Major comments:

Unfortunately, it's quite obvious that large parts of the manuscript are either written by AI or at least strongly formulated by AI. There seems to be extensive use of "excess vocabulary". If you use AI for formulating or fixing your grammar (as acknowledged in the acknowledgments), it shouldn't be that obvious at least. In my opinion, the credibility of the results suffer, even if all results are valid and correct. For the revision, I advise the authors to rely less on AI or actually only use it to fix grammatical issues without letting ChatGPT write whole sentences.

- We thank the reviewer for this comment, though we find it speculative to assume that the manuscript was “generated” or “written” by AI.
- We point our that chatGPTzero indicates 70 % human, 12 % mixed and 18% AI The latter comes from Grammarly, as we explain below with an example
- We do acknowledge the use of AI, which is allowed by the journal’s policy. As an example, we copied one section of an old article by the first author Tedesco (reported below) and we used Grammarly (not even ChatGPT in this case) to improve style and grammar. Upon this, the ChatGPTZero detector suggested 61 % AI-generated and 39 % human-generated, whereas before it was 100 % human-generated. Again, we do acknowledge the use of AI but we would kindly ask the anonymous reviewer to give us credit for our original work.

Original text: [26] A likely cause of this extreme melt event was an anomalous ridge of warm air, acting as a strong heat dome that became stagnant over Greenland. The heat dome is identifiable in the 500-hPa height anomaly from the National Center for Environments Prediction (NCEP) Climate Data Assimilation System (CDAS,http://www.cpc.ncep.noaa.gov/products/intraseasonal/). Note that the 500-hPa geopotential height anomaly over Greenland, defined as the Greenland Blocking Index [*Hanna et al.*, 2012], was the strongest for June 2012 in the 1948–2012 NCEP Reanalysis record for June [*Overland et al.*, 2012]. The ridge was one of a series that has dominated the weather across Greenland since the end of May, with each successive ridge being stronger than the previous one. The heat dome began to dissipate by 16 July 2012. Then, another ridge that was not as strong as the earlier one came in and dominated mainly in southern Greenland. This later dome coincided with the melt on 29 July, which was not as extensive as the earlier extreme melt event.

[27] Historically, melt is rare in cold polar areas at high altitudes like Summit on the Greenland ice sheet. A pronounced ice layer from a significant melt event, which is

clearly evident in documented firn cores at many sites in Greenland, is the 1889 ice layer [*Clausen et al.*, 1988]. Records from the GISP2 deep ice core, which was retrieved at Summit, show that "thin ice layers which reflect melt from a single summer" occurred in the ice core only eight times between 500 and 1994 [*Meese et al.*, 1994].

[28] From the same core, *Alley and Anandakrishnan* [1995]studied melt layers from the upper 1565 m of the GISP2 core over a time period of 10,000 years. However, the frequency of melt occurrence varies widely in time as identified by ice layers in ice cores. Prior to the 19th-century event, another significant melt event occurred about 680 years earlier [*Meese et al.*, 1994] preceded by several events in the Medieval Warm Period (a.k.a. the Medieval Climatic Anomaly). Melt occurred once in about 250 years from 1000 to 4000 BP (referenced to 1950) and once in about 82 years from 5000 BP to 8500 BP according to *Alley and Anandakrishnan* [1995]. These significant melt events are widely sporadic in different periods of the Holocene, clearly exhibiting their non-stationary behavior. Thus, a single average value of melt frequency is not necessarily applicable to represent climate change at a given time period in the past, the present, or the future.

[29] In summary, this paper highlights the satellite capability for melt detection, combining data from multiple satellites to provide full coverage without gaps across the entire ice sheet. The satellite observations captured the GIS extreme melt event in its entirety within a short latency, allowing scientists to plan for timely science investigations given the rarity of the event. Given the vast GIS extent of 1.71 million km^2, impacts of this extreme melt event remain to be investigated; for example, the mass balance of the ice sheet, surface heat exchange in the boundary layer across GIS, or atmospheric chemical processes involving different states of the snow cover on GIS. Thus, the 2012 melt event is a historic record that may excite many new scientific research studies.

Text after Grammarly:

[26] A likely cause of this extreme melt event was an anomalous ridge of warm air, which acted as a strong heat dome that became stationary over Greenland. This heat dome is evident in the 500-hPa height anomaly data from the National Centers for Environmental Prediction (NCEP) Climate Data Assimilation System (CDAS, http://www.cpc.ncep.noaa.gov/products/intraseasonal/). Notably, the 500-hPa geopotential height anomaly over Greenland, defined as the Greenland Blocking Index [Hanna et al., 2012], reached its highest value for June 2012 in the 1948–2012 NCEP Reanalysis record [Overland et al., 2012]. The ridge was one of a series that dominated

Greenland's weather since the end of May, with each successive ridge stronger than the last. The heat dome began to dissipate by 16 July 2012. Subsequently, another ridge, not as strong as its predecessor, developed and primarily affected southern Greenland. This latter dome coincided with the melt on 29 July, although it was not as extensive as the earlier extreme melt event.

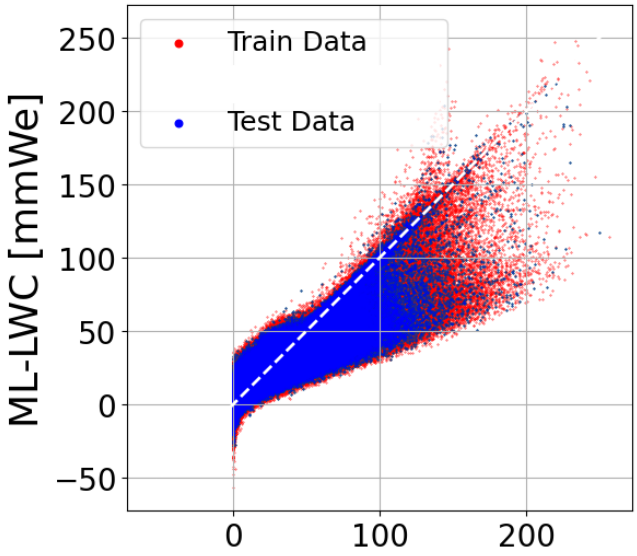
[27] Historically, melt events are rare in cold polar regions at high altitudes, such as Summit on the Greenland ice sheet. One prominent ice layer resulting from a significant melt event, clearly visible in documented firn cores across Greenland, is the 1889 ice layer [Clausen et al., 1988]. Records from the GISP2 deep ice core, retrieved at Summit, indicate that "thin ice layers reflecting melt from a single summer" were found only eight times between 500 and 1994 [Meese et al., 1994].

[28] Using the same core, Alley and Anandakrishnan [1995] analyzed melt layers from the upper 1565 meters of the GISP2 core over a span of 10,000 years. The frequency of melt occurrence, as identified in ice layers, varies significantly over time. Before the 19th-century event, another major melt occurred roughly 680 years earlier [Meese et al., 1994], preceded by several occurrences during the Medieval Warm Period (also known as the Medieval Climatic Anomaly). According to Alley and Anandakrishnan [1995], melt events happened once every 250 years from 1000 to 4000 BP (relative to 1950) and once every 82 years from 5000 BP to 8500 BP. These significant melt events are sporadic throughout different periods of the Holocene, clearly demonstrating their non-stationary nature. Therefore, a single average value for melt frequency does not adequately represent climate change during any particular period in the past, present, or future.

[29] In summary, this paper emphasizes the capabilities of satellites for melt detection, combining data from multiple sources to ensure complete coverage of the entire ice sheet. Satellite observations captured the extreme GIS melt event in its entirety with minimal latency, enabling scientists to plan timely investigations given the rarity of such an occurrence. Considering the vast GIS area of 1.71 million km^2, the impacts of this extreme melt event remain to be fully explored—for instance, changes in the ice sheet's mass balance, surface heat exchange in the boundary layer, and atmospheric chemical processes involving various snow cover states. The 2012 melt event stands as a historic record, likely to inspire numerous new scientific research efforts.

While it's true that ML approaches in ice sheet context are still quite sparse, I think this study has to make more clear why it's novel. At least, it is not clear to me how the ML-based approach is better than for example a simple linear regression. I am missing a comparison with some baseline model. How does it compare to simply doing linear regressions for example? Is there even a clear gain of information? For the revision, I think it's necessary to include such a comparison.

- We thank the reviewer for this important suggestion. We addressed this comment in two ways. First, to clarify the predictive value of the proposed approach, we added a comparison with a simple ordinary least-squares linear-regression baseline in Section 3.1. Using the MAR-IA1 predictor set, the linear model achieved $R^2 = 0.90$, and RMSE of 3.35 mm w.e. day^{-1} compared with $R^2 = 0.98$, and RMSE = 1.34 mm w.e. day^{-1} for the corresponding XGBoost-based MAR-IA1 model, indicating a clear gain in predictive skill from the non-linear emulator.



- Second, we revised the Introduction to clarify the novelty of the study more explicitly. In particular, we now distinguish our work from recent ML-based downscaling studies and emphasize that our contribution is the development of a direct emulator of daily surface meltwater production from MAR together with interpretable attribution of melt drivers using SHAP. (section 2.2) We hope these revisions make both the added predictive value and the novelty of the manuscript clearer.

I think the introduction needs quite a lot of work. I am missing a discussion and context/comparison to recent advances of ML approaches in ice-sheet modelling/observations. Just some examples that should/could be discussed in the introduction at least: Lütjens et al. 2025 (https://arxiv.org/abs/2512.12142), Bochow et al.

2025 (https://egusphere.copernicus.org/preprints/2025/egusphere-2025-3927/). Especially a comparison with Schlager et al. 2026 (https://egusphere.copernicus.org/preprints/2026/egusphere-2026-7/) is necessary. To be fair, that preprint was posted after this paper but it seems like there is a very similar approach/idea described but instead of XGBoost a neural net is used. How does your model compare with theirs? What are the differences/similarities?

- We thank the reviewer for this helpful suggestion. We revised the Introduction to discuss Lütjens et al. (2025), Bochow et al. (2025), and Schlager et al. (2026), and to clarify how our study relates to and differs from each.
- Specifically, we now note that Lütjens et al. (2025) focuses on deep-learning-based spatiotemporal downscaling of daily 100 m surface meltwater maps by combining remote sensing with physics-based model output over the Helheim Glacier, while Bochow et al. (2025) develops a physics-constrained generative ML framework for high-resolution downscaling of monthly Greenland surface mass balance and surface temperature fields. In contrast, our study develops a direct emulator of daily surface meltwater production from MAR and emphasizes interpretable attribution of melt drivers using SHAP.
- We also added an explicit comparison with Schlager et al. (2026), which is conceptually the closest study to ours. Both studies use machine learning to emulate daily Greenland surface melt from polar regional climate model output. However, Schlager et al. use a neural-network emulator trained on HIRHAM5 and its firn model DMIHH forced by ERA-Interim, whereas we use XGBoost trained on MAR output. In addition, our manuscript places particular emphasis on SHAP-based attribution of melt drivers and their spatial and temporal evolution, whereas Schlager et al. emphasize a physics-informed neural-network design combining short-term weather patterns with long-term climate memory.
- These changes are reflected in the introduction.

In general try to make some sentences short and more on the point. There are quite a lot of nested sentences that make it hard to follow thoughts.

- We have revised the sentences to the best of our ability, as no specific indication was provided by this reviewer. We hope this is satisfying.

Specific comments:

L.19 What is a low MSE? I think a relative error would be better here.

- Should we just do RMSE/RMS for MW?

L.20 SHAP is not clear to everyone (acronym).

- We have revised the abstract with the full form of SHAP.

L.21 The long dash is a dead giveaway for AI use.

- This has been removed from the abstract. We appreciate the suggestion, but we would kindly ask the reviewer to be respectful of our work. AI is generally used to help with grammar and other topics, and we use it for this purpose. This does not impact the originality of our work and its quality. As per the journal's policy, we acknowledge the use of AI to revise text and suggest improvements, in accordance with the journal's guidelines.

L.26 The last sentence seems out of place and reads more like an opinion.

- This has been removed from the abstract.

L.31 I would say increased melt instead of enhanced.

- This has been changed in the abstract.

L.33 The MAR acronym is written out, RACMO and HIRHAM are not.

- The acronym has now been written out.

L.35 Driving processes?

- This has been changed to "surface energy- and mass-balance processes that govern meltwater production."

L.37 "Meltwater production" sounds odd.

- This have been changed to surface melt

L.37 In Pirk et al., neither Greenland nor the word "melt" is mentioned even once. Are you sure it's the right reference here?

- This reference has been removed.

L.44-45 Where are the references for that claim?

- We have modified that sentence and the reference is not required anymore

L.45-46 Was there a previous study emulating melt dynamics, or is this something you do for the first time? If the latter, I do not understand the sentence.

- We have revised the sentence to clarify that it refers to the general potential of ML models trained on RCM output and added prior literature on melt emulation.

L.51 What does it mean that predictands are "drawn" from SEB components?

- We have changed the wording to, hopefully, clarify.

L.54 Do you use only ERA5, or some other dataset as well? If only ERA5, I would remove "such as."

- Removed "such as"

L.55 What is "evolving importance"?

- Removed the word "evolving"

L.57 In the ML community, "benchmark" is used in a different context, it could be confusing here.

- We revised the sentence to avoid the term "benchmark" and now describe the study as demonstrating the potential for interpretable ML models in cryospheric research.

L.86 Maybe add example tasks, especially in the context of ice sheets/climate modelling.

- We revised the opening of Section 2.2 to avoid attaching a gradient-boosting reference to a broad statement about machine learning in general. Bentéjac et al. (2021) is now cited in the XGBoost-specific context, where it is more appropriate. We also added recent cryosphere-focused ML examples in the Introduction to provide broader context for the types of ML tasks relevant to this study.

L.92 I don't think it's necessary to list boosting and decision trees as (a) and (b) here. I would suggest rephrasing the sentence.

- This sentence has been rephrased to "XGBoost combines two key ideas: boosting and decision trees."

L.92f I also think the description of boosting, and especially trees, is not very clear. For example, in line 96, what kind of thresholds? For the audience of TC, where I assume there are not many ML experts, I think this should be described a bit more extensively. For example: what kind of regularisation term? What does "efficiently" mean, compared to what? Neural networks?

- We revised this part of Section 2.2 to improve clarity for a broader cryosphere audience. Specifically, we rewrote the description of boosting using regression-appropriate language, clarified how decision trees partition the predictor space through binary splits, and expanded the explanation of regularization and computational tractability.

L.100 The term “features” was not introduced.

- We revised this paragraph and now use the terminology “predictors” and “input variables” consistently, so the previously undefined term “features” no longer appears.

L.105 I’m not sure why the metaphor of players is introduced. I don’t think it’s necessary. I think it’s easier to understand if you phrase it in terms of variables/predictors rather than “players.”

L.109/110 Rephrase the sentence, or split it into two sentences.

- Please see below, response to Line 123

L.122f Split the sentence.

- Please see below, response to Line 123

L.123 What does “assign” mean? Is this also computed?

- We have revised section 2.3 to make it easier to read.

L.140 It is not necessarily clear what you mean by “distribution of predictors.”

- We revised the text to clarify that Figure 1 shows the statistical/value distributions of the predictor variables, displayed using kernel density estimates.

L.144f Split the sentence

- We have revised this sentence.

L.147 Missing)

- This has been fixed.

L.148 A word is missing.

- We revised the sentence

L.149 Why were shortwave and longwave radiation removed?

- As replied to another reviewer, we point out that the emulator is not aiming to create a new physical model or develop new theories about melting and SEB, but simply to replicate the model’s performance while remaining consistent with energy-balance

knowledge. We decided to remove the longwave because we thought the relatively small improvement from its inclusion might not be worth it, given the potential impact on data availability for extending the emulator's use to colleagues. Again, the point is to have a model that balances computational efficiency, the quality of outputs that are close to the original climate model, and applicability to others using reanalysis or other datasets.

L.164 Explain five-fold cross-validation.

- A brief description has been added

L.189 80% + 20% + 10% = 110%

- This have been fixed to 70+20+10

L.197 Here you introduce the term “bias”, however, you used it earlier. Maybe move the definition of bias, MSE, etc. to when you first mention these terms.

- This was moved to the previous section, 2.5 hyperparameter optimization

L.200 I’m not a fan of using superlatives like “extremely”.

- The word has been removed.

L.217 This sounds odd: “explanatory nature”

- This has changed to interpretability

L.283 Why is there a “["?

- This has been removed.

L.340 Split the sentence; it’s also not clear what you want to say.

- The sentence has been improved

L.353 Again, split the sentence. Please avoid sentences that run over four lines or more.

- These sentences have been split.

L.373f What does “such as” refer to?

- We rewrote the sentence.

Figures and Tables

The caption are generally not descriptive enough. Please extend them. The figures and tables should be self-explanatory without looking into the text.

Tab. 2 Metrics on what? The test set?

- This is on the test set and has now been mentioned in the text.

Tab. 3 Avoid refering to other figures/tables in the caption.

- This has been fixed.

Fig. 2 Did you check if the correlation of the variables is approximately the same in the training/test/validation data set? Colorbar has no label

- A label has been added to the colorbar. As mentioned, we shuffled and randomly split the dataset. We hope this answers the question.

Fig. 3 last colorbar is missing the label

- We added “unitless” to make it clear that albedo has no unit

Fig. 4 labels are missing (a,b,c ...). Maybe make the data points transparent. Currently the blue points overlay everything. The quality is not the best; consider using a vector-based figure.

- We improved the figure as requested by this and another reviewer

Main Concern #1: The train/validation/test split

Lines 189-190 describe that the data was split into 80% training, 20% validation, and 10% test set. Besides the effect that this sums up to 110%, and that the use of the 3 subsets should be explained more closely, my major concern is regarding the splitting strategy, which is not explained. The data splitting strategy for timeseries data needs to account for the high temporal correlation of samples, as a random splitting strategy for timeseries data yields a test set which is not independent of the training and validation set, and can thus result in overoptimistic performance score on the test set. The authors thus need to clearly describe their train/validation/test splitting strategy, and how it sufficiently takes into account the temporal correlation of the data. Since the cross-validation uses simple k-fold cross validation, which is not appropriate for timeseries data, the associated cross-validation scores are not trustworthy. Moreover, if the authors did use a split strategy that does not take into account the temporal correlation for the following model trainings, all the scores cannot be trusted either, and the experiments would need to be repeated with a more appropriate split. Literature for timeseries forecasting discusses splitting methods and pitfalls when working with timeseries data, which need to be considered also when regressing on timeseries data (see e.g. Hyndman et al., 2018).

- We agree that the original description of the data split was incomplete and that the percentages were incorrectly stated. We have corrected the text to 70% training, 20% validation, and 10% testing, and we now state explicitly how the split was performed.
- In the experiments, the full dataset was randomly shuffled prior to splitting. To assess whether temporal correlation was inflating model skill, we performed an additional chronology-preserving experiment in which the models were trained using data up to 2014 and evaluated on data from 2015–2024. We carried out this test for MAR-IA2 and MAR-IA2-ERA. The resulting skill was unchanged: MAR-IA2 remained at $R^2 = 0.97$and MAR-IA2-ERA remained at $R^2 = 0.83$. This indicates that, for these configurations, the main conclusions are not sensitive to whether the split is random or temporally separated.
- We have added this clarification to the revised manuscript and now note explicitly that the reported performance is robust to a temporally separated train/test split.

Main Concern #2: Attribution analysis and SHAP values: The authors do not sufficiently separate the use of SHAP values to explain the model predictions on the one hand and explain the physics on the other hand. While the SHAP values can serve as a “sanity check”

for the model and to explain the reason for the model's predictions, care needs to be taken to translate that into interpretations of underlying physics.

The explanation of the SHAP implementation is not clear to me. Some more clarification on the implementation and mentioning the used libraries are needed. I wanted to find out more via the code as there is some Code/Data Availability given. However, the given Zenodo records include only data, not the code.

Lines 109-112, 219, 326-327 explain the use of SHAP values for model interpretation, by quantifying the contribution of each input feature to the model's prediction. However, lines 57 and 124 then indicate that these SHAP values are used to draw physical conclusions and find causal relations; and in line 334 it is claimed that the physical interpretability is now confirmed. However, just because the SHAP values do seem (mostly) reasonable, it does not mean that they explain causal relationships. Based on the current explanation, I am not convinced that the SHAP analysis supports conclusions exceeding the explainability of the ML model itself. For example, in line 145 the issue of using correlated inputs for attribution analysis is mentioned, and in line 296 the fluctuation of the SHAP values is explained by the remaining correlation of the features. However, the discussion beforehand used the SHAP values to draw physical conclusions, not discussing how these correlations may influence the SHAP values, and thus the trends observed. Also, direct melt drivers are identified using SHAP values, although input variables were used, which are not direct melt drivers. Furthermore, the fluctuation of the SHAP values for longitude in Figure 8 is not discussed at all.

Besides my doubts on the validity of the attribution study using SHAP values, the use of an ML emulator for such an attribution study is not properly motivated. The MAR model delivers all the radiation and turbulent heat flux values to calculate the surface energy balance and explain the melt drivers, see also Wang et al. (2021), Zhang et al. (2023), and Hofer et al. (2017).

The terminology related to SHAP is currently inconsistent (e.g., "Shapley values", "SHAP values", and "Shapley coefficients"), which is rather confusing. In the ML literature, "SHAP values" is typically used for the specific method, while "Shapley values" refers to the underlying game-theoretic concept.

- We thank the reviewer for this comment. We agree that SHAP values should be interpreted primarily as a tool for model interpretation, rather than as direct evidence of causality or as a standalone diagnosis of the underlying physics. In response, we revised the manuscript throughout to clarify this distinction.

- First, we now state explicitly in the Methods section that SHAP values are interpreted as model-based attributions of the trained emulator. They quantify how the emulator distributes predictive importance across the provided predictors, but they do not by themselves establish causal physical relationships. We have accordingly softened several statements in the Introduction, Results, Discussion, and Conclusions that previously implied stronger causal or physical claims than warranted.
- Second, we added an explicit limitation statement noting that correlations among predictors can redistribute SHAP attribution among related variables, even after removal of highly collinear inputs. We now clarify that the SHAP results are discussed as diagnostics of emulator behavior, to be used towards physical understanding, rather than as direct proof of physical melt drivers. Relatedly, we revised the discussion of geographic variables such as latitude and longitude to avoid interpreting them as direct physical drivers and instead describe them as proxies.
- Third, we clarified the SHAP implementation in the Methods section, including the library used. Specifically, SHAP values were computed using the Python shap library applied to the trained XGBoost models; in our implementation, shap.Explainer was used, which for tree-ensemble models applies the corresponding tree-specific SHAP method.
- Fourth, we standardized the terminology throughout the manuscript. We now use "SHAP values" consistently for the method outputs, while reserving "Shapley values" only for the underlying game-theoretic concept. Terms such as "Shapley coefficients" were removed to avoid confusion.
- Finally, we clarified the motivation for the emulator-based attribution analysis. Our objective is not to replace direct process-based diagnosis from MAR's native SEB output. Rather, the emulator and SHAP analysis provide a complementary and computationally efficient framework for interpreting the learned melt surrogate.

Main Concern#3: The selection of input variables

While the conclusion highlights that predictor selection was guided by physical relevance and statistical analysis, I don't see that this was done in a sufficient manner.

Line 80 mentions 2-meter air temperature to be a driver of melt energy. However, while 2-m air temp. is closely related to heat fluxes, considering the surface energy balance, it is not a driver of melt itself, as the surface energy balance is directly driven by shortwave and longwave net radiation, sensible heat flux. latent heat flux, and ground heat flux (Lenaerts

et al., 2019). The use of additional inputs, especially when trying to interpret the drivers physically, needs some more explanation.

Specifically, the motivation for using topographic variables needs more explanations. Line 83 claims that topographic variables modulate the local energy balance and atmospheric conditions. However, when using the atmospheric conditions themselves as model inputs, why use the topographic variables that contribute to those atmospheric conditions?

- We thank the reviewer for the comment. We are not sure what the reviewer means "done in a sufficient manner". We considered the drivers of melting from a Surface Energy Balance perspective and included air temperature because it is correlated with skin temperature, but, as we explain, it does not saturate at 0ºC when melting occurs. The air temperature, therefore, can provide further information to the emulator about the intensity of melting or any other information the data could provide to the model. Same for topographic variables. We wanted to explore the information content that such proxies could provide to the models' performances and also add an input that could potentially be explored in the future by others. We point out that the emulator is not aiming to create a new physical model or develop new theories about melting and SEB, but simply to replicate the model's performance while remaining consistent with energy-balance knowledge. We also decided to remove the longwave because we thought that the relatively very small improvement due to its inclusion might not be worth it in terms of potential data availability for extending the use of the emulator to colleagues. Again, the point is to have a model that compromises computational efficiency, quality of the outputs that are close to the original climate model, and applicability from others using reanalysis or other datasets.

Main Concern #4: Conflicting model scores

The abstract claims an R2 of 0.99, but none of the resulting models reach such a high score according to Table 2. Also, the formulation implies that this high score is reached for the model using ERA5 inputs, which seems misleading.

- We agree that the abstract wording was misleading. The value of $R^2 = 0.99$ referred to the optimum identified during hyperparameter tuning, not to the reported models in Table 2. We have therefore revised the abstract to report the performance of the final models consistently with Table 2, and we now state that the best-performing configuration reached $R^2 = 0.98$. We also revised the wording to avoid implying that this performance was achieved by the ERA5-based model.

The discussion, conclusion, and Table 2 do not mention if these are the scores on the test set. Also, the scores in the conclusion are different from those in the discussion and in Table 2.

- We have revised the manuscript to state explicitly that the reported quantitative performance metrics are test-set scores. In particular, we clarified this in Section 3.1 and in the caption of Table 2. We also corrected the Conclusions so that all reported R^2, MSE, and related values are fully consistent with the final test-set results shown in Table 2. Thanks for pointing this out.

Inconsistent use of metrics: in 2.5 MSE is reported, in 3.1 RMSE, and in 3.3 mean absolute difference and mean difference are used.

- To improve consistency in the presentation of model performance, we added RMSE values to the site-based evaluation in Section 3.3, while retaining the previously reported summary statistics for direct comparison with the original text. RMSE is now used alongside the existing reported quantities at Swiss Camp and K-transect S6.

Main Concern #5: Related literature

lines 41-43: The authors only mention one paper (Doury et al., 2023) for a specific ML application, which refers to a downscaling method and thus seems somewhat loosely connected to the melt emulation task, as there exists more closely related literature. Furthermore, the authors mention attribution analyses based on ML-emulation, which are not backed up by the two given references.

On the other hand, the ML approaches for downscaling climate fields (like Doury et al., 2023) should be discussed in the context of the MAR-IA2ERA models. Since these emulators rely on ERA5 data reprojected onto the MAR grid, it would be helpful to clarify that first downscaling ERA5 to the MAR configuration would ensure spatial consistency between the emulator inputs and the resulting melt predictions, and that a lot of research is being done in developing such downscaling via ML.

Lines 89-90: The given literature for the applications of XGBoost seems quite unrelated to the task in the paper too. Some additional references to work in climate/cryosphere research would be interesting, e.g. Veldhuijsen et al. (2025).

References should be double checked: Nghiem et al. (2012) was cited four times in the paper, but does not appear in the reference list. In line 291 it is used as only reference for a claim that refers to multiple studies. In contrast, Tedesco et al. 2016a and 2016b seem to be the same reference.

- We have substantially revised the Introduction to better position the manuscript within the recent literature on ML applications in cryosphere and climate science. In particular, we now discuss recent related studies by Lütjens et al. (2025), Bochow et al. (2025), and Schlager et al. (2026), and clarify how our work differs from downscaling-focused approaches and from recent neural-network-based melt emulators.
- We also clarified the role of the MAR-IA-ERA configuration. Specifically, we now note that ERA5 variables are reprojected to the MAR grid for use in the emulator, but are not dynamically downscaled to MAR-consistent fields, and we acknowledge that ongoing ML-based downscaling developments are highly relevant for improving such applications.

Further comments:

Inconsistent terminology, e.g. surface temperature and skin temperature are used interchangeably; same with inputs, variables, features, and predictands

- We have revised the manuscript to make the terminology more consistent throughout. In particular, we now use surface temperature consistently in place of mixed use of “surface temperature” and “skin temperature”; predictors for model input variables; target variable or meltwater production for the model output; and SHAP values for the attribution results. We also corrected several remaining instances where terms such as “predictands” and “features” were used inconsistently.

While the formulation “approximately normal distribution” in line 142 is odd in general, especially albedo and upward longwave radiation show distributions that are not comparable with a normal distribution at all. It would therefore be more accurate to state that the variables are not strongly concentrated within specific value ranges, rather than characterizing them as approximately normal. Furthermore, the x axis (or the data itself) seems cropped for most data, and the existence of long tails and extreme values cannot be judged.

- We agree that the phrase “approximately normal” was too strong and not accurate for several predictors. Since Figure 1 is intended primarily to compare the overall distributional shapes of the predictor variables rather than to characterize tail behavior in detail, we revised the text accordingly. The new wording avoids referring to them as approximately normal.

The hyperparameter optimization does not state the ranges that were chosen for the different parameters to be optimized, and how many different combinations were tried in total.

- We have revised Section 2.5 to state explicitly the hyperparameter ranges used in the Bayesian optimization. Specifically, the search space was defined as n_estimators = {100, 200, 500, 700, 1000, 1500, 2000, 2500}, max_depth = {2, 3, 5, 10, 15}, learning_rate = {0.05, 0.10, 0.15, 0.20}, min_child_weight = {1, 2, 3, 4}, and gamma = {0, 0.25, 0.5}. This corresponds to 1920 possible combinations in the full discrete search space. Because we used Bayesian optimization (BayesSearchCV) with n_iter = 50, 50 candidate configurations were evaluated adaptively rather than exhaustively testing all combinations.

I do not understand the claim in 365-367: “the MAR-IA emulator allows running past and future scenarios without forcing the atmospheric model in MAR with fields that are not always available for past periods and future simulations.” - Which fields are not available, which emulator product to you mean, and how do you justify the validity of extrapolating past and future data?

- We have added the following to clarify this: Moreover, the emulator offers a practical alternative when the full MAR atmospheric model cannot be run. MAR requires lateral boundary forcing over Greenland with meteorological variables at several vertical levels, and these inputs are not always available with sufficient completeness for historical reconstructions or future climate simulations. The emulator circumvents this requirement by predicting meltwater directly from a reduced set of atmospheric and surface predictors, without integrating the full atmospheric model. Its use for past or future applications is therefore promising, but should be restricted to conditions that remain within, or close to, the range represented in the training data, with further validation needed for true extrapolative cases.

Table 2 includes RMSE and bias for the 95% CI which was not explained nor mentioned or interpreted in the main text.

- We have revised the manuscript to explain that the 95% confidence intervals are reported in Table 2 and to mention them explicitly in the main text.

Figure 4 includes so many datapoints, that the distribution of errors is not visible; a density plot would be a better fit here. Furthermore, plotting test, validation and train data on top of each other is not of any use here, as the data points from train and validation set are mostly not visible anyway; and the error distribution of the subset was not discussed anyway. And again, it's unclear on which subset (train, validation, or test) the scores in the table were calculated.

- Figure 4 now only shows the test datapoints with color representing density

Figure 4 also shows the presence of negative melt predictions. In further applications, such values would likely be set to zero. It is therefore reasonable to truncate the predictions at zero and report performance metrics based on the truncated results.

- We thank the reviewer for this suggestion We have now modified the model to set to 0 any negative value. We have also created new plots and statistics that are included in the paper

General Comments:

The Introduction should be expanded to include recent work on RCM emulators. For instance, the study by Bochow et al. (2025) develops an RCM emulatordirectly relevant to this work. A thorough comparison with such studies would better position the contribution of the manuscript and clarify its novelty.

- We thank the reviewer for this helpful suggestion. We have expanded the Introduction to include recent related work on RCM-based machine-learning approaches, specifically Bochow et al. (2025). We now clarify that Bochow et al. developed a physics-constrained ML framework for downscaling monthly Greenland surface mass balance and surface temperature fields from MAR, whereas our study focuses on the direct emulation of daily surface meltwater production and on its interpretable attribution using SHAP.

XGBoost is widely used but relatively old, and all cited references are more than five years old (Lines 85-100). Could the authors clarify the rationale for

selecting this specific model? Additionally, the manuscript would be strengthened by including comparisons with other commonly used machine learning

approaches, such as Support Vector Machine (SVM), and potentially with more advanced deep neural network models

- We thank the reviewer for this helpful comment. We have revised the manuscript to clarify the rationale for selecting XGBoost. Specifically, we now note that our emulator is trained on a large tabular regression dataset composed of surface energy balance, meteorological, and geographic predictors, for which gradient-boosted tree methods remain strong and computationally efficient baselines. We also clarify that interpretability is a central objective of this study, and that XGBoost is particularly advantageous in this context because it enables efficient SHAP-based attribution analysis (line 92).

Specific Comments:

Providing a conceptual overview diagram of the RCM emulator would greatly enhance the clarity and readability of the manuscript.

- This is now fig 1

Lines 29-30, 119, 191: The references are neither chronological nor alphabetically arranged by first author.

- References are now ordered alphabetically

Line 37: Remove the extra space before the semicolon.

- This has been removed

Line 43: Spell out "XGBoost" in full at first mention.

- This have been changed to eXtreme Gradient Boosting (XGBoost)

Lines 102-125: The attribution analysis using Shapley coefficients is a key novelty but currently quite dense and may be difficult for readers to follow. Consider breaking it into smaller paragraphs for readability. Additionally, including a schematic or figure illustrating how SHAP values are computed and interpreted would help readers unfamiliar with the method.

- We thank the reviewer for this suggestion. We have revised the SHAP/attribution section to improve readability by breaking the explanation into smaller paragraphs and streamlining the description of the method. In particular, we now separate (i) the motivation for using SHAP, (ii) the interpretation of Shapley values, and (iii) the specific attribution metrics used in this study. (section 2.3). Concerning the schematic: we would prefer not to add this figure as Shapley is widely used in the literature and there are several schemes that can be accessed with an easy search on the web. We feel that adding a single figure without having the proper space to contextualize would imbalance the paper.

Line 189: Clarify the exact years used for each data subset (training, validation, test).

- We thank the reviewer for this comment. We have clarified that the training, validation, and test subsets were obtained from a randomly shuffled split of the full dataset, rather than corresponding to specific years. The revised text now states the exact proportions used for training, validation, and testing.

Figure 2 shows linear correlations between predictors and predictand. However, the relationships among surface energy balance components are highly non-linear. Using simple linear correlation may underestimate these dependencies. Consider using more robust measures that capture non-linear relationships

- We thank the reviewer for this comment. We have clarified the text around Figure 2 to state that the correlation analysis was used to identify highly correlated predictors and reduce redundancy in the input feature set prior to training.