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.

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.

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?

L85 – “… according to the geographical location of each station.” What does this mean? Where different aggregation methods used in different locations?

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.

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

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).

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.

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.

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

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.

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

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).

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?

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.