Review of "**Improving forecasts of snow water equivalent with hybrid machine learning**"
Pomarol Moya et al. 2025

General and Major comments

The manuscript is improved from its previous iteration; however I still have several concerns. I especially appreciate the extra effort and detail put into the feature importance analysis. In the first round of revisions, between myself and the other reviewer, the two most major concerns were (1) the data scarcity issue, and (2) the method's applicability for forecasting. While these issues were discussed in the manuscript, neither were truly addressed. I believe that actually addressing either one of these major issues (either by implementing more training data or by testing the application with model forecast data instead of in-situ meteorological data) would greatly improve the manuscript. As such, I am a bit disappointed by the changes made from the previous iteration and questions what this manuscript adds, especially given the similarities with Steele et al (2024). At a minimum, the framing of the manuscript should be changed from SWE forecasting to maybe just comparing hybrid SWE model setups. Testing these setups with model forecast data to demonstrate their applicability for SWE forecasting would greatly strengthen the manuscript.

Dear reviewer, thank you for your review feedback. We hope to tackle your major concerns in two ways:

1) Make our goal clearer in the manuscript and reduce the use of the "forecasting" keyword.

   The main objective was to evaluate if hybrid models can enhance SWE predictions compared to process-based (Crocus) and pure ML (MSB) models. So, it is important for us to use *in-situ* data to ensure that our results reflect the errors in SWE modelling rather than those coming from inaccuracies in the meteorological data. If we would train our models using a modelled meteorological product, the ML-based models may learn to correct biases specific to that product, making our conclusions less generalizable.

   Because that might not have been clear enough in the manuscript, we have changed the title to "Improving snow water equivalent modelling: a comparative study of hybrid machine learning techniques". Furthermore, we have modified multiple sections of the main text clarifying our objective and specifically avoiding the "forecasting" keyword except where it truly makes sense.

2) Provide an analysis of the performance of the models when using modelled meteorological data instead of *in-situ* data (refer to the relevant section below).

   Unfortunately, it is not feasible to re-run Crocus given the short time frame, so we cannot test the performance of the post-processing (PPC) setup, which requires Crocus as an input. Therefore, the results below refer only to the data augmentation (AUG) and MSB setups.

   Because of the incomplete nature of this analysis and given that these results are out of the scope of our study (which focuses in the modelling improvements rather than forecasting), we propose not to include it in the manuscript. Instead, we have stressed even further in the discussion that the performance of the models is expected to decrease in practical applications using modelled meteorological data. If deemed important, we could add this analysis as a supplement and discuss it briefly in the main text.

Finally, we wanted to stress the main differences between Steele et al. (2024) and our paper. While it is true that their methodology is very similar to our PPC setup, we also introduce a new approach, AUG, particularly suited for data-scarce scenarios. Moreover, we test more aspects of hybrid model creation, e.g., different ML algorithms, feature importances and impact of adding lagged features. Finally, the dataset we used covers a much larger geographical extent than the one used in Steele et al. (2024) and consists of *in-situ* meteorological variables instead of modelled, which is important for model intercomparison as we argued above. In conclusion, we believe that our contribution is sufficiently distinct from that of Steele et al. (2024) to justify the publication of this study.

Analysis of model performance with modelled meteorological data

As requested, we performed an analysis to study the impact of predicting SWE with ML and hybrid models using modelled meteorological variables as inputs instead of *in-situ* data.

As modelled meteorological dataset, we used the bias-corrected reanalysis forcing data provided in the ESM-SnowMIP dataset (Menard & Essery, 2019), corresponding to GSWP3 (Global Soil Water Project Phase 3) run at 0.5° spatial resolution for LS3MIP (Kim, 2017). This data was used to simulate SWE using the MSB and AUG models trained with *in-situ* data, as described in the manuscript.

First, we discuss the differences between GSWP3 and the *in-situ* observations. Figure 1 (below) shows the relations between GSWP3 and *in-situ* data. Air temperature shows a good correlation, but the errors increase for the downward shortwave radiation. This is even more extreme for the snowfall, where the correlation suffers heavily. In particular, there are many instances where one source indicates there is snowfall while the other does not. Since the timing of the shortwave radiation and snowfall are critical for accurate simulation of SWE, this is already an indication that the performance will be worse.

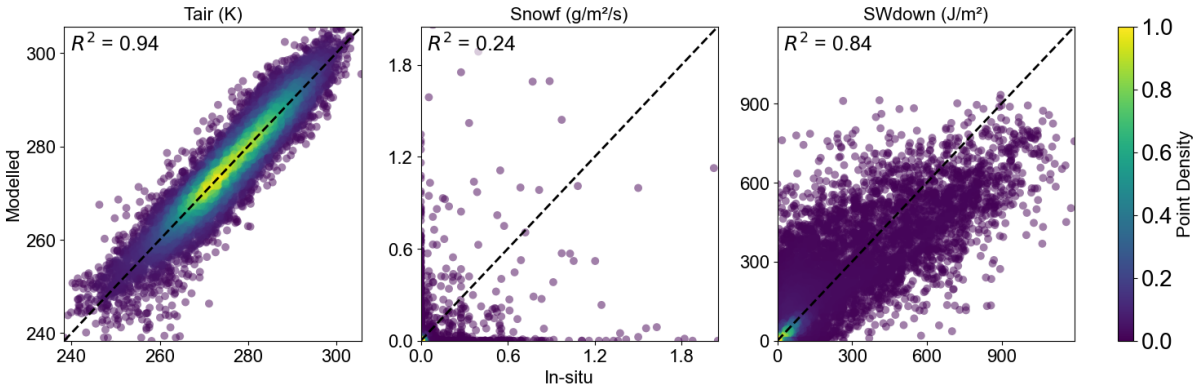

Figure 1: Scatter plot of daily GSWP3 values against *in-situ* data of the most important meteorological variables for the ML-based models, namely air temperature (left), snowfall (centre) and downward shortwave radiation (right). A selection of 10000 out of 938631 points have been randomly selected for plotting purposes.

The results of simulating SWE using the GSWP3 forcing data (Table 1) show a considerable decrease in model performance compared to the same simulations using *in-situ* data. Particularly, for the station split AUG almost doubles its error, which reaches a similar value to that of MSB.

	Temporal split			Station split		
	MSB	AUG	Difference	MSB	AUG	Difference
In-situ	38.9	48.3	24.3%	159.0	109.3	-31.3%
Modelled	72.9	84.8	16.3%	201.5	211.2	4.8%
Difference	87.7%	75.5%		26.8%	93.2%	

Table 1: Root mean squared error for the data augmentation (AUG) and measurement-based (MSB) models when using modelled compared to *in-situ* meteorological forcing data.

In Figure 2 (below) we can notice different patterns of accumulation for some years (e.g., 1991-1992), which can be explained by the low correlation of the snowfall in the modelled meteorological data. However, it is the acceleration of the snowpack ablation that results in an early snow disappearance (e.g., 1994-1995) that has the largest impact in model performance, especially for AUG. Nonetheless, the seasonal pattern is still well captured by both approaches.

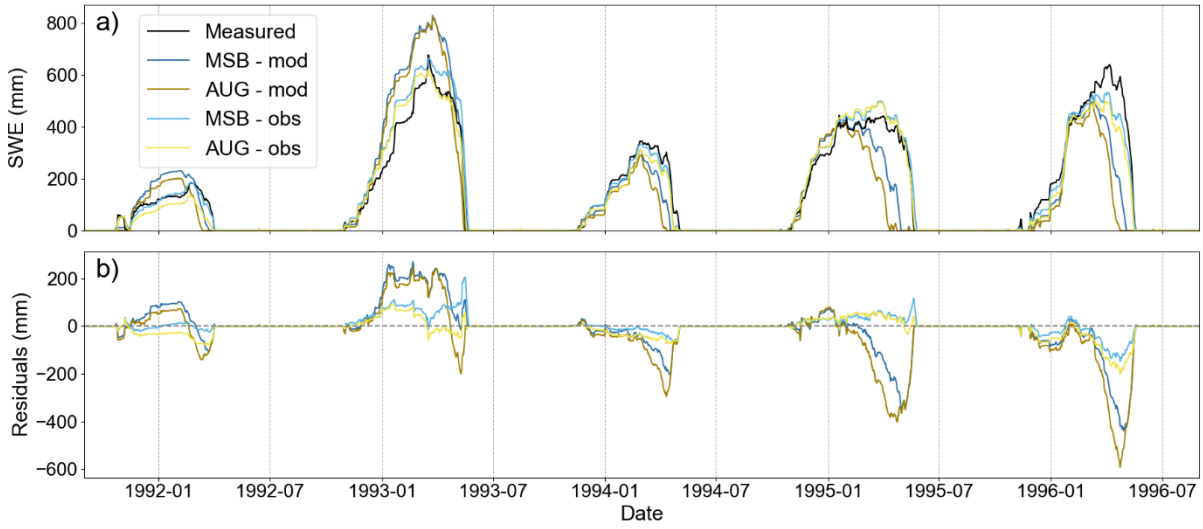

Figure 2: Time series of an example five-year range in the Reynold Mountain East station including MSB and AUG using *in-situ* (obs) or GSWP3 (mod) data for the temporal split. The station and time range correspond to Figure 4 in the original manuscript.

Overall, we can observe that the change in performance when using GSWP3 is of much larger scale than the differences between modelling approaches, which is to be expected given the poor fit of GSWP3 with the *in-situ* data. This highlights the importance of having high quality meteorological data, which can have a larger effect on model performance compared to improvements related to snow modelling. Nevertheless, it is important to note that in forecasting applications, where the ML models would be trained with the same data source used for inference, the errors would likely be much lower (e.g., Steele et al., 2024).

Kim, H. (2017). *Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1)* (Version 1) [Dataset]. Data Integration and Analysis System (DIAS). https://doi.org/10.20783/DIAS.501

Menard, C., & Essery, R. (2019). ESM-SnowMIP meteorological and evaluation datasets at ten reference sites (in situ and bias corrected reanalysis data) [Dataset]. In *Supplement to: Menard, Cecile; Essery, Richard; Barr, Alan; Bartlett, Paul; Derry, Jeff; Dumont, Marie; Fierz, Charles; Kim, Hyungjun; Kontu, Anna; Lejeune, Yves; Marks, Danny; Niwano, Masashi;*

Raleigh, Mark; Wang, Libo; Wever, Nander (2019): Meteorological and evaluation datasets for snow modelling at 10 reference sites: Description of in situ and bias-corrected reanalysis data. Earth System Science Data, 11(2), 865-880, https://doi.org/10.5194/essd-11-865-2019. PANGAEA. https://doi.org/10.1594/PANGAEA.897575

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

Minor comments (line numbers from tracked changes version)

L30 – "water resources" -> "water resource"

L31 – "its" -> SWE

L74 – "2) targeting SWE prediction at ungauged stations" – How is this done? I guess this means the model was tested on sites not used in the training? But these stations used for evaluation were still gauged so I find this a bit misleading.

When you use the word "significantly" does this mean "statistically significant"? Did you do tests of significance in these cases? If not, avoid using this term.

L379 – remove "tremendous"

L398 – "its measured equivalent" – what is meant by this here?

L401 – What does "they" refer to in this sentence?

Thank you for your suggestions, we have incorporated them into the manuscript.