The manuscript extends a physics-guided LSTM emulator of SnowModel-3D to the pan-Arctic scale by applying it along the Lagrangian parcel trajectories of SnowModel-LG, with sea ice topography derived from CryoSat-2 surface roughness. The topic is important, and the manuscript is generally clearly written. However, I am not quite convinced that the present analysis shows that the better agreement with airborne observations comes from the representation of sea ice topography, which is the main claim of the paper, or that the emulator is sufficiently validated against its own training target.

My main concerns are as follows.

1. The central claim of the manuscript is that the emulator adds the effect of sea ice topography on snow redistribution, and that this brings the simulated snow depth closer to the airborne observations. However, I do not see any analysis that isolates the role of topography. The emulator differs from SnowModel-LG in several: it is trained on a different model, it is a statistical model that tends to smooth its predictions towards the mean, and its output is gridded and sampled differently. In my opinion, the improvement over SnowModel-LG could come from any of these, and the manuscript does not show that topography is the reason. It is also not shown whether the difference between the emulator and SnowModel-LG is related to CryoSat-2 roughness at all. In addition, I am not sure that the differences are physically consistent with the proposed mechanism. The Discussion argues that redistributed snow becomes denser and therefore shallower for the same snow water equivalent. If I read the colour bars in Fig. 6 correctly, the emulator is both denser and slightly deeper than SnowModel-LG over most of the central Arctic, which would imply more snow water equivalent from the same snowfall. In my view, the attribution of the improvement to sea ice topography is therefore not yet supported by the results.
2. The use of an LSTM is motivated by its ability to retain memory of past conditions, but I am not convinced that this memory is actually needed or understood here. The previous snow density and depth are already fed back explicitly at each time step, so the snow state itself is carried outside the network. It is therefore not clear what additional information the hidden and cell states store, and the choice of an LSTM is not compared with a simpler model without memory. More importantly, it is unknown how this memory behaves through the year. The relevant history of the snowpack is very different during autumn accumulation, winter compaction, spring melt and the summer period when most of the snow disappears. It is not shown whether the memory of the LSTM follows these changes, resets after the melt season, or carries information from one year into the next. The training sequence length is not given, and neither is whether the network was trained with the true previous state or its own predictions, so it is not clear over which time scales the memory was learned. At inference, by contrast, the model runs freely for eight years from a zero hidden state. In addition, parcels that become active again receive the hidden and cell states of a neighbouring parcel, so their memory partly belongs to a different ice history. In my

opinion, the statement that the autoregressive framework captures the persistence of snow states across multi-year timescales is therefore not supported by the results.

3. I am also not clear how well the emulator reproduces SnowModel-3D itself. Leave-one-region-out cross-validation is described in Section 2.7, but the errors on the held-out regions are not reported anywhere in the manuscript. Similarly, Lines 240-241 announce an assessment of the cumulative error in the autoregressive forecasts, but I could not find it in the Results. This is important because, without knowing the emulation error, I cannot tell how much of the 77 density RMSE against SnowModel-LG (Table 2) comes from the topographic effect and how much simply comes from the emulator not reproducing SnowModel-3D. Likewise, I am not sure that the smoother fields of the emulator can be attributed to snow redistribution, since part of this smoothing may come from the method itself.
4. MERRA-2 winter precipitation over the Arctic Ocean is a known source of uncertainty, which the authors acknowledge, and the underestimation of snow depth at MOSAiC is attributed to MERRA-2. However, this attribution is not tested, and in its current form it remains a hypothesis. It is also not clear how robust the improvement against the airborne data is to the choice of forcing, since both the emulator and SnowModel-LG use the same reanalysis. Moreover, the emulator is trained on MERRA-2 forcing from only five regions. It is not clear whether each region corresponds to a single grid point, or how many independent training sequences this provides for a two-layer LSTM with 128 hidden units. My impression is that the range of atmospheric conditions seen in training is rather narrow compared with the pan-Arctic conditions to which the emulator is applied.
5. I am not convinced that the topographic input used at inference is equivalent to the topography used in training. It is not explained how the topographic input is computed from the synthetic ridge fields in training. At inference, it is derived from CryoSat-2 roughness with a single value of μ from N-ICE2015, applied to the whole Arctic, to all ice types and to all months. With μ fixed, the input is simply a monotonic function of r0. The source of the training ridge statistics is also inconsistent: You cite Sudom et al. (2011), whereas Fig. 1 refers to N-ICE2015. Moreover, as far as I know, CryoSat-2 roughness is usually not available from May to September or north of 88° N, and the manuscript does not explain what is used there. Similarly, the title refers to “high-resolution distributions”, whereas the emulator only provides floe-averaged values on a 25 km grid.
6. The emulator is only compared with SnowModel-LG. NESOSIM, the Warren et al. (1999) climatology and the multi-frequency altimetry snow depth products that the authors cite themselves (Carret et al., 2025; Landy et al., 2026) are discussed in the Introduction but not included in the evaluation. As a result, I do not think the claimed added value for altimetry-based thickness retrievals and climate studies is supported. I also find the statement in Line 425, that existing pan-Arctic snow products do not include the ice surface beneath the snow, not quite fair to observation-based products. These products measure the actual snow cover and therefore already include any effect of topography.

7. In Fig. 7 the two models are evaluated on different numbers of grid, so the RMSE values in Table 4 are not strictly comparable. In addition, only RMSE is reported, although the emulator values appear more compressed than those of SnowModel-LG. Please also give bias and correlation, as in Table 6.

Specific comments:

1. Line 205: Please state where the initial snow state comes from
2. Lines 212-215: Reactivated parcels inherit the snow state and hidden state of the nearest active parcel. Does this also apply to new ice forming in autumn?
3. Lines 237-238 state that model performance is evaluated against SnowModel-LG, whereas Line 308 states that this comparison does not serve as evaluation.
4. Lines 293-295 describe lower densities in the central Arctic and higher densities near coastal and marginal ice zones. Fig. 3a and Lines 376-377 seem to show the opposite.
5. Lines 312-315: Year-to-year variability in RMSE does not show that natural fluctuations are “captured effectively” by the emulator.
6. Fig. 9b: From June onward the observed snow depth remains around 0.1 m while both models are close to zero. The Magnaprobe dataset also contains melt-pond depths (Itkin et al., 2021). Please explain how summer measurements were filtered.
7. The caption of Table 1 implies that CryoSat-2 topography drives SnowModel-3D?
8. Line 329 refers to the “combined Jutila et al. IceBird campaigns”, which also include PAMARCMIP2017 and RESURV79?