Reply to RC.1

General comments

The authors of "Learning to filter: Snow data assimilation using a Long Short-Term Memory network" have made efforts to address previous comments. Please see my comment below for further clarification and improvement.

We thank the reviewer for the time dedicated to the evaluation of our manuscript to further improve its quality and to make it better understandable to a broader audience. We modified the manuscript following reviewer's comments, as we detail below.

Specific comments

1. Please consider adding longitude lines for the European sites, and one more latitude line for both European and Asian sites in Figure 1.

Comment addressed.

2. If my understanding is correct, the LSTM model—at least in the operational testing configuration—uses only a single previous timestep as input. The main difference compared to a basic model appears to be that the LSTM's memory component includes meteorological forcings from the previous timestep (as shown in Equation 4). This raises the question: is the use of an LSTM necessary in this case? Would a simpler ANN model suffice? Additionally, the authors state that this setting is to mimic the real forecast setting, but it is unclear why this is the case. Typically, historical meteorological data should be accessible for forecasting applications. Could the authors clarify this point? Finally, the response mentions that the training phase uses multi-time-step input sequences. What is the specific timestep length used during training? Does this imply that the model is trained using a different temporal input structure than what is used during testing? If so, a more explicit explanation of this mismatch would be helpful.

As correctly understood by the reviewer, the operational testing configuration of the LSTM model indeed uses only the previous single timestep as input. However, contrary to a basic Artificial Neural Network, the LSTM can retain and implicitly utilize information from a longer sequence of past time steps through its internal cell states and hidden states previously trained. This ability to maintain a "memory" of historical conditions, even when only a single previous timestep is given as input, is why we choose to use a LSTM even while relying on this alternative single-step testing setup. Additionally, during the training phase, the model was trained using multi-time-step input sequences. Specifically, approximately 10 years of hourly data records for each site were used during training. Acknowledging that the different approach between training and testing may not be satisfactory explained, we added an explanatory comment on line 304 after "... observations and model predictions.:

> "It is important to stress that, while the training phase was performed in the conventional way of training neural networks -meaning multiple timestep as input

to obtain a sequence of outputs - the operational testing phase was performed giving to the LSTM trained models only one timestep at a time, to be coupled with the forward step of the cryospheric model."

Finally, we agree with the reviewer that historical meteorological data is accessible for model development and training, in a true operational forecast scenario; however, we would like to point out that the framework we built was to couple our LSTM as an online sequential data assimilation step (i.e., so-called filtering) to re-initialize our cryospheric model, which relies only on the most current observations and its internal learned dynamics during each computational step to project snow conditions forward. This argument is particularly important in terms of reducing the amount of storage needed to run a model operationally for forecasting purposes. This is why we chose to only rely on the most recent timestep in testing.

3. Figure 5. Please check the colors in the legend. It seems to me the color for LSTM-DA and LSTM-DA with memory in the legend is too light and do not match with the figure.

Comment addressed.

Reply to RC.2

General comments

I think the authors have done good revision work and improved the quality of this manuscript. My concerns have been addressed or explained.

We appreciate the help of the reviewer in further improving the quality of the manuscript and we will perform an in-depth grammar and typo check before submitting the new version.

Specific comments

I only have two minor comments left:

1. The colors of the result figures are difficult to distinguish between the different models. Please further improve the clarity of these figures. Figure A1: The legend is hard to read.

Comment addressed.

2. Please correct the grammar and typo errors throughout the manuscript.

For example:

Line 403: "KGE" should not have a unit.
Line 498: "without loss in performance."

There are more besides these two examples.

Comment addressed.

Reply to RC.3

General comments

The topic of the manuscript is interesting and relevant for the readers of Cryosphere, as well as for the larger international hydrology community. It is also in line with the actual popularity of machine learning in hydrology (and everywhere, really!). If I understand the situation correctly, the manuscript has already been reviewed by two anonymous referees, who might not be available anymore, or the editor wanted a third opinion. Considering this, I have tried my best to prioritise verifying that the authors have addressed all the comments made by the two initial referees and I will refrain from starting the review process all over again. I did, however, notice a few minor points that were not raised during the initial review that I think would be wort addressing. Overall, I would consider this a minor revision

We appreciate the reviewer comment on the impact of our paper on the Cryosphere community and we thank them for the further effort made to help us improve our manuscript. We provide here more in-detail answers to reviewer's comments and modified the manuscript following the reviewer's suggestions.

Specific comments

1. Introduction: too much emphasis on mountains- The introduction of the manuscript start with emphasising the importance of snow in mountain regions. As someone who lives in a non-mountainous area where snow plays a very important role in the hydrological cycle, I tend to nd this a little bit annoying. The data assimilation method you propose is presented as a point-based method, at least for now. One of the challenges of modelling (or assimilating) snow in mountainous regions is the high spatial variability, and this is something that the current version of your method is not addressing. However, your method is still interesting an useful, in general, for snow dominated areas. All of this to say: I think it would make more sense to start the introduction from a more general point of view, and not "mountains " specially. For instance, you could just start the introduction with "reliable estimates of snow water equivalent (SWE) and snow depth in snow-dominated environments are essential (. . .) ",explain why, and then maybe mention the specific case of mountains.

We understand the reviewer's concern. We addressed this comment by changing the introduction as follows:

> "When studying the hydrological cycle, one cannot underestimate the key role played by snow (Pagano & Sorooshian 2002); indeed, for snow-dominated catchments, today's snow is tomorrow's water. Information on the state and distribution of snow cover provides helpful information to characterize seasonal water storage (Zakeri et al. 2024), seasonal to annual water availability (Metref et al. 2023), and several cascading socio-hydrologic implications (Avanzi et al. 2024).

Especially in cold regions, which are heavily affected by climate change (Hock et al. 2019), the snowpack often functions as the primary source of streamflow, particularly during spring and summer (Bales et al. 2006). Moreover, considering the high spatial variability in these regions, the scientific community agrees on the needs of reliable estimates of Snow Water Equivalent (SWE) and snow depth in snow-dominated environments, which are essential for effective and timely management of water resources (Hartman et al. 1995)."

2. Data quality: how do you measure it? Section 2.1 mentions that you used '(...) high-quality, pre-processed datasets from long-term, internationally acknowledged snow research stations across the northern hemisphere (...) '. Can you specify how data quality was measured and provide some details?

To assess the quality of the dataset used in our study, we relied on the quality control procedures and documentation provided by the original data providers, who typically apply standardized protocols for data collection, sensor calibration, and outlier removal. These datasets are widely used and cited in the scientific community, and their quality assurance practices are well established. While we did not perform an independent quality assessment, we selected stations with a long history of data availability and peer-reviewed documentation to ensure the reliability of the input used in our analysis.

3. Data quality: how robust is the proposed method? Maybe this comment is more for the discussion, and it also depends on your reply to my previous question, but I think it would be interesting to reflect on how robust you think the method will be to lower quality data? Operationally, one would want to apply your method using data of slightly lower quality, if that is what is available to them. How will this affect the results? Would the method be able to 'ignore' a portion of the data if the quality is too low?

We appreciate this timely and relevant question, which indeed opens a broader discussion around the applicability of artificial intelligence-based (AI) methods in real-world operational contexts. It is well known that AI and, in particular, deep learning techniques are highly sensitive to data quality and quantity. The performance of these models tends to degrade when trained or applied to noisy, incomplete, or low-resolution data.

However, this limitation can often be addressed through robust data pre-processing and augmentation strategies. Indeed, a significant portion of time in AI workflows is devoted to improving data quality before model training. Moreover, there is growing interest in hybrid approaches that combine traditional statistical techniques with deep learning, offering a potential pathway to enhance robustness in low-quality data settings.

For instance, recent work by Gauch et al. (2025) provides valuable insights into handling missing data in operational environments, suggesting imputation and correction methods that could be integrated into AI pipelines. Additionally, On the AI side, generative models (e.g., GANs or diffusion-based models) have shown promise in enriching and recovering incomplete datasets, as illustrated by Dhoni (2023).

In order to add a discussion of this timely topic to our paper, we have added a sentence after line 557:

While our results are based on high-quality forcing and observational datasets, we acknowledge that operational applications may involve lower-quality inputs. In such cases, pre-processing strategies (e.g., bias correction, gap-filling) and hybrid DA–AI frameworks could help mitigate performance loss, with the potential to selectively down-weight unreliable inputs rather than propagating their errors through the model. Recent work by Gauch et al. (2025) demonstrates the effectiveness of imputation and correction methods for handling missing or degraded data in operational environments, while generative models such as those explored by Dhoni (2023) offer promising avenues for enriching and augmenting incomplete datasets.

4. Reviewer 1's comment about section 3.4 (spatial transferability vs multisite training) The original Reviewer 1 commented that ' instead of presenting the spatial transferability of a single model, it might be more meaningful to compare and discuss the site-specific LSTM and the multi-site LSTM '. I completely agree with them, and I don't think the authors have addressed that comment satisfactorily. I understand your concern and interest with exploring the lower bounds of snow model performance, etc., but this comparison between single-site and multi-site training is one of the most relevant aspect of the paper, considering the current literature in hydrology and LSTMs. In that sense, I really don't think that the addition you made after line 360 is sufficient. I would strongly encourage you to provide a more thorough comparison, as it would be of very high interest to the scientific community. On the same topic, I found Figure 8 extremely difficult to interpret, even just on its own, and impossible to compare with Figures 3 to 6. Please provide a figure that will allow the readers to directly compare the results of single site vs multi-site training. Theoretically, it seems like a good idea to train the LSTMs on multiple sites, as it provides them with more data, and more diversity. Is it the case in practice for data assimilation? This should be discussed. Also regarding Figure 8: please write ' probability density function ' instead of the acronym ' PDF ' or, even better, remove the title above the figure altogether, keeping only the caption below.

We acknowledge the point of the reviewer and we now provide a comparison between the best site-specific LSTM and the multiple site LSTM. We added a new section in the Results :

Comparing the multi-site LSTM DA with the site-specific LSTM DA trained over KHT, results show comparable performance for SWE, with neither approach consistently outperforming the other(see Fig. 9). In some cases, the site-specific model achieves lower errors, while in others the multi-site model performs equally well or slightly better. For snow depth, however, the multi-site LSTM DA tends to outperform the site-specific LSTM DA across most sites, although the improvements are generally modest (e.g. see Fig. 9 pannel d)

Then, in the discussion section, after line 544, we added this paragraph :

Additionally, considering a comparison between two approach, neither the site-specific nor the multi-site LSTM-DA consistently outperforms the other. While

multi-site training is theoretically expected to improve generalization by exposing the model to a broader range of conditions, this benefit is not clearly observed for SWE. A likely reason could be an uneven representation of sites, combined with variability in snowpack properties, meteorological drivers, and measurement methods, which may bias the model and introduce noise, leading to underfitting. Snow density also plays a crucial role; SWE is defined as $W = d\rho$, where W is SWE [kg/m^2], d is snow depth [m], and ρ is bulk snow density [kg/m^3]. A site-specific model such as KHT may implicitly capture a representative density evolution that transfers well across sites, whereas a multi-site model must attempt to generalize density dynamics across all environments, often with less accuracy. Overall, the multi-site LSTM-DA and the EnKF-DA perform similarly, with the latter only marginally better. This is encouraging, as it highlights the potential of the multi-site LSTM-DA to achieve comparable performance while substantially reducing the computational cost associated with ensemble-based methods.

Additionally, We modified the title on figure 8 (Now figure 9) accordingly to reviewer's suggestion.

5. Reviewer 2's general comment about the lack of details in the description of your experiments. I agree with this comment and I acknowledge the efforts that you already made to improve this. However, I think some important details are still missing or too vague:

- What is the programming language that was used to build the LSTMs? Any specific toolboxes or libraries?

To address this comment we added a sentence after line 282:

To develop the LSTM algorithm, we used Python 3.9.21 programming language and the open source libraries `Keras v.2.10.0` (Chollet et al. 2015) and `Scikit-learn v.1.1.1` (Pedregosa et al. 2011).

-In section 2.4.4: what is the length of the lookback window? Related to that, Rev 2 asked to ' please clarify the memory component (...) '. Normally, LSTMs do not need to be provided with data from previous timesteps, because of the lookback window (sometimes called lookback period, or just lookback). This is an important advantage of LSTMs compared to multilayer perceptrons, for instance, which have no memory. As their name indicates, LSTMs were designed specifically to have a memory. It is difficult to understand why you are adding a supplementary memory component. Including input variables form previous time steps is something we would typically do with a MLP, because of their lack of memory. Therefore, in addition to specifying the length of the lookback window, can you explain how adding more time steps for input variables is not redundant with the lookback window?

The lookback in the training phase is on average 10 years of hourly data (24x365x10), following the comments of Liu et al. (2015) who raised the concern that for long sequences the important information from the beginning of the sequence has to be dragged through the

whole sequence. Then during the testing the lookback is ether 1 or 2 hours (in the memory component configuration). So in that sense the memory component configuration name is used to stress a "longer" memory call rather than a shorter memory component usage. We acknowledge this may still be not so clear so we modified the sentences at line 311 as follows:

> "The second test configuration introduced an additional feature component to call back on the use of the "long" memory component of the LSTM during the operational test phase."

- Rev 2 had a specific comment about 'Furthermore, any LSTM prediction that fell below zero was forced back to the zero, effectively managing intermittent nature of snow data. '. You modified the sentence and moved it to lines 271-272, but in my opinion this is still not clear. How exactly was this 'forcing back to zero' performed? I guess that you determined specific dates for each one of the 7 study sites where there should not be any snow, and then for each year, between those dates you replaced the LSTM prediction by zeros? If this is the case, can you please explain how the dates were determined for each site?

The 'forcing back to zero' was a direct post-processing step applied to the LSTMs output. Any negative predictions for Snow Water Equivalent or snow depth was forced back to zero by applying the ramp function. This was a simple, per-timestep constraint on the model's output, not a process based on predefined snow-free periods or specific dates for each study site.

6. A point for the discussion: how to move from the current model to a spatialized version. As you mention, S3M is a distributed model. In the discussion, you briefly mention that ' These results open a window of opportunity for spatially distributed deep data assimilation; hence future work should focus on testing such a spatio-temporal water configuration. ' Could you please expand on what would need to be modified in the proposed data assimilated method in order for it to be applied in a spatially distributed way?

We thank the reviewer for the opportunity to further expand on this point and improve the clarity of our manuscript. We have added a comment after line 570 after the sentence -"LSTM robustness during dry and average water years further underscores the generalization capacity of such a framework.".

> "Using the LSTM as an emulator of the ensemble Kalman Filter allows us to significantly reduce the computational cost of ensemble-based data assimilation. This is particularly advantageous in a spatially distributed configuration, where running a full ensemble over large domains could otherwise be prohibitively expensive. In our current setup, ensembles are required only during the LSTM training phase, not at inference time — resulting in a more efficient approach for operational use.
>
> Preliminary tests with multi-site LSTM configurations have shown promising results: a single LSTM model trained on data from multiple locations can generalize well to other sites. Building on this idea, we envision extending the application of LSTMs from single-point setups to multiple representative points within a catchment. This spatially sparse assimilation could then be combined with an inter-

polation or spatial mapping strategy to propagate the correction across the entire domain.

Such an approach would provide a practical compromise between the need for spatially distributed corrections and the computational limitations of full-domain deep data assimilation. This transition from point-based to distributed correction —leveraging spatial generalization and interpolation — will be a key focus of our future work."

Minor comments, typos

- Figures A1 and A2: the text on those figures is much too small. Please make it bigger
- Line 1999 there is a space missing here ‘ by Reichle et al. (2007),Lanoy et al. (2010) ’
- Line 285: there is a missing space before the new part that was added in blue.
- Line 299: there is a missing comma before ‘ cycling between ’
- Figures 3-4-5-6, replace ‘ pan ’ by ‘ panel ’
- Beginning of section 3.1, line 386: remove one of the ‘ whiles ’ in ‘ while while ’

Comments addressed.

References

Avanzi, F., Munerol, F., Milelli, M., Gabellani, S., Massari, C., Girotto, M., Cremonese, E., Galvagno, M., Bruno, G., Morra di Cella, U. et al. (2024), 'Winter snow deficit was a harbinger of summer 2022 socio-hydrologic drought in the po basin, italy', *Communications Earth & Environment* **5**(1), 64.

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R. & Dozier, J. (2006), 'Mountain hydrology of the western united states', *Water Resources Research* **42**(8).

Chollet, F. et al. (2015), 'Keras', `https://keras.io`.

Dhoni, P. S. (2023), 'Enhancing data quality through generative ai: An empirical study with data', *Authorea Preprints* .

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Cohen, D. & Gilon, O. (2025), 'How to deal with missing input data', *EGUsphere* **2025**, 1–21.
URL: *https://egusphere.copernicus.org/preprints/2025/egusphere-2025-1224/*

Hartman, R. K., Rost, A. A. & Anderson, D. M. (1995), 'Operational processing of multi-source snow data', *Proceedings of the Western Snow Conference* **147**, 151.

Hock, R., Rasul, G., Adler, C., Cáceres, B., Gruber, S., Hirabayashi, Y., Jackson, M., Kääb, A., Kang, S., Kutuzov, S. et al. (2019), 'High mountain areas supplementary material', *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* .

Liu, P., Qiu, X., Chen, X., Wu, S. & Huang, X.-J. (2015), Multi-timescale long short-term memory neural network for modelling sentences and documents, *in* 'Proceedings of the 2015 conference on empirical methods in natural language processing', pp. 2326–2335.

Metref, S., Cosme, E., Le Lay, M. & Gailhard, J. (2023), 'Snow data assimilation for seasonal streamflow supply prediction in mountainous basins', *Hydrology and Earth System Sciences* **27**(12), 2283–2299.
URL: *https://hess.copernicus.org/articles/27/2283/2023/*

Pagano, T. & Sorooshian, S. (2002), 'Hydrologic cycle', *University of Arizona. Tucson. AZ. USA* .

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Zakeri, F., Mariethoz, G. & Girotto, M. (2024), 'High-resolution snow water equivalent estimation: A data-driven method for localized downscaling of climate data', *EGUsphere* **2024**, 1–30.