Response to Referee 1 on egusphere-2023-1556

First, we would like to thank the Referee for reviewing and commenting on the manuscript, which will improve the quality of the manuscript. Please find the item-by-item reply below, with the original comments in *italics* and the responses in blue. All the suggested changes will be implemented in the revised text that will be uploaded.

*This paper details a study using machine learning (ML) to examine Antarctic firn density. The paper is interesting and needs some further revisions before it is suitable for publication. I have put some suggestions and questions below.*

*Major comments:*

*Introduction, I suggest you start bigger, why does Antarctica ice sheets matter to the globe? Also, I think you need to define firn for folks who are not clear on what it is.*
We appreciate the suggestion, and we will add the importance of Antarctic mass loss and sea-level rise. We will also add an explanation of firn in the revised manuscript.

*On line 142, you say that the firn model has a resolution of 27 km – is that sufficient to capture the firn variations? This is quite coarse, in my opinion. Is this 27 km by 27 km grid cells? I think this needs to be stated more clearly.*
The 27km model resolution is indeed coarse as it corresponds to the resolution of Antarctic wide state-of-the-art climate models that typically drive firn models. This coarse resolution is therefore not expected to capture the fine scale variations on the steep slopes of the Antarctic Peninsula or along grounding lines as the 27x27 km horizontal resolution is too coarse to resolve atmospheric variables. However, this study focuses on dry pixels, which are mainly located in regions of the AIS where climatic gradients, and thus firn property gradients, are not that large.

Moreover, we want to stress that our study is also based on/limited by the coarse resolution of the satellite radiometer (25 km). According to Picard et al. (2014), who compared the metre-scale ground-based brightness temperature measurements to the coarse-resolution satellite brightness temperature measurements around Dome C in Antarctica, there is indeed metre-scale density variation, but "the study also shows that, for the hectometre to kilometre scales, the variations are much smaller. The average of the ground-based brightness temperature is close to the SSM/I and WindSat satellite observations meaning that the investigated area was representative of the pixel of the satellites including Dome C. An important consequence is that spaceborne passive microwave sensors cannot spatially resolve these wind-formed features, but they are very sensitive to the areal proportion of these features." Given the gentle slopes in the interior of Antarctica, we expect this representativeness also to apply to the dry region pixels we studied.

Nevertheless, based on the previous arguments for the representativeness of coarse resolution for both models and satellite observations, we do agree that the coarse resolution may raise questions. To address these, we will adapt the discussion to clarify the impact of resolution.

*I think you need at least one study site figure that has all of the locations you refer to in the paper on one introductory map. See my comment from Line 152, for example.*

We will try to improve the indication of locations in the revised manuscript. To address the concerns of the reviewer, we refer to Fig. 1 of this document (below) for the locations. However, following both reviewers' suggestions, we will increase the training dataset and should assess how to better present the figures.

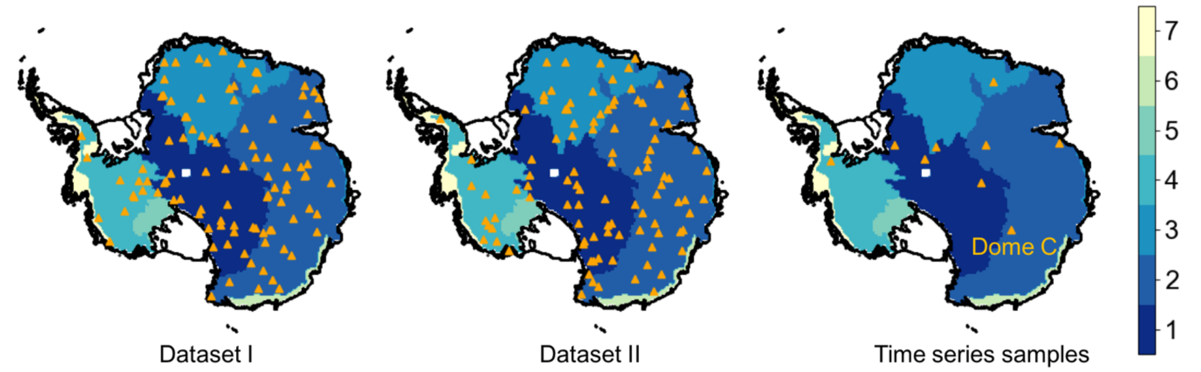


Figure 1. Indication of the mentioned locations.

*Overall, the study design seems confusing. You take the time to cluster the data, but then you do not use it for the analysis, really. Why would you not use that to identify the dry-snow zones, and then perhaps build multiple RF models to see what zone could be best captured? This seems like an interesting approach to take but was not used. I think that this would also eliminate the need to only model the non-wet areas if you simply remove the regions that do poorly in satellite observations.*

We admit that the description of the study design could be better elucidated. To simply answer the reviewer's question, the purpose of clustering was indeed to identify the dry-snow zones. Then, the clusters are used to ensure that different regions are represented sufficiently.

Overall, we hope the following flowchart (Fig. 2) is helpful in resolving the confusion, which we also noticed in the other comments. In this flowchart, the rectangles represent original parameters consisting of: i) satellite parameters (TB and sigma0), ii) IMAU-FDM densities, iii) external datasets used for result analysis, and iv) a set of hyperparameters to define the RF regressor. The ovals represent derived parameters. The rounded rectangles represent steps of our study. To be specific, the time series anomalies from TB and sigma0 are clustered to identify dry snow zones. Four distinct dry snow zones have been identified, but we have to admit that we could not relate the separation of dry snow zones to actual physical phenomena. Then, for the dry snow zones, estimation of firn densities using RF regressor is performed.

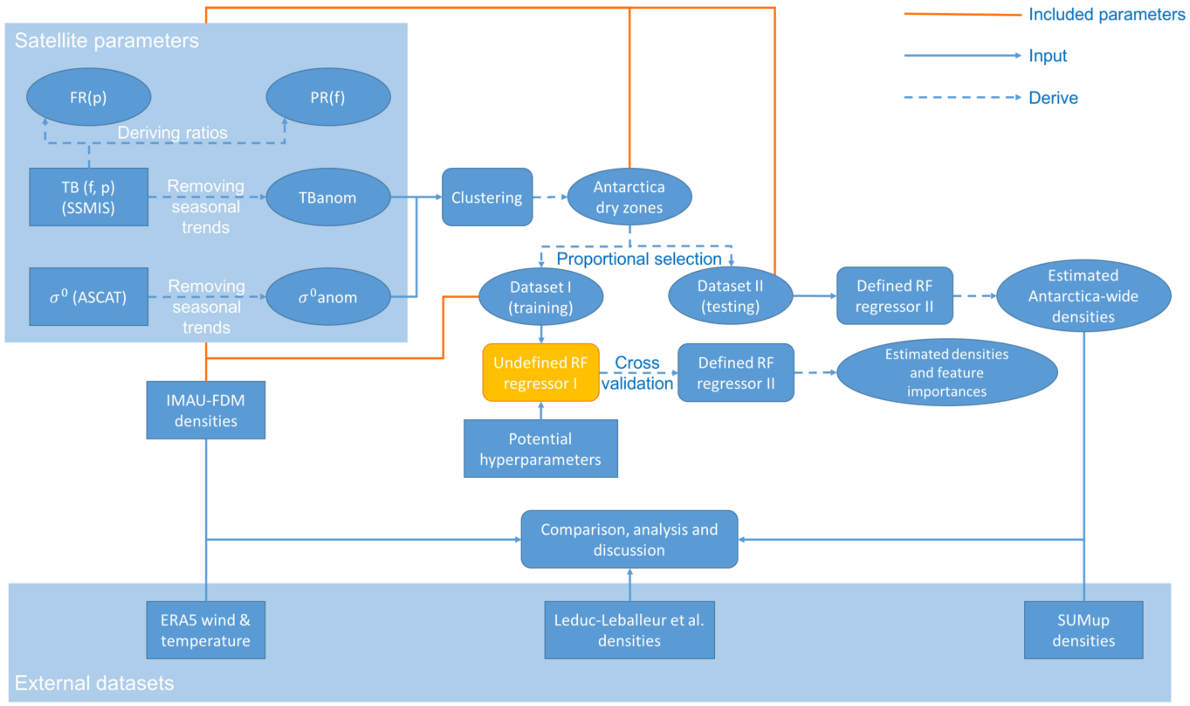


Figure 2. Flowchart of the study design.

The application of the RF regressor consists of three steps (Lines 231—243 of the manuscript). To reduce overfitting, the first step is to use a training dataset (Dataset I in Fig. 2) to perform a hyperparameter tuning through a 5-fold cross validation process (orange rounded rectangle in Fig. 2). The selection of pixels for Dataset I is proportional to the total pixels in each of the clustered dry snow zones, and the actual numbers of pixels are shown in Table 1 below. Please note that both 19027 and 100 are the numbers of pixels, but the features include 10 years of satellite parameters with a temporal resolution of 10 days, therefore the training dataset consists of 100pixels*366time_steps = 36600 samples. RF is trained with the IMAU-FDM densities.

Table 1. Statistics of pixels per cluster and pixels used for further RF estimating.

| Cluster | Number of pixels | Number of training pixels |
|---|---|---|
| Firn 1 | 4540 | 26 |
| Firn 2 | 7360 | 42 |
| Firn 3 | 3465 | 20 |
| Firn 4 | 2284 | 12 |
| Firn 5 | 429 | 0 |
| Firn 6 | 325 | 0 |
| Firn 7 | 624 | 0 |
| Total | 19027 | 100 |

The second step of the application of the RF regressor is to provide a simple visualisation of the performance of the tuned RF regressor, and the importance of each feature. In this step, another 100 pixels (Dataset II) are used. They are again proportional to the number of pixels per cluster, but the locations are different from Dataset I. The target parameter is the densities of Dataset II, again consisting of 100pixels*366time_steps.

The third step is using the tuned RF regressor to estimate the densities over the entire dry snow zones in Antarctica. Please note that after the hyperparameter tuning in the first step, we use the identical set of hyperparameters for the RF regressor in both the second and the third steps. The training dataset is also identical, which remains the samples from Dataset I. We would like to point out that the proportional selection of Dataset I is important, because we also tried using 100 random pixels not restricted by the clusters, and the result degraded in central Antarctica in terms of RMSE (see figure below).

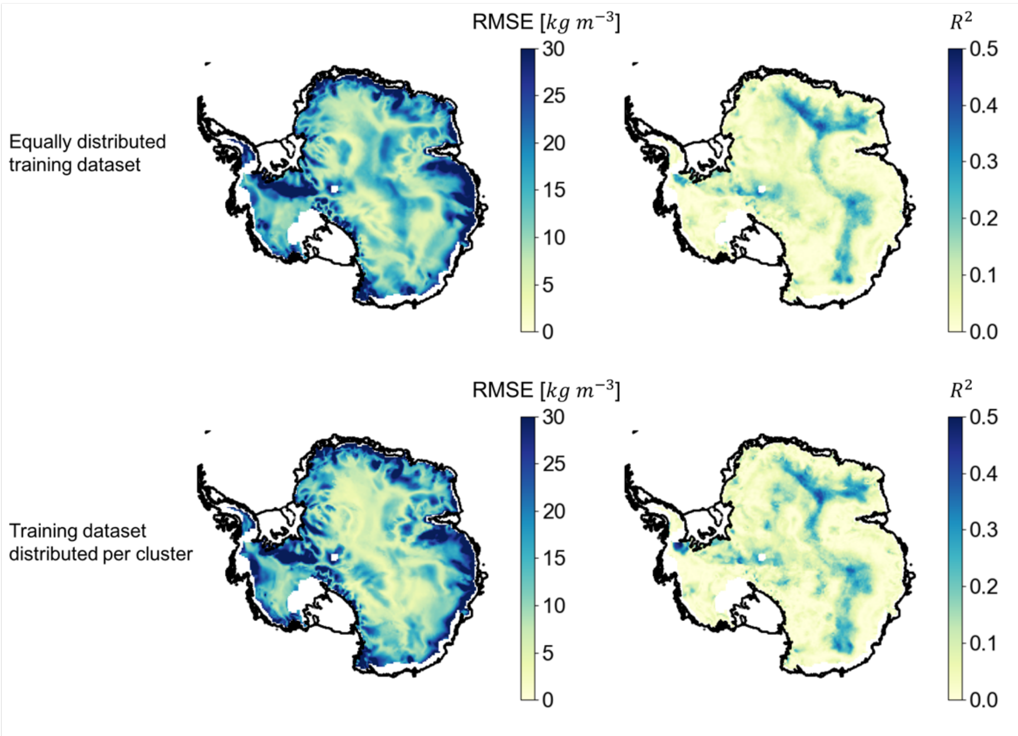


Figure 3. Comparison of performance between using randomly selected pixels (upper row), and proportionally selected pixels (lower row).

Therefore, the clusters are used to ensure that the training samples are selected in a way where different regions are sufficiently represented. We did not train different RF models for different clusters although this should be feasible and interesting, but is outside of the scope of the current paper. Nevertheless, we can add the suggestion to the discussion.

*I do not understand why you didn't use the RF and importances to reduce your model variables. As you show in Figure 5, it looks like these anomalies are not adding much to the RF model. I think you might be able to remove them in the analysis.*

We appreciate the suggestion. However, the hyperparameters are already tuned based on the whole set of parameters. Changing the combination of parameters requires tuning another set of hyperparameters. Therefore, we can add another section to the manuscript regarding changing the combination of the parameters.

*Did you consider other types of ML models, or did you just decide to use RF approaches? Why not consider other approaches?*

We considered using support vector machines (SVMs), but as the previous major comment pointed out, we would like to take advantage of the importances from the RF regressor to understand which parameters are the most influential factors. Moreover, we would like to stress that the scope of this study is to assess the feasibility of "combining radiometer and scatterometer remote sensing data to assess Antarctica-wide dry firn density by using a state-of-the-art ML method” and not to compare different ML algorithms. Therefore, discussing the performances of different supervised ML algorithms is beyond the scope of our study. However, we appreciate the reviewer’s suggestion, and agree that a comparison between different machine learning algorithms can be an interesting scope for future studies and we will stress this explicitly in the discussion.

*On lines 325, you say “that do not correspond to changes in densities in dry-firn regions?. This line has me wondering about the objective of your work. Are you interested in the firn estimation or are you interested in the change in firn over time? Is the RF model developed for this? Or, are the clusters? You say in the beginning of the paper (Line 71) that the objective of this paper is to “assess the feasibility of combining radiometer and scatterometer remote sensing data to assess Antarctica-wide dry firn density.” But, you also say on Line 220 “As our goal is to relate the satellite time series to assess spatio-temporal variations in firn density, we adopt an alternative approach that uses the output of IMAU-FDM as training data instead of relying on in situ data.”. What is the objective of this work? If it is average firn, then you can develop your model in one way, but if it is not, then you should develop it in another.*
The main objective is to propose and assess a methodology to derive firn density and its spatial and temporal variations over the Antarctic ice sheet based on on daily satellite observations (and not on changes in these observations). More specifically, assuming firn densities in several locations are known, our study tries to assess firn densities of the unknown regions in space and time using a combination of satellite observations, namely brightness temperature (Tb) from SSMIS, and backscatter intensity (sigma0) from ASCAT. The motivation is that multiple drivers (e.g. wind velocity, firn temperature) of changes in satellite observation can also drive the changes in firn densities, but the mechanism has not been explicitly quantified or modelled. The “known densities” in our study, are assumed to be the modelled firn density from IMAU-FDM, which is a firn model. Therefore, this paper focuses on both the spatial estimation of firn density, which seems to work well, but also on the temporal variations, which performs less well. Since RF method is based on the daily observations, it does not directly account for changes (e.g. by including change parameters in the RF model), but it does so indirectly by assuming that the satellite data reflect these changes as well. We will clarify that better in the revised version.

*Minor Comments:*

*Line 17, short-term (or seasonal) variations, is it both or do you just mean seasonal?*
It should be both. The parentheses will be removed in the revised manuscript.

*Line 30, This statement needs a reference.*
We assumed that the references in the previous sentence (Macelloni et al., 2007 and Champollion et al., 2013) would also be applied here. This will be clarified in the revised manuscript.

*Line 63 However, the precise mechanisms underlying the interaction between firn densities and satellite observations cannot always be fully understood (Champollion et al., 2013; Fraser et al., 2016; Rizzoli et al., 2017). What do you mean by this? Interaction implies they are interacting, which they are not…*
Indeed. We will clarify and adapt it in the manuscript that we look at the effects of density on satellite observations.

*Line 67, "to other areas or time periods therefore requires further assessment (Tran et al., 2008; Fraser et al., 2016; Nicolas et al., 2017; Rizzoli et al., 2017)". What did they find? Was it successful, i.e, did it work?*
Tran et al. (2008): this study classified snow facies over both Greenland and Antarctica in 2004 based on passive microwave data (brightness temperature) and altimeter data (backscatter intensity) using an unsupervised ML method. The study regarding Antarctica did not capture melt zones, but indicated "a strong topographic control on the class distribution". This is already different from our study, as we managed to detect melt zones in the more recent decade.

Fraser et al. (2016): this study discussed the scatterometer "backscatter response to surface forcing parameters (wind speed and persistence, precipitation, surface temperature, density and grain size)" by comparing the backscatter with modelled parameters between 2007 and 2012. The study shows that sigma0 is affected by surface temperature and wind speed, hence provides theoretical background for our study.

Nicolas et al. (2017): this study identified a melt region in West Antarctica, close to the Ross Ice Shelf, hence provides theoretical background for our study.

Rizzoli et al. (2017): this study characterised snow facies over Greenland using interferometric synthetic aperture radar (InSAR) acquisitions. The study identified melt zones using an unsupervised ML algorithm, hence provides theoretical background for our study.

We will try to add it to the introduction concisely.

*Generally, italicize In situ.*
Perhaps it is not necessary for The Cryosphere; see Orsolini et al. (2019), for example.

*On line 70, you talk about calibration. You did not mention calibration previously, and it is unclear what this is referring to. Models? The satellites? Fusion methods? I think this needs to be tied to modeling and why calibration is needed. Otherwise it seems to be coming in the text out of the blue.*
We agree and this statement will be rephrased.

*Line 72, you talk here at three experiments, did you compare /use the observations in situ ever? It seems like the SUMup is not used (or mentioned) in any one of the experiments. I think if you are going to mention SUMup, you need to say where it was applied in the experiments.*
SUMup is not used for setting up the experiments, but for the validation and analysis of where the potential errors come from. This will be clarified in the revised manuscript.

*Line 132, "outputs of the regional atmospheric climate model RACMO2.3p2" These scales seem really different… What resolution is the model run at?*
We will clarify this as follows:

Lines 131-133:
IMAU-FDM simulates the transient evolution of the Antarctic firn column, and is forced at the upper boundary by outputs of the Regional Atmospheric Climate Model (RACMO2.3p2) at a 27 km horizontal resolution (van Wessem et al., 2018).

*Line 140 – move these two sentences up to say this earlier (perhaps line 131), that will assist with my previous comment. The first sentence of this paragraph could be combined with the previous one.*
This will be implemented in the revised manuscript.

*Line 132, RACMO2.3p2 – define?*
We will give the definition with capitals. Please see the comments above for the definition.

*Generally, through the text, you refer to "the model output", or "models". As you have multiple models, I suggest calling the models by their names, or ensuring they are referenced clearly to distinguish the model.*
This will be better clarified in the revised manuscript.

*Line 135, we focus on the density of the… How many layers are there in this model in total?*
The model employs up to 300 layers in total of 3 to 15 cm thickness, which represent the firn properties in a Lagrangian way. The output is resampled to a regular grid with layers of 4 cm.

*Line 142, the firn data are reprojected – this is modeled data, correct? I think you want to make sure to differentiate the model from the observations.*
Yes, this will be clarified as "the firn density model data from IMAU-FDM".

*Line 138, "…have been acquired at approximately this depth…" Why? This seems kind of arbitrary. Also, 4 cm seems very shallow for firn. Is this because it is in Antarctica?*
It is mainly because in Fig. 9, we compare IMAU-FDM with in situ measurements acquired in 2014—2015. The in situ measurements were acquired within 0—2 cm depth. This is comparable to the highest vertical resolution of IMAU-FDM dataset we are using (4 cm).

*Line 145, Surface Mass Balance and Snow on Sea Ice Working Group (SUMup) dataset. You have already used this acronym, define it earlier.*
This will be improved in the revised manuscript (together with the issue on Line 72).

*Line 146, "at the smallest mid-point depths" More clarity please, what is 'small' and what is the mid-point of?*
Mid-point refers to the mid-point of the ice sample. SUMup provides information on start-point, end-point and mid-point. We use the mid-point here to define the depth of the reference data. Sometimes multiple samples are taken at each location. To make the visualisation clearer, we only use the shallowest depth of the samples at each location. This will be clarified in the revised version.

*Line 151- For each date of measurement at each location, talk about the locations and dates first… What locations are these dates at?*
We are sorry but we did not really understand the nature of this comment. However, we have SUMup data at specific locations (shown in Fig. 6a and 6b of the manuscript) sampled at different moments in the period between 1984 and 2017 and a time series at Dome-C (shown in Fig. 3 of the manuscript).

*Line 152, Dome C, where is this? Map?*
This is mentioned in Line 269 and shown in Fig. 3. We will improve the manuscript to refer to the figure which shows the location (Label E in Fig. 3).

*Line 159, By incorporating this information… I don't understand how the ERA5 data was used and why it was used. This needs to be better explained.*
As mentioned in Line 156 of the manuscript: To assess the difference between the measured, modelled and estimated densities, it is important to understand the effects of climate conditions. Therefore, we use the climatic data as a comparison. This should also resolve the comment below. We will explain it better in the revised manuscript.

*In Section 2.5, are you talking about comparing model and observations (at points?).. and the satellites? I think this needs to be thought through and justified in the text. Comparing satellite and model data with single point measurements is tricky. There are a lot of references out there about how to do this, particularly in the climate modeling realm. I suggest the authors read some of these papers and at least add a discussion in the text around this.*
The point of this section is to point out that potential errors with IMAU-FDM are linked to certain climate conditions, which can be propagated through the training process to further bias the results. ERA5 serves to help understand in which conditions IMAU-FDM leads to more ideal results. This analysis is done Antarctica-wide, and has nothing to do with comparing model and observations at points.

*Sometimes you say "firn data" and other times you talk dry firn. Should this be defined? Can you make sure you are being consistent through the text?*
Yes, this will be made clear in the revised manuscript.

*Line 168m dry-snow zones, what are these?*
Section 3 (until 3.1) is a high level description of the next section to provide an overview of the approach. The dry-snow zones are therefore explained in Section 3.2, as is also indicated between brackets. It is a preview that will be explained in Section 3.2.

*Line 184, model training procedure. Which model are you talking about?*
Random forest model. This will be clarified in the revised manuscript.

*Line 190-195, this is not very clear. Can you rephrase?*
We will rephrase it into:
We expect that clustering the time series of satellite observations will effectively differentiate pixels experiencing melting from those unaffected. By identifying and excluding melt-affected pixels, we can ensure the accuracy of density estimations using the RF regressor. Additionally,

to enhance the RF regressor's ability to capture the characteristics of various dry snow types, we choose training samples based on the identified dry snow types. This approach enables the representation of diverse snow types in the training dataset, improving the RF regressor's accuracy in estimating density across different snow types.

*Line 196, variations of other properties. What other properties?*
This statement intended to tell that by removing the surface temperatures, the non-annual variations such as melt—refreezing cycles, potential precipitations and density or snow grain size variations could be kept, which in turn helps us distinguish different snow regions especially distinguish melt from non-melt regions and facilitates the following steps (please refer to Fig. 2 of this document as well). This will be better clarified in the revised manuscript.

*Line 196, In addition, although may not have such large dependence on firn temperature as TB, we use its time series anomalies to maintain consistency with TB. This is unclear, can you rephrase?*
There was a mistake. sigma0 is also affected by firn temperature. We will rewrite the whole concept in the revised manuscript.

*Line 204, 'distance' between pixels. Make sure to clarify that this is not spatial distance.*
Should be distance between features of the pixels. We will clarify that in the revised manuscript.

*Line 205, between the parameters of different pixels. What parameters?*
All parameters mentioned in Lines 195—200. Please refer to the following answer as well.

*Line 210, different satellite parameters, together with the IMAU-FDM density for each cluster. I do not understand what this means. What parameters? I think you need to make a table of parameters.*
The satellite parameters include brightness temperatures and derived ratios, as well as scatterometer backscatter intensity. We will clarify it better in the revised manuscript.

*Line 217, RF regressor. Add more references since this approach is now widely used in climate science, for snow distribution mapping and other work.*
This will be improved in the revised manuscript. For example, we will add Vafakhah (2022) and Viallon-Galinier (2023).

*Line 225, for pattern recognition in noisy datasets. Add a reference here.*
This will be added in the revised manuscript.

*Line 226, reduce the variance of the model and prevent overfitting. Add a reference.*
This will be added in the revised manuscript.

*Did you consider other ML approaches?*
Please refer to the major comments.

*Line 230-onward. Are you building a RF for each time step to estimate the timeseries at each grid cell? How many samples in total went into the model? Line 244 talked about pixels, and* *the resulting sample size. Is this the total number of samples? Do you think this is enough to train a RF, especially considering the results?*

No, we do not rebuild an RF for each time step. We build one RF model that can be used for multiple time steps. The RF is tuned for multiple pixels (100 and please refer to Fig. 1 of this document) and multiple time steps (3,66 daily samples between January1, 2011, and December31, 2020) using the training dataset through a 5-fold cross validation process, and the tuned RF is then used throughout the other experiments. For the details please refer to the major comments.

However, we do appreciate the suggestion to increase the sample size, and also considered using 10% of all pixels instead of 100 pixels for training. Please note again that when we use 10% of the pixels, the total number of samples consists of 1739pixels*366time_steps. The performances (RMSE and R^2) are shown below. From this comparison, we see that increasing the training samples slightly improves the RMSE overall, and enhances R^2 in some regions. We will include it in the revised manuscript.

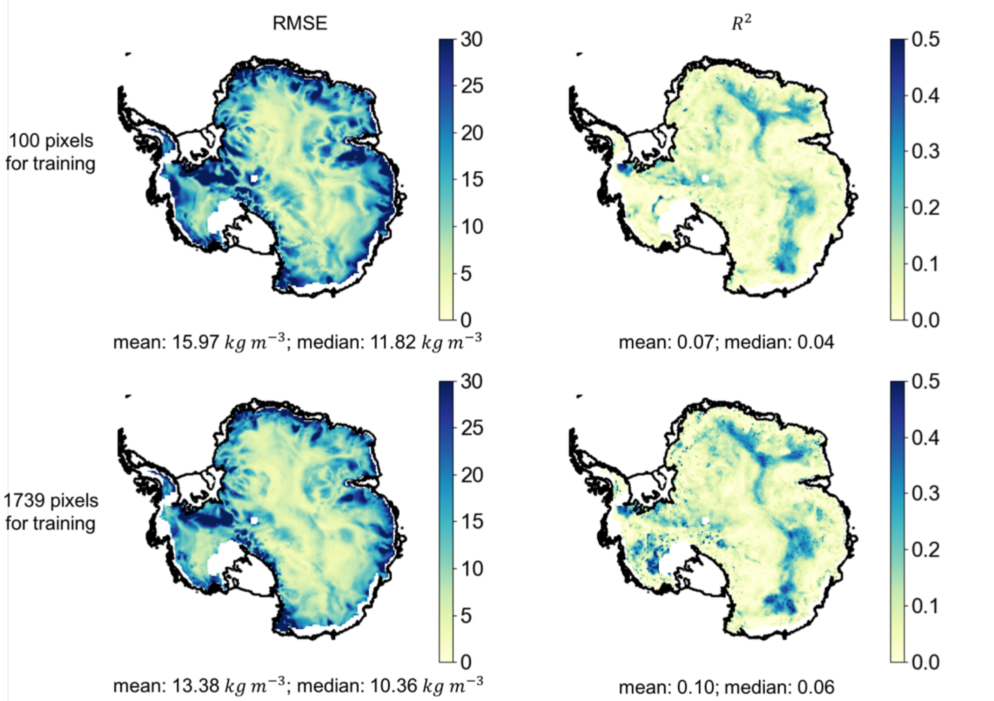


Figure 4. Comparison between using 0.5% of the data and 10% of the data for training.

*Table 1. These are all hyperparameters, do you call them parameters (which you use other times in the paper for the actual input parameters to the model, which in itself is confusing).*

They are not the same. Hyperparameters are hyperparameters used by the RF regressor, as stated in the caption of this table. It is also a typical element of supervised machine learning (see Anilkumar et al. (2023)). Input parameters are referred to as parameters, including radiometer and scatterometer measurements and the derived products. Please refer to the major comments (Fig. 2 of this document).

*Line 258. Why did you only use Gini importance vs other importance metrics? Did this choices affect any of your importance rankings?*

We considered the permutation importance, and the ranked importance (using 10% of the pixels) is shown below:

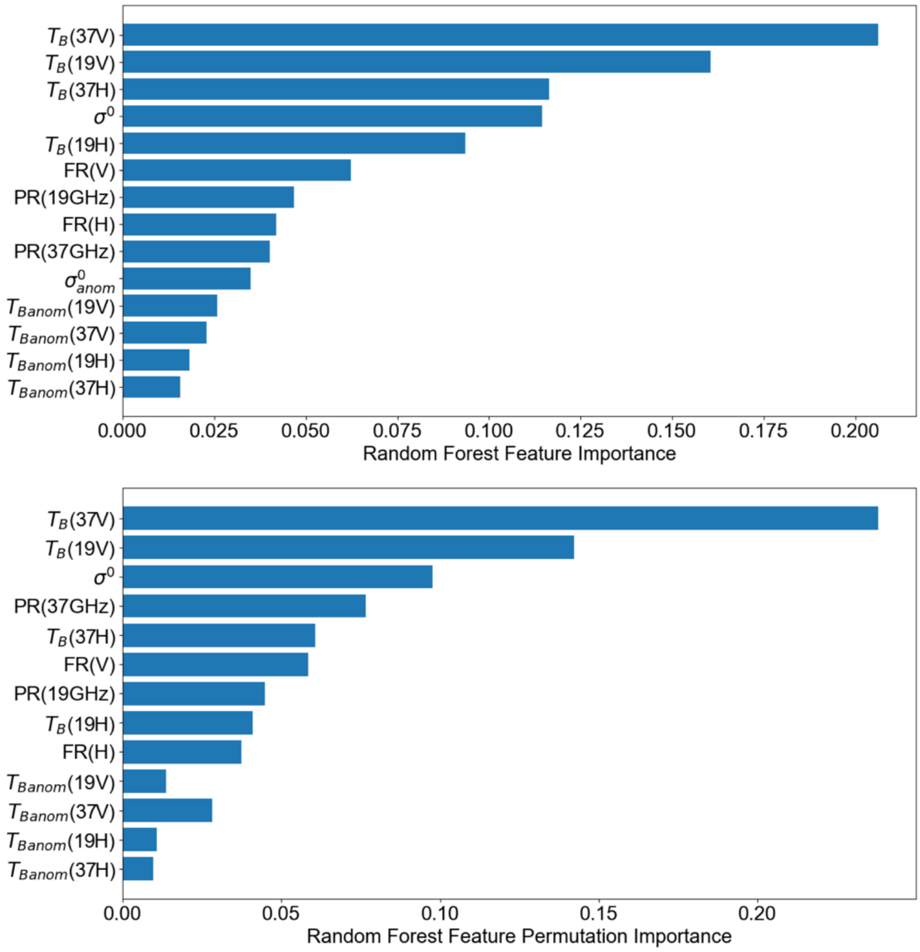


Figure 5. Comparison between upper: Gini importance, and lower: permutation (Breiman) importance.

The rankings indeed changed. We will include both importances in the revised manuscript.

*Line 264, by means of the RMSE. Do you mean averages or do you mean via using RMSEs?*
Sorry for the confusion. We mean by using the RMSEs. This will be clarified in the manuscript

*Line 269, this is the first time you refer to Figures… why do it here and not in the rest of the methods?*
This figure will be updated (see major comments) to reflect all sample locations and will be used throughout the methods section to better show the locations of the different sample locations.

*Line 277, 'cluster Firn 5', you have not introduced the cluster results yet so we do not know what these are.*
We will better refer to the clusters in the revised manuscript.

*Figure 1, can you tell us which ones are from the satellites parameters and which ones are from the model? Again, a table might help with this.*

We will improve the manuscript. To address the reviewer's concern, the datasets and the differences of data sources are: 1) the observable Tb from SSMIS satellite mission, 2) sigma0 from ASCAT satellite, and 3) densities from IMAU-FDM, which is a modelled parameter.

*Line 283, especially at the location of Dome C, again, need to show on the maps or include a figure.*

Please refer to the major comments.

*Line 287, There are a lot of other good reasons why the RF should be used, I do not think that this is the strongest one.*

We refer to Anilkumar et al. (2023) for the performance of the RF. Moreover, the advantages of RF include the capability of capturing non-linear relationships between input features and the target variable; less sensitivity to outliers compared to simple linear regression; handling the correlated predictor variables; providing a unique feature importance; no assumption about data distribution; as well as less overfitting. This will be clarified in the revised manuscript.

*Section 4.2 Do you think that you could produce different RF models for each cluster, perhaps? I think this would be very interesting to understand the difference between the performance of each of these models. For instance, if the dry firn can be modeled with lower RMSE /error than some of the other clusters. Honestly, I am still unclear if you did it this way or not.*

We are sorry that the approach was not 100% clear. The clusters are used to ensure that the training samples are selected in a way where different regions are sufficiently represented. We did not train different RF models for different clusters although this should be feasible, but is outside of the scope of the current paper. (Please also see the major comments).

*Figure 4 and Figure 5a*

*These figures are making me wonder what it is that you are trying to do. For instance, are you trying to estimate the time series of the seasonality and variability you see in Figure 4 with the RF? What is the RF estimating, exactly…? You say "firn densities based on satellite parameters" and you talk about a time series, but I am wondering how you are doing this. What is X in Equation 4, actually? I do not know if this is ever said.*

Figures 4 and 5 are from separate experiments. First, we would like to refer to the answer to the major comments. So, Fig. 4 shows that after clustering, dry firn zones and firn zones that experience melt can be distinctively recognised. However, between different dry firn zones, we cannot intuitively relate the time series to actual physical firn properties (mainly due to lack of field measurements). Nevertheless, by proportionally choosing training points within each cluster, we do observe an optimal performance of the experiment (see Fig. 3 of this document). We attribute this to the reason that all types of regions are represented sufficiently in this way.

X refers to the set of features.

*Figure 5b, are all these parameters standardized? Is the importance based on the standardized inputs? I just wonder because the anomalies appear to be the least important, which makes* *me wonder if perhaps the other parameters are not. Again, if these are not contributing much to the model, did you play around with them being removed? Does the model improve with fewer parameters? Are there any strong correlations between these parameters at all? How are they related or not related to each other?*

The parameters are not standardised. We assumed that random forest does not require standardising, as the tree partitioning depends on the scales of the independent variables. Moreover, Fig. 1 of the manuscript shows that sigma0 varies between -25dB and 0dB, yet ranks as an important feature.

Typically, all TB values are highly correlated to each other, as they are mainly affected by firn temperature. We understand that one may be concerned to have multiple correlating features, hence performed another experiment using only vertical channel of 19 GHz brightness temperature, sigma0, polarisation ratios and frequency ratios. The comparison is shown below. It is interesting to see that the original setting still outperforms in terms of RMSE.

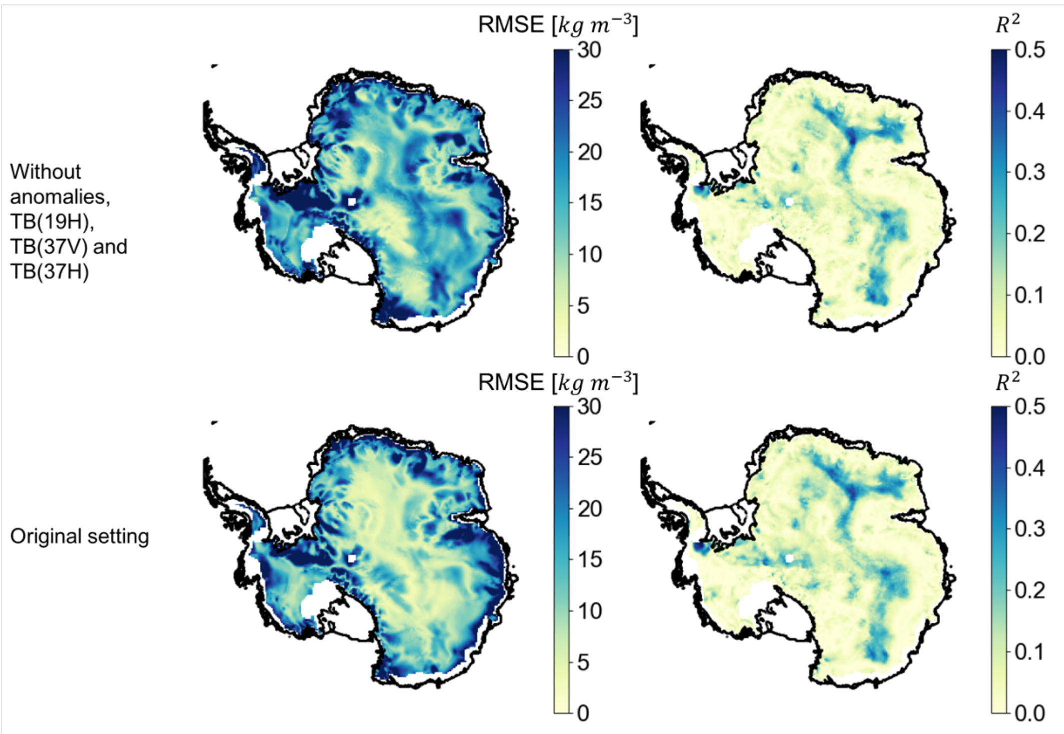


Figure 6. Comparison between using selected parameters (upper row) and the original setting (lower row).

*Line 301, The differences between these clusters mainly arise from deviations in TBanom and, to a lesser extent, sigma0anom. What is different about them?*

We notice that for cluster 1, TBanom varies between -5K and 5K, for cluster 2 and 3, TBanom varies betwee -5K and 10K, and for cluster 4, TBanom varies between -10K and 10K. Moreover, compared to cluster 2, TBanom of cluster 3 experienced a decreasing trend over time. What we could assume is that cluster 1 consists of most interior regions, hence is overall most stable, whilst cluster 4 is located in West Antarctica, hence is least stable (with the largest variations). The separation between cluster 2 and cluster 3 resembles Fig. 4 in Stokes et al. (2022), in which cluster 2 tends to lose mass while cluster 3 tends to slightly gain mass. However, we can only infer that this result might indicate that cluster 2 has a less stable condition than cluster 3, but the conclusion is not solid.

*Line 298, If 1-4 are the basically the same, why are they not being treated as a single cluster?*
We would not conclude that clusters 1—4 are "basically the same". Rather, at the moment we cannot relate the differences to actual physical phenomena.

*Line 303, Are the melt events shown /evidenced in time in the region? Can you talk about this a little bit? You refer to a paper, but don't go into detail otherwise.*
We refer to de Roda Husman et al. (2022), where it shows that satellite-based melt events are commonly well recognised.

*Line 305, Can you describe why how density would change under these melt events, and why? You do not give much background on that.*
After melt events, the density increases by refreezing. This is a typical phenomenon, which is also documented in Fig. 4 of Nilsson et al. (2015) that showed the high-density melt layers during the famous melt over Greenland in 2012.

*Line 306, Firn 5, where the melt event of 2016 shows a prolonged effect on the anom time series due to the formation of a sub-surface refrozen high-density layer in IMAU-FDM. Again, what I the implications for this, and what does it mean for firn?*
We refer to Nilsson et al. (2015), where it shows that a sub-surface refrozen layer drastically changes the volume scattering mechanism hence changes the backscattering signals.

*Figure 6. Again, this figure only shows results as temporal averages. How did the time series of the RF do?*
We cannot show all the time series as there are 17649 pixels all together. That is the reason why we took 9 sample pixels to visualise the time series in Fig. 8, and conclude that the precise temporal performance of RF is compromised.

*Line 383, It is important to note that the wet firn clusters are not used in the following RF steps due to the complex impact of the melt–refreeze cycle on satellite observations. Again, I am thinking that the RF and this cluster analysis is not related.*
Please also refer to the major comments and Fig. 2 of this document. To briefly address this question, the clustering separated the wet firn from the dry firn, so it helps the following analysis.

*Line 317, Exhibiting a linear relationship between predictors and the predicted variable – predictand? Saying it this way is confusing.*
We will check the consistency of descriptions here.

*Figure 5, add units.*
This will be corrected in the revised manuscript.

*Figure 5a, Why do no values exceed this amount? I wonder if perhaps your training data set somehow selects lower firn values… are you randomizing between your training and test sets?*

As y-axis (IMAU-FDM densities) of this figure shows, we have selected density values up to 500kg/m^3. Please note that within 4cm depth of the snow in Antarctica, it is normal to have most of the density below 400kg/m^3, so it is possible that the values that exceed 400kg/m^3 are less represented in the RF training process, which could indeed indicate a sampling issue. Therefore, by using 10% of the pixels as training samples, we hope to better resolve this and add the analysis to the revised manuscript accordingly.

*Figure 6 shows that the model is basically as good as the RF (if not better at anything other than the mean). So, why do you need an RF model in this case? How difficult is the model to set up and apply? Again, is there a good reason for the RF here if it doesn't perform that well, is not finer in scale, or it doesn't really do that well except on average?*
Please refer to the major comments. The objective of our study is to assess the ability of using a combination of ML algorithms and satellite parameters to estimate firn densities, not to reproduce the modelled density. To do this, we require sufficient training data. However, due to the limitation of the in situ measurements, we use IMAU-FDM as an assumption of "real densities". Actually, as Fig. 6 and Fig. 9 indicate, IMAU-FDM does not capture many variations in the in situ data, resulting in temporal gaps in the RF estimations. This has been pointed out and analysed in Lines 390 onwards.

*Figure 6d, can you show which is which? Use different symbols instead of colors? They are difficult to differentiate. This intercomparison with observations is likely very challenging to achieve (which I think is what you are attempting to do). I might suggest some sort of spatial upscaling for single point /insitu observations.*
This will be improved in the revised manuscript.

*Figure 8 illustrates how poor the RF is for a time series. But, I am unclear if you are doing this in the right way. Clarity of methods is required.*
Please refer to the previous comments.

**Reference**

Anilkumar, R., Bharti, R., Chutia, D., and Aggarwal, S. P.: Modelling point mass balance for the glaciers of the Central European Alps using machine learning techniques, The Cryosphere, 17, 2811–2828, https://doi.org/10.5194/tc-17-2811-2023, 2023.

de Roda Husman, S., Hu, Z., Wouters, B., Munneke, P. K., Veldhuijsen, S., and Lhermitte, S.: Remote Sensing of Surface Melt on Antarctica: Opportunities and Challenges, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 1–20, https://doi.org/10.1109/jstars.2022.3216953, 2022.

Nilsson, J., Vallelonga, P., Simonsen, S. B., Sørensen, L. S., Forsberg, R., Dahl-Jensen, D., Hirabayashi, M., Goto-Azuma, K., Hvidberg, C. S., Kjaer, H. A., and Satow, K.: Greenland 2012 melt event effects on CryoSat-2 radar altimetry, Geophysical Research Letters, 42, 3919–3926, https://doi.org/10.1002/2015gl063296, 2015.

Orsolini, Y., Wegmann, M., Dutra, E., Liu, B., Balsamo, G., Yang, K., de Rosnay, P., Zhu, C., Wang, W., Senan, R., and Arduini, G.: Evaluation of snow depth and snow cover over the

Tibetan Plateau in global reanalyses using in situ and satellite remote sensing observations, The Cryosphere, 13, 2221–2239, https://doi.org/10.5194/tc-13-2221-2019, 2019.

Picard, G., Royer, A., Arnaud, L., and Fily, M.: Influence of meter-scale wind-formed features on the variability of the microwave brightness temperature around Dome C in Antarctica, The Cryosphere, 8, 1105–1119, https://doi.org/10.5194/tc-8-1105-2014, 2014.

Stokes, C.R., Abram, N.J., Bentley, M.J., Edwards, T.L., England, M.H., Foppert, A., Jamieson, S.S.R., Jones, R.S., King, M.A., Lenaerts, J.T.M., Medley, B., Miles, B.W.J., Paxman, G.J.G., Ritz, C., van de Flierdt, T., Whitehouse, P.L. Response of the East Antarctic Ice Sheet to past and future climate change, Nature, 608, 275–286. https://doi.org/10.1038/s41586-022-04946-0, 2022.

Vafakhah, M., Nasiri Khiavi, A., Janizadeh, S., Ganjkhanlo, H. Evaluating different machine learning algorithms for snow water equivalent prediction, Earth Sci Inform 15, 2431–2445. https://doi.org/10.1007/s12145-022-00846-z, 2022.

Viallon-Galinier, L., Hagenmuller, P., and Eckert, N.: Combining modelled snowpack stability with machine learning to predict avalanche activity, The Cryosphere, 17, 2245–2260, https://doi.org/10.5194/tc-17-2245-2023, 2023.