Authors' response to the reviews of:
"DCG-MIP: The Debris-Covered Glacier melt Model Intercomparison exPeriment"

RESPONSE to EDITOR and GENERAL REVISIONS

We would like to thank the two referees very much for their constructive reviews of our manuscript. We found their comments very useful and we have addressed them all. As a result, the revised manuscript has improved in clarity and impact. We provide here a short summary of the main points raised by the reviewers (all minor) and of our corresponding revisions.

The issues in the two reviews were:

- To discuss the limitations associated with using only one season for model evaluation (Rev 1)
- To slightly adjust the terminology and classification of model KO2 and of model Hyperfit to better reflect their characteristics (Rev 1)
- To use for the evaluation of model performance a different metric than the median (Rev 2)

We have carefully addressed each of them, and as a result we have:

- Included a paragraph at the end of Section 2, and a paragraph in Section 4.3 to explain the limitation of using only one season for model evaluation.
- Classified KO2 as an enhanced temperature index model (described in section 4), and indicated what was before designated as the uncalibrated version of Hyper-fit as a version of Hyper-fit with estimated parameters from the literature. We agree entirely with the reviewer's point here.
- Conducted a careful analysis to include a metric of absolute accuracy in our ranking. We have thus included the Mean Absolute Error in our model evaluations and ranking (Tables 3 and 4) and we show that our results do not change if we use the metrics suggested.

As part of the discussion in point 2 above about the best designation/definition of some of the empirical models, we decided to indicate models KM1 and KP1 as two separate models, to accommodate the KO2, KM1 and KP1 modeller wishes. We have thus changed Figure 3 and Table 2 to accommodate 15 instead of 14 models. This is only a matter or reclassifying a model (from a model in its calibrated and uncalibrated form to two models, one calibrated and the other one uncalibrated) and does not affect the way we consider that model (now two) in the paper, their ranking or our conclusions. A description of this change is attached at the end of this response.

As a result of the modifications summarised above, none of the main results, discussion points and conclusions of the paper changed.

We hope the editor and referees will be satisfied with the revision of our manuscript. We are very happy with its revised form, and look forward to the editor's decision.

Thank you very much.

Adrià Fontrodona Bach and Francesca Pellicciotti on behalf of the authors of the DCG-MIP.

We indicate in Black the reviewer's comments, and in Blue the authors' response

In the case of revised text shown in this response: "Normal blue for original text, ~~red strikethrough for removed text~~ **and bold purple for revised text"**

Line numbers of our responses refer to the **revised track changes manuscript**.

RESPONSES TO REVIEWER 1

RC1: 'Comment on egusphere-2025-3837', Anonymous Referee #1, 18 Sep 2025

RC1.1 This is an eagerly awaited report of results from the debris-covered glacier intercomparison project. It has a good geographical range and number of participating models, but is limited to only one melt season per site. This does not allow spin up of debris and ice temperatures or division into calibration and validation periods. Repeating the experiments with more years of data would be beyond feasible modifications for this paper, but reasons for and implications of this restriction should be discussed in more detail.

We thank the referee very much for their review of our manuscript and thoughtful comments, which have substantially improved the paper. We provide below a detailed response describing the revisions of our manuscript.

We agree with the reviewer that a limitation of our study is that the model comparison was carried out over only one melt season at each site. This was, as the reviewer notes, dictated by data availability, as very few sites had measurements available for more than one season. We thus appreciate that the reviewer finds it unfeasible to repeat our experiments for more years at this stage, which would indeed be impossible. Instead, we have, as suggested, expanded the discussion of this limitation in the revised manuscript. In the original submitted paper we had devoted a sub-section of the discussion (Sub-section 6.3, *Limitations and recommendations for Phase II*) to the experiment's limitations, where we included a paragraph on this and a detailed recommendation for a follow-up experiment addressing this issue (Lines 949-957). In addition to that section, we have included a description of this issue in the model evaluation section (Lines 425-429):

"**Model performance was only evaluated over one ablation season because data were not available for additional seasons at most sites. This limited our capacity to assess model robustness over multiple seasons. For the models requiring calibration, i.e. temperature-index, enhanced temperature-index and simplified energy-balance models (MCC19, DETI, Hyper-fit, KM1, DDF_{debris}), in particular, this does not allow a separate validation period, as all models requiring calibration were optimised for the entire period of data available.**".

We have also indicated in Section *2. Experimental setup* that energy balance models were not spun up prior to the simulations for the experiment, given the lack of data (Lines 175-178):

"Given that data were available for one melt season only, no spin-up was possible for energy balance models. Given that no snow was present at the beginning of our experiment, that the time required to spin up within-debris temperature should be in

the order of hours or a few days, and that only one model allows ice-temperatures to go below zero, we expect the lack of a spin up period to have a minimal effect on the study."

RC1.2 For annual mass balance simulations, temperature-index models also require precipitation as input.

It is true that temperature-index models also require precipitation as input when used to simulate glacier mass balance. However, in the paper we focus only on melt processes and the ability of models to simulate melt under debris during the melt season only. For the simulations carried out in the paper, precipitation is therefore not required.

We have specified that this sentence only regards melt modelling as follows (line 135):

"Despite this, temperature-index models have seen successful applications at the glacier and regional scale because they are simple, computationally efficient and require only air temperature (occasionally incoming shortwave radiation) as input **to model melt** and a low number of parameters (e.g., Kraaijenbrink et al., 2017)."

RC1.3 Table 1: Position to four significant figures locates the glaciers to within about 10 m.

The coordinates refer to the automatic weather station, and not a general glacier location. We have revised the caption to make this clearer:

"Table 1. Overview of study sites. Validation data indicates what kind of melt observations are used at each site: ultrasonic depth gauge (UDG), ablation stakes, a draw-wire, and debris surface temperature (Ts). hd = debris thickness. **The latitude and longitude coordinates refer to the locations of the automatic weather stations.**"

RC1.4 Table 2: With net solar radiation used as an input, why is KO2 not classified as an enhanced temperature index model?

This is a very valid point and the reviewer is right that KO2 should be considered as an enhanced temperature index model, despite not resolving the daily cycle of shortwave radiation because of its daily time scale. We have revised the manuscript as suggested, applying the appropriate changes to text, figures and tables, to reflect the change from TI to ETI for model KO2 (e.g. abstract, "enhanced temperature index models" subsection, Table 2, Figure 3, Figure 5, Table 3, Figure 7, Table 4, Figure 9 and Figure 10). Our rank in terms of model complexity remains unchanged. According to our definition of complexity (Figure 3), $DETI_m$ and KO2 are both enhanced temperature models, both using solar radiation as input, but as $DETI_m$ is run at hourly resolution and we prioritised this in terms of preferred model complexity, we regarded $DETI_m$ as more complex than KO2.

In Section 4, modified "Enhanced temperature index models" subsection:

"Enhanced temperature index models

The debris enhanced temperature index (DETI; Carenzo et al., 2016) model was developed as a model of intermediate complexity between a temperature index model and an energy balance model, building on similar developments for clean ice (the ETI model, Pellicciotti et

al., 2005). It includes the shortwave radiation balance, and a term dependent on air temperature that represents empirically all other fluxes in the energy balance equations. The model's empirical parameters are a function of debris thickness, to account for the time needed to transfer energy from the surface to the ice, and were derived through functional relationships between the shortwave radiation flux and temperature with sub-debris melt simulated by an energy balance model at different thicknesses. It is designed to run at hourly resolution.

Winter-Billington et al. (2020) introduced an enhanced temperature-index model based on a mixed-effects approach. Fitted using data from 27 glaciers, the model predicts degree-day factors as an exponential function of debris thickness and, combined with air temperature data and net shortwave radiation as a second fixed effect, estimates melt at a daily time step. The mixed-effects framework allows these models to be applied to new sites without recalibration, while providing prediction uncertainty based on the original training data. With two fixed-effect predictors, the model KO2 is considered an enhanced temperature-index model under our classification scheme."

RC1.5 Figure 4: Notation for radiation fluxes differs from Equation 1.

Thank you for spotting it. We have standardised these notations.

RC1.6 272: Elsewhere it is stated that net shortwave radiation is given, not calculated.

It is indeed given, not calculated. We have revised the text as follows (line 277):

"All energy balance models **directly use the provided observed** ~~calculate the~~ net shortwave **radiation flux**, and **calculate the** longwave radiative fluxes and the turbulent sensible heat flux at the surface (Fig. 3).

RC1.7 282: Relative humidity is a property of air. Are the assumptions rather on the wetness of the debris surface?

Yes, the reviewer is right: the assumption is on the wetness of the debris surface, based on the relative humidity of the air. We have revised the text to clarify this (lines 285-289):

"Since no data on the water content within the debris were available at any of the sites, modellers either neglected this flux or made assumptions on the actual relative humidity of the **air at the** debris surface (RHs) **based on the relative humidity of the air or precipitation occurrence** (Table S16). These vary from assuming that the **air at the** surface is saturated when it rains (DEB_{CF}, ROU15) to assuming that RHs=100% when the air relative humidity **at the measurement height** (RHa) is 100% (DEB_{PG})."

RC1.8 Figure 5: (a) and (b) labels are missing from the figure.

Thank you, we have added these.

RC1.9 458-474: With this much discussion of Figure S3, it would be better to include it in the paper.

Thank you for the suggestion. We agree that Figure S3 is extensively discussed and could be included in the paper. We have included it together with Figure 8 arranged as two panels, as shown below:

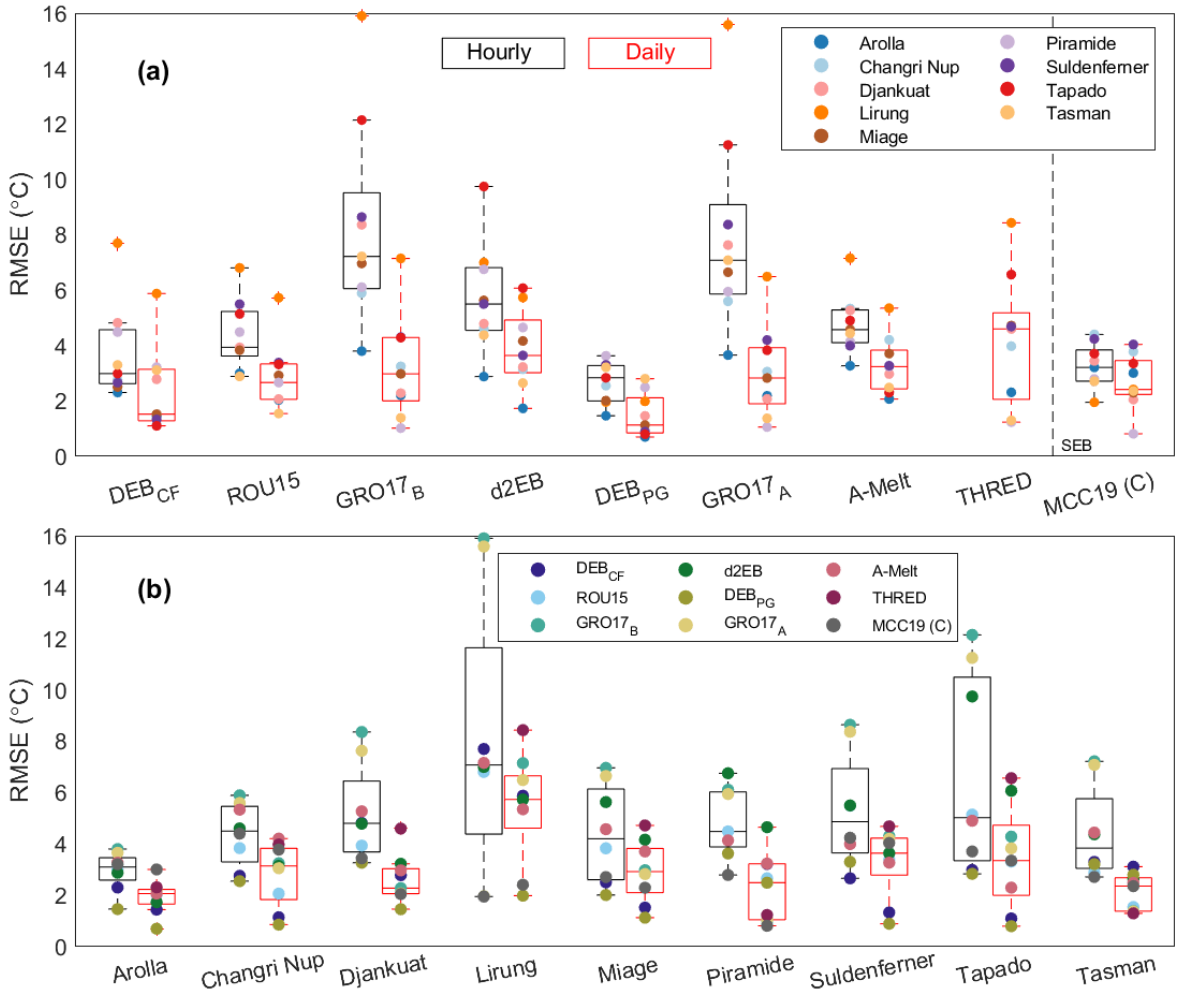

Figure R1.1. New revised Figure 8 including Figure S3 as bottom panel.

RC1.10 577: Djankuat is just Fig. 10c.

We have corrected this.

RC1.11 Figure 11: A reminder that fluxes are negative when away from the surface would be useful in the figure caption.

We have added this in the caption: "**Note that fluxes are negative when going away from the surface.**"

RC1.12 795: The “uncalibrated” version of Hyper-fit with parameters for the same glacier but a different time period would be regarded as a calibrated model in any other study.

We agree that the term “uncalibrated” is ambiguous here and might be misleading. Since we left the definition of uncalibrated up to the modellers, the Hyper-fit uncalibrated parameters were estimated by the modeller using Ostrem curves derived from debris thickness and melt data from the same glacier (except for two glaciers, Suldenferner and Tasman for which global data were used), but without using the specific dataset from this experiment. To avoid confusion, we have replaced the “Hyper-fit uncalibrated” version with “Hyper-fit estimated” version. We have replaced the (U) from uncalibrated for Hyper-fit with (E) from estimated in all figures in the manuscript. We have also corrected this in the description of this model in the supplement, and provide the definition of estimated in the main text (lines 398-402), as follows:

“The uncalibrated version of Hyper-fit used previously-published~~, independent parameters~~ **melt and debris thickness** data from six of the nine sites**, outside the study period of this experiment, to estimate values for h*** (see Table S24). For the other three sites, the global mean h* value (Anderson and Anderson, 2016) was used to represent h*. **Therefore, these Hyper-fit model runs were not defined as uncalibrated but as estimated (E).”**

RC1.13 I have not checked the reference list in detail, but the authors should. Kuzmin (1961) at least is missing.

We have checked and revised the reference list and we have included Kuzmin (1961).

RC1.14 Table S1: Although the models are not required to calculate albedo, it would be interesting to know the measured debris albedo for each site.

As albedo is a property derived from meteorological observations we have added the mean measured outgoing shortwave radiation and mean measured albedo at each site in Table S3, together with the other mean measured meteorological forcing variables.

RC1.15 The paper is well written, with few errors that I noticed: 73: “which has has”; 821: “This suggests suggests”

Thank you for spotting these, we have corrected them.

RESPONSES TO REVIEWER 2

RC2: 'Comment on egusphere-2025-3837', Duncan J. Quincey, 30 Sep 2025

RC2.1 This is a comprehensive analysis of the available debris-covered glacier melt models and their performance across a range of sites. It represents a considerable effort on behalf of the DCG community and despite the complexity of the experimental approach, the paper is relatively easy to follow and the key points are simple to digest. The summary figures are well-constructed and convey a clear message, and the transparency in describing the limitations of the approach is particularly appreciated. I have few substantial comments overall, and I am in support of the paper being published, but I would like to raise one issue (see comments under 'Table 3' below) that may require some consideration and might change some of the interpretation of which models are the 'best' performers, if it is addressed.

We thank Dr. Duncan Quincey for the positive and constructive review of our manuscript, and for the appreciation of the effort involved in this experiment. We are happy that the manuscript is easy to follow and conveys clear messages, while we have aimed to improve it further for publication based on the reviewer's very useful comments. We respond to these comments below, and provide detailed revisions for each of the points raised by the reviewer, including the point on Table 3 on model ranking.

General comments

RC2.2 Throughout, I find the bold lettering in the main text to be distracting. In some places it doesn't highlight anything that seems to be particularly important – maybe just let the reader make their own judgement, as in most other papers?

The reviewer is very right about this. We have removed the bold lettering throughout the manuscript.

RC2.3 L115: it is unclear here whether the phrase 'where most previous research has been carried out' refers to the European Alps, or to the remote sites outside the European Alps – it can be interpreted both ways.

It refers to the European Alps. We have rephrased it to clarify it (line 116):

"... are difficult to constrain spatially and temporally even for individual glaciers and short (sub-annual) periods, especially at remote sites outside of the European Alps, **as** ~~where~~ most previous research has been carried out **in the European Alps** (e.g. Nicholson et al., 2006; Brock et al., 2010)."

RC2.4 Figure 1: At first look the colour bar seems to relate to the shading on each of the inset glaciers, rather than to the data in the map. I think it's because the inset glaciers and the colour bar are the most distinctive elements (with the map data being more subtle). Maybe shrinking the colour bar, or having a more complete legend that includes the two glacier shades as additional items, rather than describing them in the caption, would make it clearer? The darkest colours in the 1 x 1 degree cells (i.e. those approaching black) also seem to be beyond the minimum value in the colour bar?

We appreciate the reviewer's suggestion to improve the readability of Figure 1. We have revised it according to the suggestion, as shown below. The colour bar was shrunk and clean ice, debris cover and AWS have been added as additional legend items. Furthermore, we have increased the font size for better visualization.

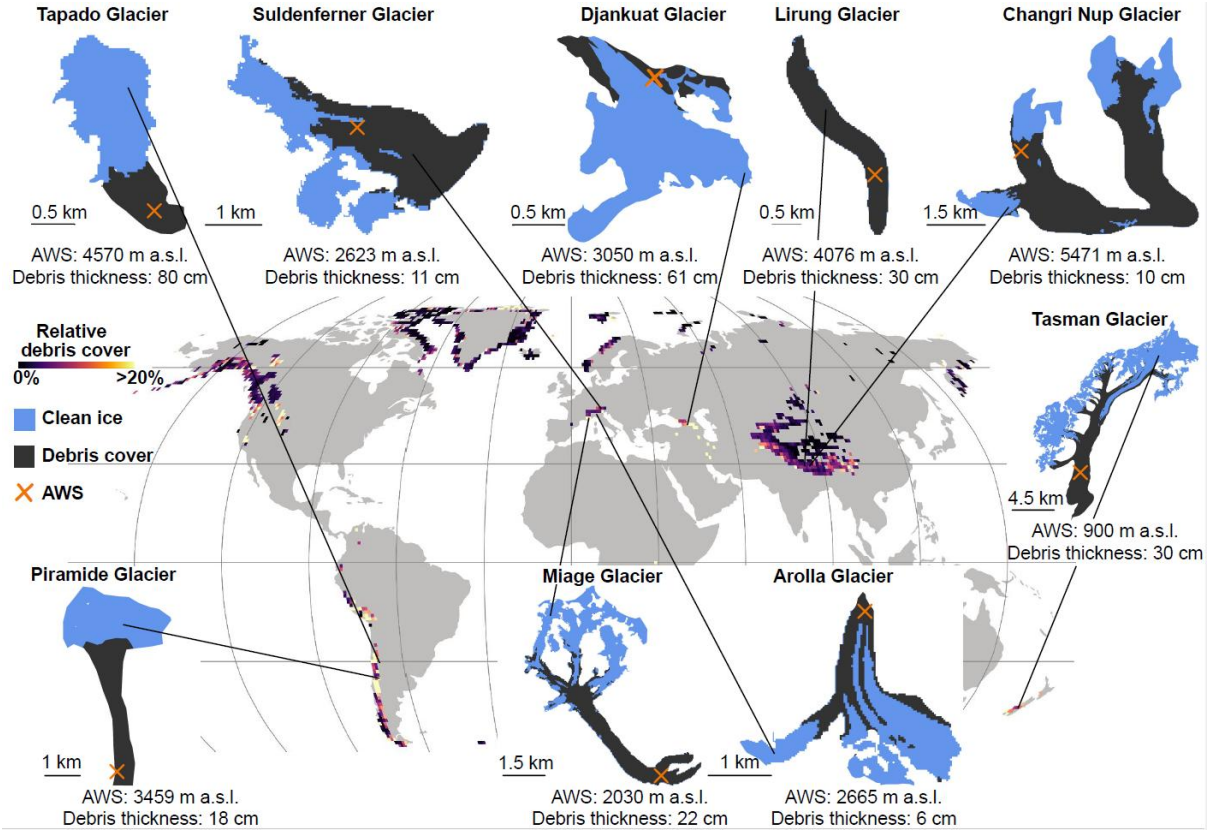

Figure R2.1. Revised Figure 1 including the reviewer's suggestions.

Table 2: Kayasha = Kayastha?

We have corrected this.

RC2.5 Table 3: I have a doubt about using the median of a range of signed values to evaluate how well a model is performing, since the positive and negative values effectively cancel each other out such that the median tends to zero even where a model is performing poorly. The median therefore tells us only of the model is biased towards under- or over-prediction, rather than anything about the accuracy. It seems to make more sense to treat the values as absolute, and then take the median of those. Or better still, use a magnitude-based metric like MAE or RMSE – which would also give consistency with the subsequent analysis of surface temperatures that follow. If you do this re-calculation, the ranking changes, and this will feed through into some of the interpretations and might impact the discussion as well.

This is a very useful comment, which we have extensively considered. The reviewer is right that the median of signed values can be close to zero when values cancel each other. This is why our Table 3 also shows the interquartile range of model performance across sites. Our ranking in Table 4 uses the median and also the IQR, to penalise those models with a large spread in model performance.

Nevertheless, we agree that including a metric that better indicates absolute model accuracy is useful, and we have therefore included the Mean Absolute Error (MAE) of model performance across sites as an additional metric to Table 3 as suggested by the reviewer. We copy below the revised Table 3, and note the bolding of absolute values < 20% in response to the reviewer's comment RC2.6.

Table 3. Modelled melt error across models and across sites in percentage. Models are ordered by complexity as defined in Figure 3. The last three columns correspond to the median (Med.), interquartile range (IQR) and mean absolute error (MAE) across sites per model, and the last three rows correspond to the median, IQR and MAE across models per site. We provide the IQR as a measure of the spread in model performance and the MAE as a measure of absolute error. The values of median, IQR and MAE in the bottom right of the table indicate the overall respective value across sites and models altogether. "C" indicates calibrated, and "U" indicates uncalibrated. Values between -20% and +20% are shown in **bold**.

Cat.	Model	ARO	CNU	DJA	LIR	MIA	PIR	SDF	TAP	TAS	Med.(%)	IQR (%)	MAE (%)
EB	DEB_{CF}	**-17.7**	-23.3	45.4	103.7	**4.6**	**-7.2**	-25.7	-41.0	46.5	**-7.2**	69.6	35.0
EB	ROU15	**-9.6**	-22.4	87.8	103.1	**18.4**	**-7.5**	**-5.8**	**-7.7**	67.1	**-5.8**	80.4	36.6
EB	$GRO17_B$	**-6.9**	-23.4	120.6	154.8	46.5	29.0	**15.0**	48.5	98.1	46.5	94.2	60.3
EB	d2EB	**-16.0**	**-15.8**	90.8	105.7	24.2	**-11.7**	**-5.3**	-28.3	57.9	**-5.3**	82.0	39.5
EB	DEB_{PG}	**14.9**	**-19.7**	95.2	54.4	23.7	**18.3**	**-9.5**	-29.8	76.3	**18.3**	72.0	38.0
EB	$GRO17_A$	**0.2**	**12.2**	112.4	145.1	44.8	28.5	**13.9**	42.7	97.3	42.7	87.6	55.2
EB	A-Melt	**-9.2**	**-9.2**	104.6	101.8	25.4	**-13.2**	**-7.3**	**-14.8**	92.2	**-7.3**	104.8	42.0
EB	THRED	**-7.1**	**-5.7**	140.8	146.1	34.7	24.8	**4.0**	22.9	82.9	24.8	95.8	52.1
SEB	MCC19-C	**-1.6**	**-7.1**	100.9	55.2	**12.1**	**9.3**	**-1.1**	**-7.4**	52.2	**9.3**	55.9	27.4
ETI	DETIm-C	**-13.0**	-56.2	34.6	74.0	**-0.5**	-20.7	-27.1	-58.0	43.6	**-13.0**	71.2	36.4
ETI	DETIm-U	**-1.0**	-21.5	-46.0	-23.3	**-0.5**	21.4	**6.2**	-56.1	**-18.0**	**-18.0**	30.2	21.6
ETI	KO2-U	-28.3	-81.1	-50.7	-85.5	70.2	46.9	**12.2**	-77.2	157.8	-28.3	130.9	67.8
TI	KM1-C	**-3.7**	**-2.1**	**2.6**	-67.3	-23.5	**7.2**	**-4.0**	**4.9**	**-18.2**	**-3.7**	22.7	**14.8**
TI	KP1-U	-53.6	-94.3	**-11.3**	-34.8	134.9	154.2	-20.5	**-7.5**	542.6	**-11.3**	179.2	117.1
TI	Hyper-fit-C	**6.1**	**3.1**	188.5	**1.2**	**7.3**	**-13.6**	**2.5**	63.9	**2.1**	**3.1**	**19.5**	32.0
TI	Hyper-fit-E	22.6	-45.8	41.3	**4.3**	39.5	**-5.0**	**8.8**	76.2	**-14.7**	**8.8**	47.4	28.7
TI	DDF_{debris}-C	**6.0**	20.8	187.6	**1.7**	**6.9**	**-13.3**	**1.9**	64.1	**2.3**	**6.0**	29.8	33.8
-	Median (%)	**-3.7**	**-15.8**	90.8	55.2	23.7	**7.2**	**-1.1**	**-7.5**	57.9	**4.3**	-	-
-	IQR (%)	**15.4**	24.2	87.8	109.1	34.5	37.8	**14.7**	76.7	91.2	-	59.5	-
-	MAE (%)	**12.8**	27.3	85.9	74.2	30.5	25.4	**10.1**	38.3	86.5	-	-	43.4

We have also incorporated the MAE in the ranking calculation with the median and IQR (new Table 4, Table R2.1 below) and revised the ranking accordingly. In Table R2.1 below, we show the revised Table 4, and compare the original ranking with the new revised ranking including the three metrics (median, IQR and MAE). The changes are minimal compared to the original ranking and strongly support the discussion and conclusions of our study. Supporting point 1 of our conclusions, temperature-index models in their calibrated or estimated parameter version top the ranking even more clearly than before, while uncalibrated models dominate the last positions even more clearly than before too (with the exception of the $DETI_m$). Energy-balance models have lowered positions in the ranking, penalised by the large absolute errors at the three sites with inaccurate debris properties. However, ranking positions among energy balance models have not significantly changed, and the two best models remain DEB_{CF} and ROU15, in the same order, supporting point 2 of our conclusions.

To further support the fact that our choice of ranking method and way in which we visualize model performance does not change the conclusions of our study, we have also added a scatter plot of

the MAE and IQR in the Supplement (see Figure R2.2 below), showing the same clear groups of model performance that Table 4 indicates.

RANK-ING	Model Type	Model	Median (Rank)	IQR (Rank)	MAE (Rank)	ORIGINAL MEAN RANK	REVISED MEAN RANK	RANK CHANGE
1	TI	KM1 (C)	2	2	1	2	1.7	+1
2	TI	Hyper-fit (C)	1	1	5	1	2.3	-1
3	TI	DDF_{debris} (C)	5	3	6	4	4.7	0
4	TI	Hyper-fit (U)	8	5	4	6.5	5.7	+1
5	SEB	MCC19 (C)	9	6	3	7.5	6.0	+3
6	ETI	$DETI_m$ (U)	12	4	2	8	6.0	+3
7	EB	DEB_{CF}	6	7	7	6.5	6.7	-3
8	EB	ROU15	4	10	9	7	7.7	-2
9	EB	d2EB	3	11	11	7	8.3	-2
10	ETI	$DETI_m$ (C)	11	8	8	9.5	9.0	0
11	EB	DEB_{PG}	13	9	10	11	10.7	+1
12	EB	A-Melt	7	15	12	11	11.3	-1
13	EB	THRED	14	14	13	14	13.7	+2
14	EB	$GRO17_A$	16	12	14	14	14.0	0
15	TI	KP1 (U)	10	17	17	13.5	14.7	-2
16	EB	$GRO17_B$	17	13	15	15	15.0	0
17	ETI	KO2 (U)	15	16	16	15.5	15.7	0

Table R2.1. Revisions to Table 4 from the manuscript. “Median”, “IQR” and “Original mean rank” are the same columns as Table 4 from the original submission. Mean Absolute Error (MAE) is now added and the revised mean ranking is calculated based on the mean of the median rank, IQR rank and MAE rank. Rank change indicates the positions a model escalated or de-escalated in the ranking. Note **our actual revised Table 4 does not include the columns “original mean rank” and “rank change”.**

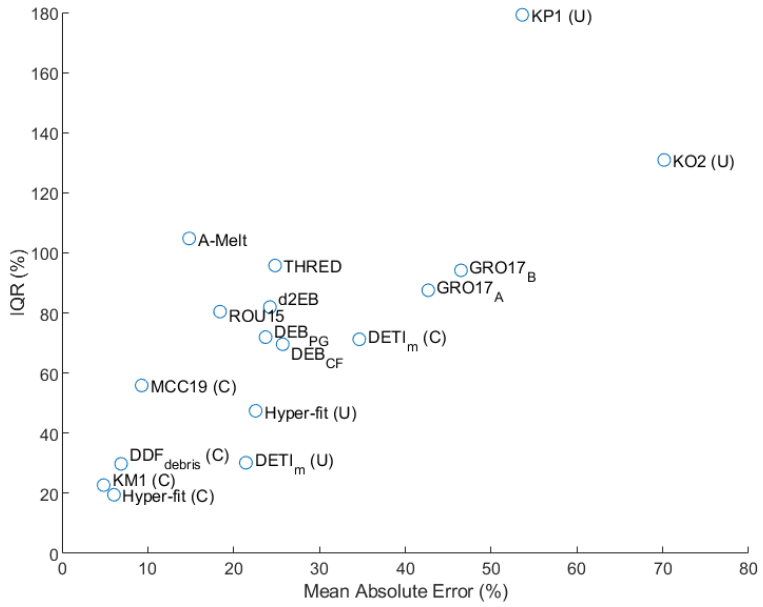

Figure R2.2. Model performance visualisation (MAE vs. IQR). Now included in the Supplement as Figure S6.

Summary of revisions: We have revised Table 3 and the ranking of models (Table 4) as outlined above (in Table R2.1), and have incorporated the scatter plot of MAE and IQR (Figure R2.2) in the Supplement as Figure S6, to show the reader how the use of different metrics and visualizations may slightly change the order of the ranking while not changing the conclusions of our study. We have added a sentence to this respect at the end of the paragraph describing Table 4 (lines 550-552): **"Figure S6 presents an alternative visualization of model rankings, demonstrating that the method of visualization does not alter the overall results or ranking. The figure produces results and interpretations that are consistent with those reported in Table 4."**

The text has been revised to reflect the changes described above (e.g. Section 4.3.1). However, changes to the text are minimal because the ranking results are minimally changed, and the discussion and conclusions remain unchanged.

RC2.6 Consider shading the cells in Table 3 according to how close to zero they are? I think that would help the reader to more quickly identify the patterns across models and across sites.

Although we agree it could help the reader, the journal guidelines explicitly state that "Coloured table cells should be avoided." and we therefore instead bold all values in Table 3 that are smaller or equal than 20% (in absolute value).

RC2.7 Line 504: this value of 4.3% is misleading for the same reason given above. Indeed, the models rarely perform as well as this (Table 3); the uninitiated reader could easily read the paper and think the models overall perform exceptionally well, when the truth is somewhat different. If this is instead calculated using absolute values, the median error is 23.3%, which I would suggest is a much more realistic estimation of the overall model performances.

We agree the median may be misleading without an accompanying metric to provide more context, and this is why we also provide interquartile ranges in Table 3. We have added the mean absolute error as a metric to Table 3 (see reply to the reviewer's comment RC2.5 above) and in the text, as suggested by the reviewer. We have also indicated in the text the overall MAE once the three worst performing sites (Djankuat, Lirung, Tasman), as well as the uncalibrated models, are not taken into account. The revised text reads as follows (lines 526-529):

"Overall, models tend to overestimate melt, with a median final melt error of 4.3% **and mean absolute error of 43.4%** across all model runs (Table 3), dominated by the three sites where melt is consistently overestimated. However, there are important differences between models **and sites. When removing the effect of the three worst performing sites, and the uncalibrated models, the median error across all model runs is -2.8%, and the mean absolute error 17.5%."**

RC2.8 Figure 9: It's interesting that the error exceeds the uncertainty in the majority of cases. What is the uncertainty analysis missing here? Does it come down to the factors discussed in Section 6.2, or is it something more systematic? Maybe you can just add a line or two about that here?

In fact, the uncertainty is low because knowledge of debris properties and their expected ranges was limited at the time this project was established and its experimental setup designed. Standard ranges (10% for thermal conductivity, surface roughness and porosity and 5% for emissivity, following Reid and Brock., 2010) were applied to each debris property, without a full appreciation of the real magnitude of variation of these. A discussion regarding the debris properties knowledge gap is in Section 6.1.5, but we agree we did not specifically discuss here that errors were larger

than uncertainty in most cases, and why. We have expanded this discussion in lines 707-710, as below:

"From all the cases considered in this intercomparison, and from the variety of approaches adopted to determine debris properties, it is apparent that debris properties are not well constrained at most sites, estimates from literature are often not appropriate, and even more importantly, that published methods to determine conductivity in the field (Conway and Rasmussen, 2000; Nicholson and Benn, 2012; Reid et al., 2012) may not agree, as exemplified by the case of Lirung, and confirmed by a separate study (Melo-Velasco et al., 2025). In addition, even if we are able to constrain debris properties at an individual automatic weather station, their values are affected by differences in porosity and pore water content across the debris**-covered** areas of a glacier, and therefore are also likely to vary considerably in time and space. Neither aspect has seen much investigation to date. **Limited knowledge of the variability of debris properties also led to applying literature-derived uncertainty estimates to the debris properties (~10% for surface roughness and thermal conductivity), which are likely too narrow, leading to a resulting uncertainty in modelled melt smaller than the actual model error (Fig. 9).**"

RC2.9 L633-647: this seems like a key point – why not run all of the models with and without these modifications given that they were 'known' to be incorrect, at least in terms of the debris thickness? I appreciate that would be a pain, but it doesn't seem right to have knowingly run the models with incorrect values…?

The reviewer points to an important aspect of the modelling that we have discussed at length. There are two aspects here to be considered for Lirung: 1) debris thickness at the AWS was different from the one at the closest stake; and 2) conductivity was likely overestimated.

As for 1), we would not call the debris thickness wrong - this was the debris thickness measured at the site of the AWS, and we think it is important to run the model with the actual debris thickness of the site where the measurements of meteorological forcings were taken - anything else could lead to physical inconsistencies. Indeed, for this experiment, it would be wrong to run the model with a debris thickness that did not correspond to the meteorological and surface conditions at the AWS. The difficulty was that the stakes were several meters away with a distinct debris thickness, which gave an apparent large modelling error. We were able to show however that this error could be partly explained by the difference in debris thickness between the AWS and stake location. We thus thought that the analysis described in this section was a good compromise to have physically consistent simulations and at the same time point to the difficulties of comparing simulations at AWS with measurements at stakes that are often several meters away, and at distinct debris thicknesses.

As for 2), it was difficult to decide that debris conductivity was wrong. This awareness grew with the development of the experiment and analysis of the simulations. Limited time and resources made it challenging to ask modellers to re-run their models, and we therefore performed the additional sensitivity analysis described in these lines and Table 5 with the model used in the research group of the core team of authors (which had turned out to be among the best performing of the EB models). This resulted in the analysis presented in this section, which qualifies the simulations conducted with the standard values provided to modellers. We would not call those simulations wrong though, for the reasons described above.

We think the additional model runs with the best performing EB model of this experiment provide interesting results that caution against use of standard literature values, and call on the need to carefully measure debris properties (including thickness) at the sites of AWS. This indeed became one of the paper's main results.

From this experiment, it is interesting to note that the debris thickness difference only partly explains the melt overestimation, while the combined effect of a reduced thermal conductivity and increased debris thickness improves performance considerably. Our sensitivity analysis confirmed that the difference in the debris thickness between the station and UDG location mattered, something that is often overlooked. We have revised the text to make these points clearer in lines 671-673:

"It is therefore a combination of at least these two factors (heterogeneous debris thickness between the automatic weather station and validation site, and uncertainty in the site conductivity) that likely explains the poor performance of all models at this site. **Our sensitivity analysis thus confirms that deviations between the debris thickness at the location of the automatic weather station and the location of the sub-debris melt observations can have substantial implications for sub-debris melt modelling assessments**."

RC2.10 L940: One clear outcome from the analysis is that TI models perform surprisingly well – indeed four out of the top five models are temperature index based (Table 4). I wonder if you can highlight this a bit more clearly in the conclusions? L941 states they perform 'very well' once calibrated, but then goes on to point out that they don't perform well if not calibrated. I don't think this does them justice, given Figure 7 very clearly shows that the calibrated versions of the TI models comfortably outperform the EB models in the scenarios tested! I think it's worth flagging this up more explicitly to the reader here.

We agree with this comment and have now highlighted more that the best performing models are calibrated TI models. We want to note here again though for the reviewer that only the **calibrated** TI models perform so well, and the uncalibrated are the worst performing models. As noted by the reviewer above, TI models (and other calibrated models) were not evaluated over a validation period separate from the calibration, which is a fundamental test to assess their performance.

Nevertheless, we have taken the reviewer's point on board, and have revised the text for the first paragraph of the conclusions starting in line 970:

"Energy balance models and empirical temperature index models perform in a distinct manner and serve distinct purposes. ~~In general, t~~**T**emperature index models perform ~~very well (median performance)~~ **best in this experiment, and better than energy balance models,** when calibrated, and poorly **(worse than energy balance models)** when applied in their uncalibrated form. However, ~~the Hyper-fit model and~~ the DETI model show**s one of the best performances with parameters from one site applied to all other sites. The Hyper-fit model shows** that site-specific ~~literature~~ **estimated** parameters ~~determined outside of the study period~~ **(i.e. parameters from the same site but outside of the study period)** can produce satisfactory melt estimates."

RC2.11 L944: On a similar point, it's clear that the EB models can quite seriously overestimate melt, but the second bullet has quite a positive spin – highlighting that increased complexity improves process-based representation at the debris-atmosphere interface. That may be so but

do the results not also suggest that increased complexity can lead to overestimation? The bullet point might just be better balanced by also clearly stating where they are also deficient.

Our conclusion arises from the fact that the most complex EB models, according to our definition of model complexity, perform best (among EB models only). While it is true that overall energy balance models overestimated melt, this is **only the case when the three sites with inaccurate debris properties are considered**. The median error among energy balance models when all sites are considered is 20.6%, but this error is reduced to -6% when the three poor performing sites (Djankuat, Lirung, Tasman) are not considered. As the reason why those sites overestimate melt is very clear, we do not think that EB models on average seriously overestimate melt. However, errors can still be large, and we agree this should be more clearly stated, including the differentiation of overestimation vs underestimation depending on whether the three bad performing sites are considered or not. We have revised the text for bullet point 2 of our conclusions:

"Energy balance models show a range of performance and model skills, **with an overall overestimation of melt and relatively large errors (20.6% median, 44.8% MAE) when all sites are considered, and a slight underestimation of melt and much smaller errors (-6.3% median, 18.7% MAE) when the three sites with poorly constrained debris properties are not considered.** A clear finding from this work is that th**e energy balanc**e models that perform best are those with the highest degree of complexity at the debris-atmosphere interface. We were not able to demonstrate the added value of additional complexity within the debris, because of lack of data representative of processes within the debris layer. The use of simplifying assumptions (and of a linear temperature profile within the debris in particular) within the model that included convection in the debris made it difficult to disentangle the importance of this process."

We have also revised the text in the results (lines 542-544), to include the overall performance of energy balance models with and without the bad performing sites:

"Clear differences are also evident among energy balance models**, which show an overall overestimation of melt and relatively large errors (20.6% median, 44.8% MAE) when all sites are considered, and a slight underestimation of melt and much smaller errors (-6.3% median, 18.7% MAE) when the three sites with poorly constrained debris properties are not considered.** The models that perform…"

RC2.12 Just to note that the data and simulations are not currently publicly accessible – I presume they will be made so if/when the paper is published.

Yes, the "restricted access" on Zenodo will be changed to "open/available" with the publication.

RC2.13 Consider also giving the links to the model codes (rather than just their references)?

We find this a very good suggestion. We agree with the principles of publishing all model codes, and most of them are already available at the model references provided, or directly requesting them from the authors. However, enquiring to all modellers about their codes and requesting to make them suitable for publication (understandable, with comprehensive comments) is not something we can quickly do in the timeframe of this manuscript revision. We therefore encourage the readers to contact the modellers of this article (indicated in author contributions) to request the model codes.

CHANGES IN KP1 AND KM1 DENOMINATION

As part of the discussion about the best designation/definition of some of the empirical models, we have decided to indicate models KM1 and KP1 as two separate models, to accommodate the KO2, KM1 and KP1 modeller wishes in that discussion.

We have thus modified a few places in the text to indicate 15 instead of 14 models (Abstract, Intro and Models sections); adjusted Figure 3 and Table 2 to split KM1 and KP1; and we have made tiny adjustments to three sentences in the text (pasted below as in the track changes document):

Page 13: For some of the latter models ($DETI_m$~~, KM1~~ and Hyper-fit), the uncalibrated versions were also run to assess their transferability.

Page 26: Finally, at Djankuat, even the more empirical models, ~~in~~ both ~~their~~ calibrated and uncalibrated ~~versions,~~ fail to match the observed melt, with the exception of the calibrated KM1.

Page 30: The KM1 model performs considerably better than ~~its uncalibrated version~~ KP1, a very similar model which is uncalibrated (Fig. 7, Table 3).

None of our results, conclusions or any major aspect of the paper has changed as a result of these adjustments. We apologise for the late change.

We have also revised the text regarding model description and calibration for models KP1, KM1 and KO2, in the main text and in the Supplement, as shown below (pasted as in the track changes document), with no effect on the results or conclusions of the study:

Temperature index models

Temperature-index models assume that melt is linearly dependent on air temperature and use a degree-day factor to estimate melt. The degree-day factor is generally calibrated to reproduce observed melt under debris (e.g. Kayastha et al., 2000), and likely cannot be transferred to sites with a different debris thickness or different climates ~~(Winter-Billington et al., 2020)~~. Anderson and Anderson (2016) developed a sub-debris melt model (Hyper-fit; Anderson et al., 2021) where a degree-day factor for clean ice is used to estimate a hypothetical bare ice melt rate at each site. To estimate sub-debris melt, the bare ice melt rate is reduced based on local debris thickness and a characteristic debris thickness length scale, h*. The characteristic length scale controls how rapidly sub-debris melt asymptotes toward zero melt as debris thickens, via a hyperbolic relationship. The parameter h* can be calculated as a function of debris properties (conductivity and porosity, and ambient conditions) but the model performs best by constraining h* directly using empirical debris-thickness melt data. The model has two parameters: DDF_{ice} and h* (Fig. 3, Supplement Sect. 2.4).

Winter-Billington et al. (2020) introduced two modifications of the temperature-index model ~~. by Kayastha et al. (2000), both~~ designed for daily simulations. Models KM1 and KP1 differ from KO2 in that they do not use shortwave radiation and use debris thickness as the sole fixed effect. The models KM1 and KP1 share the same structure but are considered different models because they differ in their training dataset and parameter values, as well as calibration scheme (see Sect. 4.1). ~~In the first one (KP1/KM1), the degree-day factor is computed as a~~

~~function of debris thickness through an empirical relationship with seven parameters, while in the second model (KO2) the degree-day factor is a function of both debris thickness and potential incoming shortwave radiation that has four empirical parameters.~~ The last model included in this intercomparison is the DDF_{debris}, the simplest degree-day factor approach calibrated for sub-debris melt reduction (Fig. 3).

And in section 4.1 on model calibration:

~~The KP1 model (in which the degree day factor depends on only the debris thickness) was run in its calibrated (KP1) and uncalibrated (KM1) version, with the calibrated version optimising the threshold temperature T_0 using cumulative melt data while all other empirical coefficients are as in Winter Billington et al. (2020). The KO2 model (in which the degree day factor depends on both debris thickness and net shortwave radiation) is uncalibrated, as it uses the empirical coefficient from Winter Billington et al. (2020).~~ In models KP1, KM1 and KO2, the parameters b_0, b_1 and b_2 were not recalibrated. The model parameter values are applicable to any site without recalibration by definition of the mixed-effects modelling approach (Winter-Billington et al., 2020). However, the value of the melt onset threshold air temperature (T_b) that was used to calculate positive degree days (PDD) as input to model KM1 was recalibrated for each site. Due to the original data used to fit the models, it was not possible to recalibrate the value of T_b to compute PDD for input to models KP1 or KO2 (see Winter-Billington et al. (2020) for details), and therefore these models are considered uncalibrated.