Thank you for this positive review of our work and for the careful proofreading. We answer each point below in blue text .

Yours sincerely,

Baptiste Vandecrux on behalf of the co-authors

General comments

1) When I first read the title of the manuscript, I thought the documented warming came directly from observations. Instead, the warming trends come from output of the ANN. I understand that documenting widespread warming from observations is challenging because observations from a single location do not exist over long time periods and there is a lack of spatial coverage across the ice sheet. However, I have two recommendations to address this issue. The first would be to slightly modify the title. Maybe something along the lines of "Historical firn and ice temperature observations inform modeled warming trends across the Greenland ice sheet" (something along those lines).

The title was updated to: *Recent warming trends of the Greenland ice sheet documented by historical firn and ice temperature observations and machine learning*

My second suggestion is to actually try to evaluate how well ANN, HIRHAM, RACMO, and MAR temperatures compare to observations over time.

We have done our best to do precisely this in section 3.2 "RCM evaluation and comparison with the ANN". The evaluation of the ANN done in section 3.1 is done through the evaluation of spatial cross-validation. We also added an evaluation of the trends calculated by the ANN against observations which was also a suggestion from the other reviewer.

Perhaps creating a temperature vs. time plot (maybe breaking it into panels by dry snow, percolation, and ablation zones) and plotting each temperature value with its corresponding date. You could then add the temperature simulated by the ANN, HIRHAM, RACMO, and MAR for that same location at the same time. In the paper you mention that all models diverge when estimating past history of sites when observations are not available. While it seems like the ANN is probably the most trustworthy model, making a plot this way might help to show that the ANN can model temperatures most accurately through time across the three regions of the ice sheet.

Our figure 5 plots temperature, observed and modeled by the ANN and RCMs, versus time, at sites representative of different climate zones. It also shows the spread between models and how the ANN matches best with observations. But maybe the suggested plot was meant to display all observations from a given zone along with the model prediction (like adding observations to Figure 6). In that case we are afraid that the seasonal variation of temperature and their spread within their respective zones would make the figure hard to read. The cloud of observations from a given zone would be uncorrected from observations' sampling bias: in the percolation areas, the observations from western Greenland outweigh those from eastern Greenland and such a cloud would misrepresent the true average temperature for that climate zone.

It could also allow you to compare observational temperature trends with model trends by performing regressions on the temp vs. time plot. Not sure – this is just my initial thought here, but I'm happy to hear other suggestions if the authors have them.

Prompted by another comment from the second reviewer, we added in Section 3.1 a more systematic evaluation of the T10m trends estimated by our ANN at sites where enough observations are available:

To evaluate the capacity of the ANN to capture the recent evolution of T_{10m}, we select 10 sites where more than 60 monthly values are available between 1998 and 2022 and compare the trends calculated from the ANN and the observations over the periods 1998-2010 and 1998-2022 (Table 2). These periods were chosen because of a general lack of measurements between 2011 and 2020. Trends calculated from the ANN only consider the months where observations are available. We note that due to the missing months, these trends are not reliable for general inference on the true T_{10m} evolution: depending on which months are missing it might overestimate or underestimate the true T_{10m} trend for these periods. The median T_{10m} trends for 1998-2010 are 0.9 and 0.8 °C decade^{-1} for the ANN and for the observations respectively (Table 2). For the period 1998-2022, the median T_{10m} trends for 1998-2010 are 0.4 and 0.6 °C decade^{-1} for the ANN and for the observations respectively (Table 2). The ANN therefore slightly overestimates the T_{10m} trend during 1998-2010 and underestimates it during 1998-2022. We conclude that the ANN reproduces the magnitude of the T_{10m} increase seen in observations although this aptitude varies with the location and the time period considered. From this assessment and because the ANN does not suffer temporal nor spatial gaps, the ANN appears as a suitable tool to study the trends in T_{10m} over the entire Greenland ice sheet.

Table 2: Trends in 10 m subsurface temperature (T_{10m}) calculated from the ANN and observations (obs.) at 10 sites for the periods 1998-2010 and 1998-2022. ANN trends are calculated only from the months where observations are also available. The difference between the two calculated trends as well as the number of monthly values used for the calculation (N) are also given for each site.

	Trends in T10m (°C decade-1)							
	1998-2010				**1998-2022**			
Site	**ANN**	**obs.**	**ANN - obs.**	*N*	**ANN**	**obs.**	**ANN - obs.**	*N*
NASA-SE	*1.0*	*0.7*	*0.3*	*115*	*0.4*	*0.5*	*-0.1*	*171*
NASA-E	*0.5*	*0.5*	*0.1*	*140*	*0.6*	*0.5*	*0.0*	*270*
Summit	*0.4*	*1.0*	*-0.6*	*133*	*0.3*	*0.6*	*-0.3*	*172*
Tunu-N	*0.7*	*0.3*	*0.4*	*140*	*0.6*	*0.5*	*0.0*	*150*
South Dome	*1.4*	*0.8*	*0.5*	*97*	*0.2*	*0.5*	*-0.2*	*116*
Saddle	*1.4*	*0.7*	*0.7*	*125*	*0.2*	*0.6*	*-0.4*	*156*
Humboldt	*0.5*	*1.0*	*-0.4*	*66*	*0.4*	*0.3*	*0.1*	*71*
Crawford Point 1	*1.3*	*3.0*	*-1.7*	*63*	*0.4*	*0.7*	*-0.3*	*120*
DYE-2	*1.2*	*0.8*	*0.4*	*139*	*0.3*	*1.1*	*-0.7*	*220*
Swiss Camp	*0.7*	*0.8*	*0.0*	*83*	*0.3*	*1.8*	*-1.5*	*172*

2) One interesting aspect of this paper is that it helps to show where future field efforts should be concentrated to collect subsurface temperature measurements. I'm struck by the fact that air temperature, snow accumulation, and temperature amplitude at observation sites are generally pretty normally-distributed compared to the distribution of these parameters across the entire ice sheet (Figure 2). Field researchers need to target locations that will skew the

overall temperature and accumulation distributions. In general, sites that have lower air temperatures and snow accumulation need to be targeted for additional measurements. However, the uncertainty of the ANN is actually highest in the southeast portion of the ice sheet, which to my understanding is characterized by higher air temperatures and accumulation rates. Overall, we have decent coverage of these conditions at observation sites (Figure 2). However, we have few measurements in the specific firn hydrologic regime that produces firn aquifers. Additionally, ANN uncertainty is elevated in northern Greenland (particularly Northwest – Figure 3c). This is where we have radar observations of thick ice slabs forming (a different firn hydrology regime from firn aquifers). My question is: is it more important to collect subsurface temperature observations in locations that improve representativity of accumulation/air temperature conditions or should our focus be on improving representativity of observations from different firn hydrology regimes? If you create a map of accumulation/temperature conditions where our observations are deficient, how do they compare with where ANN uncertainty is highest?

Thank you for this interesting reflection. We agree that Figure 2 shows that the cold and dry, high elevation areas are underrepresented in our dataset and that simultaneously Figure 3 shows highest uncertainty in the southeast. The low uncertainty (Figure 1c) and good performance of the ANN (Figure 1ab) in the cold regions show that the available observations apparently are enough to simulate subsurface temperature there. So we do not think that additional measurements are needed there, especially if continuous monitoring sites like GC-Net weather stations are maintained. However, although there are measurements in the firn aquifer region, they are still few and not covering the entire southeast. The high uncertainty in Figure 3c also reflects that the ANN depends on few observations there. Where we "should focus" or where "observations are deficient" depends on our objective. If we aim at documenting the Greenland ice sheet as a whole, then the extensive high elevation plateau is of high importance. If we aim at understanding processes or at reducing the uncertainty in less representative regions like the firn aquifer region, then that is where we need to collect more measurements. We added to our discussion:

More observations are needed from these less-visited parts of the ice sheet to further train the ANN. These new measurements could either focus on the coldest parts of the ice sheet, where our compilation currently lacks measurements (Figure 2a) or on the areas where our uncertainty is the highest, in the Southeast (Figure 3c).

and to our conclusion:

Our evaluation shows highest ANN uncertainty in the southeast and in the lower percolation area in northern Greenland (Figure 3). Those are regions where few observations are available (Figure 1) and consequently, any additional measurements there will help to constrain models and understand the relevant processes.

Specific comments:

1) This comment may just be a lack of my detailed understanding of ANNs. You state on line 132- 133 that "ANNs have proven their ability to capture these non-linearities and interactions between input parameters in numerous glaciological and meteorological applications". However, on lines 213-214 you state that you use a rectified linear unit activation function. Isn't it the activation function that makes ANNs either linear or nonlinear? So although you are attempting to capture nonlinearities, are you not just using a linear ANN? Again, this question is probably due to me not fully understanding the structure of ANNs, but maybe you could add a sentence that describes how the ANN is nonlinear.

We see that it can be counter intuitive that the ANN is a non-linear model made up of many piecewise linear functions. First, it needs to be said that the activation function we use is a piecewise linear function that gives 0 if the input is below 0, f(x)=x if input x is above 0, and is not linear strictly speaking. But more importantly, it is the interactions between the neurons that makes ANNs able to reproduce highly nonlinear functions. Taking for example an input x and a target function f=exp(x), for an ANN of one layer and three neurons, f is approximated by a combination of three piecewise functions, meaning a line with three breaks and four segments. If we add another layer to the ANN, each neuron on the second layer can make its own weighted sum of the output of the three previous neurons, apply an offset (called bias) and apply their activation function to the result. These weights and biases are as many degrees of freedom that allow to tune the network to fit even non-linear functions.

2) I think that it is important to mention that in Figures 3 and 5 that the observations should generally be well-reproduced by the ANN, because the model has been trained explicitly "seeing" these data, which is probably why, especially in Figure 5, the ANN appears to outperform other models over periods where observational data exist.

When reaching Figure 5, the reader is aware that the ANN is trained on the observations. This is one of the strengths of the ANN, and we describe at length in Section 2.2 and 2.3 how the ANN and RCM are of different nature. In our result section, we did our best to be nuanced in our analysis and avoid one-to-one comparison of the ANN with the RCM for exactly that reason. That is why for instance, the scatter plots of the ANN (Figure 3) are not presented along with similar plots from the RCM (Figure 4). We therefore prefer to avoid further repetition of how the ANN is built while describing its performance. Last but not least, the presented RCMs are regularly evaluated and updated to match the available observations (see f.e. Langen et al., 2017, Vandecrux et al., 2022, Brils et al., 2022) and therefore also "see" the observations, just in a different and less direct way than the ANN.

Technical corrections/comments:

These are some typos that I have caught or comments that are meant to improve clarity of the text.

L1: Should "firn" replace "snow" in the title, since by 10 m you are observing firn temperatures rather than snow?

We updated the title to "*Recent warming trends of the Greenland ice sheet documented by historical firn and ice temperature observations and machine learning"*

L22: Specify what about "melting". Melt production? Melt rates? Melt extent, duration?

We added *in intensity and extent*

L42: change "undergoing" to "experiencing" or "exposed to"?

We changed to *exposed to.* Thanks.

L51: Perhaps add "and grain coarsening" after "liquid water within snow"? You can cite Nolin and Stroeve (1997) for this.

Added, thanks.

L53: Is Humphrey et al. (2012) not an appropriate reference here?

Added, thanks.

L57: note that the "retaining meltwater" is retention through refreezing

We updated to *refreezing and retaining meltwater*.

L58: Maybe caveat this statement – the degree to which cryo-hydrologic warming will increase dynamic mass loss depends on the glaciological setting (e.g., Poinar et al. 2017).

Thank you for pointing out this study. We rephrased to:

Subsurface warming could also [...]and increase the ice viscosity (Phillips et al., 2010, 2013, Colgan et al., 2015) although with limited impact on dynamic mass loss (Poinar et al., 2017).

L63: Just noting the inconsistency. You state 4500 in the abstract and here you say 4600.

Thank you for spotting this. These numbers were actually both inaccurate and we updated them to *4612* for the number of observations in our compilation and *4597* the number of observations being used to train our ANN. We

added details about these two numbers in Section 2.1:
Among these 4612 T_{10m} observations, 15 measurements are either outside of the current ice sheet extent as defined by the GIMP ice sheet delineation (Howat et al., 2014) or outside of the 1950-2022 period we consider for our T_{10m} reconstruction. There are therefore 4597 T_{10m} observations in our compilation that can be used for the reconstruction of T_{10m} on the ice sheet between 1950 and 2022.

L65: I know you state this in the methods section, but it may be worth just adding a sentence as to why you choose an ANN over other machine learning models in the introduction.

We would like to keep the introduction focused on the motivation of the study and keep the technical details for the Methods section. Adding a sentence about why we choose it over other methods would require that we define, also in the introduction, what the ANN is, how it is being used and eventually why we prefer it over other techniques (Section 2.2). For the sake of concision, we keep model description to the methods.

L71-72: I'm not entirely sure what you mean by "We put special emphasis on T10m magnitude and trends in various areas of the ice sheet". Maybe something like: "We evaluate the differences between observed and modeled T10m across the entire ice sheet and specifically in the ablation area, and we evaluate modeled temperature trends in bare ice regions, the percolation zone, and the dry snow zone of the ice sheet."

Thank you for the suggestion. We updated to:

Using our observational dataset of subsurface temperature as well as our ANN, we evaluate three regional climate models (RCMs) widely used to estimate the surface mass balance of the Greenland ice sheet [...]. We then evaluate the ANN and RCMs' T10m magnitudes and trends in the bare ice, percolation, and dry snow areas of the ice sheet. Lastly, we discuss the impact of this subsurface warming on the ice sheet mass balance processes.

L78: change "thermistor strings," to "thermistor strings:"

Updated, thanks.

L81: change "was operating" to "operated"

Updated, thanks.

L81: change "and equipped" to "and was equipped"

Updated, thanks.

L83: change "and active" to "and was active"

Updated, thanks.

L84: change "along" to "over"

Updated, thanks.

L85: delete "yet"

Updated, thanks.

L88: add the sensor spacing here.

Updated, thanks.

L91: change "was" to "were"

Updated, thanks.

L95-96: Maybe clearer to write "Using firn temperature observations reported by Samimi et al. (2012) and Heilig et al. (2018) at Dye-2 as a reference..."

Updated, thanks.

L105: change “was” to “were”

Updated, thanks.

L106-107: Why can Benson not be used – too shallow? Also, why do you include observations that were not collected on the Greenland ice sheet?

Thank you for pointing this out. Those measurements fell outside of the GIMP ice mask (Howat et al., 2014) so we meant that we do not use them for the T10m reconstruction. They are still in the compilation of T10m measurements. We moved this part to a different paragraph and rephrased it to:

Among these 4612 T_{10m} observations, 15 measurements are either outside of the current ice sheet extent as defined by the GIMP ice sheet delineation (Howat et al., 2014) or outside of the 1950-2022 period we consider for our T_{10m} reconstruction. There are therefore 4597 T_{10m} observations in our compilation that can be used for the reconstruction of T_{10m} on the ice sheet between 1950 and 2022.

L148-149: A little clearer to say “Additionally, for each grid cell and monthly time step we calculate…”

Updated, thanks.

L152-153: May be clearer to write: “Lastly, to assist the ANN in capturing the annual periodicity, we input the cosine of the month (assigning 1 in January and -1 in July).

Updated, thanks.

L157: change “ANN” to “ANNs”?

Updated, thanks.

L163: change “will be” to “is”

Updated, thanks.

L181: “and inversely” is a little unclear. May be better to just say this explicitly… “while the weight will be less than one if the observation histogram is greater than target histogram.”

Updated, thanks.

L222: change “each the” to “each of the”

Updated, thanks.

L305: change “evaluate directly” to “directly evaluate”

Updated, thanks.

L314: change “is” to “are”

Updated, thanks.

L335-336: Should this sentence reference Figure 5?

This paragraph is dedicated to Figure 4 which supports the issue MAR has in the entire ablation area. Figure 5 indeed illustrates this statement at a few ablation sites (but not for the entire ablation area) and we would like to wait until Figure 5 is properly introduced in the next paragraph, before we reference it.

Figure 5 caption: change “y-axis” to “y-axes”

Updated, thanks.

L386: Isn’t MAR overestimating between 1950-2000 compared to the ANN?

Well spotted, thank you.

L412: Maybe instead of "decrease" just "trend" or "change" since the negative sign is already included

Updated, thanks.

L425-426: Clearer to write "This indicates that the ANN successfully learns which areas are susceptible to undergo meltwater infiltration and refreezing from the ANNs training data"

Updated, thanks.

L429: change "ANN" to "ANNs"?

Updated, thanks.

L441: change "Neither" to "Nor"?

Updated, thanks.

L488: note that "retention capacity" is retention capacity from refreezing rather than storage in perennial or ephemeral firn aquifers.

In the following sentence, we elaborate on retention through refreezing so we think that this part is clear.

L495: note that the averages are from the ANN.

We now mention that they are from the ANN.

References cited in review:

Nolin, A. W., & Stroeve, J. (1997). The changing albedo of the Greenland ice sheet: Implications for climate modeling. *Annals of Glaciology*, *25*, 51-57. https://doi.org/10.3189/S0260305500013793

Poinar, K., Joughin, I., LENAERTS, J. T., & Van Den Broeke, M. R. (2017). Englacial latent-heat transfer has limited influence on seaward ice flux in western Greenland. *Journal of Glaciology*, *63*(237), 1-16. https://doi.org/10.1017/jog.2016.103

Brils, M., Kuipers Munneke, P., van de Berg, W. J., and van den Broeke, M.: Improved representation of the contemporary Greenland ice sheet firn layer by IMAU-FDM v1.2G, Geosci. Model Dev., 15, 7121–7138, https://doi.org/10.5194/gmd-15-7121-2022, 2022.

Howat, I. M., Negrete, A., and Smith, B. E.: The Greenland Ice Mapping Project (GIMP) land classification and surface elevation data sets, The Cryosphere, 8, 1509–1518, https://doi.org/10.5194/tc-8-1509-2014, 2014.

Langen, P. L., Fausto, R. S., Vandecrux, B., Mottram, R. H., & Box, J. E. (2017). 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*, 110. https://doi.org/10.3389/feart.2016.00110

Vandecrux, B., Mottram, R., Langen, P. L., Fausto, R. S., Olesen, M., Stevens, C. M., Verjans, V., Leeson, A., Ligtenberg, S., Kuipers Munneke, P., Marchenko, S., van Pelt, W., Meyer, C. R., Simonsen, S. B., Heilig, A., Samimi, S., Marshall, S., Machguth, H., MacFerrin, M., Niwano, M., Miller, O., Voss, C. I., and Box, J. E.: The firn meltwater Retention Model Intercomparison Project (RetMIP): evaluation of nine firn models at four weather station sites on the Greenland ice sheet, The Cryosphere, 14, 3785–3810, https://doi.org/10.5194/tc-14-3785-2020, 2020.