Response to Reviewer 1:

First of all, we would like to thank Devon Dunmire for the time for reviewing our manuscript. We appreciate the constructive and insightful feedback we received. Following the feedback we have improved the explanation of our methodology and the clarity of the text. Below, we address your specific suggestions. Responses to the comments of the reviewers are written in red and citations of the manuscript are written in blue.

Kind regards, Sanne Veldhuijsen

Major comments

First, L80 reads: "The bucket method is computationally efficient but does not allow for saturated pore spaces, preferential flow, standing water over ice layers, or horizontal flow". Since a firn aquifer is quite literally "saturated pore space", I think a more in depth explanation about why the bucket scheme can be used in this work is required.

Thank you for this suggestion. We already elaborated on this in the discussion, but we add an explanation: "Since PFAs are defined as saturated firn, this is something different from what IMAU-FDM can simulate. However, the model does simulate the insulation of downward percolating liquid water by sufficiently high accumulation rates, that prevent the meltwater from refreezing in the winter. Therefore, the presence of year-round or spring liquid water in the firn has been successfully used to identify the presence of aquifers (Forster et al. 2014, Kuipers Munneke et al. 2014, Van Wessem et al. 2018, Brils et al. 2024)."

And we already did elaborate on this in the discussion: "In addition, the use of irreducible water content hinders the ability to estimate the volume of meltwater stored within a PFA. … While it is not possible to estimate the volume of meltwater stored within a PFA with the bucket method, the presence of year-round LWC in the firn can be used as an indication of PFA presence (Munneke et al. 2014). Brils et al. (2024) simulated aquifers in Greenland using IMAU-FDM, agreeing with 62 % of the observed aquifers by airborne and ground-penetrating radar measurements. The mismatch can be explained by. …."

The paragraph in lines 115-119 is very confusing. I don't understand the reasoning for why it is necessary to introduce a negative value to counterbalance the positive value. Also how are these values 'of the same order of magnitude'? In Figure 2a,b there are many points with LWC above 1000; however I would think the minimum possible value for the number of days without LWC is -365. When you say "we introduce a negative value…", what is meant by this? Do you average this value with the perennial LWC? Or is your model then predicting 2 separate target variables?

Thank you for these concerns. We agree that this explanation can be confusing and lacks some information, and therefore add more explanation about the rationale behind this approach. We rephrase this entire paragraph as follows: "Our target variable is the annual perennial LWC, which is defined as the minimum vertically integrated LWC over a year, based on model output at 10-day intervals. A yearly perennial LWC of zero indicates the absence of LWC at some point during a year, meaning there is no PFA. A drawback of this approach is that it does not differentiate between years with brief periods without LWC (e.g. 10 days) and years with long periods without LWC (e.g. 10 months). Since the target variable, the annual perennial LWC, cannot take negative values, this prevents negative values from counterbalancing positive biases in the predictions (e.g. an asymmetric distribution of the errors). To address this, we introduce negative values to represent the number of days without LWC during periods when PFAs are absent. This allows us to capture both the presence and absence of LWC in a unified way. For example, if LWC is absent for 10 days, the target value is -10; if LWC is absent for

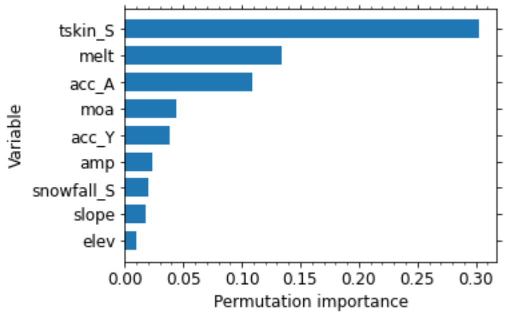
10 months, the target value is approximately -300. Although the negative and positive quantities are conceptually different, their absolute magnitudes are comparable in scale, which provides a relatively smooth transition around zero. To further clarify, this method does not involve averaging LWC values with negative values. Instead, the target variable is a single, unified measure that can take either positive values (indicating LWC amount) or negative values (indicating days without LWC). If the target variable exceeds zero, it is considered to indicate the presence of a PFA. "

Also, why can you not just predict the average LWC during the winter months, instead of the annual minimum? As winter is the least-likely month for liquid water to exist, maybe this would be a simpler way to predict PFA presence?
In some studies, the presence of LWC in winter or spring is indeed used to identify PFAs. However, in some regions winter melt events occur, so to minimize the risk of classifying this as a PFA, we use the minimum amount of LWC during a year. Also, if for one month/week there is no liquid water, it means the aquifer is not perennial. Therefore, this approach is a more correct way to identify PFAs.

Finally, from the target variable, what do you actually consider to be a PFA? This is unclear from the text. We clarify this with the following sentence: "If the target variable exceeds zero, it is considered to indicate the presence of a PFA."

Further, in section 4 (L184): It does not make sense to me to report the RMSE as 86 mm or days. I think I just don't understand the target variable but I believe it to be some combination of LWC and days without PFA presence it which case the RMSE would not strictly be 86mm or 86 days but some combination of the two?
We hope that our explanation to the comments above regarding the target variable helps to clarify this point.

Finally, a look into feature importance would be a very valuable addition to this work. Are there features that are more heavily utilized by the model than others for predicting LWC? Which features of the model are most important in cases of firn aquifer presence?
The list of permutation importance metrics is shown below. However, a part of our training dataset, especially for SSP1-2.6 and SSP2-4.5 represents conditions that are not conducive to aquifer formation (e.g. relatively cold regions). As a result, some features might have a high importance, just because they exclude those regions. For features like aquifer presence, we therefore think the permutation importance is not meaningful as it primarily shows which parameters are needed to predict the presence of a PFA, not its magnitude. This is in contrast to e.g. firn air content, which is non-zero for most conditions, and subsequently the feature importance highlight the parameters relevant for the FAC.



Minor and technical comments

Abstract

As you redefine all acronyms again later in the text, I think it would be better if the abstract did not contain acronyms. This will help make the abstract more clear and approachable.

We remove the acronyms from the abstract except for PFAs, and also redefine some of the major acronyms at the beginning of new sections.

L5 – "to approximate a firn model" I would recommend being more specific here since your emulator does not approximate the entire firn model but merely a part of it.

Indeed, we suggest rephrasing the following text: *"To overcome this, we developed an XGBoost emulator, a fast machine learning model, to approximate a firn model. The PFA emulator was trained with simulations from the firn densification model IMAU-FDM, forced by three emission scenarios (SSP1-2.6, SSP2-4.5 and SSP5-8.5) of the combined regional climate model (RCM) RACMO2.3p2 and general circulation model (GCM) CESM2."*

into: "To address this, we develop an XGBoost perennial firn aquifer emulator, a fast machine learning model that is trained on PFA output of simulations from the firn densification model IMAU-FDM. The firn simulations were forced by the combined regional climate model RACMO2.3p2 and general circulation model CESM2 for three emission scenarios (SSP1-2.6, SSP2-4.5 and SSP5-8.5)."

L11 – "For SSP5-8.5, PFAs expand to Ellsworth Land in West Antarctica and Enderby Land in East Antarctica" – This is only true for some cases of SSP5-8.5 right?

For completeness, we rephrase this as follows: "For SSP5-8.5, PFAs expand to Ellsworth Land in six out of the seven simulations and to Enderby Land in East Antarctica in five out of the seven simulations."

L12 – "For climatic forcings from RACMO and MAR, we find that liquid water input (melt and rain) and snow accumulation are good predictors for PFA occurrence." – How do you quantify this? I would recommend computing some feature importance metrics to quantify which features are the most important in your emulator.

To clarify, we rephrase the following text: "*For climatic forcings from RACMO and MAR, we find that liquid water input (melt and rain) and snow accumulation are good predictors for PFA occurrence. However, HIRHAM predicts considerably less surface melt and accumulation for a given air temperature than MAR and RACMO do, resulting in less realistic PFA predictions.*"

into: "The emulator results for RACMO and MAR agree on the snowmelt and accumulation conditions required for PFA formation. While these results for HIRHAM are slightly different, caused by different modelled relations between temperature, accumulation and melt compared to RACMO."

This is also rephrased in the conclusion section.

L14 – specify "air" temperature.

Done. Thank you for noticing this.

Introduction

L20 – "reduces their buttressing effect" → "reduces ice-shelf buttressing effect"

We have changed this accordingly.

L20 – "Another process to reduce the…" →"Another process that can potentially reduce the…".
I think the phrasing of this sentence should be modified a bit because the process of hydrofracture does not directly impact ice-shelf buttressing. Hydrofracture has more of an indirect effect because it can impact the ice-shelf volume (i.e by causing ice-shelf disintegration). This change in ice-shelf volume then reduces ice-shelf buttressing.
We agree and clarify this as follows:
"Another process that can potentially indirectly reduce the buttressing effect of ice shelves is melt pond-driven hydrofracturing. This mechanism contributes to ice-shelf disintegration thereby reducing the ice-shelf volume, which in turn reduces the ice-shelf buttressing. This process is expected to increase under future warming (Lai et al. 2020, van Wessem et al. 2023)"

L25 – Maybe it would be better to write this as: "Currently, 60% of the ice-shelf area buttresses upstream ice and…"
We agree and have adjusted this accordingly.

L28 – "Perennial firn aquifers…, potentially causing ice shelves to break up". Please cite Montgomery et al 2020 or the Firn on Ice Sheets review paper (https://doi.org/10.1038/s43017-023-00507-9) here.
Done, see our response to the Remark 1 of the Community comment.

L42 – Fix "climaticx"
Done.

L43 – I think the statement "The absence of ice shelves along most of the north-western AP coast also suggests that the combination of ice shelves and firn aquifers is not viable" is misleading… There is no evidence that suggests that the NW AP has no ice shelves because of PFAs (which is what is indicated by this statement). Instead, the atmospheric and oceanic conditions on the NW side of the AP contribute to the lack of ice shelves. It may be that the climatic conditions which make ice shelves unviable are the same as those which allow PFAs to form. I find this statement misrepresentative as currently written.
In hindsight, the lack of ice shelves could indeed also be caused by the oceanic conditions. Therefore, we decided to remove this statement.

L60 – add "properties" after "firn"
Done.

Methods
Some details in the IMAU-FDM firn modeling section are missing. What sort of spin-up do you do for the firn model? How thick are the firn layers in your model? Are they add steady state when you begin your simulations?
We add more details about the IMAU-FDM simulations: "An initial firn layer is obtained by looping over the forcing of the 1950-2000 reference period until the firn layer is in equilibrium with the surface climate. The model employs up to 3000 layers of 3 to 15 cm thickness, which represent the firn properties in a Lagrangian way."

L95 – "see next section" → "(see section 2.2.1)"
Done

L123 – "These are the most…" → "These input features describe the most…"

Done

L125 – "For the summer, we consider snowfall instead of snow accumulation due to the likely presence of evaporation along side sublimation, which has different implications for PFAs". I don't follow this reasoning for including DJF snowfall, but annual and MAM snow accumulation. Why does this need to be different for different periods of the year?
This statement was indeed not clearly formulated. Liquid and solid evaporation is largest during summer due to the higher atmospheric temperatures. However, we would like to include only how much the surface firn is replenished by new snow during the summer, as this could influence the overall densification of the firn layer and the evolution of PFAs. For the fall months and the annual averages, the net accumulation is deemed more relevant. We didn't test this, however, in detail. We reformulate this sentence to: "For summer, we consider snowfall instead of snow accumulation due to the larger magnitudes of liquid and solid evaporation. This is considered to better capture the summer replenishment of firn air content and hence pore space for PFAs."

L130 – How many total input features are there then for your model?
In total, the emulator requires 23 input features. 5, 10 and 30 year data for 7 atmospheric variables in combination with elevation and slope. We add: "This amounts to 23 input features in total."

Section 2.2.3 – Which parameter is optimized? R2? RMSE?
We optimize based on R^2, this is clarified as follows: "Firstly, we evaluate the performance using spatial blocking cross-validation, with R^2 as the scoring metric."

Emulator tuning and evaluation
L178 – This reference should be Table 2
Thank you for noticing this, done.

L179 – Do you also leave out the 30-year rain input feature?
Yes, to clarity this, we have replaced this: "30-year melt input feature" by: "30-year total annual liquid water input (surface melt plus rainfall) input feature"

L184 – I think the R2 value here should be 0.89.
Indeed, this has been updated.

L187 – What is meant by "the individual locations" in this sentence?
To clarify, we change this: "The emulator yields an R^2 value of 0.90 for the years with PFAs of the individual locations (Fig. 2c)."
Into: "The emulator yields an R^2 value of 0.90 for the years with PFAs of each grid point (Fig. 2c)."

L188 – "The performance is poorest on and around the Larsen C and D ice shelves…" This makes sense to me because you are asking the model to extrapolate to warmer climate conditions that are likely not seen during the training which utilizes lower emission scenarios.
These results are for the spatial blocking evaluation. Region 7 has a distinct climate in terms of its relatively high melt and rather "low" accumulation. To avoid this confusion, we add this, so the sentence will be changed into: "The performance, while using spatial blocking, is poorest on and around the Larsen C and D ice shelves…"

Results

L226 – "Notably, HIRHAM-EC-Earth predicts at least twice the initial (2015)…" Is ECEarth much warmer in the present climate? Higher precipitation?

Yes, this was already explained in L236 (Section 4.3 Climatic drivers): "For HIRHAM-EC-Earth, the initial temperature is high, related to the warm bias over Antarctica in EC-Earth3 (Boberg et al., 2022), which explains the high PFA extent at the start of the simulation." We add a reference to Section 4.2 at this point: "Notably, HIRHAM-EC-Earth predicts at least twice the initial (2015) PFA extent (72,000 km2) compared to the other simulations, related to the higher initial temperatures (see Sec. 4.3). "

L242 – "For example, on Larsen C ice shelf most aquifers are predicted by MAR…" What is meant by "most aquifers" here. "Most" compared to what?

To clarify we rephrase this: *"For example, on Larsen C ice shelf most aquifers are predicted by MAR, which is related to high surface melt and high accumulation at the foot of the mountains, which are absent in the other two models."*

As: "For example, on Larsen C ice shelf the emulator predicts more aquifers for MAR compared to RACMO and HIRHAM, which is related to high surface melt and high accumulation at the foot of the mountains, which are absent in the other two regional climate models."

Section 4.4 – Why are some firn aquifers transient? From the snow modeling output, what happens to make the aquifers disappear? We add a more physical explanation for this: "As Figure 4b shows, PFAs can also develop, shrink and subsequently disappear, henceforth referred to as transient PFAs. This shrinkage and disappearance occurs when the firn layer becomes too thin (i.e. the firn air too depleted) to host an aquifer."

L291 – "As future warming leads to increased melt accumulation, the emulator is expected to produce more accurate PFA predictions…" This seems highly speculative and I don't fully understand the logic behind this statement.

To clarify, we rephrase this: *"As future warming leads to increased melt and accumulation, the emulator is expected to produce more accurate PFA predictions, as more locations will shift into distinct PFA climate regimes rather than remaining in transitional states."*

As: "As future warming leads to increased melt and accumulation, the emulator is expected to produce more accurate PFA predictions compared to the contemporary estimates. This is because warming is likely to shift more locations into well-defined PFA climate regimes characterized by high melt and accumulation, where aquifer formation is relatively certain, rather than into regimes with lower melt and accumulation, where aquifer formation is less certain (Fig. 11)."

Discussion

L325 – CESM2 also has a high precipitation bias

This is now included, and another comment is added: "However, these biases do not necessarily persist in the downscaled RCM output (Veldhuijsen et al. 2024)."

L331 – "Thus, the importance of including PFAs when assessing the timing of ice-shelf vulnerability also decreases" Can you elaborate on this? I don't fully understand.

To clarify, we propose to change this: *"The duration of such transient PFA presence will decrease with stronger warming rates or lower accumulation rates. Thus, the importance of including PFAs when assessing the timing of ice-shelf vulnerability also decreases."*

Into: “The duration of such transient PFA presence will decrease with stronger warming rates or lower accumulation rates due to relatively quick firn air depletion. Thus, the difference in the timing of ice-shelf vulnerability because of PFAs and firn air depletion becomes smaller.”

Figures

Figure 2 c/d – The x and y-labels are “Years with PFAs” but the units are “mm”.
Changed.