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 referrees, 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.

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 find 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 » specifically. 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.

2. Data quality: how do you mesure 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 mesured and provide some details?

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?

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 extrêmely 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.

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?
- 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?
- 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?

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?

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 »