Author replies to Anonymous Referee #1

General comments

This paper introduces a newly developed neural network-based emulator that predicts the temporal evolution of Greenland ice sheet surface melt. The emulator was trained on the output from the polar regional climate model HIRHAM5 and its firn model DMIHH, forced by the ERA-Interim reanalysis. It is clearly shown that the Modular NN configuration of the emulator, the standard setting developed in this study, can provide realistic information on the spatiotemporal evolution of ice-sheet surface melt, along with the daily melt amount. My impression is that this is a unique study that can provide useful information on the synergy between machine learning and cryosphere science. Although I think the information provided, in particular on the methods, can be improved, the results and discussion sound reasonable and sufficient to me. Therefore, I suggest that this paper can be published after revisions. I list some specific comments below.

A: We thank the reviewer for the positive and constructive assessment of our manuscript and appreciate the recognition of the novelty and relevance of combining machine learning with cryosphere modeling. After carefully considering the reviewer's comments, we recognise that the manuscript is not as clear as it could be, and we appreciate the opportunity to improve its presentation. Our explanations on how we will address the single comments are given below.

Specific comments

L. 9 "mean absolute error below 0.23 mm w.e.": Compared to what? What is the reference data for this comparison? Please explain.

A: We agree that the reference for this metric is not clear. We will rewrite this line: "Compared with DMIHH-simulated surface melt, the emulator achieves a mean absolute error below 0.23 mm w.e. per day on the independent test set across all six Greenland drainage basins, with errors largely attributable to spatial over-smoothing."

L. 45 ~ 58: It is worth reviewing and citing the paper by Hu et al. (https://doi.org/10.5194/tc-15-5639-2021) in this part.

A: We would like to thank you for bringing this paper to our attention. We will extend the related literature discussion in the introduction with this paper alongside some other related papers.

L. 59 “high temporal variability”: Can the authors explain this point quantitatively and add a reference for this argument if possible?

A: We agree that the term was ambiguous. In the revised manuscript, we will clarify that “high temporal variability” refers to the strong day-to-day fluctuations in melt at daily resolution compared to temporally aggregated (monthly/annual) quantities. We will emphasize that this reflects reduced temporal smoothing and increased modeling complexity compared to the works discussed above (lines 45-58), rather than a specific statistical measure.

L. 60 “temporal context”: I don’t think this technical term is widely recognized in the cryosphere community. Can the authors introduce additional explanations about the term so that more readers can easily understand?

A: We will explicitly define “temporal context” as the inclusion of atmospheric conditions from previous days as input features to the neural network, enabling the model to represent short-term memory effects.

L. 60 “While the models predicting annual ~”: Do the authors mean that the models refer to “ML” emulator? Or RCMs? Please clarify.

A: Here we refer to the ML emulators discussed in lines (45-58). We will clarify this together with the comments above.

L. 63: What do the authors mean by “lag effects”? Please explain in more detail.

A: We mean that surface temperature and melt onset depend not only on current atmospheric forcing but also on the previous thermal state of the surface and energy input to the surface, introducing a short-term memory. We will explain this more explicitly in the revised version.

L. 67: What do the authors mean by “model generalization”? Please explain.

A: We will clarify that “model generalization” refers to the ability of the trained emulator to perform well on unseen data sets. In the context of our work, this would include HIRHAM simulations driven by different GCM forcings (also under different SSP scenarios) and different (future) time periods. However, we do not claim that the present model has already demonstrated robust performance under such conditions. Rather, our point is conceptual: a firn emulator that operates independently of geographic location is structurally better suited for generalization than a location-specific model. This is

particularly relevant because the spatial extent and distribution of melt under future climate forcings or different SSP scenarios may differ substantially from those represented in the training data. A model that relies solely on atmospheric inputs, rather than location-specific features, is therefore less constrained by the spatial patterns present in the training period and may be better positioned for retraining under altered climate conditions.

L. 73 “Our model can be re-trained on data for future scenarios ~”: If the NN will be used for the future simulations of the ice sheet surface melt, do the authors have to train the NN using the output from the future climate simulations by an RCM such as HIRHAM5? Please explain more explicitly.

A: Yes, applying the emulator to future climate scenarios requires retraining on RCM simulations produced under those specific forcings. Since the neural network learns a mapping from atmospheric input to firn model output, reliable extrapolation beyond the training distribution cannot be assumed. Validity of temporal extrapolation or interpolation (e.g. to a lower SSP scenario) need to be investigated and are out of scope for this work. We consider it essential to clarify this point in the introduction to avoid suggesting that the emulator is universally applicable to all HIHRAM atmospheric data, irrespective of the applied forcing. The issue is briefly revisited in the conclusions (lines 337–339), and we intend to expand the discussion there, as it primarily concerns directions for future work.

L. 78 ~ 79: Please explain all the properties included in the daily output of the polar RCM HIRHAM5 with its firn model DMIHH.

A: We see that this subsection is not described clearly enough in general and causes confusion. We will revise the whole subsection, addressing this and all the comments below. In addition, Subsection 2.1 will be renamed to make explicit that it describes exclusively the generation of the DMIHH simulation data, rather than the emulator’s inputs/outputs.
All output variables from the polar RCM HIRHAM5 that are used as input to DMIHH are listed in lines 80/81. The DMIHH output comprises several SMB variables, of which only surface melt and albedo are considered in this study.

L. 79: What is the total snow and ice model layer thickness that DMIHH considers with the 32 model layers?

A: The 32 layers represent a firn/ice column of 60m w.e. thickness in total.

L. 84: It is better to explain how bare ice is determined in the DMIHH model.

A: The firn model scheme DMIHH operates on all grid cells that are defined to belong to the ice sheet by an ice sheet mask. Bare ice is defined as the top layer having zero snow fraction.

L. 85: Atmospheric forcing for what? For DMIHH? Or for the newly developed emulator? Please clarify. In addition, please list all the properties included in the atmospheric forcing.

A: Here we refer again to the atmospheric forcing for the DMIHH model, listed in lines 80/81.

L. 90: It is unclear what the "input data" are. Input data for DMIHH? Or input data for the emulator?

A: Same as above: Input data for the DMIHH model.

L. 97 "they can be problematic when training ML models.": Please explain the reason for this argument in more detail.

A: We will expand the explanation to clarify that extreme outliers can disproportionately influence gradient-based optimization, distort the loss function, and impair stable convergence during neural network training.

L. 107: Does the negative sensible heat flux mean that the heat flux directs from the ice sheet surface to the atmosphere? Or opposite? Please explain.

A: In HIRHAM5 the sign convetnion for heat fluxes is defined such that positive values indicate energy directed towards the surface, while negative values represent energy transfer from the surface to the atmosphere. Thus, the extreme sensible heat flux down to –400 W/m² corresponds to intense surface cooling, causing the surface temperature approaching 0 Kelvin.

L. 129: Why is the number 5000 selected here? A more detailed explanation is needed.

A: We will add a justification for selecting 5000 grid cells, explaining the trade-off between computational feasibility and maintaining sufficient spatial representation across melt zones, as well as noting that this number provided stable training performance in preliminary experiments.

L. 178 “we choose the hidden layers of the network to be 64-128-128-64-32-16-16”: Please explain the meanings of each number, in particular for non-specialists in NN.
L. 182, L. 184, and L. 185: Same as the comment on L. 178.

A: We will introduce neural networks in general by explaining their layered structure, where each layer is composed of multiple neurons. The architecture can be summarized by listing the number of neurons in each layer, such as “number of neurons in layer 1 – number of neurons in layer 2 – number of neurons in layer 3,” and so on.

L. 188: Please explain in more detail about “LeakyReLU activation function.”

A: We will also explain the functioning of nonlinear activation functions in neural networks in general, and LeakyReLU specifically. Although many activation functions are available, LeakyReLU is computationally efficient while still allowing a small, non-zero gradient for negative inputs. This contrasts with the standard ReLU activation function, which can completely block gradient flow for negative values.

L. 193 “the optimal number of days to be used in the short-term module”: What do the authors mean by “optimal”? Please explain in more detail.

A: In this setting “optimal” refers to the number of preceding days yielding the lowest validation error during hyperparameter tuning, and that performance gains became marginal beyond a certain number of days.

L. 215~216 “the total computational cost remains far lower than physical firn models”: Can the authors add quantitative information for this explanation? I think such information is useful for other emulator developers.

A: DMIHH needs about 2,5 hours per simulation year on 16 CPU cores. We will add this information to the revised manuscript.

Technical corrections

L. 89: It is better to add something like “within DMIHH” at the end of this sentence.

L. 111: It is better to add the mathematical symbol “x” after “heat flux values.”

Table 1 caption: Please add “Autoreg” after “the autoregressive element.”

Figure 3 caption: It is better to explain the numbers in Gt listed in each panel.

L. 319: Suggest adding “surface” before “atmospheric variables.”

A: We will include these technical corrections in the revised version.