Dear Reviewer #2,

Thank you very much for your thorough review and questioning of our manuscript. We are grateful for your constructive discussion, which helped us improve the manuscript a lot. We hope we can answer all your questions to your satisfaction.

Best wishes,

Viola Steidl and co-authors

This work applies a Physics-Informed Neural Network (PINN) as a data-driven tool to estimate ice thickness across glaciers located on Spitsbergen in Svalbard. A physics-based loss function used in training the PINN is designed to penalize solutions diverging from a modified form of mass conservation. Additional physics-inspired loss functions are added relating components of the surface velocity arising from deformation of the ice and sliding at the base. The neural network is also provided with a number of other data inputs, including surface velocity, surface slope, elevation, positional parameters, and values providing assumed relationships between surface and depth-averaged velocities. The authors explore their results using a cross-validation scheme designed to avoid problems with spatial correlation. PINNs have seen an increasing number of uses within glaciology in the past few years. Their intrinsic ability to mix together known physics with poorly calibrated constants and sparse and/or noisy measurements make them an appealing modeling tool for underdetermined problems. This work is novel in training a single PINN over an extremely large domain (effectively all of Spitsbergen) and in mixing a large number of physical constraints, physically-inspired constraints, and plausibly related data sources.

Two related challenges complicate the evaluation of results in this work. First, as with almost all ice-covered regions, direct measurements of ice thickness are sparse. Second, the dynamics of the glaciers are complex and poorly understood. In Svalbard, a number of complicating factors are at play:

1. Many glaciers are topographically constrained, making lateral drag important and complicating simplified models of ice dynamics.

2. Many glaciers are thought to be polythermal, often with a significant layer of temperate ice at the base overlain by cold ice (Sevestre et al., 2015)

3. Cold surface temperatures allow for the accumulation of thick firn layers with poorly constrained density (Pälli et al., 2017)

These complications are not unique to Svalbard, of course, but aspects of Svalbard's topography and geographic location make them especially notable here. The authors frame the cross validation results in a way that seems somewhat disappointing. I am perhaps more optimistic than the authors about the results. In particular, I think the evaluation of a physically-based model on a glacier where no ice thickness data was provided is an unfair assessment of the model. The PINN proposed in this work is something of a hybrid between a data-driven estimator and a PDE solver. These two types of tools would be accessed in different ways. Additional consideration of appropriate evaluation mechanisms is probably needed.

Architecturally, I think this work is very interesting. There is a novel fusion of physics-based, physics-inspired, and non-physical relationships at work here. Unfortunately, the lack of explainability and the lack of a good ground-truth data source make it difficult to see a path to the results presented here significantly updating our thinking about Svalbard's glaciers. Given this combination, I would

encourage the authors to consider leaning into exploring the design of the PINN by, for example, exploring the importance of the various input fields or designing an experiment to consider the use of different ice physics approximations within this framework.

Thank you for your assessment of our work. We agree that there is a lot of potential that can be explored in follow-up studies and you make a number of useful suggestions

However, we want to reiterate that this is a proof-of-concept (PoC) study to assess the viability of a PINN approach for a long-standing challenge in glaciology, and we use Svalbard as a test case, partly for the reasons you mention above related to the range of flow conditions and glacier geometries and partly because it is one of few areas with relatively good coverage from observations and other estimates. Indeed, we compare our solution with three others and show that it is not so easy to infer that one of those four is "preferable". Our intention is not to shed new light on Svalbard ice thickness but on the potential of physics-informed ML for this problem. We have made this more explicit and clearer in the Introduction to avoid any ambiguity about our aims and focus:

"*As a proof of concept, we include all non-surging glaciers in Spitsbergen, Barentsøya, and Edgeøya in Svalbard to show that it is possible to use a PINN architecture for an entire region.*"

We also repeat the statement in the Conclusion:

"*This serves as a proof of concept that physics-informed models can not only be applied to one single closed system but, together with auxiliary data, can make meaningful predictions for entire regions.*"

Specific comments below:

PINN

- It is not clear to me what the coordinate system is used to feed the network. Is it a standard projection? Are the coordinates consistent across all of Spitsbergen or are glaciers each on their own local coordinate system in some way?
  We apologize that this was not clear in the manuscript. The coordinates are consistent across the entire study region (now mentioned in the manuscript: "*The coordinates of the individual grids are all transformed to the same projection.*"), we used the EPSG:25832 projection (which is mentioned in the model configuration file in the linked github repository).

- Why the current set of inputs to the neural network? It is not obvious to me, for example, why the area of the glacier should be included. In general, it would be interesting to know how including each input impacts the results.
  In general, the features are added because they have some relation to the glacier ice thickness and are available through OGGM. It is left to the neural network to find relations between the input and the target.
  The area of the glacier is added to the input vector as area/volume scaling is a well established approach that has been used in the past to estimate thickness (e.g. Bahr et al., 1997).
  We agree that a detailed analysis of feature importance would be interesting for explainability of the results. Therefore, we conducted a quantitative analysis of the feature importance through SHAP (SHapley Additive exPlanations) using Captum's (Kokhlikyan et al., 2020) implementation of the DeepLIFT SHAP algorithm (Lundberg and Lee, 2017) and included this in the Appendix.

The framework explains feature contributions to the model prediction, usually for purely data-driven machine learning. It comes with some limitations: all features should ideally be independent of each other. This is clearly not the case for our features. Nevertheless, we used this framework as it is a standard approach often used in machine learning.

We conducted the SHAP analysis on all of the seven LOGO CV models. Figure 1 shows the mean absolute SHAP values: high SHAP values signify a high impact on the output of the model; low absolute values signify a low impact. For our PINN, the spatial coordinates are by far the most important. This is expected as they define the domain of our solution to the mass conservation PDE.

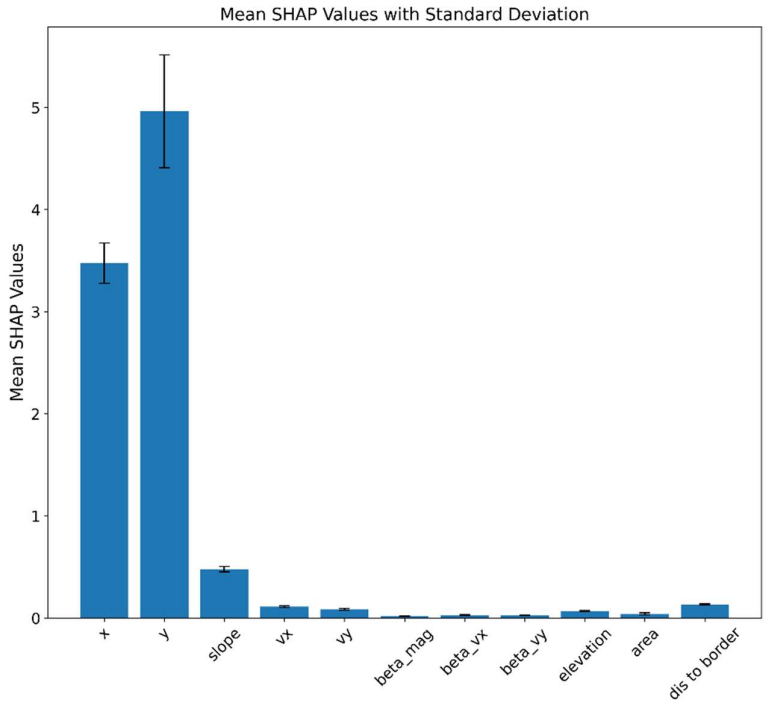


Figure 1 Mean of the approximated SHAP values for all seven LOGO CV m*odels.*

Besides the spatial coordinates the slope has the biggest impact on the prediction. Figure 2 provides a more detailed view of the impact of features. For every point in the dataset it shows how it impacts the predicted thickness.
The colour signifies the feature value of the individual data points: red signifies a relatively high value (within the range of the feature values), and blue signifies a relatively low value. For example, datapoints with low values for slope are more likely to lead to high values for the predicted ice thickness, while high values are less likely to impact the thickness prediction or rather decrease it.
This is what we would expect given that ice thickness and slope are indirectly proportionally related in the SIA: flat slopes lead to thicker ice.
The SHAP values for the distance-to-border feature tell us that the model thinks that at the border the ice thickness should be smaller than within the glacier.
For the surface velocity values the interpretation is less clear, also because we only see the component-wise features. High surface velocities do not seem to have much impact on the

ice thickness prediction, although, following glacier physics, they should have a strong influence on ice thickness.
Overall, the SHAP analysis is a good tool for deeper insights into the correlations between feature values and model prediction. However, we should be careful not to overestimate the explanatory power, especially when dealing with correlated features.
The analysis depends very much on the dataset (we chose the validation datasets to conduct the analysis). Therefore, the results can only show the impact of the features on the ice thickness prediction for our specific dataset and model setup. We can not derive universal feature importance from the analysis, let alone find causal relationships.

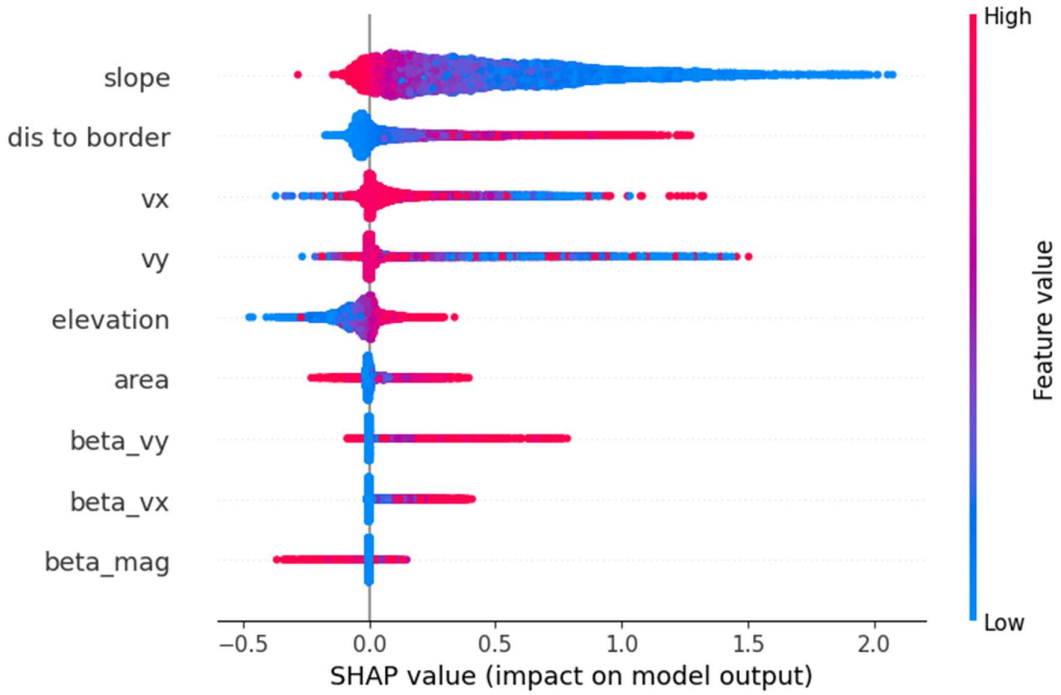


Figure 2 SHAP values for each datapoint by feature. The colour shows the relative value the feature takes for each datapoint.

Physical Model

- The way that deformation velocity, sliding velocity, depth-averaged velocity, and surface velocity are explained is somewhat confusing to me. My interpretation is that the authors are using a simplified physical model (Appendix B) to set a relationship between surface and depth averaged velocity, which you then qualitatively decide to loosen. Separately, they assume that sliding velocity and surface velocity are related by a pre-determined field. The network predicts deformation velocity only and evaluation of the mass conservation loss term is done by adding in sliding velocity according to the defined constant and the surface velocity.
  Thank you very much for your comment on this section. Reviewing the section we found that there was some confusion in it. Therefore, *we rewrote the whole Section 2.2*. The amount of basal sliding is introduced in the equation with the factor β that is estimated from the ratio of surface slope and surface velocity, following Millan et al., (2022).

If there is no basal sliding at all we estimate the depth-averaged velocity to be smaller than the surface velocity. We set a lower boundary assuming that depth-averaged velocity won't be any smaller than 70% of the surface velocity.
If the entire surface velocity would be due to basal sliding then the depth-averaged velocity is equal to the surface velocity. The estimate should reflect that the more basal sliding we have the closer the depth-averaged velocity will be to the surface velocity.

- What is the significance of the network outputting deformation velocity rather than directly depth-averaged velocity? Lines 61-62 seem to imply this is important, however it is not clear to me why. It seems to me that it is simply a choice between an extra calculation to compute mass balance and an extra calculation to compute the depth-averaged velocity bounds loss.
  We apologize for the confusion in this subsection. In the revised model we directly estimate depth-averaged velocity and updated the description of the calculation of the depth-averaged loss:

$$\mathcal{L}_{vel} = \begin{cases} (v_s - \bar{v})^2 & \text{if } \bar{v} > v_s \\ (v_s(l_{\text{lower}} + (1 - l_{\text{lower}})\beta) - \bar{v})^2 & \text{if } \bar{v} < v_s(l_{\text{lower}} + (1 - l_{\text{lower}})\beta) \\ 0 & \text{else} \end{cases} \quad \begin{array}{l} \text{with } \bar{v} \in \{\bar{v}_x, \bar{v}_y, \bar{v}_{\text{mag}}\} \\ \text{and } \beta \in \{\beta_x, \beta_y, \beta_{\text{mag}}\} \end{array}$$

- Lines 69-70 state that depth-averaged velocities are calculated for the x direction, y direction, and magnitude separately using different values of beta. Beta relates surface velocity to sliding velocity. In the simplified model of Appendix B, the sliding velocity must be in the same direction as the surface velocity, but different values of beta for x and y implies that the sliding velocity is in a different direction.
  We have 3 different β values, that are used separately in the component-wise/magnitude estimate of depth-averaged velocity. The β values are derived from the velocity in x- and y-direction and the magnitude separately, so it is ensured that the estimate of the amount of sliding is along the same direction as the surface velocity. To clarify this we added it to the description in Section 3.2:

  "*For each point, we compute three β values from the surface velocities in the x- and y-direction and the magnitude of the surface velocity.*"

- Apart from stating that ice is assumed to be incompressible (Line 50), I saw no mention of the effects of unknown density of snow and firn. To my understanding, glaciers in Svalbard may have significant firn layers (Pälli et al., 2017). This contributes to uncertainty in the radar measurements (as the dielectric permittivity is dependent on density) and impacts the implied mass flux. This source of uncertainty should at least be discussed.
  Thank you for mentioning that, we added it as a source of uncertainty in the discussion part.
  "Ice thickness measurements from ice-penetrating radar, for example, are subject to errors due to varying density of glacier ice but also due to unknown thickness of snow and firn layers (Lindbäck et al., 2018)."
  In general, we think that the uncertainties of the measurement data play a big role in the model performance and we emphasize that throughout the manuscript. Some of the in situ data points are also with a given uncertainty that in a follow-up would be interesting to include.

- In my view, the simplified ice dynamics of Appendix B may be insufficient for glaciers in Svalbard. I believe that the model selected ignores stresses from drag against the sidewalls, which seem significant for the topographically constrained glaciers on Svalbard. Additionally, assuming A to be constant with depth seems like a stretch. Many glaciers are suspected to be polythermal and this has been proposed as a mechanism for the surge behavior seen in Svalbard (Sevestre et al., 2015). While the authors have excluded currently surging glaciers, the presence of this phenomenon implies to me that depth-dependent temperature may be an important part of glacier dynamics in this region. At a minimum, further discussion of this point is needed.
  We agree that our model neglects a lot of physical processes and uses very simplified physical descriptions. You are right that our model does not specifically account for drag against the sidewalls. The estimate of the ice flow and velocities is not very restricted in the sense that we only set an upper and a lower limit for the depth-averaged velocity. Within those boundaries, the model can freely estimate the depth-averaged velocity.
  This is a design choice that we explicitly made to avoid choosing a physical model with parameters like the viscosity of ice that we cannot be certain about and would introduce new uncertainties.
  We added this to the Discussion section to better emphasize this point. Also, we included your point that A is actually temperature-dependent in the Discussion and not only in the derivation of the lower bound for the velocity estimate.
  "This approximation does not account for lateral drag and assumes the creep coefficient A to be temperature-independent. However, many glaciers in Svalbard are believed to be polythermal (Glasser, 2011). Therefore, the estimate of the depth-averaged velocity might have another source of uncertainty that is challenging to quantify."

OGGM-Processed Inputs

- Are any of the input fields that are processed with OGGM interpolated by OGGM in any way? If they are interpolated following a similar physical model to yours, does this introduce a circularity?
  Yes, the input collected through OGGM are interpolated to the grids of the glaciers. To our best knowledge, the source code of OGGM uses transformations that reproject and scale the data to the glacier grids but do not use a physical model. Mostly the data is reprojected using methods from rasterio or salem libraries. (Example for reprojection of the dh/dt data from Hugonnet et al., 2021: https://github.com/OGGM/oggm/blob/master/oggm/shop/hugonnet_maps.py#L12).
  We clarified that in the manuscript:
  "OGGM reprojects and scales the data for each glacier to the glacier grids. We collect these data and transform the coordinates from the individual grids into a common projection."

- I think it would be helpful to discuss how the surface mass balance input is derived. It sounds like a model-derived value? There are quite a few weather stations in Svalbard. Has the model been validated? How does it perform?
  You are right; the surface mass balance is modelled with OGGM's ConstantMassBalance model. The model is calibrated from geodetic mass balance measurements and computes mass balance according to the elevation of the data point (https://docs.oggm.org/en/stable/mass-balance-monthly.html). We have not validated the model on in situ data but already the OGGM documentation states that "more physical approaches are possible". Therefore, in the Discussion we mention that improving the estimate of the apparent mass balance calculation is one of the more important tasks to improve the performance of our PINN model.

Training and Evaluation

- The authors point out that data is highly correlated in space and thus they have used a cross-validation scheme based on leaving out an entire glacier at a time. I think that's a good approach to a challenging issue.
  Thank you.

- With the above said, however, I do wonder if this is an overly harsh method of evaluation. The effect is that, in looking at Table 2, we're looking at glaciers where no ice thickness data was available, greatly diminishing the value of the mass conservation approach. Another approach might be to leave in only the highest (elevation) 20% of the ice thickness data and explore how well the PINN can use mass conservation to extrapolate this downstream.
  We agree on the fact that it might be a harsh method to evaluate the model, but given that we are also predicting ice thicknesses for glaciers where no in situ thickness measurements are available it is crucial to know what the expected accuracy is on those glaciers. The LOGO cross validation, therefore, tells us that we should not be too certain about the predictions the PINN makes on any glacier without any given in situ measurements.
  Training on only ice thickness measurements above a certain altitude is an interesting approach to evaluating performance additionally. For generalizability, we think the LOGO CV is the most important.

- On Line 201, the authors state that the results suggest the model is overfitting. While this would be the conventional interpretation for a neural network, I think this is an overly critical interpretation for a PINN. Evaluating a PINN with no training data for the data loss function is sort of like evaluating a PDE solver with no boundary conditions. The analogy does not fully hold as the authors have also introduced some other inputs which can perhaps be used to guess at the ice thickness, but, in general, I think the authors may be too critical of their own results here.
  We agree that the task is challenging if there are no boundary conditions or measurements to constrain the model predictions more. However, we want to be clear that the model cannot yet predict ice thickness for glaciers without in situ measurements with the same accuracy as for glaciers where we provide ice thickness measurements.
  As you correctly mentioned, the model actually has other input features that it can use to learn the distribution of ice thicknesses. To us this seems very much like an overfitting problem where the model fits very well to the training data and does not generalize well to regions where the labels have not been in the training data.
  In theory the physics-aware loss components should take over in these areas. Therefore, improving this overfitting problem is a bit trickier than in purely-data driven machine learning models. On top of all the machine learning reasons that could lead to overfitting, there are also a couple of issues related to the physics-aware part of the model, like finding the optimal balance between loss components to enforce physical consistency that the noisy training data probably cannot even provide. Apart from that, we think 'overfitting' describes the situation quite well.

- Later (Line 177), the authors mention a random split between training and validation data. Given the aforementioned spatial correlation problem, how is this validation dataset used? Is it meaningfully independent of the training data?
  The random split is not meaningfully independent of the training data because of their spatial closeness to the training data. We clarified this in the manuscript by adding

"The training and validation data are spatially correlated. Therefore, the in-sample evaluation of the model probably overestimates its performance."
However, the split is useful to compare the results from the LOGO CV against each other and to the performance on the test glaciers. Since the in-sample performances do not significantly differ from model to model, we can be sure that the method is at least robust to leaving out thickness data of entire glaciers.

Interpretation and Applications

- It would be good to discuss the importance of ice thickness on Svalbard. This might depend on what you think your model is good at. For example, an improved estimate of total ice volume would be impactful for sea level rise projects. Improved fine-scale ice-free topography might have more relevance to projecting the evolution of specific glaciers that are relevant to local communities.
  We think that, since our work is more on the methodological side of estimating ice thickness, discussing the importance of ice thickness on Svalbard's glaciers would shift the focus too much away from the core of the work. Please see our response to your first comment above. In addition, our paper has been submitted to TC where we believe readers will understand why ice thickness is important. Nonetheless, we have added the following sentences to the intro which we believe is a strong justification:
  "Ice thickness is the single most important input for modelling the dynamics of an ice mass because surface velocity is proportional to the fourth power of thickness. Combined with surface elevation, it provides bed topography, also key for modelling flow."

  Once the method produces robust and reliable results for unseen glaciers, it could help both in improving the estimate of total ice volume and also the fine-scale ice-free topography (as we can choose the grid resolution as fine as we need them).

- I would like to see discussion of what components of the inputs and loss function are most important. Many applications of PINNs largely use them as tools for solving PDEs where constraints, regularizations, or boundary conditions do not easily fit in conventional solvers. This work goes beyond that, feeding in multiple layers of data that is not directly incorporated into a physics-based loss term. This, of course, raises the question of which parts are most informative. A careful set of experiments exploring this would be very interesting.

  We agree that the importance of the individual loss components is also interesting to quantify, just as the importance of the loss components (already discussed earlier). However, it is tricky to evaluate, as we are dealing with unevenly distributed, correlated, noisy in situ measurements as labels to evaluate the PINN performance.
  Despite this, we ran experiments in which we set the weight of each of the loss components to zero one after another. We then compared the scores to the scores of the reported model. To do so we calculated a relative RMSD as relative RMSD = $RMSD_{reported}$-RMSD/$RMSD_{reported}$. The relative RMSD will be positive if the score improves, and negative if the score gets worse by setting the weight of a loss component to 0.
  The relative differences vary among the scores for the test glaciers but are below 5% on average, as shown in Table 1. However, it is interesting to see that, while the scores on the **in-sample** validation data improve on average when switching off the physics-aware loss components (see Table 2), the scores on the **out-of-sample** test glaciers get worse on average.
  This matches our intuition that the model is overfitting on the in situ ice thickness data that we provide it with during training. The physics-aware loss components act like a regularization while

demanding physical consistency. From this experiment, it looks like the loss to bound the estimate of the depth-averaged velocity is the most important component.
This intuitively makes sense as a wrong estimate of depth-averaged velocity directly influences the mass conservation loss as well. With corrupted depth-averaged velocities, the mass conservation loss will not be able to enforce physical consistency.

Nevertheless, we want to emphasize again that the configuration of the loss weights is certainly not optimal, as we discussed in Section 5.3, so there might be another distribution of importance if all the loss components are better balanced. Also, as already mentioned, this experiment depends a lot on the dataset, so the importance of loss components also only applies to this specific study.
We added this Discussion to the Appendix.
For the Discussion of the importance of input features, please refer to our answer to your earlier question in the Section "PINN".

Relative RMSD for the scores from the out-of-sample test glaciers of the seven models (no in situ ice thicknesses for those glaciers were in the training dataset)

Relative RMSD For test glaciers	No MC	Vel_loss	Smoothness	Negative thickness
RGI60-07.00240	-0,091	-0,195	-0,013	-0,013
RGI60-07.00344	-0,018	-0,036	0,000	0,018
RGI60-07.00496	-0,132	-0,184	0,105	-0,053
RGI60-07.00497	0,000	-0,087	0,000	-0,022
RGI60-07.01100	0,025	-0,125	-0,150	0,000
RGI60-07.01481	0,000	0,072	-0,157	-0,024
RGI60-07.01482	0,008	0,032	-0,048	-0,048
Relative RMSD Mean	**-0,030**	**-0,075**	**-0,038**	**-0,020**

Table 1 Test glacier relative RMSD

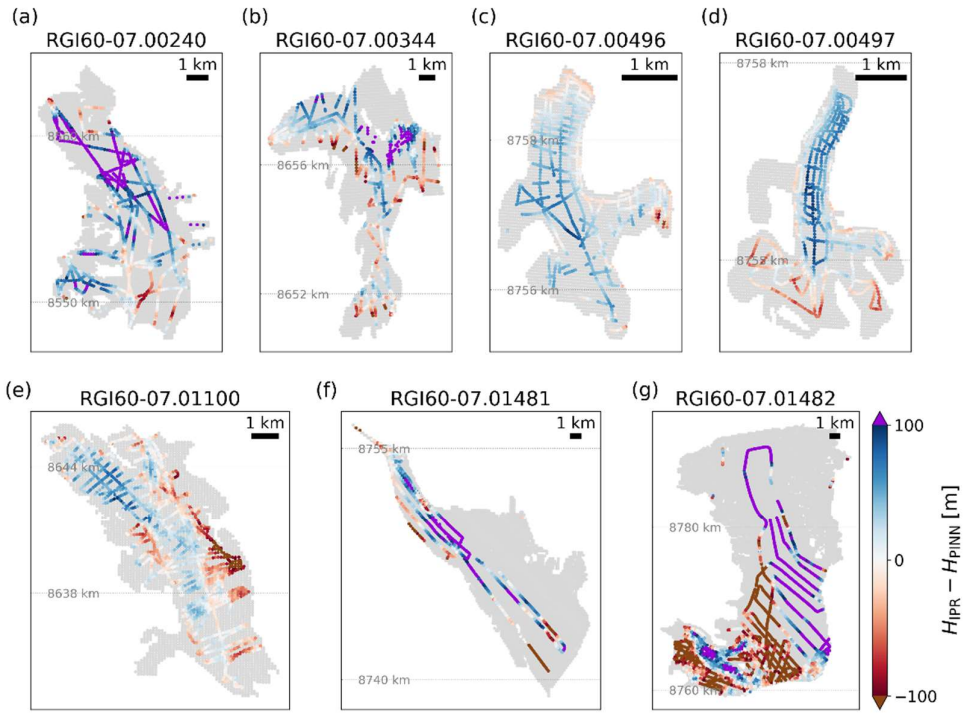
Relative RMSD for the scores from the in-sample validation data for the seven models

Relative RMSD In-sample Val	No MC	Vel_loss	Smoothness	Negative thickness
RGI60-07.00240	-0,033	0,000	-0,033	-0,033
RGI60-07.00344	0,032	0,032	0,065	0,000
RGI60-07.00496	0,000	0,030	0,061	0,000
RGI60-07.00497	0,033	0,000	0,033	0,000
RGI60-07.01100	0,032	0,032	0,032	0,000
RGI60-07.01481	0,000	0,033	0,033	-0,033
RGI60-07.01482	0,034	0,034	0,034	0,000
Relative RMSD Mean	**0,014**	**0,023**	**0,032**	**-0,010**

Table 2 In-sample validation relative RMSD

Typos and minor corrections

- Line 10-11 - Ice flux is determined by more than simply ice thickness and surface slope under real world conditions. This should be clarified to not suggest that those two variables alone are sufficient.
  Thanks for the remark, we meant to emphasize that ice thickness is most important to reliably **model** ice dynamics. Therefore, we changed the sentence to
  "Glacier ice thickness is a fundamental variable required for modelling the evolution of a glacier."
- Line 69 - bracket is the wrong way around
  The bracket opening to the left should signal that 0 is outside the interval as it is not a possible value for the parameter. The section was rewritten without the bracket now.
- Figure 4 - Are the color scales saturating? If so, it would be good to show the clipping in a different color so we can see where the error exceeds +/- 100 m.
  Thanks for the remark; we updated the figure:



- In Table 2, comparing the first glacier's performance in-sample versus LOGO, the RMSD more than doubles while the MAPD decreases. Is this correct?
  Yes, this is correct. This is because the glacier is one of the thicker glaciers. Therefore, a high RMSD might not directly lead to a high MAPD as the MAPD is the error relative to the value of the true ice thickness.

I enjoyed reading this work and believe it to be a promising avenue. I hope that these comments can help improve this manuscript.

Thank you again for all your comments, we think it greatly helped to improve the manuscript.

Bahr, D. B., Meier, M. F., and Peckham, S. D.: The physical basis of glacier volume-area scaling, J. Geophys. Res. Solid Earth, 102, 20355–20362, https://doi.org/10.1029/97JB01696, 1997.

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., Farinotti, D., Huss, M., Dussaillant, I., Brun, F., and Kääb, A.: Accelerated global glacier mass loss in the early twenty-first century, Nature, 592, 726–731, https://doi.org/10.1038/s41586-021-03436-z, 2021.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch, https://doi.org/10.48550/arXiv.2009.07896, 16 September 2020.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, 2017.

Millan, R., Mouginot, J., Rabatel, A., and Morlighem, M.: Ice velocity and thickness of the world's glaciers, Nat. Geosci., 15, 124–129, https://doi.org/10.1038/s41561-021-00885-z, 2022.