https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026

Learning to melt: Emulating Greenland surface melt from a polar RCM with machine learning

Elke Schlager[1,2], Sebastian Scher[3], Ruth H. Mottram[2], and Peter L. Langen[1]

[1]Department of Environmental Science, iClimate, Aarhus University, Denmark
[2]National Centre for Climate Research (NCKF), Danish Meteorological Institute, Denmark
[3]Wegener Center for Climate and Global Change and Department of Geography and Regional Science, University of Graz, Austria

Correspondence: Elke Schlager (eschlager@envs.au.dk)

Abstract.

Predicting surface melt on the Greenland ice sheet is critical for understanding surface mass balance (SMB) and sensitivity to climate change. Polar regional climate models are the primary tools for simulating melt and projecting future SMB, but different models produce significantly different results. However, they are too computationally expensive to create the large ensembles needed to quantify this uncertainty. We develop a neural network based emulator that predicts daily surface melt from atmospheric variables, trained on output from the polar regional climate model HIRHAM5 and its firn model DMIHH forced by ERA-Interim reanalysis. The emulator uses a physics-informed design combining short-term weather patterns with long-term climate memory, capturing both immediate atmospheric forcing and accumulated firn characteristics. The emulator achieves mean absolute error below 0.23 mm w.e. per day across all six Greenland drainage basins, with the errors primarily attributable to spatial over-smoothing. Our work demonstrates that machine learning can successfully emulate firn model behavior from climate forcing alone with computational costs orders of magnitude lower than traditional simulations. Once retrained for specific climate forcings, the emulator thus enables extensive ensemble projections. Furthermore, the modular architecture can be readily adapted to emulate other SMB quantities such as runoff. This represents a crucial first step toward computationally efficient emulation of polar regional climate models and surrogate modeling of SMB components in Earth system modeling.

1 Introduction

The Greenland ice sheet is losing mass today, and it will continue to do so in the future. Runoff caused by surface melt is a key component of surface mass balance (SMB), with increased melt leading to a negative SMB and subsequently a decrease in mass. With ongoing atmospheric warming and the positive feedback of surface melt with albedo and the lowering of elevation, surface melt will increase further in the future (Meredith et al., 2019). While future projections agree that melt will increase, the predictions are inconsistent about the rate of this melt increase and associated SMB loss (Glaude et al., 2024).

Polar regional climate models (RCMs) combined with a firn model show the best agreement with observational data among different modeling approaches for predicting SMB (Fettweis et al., 2020). But they are also the most complex and computationally intensive approach, as they require running two numerical models: (a) the RCM to downscale atmospheric data from a forcing global climate model (GCM), and (b) the firn model to infer surface mass balance based on the surface energy balance and firn properties that evolve from local atmospheric conditions. Both models are based on physical processes, modeled with computationally expensive numerical schemes. Despite their common framework, polar RCMs such as HIRHAM (Mottram et al., 2017; Langen et al., 2017), MAR (Fettweis et al., 2013, 2017), and RACMO (Noël et al., 2018) differ considerably in their assumptions, physical process representations, parameterizations, and numerical schemes, producing notable discrepancies in their SMB estimates (Fettweis et al., 2020). These discrepancies increase even more in future scenarios, as the models exhibit different sensitivities to atmospheric warming, leading to a varying increase in melt water production, and in turn to a positive feedback amplifying the models' discrepancies even more (Glaude et al., 2024). The use of a diverse ensemble of simulations is therefore crucial to mitigate model-specific biases and to improve robustness and reliability of projections, as well as to evaluate the projections statistically (Glaude et al., 2024; Mankin et al., 2020). However, the high computational costs of polar RCMs limit the generation of large ensembles.

In recent years, Machine Learning (ML) has shown great prospects in being used for simulating various parts of the Earth system (Pan et al., 2025; de Burgh-Day and Leeuwenburg, 2023; Reichstein et al., 2019). A major advantage compared to numerical models is that ML models can produce predictions at a fraction of the cost of the respective physical models, allowing for the production of large ensembles (Tebaldi et al., 2025). More specifically, an emulator of a polar RCM could be used to extend existing simulations over longer time periods, to complement projections under various Shared Socio-economic Pathway (SSP) scenarios which were not covered by the original simulations, or to produce simulations under new climate forcings. Additionally, emulators can be used not only for creating more simulations, but also for rapid hypothesis testing and sensitivity analysis. Moreover, ML emulators can not only be used for standalone emulation of Earth system components, but also for surrogate modeling within numerical Earth system models.

But what might an emulator for a firn model look like? Veldhuijsen et al. (2025) trained an XGBoost model on data from the polar RCM RACMO to infer perennial firn aquifers in Antarctica. To account for the slow evolution of firn properties, the model predicts annual liquid water content from local atmospheric forcings at multiple temporal scales: they used annual values as well as temporally aggregated averages over 5, 10, and 30 years. Similarly, Vandecrux et al. (2024) trained a neural network (NN) on monthly 10 m depth temperature observations to reconstruct firn temperatures across the entire Greenland ice sheet based on ERA5 reanalysis data. They incorporated input data at a monthly, annual, 5-yearly, and 10-yearly resolution to represent the firn pack's capacity to respond to conditions at different time scales.

Taking a more integrated approach, Van Der Meer et al. (2023) used ML downscaling to infer monthly SMB over Antarctica directly from forcing GCM data, effectively emulating both the RCM and the firn model simultaneously. Their work builds on ML techniques for image super-resolution methods that reconstruct high-resolution images from low-resolution counterparts, which have been adapted for downscaling meteorological variables in various ways (Sun et al., 2024). While some of these downscaling approaches add a temporal dimension for highly heterogeneous variables (e.g., wind) or when high temporal resolution is required, the spatial dimension remains dominant. Yet Van Der Meer et al. (2023) did not include a temporal dimension but computed monthly SMB values based solely on the atmospheric conditions of the current month.

Surface melt at a daily resolution shows high temporal variability which an ML emulator needs to be able to account for. However, the amount of temporal context that ML emulators require remains unknown. While the models predicting annual liquid water content and monthly firn temperature discussed above incorporate up to 30 year averages of climate conditions to account for properties deep inside the firn pack, daily surface melt operates on shorter time scales. Yet surface temperature exhibits lag effects, suggesting that past conditions of at least the previous day are needed to determine whether the surface has reached the melting point.

While emulating the entire polar RCM with its firn model also requires the downscaling from GCM to RCM, it is advantageous to emulate the RCM and the firn model separately. Unlike the RCM, a firn emulator can be trained independently of location, which reduces model complexity while simultaneously increasing model generalization and robustness. Separating the emulators also helps disentangle atmospheric driven and firn pack related biases and uncertainties.

In this study, we develop a neural network (NN) for emulating Greenland ice sheet surface melt at daily resolution, based on model output from the polar RCM HIRHAM5 (forced by ERA-Interim). The NN is designed to be modular and physically informed, facilitating future extension to emulating runoff and other firn processes. We optimize the choice of atmospheric variables and their temporal ranges for predicting melt by optimizing the NN over various subsets of input features, to assess the importance of the temporal dimension. Our model can be re-trained on data for future scenarios or on other polar RCMs, to facilitate the creation of future projections, multi-model ensembles, and comparative studies on bias attribution. This work is the first step of emulating SMB processes from polar RCMs as a whole.

2 Materials and methods

2.1 Data

For the creation of our emulator we use daily output of the polar RCM HIRHAM5 with its firn model DMIHH, forced by ERA-Interim for the period 1980–2016 (Langen et al., 2017). DMIHH is a one dimensional model organized in 32 layers. The surface layer is updated by snowfall, rainfall, deposition/sublimation, and energy fluxes from radiation, turbulent fluxes, and heat fluxes from the layers below. The surface state is then determined via the energy budget, with surface temperature being bound above by 0°C, and any excess energy producing surface melt. Surface albedo is calculated internally based on snow depth and surface temperature: it falls as the surface warms toward the melting point and decreases further when snow is shallow (bare ice exposure).

Atmospheric forcing is provided at hourly resolution through temporal interpolation of 6-hourly HIRHAM5 output fields. The firn model runs offline, which means that atmospheric data applies forcing to the surface without feedback from the surface back to the atmosphere. Consequently, the latent and sensible heat fluxes are prescribed by the RCM without adjustment for actual surface characteristics. While downward short- and longwave radiation are also prescribed by the RCM, their upward fluxes are calculated based on the dynamically calculated albedo and surface temperature.

The SMB outputs are then postprocessed and, together with the input data, aggregated to daily values. This temporal aggregation potentially constrains melt emulation accuracy by smoothing sub-daily variability and short-lived extremes that control timing and peak rates of melt. However, daily resolution enables broader utility in future applications, aligning with the typical daily temporal resolution of both GCM output and current ML downscaling approaches.

Data cleaning

Although data from physical simulations are generally consistent, they may still contain extreme values and outliers caused by numerical instabilities, although these are very rare. While these individual outliers do not affect the overall assessment of modeled melt or other properties, they can be problematic when training ML models. In order to decide how to treat these extreme values and outliers we must consider their source, their impact, and relation to other variables.

Aggregating rainfall and snowfall to daily values during postprocessing, some negative rainfall values arise as numerical artifacts. We set these values to zero for consistency. Furthermore, rainfall and snowfall show some suspiciously high values of up to about 700 and 1000 mm w.e. per day, respectively. Although they are likely caused by numerical instabilities in high relief topography, we do not correct these values to preserve consistency between the precipitation and the firn model output data used as target. However, we transform both rainfall and snowfall data by applying the logarithm (after adding 1, ensuring the transformation is well-defined for 0 values), which compresses these very high values. While this transformation neither adds nor removes information, it improves the data's usability for ML model training by preventing small values from vanishing in numerical noise relative to large outliers.

In rare instances, numerical instabilities also produce events of surface temperature runaway, which lead to unrealistic surface temperatures approaching 0 Kelvin and associated sensible heat flux values as low as -400 $\mathrm{W\,m^{-2}}$. We correct these heat flux values to a lower bound of -140 $\mathrm{W\,m^{-2}}$, which was the lowest sensible heat flux observed in simulations unaffected by the runaway. The ranges for sensible and latent heat flux span several hundred watts per square meter, yet the majority of the data are concentrated in a very narrow range. To expand the narrow and compress the large range of heat flux values, we apply a symmetric logarithm transformation $symlog(x) = sgn(x) \cdot log(|x|/\lambda + 1)$ with $\lambda = 5$ for latent heat flux and $\lambda = 15$ for sensible heat flux, corresponding to approximately half of their respective inter-quartile ranges.

Data preparation for training

We split the data spanning from 1980 to 2016 into three separate periods: a training period (1990–2013), a validation period (2014), and a test period (2016). The first 10 years of data (1980–1990) are included indirectly in the training set as decadal averages for the long-term module. To prevent information leakage between training and final evaluation, we introduced a one year gap between the validation and the test period. Although the time windows for calculating the decadal mean conditions still overlap, this overlap is considered to have negligible impact on the validity of the tests because this study focuses on a relatively short period lacking significant trends. Therefore, the decadal means primarily capture location specific characteristics rather than temporal development.

In addition to high quality, we also need sufficient data quantity for training, or more precisely: sufficient quantity of data that is relevant for the task we aim to solve. HIRHAM5 operates at a spatial resolution of 5.5 km, resulting in 58391 grid cells for the Greenland ice sheet, and thus just as many samples for every single day in the data set. However, not all of this data is rich in information for solving our task, since surface melt is zero, or very close to zero, for large areas of the ice sheet and a substantial part of the year. We significantly reduce the portion of low-relevance data samples through strategic sub-sampling in time and space. The temporal sub-sampling reduces the number of no/low melt days by randomly sampling 100 days per year from a normal distribution centered around the 24th of July (as peak of the melt season) and a standard deviation of 60 days. The spatial sub-sampling, on the other hand, favors grid cells in high melt areas over dry areas by selecting 5000 grid-cells according to predefined zone specific probabilities (see Appendix A).

By sub-sampling 5000 grid cells and 100 days per year from the training period, we create a training set of 12 million samples, reducing training costs significantly compared to the full data set of 511 million samples. While only about 6% of the full training set show melt above 1 mm w.e. per day, the temporal and spatial sub-sampling increase this ratio to approximately 26%. This sub-sampling stabilizes the training process; without it, the network tends to converge to the trivial local minimum of constantly predicting zero melt. We monitor training progress on the likewise sub-sampled validation set, but conduct model comparisons on the entire validation period data set to inform design choices, and report final performance on the entire test period data set.

As input data, we select atmospheric variables that dominate the surface energy budget: near-surface (2m) air temperature, rainfall, snowfall, sensible heat flux, latent heat flux, the downwelling longwave, and the downwelling shortwave radiation. We denote these input variables at daily-aggregated resolution X_d. In addition, we include the cyclic features $C = \cos(\frac{2\cdot\pi}{365}DOY)$ and $S = \sin(\frac{2\cdot\pi}{365}DOY)$ (with DOY the day of year) to encode seasonality in our model. Lastly, long-term history is represented by the previous 10 year average near-surface air temperature and snowfall, denoted by X_l. The target variable is daily surface melt. Since absorbed shortwave radiation is highly sensitive to surface albedo, we also test model setups that incorporate albedo. As a last preprocessing step, all the data is standard scaled to zero mean and unit variance with respect to the sub-sampled training data.

2.2 Emulator design

Machine learning encompasses a variety of algorithms. There is no single best algorithm, rather, the choice depends on factors such as the type of the problem, the type and complexity of the data, as well as its quality and quantity available for training. Given the highly nonlinear characteristics of the data set and the large amount of available data we have chosen a NN for regressing the surface melt based on atmospheric variables.

Since we aim to emulate surface melt modeled by the HIRHAM5 firn model, we refer to that modeled surface melt as the 'true' melt. The output of our ML model is called the 'predicted' melt or 'prediction'. With $M(t)$ denoting the true melt, and $\hat{M}(t)$ the predicted melt for a specific day t, we formulate our problem as finding a NN f such that

$$M(t) \approx \hat{M}(t) = f(X_d(T), X_l(t), \hat{M}(t-1)), \tag{1}$$

for all days t, where $X_d(T)$ represents the daily input variables for $N+1$ days $T = \{t-N, ..., t\}$, $X_l(t)$ the long-term inputs leading up to day t, and $\hat{M}(t-1)$ the melt prediction of the previous day.

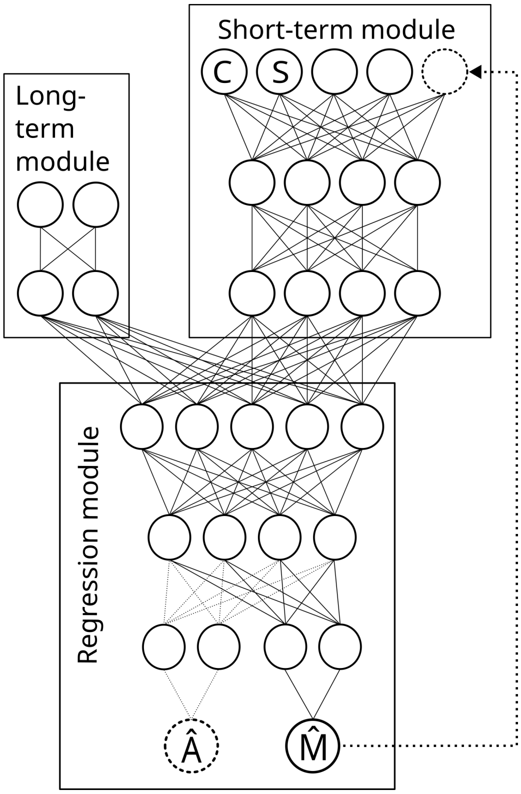


Figure 1. Schema of the modular neural network with long-term module, short-term module, regression module, the auxiliary target $\hat{A}$ and the autoregressive melt element (dashed elements).

We designed the neural network in two feature extraction modules and one regression module, as depicted in Fig. 1. The first extraction module, the short-term module, takes the daily inputs X_d of days $t-N,...,t$ as inputs to determine the current forcing on the surface layer. Furthermore, it incorporates the seasonal encodings C and S for day t, intended to approximate the firn cold content through seasonal indicators.

The second module, the long-term module, uses the long-term inputs X_l. These long-term inputs are motivated by the spin-up procedure common to firn models and are meant to describe the prevailing firn characteristics at a site. We include these inputs alongside the seasonal encoding to provide location-specific information on firn cold content and bare-ice exposure risk—factors that affect surface albedo and, consequently, surface melt.

The outputs of the two modules are then concatenated and fed into the final regression module which outputs the melt prediction $\hat{M}(t)$ for that day. While our proposed model Modular NN consists of these three modules only, the network can be extended further by an autoregressive element, or by additional target variables. The autoregressive element (dashed arrow in Fig. 1) feeds the melt of the previous day back into the daily module of the network to include the self–enhancing effect of surface melt.

Table 1. Overview of network configurations ordered by complexity, with their respective use of number of previous days N in the short-term module, the long-term module, the autoregressive element, and albedo as auxiliary target variable. Our main model Modular NN is indicated in bold. While Modular NN does not use albedo, we also trained a version of Modular NN with albedo as input as an upper threshold for performance.

	N	long-term	autoreg	albedo
Regression NN	0	no	no	no
Short-term NN	9	no	no	no
Modular NN	9	yes	no	no
Autoreg NN	9	yes	yes	no
Albedo NN	9	yes	no	yes

Alternatively, we use albedo as additional target variable since simultaneously learning albedo might improve melt predictions. In this case, Eq. (1) holds true not only for melt M but simultaneously for albedo. While the weights of the NN f are shared for predicting melt and albedo throughout most of the network, the regression module branches before its final layer, with separate last hidden layers for the two output neurons predicting melt and albedo (indicated by the dashed neuron connections for albedo output in Fig. 1).

Table 1 gives an overview of the different network configurations tested in this work. The simplest configuration, **Regression NN**, consists only of the short-term module, with using climate conditions of only the current day t as input (i.e. N=0), representing a pure regression model without any short-term or long-term historical information. This results in 9 input features (7 climate variables + 2 seasonal encoding variables), and we choose the hidden layers of the network to be 64-128-128-64-32-16-16, terminating in a single output neuron for melt prediction.

Short-term NN is an extension of the Regression NN by including $N = 9$ past days of the input variables, resulting in 72 input features. Due to the larger amount of input features, we expand network capacity by defining the hidden layers to be 128-128-256-256-128-64-32-16-16. Extending this configuration further by the long-term module yields the **Modular NN**. The long-term module has two input neurons (for the 10 years average of temperature and snowfall), and we choose two hidden layers of 32 neurons each. The hidden layers of the Short-term NN are split up into 128-128-256 for the short-term module, and 256-128-64-32-16-16 for the regression module. **Autoreg NN** is based on the configuration of Modular NN too, but the melt of the previous day $t-1$ is used as additional input. In contrast, **Albedo NN** does not use an additional input, but uses albedo as additional target and thus terminates in two output neurons.

For all hidden layers the LeakyReLU activation function is used.

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026

2.3 Emulator training

To determine the necessary yet sufficient subset of input and target variables, we developed our network iteratively, by subsequently tuning the network configurations listed in Table 1.

First, Regression NN was tuned to serve as a simple baseline model. Then, Modular NN was tuned, with varying the number of the of preceding days $N \in [1, 10]$ to find the optimal number of days to be used in the short-term module, with $N = 9$ yielding the best performance on the validation set. To investigate the necessity of the long-term module, we then tune the network without it (Short-term NN). Next, we test whether incorporating an autoregressive step in the modular NN improves model performance. Autoreg NN is tuned using the true previous melt as input during training (called teacher-forcing), although we also experimented with autoregressive learning approaches using the previous prediction directly during training (for more details see Appendix B). Finally, we explore the effect of including surface albedo. By including daily albedo values as input variables in the Modular NN we first establish an upper bound on the information content available from albedo. However, since albedo is an output from the firn model and not available when making new predictions based on atmospheric data alone, we also train Albedo NN, where albedo is included as an auxiliary target alongside surface melt.

Each configuration is trained multiple times for 300 epochs, respectively, tuning the learning rate using the Python library Optuna (Akiba et al., 2019) for Bayesian optimization. We use Adam optimizer (Kingma and Ba, 2014), a batch size of 256, a learning rate decay factor of 0.9 every 50 epochs, and gradient clipping to a norm of 1 to stabilize training. For a more detailed description and the results of the tuning procedure, see Appendix B.

We performed the training on an NVIDIA GRID A100D-40C GPU with 16 vCPUs (128 GB RAM). Individual training runs required approximately 25–45 minutes, depending on the complexity of the configuration. While our network configurations are regarded small in a deep learning context, the large volume of data is the critical factor in training time, and the data loading process remains the bottleneck in our pipeline despite heavy optimizations: For efficient data loading, we saved the training data in zarr chunks by date, with each chunk containing the samples of all 5000 sub-sampled grid cells, to minimize number of chunks that need to be opened and loaded. Thus, during training we read batches of 256 chunks, which leads to an effective batch size of $256 \cdot 5000 = 1280000$ samples. The batch size of 256 is thereby limited by the GPU memory, and the data loading is distributed across multiple CPUs in parallel to achieve high data throughput. After the model is trained, generating one year of melt predictions from preprocessed input takes approximately one minute on CPU. Although preprocessing (computing 10-year averages, data cleaning, scaling, and reformatting to zarr files) requires up to two hours, the total computational cost remains far lower than physical firn models, which additionally require extensive spin-up periods.

2.4 Evaluation

For final evaluation of our models, we use the test set (year 2016), which has not been used during training or to inform any further modeling decisions. For each configuration, we select the model from the tuning procedure that performed best on the validation data set, and report its root mean square error (RMSE), mean absolute error (MAE), mean bias error (MBE), and the coefficient of determination (R^2). Alongside the performance of our different model configurations, we provide a

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026

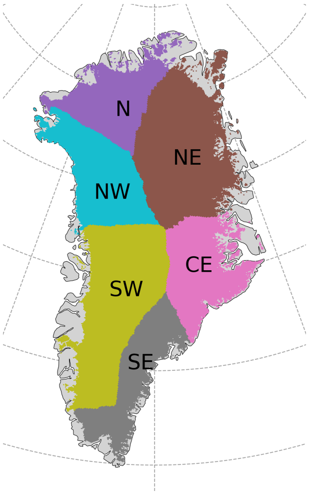


Figure 2. Map of basins used in the evaluation.

reference benchmark consisting of a running climatology of surface melt. This climatology is computed from the training set and smoothed with a 15 days moving average. Comparing the emulator skill to this benchmark, we can differentiate whether the emulator simply learned the climatology, or whether it also learned to predict the variation and anomalies relative to climatology, which is the actual purpose of the emulator. We therefore also report the R^2 on the anomalies with respect to this climatology (R^2_{anom}). The error tables are further complemented by visual evaluations to discuss error patterns qualitatively.

SMB outputs are often used in aggregated form per basin, which exhibit very different atmospheric forcings. To assess whether our location-agnostic model performs consistently across the ice sheet or exhibits basin-specific biases, we evaluate model performance separately for each basin, using the basins definitions shown in Fig. 2 (from Fettweis et al. (2020)). Because internal variability produces large inter-annual differences, an evaluation based on the test year alone can be misleading. Therefore, we perform the basin-wise assessment over the full period 1990–2016. Although this multi-year evaluation can yield overly optimistic performance scores compared with the single test year, since it includes the training set, it provides a more representative picture of the error distributions within and between basins. Along with RMSE, MAE, MBE, R^2, and R^2_{anom} we also report normalized RMSE and MAE (NRMSE and NMAE) to enable fair comparisons across basins. The normalized errors are computed by dividing the error totals by the basin mean annual melt—i.e., the average annual melt at each grid cell summed across the basin—and multiplying by 100 to report values as percentages.

3 Results and discussion

We start by analyzing overall mean performance over the entire ice sheet. The performance of the best models from the tuning process (Appendix B) on the test set are summarized in Table 2. All five configurations outperform the climatology benchmark, and performance increases with model complexity from a MAE of 0.40 mm w.e. per day for Regression NN, to

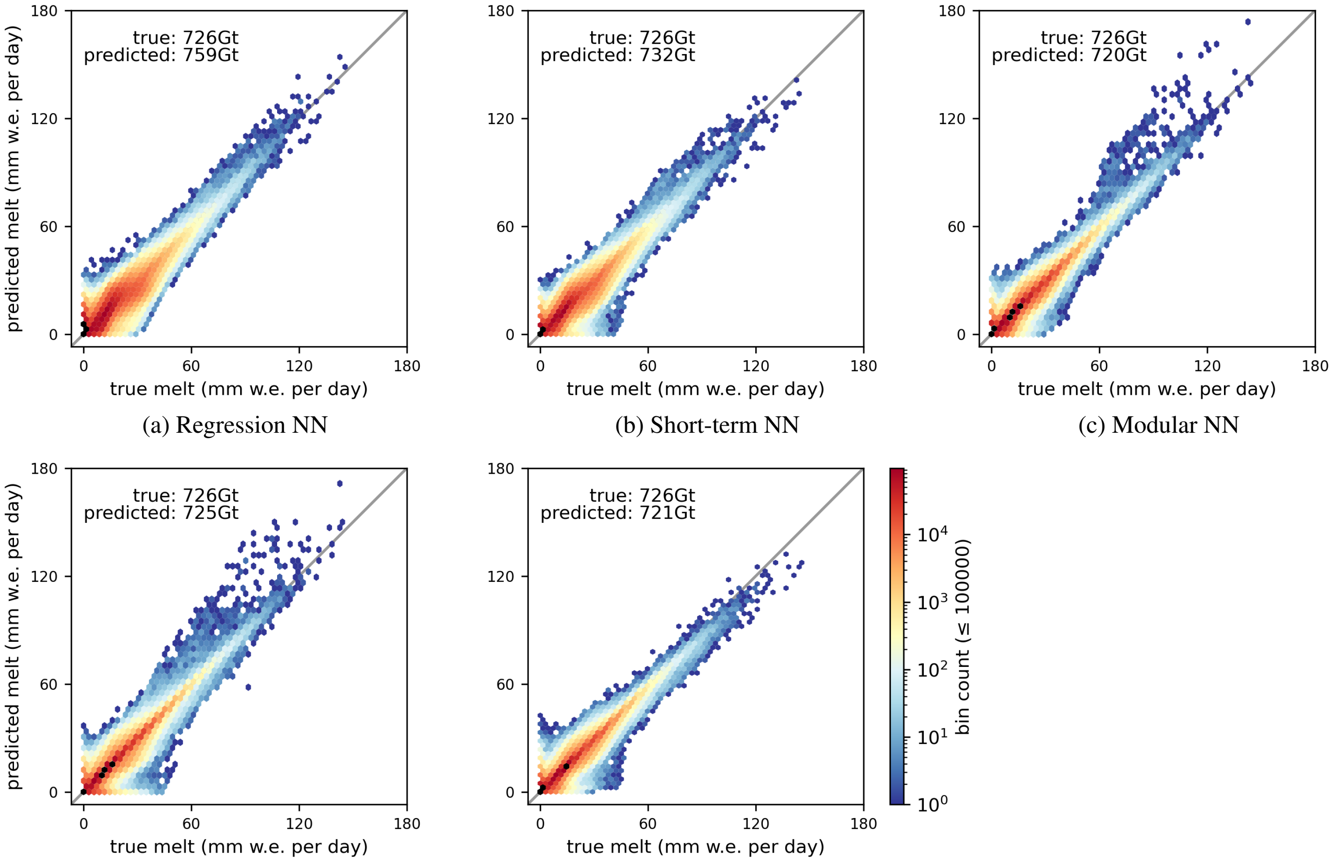
Table 2. Performance of final models on the test set. RMSE, MAE, and MBE in mm w.e. per day. The five models are ordered by increasing model complexity, with an ablation study on seasonality encoding for Modular NN. Autoreg (teacher) and Mod. w albedo use firn model output as input variables and are only listed for comparison.

	RMSE	MAE	MBE	R^2	R^2_{anom}
Climatology	2.30	0.56	-0.17	0.78	–
Regression NN	1.60	0.40	0.05	0.89	0.51
Short-term NN	1.23	0.26	0.01	0.94	0.71
Modular NN	0.90	0.18	-0.01	0.97	0.85
w/o seasonality	0.96	0.19	-0.02	0.96	0.83
Autoreg NN	0.90	0.15	-0.00	0.97	0.85
Albedo NN	0.90	0.17	-0.01	0.97	0.85
Autoreg (teacher)	0.40	0.08	0.01	0.99	0.97
Mod. w albedo	0.24	0.05	-0.00	1.00	0.99

0.26 mm w.e. per day for Short-term NN, 0.18 mm w.e. per day for Modular NN, 0.17 mm w.e. per day for Albedo NN, and 0.15 mm w.e. per day for Autoreg NN. The other metrics show improvement from Regression NN to Short-term and Modular NN, but then plateau. Thus, information about the past few days and long-term history is essential. Retraining of Modular NN without the seasonal encodings C and S yields a decrease in performance with RMSE 0.96 and MBE -0.02 mm w.e. per day, demonstrating the added value of including seasonality. The autoregressive element does not significantly improve performance further in inference mode (i.e., when using the previous day's prediction as input), although the knowledge of the previous melt proves valuable, as evidenced by evaluating Autoreg NN in teacher-forced mode. The additional experiment Mod. w albedo based on the Modular NN shows the importance of surface albedo for accurately deriving surface melt from atmospheric conditions, as including albedo as input results in excellent performance with almost perfect R^2 scores.

Point-wise evaluation

Across all models, the RMSE substantially exceeds MAE, indicating considerable under- and overestimation of melt across melt events of all magnitudes, as shown by the dark blue colored hexagonal bins in Fig. 3. The superior performance of Modular NN, Autoreg NN, and Albedo NN compared to Regression NN and Short-term NN stems primarily from improved predictions for a majority of data points for melt events up to 50 mm w.e. per day (narrower dark red band in Fig. 3 (c)–(e) compared to (a) and (b)). For Modular NN, 64% of the RMSE is attributable to absolute residuals up to 5 mm w.e. per day, a further 34% to absolute residuals between 5 and 15 mm w.e. per day, and only 2% of the RMSE to absolute residuals exceeding 15 mm w.e. per day.



(a) Regression NN

(b) Short-term NN

(c) Modular NN

(d) Autoreg NN (inference)

(e) Albedo NN

Figure 3. 2D hexagonal binning plots of true versus predicted surface melt of the test set of the five models. The logarithmic color bar is valid for bins containing up to 10^5 data points; bins containing more than 10^5 points are indicated in black for better visibility.

Modular NN and Autoreg NN show some notable overestimation pattern of melt events above 50 mm w.e. per day. This systematic deviation originates from an unusually early melt event in April in the SW basin, where severe overestimation occurs on only one single day (Fig. 4). While the model successfully predicts the unusually early surface melt for most days during this event, one particular day exhibits sensible heat flux values up to 460 $\mathrm{W\,m^{-2}}$. While such extreme sensible heat flux values also appear in the training set, the combination of an anomalously large heat flux with unusually early melt creates conditions that are effectively out-of-sample.

Qualitative assessment

But what causes these over- and underestimations? For this, we look at a typical day from the peak melt season in the test set in more detail. Figure 5 shows true and predicted melt, alongside residuals, for different models for a day in July. Modular

NN predicts spatially over-smoothed melt fields, evident in both the predicted field itself and the residual map (b). Autoreg NN in teacher-forced mode (d) shows substantial improvement in the spatial structure, with more pronounced contours in the predictions. However, the true previous melt used to make teacher-forced predictions is not available when applying the emulator to new climate data, where the model is run in inference mode, using its own prediction of the previous day. While the predicted field in inference mode (c) still exhibits sharper contours compared to Modular NN, the residual plot shows that there is uncertainty on where exactly these sharp contours should be.

Modular NN with daily albedo as an additional input shows that knowledge of surface albedo improves the spatial patterns significantly (e). This suggests that the autoregressive element's importance lies mainly in its role as a proxy for surface albedo. Yet, also the model with albedo as input cannot be used when making new predictions, since albedo is an output of the firn model. Therefore, we train the model configuration Albedo NN, which uses albedo as an auxiliary target instead to guide the network by learning albedo alongside surface melt. Unfortunately, the predicted albedo fields are also over-smoothed (f, g), and do not lead to noteworthy improvement (h).

These results reveal the fundamental challenge: the simultaneous over- and underestimations arise from the model's inability to accurately reconstruct the sharp spatial structures of the surface state. Without explicit knowledge of surface albedo, the model produces smoothed fields that systematically overestimate melt in some locations while at the same time underestimating it in others, creating the characteristic spatial error pattern observed in the residuals.

Basin-wise evaluation

In the remainder of this section we investigate the performance at the basin level, including data from the entire period 1990 to 2016. Since Modular NN, Autoreg NN, and Albedo NN show very similar performance on the test set, we continue the evaluation with Modular NN as it is the least complex model among these three. To enable performance comparison across the six basins, we present the average annual melt and performance scores for each basin and the whole ice sheet in Table 3. While these scores are based on the period 1990–2016 to reveal underlying patterns instead of scores dominated by interannual variability, Table C1 shows the corresponding scores for the test year only.

The model does not exhibit any severe basin-dependent bias, with MBE ranging from -0.01 to 0.01 mm w.e. per day for all basins except for SW basin, which shows a slightly higher bias of -0.02 mm w.e. per day. Correlation is also high across all basins, with the northern basin having slightly lower R^2 (0.95) than the southern basins (0.97-0.98). The R^2_{anom} computed on anomalies from the climatology is highest with 0.90 and 0.87 for the two southern basins (SE and SW), and lower for the other basins with an R^2 of approximately 0.80. This indicates that the emulator captures variability with respect to the climatology particularly well in the SE and SW basins, which have stronger, higher-signal anomalies that the model can learn, while the other basins show lower anomaly skill.

RMSE are highest for basins N and NE, followed by basin SE. As with the whole ice sheet, MAE is significantly lower than RMSE for all basins, indicating that a few large residuals persist across all regions. The MAE-to-RMSE ratio indicates that basin NE is affected most by a small number of high residuals, while basin SE is less dominated by such outliers, and more by small and medium errors. The normalized errors NRMSE and NMAE show error scores relative to the average annual melt in each basin. The north basins N and NE are worst in both absolute and normalized RMSE. In contrast, while basin SE shows similar absolute RMSE to basin NE, its NRMSE is only 0.53% of the basin's average annual melt compared to 0.62% for basin NE basin. The highest melt basin SW shows the lowest NRMSE of 0.43%, while basin NW has higher NRMSE (0.49%) but lower NMAE, indicating that basin NW is more affected by few high residuals than basin SW.

Table 3. Performance of Modular NN per basin and for the whole Greenland ice sheet (GrIS) over the whole period 1990–2016. Average annual melt in mm w.e. per year; RMSE, MAE, and MBE in mm w.e. per day; NRMSE and NMAE as percentage per day of the average annual total.

	mean melt	RMSE	MAE	MBE	R^2	R^2_{anom}	NRMSE (%)	NMAE (%)
N	122	1.03	0.18	-0.01	0.95	0.79	0.84	0.15
NE	137	0.85	0.13	-0.01	0.95	0.82	0.62	0.09
CE	139	0.78	0.14	0.01	0.95	0.79	0.56	0.10
SE	160	0.84	0.20	-0.01	0.98	0.90	0.53	0.12
SW	186	0.80	0.16	-0.02	0.98	0.87	0.43	0.09
NW	146	0.72	0.11	-0.01	0.97	0.80	0.49	0.08
GrIS	159	0.83	0.15	-0.01	0.97	0.85	0.53	0.09

Figure 6 shows the basin-wise integrated true and predicted melt for the test year alongside their climatologies, with residuals of predicted versus true melt for the test year (middle rows) and averaged over all years 1990–2016 (bottom rows). For the year 2016, basins NE, CE, and SE show nearly equal amounts of over- and underestimation, leading to total errors between -1 and 2 Gt over the entire year. Basins N and NW, on the other hand, show a tendency toward overestimation, resulting in 5 Gt and 4 Gt of excess melt for the year 2016, respectively (Fig. 6 N, NW middle rows). Compared to the mean annual over- and underestimations (Fig. 6 N, NW bottom rows), this indicates that the overestimation is specific to that year rather than a systematic model behavior in those basins. In contrast, severe underestimation dominates in basin SW, with the total year's overestimation amounting to 10 Gt versus underestimation totaling -23 Gt. While the average annual underestimation is significantly less extreme at -14 Gt per year, this still represents a slight negative bias for basin SW.

The timing of over- and underestimating melt is largely synchronous across basins. However, basins CE and SE show a tendency toward early-season underestimation and late-season overestimation. While the northern basins N and NE have a shorter melt season, the spatially aggregated daily residuals are larger than for basins CE, SE and NW.

4 Conclusions

We have developed a machine learning emulator that successfully predicts daily surface melt on the Greenland ice sheet from atmospheric variables alone. By training a neural network on 24 years of output from the polar regional climate model HIRHAM5 and its firn model DMIHH, we demonstrate that surface melt can be accurately emulated with a mean absolute error of 0.18 mm w.e. per day, significantly outperforming climatological benchmarks. Basin-level evaluation demonstrates that our location-agnostic approach generalizes well across the diverse climatic regimes of Greenland. The emulator maintains high correlation (R^2 = 0.95–0.98) across all six major basins with minimal systematic bias.

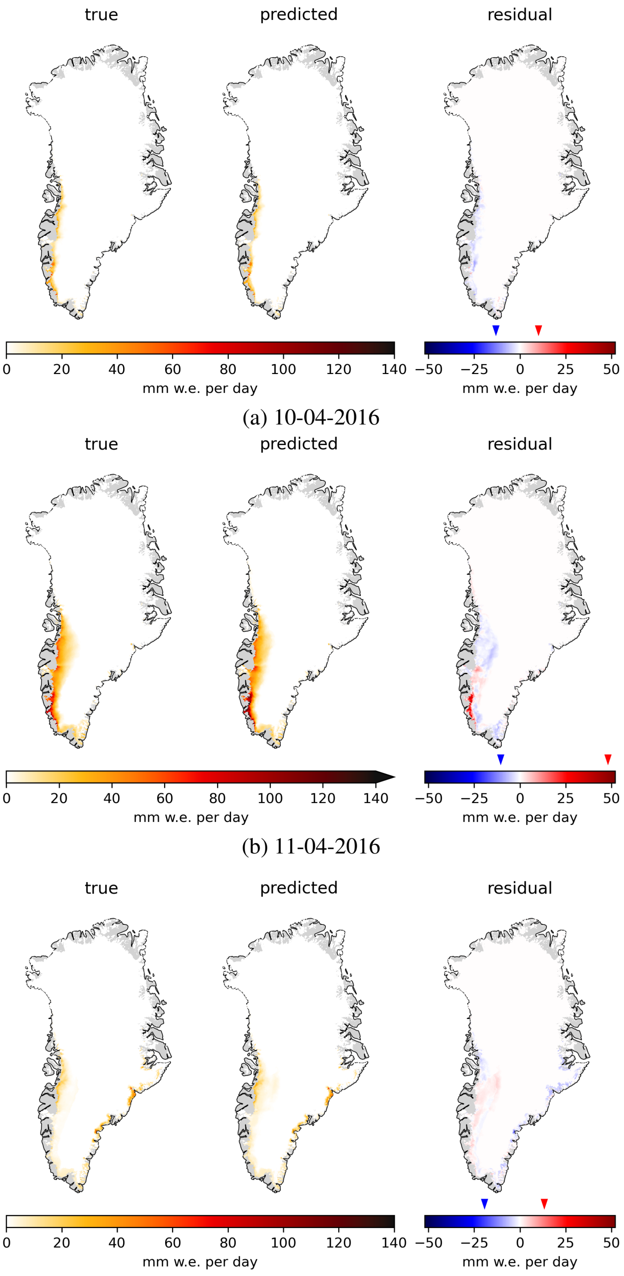
Our iterative model development reveals several key insights about the role of temporal information. Including atmospheric conditions from the previous nine days substantially improves performance over using only current-day conditions, demonstrating that temporal context matters. Furthermore, long-term climate memory in the form of decadal averages of temperature and snowfall improve model performance by providing crucial information about location-specific firn characteristics that affect the surface energy balance. Thus, the model profits from short- and long-term memory from these past conditions.

However, predicted melt fields tend to be spatially over-smoothed compared to the firn model output, lacking sharp transitions between regions of different melt extent. We addressed this issue through two different approaches: the autoregressive approach (Autoreg NN), and a multi-target approach (Albedo NN). While autoregressive models that incorporate previous melt show improved spatial structure in teacher-forced mode—where they use the true melt of the last timestep as input—they struggle in inference mode, when true previous melt is unavailable. The multi-target approach with albedo as an auxiliary target also does not resolve this issue, as predicted albedo fields remain over-smoothed too. This suggests that capturing sharp spatial gradients in surface conditions remains a fundamental challenge for data-driven approaches using atmospheric input data only.

Future work includes extending the domain of applicability: this model is developed on HIRHAM5 reanalysis data and trained to emulate DMIHH firn model behavior. To apply the emulator to climate data under different forcings, from different time periods, or entirely different polar RCMs, retraining is necessary, since extrapolation beyond the training distribution can yield unreliable results in data-driven approaches. Furthermore, the emulator can be extended to predict additional firn model outputs, such as runoff, to create a more comprehensive tool for Greenland SMB estimation.

In conclusion, this work demonstrates that machine learning can successfully emulate firn model behavior with spatially and temporally consistent accuracy and computational efficiency, while also revealing fundamental challenges in capturing sharp spatial patterns driven by surface characteristics. This emulator, when coupled with downscaling emulators that bridge the gap between global climate models and regional applications, enables the large ensemble projections needed to quantify uncertainty both within individual RCMs and across the divergent projections from different polar RCMs. Furthermore, such a firn emulator can be used as a surrogate model for SMB processes in Earth system models, enabling interactive ice sheet-climate coupling at scales previously computationally infeasible.

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026



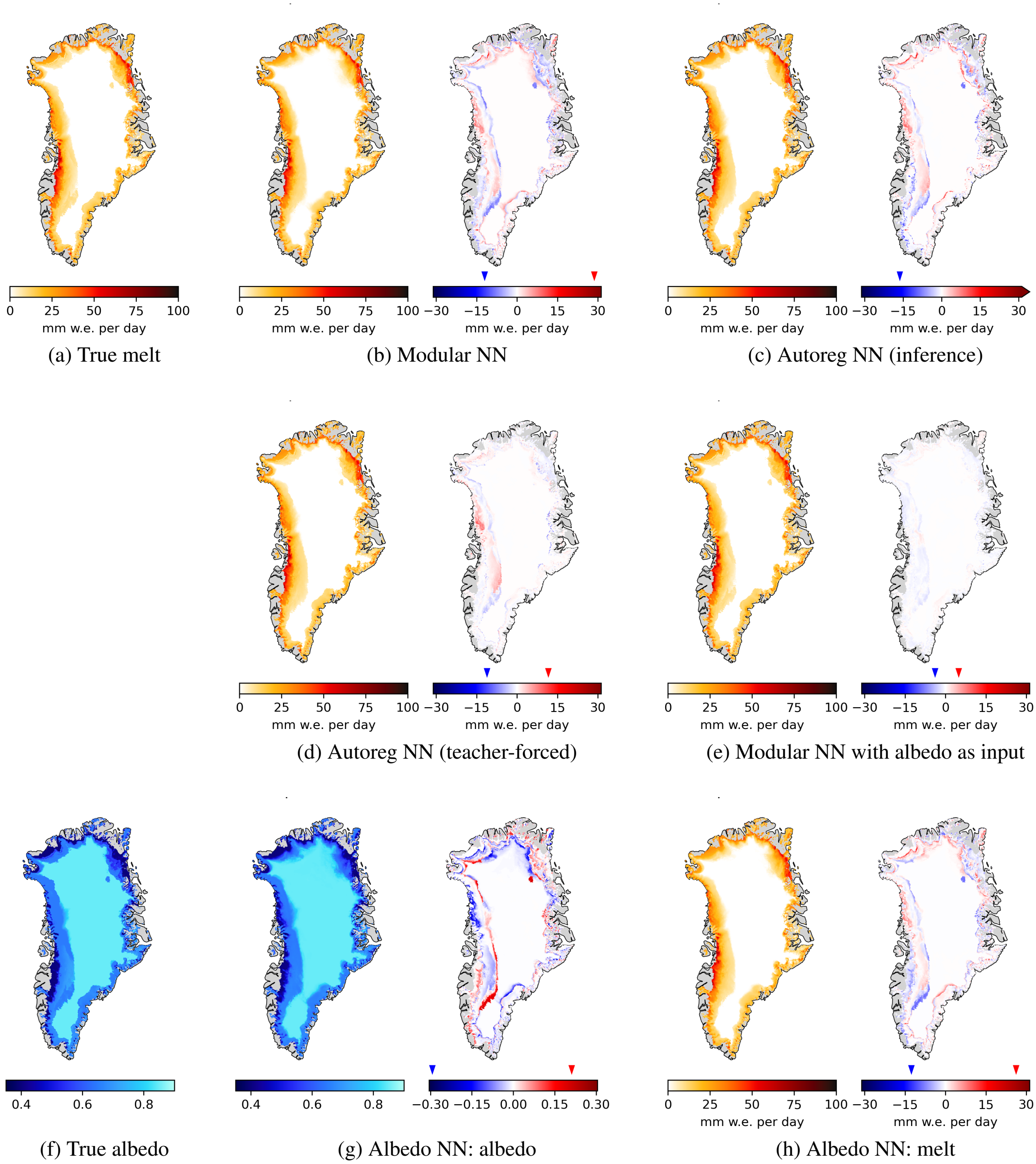
(a) 10-04-2016

(b) 11-04-2016

(c) 12-04-2016

Figure 4. True melt, and predicted melt by Modular NN with associated residual for three consecutive days in April 2016, with (b) causing the positive residual outliers shown in Fig. 3(c). The blue and red triangles on the residual color bar indicate the highest negative and positive residual of that day, respectively.

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026



(a) True melt

(b) Modular NN

(c) Autoreg NN (inference)

(d) Autoreg NN (teacher-forced)

(e) Modular NN with albedo as input

(f) True albedo

(g) Albedo NN: albedo

(h) Albedo NN: melt

Figure 5. Surface melt for 21st of July 2016. (a) true melt, (b) and (c) predicted melt (left panels) and residuals (right panels) of Modular NN and Autoreg NN (in inference mode). (d) and (e) show the predictions and residuals of Autoreg NN in teacher-forced mode, and of Modular NN with albedo as additional input; both these models cannot be used to produce predictions from climate forcing only, as they use firn model outputs as inputs. (f) shows the true albedo, (g) the albedo prediction and its residual, and (h) the melt prediction and its residual of the multi-target model Albedo NN.

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026

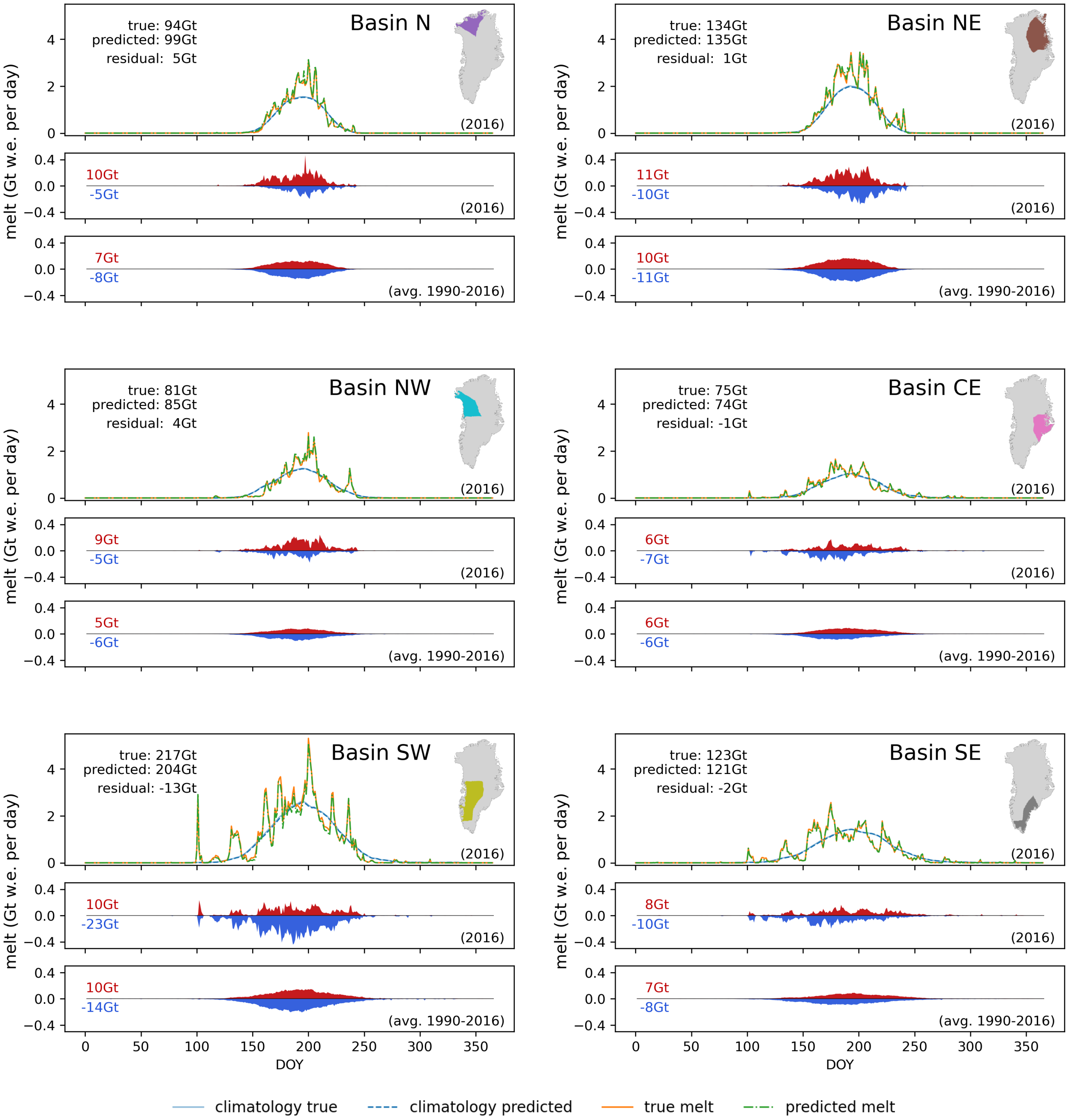


Figure 6. Temporal distribution of over- and underestimations of Modular NN per basin. The respective upper subplots show the basin-wise total true and predicted surface melt for the test year together with the true and predicted melt climatologies (1990–2013). The middle subplots show the total amount of overestimated (red) and underestimated (blue) melt per basins for the test year. The lower subplots show the average annual overestimate and underestimated for the whole time period 1990–2016.

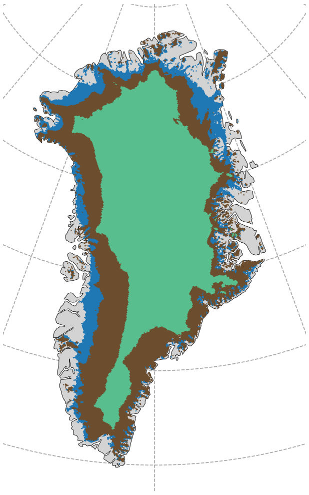


Figure A1. Ablation (blue), percolation (brown), and dry-snow (green) zone used for spatial sub-sampling.

Code and data availability. Code is available at the GitHub repository https://github.com/eschlager/MeltEmulation under the MIT License (the repository will be archived at Zenodo upon acceptance). The data produced in this study is available at https://doi.org/10.5281/zenodo.17913228 (Schlager, 2026). The HIRHAM5 simulation data is freely available upon request (Langen et al., 2017).

Appendix A: Spatial Sub-sampling

The spatial sub-sampling of 5000 grid-cells is performed according to predefined zone specific probabilities. The probabilities are chosen to be of 65% for the ablation zone, 30% for the percolation zone, and 5% for the dry-snow zone. Here, the zones are defined based on the SMB from the 10 year time range 1990–1999, with the ablation zone having negative SMB, the dry-snow zone showing melt below 100 mm w.e. per year for each year, and the percolation zone covering the remaining grid cells of positive SMB but non-negligible melt (Fig. A1).

Appendix B: Network tuning

Hyperparameters in ML algorithms are settings that control the algorithm's behavior, but are not adapted by the algorithm itself. This includes the choices of the architecture itself, its capacity, the activation function, regularization techniques, initialization, optimization algorithms and their specific setting, and more. Since it is unfeasible to tune all these hyperparameters, well-informed choices must be made to prioritize the most impactful parameters. The Modular NN architecture was defined based on physical principles, and we explore different architectural configurations to identify the most suitable design. The network capacity (number of layers and neurons per layer) was chosen to be sufficiently large, as evidenced by overfitting observed during training. To prevent using an overfitted model, we select the model weights that yield the lowest validation loss, which effectively corresponds to regularization via early stopping. The batch size is fixed at the maximum value permitted by available computing resources. Given these design choices, we primarily focus on tuning the learning rate for each configuration, as it is the most critical factor for training convergence and optimization performance (Goodfellow et al., 2016).

Table B1 presents the overview of the NN configurations, the tuned hyperparameter ranges, alongside RMSE, MAE, and MBE of the validation set of the best performing model for each configuration across all tuning trials. We define the best performing model by the sum of the relative MAE and MBE w.r.t. the seasonal anomaly errors. Fig. B1 shows the performance of all the trials, with the more complex models having consistently better performance than the less complex models. The results of Albedo NN are not plotted, since they strongly coincide with the performance of Modular NN.

For each of the five configurations, we tuned the learning rate. For Modular NN, we additionally tuned the number N of preceding days used as input, to determine the optimal number of input days, which was then fixed when tuning the subsequent configurations. As Albedo NN has a composite loss function consisting of both melt and albedo terms, the weighting between those two components is crucial. Therefore, while tuning the learning rate we simultaneously varied the weighting factors of melt MSE and albedo MSE for calculating the total loss, testing the following melt:albedo weight–ratios: 1:1, 7:3, 9:1, and 3:7. Further, we took a second attempt using trainable weights as proposed in Cipolla et al. (2018).

We trained the configurations one after another, to learn from the results for the next configuration. While performance generally improves with more preceding days, gains become marginal beyond 8 days, with 9 days achieving the best score. Therefore, the subsequently trained configurations Short-term NN, Autoreg NN, and Albedo NN use 10 input days (i.e., $N = 9$ preceding days), with the learning rate being tuned for 33 trials. We also narrowed the range of possible learning rate values when progressing through the different configurations as we gained insight in which ranges make sense.

When training the autoregressive NN under teacher-forced mode, we use the true melt with random noise, i.e., $M(t-1)+\varepsilon$ with $\varepsilon \in \mathcal{N}(0, 0.1)$ to get more robustness and not rely too much on the previous melt input. The results of Autoreg NN (teacher) show that the model benefits significantly from knowing the previous day's surface melt. However, this advantage diminishes when the model is evaluated in inference mode, i.e., when using the previous prediction instead of the true melt. We alternatively tested different strategies of training autoregressively with using the previous predicted melt $\hat{M}(t-1)$ during training, using different ratios of teacher-forced versus true melt, and different lengths of rollout windows. Although RMSE slightly improved, MAE and MBE did not decrease, and the same error patterns observed for the modular NN (which are discussed in the qualitative assessment in section 3) remained. Furthermore, the autoregressive training requires much higher computing resources, since a prediction for a specific day requires to make the prediction for the previous day(s), which also required a decreased batch size during training.

As baseline for information gain from the variable albedo, we retrain Modular NN including albedo in the set of daily input variables. This leads to a RMSE of 0.22, MAE of 0.05, and MBE of 0.25e-2 mm w.e. per day.

Appendix C: Basin-wise evaluation

Supplementary to Table 3, which presents the performance scores for across the basins for 1990–2016, Table C1 summarizes the scores for the test year 2016 only. Comparing the mean melt values of 2016 with the average annual melt amount of all 27

Table B1. Network Tuning: Overview of the five network configurations with their use of the separate modules, the tuning parameter ranges, and their validation scores in mm w.e. per day. The performance of Autoreg NN is stated in teacher-forced and in inference mode.

	Optuna study	RMSE	MAE	MBE
Climatology		2.22	0.53	-5.29e-2
Regression NN	10 trials: $lr \in (10^{-4}, 10^{-1})$	1.48	0.34	5.27e-2
Modular NN	50 trials: $lr \in (10^{-4}, 10^{-1})$, nr days N$\in [1, 10]$	0.86	0.16	0.15e-2
Short-term NN	33 trials: $lr \in (10^{-3}, 10^{-1})$	1.13	0.22	-0.08e-2
Autoreg NN (teacher)	33 trials: $lr \in (10^{-3}, 10^{-1})$	0.34	0.06	-0.06e-2
inference		0.86	0.14	0.01e-2
Albedo NN	33+33 trials: $lr \in (10^{-3}, 10^{-2})$, loss weights (manual+trainable)	0.84	0.15	0.11e-2

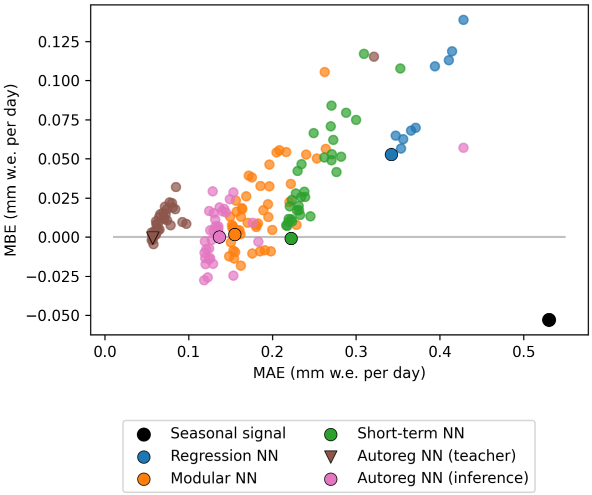


Figure B1. MAE and MBE on the validation set of the tuned networks Regression NN, Modular NN, Short-term NN, and Autoreg NN. The seasonal signal (black circle) is the melt climatology over the years 1990–2013, smoothed with a 15 days window. The best performing model for each configuration is indicated by the black outlined shapes. Note that Autreg NN (teacher) cannot be interpreted as applicable model, since it relies on the true melt and can only be used in its inference state.

years (Table 3), shows that 2016 is a strong melt year with melt exceeding the 27 year average in each basin. Except for MBE, scores for the 2016 test set are comparable to or better than the 27-year evaluation, indicating the basin-wise evaluation on the entire period in section 3 is not overly optimistic. The larger MBE magnitudes and the differing basin rankings across metrics

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026

underline the importance of the multi-decadal evaluation to avoid assessing model performance based on conditions specific to a single year.

Table C1. Performance of modular NN per basin for test year 2016. Average melt in mm w.e. per year; RMSE, MAE, and MBE in mm w.e. per day; NRMSE and NMAE as percentage of the annual total.

	mean melt	RMSE	MAE	MBE	R^2	R^2_{anom}	NRMSE (%)	NMAE (%)
N	143	0.97	0.18	0.06	0.96	0.79	0.68	0.13
NE	174	0.77	0.13	0.01	0.97	0.82	0.44	0.08
CE	159	0.82	0.15	-0.01	0.95	0.79	0.52	0.10
SE	177	0.97	0.23	-0.03	0.97	0.90	0.55	0.13
SW	221	1.02	0.22	-0.09	0.97	0.87	0.46	0.10
NW	159	0.83	0.14	0.04	0.96	0.80	0.52	0.09

Author contributions. ES, PL, and RM conceptualized the study and designed the methodology. ES did the code implementation, computational experiments, and visualization. ES performed the analysis and validation with consultation from SS. ES prepared the original draft of the manuscript, with help from SS, PL, and RM. RM provided the financial support and access to computing resources.

Competing interests. One of the coauthors is a member of the editorial board of *The Cryosphere*.

Acknowledgements. ES received support for this study from the National Centre for Climate Research (NCKF) and some additional support from the Novo Nordisk funded PRECISE project (NNF23OC0081251). We gratefully acknowledge the computing resources provided by the European Weather Cloud under special project dkmottr2. The original climate model simulations used in this project were carried out under the framework of the European Union's Horizon 2020 research project PROTECT under grant agreement 869304. Claude Sonnet 4.5 was used for grammar check and to improve readability.

https://doi.org/10.5194/egusphere-2026-7
Preprint. Discussion started: 27 January 2026

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623–2631, 2019.

Cipolla, R., Gal, Y., and Kendall, A.: Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7482–7491, ISBN 2575-7075, https://doi.org/10.1109/CVPR.2018.00781, 2018.

de Burgh-Day, C. O. and Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review, Geoscientific Model Development, 16, 6433–6477, https://doi.org/10.5194/gmd-16-6433-2023, 2023.

Fettweis, X., Franco, B., Tedesco, M., van Angelen, J. H., Lenaerts, J. T. M., van den Broeke, M. R., and Gallée, H.: Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR, The Cryosphere, 7, 469–489, https://doi.org/10.5194/tc-7-469-2013, publisher: Copernicus Publications, 2013.

Fettweis, X., Box, J. E., Agosta, C., Amory, C., Kittel, C., Lang, C., van As, D., Machguth, H., and Gallée, H.: Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model, The Cryosphere, 11, 1015–1033, https://doi.org/10.5194/tc-11-1015-2017, publisher: Copernicus Publications, 2017.

Fettweis, X., Hofer, S., Krebs-Kanzow, U., Amory, C., Aoki, T., Berends, C. J., Born, A., Box, J. E., Delhasse, A., Fujita, K., Gierz, P., Goelzer, H., Hanna, E., Hashimoto, A., Huybrechts, P., Kapsch, M.-L., King, M. D., Kittel, C., Lang, C., Langen, P. L., Lenaerts, J. T. M., Liston, G. E., Lohmann, G., Mernild, S. H., Mikolajewicz, U., Modali, K., Mottram, R. H., Niwano, M., Noël, B., Ryan, J. C., Smith, A., Streffing, J., Tedesco, M., van de Berg, W. J., van den Broeke, M., van de Wal, R. S. W., van Kampenhout, L., Wilton, D., Wouters, B., Ziemen, F., and Zolles, T.: GrSMBMIP: intercomparison of the modelled 1980–2012 surface mass balance over the Greenland Ice Sheet, The Cryosphere, 14, 3935–3958, https://doi.org/10.5194/tc-14-3935-2020, publisher: Copernicus Publications, 2020.

Glaude, Q., Noël, B., Olesen, M., Van den Broeke, M., van de Berg, W. J., Mottram, R., Hansen, N., Delhasse, A., Amory, C., and Kittel, C.: A factor two difference in 21st-century Greenland ice sheet surface mass balance projections from three regional climate models under a strong warming scenario (SSP5-8.5), Geophysical Research Letters, 51, e2024GL111 902, https://doi.org/10.1029/2024GL111902, iSBN: 0094-8276 Publisher: Wiley Online Library, 2024.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, www.deeplearningbook.org, 2016.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Langen, P. L., Fausto, R. S., Vandecrux, B., Mottram, R. H., and Box, J. E.: Liquid Water Flow and Retention on the Greenland Ice Sheet in the Regional Climate Model HIRHAM5: Local and Large-Scale Impacts, Frontiers in Earth Science, 4, https://www.frontiersin.org/articles/10.3389/feart.2016.00110, 2017.

Mankin, J. S., Lehner, F., Coats, S., and McKinnon, K. A.: The value of initial condition large ensembles to robust adaptation decision-making, Earth's Future, 8, https://doi.org/10.1029/2020EF001610, 2020.

Meredith, M., Sommerkorn, M., Cassotta, S., Derksen, C., Ekaykin, A., Hollowed, A., Kofinas, G., Mackintosh, A., Melbourne-Thomas, J., Muelbert, M., Ottersen, G., Pritchard, H., and Schuur, E.: Polar Regions. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate [H.-O. Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, N.M. Weyer (eds.)], Tech. rep., Cambridge University Press, Cambride, UK and New York, NY, USA, https://doi.org/10.1017/9781009157964.005, 2019.

Mottram, R., Boberg, F., Langen, P., Yang, S., Rodehacke, C., Christensen, J. H., and Madsen, M. S.: Surface mass balance of the Greenland ice sheet in the regional climate model HIRHAM5: Present state and future prospects, , 75, 105–115, https://doi.org/10.14943/lowtemsci.75.105, publisher: 75 , 2017.

Noël, B., Van De Berg, W. J., Van Wessem, J. M., Van Meijgaard, E., Van As, D., Lenaerts, J., Lhermitte, S., Kuipers Munneke, P., Smeets, C. J. P., and Van Ulft, L. H.: Modelling the climate and surface mass balance of polar ice sheets using RACMO2–Part 1: Greenland (1958–2016), The Cryosphere, 12, 811–831, https://doi.org/10.5194/tc-12-811-2018, iSBN: 1994-0416 Publisher: Copernicus GmbH, 2018.

Pan, X., Chen, D., Pan, B., Huang, X., Yang, K., Piao, S., Zhou, T., Dai, Y., Chen, F., and Li, X.: Evolution and prospects of Earth system models: Challenges and opportunities, Earth-Science Reviews, 260, 104 986, https://doi.org/10.1016/j.earscirev.2024.104986, 2025.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Schlager, E.: Output of Learning to melt: Emulating Greenland surface melt from a polar RCM with machine learning, https://doi.org/10.5281/zenodo.17913228, 2026.

Sun, Y., Deng, K., Ren, K., Liu, J., Deng, C., and Jin, Y.: Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review, ISPRS Journal of Photogrammetry and Remote Sensing, 208, 14–38, https://doi.org/10.1016/j.isprsjprs.2023.12.011, 2024.

Tebaldi, C., Selin, N. E., Ferrari, R., and Flierl, G.: Emulators of climate model output, Annual Review of Environment and Resources, 50, https://doi.org/10.1146/annurev-environ-012125-085838, iSBN: 1543-5938 Publisher: Annual Reviews, 2025.

Van Der Meer, M., De Roda Husman, S., and Lhermitte, S.: Deep Learning Regional Climate Model Emulators: A Comparison of Two Downscaling Training Frameworks, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 593, https://doi.org/10.1029/2022MS003593, 2023.

Vandecrux, B., Fausto, R. S., Box, J. E., Covi, F., Hock, R., Rennermalm, K., Heilig, A., Abermann, J., van As, D., Bjerre, E., Fettweis, X., Smeets, P. C. J. P., Kuipers Munneke, P., van den Broeke, M. R., Brils, M., Langen, P. L., Mottram, R., and Ahlstrøm, A. P.: Recent warming trends of the Greenland ice sheet documented by historical firn and ice temperature observations and machine learning, The Cryosphere, 18, 609–631, https://doi.org/10.5194/tc-18-609-2024, publisher: Copernicus Publications, 2024.

Veldhuijsen, S. B. M., van de Berg, W. J., Kuipers Munneke, P., Hansen, N., Boberg, F., Kittel, C., Amory, C., and van den Broeke, M. R.: Emulating the expansion of Antarctic perennial firn aquifers in the 21st century, The Cryosphere, 19, 5157–5173, https://doi.org/10.5194/tc-19-5157-2025, 2025.