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.

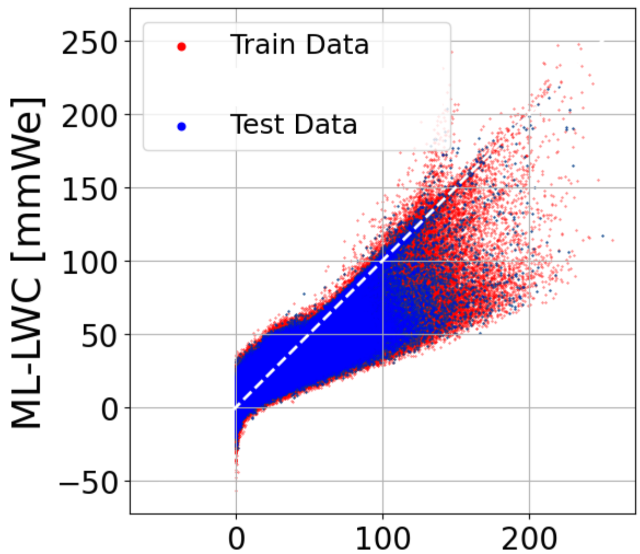
[27] Historically, melt events are rare in cold polar regions at high altitudes, such as Summit on the Greenland ice sheet. One prominent ice layer resulting from a significant melt event, clearly visible in documented firn cores across Greenland, is the 1889 ice layer [Clausen et al., 1988]. Records 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