Review of "**Improving forecasts of snow water equivalent with hybrid machine learning**"

Pomarol Moya et al.

This manuscript presents a hybrid ML approach that combines SWE and meteorological observations with output from a physical snow model to provide enhanced SWE forecasting. I believe these types of hybrid approaches represent an exciting future for snow modeling. The manuscript is well written and the figures are generally clear. However, I have some major concerns with this approach, especially regarding the inclusion of in-situ meteorological data and how this impacts the application of this approach to global SWE forecasting, which is described by the authors as a main motivation for this work.

Dear reviewer, thank you for your kind comments and also for raising some interesting points of discussion about our manuscript. We hope to provide a comprehensive answer to them in the following text.

Major concerns

My main concern is that I struggle with the application of this approach. In the abstract, the authors say "potential to improve forecasts of SWE at unprecedented spatio-temporal scales". However, in the manuscript the ML models are tested only in areas with weather station data and the in-situ meteorological data at times t and t+1 (forecast time) are used as input features for the model. This severely limits the applicability of the model from a forecasting perspective (in-situ meteorological conditions are not available at time t+1) and to specific locations with in-situ meteorological data. Ideally for a large-scale SWE forecast application, such a model would be applied with meteorological output from a model forecast. However, this manuscript does not evaluate how such an approach would perform for this application. In its current form, the methods and models presented in this manuscript are limited, and I don't believe they have much use for forecasting SWE, especially at the global scale as only a handful of sites are utilized in this study.

Our conclusions may have been articulated more ambitiously than the scope of the paper permitted, so we propose to rephrase the relevant text in the introduction and conclusions to avoid overpromising, and to raise this point more explicitly in the discussion.

Nevertheless, we believe that the paper still adds significant value to the community and is therefore worth publishing. Admittedly, modelled forecasts of meteorological data would be required for forecasting applications, and that was not tested in our work. However, we believe that evaluating these hybrid setups extensively with higher quality in-situ data is a valuable first step in achieving that goal. Especially, our work highlighted the value of physical model data as a complement to in-situ observations for training a machine learning algorithm. In particular, using it for data augmentation significantly improved its spatial transferability.

Methodological comments

L72 - Why were only 10 stations utilized? It seems that this approach could benefit greatly from an increase in training data and there are certainly more stations in the NH with timeseries of SWE data that could be used for training. Even if a site has data from only a few years surely this would still be useful, no?

We agree that 10 stations is a small dataset, as elaborated upon in the Discussion section. A more elaborate justification will be provided in the manuscript, in line with what we outline below.

There were multiple reasons for choosing this dataset, most importantly, its quality. It consists of in-situ SWE measurements at high temporal resolution covering a large diversity in geographical locations and station characteristics. It also contains several in-situ meteorological variables which had been used to generate snow forecasts using Crocus. To the best knowledge of the authors, there are no standardized datasets which satisfy these characteristics, and creating our own or significantly expanding it would be an arduous and time-consuming task.

Furthermore, one of the purposes of this paper is to show the performance of hybrid models under data scarcity conditions, since (even globally) only limited daily SWE measuring stations are available. Lastly, this dataset has been previously used for model intercomparison purposes and is well-known in the field, establishing a controlled setting for evaluating our hybrid setups.

L85 – "… according to the geographical location of each station." What does this mean?

Where different aggregation methods used in different locations?

This fragment refers to the calculation of the daytime average, for which the daytime hours are calculated for every day of the year according to the geographical location of the station. We will rephrase the sentence for improved clarity.

Table 1 – Why do you use both SWdown_avg and SWdown_day? These features will be nearly perfectly correlated and I doubt both are necessary.

The explanation and justification of the predictors will be further outlined in the manuscript. Regarding the variables derived from the shortwave radiation, while both are certainly correlated they only obtain an R^2 value of 0.67. This is because the first one reports the average shortwave radiation over the 24h, while the second the average over the hours that fall between dusk and dawn, which depends on the day of the year and latitude. The 24h average is more sensitive to seasonality, while the daytime average more directly encapsules the atmospheric conditions, so both were deemed potentially useful. Lastly, the daytime average is not very important according to the SHAP analysis, so it is unlikely that it has a strong negative effect on the performances of the machine learning models.

Table 2 – What is RAM_SONDE_avg and why is it a useful feature for the ML?

The explanation and justification of the predictors will be further outlined in the manuscript. Regarding the ram sonde variable; as described in the table, this Crocus state variable accounts for the "average of the penetration of ram resistance sensor", which is a cone-tipped metal rod designed to be driven downward into deposited snow or firn (American Meteorological Society – glossary of Meteorology, https://glossary.ametsoc.org/wiki/Ram_penetrometer, last access 11 July 2025). The penetration distance of the rod into the snow of firn for a given amount of force is an indication of one important physical (mechanical) property of the snowpack, namely its hardness, much related to the snow density and microstructure. Both properties have important implications for heat transfer within the snowpack (snow thermal conductivity is typically much related to density, e.g. Calonne et al, 2011) and to a certain extent, for snow melt. Therefore, this variable is a good candidate to consider in relation to SWE prediction and snowmelt behaviour.

Calonne, N., Flin, F., Morin, S., Lesaffre, B., du Roscoat, S. R., & Geindreau, C. (2011). Numerical and experimental investigations of the effective thermal conductivity of snow. Geophysical Research Letters, 38(23).

Figure 1 – The formatting for this diagram is a bit confusing. Why are [and) brackets used? I also think that $\Delta SWE_{(t)}$ is confusing. I see that it is defined in the caption, but it is not immediately intuitive as really the target variable is the change in SWE at time t+1. Also Measured is shortened to Mea. In b) but not a) or c).

Admittedly, that figure lacks some explanation regarding the use of brackets and parenthesis, which will be added to the manuscript. These refer to the aggregation method; each daily value is computed from the hour corresponding to the prior SWE measurement (t) up to, but not including, the same hour next day (t+1). We will also incorporate the other proposed improvements for the final version.

L125 – "Additional Crocus-based predictors, such as the ones described in Table 2, may also be added…" What is meant by this? More details are necessary on this.

This sentence is indeed unclear and would benefit from re-writing. The meaning is that besides including only the model-simulated SWE as an additional predictor, one could also add other Crocus-generated state variables, such as those described in table 2. This directly relates to the contents of section 3.4.2, where we compare the results with and without these additional variables.

L140 – "Three different ML algorithms were compared:" Why were these three chosen? How was the LSTM model set-up? It's not surprising that the LSTM does not perform optimally as these are typically better with longer time series of data. Perhaps a GRU model would be preferrable? For the NN and the LSRM, how were the hyperparameters tuned? RFs typically can perform better 'out of the box'. In contrast NNs typically require much more substantial hyperparameter tuning. From Table A1 it's not surprising that the NN and LSTM did not perform as well as it seems that not very many hyperparameters were tested.

While we fully agree that testing other ML algorithms such as GRU would be a great addition, the aim of the paper was not to provide a thorough comparison of different ML algorithms as the focus is on comparing different hybrid modelling setups. The three proposed algorithms are amongst the most popular; RF and LSTM have been used for hybrid SWE prediction in the literature (e.g., King et al., 2020; Steele et al., 2024) while a feedforward NN offered an intermediate step in terms of complexity. We considered that a sufficient subset of the available options.

The implementation of the LSTM model will be further expanded in the manuscript. The implementation was done by taking the lag time window (14 days) of meteorological variables as the sequence length where the LSTM units unfold. After, a dense layer takes the outputs of the LSTM layer and any additional variables at the last time step to produce the predicted ΔSWE from the current step until the next one. When applied for inference sequentially, the same procedure was followed after shifting the time window one day forward and updating the current SWE (which is also a predictor) to the last predicted value.

Finally, we agree that more tuning would likely improve the performance of NN and LSTM, but would also require much higher run times. For this paper we decided to use a fixed budget for tuning, finding a model that strikes a balance between accuracy and usability. The goal was not to claim what algorithm works best, but rather to find a good performing one to test the application of hybrid models. We believe this needs to be more explicitly mentioned in the Discussion section and we will do so when revising the manuscript.

King, F., Erler, A. R., Frey, S. K., & Fletcher, C. G. (2020). Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. *Hydrology and Earth System Sciences*, *24*(10), 4887–4902. https://doi.org/10.5194/hess-24-4887-2020

Steele, H., Small, E. E., & Raleigh, M. S. (2024). Demonstrating a Hybrid Machine Learning Approach for Snow Characteristic Estimation Throughout the Western United States. *Water Resources Research*, *60*(6), e2023WR035805. https://doi.org/10.1029/2023WR035805

L168 – "Nash-Sutcliffe efficiency" I'm not immediately familiar with this metric. Maybe explain briefly?

The NSE is calculated as one minus the ratio of the error variance of the modelled time-series divided by the variance of the observed time-series. It is a commonly known metric in hydrology, so we did not explicitly define it, but we could add it to accommodate for researchers from other domains.

Feature importance

To me, it is not expected that downwards shortwave radiation would be the most important feature. You are modeling both the accumulation and ablation season correct? I would expect SW radiation to be very important but only during the ablation season. I'm curious if the feature importances change temporally? This might be interesting insight to include. My guess is that SW radiation has high magnitude SHAP values during the ablation season because ΔSWE is generally much higher during the ablation than the accumulation season. I'm curious what you would see if you compute relative SHAP values (by normalizing by ΔSWE). I would expect other features (precipitation) to be relatively more important.

Shortwave radiation is indeed most impactful during the ablation period, but it does have some impact on the ML model predictions for the remainder of the year as well.

To test this, we calculated the mean absolute SHAP values for the accumulation and ablation time steps separately (Figure 1), as defined by the sign of the corresponding ΔSWE prediction. The shortwave radiation is not only the most important feature (on average) during ablation, but also the second most important feature for the accumulation time steps, only below the snowfall rate.

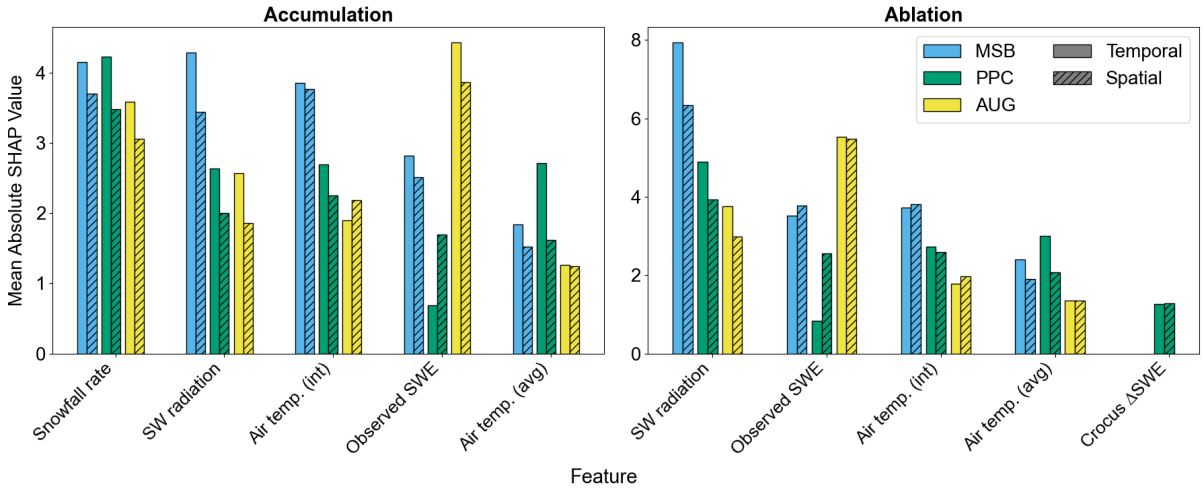

Figure 1: Feature importances of the five top ranking variables, calculated as the mean absolute SHAP value aggregated for all lagged variables, for each ML-based setup and split type. They are ordered according to their average importance for the three ML-based setups. The left subplot shows the results for the accumulation period, that is, for time steps where $\Delta SWE > 0$, and the right

subplot for the accumulation, containing the remaining ones. SW radiation refers to the downward shortwave radiation, 'avg' to the daily average, and int to the daily time integral of positive values.

We have also computed the relative mean absolute SHAP values (Table 1), and despite small changes in the feature importance order, the shortwave radiation again features among the most important variables for both split types.

Table 1: Mean relative absolute SHAP values for the top five values according to their average values across the different setups, for both types of split. MSB, PPC and AUG refer to the measurement-based, post-processing and data augmentation setups described in the paper, respectively, and the column mean refers to the average of the three. Regarding the rows, SW radiation refers to the downward shortwave radiation, temp to the temperature, 'avg' to the daily average, and int to the daily time integral of positive values.

Temporal split:

	MSB	PPC	AUG	mean
Observed SWE	24.59	2.99	66.69	31.42
SW radiation	33.23	13.08	44.53	30.28
Air temp. (int)	20.35	12.23	13.57	15.38
Air temp. (avg)	10.85	12.47	14.86	12.73
Snowfall rate	8.57	4.30	15.56	9.48

Station split:

	MSB	PPC	AUG	mean
Observed SWE	7.86	9.60	28.30	15.25
Air temp. (int)	10.31	13.13	14.01	12.49
SW radiation	12.11	11.00	14.29	12.47
Air temp. (avg)	4.40	9.17	8.56	7.38
Snowfall rate	4.00	6.59	10.04	6.88

A plausible hypothesis is that its importance could be overestimated compared to physical models since it is a good indicator of the seasonality, which the ML model may be using to guide its predictions. Another hypothesis is that at some low-altitude sites like the Col de Porte, that are frequently close to the rain-snow transition in terms of winter temperatures, snowfall and melt may happen in the same day as a result of rapid variations in weather conditions. At such sites, incoming shortwave radiation can hence also modulate the accumulation of SWE (by reducing it at daily scale when there is melt just after) and be therefore a relevant predictor in the accumulation phase.

We will include the above figure and table in the revised manuscript providing a short explanation along the lines of this rebuttal.

L229 – "reaching above a SHAP value unit." What is meant by this?

What was meant is that those variables achieve a mean absolute SHAP value higher than 1. We will re-write that sentence for improved clarity.

Figure 6 – Why did you choose to plot the mean absolute SHAP values? For some features, it may also be interesting to see how the feature impacts the predictions (i.e., increasing or decreasing predicted SWE).

The purpose of this figure was to show the most important variables for SWE prediction and their distinction per hybrid setup and split type, without delving into the more complex relationships that would make the figure less readable. It could even be further compacted by combining the two subplots of that figure into one for easier comparison between spatial and temporal splits, similar to the previous figure on accumulation and ablation, or even replaced by that figure.

For more information regarding the correlation between each predictor and the target, we computed the SHAP violin plots, which show how the values of each variable influence the target. When the predictor goes from blue to red (left to right), it indicates a positive correlation, and from red to blue a negative one. This is most clear for the air temperature, which is red for negative SHAP values and blue for positive ones. Snowfall rate contains very strong positive SHAP values when it is high (meaning it produces a large positive effect to ΔSWE), while its lower values have little influence, as we might expect. These plots will be added to the appendices along with a short discussion of the influence of each variable. This could be even further enriched with scatterplots of specific variables against their SHAP values, as in Figure 7 from the paper.

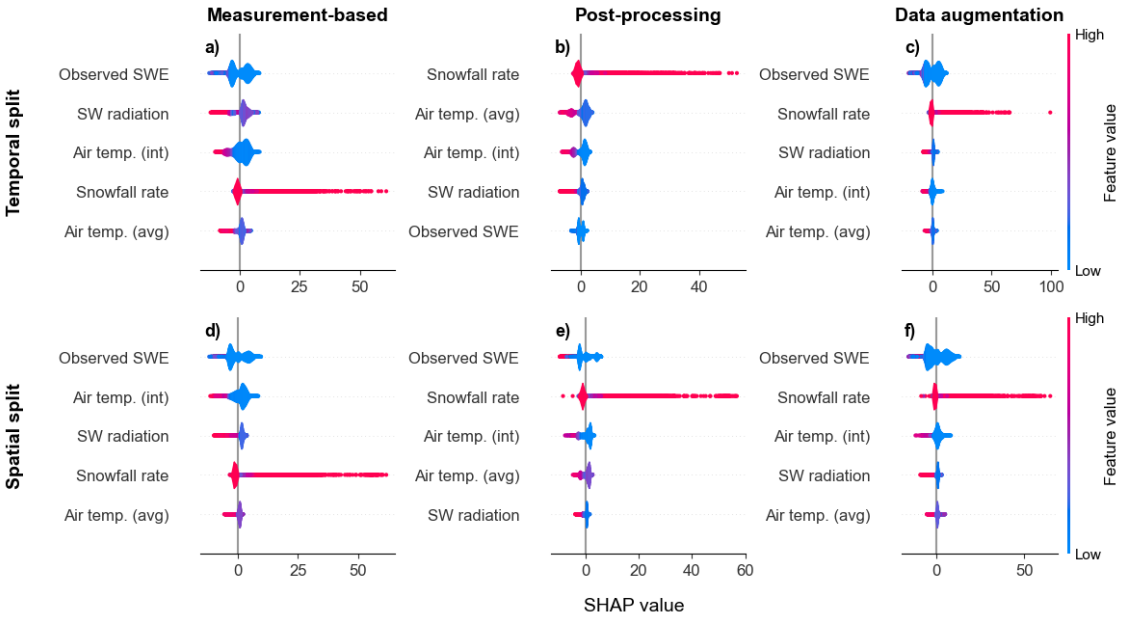

Figure 2: Violin plots of the SHAP values for the five highest ranking variables in terms of mean absolute value for each type of split and setup. The colour represents the value of the feature from high to low compared to their average. The sign and magnitude of the SHAP values indicate whether the variables have a positive or negative impact on ΔSWE and how strongly that impact is.

Discussion

L281/2 – "The differences in performance concentrate towards the end of the snow period, where the ML-based models particularly improve the timing of the snow melt." Does this indicate that there are substantial errors in the melt dynamics in the physical model?

Our results suggest that there are indeed non-negligible errors in the melt dynamics of Crocus, which the ML models seem to improve. An in-depth analysis of the main causes for that would be highly interesting, although out of scope for our paper. We will add that comment to the Discussion.

Technical comments

Mind consistency with 'an ML model' vs. 'a ML model' (for example in the abstract both are used). I personally don't know which is correct but try and be consistent with your usage!

Table 2 – Type "soild" in row 3

L249/250 – Different tenses are used in the same sentence here ('reduced', 'achieves').

L260 – Maybe ‘slowest’ instead of ‘softest’ here?

L300 – I’m not sure that ‘impoverished’ is the best word choice here. Maybe just ‘poor’ is better.

Thank you for your suggestions, we will incorporate them into the manuscript.