensitivity test Figure 12). As the second reviewer had similar concerns, we expanded the discussion section in 5.1 on these issues in the revised version. At this point in the text, we refer the reader to the discussion section but refrain from

getting into too much detail before showing results. Added sentence: "We discuss the causes for these unphysical solutions in detail in section 5.1."

L230: I think you mean training

Changed accordingly

RESULTS

L241: How many scenes make up the corner of (b) that is incorrect? The geometry of the erroneously cold area seems odd. Do you think the >0.5C data is reliable nearby? I'm curious why the bare ice / ice cliff geometry isn't consistent between images? That would be a data quality red flag if stable thermal features aren't captured in a repeatable fashion.

Around 20 scenes. The issue likely lies in the calibration being less reliable towards the beginning and end of the flights when no ice cliffs are available for calibration.

Thanks for pointing out the inconsistency in ice cliffs – that's an important point! The inconsistency in the bare ice geometry is due to the residual uncertainties of the ice surface temperature. Figure 2 shows that the ice surfaces can vary by several degrees (~6 K). This means that our threshold of 0.5 K in figure 3 does not capture all ice surfaces. We added to the caption of figure 3: "Due to the residual uncertainties of the LST, ice surface geometries appear inconsistent in time."

L247: Is 0/360 degrees north?

Yes. Changed figure caption to: "… colorized for terrain aspect with 0/360 ° facing north."

L280: Fig. 5. The plot of time of LST_max is really clever! But how is it a function of terrain aspect? Wouldn't it be a combination of aspect and the thermal inertia of the debris? Isn't bar{T_d} just LST/2 per definition in this paper (LST + 0)/2)(L211)? Does this quantity really have more meaning than LST alone?

Thanks for the credits! We agree that the LST is a function of both aspect and thermal inertia. The influence of aspect is nicely shown in the figure and examples. The influence of thermal inertia is more difficult. As thermal inertia is a material property, it may change across the scene. However, our impression during fieldwork was that the rock type making up the debris layer does not vary across the surface. In any case, what we show in the figure is only the effect of aspect, which does not mean that thermal inertia plays no role. To make it clear that we only refer to the influence shown in the figure, we changed the sentence in the caption to: "The panel shows the time at which $\bar{T}_d$ reaches its maximum diurnal temperature and thereby emphasizes the effect of terrain aspect."

It is true that $\bar{T}_d$ is directly connected to the LST. As we don't know the mean debris temperature, we approximate $\bar{T}_d$ using the LST. We added an explanation to the figure caption, but we prefer to keep the figure as it is, because it connects

immediately with Eq. 6, which describes the computation of delta-S. Caption text changed to: "Sinusoidal regression of mean debris temperature ($\overline{T}_d$), estimated from LST (**Fehler! Verweisquelle konnte nicht gefunden werden.**) for each pixel..."

L289: Somewhere can you add a histogram of debris thickness please? Fig. 6(e,f) suggest many of the measurements were 5cm or less. L101 said the mean was 9cm but maybe also state the median? Or perhaps a different drawing order in the fig or different data visualization if there are too many lines to show the whole dataset. I'm just concerned that the methods used here are better suited (meaning more stable for method evaluation) to slightly thicker debris.

That's a great suggestion! We added a histogram of the field measurements to figure 1. We agree the median information would be helpful too. The sentence changed to: "The debris cover on the TNG is generally thin: measured thicknesses are below 30 cm, with a mean of 9 cm, a median of 5 cm and a standard deviation of 10 cm."

L290: methods not results

True, we moved the sentence to section 3.6

L305: I don't get this, are you saying G is actually large for thin debris or just a very subtle max over a time series of low values? G might be sensitive to inaccurate LST over thin debris cover because the ice should keep the numerator of G low or near zero, but a thin debris thickness denominator might make G sensitive to LST error. Debris thickness measurements can also be quite sensitive over thin debris covers, where getting the local mean thickness wrong by 1 cm could be a high percentage error, while a 3 cm error over a 50 cm debris cover will propagate less error though these sorts of calculations.

These are good points, thanks! In fact, the absolute values of G can be large. The values on the y-axis are scaled by 1e3, which was unfortunately hidden above the figure. In the revised version, we made sure that the values are shown on the axis. Yes, we absolutely agree with the reviewer. G is sensitive to inaccurate LST, which is also one of the reasons why we obtain unphysical solutions. We expanded our discussion in section 5.1 by the following:

"The most likely reason for no physical solution to Eq. 7 are inaccurate values of LST and reanalysis-derived variables. Mostly during the morning, even small deviations from true values are sufficient to find no physically meaningful debris thickness solution. For thin debris (< 2cm), G is very sensitive to uncertainties in LST and leads to large negative numbers. This is a major drawback of the SEB approach, and it highlights the sensitivity of the approach to uncertainties in the input data. We note that these uncertainties prevail, even if a solution is found. However, compared to the ground observations of debris thickness the model predictions show a positive correlation."
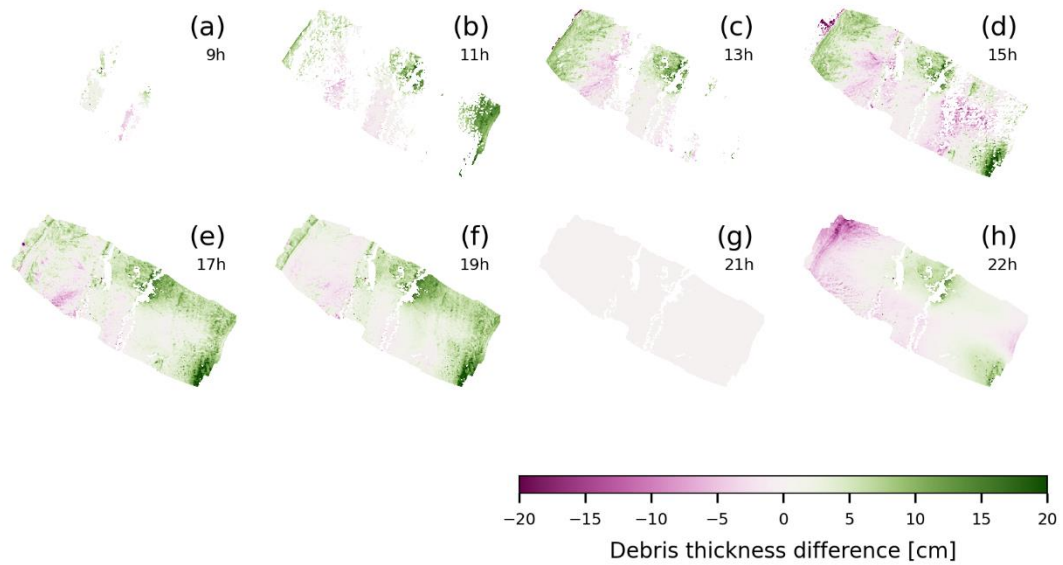
L316: I don't find this sufficient. Reality doesn't return an undefined solution, I think if the model fails and you understand why, you should write a piecewise function that either returns continuous results or at least sets constraints on where the method is applicable.

> We understand that discontinuous maps are not satisfying. However, we think it is an interesting behavior of the model that highlights some difficulties that might be hidden otherwise. By extending the common SEBM approaches by the component of ΔS, we introduce more complexity as now both G and ΔS are a function of debris thickness. SEB approaches as applied by e.g. Rounce and McKinney, 2014 or Foster et al., 2012 would result in continuous maps, which does not necessarily mean that these results are closer to reality. We are aware of the high uncertainties of our LST measurements and the reanalysis data, however, the shape of the diurnal temperature cycle (Fig 5 c, d, e, f) still shows the expected diurnal variation which is encouraging. We decided to take advantage of the sub-daily time intervals and show the discrepancies in the model results, which we also discuss in detail (section 5.1).
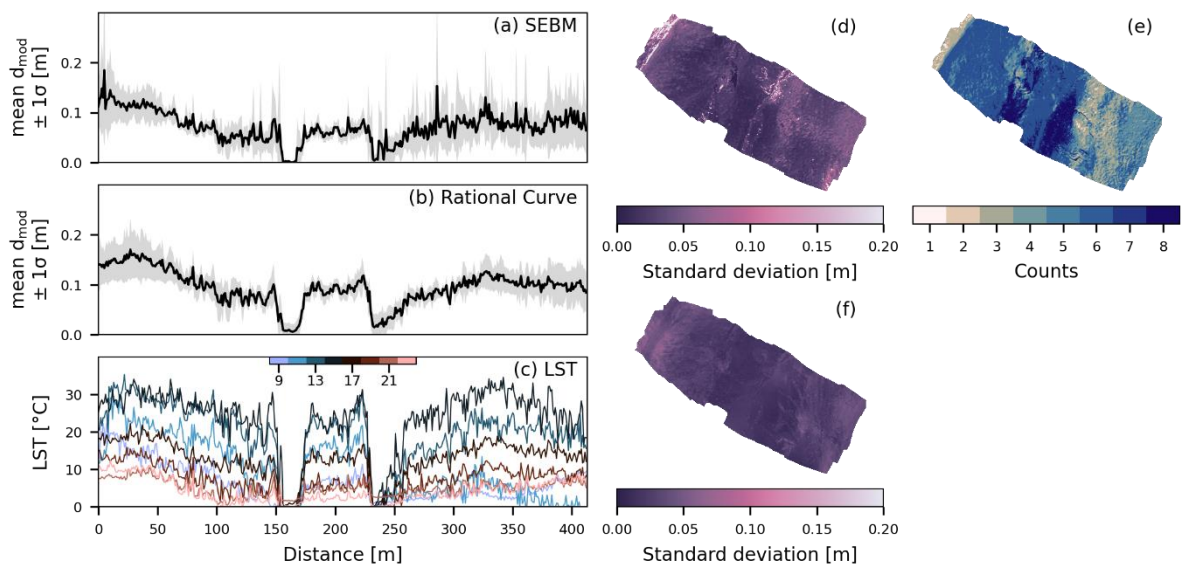
L313: Fig, 7: If the debris thickness estimates were continuous over the same domain the histograms would be informative, but with data voids they aren't very comparative. I would like to see difference maps, with respect to perhaps one reference time that agreed particularly well with measurements. The main challenge in this problem is taking variable input datasets and returning a constant value. Based on the images alone in Fig. 7 it looks like a general trend is preserved but the pixel-to-pixel debris thickness will be shown to vary quite a bit unrealistically over hours and I also see what look like artifacts in the data. Likely the same artifacts that are discussed in the text but I would be hesitant to keep them for use in the later analysis of this paper. It's possible the quality of thermal sensor that can fit in a DJI UAV isn't quite high enough to return clean data.

> We agree on the relevance to evaluate the debris thickness maps with respect to time. We already addressed this topic in Fig 9, by showing the mean +- 1σ of a profile line. We tested the reviewer's suggestion and created difference maps, with respect to the 21h flight time. However, we have difficulties picking a reference time for the difference maps and also don't want to overload the manuscript with two additional large figures (SEBM and rational curve). Therefore, we suggest expanding Figure 9 with a map of the standard deviation of the predicted debris thickness values and a map that shows the number of valid solutions of the SEBM to evaluate the consistency in time.
>
> Option 1:

(a) 9h (b) 11h (c) 13h (d) 15h
(e) 17h (f) 19h (g) 21h (h) 22h

Debris thickness difference [cm]

Option 2:

(a) SEBM
(b) Rational Curve
(c) LST
(d)
(e)
(f)

mean d_{mod} ± 1σ [m]

LST [°C]

Distance [m]

Standard deviation [m]

Counts

Standard deviation [m]

L317-318: I would like to read more quantitative results than "relatively consistent in time"

Changed to: "Predictions of thicker (>10 cm) debris are primarily found in the afternoon and evening hours (17 h to 22 h) and the pattern of thin debris (<10 cm) predictions, primarily in the central part of the glacier, is relatively consistent in time."

Line 320: Line 101 states the mean debris thickness is 9 cm so a RMSE of 6 to 8 cm means nearly 100% error for the average case.

> Yes, we agree, and starting this research we wished for a better match between observed and modelled debris thicknesses.

L320-325: Fig, 8. I think this is just about how well debris thickness can be predicted from mostly or exclusively thermal data, and I think one should be careful from trying to see what we all wish it would show. "Correlate well with observations even if they do not follow the 1:1 line" sounds like seeing what you want to see, a successful predictive model really should be clustered around the 1:1 line.

> These are valuable points! We agree with the reviewer and rephrased the sentence to: "For most of the flights we find a positive correlation between the predictions and observations, even if they do not follow the 1:1 line."

L328: I'm not sure it's so clear about the thin debris, if you change the plot limits to 5 cm I think it might look like random scatter. The distance away from 1:1 might be less, but with respect to the magnitude of the debris thickness there may not be clear evidence of predictive capability. This tight cluster of thin debris cover near the origin likely also weights the RMSE favorably. If you consider only debris greater than 5 cm thick, for example, the RMSE will likely increase.

> These are valid points. In the sentence this comment referred to, however, we address how sensitive predictions of thin debris are to the time of day: "Predictions of thin debris cover are less sensitive to the time of the day, compared to thick debris." Although the percentage change may indeed be still large, the absolute change in predicted thin debris cover is smaller than that for thick debris. To make this clearer, we changed the sentence to: "Absolute values of predicted thin debris cover are less sensitive to the time of the day, compared to thick debris."

L335: Are you computing the standard deviation from 7 or 8 measurements? Is that enough?

> Well, we see the point of the reviewer. However, we show the standard deviation simply to indicate the variability, and for this purpose, we think it serves the case.

L347: How did you divide the dataset?

> We used the python scikit learn function: sklearn.model_selection.train_test_split()
> https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

> Citation added to the text.

DISCUSSION

L370: comma after pattern

> Done.

L371: I think you need to show how well the model results are consistent in time as well as approach the testing dataset to make a statement about method viability.

> Good point. We toned down the sentence to: "The fact that the modelled debris thickness does not vary in unreasonable ways across the glacier surface, but in a systematic pattern shows that mapping high-resolution debris thickness with UAVs has some potential."

L405: Fig. 12: x-axis units

> Done.

L445: this could be the piecewise condition mentioned above

> See answer to comment L316

L450-452: I think this might be true with a more advanced methodology, but the results from this study show, what look like, large changes in debris thickness over a day when the quantity should be found as static. Mapping high-contrast features like ice cliffs from high resolution thermal data has already been shown to be useful, but studies relating changes in debris thickness to erosion rates or surface properties will need quite high confidence debris maps, since the changes are likely to be on the order of cm per decade.

> Yes, we agree with the reviewer that the results from this study do not allow us to study changes in debris thickness as the uncertainties in LST are too high to resolve small changes. However, we think that this study gives an orientation where the difficulties arise and discuss in detail the limits and opportunities of UAV thermal image derived debris thickness estimations.

L460: the high emissivity of rocks limit reflected radiation making the viewing angle less of a factor, but still can be accounted for. The angularity of rock clasts make even a nadir look-angle non-normal to most surfaces.

> We agree.

L462: I would add battery limitations on range as well.

> Done.

L475: What about a more expensive camera as well? I assume higher end FLIR uncooled microbolometers have a higher accuracy even in less ideal conditions.

> Probably yes, but we think the more relevant point would be a better in-flight calibration (e.g. with a mobile blackbody) that would lead to higher accuracies.