Response to Referee #2

We thank Referee #2 for the careful and constructive review. Below, we address the comments and the additional experiments carried out in response to them.

1. Isolating the effect of sea ice topography

The referee is right that the manuscript does not yet isolate the effect of topography and that a baseline without topography is needed to support this claim. In the revised manuscript, we isolate it with controlled experiments in which everything other than the topography is held fixed.

First, we compared otherwise identical SnowModel-3D simulations with and without surface topography. Every simulation in the training set was run twice under the same weather forcing, once over a flat surface and once over a synthetic ridged surface, with roughness spanning the range CryoSat-2 observes. The difference between the two runs is the effect of topography in the physical model. With ridges, the snow is deeper, less dense and much more unevenly distributed.

Second, we retrained the emulator without topography, which is the standard remedy for attributing a feature in machine learning models (Hooker et al., 2019, 2021). Two baselines were trained with the same data, architecture and training procedures as the full emulator. In the first, the roughness input to the emulator is removed, so the network sees the flat and the ridged simulations but cannot tell them apart. In the second, the ridged simulations are removed from the training set as well, so the network never sees the effect of topography. On simulations held out from training, the full emulator (the one that also includes topography) reproduces the difference between the ridged and the flat member of each pair, whereas both baselines predict the same snow on both members, and their errors in depth and density grow with the roughness of the domain (Figs. R1 and R2). The emulator therefore carries the effect of topography, and it does so through the roughness input.

We also evaluated the two baselines against the observations on the same cells as the full emulator. The effect of topography is visible where the measurements resolve the scale of the floe: on the MOSAiC floe the full emulator is closer to the measured depth and density than the baselines. On the airborne data the three emulators score about the same. We think this is because these data are compared at the grid scale, where the effect of topography on the mean snow depth is a few centimetres, and because they do not show how the snow is distributed within a grid cell. The present in situ and airborne record is too sparse to validate the topographic component of any snow-on-sea-ice model extensively. Part of the difference between the emulator and SnowModel-LG is therefore explained by topography, the rest by the other differences between SnowModel-3D, the emulator and SnowModel-LG, and the in-situ data are too sparse to quantify the share of each.

In the revised manuscript we add a section on isolating the effect of topography with these experiments, revise the abstract and the conclusions accordingly, rewrite the paragraph on the mechanism, and add a paragraph on the limits of the observational record.

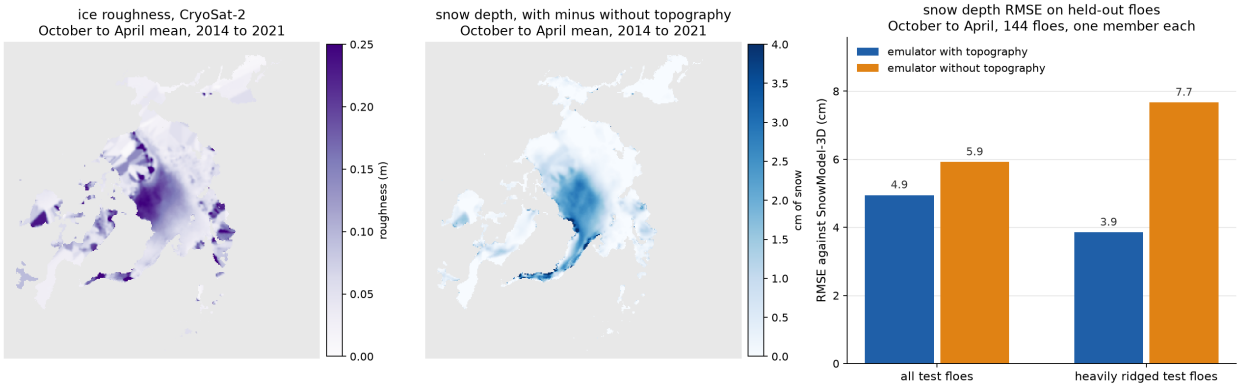

Figure R1. Left: the CryoSat-2 roughness used as the emulator input. Middle: the difference in winter snow depth between the emulator with and without the roughness input, everything else identical. Right: error against SnowModel-3D on the simulations held out from training, for all of them and for those with the roughest surfaces, with and without the roughness input.

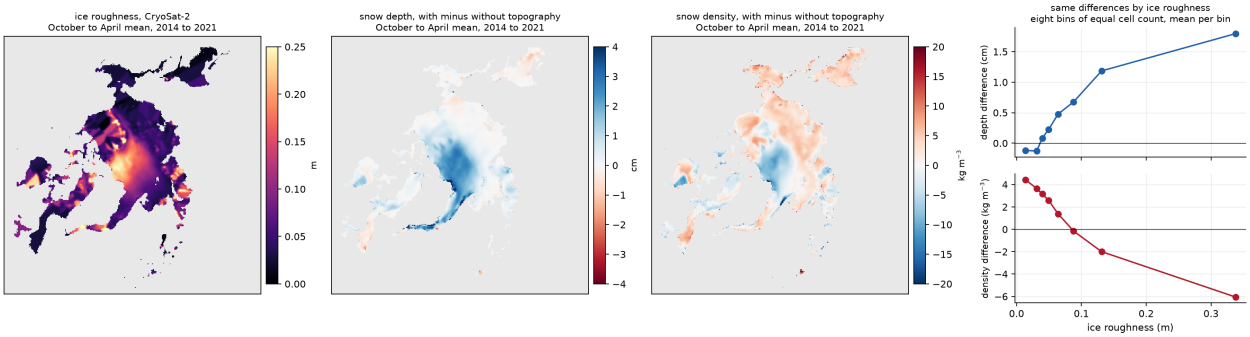

Figure R2. The effect of the roughness input in the emulator, everything else identical: the CryoSat-2 roughness, the difference in winter snow depth and in snow density between the emulator with and without the roughness input, and the same two differences binned by roughness. The effect is absent on smooth ice and grows with roughness, as deeper and lighter snow.

2. The memory of the LSTM

The referee asks what the memory of the network adds when the snow depth and density are already fed back, how it behaves through the year, over which time scales it was learned, and whether a model without memory would do as well. We give the missing training details and answer the rest with three tests on the simulations held out from training.

The network is trained in stages on sequences of about two weeks, two months, four months and a full year. The true previous snow state is used only in the first two stages and is phased out, so in the last two stages the network runs on its own predictions for a full year, the same way it is used at inference.

First, we ran the trained emulator with its memory erased on different schedules while the explicit snow state was kept and fed back as usual. Erased every day, the emulator fails: its errors in depth and density grow several times although it still receives the previous depth and density. Erased every week, the density degrades. Erased every one to three months, the results are unchanged, which also covers the start of the inference from a zero state. Second, we trained a feed-forward network without memory, with the same inputs, the same feedback of depth and density, the same physics step and the same training procedures. It performs significantly worse than the LSTM in both depth and density. Third, we analysed what the memory holds. Its variance is explained only in part by the current inputs and the fed-back state. The rest is explained by the history of the forcing: the mean air temperature of the past months, the snowfall of the season and the time since

the pack reached its maximum. The memory carries the thermal history and the age of the pack, which the bulk depth and density do not, and which govern compaction in the layered snowpack of SnowModel-3D.

In the revised manuscript we give the training windows and the schedule of the true state, describe the reactivation rule, and add a section on the memory with these tests and the model without memory.

3. Emulation error against SnowModel-3D

The referee is right that the errors on the held-out regions and the growth of the error over the free rollout are not reported in the manuscript. They were given in our earlier work on the emulator (Prasad et al., 2024) and we referred to it instead of repeating them. Since submission, we have also reformulated the emulator. As noted in Prasad et al. (2024), the physics-guided loss is a soft constraint, so there is no guarantee that the network satisfies the compaction equation at inference. We have now reformulated the emulator so that the LSTM predicts the daily change in snow water equivalent due to precipitation, the daily mass losses, and the compaction rate. The snow mass budget and the compaction equation are then applied to these predictions to compute snow density and snow depth. Density, depth, and snow water equivalent are therefore physically consistent through the model architecture, at inference as well as in training. We have also restructured the inference code, which makes the emulator significantly faster than SnowModel-LG. In the revised manuscript we describe the reformulated emulator and report its error against SnowModel-3D on the simulations held out from training, for depth and density.

4. MERRA-2 forcing

We agree that uncertainties in MERRA-2 precipitation may contribute to the underestimated snow depth at MOSAiC, but this cannot be established without additional sensitivity experiments. We will therefore revise the discussion to present this as a possible explanation rather than a demonstrated cause.

We use MERRA-2 consistently to force SnowModel-3D, the emulator, and SnowModel-LG, because it allows us to investigate differences in model representation independently of the atmospheric forcing. Assessing sensitivity to alternative reanalyses would require additional simulations and emulator training and is beyond the scope of the present study.

The five training regions are not single grid points. Each is a patch of many grid cells of the forcing, and the patches were chosen to span different ice and climate conditions, so the training covers a much wider range of atmospheric conditions than five points would. We give the number of cells and training sequences and the range of conditions in the revised manuscript.

5. Topographic Input

The topographic input is a statistic of the surface height, computed in the same way from the synthetic ridge field in training and from the field derived from CryoSat-2 at inference. The manuscript gives only the inference side, we will state both in the revised version.

With μ fixed the input is indeed a monotonic function of the CryoSat-2 roughness, and this is by design. The roughness gives the spread of the surface height, not its level, and a single level from N-ICE2015 is the simplest assumption the available in situ data allow. The variation between ice types, regions and months enters through the roughness itself, month by month, so the transform keeps all the information the satellite carries. The two sources have different roles: Sudom et al. (2011) gives the geometry of the ridges in the training fields and N-ICE2015 the level of the height distribution at inference. We make this clear in Fig. 1 and in the text.

Where CryoSat-2 has no retrieval, the last retrieved field is carried through the melt season and empty cells are filled from the nearest retrievals. Ridges persist through the summer and the accumulation season begins when the retrievals resume, so we consider this reasonable. We will describe it in more detail in the revised manuscript.

In the title, high resolution refers to the metre-scale simulations the emulator is trained on, which resolve the snow around individual ridges. The emulator carries their effect into the pan-Arctic

product on the same grid as SnowModel-LG.

6. Comparison with other datasets and models

We agree that comparisons with additional pan-Arctic snow products would provide valuable context. However, the primary objective of this study is to evaluate whether the emulator can reproduce the effects of sea ice topography represented in SnowModel-3D and transfer them to pan-Arctic simulations. SnowModel-LG provides a direct reference, while independent airborne and MOSAiC observations are used to evaluate the resulting snow properties.

Products such as NESOSIM, Warren climatology and satellite derived snow depth estimates differ in their temporal coverage, spatial resolution, and underlying methodology. An intercomparison addresses a complementary question beyond the scope of the present study.

We agree that observation-based snow products already reflect the influence of topography on snow distribution. Thank you for pointing this out. We have corrected the corresponding statement in the conclusions and clarified that potential applications to sea ice thickness retrievals and climate modelling remain to be demonstrated.

7. Fig. 7 and Table 4

The two products do not cover the same cells, since each has its own parcel coverage, and in the submitted version each was compared on every cell where it had a value. In the revised manuscript we keep, for every flight and date, only the cells where the observation, the emulator and SnowModel-LG all have a value, compute RMSE, bias and correlation of both products on that one set, and give the number of cells in Table 4, as in Table 6.

Specific comments

1. Line 205: the initial snow depth and density are taken from SnowModel-LG on the start date. We will state this in the manuscript.

2. Lines 212–215: yes, it does

3. Lines 237–238 and 308: we will make the wording consistent. The evaluation is against the observations, and SnowModel-LG is the point of comparison.

4. Lines 293–295: we will correct the text to agree with Fig. 3a.

5. Lines 312–315: agreed, we will remove the statement.

6. Fig. 9b: we state in the revised manuscript how the summer measurements were filtered and keep melt pond depths out of the comparison.

7. Table 1: that was a mistake in the caption. CryoSat-2 roughness enters only at inference, the training simulations use synthetic ridge fields.

8. Line 329: yes. The combined data set of Jutila et al. contains three AWI airborne campaigns: PAMARCMIP2017, RESURV79 in 2018 and IceBird Winter 2019. In 2019 AWI began naming all its airborne sea ice campaigns after its monitoring programme, IceBird, so the two earlier campaigns carry their own names. In the revised manuscript we refer to them as the AWI airborne campaigns and name each one.

References

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B.: A benchmark for interpretability methods in deep neural networks, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, article 873, 12 pp., 2019.

Hooker, G., Mentch, L., and Zhou, S.: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance, Statistics and Computing, 31, 82, https://doi.org/10.1007/s11222-021-10057-z, 2021.

Prasad, A., Merkouriadi, I., and Nummelin, A.: Modeling snow on sea ice using physics-guided machine learning, Environmental Data Science, 3, e28, https://doi.org/10.1017/eds.2024.40, 2024.

Sudom, D., Timco, G., Sand, B., and Fransson, L.: Analysis of first-year and old ice ridge characteristics, in: Proceedings of the 21st International Conference on Port and Ocean Engineering under Arctic Conditions (POAC), Montréal, Canada, 10–14 July 2011, POAC11-164, 2011.