Dear Professor Jouvet,

We are deeply grateful for your comprehensive review and constructive criticism of our manuscript. We value your insights and will address each of your comments in the following response.

Thank you for helping us improve the manuscript!

Best wishes,

Viola Steidl and co-authors

This paper proposes a framework based on Physics-aware machine learning to estimate the ice thickness distribution of grounded glaciers, applied to seven glaciers located in Svalbard. To achieve this, the authors incorporate mass conservation and other physical constraints into the cost function to be minimized during the training stage. They use ice thickness change and surface mass balance as primary constraints for the flux divergence. The authors conduct a glacier-wise cross-validation and discuss their results, and the cause of the rather-disappointing generalization in view to broader application.

This paper aligns with a series of recent works that explore the significant potential of machine learning techniques for ice thickness inference [e.g. Haq et al., 2021, Teisberg et al., 2021, Jouvet, 2023], including physics-informed deep learning [e.g. Riel and Minchew, 2023, 2022, Iwasaki and Lai, 2023, Jouvet and Cordonnier, 2023]. Among these, Physics-Informed Neural Networks (PINN) [Raissi et al., 2019] are particularly promising and represent an active research area.

Overall I find the approach interesting and I acknowledge the authors for their effort in presenting their results honestly. However, the work appears to be a work-in-progress, and lacks a clear picture of what should be done for making the method generalizable. This is partly due to a too a quickly-made and not very smooth writting (see e.g. my comments below about Section 2.3). The possible causes for the poor generalization results (LOGO CV) should be investigated (especially their relative importance) rather than listed, and I feel that this work could yield improvement in generalization already in this manuscript. I have a series of recommendations (encompassing the writing the paper and the methodology) that I hope will help the authors to improve the manuscript, which has the potential for a larger impact.

- In general, I find the description of the PINN rather inefficient. I focus here on Section 2.3, but my comments may be extrapolated to the entire paper. Do not expect TC readers to be familiar with all ML machinery, even less so with Physics-Informed ones. Therefore, you should provide a minimal background. Currently, section 2.3 is addressed solely to people with prior knowledge. At a minimum, briefly explain what a neural network is (a sequential composition of linear and nonlinear functions with optimizable weights). To my knowledge, this term can be intimidating, while it is not that complicated provided a minimal explanation. The current Section 2.3 mixes crucial information (I/O of the PINN) with more technical details (e.g. activation functions, which are important but not essential for most readers unfamiliar with ML). I suggest distinguishing these two levels when rewriting this part to smooth the reading and target a broader audience. Consider moving ML technicalities to an appendix, and leave te ideas in the body of the paper. Another example: you mention “Fourier layers” but do not provide any rationale (I would like to know the benefit of this). There are several ML-specific concepts (e.g., unlabeled data) that are not explained throughout the manuscript, which is a problem to maximize the audience of the paper to a general glaciological audience.
  Thank you for pointing out the necessity to clarify machine learning terms. We totally agree that we should make our manuscript understandable to anyone without machine learning background. Therefore, we made significant revisions to the section. We added a brief

description of neural networks to the section and streamlined the explanation of the I/O vectors:

"A neural network, also sometimes called multi-layered perceptron, consists of layers of connected nodes, also called neurons, where the connections each have an associated weight. At each node, the weighted outputs from each node of the previous layer are passed through a non-linear activation function (Goodfellow et al., 2016). By minimizing a loss the weights of the network are updated to make accurate predictions."

"In a PINN model the loss is given by the residual of the PDE we want to solve. In theory, PINNs only require input features that are needed to calculate the derivatives in the PDE (Raissi et al., 2018). In our work, we also provide the neural network with auxiliary data, that is related to glacier ice thickness but is not needed to solve the PDE. Therefore, we can exploit information from observable data as we would do it with a non-physics-aware neural network."

"The inputs to the model are vectors for each grid cell in the study region. They contain the spatial coordinates and surface velocities in x- and y-directions, and three β values to correct for basal sliding in x- and y-direction and in the magnitude. Additionally, the vectors contain auxiliary data like elevation, slope, the grid cell's distance to the border of its glacier, and the area of the glacier it belongs to."

To better explain the Fourier embedding of the spatial coordinates we changed the name from "Fourier layer" to "Fourier feature encoding layer" and also added a description of the Fourier embedding:

"The embedding of spatial coordinates was originally developed to overcome spectral bias in neural networks and speed up convergence in the reconstruction of images. It enables the network to learn high-frequency functions in low-dimensional problem domains."

The rationale behind using the Fourier feature embedding is to speed up the convergence of the mass conserving loss that only relies on the derivatives w.r.t. the spatial coordinates. Figure 1 (not included in the manuscript) shows that the Fourier feature embedding clearly makes the mass conservation loss drop faster, whereas it would not be improved at all without the Fourier feature embedding.

The concept of labelled and unlabelled data is now also explained:

"We refer to the points with ice thickness measurements as labelled, whereas points without being referred to as unlabelled."

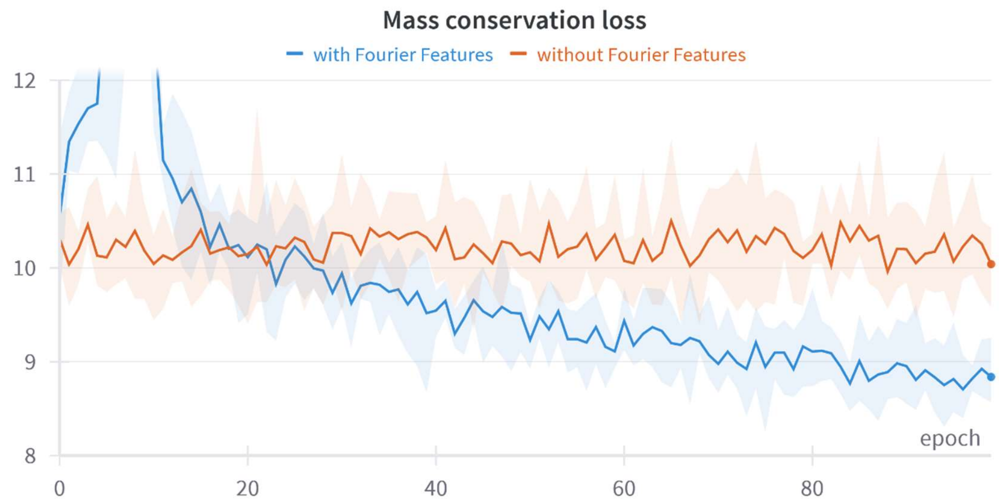

Figure 1 Comparison of mass conservation loss with and without Fourier feature embedding layer

- I am not sure I understand: Do you feed your neural network with raster data grids (as suggested in Fig 1) or with large vectors of data at each coordinate along with the coordinate data? My question is whether you exploit the spatial structure of the data (I assume you have data on a raster structure grid). If not, I understand why you use a fully connected network; if you do, why not use a convolutional neural network designed to capture spatial relationships?
  Thank you for bringing up that this could be misunderstood. As mentioned before we now clarified that the training data are vectors of data at each point of the grid.

  "The inputs to the model are vectors for each grid cell in the study region."

  The spatial structure is not exploited with a convolutional network yet, but we agree that this is an interesting follow-up.

- The description of ice flow (Section 2.2) seems rather simplified. There are a couple of assumptions behind that are not clearly written down. Including a true high-order model here would be a great added value I think. You mention in line 274 that adding momentum conservation would be "technically easy," but I am less pessimistic than you about the claim that "it would complicate the optimization of the model." Instead, the functional associated with the Blatter-Pattyn model, for example, behaves relatively well with good convexity properties [Jouvet, 2016, Jouvet and Cordonnier, 2023], and could act as a physically-consistent, welcome smoother.

Thank you for your comment. As this also came up in the second Review letter we revised Section 2.2 and included the assumptions to the SIA:

"*There are models with different degrees of approximations to the full Navier-Stokes equations to describe ice flow. The simplest one, the shallow ice approximation (SIA) assumes lamellar flow, so the driving forces are entirely opposed by basal drag. It neglects lateral shear and longitudinal stresses and the rate factor A from Glen's flow law is taken to be constant with depth* (van der Veen, 2013)."

We agree that including a higher-order model could provide better estimates of the velocity profile with depth. However, to apply these models we would need to make further assumptions, for example, about the ice viscosity and how it varies with depth or the amount of basal drag/drag from the sidewalls of the glaciers. Indeed, Rückamp et al. (2022) identify this as an issue with the Blatter-Pattyn approximation to full Stokes solutions. We want to emphasise here, that our study is a proof of concept rather than a definitive analysis. We identify several areas for improvement in future work and a higher order model for surface to depth average velocity is one possibility but, likely, not the first order issue for improving the results, which we believe are more sensitive to i) the quality of the input data, ii) the SMB estimates used and iii) estimation of basal velocities. We discuss how each of these issues could be addressed in future work.

Also we clarified our claim about adding momentum conservation being technically easy. We meant to say that adding another component in the loss function is technically easy to do, as it is just adding another term. However, supporting the correct evaluation of the loss requires detailed knowledge about parameters like the viscosity of ice. We now rewrote the sentence to make it less ambiguous:

"While this is technically easy to do, it comes at the cost of introducing uncertainties from approximating required parameters. We would have to make assumptions about ice viscosity and resistance from the bedrock, for example."

Thanks again for bringing up that the way we phrased it could be misunderstood.

- In connection with my previous point, have you considered moving the surface velocity from the input of your PINN to the data? This would make sense if you are including momentum conservation. In the present case, can this be an option too? What is the motivation to insert the "observational" data in input of the PINN or as data constrained in the loss?

We assume with 'moving the surface vel from input of your PINN to the data?' you suggest having the surface velocity in the target vector instead of the Input vector. In fact, this would be an option, too and Teisberg et al. set up their model exactly in this way. However, as the (surface) velocity is actually an important predictor of the ice thickness, we decided to leave it in the input vector.

The idea behind having the apparent mass balance only in the target vector is that we are not confident about the quality of the mass balance data as it is modelled from a simple model. Therefore, we did not want to have it as an input that would give the mass balance data more weight as compared to only introducing it with the soft constraint of the mass conservation loss.

- The comparison (Section 4.2) to the two other products [Millan et al., 2022, Farinotti et al., 2019] is not a strong point. It tells us that the PINN lies within the range, which is not surprising as the two products differ significantly. This section could be moved to an appendix.

  We agree it is not a strong point to prove the correctness of the PINN's ice thickness estimate. However, we think it is informative to show how the estimate compares to other ice thickness estimates. Therefore, we would like to keep it in the Results section.

- Maybe consider applying your method first to a synthetic case where you can create a manufactured bedrock and dataset. Then, use your method to infer the ice thickness and compare it to the ground truth. This approach would help avoid issues related to data suspicion. In general, there are many possible causes for the lack of generalization, but there are strategies to isolate these causes that you could further explore through synthetic experiments.

  We totally agree that applying the method to a perfect synthetic case would be the optimal setting to test the method and research causes for bad generalization. This would be an interesting follow-up exercise but is, by no means, a trivial exercise for the following reasons. The design of the experiment and the design of the synthetic data are crucial in our view. For example, do we use a Full Stokes model, Blatter-Pattyn or some other approximation. Which kinds of glaciers should be modelled with what kind of glacier bed? How to best sample a variety of glaciers? We would need a range of SMB profiles and, presumably, a range of bedrock thermal regimes from fully frozen, partially frozen to temperate and so on. A synthetic data approach would certainly allow us to explore how uncertainties and assumptions influence the robustness of the solution but would be a substantial effort in its own right.

  Nevertheless, we agree that applying the approach to a synthetic dataset would be ideal to better evaluate the PINN model and its strengths or weaknesses.

- Section 5 provides a list of possible causes for the lack of generalization. However, it is hard to draw any conclusions. Some causes are more important than others. It would be helpful if you could prioritize these causes (and improvement items) by order of importance, from the most significant (with the largest potential for improvement) to the least significant. I feel that "Physical constraints" should be at the top of the list.

  We agree that listing the potential causes for bad generalization is not ideal. However, it is certainly not trivial to prioritize the possible causes. We would, for example, argue that input data quality plays a huge, perhaps dominant, role. The relative weighting of data loss and the physics-aware losses, also in close relation to the amount of noise in the measurement data, has a significant impact on the convergence of the PINN (Iwasaki and Lai, 2023). Since in our model the quality/label uncertainty is not yet taken into account, we believe that this could be one way to improve the model. However, improved SMB and basal velocity estimation will also be important, as we state. For the latter, there are several approaches that could be adopted such as using winter-only velocities or by examining the seasonal cycle in velocities.

We agree that physical constraints play a significant role but the significance will likely vary by glacier. To address this concern we have indicated, qualitatively, the factors that would significantly improve the solution.

- Lines 264-266: You place a lot of trust in your Mass balance reconstruction, especially if it is not calibrated (line 264). Considering that this is a major constraint, I think this might be a significant cause of underperformance. Also, using a model for estimating the SMB (even a perfected one) is problematic, as your ” observations” are not observations but modelled reconstructions. Have you considered using in-situ sparse measurements instead?

  Yes, we thought about using observations but as the objective is to evaluate the mass conservation at each point of the grid, we need to fall back to a mass balance product that is available for the entire study area. The mass balance reconstruction that we are using is actually calibrated on observational data (https://docs.oggm.org/en/stable/mass-balance-monthly.html).

  However, maybe in a follow-up work, it would be worthwhile to include another loss component where the residual to mass conservation is calculated from in situ SMB measurements wherever they are available, just like the data loss is evaluated only where ice thickness measurements are available. Thanks for making this suggestion.

I have some additional specific comments:

- In the introduction, it would be good to ellaborate the existing literature on using ML for ice thickness inversion modeling [e.g. Haq et al., 2021, Teisberg et al., 2021, Jouvet, 2023] (line 21), as well as physics-informed deep learning applied to similar problems, such as inferring basal conditions (bedrock location or slipperiness) [e.g. Riel and Minchew, 2023, 2022, Iwasaki and Lai, 2023, Jouvet and Cordonnier, 2023] (lines 32-34). As this is a fast-evolving field, it would be good to check the latest papers, and possibly to complete.
  Thank you for providing further literature that should be included. We extended the literature review to make it more complete. We hope this meets your expectations.

- l 12: Not sure Millan et al. [2022] is the most appropriate reference for that.
  Thanks for pointing that out, we apologize for the mistake and changed the reference to Welty et al. 2020.

- l 13: “Physics-based approaches ...” This sounds to be a very personal definition, consider a more appropriate one.
  Agreed, we changed the sentence such that it cannot be misunderstood as a definition anymore: *“There are physics-based and process-based approaches that aim to reconstruct glacier ice thicknesses from in situ data and ice dynamical considerations.”*

- l 22: “One avantage of data-driven approaches is a significant speed-up compared to physics based models”: The computation speed-up has nothing to do with whether it is data-driven or physics-driven; it is the result of the efficiency of evaluating a neural network (especially on GPUs), irrepective of the training strategy: based on data [Jouvet et al., 2022] or on physics [Jouvet and Cordonnier, 2023]. Please correct.

Thank you for the correction. We changed the sentence to clarify that we are talking about data-driven machine learning methods that are fast to optimize and evaluate: "*One advantage of machine learning approaches is their efficient optimization and evaluation compared to process-based models* (Jouvet et al., 2022)*.*"

- l 30-31: These two sentences are unclear to me : i) what means "data-efficient" in the context? ii) "boundary condition to solve the PDE", I think I understand what you mean (this would be a Dirichlet BC as you can enforce the solution to be close to a certain given value somewhere), but I'm not sure this is clear for all.
  Thank you very much for bringing up that this is not clear. With 'data-efficient' we meant to describe that we are less dependent on ground truth data because we are also relying on physical constraints. We took this out to avoid misunderstanding. Also, as you rightfully mentioned the term boundary condition might be misleading as the data loss is not exactly a condition that we set on the boundary of the domain but rather an "internal constraint" that helps find a solution to the PDE. We also changed this wording in the manuscript: "*Additional ground truth data can be used to compute a data loss that acts as an internal condition to constraining solutions to the PDE.*"

- l 255: "the loss landscape is highly complex", this is an unsual way to describe the lack of convexity the loss, which is not improved – I agree – by adding the number of constraints within the loss. I am not sure I found what optimizer you used (ADAM, SGD, RMSPROP, ?).
  We apologize for not including this information in the manuscript before. We used the Adam optimizer and added the information to the new Appendix Section on the architecture of the model.

- Appendix B: I feel I have seen this exercise numerous times in textbooks, deriving a 0.8 ratio between vertically-averaged and surface velocity in the non-sliding SIA parallel slab case. I suggest you replace it a reference and use the space in the paper to better explain the ML part.
  We agree that this is often described in textbooks, but we would like to keep the derivation as an explanation of where our lower bound to the depth-averaged velocity estimation comes from and also which assumptions have been made.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, 2016.

Iwasaki, Y. and Lai, C.-Y.: One-dimensional ice shelf hardness inversion: Clustering behavior and collocation resampling in physics-informed neural networks, J. Comput. Phys., 492, 112435, https://doi.org/10.1016/j.jcp.2023.112435, 2023.

Jouvet, G., Cordonnier, G., Kim, B., Lüthi, M., Vieli, A., and Aschwanden, A.: Deep learning speeds up ice flow modelling by several orders of magnitude, J. Glaciol., 68, 651–664, https://doi.org/10.1017/jog.2021.120, 2022.

Raissi, M., Perdikaris, P., and Karniadakis, G. E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys., 378, 686–707, https://doi.org/10.1016/j.jcp.2018.10.045, 2018.

Rückamp, M., Kleiner, T., and Humbert, A.: Comparison of ice dynamics using full-Stokes and Blatter–Pattyn approximation: application to the Northeast Greenland Ice Stream, The Cryosphere, 16, 1675–1696, https://doi.org/10.5194/tc-16-1675-2022, 2022.

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