General comment

The paper "Learning to filter: Snow data assimilation using a Long Short-Term Memory network" presents a novel framework for snowpack prediction combining physical-based model and machine learning model. It could be a great fit for the journal. However, there are several aspects of the experimental setup and methodology that would benefit from additional clarification. I encourage the authors to provide more detailed descriptions of their experiments to enhance the transparency and reproducibility of the study. Please see my comments below.

We thank the reviewer for the positive evaluation of our manuscript and we acknowledge the need for additional clarification needed to sustain transparency and reproducibility of our study. We plan to improve the quality thanks to the useful feedback received. Here below is a list of answers to specific comment and planned changes.

Specific Comments

The overall data samples are limited (both years and sites), compared to https://doi.org/10.1175/JHM-D-22-0220.1 Could the authors comment on this issue?

We acknowledge the reviewer's concern. Our aim was to explore the application of pilot point approach across different locations in the Northern Hemisphere that would be fit for operational and quasi-real time applications. Despite working with a relatively limited dataset, the results are still promising and provide a solid foundation for future developments, especially in a 2D spatially distributed modeling case. Nevertheless, we acknowledge the importance to point out this difference in our work, hence we plan to add a sentence in the introduction at line and one in the discussion at line . Here the two sentences

Introduction: at line 83 after "based on topographic features.":

> As more recent exception of combining deep learning and snow data assimilation, Song et al. (2024) developed an LSTM-based framework to assimilate lagged observations of SWE or satellite-derived snow cover fraction (SCF) over the western U.S., aiming to improve seasonal snow predictions. While their approach further consolidates the potential of deep learning for data assimilation in snow hydrology, it relied on a relatively simple assimilation setup, dealing with long lagged time step rather than a consequential and quasi real time approach.

Discussion: at line 407 after "used in training".:

> Although using limited datasets in both temporal and spatial coverage compared to recent studies pursuing a similar effort (Song et al. 2024), our approach proved to be effective in speeding up traditional data assimilation techniques while maintaining comparable performance; additionally , our framework, designed to test the operational viability of a quasi–real-time pilot point method, still proved the feasibility of an alternative use of LSTM algorithm without loosing in performances. The encouraging results provide a foundation for extending this framework to broader, more diverse networks in future research.

*What is the temporal frequency of S3M? Is it 1 hour (**Line 121**)?*

The temporal frequency used in this study is 1 hour, but S3M can be run with different time steps.

What is the input time window size for the LSTM model? If my understanding is correct, only one timestep of meteorological forcings are used as input, based on Fig. 2 and Line 269. This is not a typical use of the LSTM model if multi-time steps are not involved. The architecture of the LSTM model also requires more details (e.g., hidden layers, hidden units).

We appreciate the reviewer's insightful comment. Indeed, during the operational testing phase, the LSTM model is provided with only one timestep of meteorological forcings as input. However, this design is intentional, aimed at simulating real-time forecasting conditions, where only the current timestep of meteorological data is available for prediction. During the training phase, the model is trained with multi-time-step sequences to learn temporal dependencies, consistent with traditional LSTM approaches. Therefore, although the operational phase uses only one timestep of input, the model is trained with multi-timestep data to capture temporal dynamics over time. Concerning the architecture of the LSTM model , we plan to add a clarifying section as mentioned in the answer to RC1

Related to the previous comment, please clarify the "memory components" of the LSTM model. By design, the previous time series should be used as inputs to the LSTM model. What is the model without these "memory components"? If this is beneficial, do the authors consider incorporating more previous timesteps?

The "memory components" refer to additional sets of features provided to the LSTM during both the training and testing phases. These components include:

- The meteorological forcing variables at each timestep (i.e., timestep k), and
- The state of the system at the previous timestep (i.e., at timestep $k-1$).

Regarding the reviewer's suggestion of incorporating more previous timesteps: we acknowledge that this could improve the model's predictive capability, particularly during the operational phase. We will consider exploring the use of longer input sequences of previous timesteps in future work, contingent upon operational constraints.

*Loss function. As noted in **Line 246**, the output of negative SWE is forced back to zero, why is the regularization term still necessary in **Line 260**? Is the hard cut at zero only applied after training the model?*

We acknowledge that the current structure of the paragraph may lead to a misleading interpretation of the procedure. Specifically, the hard cut to zero was applied after the LSTM prediction, and therefore after the regularization term was used. This step was introduced to preserve the intermittency of snow quantities while also helping the LSTM network learn to detect the onset of a snowpack with new snowfall on bare ground. To enhance clarity, we plan to move the sentence currently at lines 246–247 to follow line 265.

Multisite LSTM. Do the authors consider the use of site-specific information as inputs (e.g., lat-lon, slope https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023WR035009; https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031033)

We did not used elevation and coordinates of the site, but rather decide to decrease the number of input as to match those used by the EnKF in a non-distributed modelling. However, we plan to develop a 2D spatially distributed snow modeling framework of our LSTM EnKF emulator using more basins and data within each basin to leverage such info.

Line 326. *"Reduce" RMSE by "-25" seems to increase RMSE for me. Please consider rephrasing it.*

We apologize for the lack of clarity and we will improve the revised manuscript. We plan to modify the sentence at line 326 as follows:

> indeed the LSTM resulted in a reduction of 25 mm in RMSE for SWE, while the EnKF achieved a larger reduction of 31 mm.

Figure 5. *Why is the RMSE for "open loop" not shown here?*

We have modified the figures 5 and 6 to show also the RMSE of the open loop.

There are some caption inconsistencies. Please take time and revise them (e.g., the capital letters in Figure 8 caption)

We thank the reviewer for pointing out these inconsistencies, we will revise and correct the caption.

Figure 8. *Is there any particular reason to assess the performance based on different water year types? A similar and consistent RMSE as previous experiments would be helpful.*

The rationale behind evaluating performance across different water year types was to test the model's robustness under varying hydrological conditions, such as dry, normal, and wet years. This, according to our opinion and based on well established procedures in hydrology(Osuch et al. 2015), allows us to better understand how the model performs beyond average conditions, particularly in more challenging or extreme scenarios.

References

Osuch, M., Romanowicz, R. J. & Booij, M. J. (2015), 'The influence of parametric uncertainty on the relationships between hbv model parameters and climatic characteristics', *Hydrological Sciences Journal* **60**(7-8), 1299–1316.

Song, Y., Tsai, W.-P., Gluck, J., Rhoades, A., Zarzycki, C., McCrary, R., Lawson, K. & Shen, C. (2024), 'Lstm-based data integration to improve snow water equivalent prediction and diagnose error sources', *Journal of Hydrometeorology* **25**(1), 223–237.