Dear Ben,

Thank you again for giving us the opportunity to improve our manuscript. This letter combines our responses to the three reviewers' comments. As with last time, the comments are in black font, and our answers are in blue font.

We sincerely hope that the revised version meets your and the reviewers' expectations, and we look forward to your response.

Best wishes,

Viola Steidl and co-authors

Point-by-point responses to comments from Reviewer 2

Dear Reviewer #2,

Thank you for your comments and suggestions. We hope we can answer all your remarks and concerns to your satisfaction.

Thank you for helping us improve the manuscript.

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. 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 a heterogeneous domain consisting of multiple glaciers (effectively all of Spitsbergen) and in mixing physics-informed methods with purely data-driven methods.

The authors position their work as a proof of concept, largely demonstrating the ability to scale PINN-based methods to larger domains by mixing in non-physics-informed methods to produce continuous estimates in regions where no suitable boundary conditions exist. In my view, the manuscript is much improved and offers a constructive contribution to the ongoing challenges of building useful models from sparse in-situ data. The availability of the authors' code and quality documentation adds to the value of the work.

Framing of the work

- In your response to the last round of reviewer comments, you made clear that you view your work as primarily a proof of concept. In that case, I'd suggest making clear in both the abstract and the introduction that your purpose with this article is to demonstrate a new technical method.

Thank you for the comment. We added that this study serves as a proof of concept to the abstract, introduction and repeat it in the conclusion as well.

- Viewed as a technical proof of concept, I see what sets your work apart as primarily (a) the application of physics-informed techniques to a heterogenous spatial area spanning multiple catchments enabled by (b) a fusion of physics-informed and purely data-driven loss functions. If this is consistent with your view, I would suggest a few changes in your literature review:
  - Since your work fuses data-driven and physics-driven approaches, consider briefly reviewing non-physics-informed machine learning applications. Some that you might consider including:
    - Ice sheet scale thickness estimates: Leong and Horgan, 2020.
    - Glacier-scale SMB: Bolibar et al., 2019
    - Glacier-scale thickness: Haq et al., 2021

Thank you for the suggestions. We added Leong and Horgan to our literature review. Bolibar's work on SMB and Haq's work on ice thickness are both already in there.

  - In addition to reviewing the Teisberg et al., 2021 application of PINNs to ice thickness mapping, also consider incorporating variational inference-based approaches such as Brinkerhoff et al., 2016.

Thank you for the suggestions. However, we are not sure if this would add nicely to our line of arguments, leading from the popular ice thickness maps from process-based models over to machine learning to model glaciers to physics-aware machine learning and its application to glaciers, and finally for ice thickness prediction. We would like to keep this introduction a bit slimmer and therefore, decided not to include this reference.

  - I'm not sure I follow the comment about "without further consideration of bed properties" - perhaps expand on this if it's a key difference.

Agreed, this is not a key difference we want to focus on so we took this out.

  - In framing your work as covering "an entire region," I think it's important to note that the significance of this is the heterogeneity of the region in question due to it being composed of multiple glaciers separate catchments. Notably, your domain is of roughly comparable size to the Rutford Ice Stream, Byrd Glacier, and the Amery Ice Shelf, the subjects of three of the works you cite. This is not to take away from what you are doing. It's just important to note that the significance is about the heterogeneity more than the pure spatial scale.

Thank you for your comment that this can be misunderstood. We changed the sentence to "… for a heterogenous region."

Physical model

  - Section 2.2 is much clearer now.
  - Line 79: I don't think that SIA necessarily implies neglecting the temperature dependence of the rate factor. See, for example, Larour et al., 2012. I suggest separating the constant A part out as a separate assumption you are adding.

Agreed, we divided this into two sentences.

- Equation 3 feels a bit abstract without knowing the selected values of l_lower and beta. Since these are specified later in the text, I suggest specifying their value/ranges briefly here (and referring readers to the appropriate sections below for details).

Thank you for your comment. As you said the values are specified in the following sentence so we would argue, there is no significant improvement in stating them the sentence before.

PINN Evaluation

- This is an extremely tricky issue, as there are not yet well established metrics for evaluating PINNs. This application is ever more challenging due to the combination of physics-informed and non-physics-informed loss functions. In general, the authors are doing a good job of discussing these nuances.
- I am not yet convinced by the analysis of the leave-one-out results and, in particular, in the overfitting conclusion. I would suggest evaluating the physics-informed loss function alone on the results of each of the 7 test glaciers when its ice thickness data was and wasn't included in the training (when it was or wasn't the left-out glacier, that is).

  Depending on the results of this…

  - Case 1: The physics-informed loss is about the same for each glacier with the thickness data left out or not. In this case, it seems that the model has found two different thickness maps which are roughly equally physically-plausible. This would be a very interesting result. I would argue this does not suggest "overfitting". The obvious follow-on question is: if you add in one point (or some very small amount of ice thickness data), does your model now predict the right thickness map?
  - Case 2: The physics-informed loss is much higher with the thickness data left out. In this case, I would agree with your "overfitting" assessment, but it's interesting that your model is not finding a physically-plausible result without being guided by a few thickness measurements.

Thank you very much for this comment. This is a good suggestion but it is not as easily implemented. We would have to filter for the specific glacier during the training which would add a substantial computational overhead to the algorithm. Leaving out the ice thickness measurements for the glacier is easier as this is done in data preparation before the training starts. To filter for a specific glacier during the training we would need to add the RGI ID to every training data point. But we definitely agree that this would be helpful to get deeper insights in the training mechanisms.

Additional minor comments

- Line 12: I find the opening claim that surface velocity is proportional to the fourth power of ice thickness to be potentially confusing. I assume you are referring to the paragraph following Eq. 8.36 in Cuffey and Paterson, which says that surface velocity is proportional to H^4 * alpha^3. Notably, however, ice thickness is inversely proportional to surface slope (Eq. 8.9). I would recommend softening the claim that ice thickness is the single most important input and citing a specific chapter of Cuffey and Paterson (8, I assume) to make it easier for readers to find the derivation you reference.
  Thank you for the remark, we agree that this might be a bit too strong of a claim and therefore softented it

- Line 13: The importance of bed topography also needs a citation.
  Thanks for the remark, we added (van der Veen, 2013) as citation.
- Line 72: “encounter” doesn’t make sense to me here. Perhaps you meant “counter” or “balance”?
  Agreed and adjusted.
- Line 78: “lamellar” -> “laminar”
  We went with the convention to use lamellar to describe glacier flow as in (van der Veen, 2013). Therefore, we would like to leave it as is.
- Line 111: Missing period after “Appendix B”
  Thanks, we corrected this.
- Line 126 / Eq. 7: Equation 7 is not, in my opinion, a physics-aware constraint. I would argue that it is a heuristic smoothing regularization.
  We agree that this can also be seen as a regularization as it is also done in non-physics-aware machine learning architectures. However, here it only applies to one of the model outputs instead of all of them and enhances the physical correctness.
- Line 178: The units of the ratio here are a little confusing. I think the units here are (m/yr)/(m/m), which simplifies to m/yr. I recognize that is the same as yr^-1*m, however I would suggest either writing out the full un-simplified form or sticking to the more conventional m*yr^-1.
  Thank you very much for the remark. Actually, it should be yr*m^-1, we corrected this.

van der Veen, C. J.: Fundamentals of Glacier Dynamics, Second edition., CRC Press, 2013.

Point-by-point responses to comments from Reviewer 3

Dear Reviewer #3,

We highly appreciate your thorough review of our manuscript and your suggestions to improve it. We hope we met your expectations with our revisions and answers to your comments.

Thank you very much and best wishes,

Viola Steidl and co-authors

General comments

The manuscript "Physics-aware Machine Learning for Glacier Ice Thickness Estimation: A Case Study for Svalbard" presents an interesting approach for reconstructing spatially-continuous glacier ice thickness from sparse in situ measurements with a neural network. The sparsity of the ice thickness measurements is mitigated by incorporating physical knowledge into the objective function, which acts as a regularizer on the neural network predictions. The authors demonstrate that the their method produces thickness estimates that are consistent with previous studies using physical models of ice flow.

After the first round of revisions, I believe that the authors have done a good job in addressing much of the concerns regarding the methodological details and factors that impact the accuracy of the thickness estimates. I particularly liked the inclusion of the SHAP analysis and the systematic analysis of the physics-aware loss functions.

However, based on the earlier reviews, it appears that a key remaining issue holding back this work is the lack of ground truth and the limited validation via comparison with previous ice thickness estimates. I would agree with the other reviewers that a synthetic experiment would be very beneficial in instilling confidence about the thickness estimates, especially since it appears that the estimates are systematically lower than the estimates from the previous approaches (Figure 5). While the authors state in their response that such an experiment would be a substantial effort, I believe that even simplified synthetics would go a long way towards clarifying the performance of the neural network.

Concretely, I would suggest that the authors generate 1D synthetics with a Blatter-Pattyn model with some random bed topography profiles. This can be done for a relatively small number of synthetic glaciers (~20) such that the number of velocity points is on the order of a few tens of thousands. A small fraction of ice thickness values can then be used as the auxiliary data. For the neural network features, since these are 1D, the features can be limited to the x-coordinate, slope, vx, and beta (and perhaps elevation if the modeled mass balance is elevation-dependent). These features (as well as the auxiliary mass balance) can be subjected to various amounts of noise. After the neural network is trained, then the thickness can be evaluated vs. the true thickness. Crucially, the thickness can also be evaluated against the reconstructed thickness using the approach of a model like Millan et al.'s. It may turn out that the previous approaches overestimate the thickness, which would be important to know. Overall, while I understand that such an experiment is not trivial by any means, it would really help increase the impact of this work and potentially highlight its technical advantages relative to previous approaches.

Thank you for this recommendation and also the suggestion on how to build a synthetic dataset. We decided to go for a simple approach, synthesizing data for a single glacier: We assumed basal stress in the direction of flow as the only stress component:

$$\tau_b = \rho * g * H * \alpha$$

and generated the surface velocity from

$$u_s = u_b + \frac{2A}{n+1}\tau_b^n H \quad .$$

The basal sliding velocity is set to zero for simplicity and the apparent mass balance *b* is calculated from

$$\frac{\partial H}{\partial t} + \nabla \cdot (\bar{v} H) = \dot{b} \quad .$$

Figure 1 shows the model results when training with and without physics-aware losses. The model performs significantly better if trained with physics-aware losses.

We did not evaluate Millan's model on our synthetic dataset as we felt this would exceed the scope of the manuscript. To us, this experiment serves as proof that the concept of using a physics-aware model delivers physically consistent ice thickness predictions, especially when compared to a machine-learning model that is not constrained by physics-aware losses. However, we would not like to draw any conclusions on whether another product is overestimating or underestimating ice thickness.

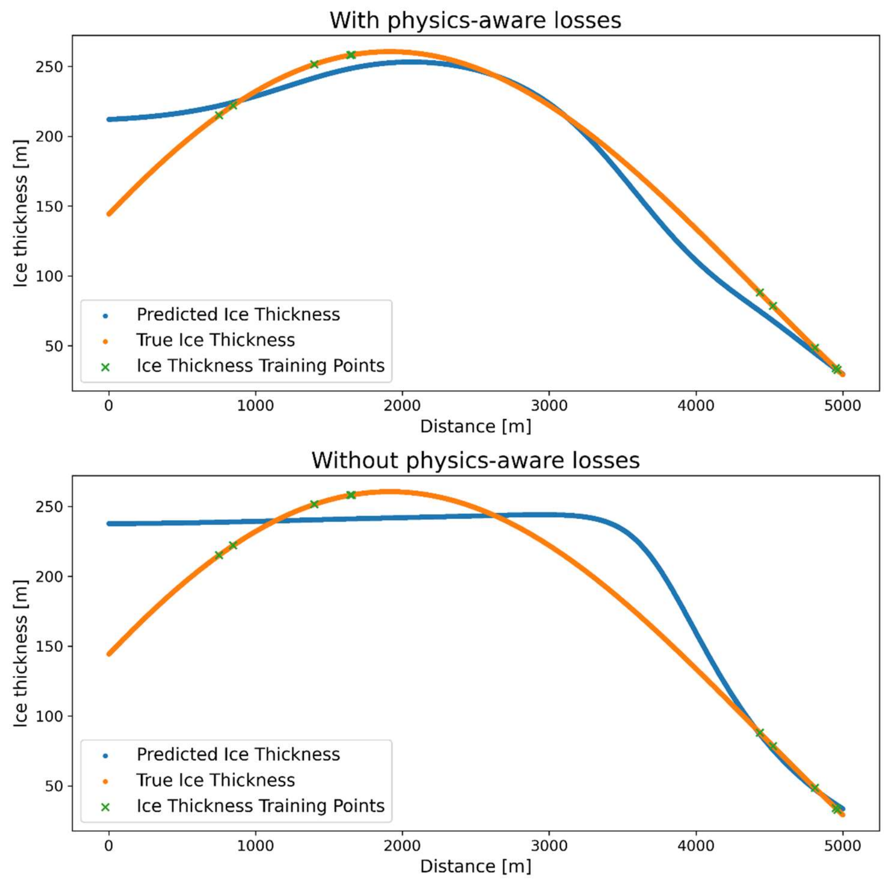

Figure 1 Predictions of models trained with and without physics-aware losses on 1D synthetic data

Other than the synthetics suggestion, I see no other major issues with this work. I've listed more specific comments below.

Additional specific comments

- The Fourier dimension and scale adds two more hyperparameters to the neural network design. How do these affect the performance of the reconstructions, and how are they chosen in this work?

We selected the hyperparameters for the Fourier embedding based on a series of non-exhaustive exploratory experiments rather than a full grid search to avoid computational costs. As described in (Tancik et al., 2020) choosing too many Fourier modes or the Fourier scale too big the predictions will be very grainy, while choosing them too small will not lead to the improvement of the mass conservation loss that we described in the answer letter to Reviewer #1.

- In Reviewer 1's comments, it was suggested to move the surface velocity data to the network outputs. Then, the data can be used as additional "auxiliary data" for constraining the predictions of the surface velocity. I think this is a good idea as it would provide a way to smooth both the surface and depth-averaged velocities in a consistent manner. This approach could also be used to predict mass balance, constrained by more accurate in situ measurements. Thus, I think the authors should at least mention this as possible follow-up work.

If we understand correctly you suggest also predicting the surface velocity. We agree that this is a valid approach, which has also been followed by (Teisberg et al., 2021). We constructed the model in this way to explore a different approach, where we have the auxiliary data together with the physical constraints. Therefore, we wanted to provide the model with as many input fields as possible.

However, we agree that predicting also the surface velocity could lead to improved training of the model weights and is a configuration that could be explored in future work.

- I'm still confused why there should be three different beta values for the different velocity components. It seems to me that beta should be calculated from the velocity magnitudes, and then the same beta is used for all three components. This would keep the partitioning consistent.

We calculate three components for beta because we do a component-wise estimate of the depth-averaged velocity. Since the influence of basal sliding on the surface velocity might also be different depending on the direction we chose to simply calculate a beta value for the two components and the magnitude. To us this would be the most consistent way.

- Line 101: It's probably better to say "calculate the terms in the PDE" since there may be non-derivatives or products of derivatives depending on the PDE

Agreed and changed.

- Figure 1: In the caption, please add a little bit more description on the inputs, outputs, and losses here. What do the colors in the loss boxes correspond to?
Agreed and improved.

- Line 130: Why not just predict log(H)

Thank you for this remark. This is actually another possibility of ensuring positive ice thickness that should be considered in future work.

- Line 153: I believe the use of the term "boundary condition" was changed to something like "internal condition" when referring to the ice thickness data, so this needs to be made consistent.

Thank you for the remark, we changed the wording to be consistent.

- Line 167: Does the varying grid resolution have an effect on generalization?

The grid resolution should not play a role in a PINN, but to be entirely sure, the LOGO CV could be run on a dataset where the grid resolution is kept the same for all the glaciers. We chose the varying grid resolutions to make sure that also smaller glaciers are represented well in the dataset with a big enough number of points.

- Line 175: Please state the size of the filter used for Gaussian smoothing (perhaps in units of fraction of average ice thickness)

The size of the smoothing filter depends on the grid resolution of the glaciers. From OGGM the smoothing window is set to 251/grid_resolution. We feel like this is a very technical detail of the data preparation that is set by OGGM and would not add to the clearness of the manuscript. Therefore, we would not like to include it in the main text of the manuscript. As all our code is publicly available, it is still well documented.

- Line 176: It's probably more accurate to say that β is introduced to incorporate the effect of basal sliding on the measured surface velocity (instead of estimate)

Agreed and changed.

- Line 189: Also here, please state the size of the filter

Please see our answer above.

- Line 234: Since the in-sample and LOGO RMSD scores are significantly different, it would be useful to show a plot (probably in the Appendix) of the training performance (loss vs. training epoch) for the train and validation data. This way, we can get a better sense on the amount of over-fitting.

Figure 2 shows the mean loss curves for the data loss and the RMSD during the LOGO CV (validation RMSD is evaluated every 10 epochs only). The last point of the validation loss is the evaluation on the left-out test glacier. This plot again shows how important it is to have the LOGO CV as a random split for training and validation data would lead us to severely overestimate the performance of the model. All the ice thickness measurements are close to each other, so the model has no difficulty interpolating from one measurement point to the next. The overfitting can only be seen through the lens of a spatial split like we did it in the LOGO CV.

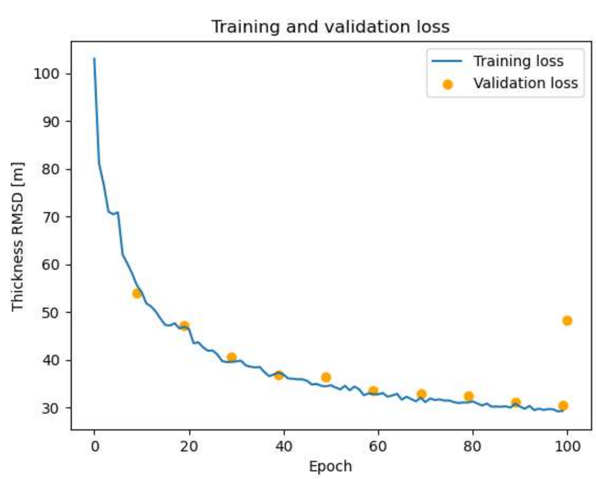

Figure 2 Thickness data training and validation loss

- Line 289: It seems like it would be good to show a comparison of velocity fields between ~2010 and 2017-18 (if such data are available).

We agree that this would be interesting to see how much of an influence the different measurement dates of velocity and ice thickness can potentially have. However, extensive velocity data from earlier periods is hardly available due to high cloud coverage in this region. NASA's MEaSUREs project (https://its-live.jpl.nasa.gov/#access), for example, collects ice velocities for every year, but for the glaciers of Svalbard, there are only very few datapoints for years before 2015. However, since glaciers are melting rapidly (Hugonnet et al., 2021) we assume that the surface velocities show a significant change compared to 2010.

- Line 298: Even though the loss weights are held fixed, you should briefly discuss how these weights affect the final thickness estimates (or perhaps show an L-curve for the most important weight).

Thank you for the suggestion. We believe that with the study of the importance of physics-aware loss components in the Appendix, we have already explained how the weights of the loss components influence the predictive performance. Also, because the work is thought to be a proof-of-concept, we think that going more in-depth to explore the parameters of the model would exceed the scope of the manuscript.

- Figure 6: I would like to see one or two additional sentences in the caption that summarize the challenges or highlight the challenges that are most consequential for the proposed method.

We agree and added our key challenge to the caption.

- Line 355: This could be straightforwardly implemented by using thickness uncertainties as weights to the loss function

We agree that this would be one way to implement it. However, the uncertainty of the measurement is not added for every data point in GlaThiDa. Therefore, retrieving measurement uncertainty and annotating every ice thickness label would still be a great effort.

- Line 375-380: This paragraph is a bit too general and doesn't really add to the rest of the manuscript. Physics-aware machine learning has been around for quite some time now and has been applied to a wide variety of geophysical problems.

Thanks you for this comment, we agree that it might sound a bit general, so we altered it a bit. However, we want to encourage the readers to also take this work as an inspiration for other problems that might be solved with physics-aware machine learning. Therefore, we also provide the complete code of the model, to make the work not only reproducible but also reusable.

- Line 415: What about the batch size?

Thanks for the remark. The batch size is 8192 and we added it now.

- Appendix D title: Please state up front what metric you are measuring importance against (validation accuracy?)

Thanks again for the remark. Yes, it is measured on the validation set. We added the information to the manuscript.

Point-by-point responses to comments from Reviewer 4

Dear Reviewer #4,

Thank you for reviewing the manuscript, your comments and suggestions. We hope we can answer all your questions and concerns to your satisfaction.

Best wishes,

Viola Steidl and co-authors

The manuscript presents the development of a new physics-informed neural network (PINN) method to infer glacial ice thickness, based on a case study of Svalbard. While the description of the method in the initial manuscript was somewhat rough, the revised manuscript shows significant improvement. The other reviewers have addressed most of my questions, and the authors provided detailed responses. Below, I list some additional questions specific to the PINN method. I recommend the publication of this paper if the authors can address these points.

Questions:

1. The authors employ Fourier feature encoding within the network, which requires setting a hyperparameter B. From Tancik et al. (2020), we know that an incorrect choice of B can lead to overfitting in the network's predictions. Could the authors provide more details on how they determined the optimal value for B?

We agree that the choice of matrix B should be explained. We draw the entries for B from a gaussian distribution. The size of B and the standard deviation of the gaussian distribution are hyperparameters of the model. We selected the parameters from a non-exhaustive search, so they might not be optimal. The shape of B is [2,32] and the standard deviation is chosen as 10.0. We added this information to Appendix B.

2. On one hand, the authors apply Fourier feature encoding to capture high-frequency features in the output; on the other hand, they include a smoothness loss term in the loss function to regulate high derivatives. However, high-frequency functions typically exhibit high local derivatives, which appears somewhat contradictory to the use of a smoothness loss. Could the authors provide an explanation to clarify and justify these settings?

We agree that this seems to be counterintuitive. We do not have a rigorous explanation of this behaviour. Empirically, the Fourier feature encoding leads to an improved optimization of the mass conservation loss, as discussed in the response to Reviewer #1. The smoothness loss solely serves to fine-tune the training. There could well be a better configuration with fewer modes in the Fourier Feature encoding that does not need an additional smoothness loss.

3. In Appendix B (line 414), the authors state that they set the loss weights λiλi to ensure all loss terms in the loss function are of the same order of magnitude. However, if the smoothness loss term is kept at the same magnitude as others, could this result in an over-smoothed thickness prediction?

Yes, if the smoothness loss is weighted too much, we assume we would have an oversmoothed prediction. However, in our case, the setting of the loss weights does not seem to lead to oversmoothed predictions.

4. Based on Figure 2, it appears that surface velocity and β-values are used solely as input features. However, both β and surface velocity also appear in the velocity loss term L_{vel} (Equation 6). It may be helpful to update Figure 2 to indicate that β-value and surface velocity are part of the data required for the loss function.

We agree that these connections should appear in the figure. As we would like to keep the figure clean, we decided to add this information to the caption.

5. As the authors mention, inferring glacial ice thickness is a highly under-constrained problem. Additionally, the sparsity of the measured thickness data complicates model validation. This may be beyond the scope of the current study, but it would be interesting if the authors could generate synthetic data to validate the robustness of the proposed PINN model.

We agree. We generated a synthetic 1D dataset to test the model performance on this dataset, as was suggested by Reviewer #3. Please refer to our answer letter to Reviewer #3 for a description of the outcome of this experiment.

6. Did the authors conduct experiments with different weight initializations or network structures to assess whether the PINN training converges to a unique thickness inference? If so, could they provide the standard deviation of these tests?

We did not conduct experiments with different weight initializations but we conducted a non-exhaustive search for the best model parameters considering different numbers of layers within the network, for example. We did not measure the standard deviation of the thickness predictions.

7. Given that the authors used various input features in the training, I suggest adding more detail on the input layers in Appendix B to help readers better understand the network structure used.

The input layer for the model is a linear layer with 256 neurons just like the other layers in the model. Therefore, there is no specific description of the input layer.

List of relevant changes

L. 7: added “proof-of-concept” to describe the experiments of the manuscript

L. 10: deleted “single” to soften the claim that ice thickness is an important input for modelling glacier dynamics

L. 13: added citation

L. 23: added citation

L.41: deleted “without further consideration of bed properties” as this detail was not adding to the clarity of the manuscript

L. 46: exchanged “entire” for “heterogenous” to better describe the achievement of this study

L. 73: exchanged “encounter” with “balance”

Figure 1: extended the caption with the sentences: “The physics-aware losses are in purple boxes. The Data loss in the blue box is the only loss depending on ice thickness measurement data. Surface velocity and β-values also add to the physics-aware losses. The connection is not shown to increase readability.”

L. 133: exchanged “boundary condition” for “internal condition”

L. 141: added: “We tested a slim version of the PINN model on a one-dimensional data set of a single glacier. The results are given in App. C. The experiment shows the added value of introducing physics-aware loss components.”

L. 155: exchanged “boundary condition” for “internal condition”

L. 179exchanged “estimate” for “incorporate”

Figure 6: extended the caption with: “Challenges for PINNs in a real-world setting like the prediction of glacier ice thickness. The separate realms are interfering with each other, complicating the optimization of the model. Weighting of the physical constraints could have the biggest positive benefit.”

L. 380: exchanged “physical law or condition provides a strong constraint” with “ physical law and multiple conditions provide constraints”

Appendix B: added description of the parameters for the Fourier feature encoding layer and the value chosen as batch size

Appendix C: added experiment on synthetic data.

Appendix E: added “calculating a relative RMSD on the validation set”