Summary

Pomarol Moya et al. in "Improving forecasts of snow water equivalent with hybrid machine learning" evaluate various machine learning (ML) based approaches in representing spatiotemporally in sample and out of sample predictions of daily snow water equivalent (SWE) across 10 measurement sites in the Northern Hemisphere, derived from the ESM-SnowMIP project, across 7-20 years. The ML-based estimates are compared to and, in some cases, informed by a physics-based snow model, Crocus. The analysis shows that ML-based models can benefit from learning about daily SWE behavior from both observations and the physics-based model, sometimes helping to offset physics-based snow model errors (e.g., snowmelt rate/snow off date). The order of importance of variables that influence the SWE prediction is also intercompared and rank ordered, many of which relate to snowpack thermodynamics, and could (potentially) be used to inform physics-based model development.

Overall, I think the paper fits within the scope of the Cryosphere and could be, given more work, a valuable contribution. ML-based methods have grown in popularity in recent years, and ML model development/sensitivity analyses like these help to inform where/when ML-based methods are or are not fit for purpose in predicting SWE spatiotemporally. However, I think there are still several major revisions that need to happen prior to this paper being accepted. While I appreciate the authors' thorough analysis, the underlying data (ESM-SnowMIP station network) is quite sparse in space/time and makes me worry about the extensibility of their findings beyond the limited station locations and years assessed. I respect that the authors provided an entire section (Section 4.1) that discusses this very point, but I feel like a more thorough decomposition (e.g., snow climate, elevation, land surface heterogeneity, etc.) of station differences is still needed (given that only a few stations are used to train and assess ML model fidelity). I also think the authors could try and provide more take home messages for the physics-based modeling community from the ML-based results on which variables/processes to target (e.g., fix the long-standing snowmelt rate/snow off date biases in physics-based snow models). Too often it feels like ML-based papers try to show how they can outperform physics-based models rather than how ML-based models/methods can be used to advance physics-based model development. This is particularly salient given that ML-based models poorly predict out of sample in space/time and under different climate scenarios and, therefore, physics-based models appear that they will be needed for the foreseeable future. I have provided, hopefully, constructive comments and suggested edits below for the authors to consider.

Dear reviewer, thank you for your interest in our paper and for your detailed and extensive review. Regarding the main limitations highlighted in this paragraph; we agree that a decomposed analysis given the small size of the dataset could provide more valuable insights, so we will expand on that in our revision of the manuscript. On the other hand, while we hope that our paper is useful for the physics-based community as well, it is important to note that the focus of our paper is to showcase the performance of hybrid models and more generally the benefits of including physical knowledge into ML models, rather than the advancement of purely physics-based modelling. We hope to sufficiently answer all points raised in more detail below.

Comments and suggested edits

Line 24 – cite "The cryosphere has a large impact on the Northern Hemisphere…", maybe with this study…

Huss, M., Bookhagen, B., Huggel, C., Jacobsen, D., Bradley, R.S., Clague, J.J., Vuille, M., Buytaert, W., Cayan, D.R., Greenwood, G., Mark, B.G., Milner, A.M., Weingartner, R. and Winder, M. (2017), Toward mountains without permanent snow and ice. Earth's Future, 5: 418-435. https://doi.org/10.1002/2016EF000514

This paper describes the critical role of the cryosphere in several aspects of the mountain regions, including human livelihood, economy, and ecosystems, and discusses the potential impact of climate change. Thank you for the suggestion; it provides a valuable reference to stress our point, so we will add it.

Line 28 – cite "…due to its spatio-temporal variability…", maybe with this study…

Alonso-González, E., Revuelto, J., Fassnacht, S. R., & López-Moreno, J. I. (2022). Combined influence of maximum accumulation and melt rates on the duration of the seasonal snowpack over temperate mountains. Journal of Hydrology, 608, 127574

This paper discusses the influence of accumulated snow (i.e., peak SWE) and melt rate in snowpack duration for the mountainous areas in the Iberian Peninsula. It does mention the interannual variability of the snowpack and some of its causes, so we will add it. Furthermore, we will expand the literature regarding that claim (Deems, Fassnacht, and Elder 2006; Grünewald et al. 2010).

Deems, Jeffrey S., Steven R. Fassnacht, and Kelly J. Elder. 2006. 'Fractal Distribution of Snow Depth from Lidar Data'. *Journal of Hydrometeorology* 7(2):285–97. doi:10.1175/JHM487.1.

Grünewald, T., M. Schirmer, R. Mott, and M. Lehning. 2010. 'Spatial and Temporal Variability of Snow Depth and Ablation Rates in a Small Mountain Catchment'. *The Cryosphere* 4(2):215–25. doi:10.5194/tc-4-215-2010.

Line 34 – add "machine learning (ML)" as this is the first time it is introduced/defined

Thanks for noticing, it will be added.

Line 36 – "find non-linear structure" – can machine learning only identify non-linear structures or both linear and non-linear?

It refers to both linear and non-linear structures. We will rephrase the sentence by stating that it is not limited to linear ones.

Line 39-40 – you might also include this citation…

Song, Y., W. Tsai, J. Gluck, A. Rhoades, C. Zarzycki, R. McCrary, K. Lawson, and C. Shen, 2024: LSTM-Based Data Integration to Improve Snow Water Equivalent Prediction and Diagnose Error Sources. J. Hydrometeor., 25, 223–237, https://doi.org/10.1175/JHM-D-22-0220.1

This paper implements an LSTM model to predict SWE where lagged observations of either SWE or satellite-observed snow cover fraction are used as predictors. Hence, it is a good addition to the provided literature on that topic and will be included in the next version of the manuscript.

Line 45-46 – this sentence needs a citation for this bold statement. Couldn't the ML models inherent and amplify biases learned from the physics-based models? Also, is there peer-reviewed evidence that ML models can skillfully produce "out of sample" predictions from one mountain/seasonal snow region to another?

There are many examples in the literature that highlight the potential benefits of using hybrid models. For instance, Karpatne et al. (2017) suggest that they may improve consistency with scientific knowledge and produce more generalizable models.

Hybrid models can certainly inherit biases from the physics-based model, so this may introduce some error, but it is precisely because they incorporate observations that bias is mitigated, so long as the observational dataset is representative of the inference domain.

Examples of the ability of hybrid models to extrapolate to untrained locations can be found for hydrological tasks such as streamflow forecasting (e.g., Konapala et al. 2020; Magni et al. 2023), but also for SWE forecasting (Steele et al., 2024).

We will expand the statement along the lines of this answer, adding more references as well.

Karpatne, Anuj, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. 'Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data'. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2318–31. doi:10.1109/TKDE.2017.2720168.

Konapala, Goutam, Shih-Chieh Kao, Scott L. Painter, and Dan Lu. 2020. 'Machine Learning Assisted Hybrid Models Can Improve Streamflow Simulation in Diverse Catchments across the Conterminous US'. *Environmental Research Letters* 15(10):104022. doi:10.1088/1748-9326/aba927.

Magni, Michele, Edwin H. Sutanudjaja, Youchen Shen, and Derek Karssenberg. 2023. 'Global Streamflow Modelling Using Process-Informed Machine Learning'. *Journal of Hydroinformatics* 25(5):1648–66. doi:10.2166/hydro.2023.217.

Steele, Hannah, Eric E. Small, and Mark S. Raleigh. 2024. 'Demonstrating a Hybrid Machine Learning Approach for Snow Characteristic Estimation Throughout the Western United States'. *Water Resources Research* 60(6):e2023WR035805. doi:10.1029/2023WR035805.

Line 60 – change "features" to "conditions"

Good suggestion, we will include it.

Line 71-72 and Line 77-79 – are 10 stations with 7-20 years of measurements enough to properly sample intra- and inter-annual variability of snowpack lifecycles across the Northern Hemisphere? Also, worryingly, only three of the stations are automatic and the others "only [have] manual measurements at irregular intervals". How many snow climates (Sturm and Liston, 2021), elevations, etc. are represented across these stations? Could the authors provide a map plot with automated/manual station lat/lon locations?

Sturm, M., and G. E. Liston, 2021: Revisiting the Global Seasonal Snow Classification: An Updated Dataset for Earth System Applications. J. Hydrometeor., 22, 2917–2938, https://doi.org/10.1175/JHM-D-21-0070.1.

We will expand the description of the station characteristics when describing the data and comment on its representativeness for the Northern Hemisphere in the discussion.

This dataset was compiled for a model intercomparison project, so it does cover a wide range of snowpack conditions. The station locations and characteristics are plotted in Figure 1 below. A vast geographical area across the Northern Hemisphere is covered, although there is some clustering and oversampling in North America. In terms of elevation, only very high-altitude regions (above 4000 m) are missing, despite the automatic stations used for training reaching only up to 2000 m. Finally, most of the snow climates from the provided reference (Sturm and Liston, 2021) are represented, although not all are covered in the manual stations used for testing.

Nevertheless, it is important to note that this is a methodological paper; it seeks to examine the potential of hybrid models, rather than creating a final, model-based SWE product. So, for this purpose, the provided dataset is sufficient. For further discussion into the reasons for choosing this dataset, please refer to the answer to question at L72 from the response to the first reviewer (https://doi.org/10.5194/egusphere-2025-1845-AC1).

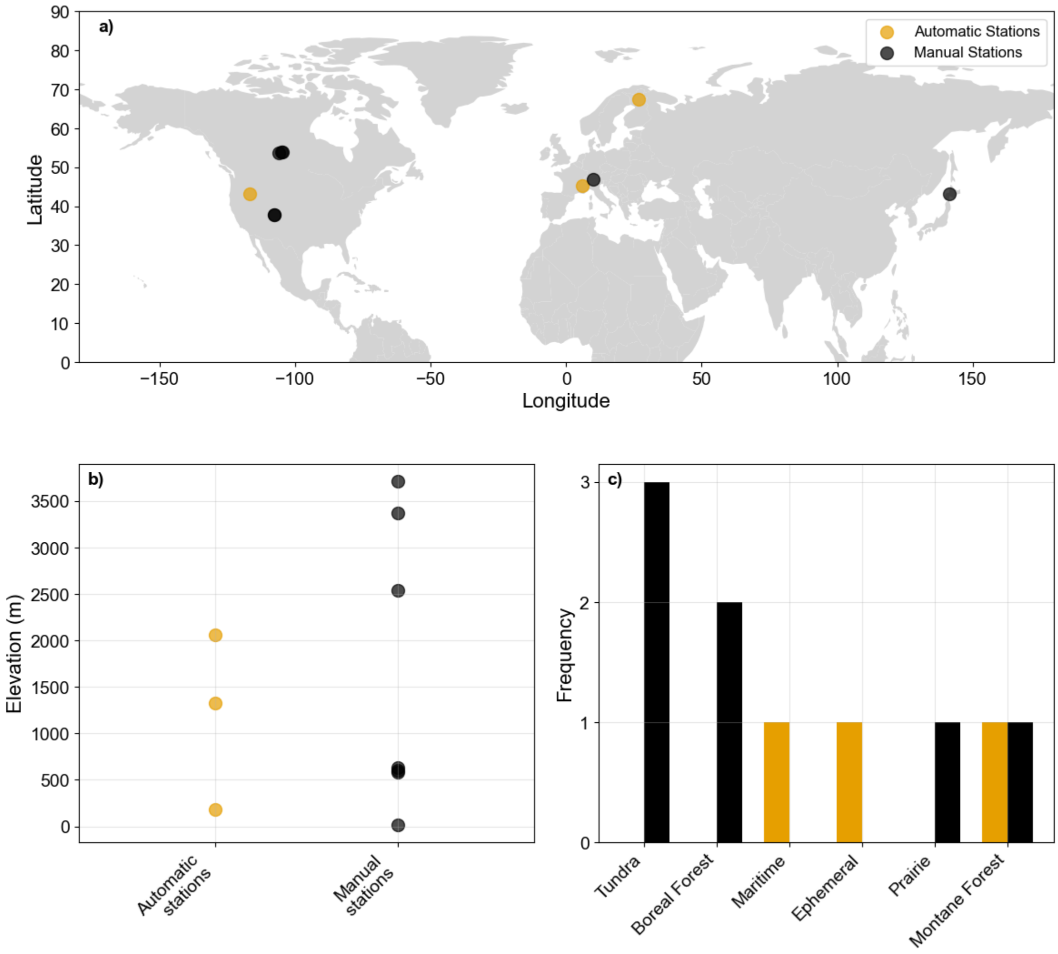

Figure 1: Representation of the a) geographical location, b) elevation and c) snow climate distributions for both automatic and manual stations.

Line 76 – change "snow water equivalent" to "SWE"

We will change it accordingly.

Line 82-83 – what does it mean that "aggregation methods were performed for some variables according to expert knowledge"? Can you provide readers with the physical basis/intuition for how each of these various aggregation methods for meteorological variables impacts a snowpack's energy/mass balance? This is needed to ensure that the ML method is learning and estimating snow physics for the right reasons.

This sentence aims to convey that variable selection and aggregation methods was not based on data-driven approaches, but rather on their expected influence in snow dynamics according to expert knowledge. We will clarify and expand the aggregation methods in the relevant section. For example, the time integral of positive air temperatures in Celsius is related to the positive degree days, used since decades in the snow modelling community (Hock, 2003).

Hock, Regine. 2003. 'Temperature Index Melt Modelling in Mountain Areas'. Journal of Hydrology 282(1):104–15. doi:10.1016/S0022-1694(03)00257-9.

Line 92 – change "50 layers" to "50 snow layers". Also, I might mention snow temperature or some other thermodynamic variable (given the mention of "energy and mass balance" in the previous sentence(s).

We propose to change that sentence to: "It dynamically adjusts up to 50 snow layers to represent a vertically discretized snow temperature, density and liquid water content profile, and provides a comprehensive evolution of the snow microstructure, thus giving a vision of the snow stratigraphy and its temporal evolution"

Line 99 – change "layer" to "snow layer". Also, does "layer information" mean the dynamic ranges of snow depth when delineating the 50 snow layers over space/time as the snowpack lifecycles evolve?

We will change it accordingly. Layer information means any of the variables reported by Crocus individually for each snow layer rather than for the whole snowpack, in which case we refer to it as "bulk information". We will clarify that in the text.

Line 101 – I would delete "2100 J/kg*K" as it seems like TMI (if an equation to compute cold content is not shown).

The equation is not explicitly written but was described in text, therefore it was deemed relevant to mention the specific heat of ice used for its calculation. However, since multiplying the predictors of a random forest model by a constant factor does not affect its performance, this is indeed not very relevant and will be removed.

Line 103 – why are most of the variables used to train the ML model daily averages? Was there a sensitivity analysis performed that is not mentioned here? For example, wouldn't minimum (e.g., nighttime) or maximum (e.g., daytime) temperatures be important too given that the snowpack might refreeze or quickly melt depending on the range of temperatures experienced in a given day? On Line 114-115 you also mention how there can be a delay in the response of the snowpack (presumably from the erosion of cold content before a phase change occurs) over a 14-day period from the given day. This also seems to be an argument that some information might be contained in minimum/maximum/etc. of meteorological variables.

As mentioned above in the answer to the Line 82-83 question, the choice of input variables was made based on expert knowledge and no specific sensitivity analysis was performed. The objective was to set the basis of a first model able to capture the most salient features of the SWE dynamics, and there are certainly several avenues of improvement that could be considered; refining the predictors is one of them.

We considered the average to be a good initial approach to aggregate the meteorological variables, but we also performed different aggregations than solely averages. For instance, Tair_int is the daily integral of positive temperatures in Celsius, a well-referenced predictor for snow melt (Hock, 2003), or in its defect, cold content erosion. This led us to discard other aggregation methods which seemed less promising like minimum and maximum, but it does not mean that they might not be useful predictors.

We completely agree that further testing of variable choices and aggregation methods, such as the ones proposed here, remains an interesting research direction for future studies, and we will discuss this limitation in the Discussion section.

Figure 1 – is there a reason that different brackets are used "[]" and "()" to describe time (t)?

Yes, they refer to how the variables were aggregated. We discussed their meaning in the comment regarding Figure 1 in the response to the first reviewer (https://doi.org/10.5194/egusphere-2025-1845-AC1). As we also state there, it certainly requires clarification.

Line 117 – "consecutive daily SWE measurements are available, that is, the automatic stations" does that mean you completely "throw out" seven of the 10 stations data? If so, I am even more worried about properly sampling intra-annual and inter-annual variability of snowpack lifecycles across the Northern Hemisphere. 1874 days (~5 years) is not very much data to train the ML model on purely observations of SWE/dSWE. A biggest question, can you more clearly state how the manual measurements are used then?

That is correct; besides the data augmentation (AUG) approach, only the three automatic stations were used for training. The manual measurements were used exclusively for testing purposes in the station split. A diagram explaining the use of manual measurement according to ML model and temporal or station split is provided in Figure 2 in the paper.

Regarding the lack of data, it is important to note that while we only have 1874 samples, those exclude periods without snow, therefore effectively represent much more than five years of data.

Lastly, the success of the AUG strategy, even with limited observational data, highlights its usefulness for forecasting data-scarce variables like SWE. However, we will add that our findings are restricted to stations/climates covered by this dataset in the Discussion.

Line 139 – so you are splitting 1874 days of data into train, validation and test? Are manual measurements used for training, validation, and/or testing too?

Both manual and automatic measurements are considered as ground truth, but the former are only available weekly or less frequently. Because consecutive measurements are required to compute our target (daily ΔSWE), only the automatic measurements can be used for training and hyperparameter tuning. However, because we only need SWE (instead of its daily change) for evaluation purposes, we can still use the manual measurements as a test set in the station split.

Regarding the concern with splitting the already small dataset even further into train, validation and test set; this is exactly what motivated our splitting strategy, which optimizes the available data by smart use of cross validation and re-training after hyperparameter selection. This is explained in Figure 2 in the paper, but we have re-written it below so hopefully it becomes clearer:

The data splits for train, validation and test are different for the temporal and station splits. In the temporal split, 3/5 of the data are used for an initial training of the model and 1/5 as validation for model selection and hyperparameter tuning. Finally, these combined 4/5 of the data are used for a final model training, which is tested in the remaining 1/5. Using cross-validation ensures that our results are robust despite the small test size as we effectively test on the whole available data. In the station split, data from 2 out of the 3 training stations are used for the initial training, and the remaining one for validation of model selection and hyperparameter tuning. This is done so that the hyperparameters are optimized for prediction in untrained stations. Again, this is performed three times in a cross-validation loop to ensure robustness. After selecting the hyperparameters, all of the data from the three automatic stations is used to train the final model, which is tested on the manual samples in the remaining stations.

We will improve the description of the data split types in the manuscript to enhance its intelligibility, as it is still slightly confusing despite our best efforts.

Figure 2 – change “a) the station split and b) the temporal split strategies” to “a) the temporal split and b) station split strategies”. Either the a) and b) in the figure is wrong or the caption is wrong.

Yes, the caption is wrong. We will change it accordingly.

Figure 3 – at the moment, a reader (who quickly glances at this plot) might infer that “Sample size auto. stations” of 171, 348, 1355 would mean the number of stations not the number of station measurements used (as I think the authors intend to convey the information). Please change this to be more specific. Also, why would NSE go down for Crocus as more information is used? Is that because model bias becomes more severe as more stations are compared with it?

We agree that the caption it is not 100% clear, we will change it for improved clarity.

Regarding the second point, because Crocus does not use training data, its performance should be independent from the number of evaluation samples in each station. So, the decrease in performance for stations with larger sample pool for the temporal split is likely coincidental. However, it is not a surprising result that Crocus shows a better SWE simulation at Col de Porte than at other sites, as Col de Porte is historically used for the development of this model and an emphasis has been put all along its development to be able to model the SWE there (Brun et al. 1989, 1992). This finding is already documented in publications (e.g. Menard et al., 2021). Note that Crocus simulations also occasionally helped to detect and correct errors in the meteorological forcing at Col de Porte. It also follows that Crocus may show poorer performance for the simulation of SWE in stations whose characteristics deviate from Col de Porte (medium altitude alpine site with mild winter temperatures, quite wet winters and low snow transport by wind).

We will shortly discuss this in the revised manuscript.

Brun, E., P. David, M. Sudul, and G. Brunot. 1992. 'A Numerical Model to Simulate Snow-Cover Stratigraphy for Operational Avalanche Forecasting'. *Journal of Glaciology* 38(128):13–22. doi:10.3189/S0022143000009552.

Brun, E., E. Martin, V. Simon, C. Gendre, and C. Coleou. 1989. 'An Energy and Mass Model of Snow Cover Suitable for Operational Avalanche Forecasting'. *Journal of Glaciology* 35(121):333–42. doi:10.3189/S0022143000009254.

Menard, C. B., and Coauthors, 2021: Scientific and Human Errors in a Snow Model Intercomparison. *Bull. Amer. Meteor. Soc.*, **102**, E61–E79, https://doi.org/10.1175/BAMS-D-19-0329.1.

Line 184-194 – are these results indicating that Crocus degrades ML model performance in the temporal and enhances ML performance across stations (e.g., comparing AUG result between the two data splits)? Why would this be the case? Also, physically, what does it mean when a model does not perform well in the temporal split but does in the station split?

This statement is true not for hybrid models in general, but only for the AUG setup, which uses Crocus simulations on the stations with manual measurements to artificially increase the number of training samples. When using the post-processing setup (PPC), which uses Crocus predictions only as an additional input, the performance is always better than the "purely" ML approach (MSB), although not by a large margin. The degradation in AUG performance in the temporal split with respect to MSB could be caused by the latter being more specialized (or in ML terms, overfitted) on its three training stations, capturing behaviours specific to them, while the increased number of training stations in AUG results in a better ability to generalize to other stations, but at the cost of station-specific characteristics.

All models performed better in the temporal split than in the station one, meaning that the interannual variability of SWE is much easier to predict than its geographical one. However, this is likely due to the characteristics of our dataset, where relatively long time series are available but only few stations, and no predictors of spatial variation (e.g., topography) were used. More generally, a model not performing well in temporal split may be systematically missing some of the specific processes explaining the snow cover dynamics at a specific station (e.g., snow transport or ablation dynamics due to foehn storms), but it may sufficiently capture most of the generally relevant physical processes of snow and its interaction to the environment so that it performs correctly at station split.

We will further expand the implications of these results in the Discussion section.

Line 196-206 – do they authors know why Crocus systematically underrepresents peak SWE (even when run at a point scale) and melts out the snowpack too early? Does it have to do with the rain-snow partitioning scheme in the accumulation season? Could this be enhanced? For example, Jennings et al. (2018) provides a potential path forward. Similarly, what might be driving the snowmelt/snow off date bias? Is there any literature to highlight this as a systematic snow model deficiency?

Jennings, K.S., Winchell, T.S., Livneh, B. et al. Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere. Nat Commun 9, 1148 (2018). https://doi.org/10.1038/s41467-018-03629-7

While an such analysis is out of the scope of this paper, and hence we prefer to leave it out of the manuscript, we hope to add some context in the following response.

The Crocus developers do not have a clear view of the reasons of this underestimation in the ESM-SnowMIP simulations. Actually, the phase partitioning was done by the site-referent researchers (Ménard et al., 2019) and not by the models or the modellers. The methods may be flawed, typically in respect to the findings by Jennings et al., 2018; but this rules out a model-specific bias on this side.

The evaluations of ESM-SnowMIP simulations in Menard et al. 2019, show that this SWE underestimation by Crocus comes with an inhomogeneous bias in albedo (that is typically overestimated at Col de Porte and Swamp Angel; whereas SWE is underestimated at these sites), so that an explanation is hard to find on that side.

Similarly, in these simulations, the SWE underestimation by Crocus comes with a usually underestimated surface temperature (Ménard et al., 2019), that generally should increase the ability of the model to maintain its SWE and is not in line with an anticipated melt.

A behaviour that has been recently highlighted for Crocus (and is not yet published), is that the heat flux from the soil is often erroneous. We observed that at Col de Porte, it can sometimes lead to a complete melt of the first snowfall of the year. This effect could explain part of the systematic SWE negative bias of Crocus, highlighted in Table B1, but does not seem to be involved for the sites WFJ and RME displayed in Fig 4 and 5.

Ménard, C. B., Essery, R., Barr, A., Bartlett, P., Derry, J., Dumont, M., Fierz, C., Kim, H., Kontu, A., Lejeune, Y., Marks, D., Niwano, M., Raleigh, M., Wang, L., and Wever, N.: Meteorological and evaluation datasets for snow modelling at 10 reference sites: description of in situ and bias-corrected reanalysis data, Earth Syst. Sci. Data, 11, 865–880, https://doi.org/10.5194/essd-11-865-2019, 2019

Line 229-235 – do these differences in meteorological variables/etc. have to do with the stations being located in different snow climates, elevations, shaded/forested regions, etc.? Do the authors think they have sampled all of these properly in training/testing the ML models?

The importances of meteorological variables are likely dependent on station characteristics, but the current analysis aimed to capture general patterns valid for all locations. As to whether station characteristics are properly sampled, please refer to the question above regarding Line 71-72. In short, the coverage is reasonably good for most station characteristics given the small sample size, which suggests that our results are at least a good indication of what one could expect when extrapolating to the Northern Hemisphere, but maybe not sufficient to provide strong claims. We will re-write our Results and Discussion to make them more nuanced, stating that these results are a good indication of the variable importances, but more station variety would be needed to provide a definitive answer.

Line 263-264 – do the authors know which stations had more or less sensitivity to lagged meterological variables at +7 day vs 7 day vs 3 day vs 1 day? Do these stations (and their sensitivities) fall into different snow climates, elevation bands, shaded/forested landscapes, etc.? This sort of information would be important to glean to guide future ML model development/application over a larger spatiotemporal set of stations.

A station-specific analysis of the sensitivity to the lagged meteorological variables would certainly be interesting but falls beyond the scope of this paper. Our current implementation concatenates all stations before computing the importances and would suppose a significant effort to add. We would like to encourage other studies to pursue that question.

Section 3.4.2 – this seems like it should be in the Data and Methods section (or Supplemental Material)

Despite that calling it feature engineering, this section showcases an important result, which is that a version of the post-processing hybrid setup (PPC) which includes additional Crocus variables, and so it has more information available about the snowpack, actually performs similar or worse than the same setup with only the Crocus-reported SWE and ΔSWE. Therefore, we believe it is best to keep it in the results section. Yet, we would like to explicit a bit more the design and purposes of the feature engineering in Material and Methods (specifically sect 2.3.1, line 125), give it a specific name (i.e., PPC-expanded), and keep the analysis of these results in the sect 3.4.2, referring to this set-up name.

Line 276-277 – in Figure 3, didn't the authors show that the AUG model (i.e., hybrid Crocus-ML model) resulted in poorer performance for temporal split (i.e., worse NSE range compared to all physics-based and ML models) and slightly better performance in station split (i.e., NSE range is more constrained and the mean NSE is slightly higher than all physics-based and ML models) than Crocus? Is the difference between Crocus and AUG performance statistically significant/appreciably different for the station split?

The performance of AUG in the temporal split is indeed lower than the other ML approaches, but it is still better than Crocus, which it improves in all test metrics analysed in the study: Nash-Sutcliffe Efficiency (NSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Bias (MB).

In the station split, AUG again improves the performance of Crocus (and more so compared to the other ML models) in almost all aspects. It improves the NSE for 6 of the 7 test stations, with an average NSE difference of 0.27 per station, besides improving it also on the entire test dataset from 0.80 to 0.85. Furthermore, it reduces the test RMSE by 12%, and the test MAE by 5%. The only metric in which Crocus achieves better results is the MB, which is almost duplicated (from -34 to -60 mm). So, despite it admittedly being a significantly more biased model, AUG's performance is clearly improved over Crocus.

Line 281-284 – could the tendencies or corrections made by the ML models be used to inform physics-based model development (e.g., to "fix" the snowmelt rate/snow off date bias)? At the very least, could the ML models be used to identify if the variable(s) driving this bias in Crocus (and other physics-based models) are mass or energy related? This could be a major value add from ML models.

This is an interesting idea, and certainly a very desirable direction for future research. However, it does not fully align with the goal of this paper, which is to show the benefits and nuances of using hybrid models. Given our results, it is difficult to assess which variables might be contributing most to biases in Crocus, but we hope more efforts are directed towards that goal in the coming years. We will reflect that by adding this recommendation in the Discussion.

Line 293-295 – what would constitute a "large, representative dataset"? How many days would be needed? How many stations? Etc.

It is hard to define in specific numbers. At the very least, it should cover well the range of the predictors, particularly its edge cases. For example, the limitations of the ML models to correctly predict ΔSWE for high snowfall values (Figure 7 in the paper) indicates that more extreme snowfall events would be needed for improving the ML model training. Similarly, the choice of stations should cover all the different characteristics or locations relevant for snow dynamics (e.g., based on Sturm and Liston, 2021), which is only partially true in our study due to the small sample size. The better performance in the temporal rather than station split indicates that having more locations would be more beneficial than longer time series for our study, but having a good climatic representativity (about 30 years) in the data is also important.

Line 298 – "greater generalization capability" Do you mean Crocus has prognostic, physical equations that can make predictions "out of sample" rather than purely diagnostic/"in sample" inferences (as an ML model arguably does)?

Yes, and we will extend that part in the discussion along these lines.

Line 309 – change "downwards" to "downward"

We will change it accordingly.

Line 313-314 – Do you mean to say something like this "…variable selection should be based on an understanding of the snow climates and geographic heterogeneity (e.g., elevation, forest cover and topographic shading) of the location or region in which the ML model is applied"?

Yes, and we will change the sentence to be more precise following your suggestion.

Line 315-338 – I appreciate that the authors explicitly stated the sample size issue here. I was looking for something like this earlier on though. Maybe a sentence or two in the Methods that references a larger discussion later on in the manuscript?

That would certainly be a great addition, we will include it in the text.

Line 325 – change “specially” to “especially”

We will change it accordingly.

Line 335 – see Song et al. (2024) citation above

It suits the aim of the sentence, so we will add it accordingly.

Line 342 – change “northern hemisphere” to “Northern Hemisphere”

We will change it accordingly.