Dear Dr. Matthieu Lafaysse,

Thank you much for your review and valuable comments. Below you will find your referee comments (in black) and our responses (in blue).

With regards,

Atabek Umirbekov, on behalf of all authors

General comments

Umirbekov et al. present a new machine learning approach to simulate snow mass with parcimonious data input and an extremely low numerical cost. The evaluation framework is really interesting as it includes independent data removed from the calibration dataset, but also the state-of-the-art ESM-SnowMIP dataset including challenging climate and environment conditions beyond those of the calibration dataset, and finally a spatialized application with more uncertain forcing data and evaluation data derived from remote sensing. Of course, the potential of machine learning has to be considered in snow modelling and I think this paper can be a significant contribution on that topic. The results clearly challenge physical models, even if obviously the output variables are not sufficient for all applications.

Nevertheless, I think the description of methods and results is sometimes a bit too fast in the current version of the manuscript and that some details are missing for an accurate understanding and interpretation of results. In general, figures are not really introduced in the main text. I would also have expected more in-depth discussion of the advantages and disadvantages of this approach compared to physical approaches and other machine learning approaches in the light of presented results and previous literature, and also discussions about the possibility to disentangle errors due to the forcing and to the algorithm itself. Maybe, the chosen structure of the paper that mixes results description and results discussion is partly responsible for this sometimes incomplete discussion. Finally the choice to try to recalibrate the Ts parameter is sometimes confusing especially when it's done on evalution datasets, as it leads to unrealistic values and overcalibration.

I also have some specific comments or questions below that can probably be addressed rather easily by the authors during the revision process.

We appreciate your overall feedback on the manuscript and are grateful for your valuable comments. In response to your suggestions, we will incorporate a comparison with physics-based snow models using proceedings from the SnowMIP2 study (Krinner et al., 2018), and include a discussion of the advantages and disadvantages of the model in comparison with other physical snow models. This discussion will be also complemented with comparisons to other machine learning applications in snow modeling. We will introduce all figures in the main text for better context. Furthermore, we will restructure the manuscript to ensure that results and their corresponding discussions are appropriately organized in their respective sections. Finally, we will provide a more detailed description in the main text regarding the precipitation-snow partitioning and the 'Ts' parameter to prevent any confusion with traditional temperature-based partitioning methods.

Detailed comments

Section 2.1 The choice of SVR relatively to other machine learning algorithms is not discussed. I would suggest to add a quick summary of advantages and disadvantages compared to the most classical algorithms available in literature (random forests, convolutional neural network, simpler regressions, etc.)

Thank you for this suggestion. We will enhance the introduction by incorporating a brief summary of machine learning application to snowpack modeling. In addition, we suggest to include a brief paragraph into the discussion section summarizing some our experiments and hypotheses. This passage would include (but might be not limited to) the followings:

"We have evaluated several other data-driven techniques for the model development, including linear regression, Random Forests, Gradient Boosting Machines, Gaussian process (not shown here). When compared to the training dataset, the performance of most models was equivalent; however, their accuracy on the evaluation dataset was worse. Experiments in other fields indicate that SVR has relatively better extrapolation potential on unseen data (Horn and Schulz, 2011; Kim and Kim, 2019), which may explain why it outperformed other algorithms. We haven't examined neural network algorithms since they take more computer resources during training, and evidence suggests that they tend to underperform relative to other machine learning ML techniques when applied to tabular data (Borisov et al., 2022). To make definitive judgments in this regard, however, would require a more extensive intercomparison experiment that is outside the scope of this paper."

Can you define more explicitely i, j, N, xi, xj, X ?

We will incorporate more explicit notations for the variables and parameters denoted in the main formula (1).

I understand from Fig.1 and Eq. 2 that when temperature is below the -1°C threshold and precipitation is zero, then dSWE is always equal to 0. Is that correct ? How often does this assumption fail in the training or evaluation dataset? Does this imply an intrinsic limitation of GEMS for transferability on steep slopes where the surface energy balance can be positive even at negative temperatures ? (I think it does.)

Thank you for these guiding questions. Yes, when temperature is below the -1°C threshold and precipitation is zero, dSWE will be automatically set by the model code to 0 mm. Due to the temperature threshold and not using/estimating solar balance, the model does not capture snow melt and sublimation which. We recognize this as one of the model limitations (particularly compared with most energy-balance snow models). We will note this limitation in the respective part of the manuscript accordingly.

Section 2.2

The authors say they « fine-tuned the hyperparameters so that the model produces similar levels of accuracy when applied to observations from the same stations for 2019 and 2020. » I understand the general idea but the detailed procedure is not accurately described. Can you describe the detailed protocol for this « fine-tuning » ?

We will complement the respective lines of the manuscript with the following details:

"The hyperparameter calibration process involved an exhaustive 'grid-search' technique, which systematically explored all possible combinations within predefined parameter ranges. Ultimately, we selected the hyperparameter configurations that resulted in the lowest root mean squared error between simulated and observed dSWE during both model training on observations from 2017-2018 and its testing on observations from 2019 and 2020."

As solid precipitation measurements are prone to large measurement errors and is one of the main predictor of the model, I would have expected more details about precipitation gauges used in the SNOTEL network, procedures applied to account for undercatch, and if possible estimated uncertainties.

We suggest to include the following passage into the Data section:

"*SNOTEL stations primarily utilize tipping bucket-type precipitation gauges, which are reported to have an accuracy range between 2% and 5%, depending on the type (USDA, 2010). It is also recognized that precipitation gauges are susceptible to solid precipitation undercatch, especially when snowfall occurs in windy conditions (USDA, 2014). Scalzitti et al., 2016 provide a comprehensive review of the issues associated with precipitation undercatch, highlighting reported undercacth ranging from 11% for snowfall under 2m/sec wind speed to more than 30% during intense snowstorm events.*"

Section 3 I think « Model evaluation » would be a more appropriate title than « model validation » as a model can never be considered as fully validated.

Thank you for this suggestion. We agree and will change the section to "Model evaluation".

The authors say « we excluded stations that exhibit precipitation undercatch, which we formulate as when SWE accumulated by March is greater than the accumulated precipitation during October to March. ». I would expect all stations to be affected by precipition undercatch and total SWE to be always higher than raw precipitation measurements. Do you apply a specific threshold to only eliminate major undercatch ? Or do you use precipitation timeseries that are already corrected for precipitation undercatch following WMO recommendations ? My misunderstanding is probably linked to the lack of details in Section 2.2 as previously mentioned.

Then, was this selection procedure also apply to the training dataset ? If not, why ?

We appreciate these comments and questions. We agree that this part needs more clarifications, and suggest to include some additional explanatory exerts. These may include (but are not limited to) the following:

"While SNOTEL stations may be susceptible to precipitation undercatch, especially during intense snowfall events and high winds (Scalzitti et al., 2016), it is essential for machine learning to have accurate training data. To ensure data accuracy, we cleaned the training dataset by removing observations with inconsistencies between daily precipitation and snow mass accumulation. These inconsistencies refer to cases when the daily increase in SWE exceeded the reported daily precipitation.

The selection approach differed for the evaluation dataset because we aimed to retain as many stations as possible for evaluation and besides that the model requires complete daily time series without missing observations. We therefore used aggregated

precipitation sums from October to March to compare with accumulated SWE by March. This approach enabled us to include more stations in the evaluation dataset while excluding only those hydrological years that exhibited inconsistencies between these variables. When using this filter, we did not set any specific threshold for the magnitude of inconsistencies, nor did we make corrections to the precipitation time series."

It should be also noted that we included a criterion that required at least five hydrological years of observations for a station to be part of the evaluation dataset (line 150). Consequently, some SNOTEL stations were excluded based on this specific requirement.

Section 3.1

L193 I would suggest to start by a sentence presenting the Figure before providing its interpretation.

Yes, we will introduce the Figure in the text before its interpretation.

In Figure 4 « actual » should be replaced by « observed ». Is there a reason to present the simulations in the X axis and not in the Y axis (that would be more common for a scatter plot) ?

Thank you for pointing at this. We will replace 'actual" with "observed". We will also swap current X and Y axis accordingly.

In Figure 5, it is not immediate to understand what is represented because the caption is not self-sufficient and the description in the text is also too vague. The definition of TAVG should be remind in the caption. Then what does represent a single point ? A station and a date ? Then, this solid fraction of precipitation does not really appear in model description, neither in Figure 1 neither in the Equations, so it is difficult to understand how this diagnostic is obtained from the provided model description. The reason for providing this Figure is also unclear as finally these outputs are not really used as a fixed temperature threshold finally replaces the values obtained by the algorithm. This needs to be clarified.

We appreciate your suggestion to incorporate additional clarifications into this section. We will add more details under the Figure caption. Furthermore, we will add more clarifications into the text, such as the following:

"The Ts constraint differs from classical temperature-based partitioning methods where the threshold define precipitation in a binary way as either 100% rainfall or 100% snow.

The model simulates snow-precipitation partitioning using its inherent learned algorithm, but only until the temperature drops below Ts. At that point, any precipitation is regarded as 100% snow. For example, when the average temperature (TAVG) is 0°C, the model will likely simulate a portion of precipitation as snow. As an example, as illustrated in Figure 5 at TAVG at around of 0°C, the model is likely to simulate a significant portion of precipitation as snow (around 75% of precipitation) even if Ts is -1°C."

As for the other Figures, introducing quickly Figure 6 would be helpful before providing the results analysis. In the description of the results of Figure 6, detailed references to the subplots would help to follow results description.

We will introduce the Figure 6 in the text.

Isn't the maxSWE score more representative of the quality of input precipitation than of the skill of the SVR model ?

Yes, given the temperature threshold, we assume that maxSWE might be more representative of precipitation input accuracy. However, since a portion of the simulated maxSWE is influenced by the model's simulation of dSWE (at temperatures above the Ts threshold), we think it is reasonable to keep maxSWE as one of the metrics.

L252-254 If removing stations with incorrect measurements is understandable, removing stations with snow drift should be avoided as snow drift is not a measurement error, it's a natural process challenging to reproduce with physical models and also maybe with machine learning models, but the general ability or unability of any model to reproduce snow conditions should account for places where snow drift happen.

Thank you for raising this concern. Instances where recorded maxSWE exceeds accumulated precipitation may be due to snow-drift, precipitation undercatch, or a combination of both factors. Unfortunately, attributing these inconsistencies to individual factors may require a separate research effort. We therefore had to exclude those stations from the evaluation, but we note the inability of the model to capture snow-drift in line 263-265 and explicitly state this as a model`s limitation in line 407-408.

L255-256 You mean that an overcalibration is obtained due to error compensation between snow drift and rain-snow transition ? Could the sentence be more clear ?

Thank you. Yes indeed, we meant that overcalibration may lead to error compensation. We will make it more explicit in a new version of the manuscript.

Section 3.3

Again, an introduction of Figure 8 in the text would be useful.

We will introduce the Figure 8 in the text.

L266-267 It is not obvious which value of NSE should be considered as « acceptable ». Indeed, NSE is easily high when dealing with variables with a high seasonal cycle. What would be the NSE value of the daily interannual mean of observed SWE ? Is the 0.7 value at Sapporo better than such a reference score ?

We intended to refer to some categorizations of NSE across multiple studies (e.g. N. Moriasi et al., 2007). However, we recognize that these classifications, designed for hydrological models, might not be directly applicable for classifying snow model outputs. Therefore, we will revise sentences with qualitative classifications like this one.

L269-270 This could be moved to the Method section

We agree with and will implement this suggestion.

L274-280 As it was already noticed with the SNOTEL dataset that local calibration of the Ts threshold leads to severe error compensations, and as the purpose of the application of the GEMS system on the ESM-SnowMIP dataset is to assess its spatial transferability beyond its training dataset, I am not really convinced of the interest to test again to recalibrate locally this threshold on each ESM-SnowMIP site. The conclusions that again this leads to overcalibration and errors compensations were rather expected, so I would suggest to remove thisanalysis.

Thank you for these insights and the suggestion. We acknowledge that calibrating the Ts threshold may result in error compensations. However, the results do not provide insights into the extent of these compensations. As previously described, the Ts threshold in the model differs from the classical temperature-based threshold method. For instance, when Ts is set at -3°C and temperature (TAVG) is 0°C, the model will likely classify a larger portion of precipitation as snow (Figure 5). Nevertheless, we recognize that calibration in general might be inappropriate when assessing the spatial transferability of the model. Hence, we ´will remove the calibration analysis for Snow-MIP stations from the manuscript.

Apart from model evaluation, calibration could still be useful during model application, particularly when local precipitation-snow partitioning patterns are known. In light of this, we suggest the following: 1) more explicitly acknowledge the risk for error compensation due to calibration in the model limitations section, and 2) recommend adhering to the default value of -1°C unless the local precipitation-snow partitioning patterns are known."

Section 3.4

Again an introduction of Figure 9 is missing.

We will introduce Figure 9 in the text.

My feeling is that the level of discussion in this section is not as advanced as for the evaluation on ESM-SnowMIP sites. How does this skill in terms of snow cover extent compare with physical models ?

Thank you for this suggestion. We acknowledge that this section's content is not as comprehensive as other sections, particularly in terms of comparison with the performance of physical models. However, the extensive computational burden for such a comparison present a significant challenge to us.

The primary objective of this section is to test and demonstrate the model's transferability to regions with complex terrain and lacking in-situ SWE data. We assume that if the extent of the simulated SWE aligns well with the remote-sensed snow cover, then the simulated SWE is likely to contain less uncertainty. This assumption is also based on fact that "*remote sensed snow cover is increasingly used for parameter calibration or uncertainty reduction in snow modules of hydrological models (e.g. Parajka and Blöschl, 2008; Gyawali and Bárdossy, 2022; Tong et al., 2022; Di Marco et al., 2021).*"

We would like to suggest to elaborate on these points in the new version of the manuscript.

Section 4.1

L312 Reference error.

This was intended as a reference to Figure 10. We will correct this in a new version of the manuscript and appropriately introduce Figure 10 in the text.

L315 Could the relatively low contribution of the heat-insolation index be possibly explained by an unsufficient variability of this predictor in the training dataset ?

Yes, this is what we intended to state. We will revise the sentence making this clearer.

In mountainous areas, shadows and slope inclinations are a major factor to explain melting. But I assume that all observations correspond to flat areas, and maybe the variability of shadows in the SNOTEL network is neither representative of the variability of topographic conditions in mountains. This is important to discuss as it could limit the possibility to apply this algorithm on areas with complex topography.

We appreciate these comments and suggestions. Indeed, the SNOTEL stations utilize flatbed pillows, but are primarily situated in mountainous regions. As discussed in Section 4.2 and illustrated in Figure 11, the model demonstrates relatively better performance in mountainous areas compared to lower elevations. However, when examining the histogram of CHILI values in Figure 2, it becomes apparent that the training dataset may be less representative of locations with lower CHILI indices. These lower indices often correspond to sites significantly shadowed by terrain or situated at higher latitudes or both. This discrepancy poses an additional potential source of model uncertainty, a point we will further discuss in the section on model limitations.

Section 4.2

I am wondering how much this conclusion is affected by the choice of NSE to quantify errors. Indeed, as this score is highly influenced by the existence of a seasonal cycle, it is rather normal to get better scores with deeper snowpacks that exhibit a very strong seasonality than on sites with more intermittent snow cover. Considering other scores (for instance a Root Mean Square Error), I would not be surprised that stations with the poorest performance would be reversed. Can you comment on that topic ?

Thank you for your guiding questions. We agree that NSE alone may not adequately distinguish between cases of 'good' and 'poor' model performance, and use of different metrics would likely result in varying compositions of these two performance groups. We assume that Root Mean Square Error may serve as a more suitable alternative for comparing the model performance across the stations. Consequently, we suggest to revisit the analysis in this section using RMSE (probably comparing lowest and highest quartiles) and incorporate it as an additional metric in the evaluation section.

L375 The authors say that « GEMS also addresses the equifinality issue that is pertinent to hydrological and snow modelling. » but the only parameter they have introduced (Ts threshold) clearly raises a very strong equifinality resulting in possible overcalibration to compensate various possible errors including snow drift, precipitation undercatch, etc.

We assume that this sentence is now justified, considering the preceding explanation of how Ts works in the model, how it differs from temperature-based partitioning methods, as well as our intention to stick to the default Ts in our recommendations.

In this sentence we refered to the challenge of calibrating multiple parameters in hydrological and snow modelling. This challenge is particularly prominent in hydrological modeling, where even relatively simple snow modules require calibration of at least two parameters: the precipitation-snow threshold and the degree factor. Considering that there are many other parameters for different components of a hydrological model, it would be easy to end up with multiple combinations of optimal model parameters. We hypothesize that replacing the snow module with a model that is based on generalizable empirical relationships may help to reduce the equifinality issue, especially when employing conceptual hydrological modeling.

L388 « GEMS can, for instance, provide information for the parameterization of physics-based models, e.g. precipitation phase partitioning and its elevational dependence ». I don't see how the results presented here suggest this conclusion and considering the strong risk of overcalibration of this Ts value (leading to clearly unrealistic values below -5°C), I am not convinced at this point that GEMS could help me to discriminate between snow and rain.

As mentioned earlier, we acknowledge that calibrating Ts poses a risk of error compensation, though considering how Ts operates in the model, the extent of overcalibration maybe not as pronounced as it would be with traditional temperature-based thresholds. Despite this, we recognize that the statement in this sentence may have been too assertive and requires further verification. We will remove this sentence from the manuscript.

There is a section 5.1 but not any section 5.2. Maybe a subtitle for the first part of Section 5 is missing.

As it was also recommended by Reviewer 2, we would like to separate section 5 into two separate sections in a new version of the manuscript: section 5 'Model Limitations' and section 6 'Discussion'.

L393-400 The authors discuss the limitations of their approach relatively to forest areas but they seem to have intentionnally remove the 3 forest sites of the ESM-SnowMIP dataset from their evaluations. This should at least be discussed if there is a valid reason for that. But even if the model skill is lower on the 3 Canadian forest sites, I would have included these sites in the evaluations to provide concrete results to support this discussion.

Indeed, we haven't evaluated the model on the three Canadian sites but because at that time we couldn't precisely locate the sites to determine CHILI parameters. We will include these sites for the model evaluation in the next version of the manuscript.

L408-410 Unfortunately, blowing snow can be an important process even at large scale especially in polar regions. So large scale applications of the system may still be affected by this limitation.

We will remove that part of the sentence.

The discussion do not compare the skill of this approach with the skill of physical models while similar metrics are provided at the same sites in Ménard et al., 2021, and other evaluations are also available in the literature for snow cover extent. I think this would be important to consider as well.

We appreciate this suggestion. We will compare the skill of the model with that of physical models that participated in ESM-SnowMIP, using model simulations presented in Krinner et al., 2018.

The discussion or final summary also lack comments about the strengths and weaknesses of their results compared to the literature cited in the introduction applying machine learning to predict snow mass.

We will present our perspective on the strengths and weaknesses of our model approach compared to other cases of snow models utilizing machine learning.

Furthermore, the outputs of the model are currently limited to SWE while several snow-sensitive applications require more variables (e.g. surface temperature for NWP and climate modelling, snow internal properties for remote-sensing retrieval

algorithms or avalanche forecasting). This limitation should also be mentioned with possibly discussions about the feasability to extend this approch to more variables.

Thank you for this suggestion. We will include this as a limitation of our model and complement it by presenting our perspective on the snow processes to which our approach may be applicable.

References:

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G.: Deep Neural Networks and Tabular Data: A Survey, IEEE Trans. Neural Networks Learn. Syst., 1–21, https://doi.org/10.1109/TNNLS.2022.3229161, 2022.

Gyawali, D. R. and Bárdossy, A.: Development and parameter estimation of snowmelt models using spatial snow-cover observations from MODIS, Hydrol. Earth Syst. Sci., 26, 3055 – 3077, https://doi.org/10.5194/hess-26-3055-2022, 2022.

Horn, J. E. and Schulz, K.: Spatial extrapolation of light use efficiency model parameters to predict gross primary production, J. Adv. Model. Earth Syst., 3, https://doi.org/https://doi.org/10.1029/2011MS000070, 2011.

Kim, M. and Kim, J.: Extending the coverage area of regional ionosphere maps using a support vector machine algorithm, Ann. Geophys., 37, 77–87, https://doi.org/10.5194/angeo-37-77-2019, 2019.

Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., Hall, A., Rott, H., Brutel-Vuilmet, C., Kim, H., Ménard, C. B., Mudryk, L., Thackeray, C., Wang, L., Arduini, G., Balsamo, G., Bartlett, P., Boike, J., Boone, A., Chéruy, F., Colin, J., Cuntz, M., Dai, Y., Decharme, B., Derry, J., Ducharne, A., Dutra, E., Fang, X., Fierz, C., Ghattas, J., Gusev, Y., Haverd, V., Kontu, A., Lafaysse, M., Law, R., Lawrence, D., Li, W., Marke, T., Marks, D., Ménégoz, M., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Raleigh, M. S., Schaedler, G., Semenov, V., Smirnova, T. G., Stacke, T., Strasser, U., Svenson, S., Turkov, D., Wang, T., Wever, N., Yuan, H., Zhou, W., and Zhu, D.: ESM-SnowMIP: assessing snow models and quantifying snow-related climate feedbacks, Geosci. Model Dev., 11, 5027–5049, https://doi.org/10.5194/gmd-11-5027-2018, 2018.

Di Marco, N., Avesani, D., Righetti, M., Zaramella, M., Majone, B., and Borga, M.: Reducing hydrological modelling uncertainty by using MODIS snow cover data and a topography-based distribution function snowmelt model, J. Hydrol., 599, 126020, https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.126020, 2021.

N. Moriasi, D., G. Arnold, J., W. Van Liew, M., L. Bingner, R., D. Harmel, R., and L. Veith, T.: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, Trans. ASABE, 50, 885–900, https://doi.org/https://doi.org/10.13031/2013.23153, 2007.

Parajka, J. and Blöschl, G.: The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models, J. Hydrol., 358, 240–258, https://doi.org/10.1016/j.jhydrol.2008.06.006, 2008.

Scalzitti, J., Strong, C., and Kochanski, A. K.: A 26 year high-resolution dynamical downscaling over the Wasatch Mountains: Synoptic effects on winter precipitation performance, J. Geophys. Res. Atmos., 121, 3224–3240, https://doi.org/https://doi.org/10.1002/2015JD024497, 2016.

Tong, R., Parajka, J., Széles, B., Greimeister-Pfeil, I., Vreugdenhil, M., Komma, J., Valent, P., and Blöschl, G.: The value of satellite soil moisture and snow cover data for the transfer of hydrological model parameters to ungauged sites, Hydrol. Earth Syst. Sci., 26, 1779 – 1799, https://doi.org/10.5194/hess-26-1779-2022, 2022.

USDA: Chapter 2 Data Parameters, in: Snow Survey and Water Supply Forecasting National Engineering Handbook, 2010.

USDA: Chapter 6 Data Management, in: Snow Survey and Water Supply Forecasting National Engineering Handbook, 2014.