

Authors Response: EGUSPHERE-2025-6066

Dear Dr. Vuyovich and manuscript referees,

We'd like to thank the handling editor and both referees for taking the time to review and provide thoughtful, constructive comments for improvement of the study. We have agreed with comments and recommendations, making several additions and revisions to the manuscript. This has strengthened specific points of the study including the use of remote sensing snow products for model verification and the consideration of precipitation phase uncertainty in creating ensembles for snow modelling.

Below is a point-by-point response to the reviews including a list of all relevant changes made in the manuscript. Please see below our responses to [in blue](#).

Sincerely,

Dave Casson

On behalf of the co-authors

RC1: '[Comment on egusphere-2025-6066](#)', Anonymous Referee #1, 09 Mar 2026

Summary

Casson et al. develop and demonstrate a framework for including uncertainty in meteorological forcing data through probabilistic snow model simulations with SUMMA in three contrasting study basins (Chena, Alaska; Bow, Alberta; Tuolumne, California). The framework first applies undercatch corrections to station precipitation data using different WMO equations for the study basins. Next, two approaches are developed (locally weighted regressions, LWR; random forests, RF) in terms of static and dynamic predictors, where it is found that RF generally improves KGE scores relative to LWR and dynamic predictors provide a modest enhancement. The SUMMA model is calibrated using a large sample emulator (LSE), and then applied with 50 RF-generated ensembles and two benchmark forcing datasets (ERA5 and CaSR). The paper reports that the probabilistic ensemble from the RF approach generates realistic SWE time series and improved metrics relative to the two deterministic approaches. The authors conclude that the framework provides a scalable approach to improve how forcing uncertainty is represented in distributed SWE simulations.

Recommendation

In my opinion, this is a potentially impactful and useful study that is within the journal's scope. This paper provides multiple novel contributions and innovations, including the LSE for parameter identification, the demonstration of RF as an effective non-linear approach for determining meteorological fields, and the probabilistic ensemble-based application across GRUs in multiple snow climates. The methods are generally thoroughly described, which justifies the use of multiple appendices. I appreciate the authors' embrace of FAIR principles and open-source frameworks, which should increase the potential usage by others in the community. Below, I offer several recommendations to improve the manuscript prior to publication.

[Thank you for the positive comment and recommendations. The recommendations are addressed in the revised manuscript and summarized in responses below.](#)

Main Comments

1. While the paper's organization is generally logical and effective, I think that there may be a need to improve the organization of Section 4 (methods). It seems that there is an opportunity to align the subsections here based on the methodology chart from Figure 3. In the current version, these are somewhat similar, but not exactly in alignment. This makes it a little hard to map the different components of the figure to the text in section 4. Additionally, it would help to refer back to Figure 3 as the seven components are described in the section 4 subsections.

This is a useful comment to improve the clarity of the Section 4 (Methods).

We have updated Figure 3 and have renamed and re-organized Section 4 headers for consistency. References back to the Figure 3 are also added in the methods description.

2. A limitation in the SWE evaluations is the spatial scale mismatch between observations and GRU estimates, which is noted by the authors (e.g., L. 558-559, 634-640, Fig. 14). It seems that there is a missed opportunity to conduct some spatial evaluation with remotely sensed snow data, such as MODIS-based snow cover metrics (e.g., Crumley et al., 2020) or lidar-based SWE. The authors indicate they do not pursue these given inconsistent availability across their study sites (e.g., L. 206-207). I would argue that this type of evaluation could be instructive, even if it is only done over a subset of the study basins/years.

We agree that the scale mismatch is a key limitation in evaluation, and the inclusion of evaluation with spatially distributed products could be instructive.

Addressing this comment has lead to several key additions to the manuscript, including evaluation with MODIS based snow cover metrics presented in Crumley et al, 2020 and in the inclusion of ASO Lidar data for the Tuolumne (Painter et al, 2016).

Please refer to the follow new sections in Data, Methods and Results and in the Appendices

- 3.5 Remotely Sensed and Lidar Snow Data
- 4.6.2 Evaluation of Remotely Sensed Data
- 5.5.2 Evaluation with Remote Sensing Snow Metrics
- Figure 14. Snow Disappearance Date (SDD) and Snow Cover Fraction (SCF) for snow model results compared to MODIS MOD10A1 derived measurement
- Addition of Lidar Data to Figure 12 and Figure E3

- Appendix D: Spatially Distributed Snow Measurement and Precipitation Phase Results

There have also been updates to the Discussion Section 6.3 Potential Application and Limitations, regarding this comment and that of the second referee.

3. The paper states that the RF ensemble improves the timing and magnitude of melt (e.g., L. 559-560, 600), but I am not convinced. First, there is a potential issue with some ensemble members where the snowpack does not completely melt over summer in the Tuolumne basin following large snow years like 2011 and 2017 (see bottom three panels of Fig. 12). The observational SWE data are unable to evaluate this aspect, and this points to another case where some type of spatial snow data (e.g., MODIS snow disappearance) could be helpful. Second, it would be more convincing if metrics were directly evaluated that isolated the melt timing and magnitude. Currently the paper focuses on SWE time series, which combines the effects of accumulation and melt, which makes it harder to understand how well these two processes are represented.

We agree that more quantitative evidence was needed for this specific conclusion, and that the observational SWE data are often unable to constrain and evaluate the melt out timing. The metrics proposed and evaluated (such as Snow Disappearance Date) are informative as to the model capability. Please refer to the sections addressed in the comment above for key additions to the manuscript.

We have updated Section 5.5.3 Evaluation with In situ Measurements, which now reads:

The impact of ensemble meteorological forcing on snow simulations was assessed by driving SUMMA with the 50-member RF ensembles and two deterministic benchmark forcings (CaSR and ERA5). Simulations were evaluated for GRUs containing NorSWE manual snow surveys (Fig.~\ref{fig:gpep_snow_station_and_tdx}), with modeled SWE compared against observed transect-based snow surveys.

Figure~\ref{fig:swe_timeseries} illustrates SWE time series for the three GRUs with the greatest number of snow surveys in each basin.

The RF-driven ensembles generally produce larger peak SWE than the deterministic forcings, with observed SWE typically falling within the ensemble spread during the accumulation season. In the Bow, the ensemble mean peak SWE is comparable to the CaSR- and ERA5-driven simulations, whereas in the Chena and Tuolumne the ensembles better capture the range of low- and high-accumulation years, including years when the deterministic forcings substantially under- or overestimate peak SWE.

Quantitative evaluation of SWE bias and RMSE (Fig.~\ref{fig:skill_and_taylor_plots}) also illustrate that the RF ensembles provide comparable or improved representation of peak SWE magnitude relative to the deterministic forcings.

Agreement at individual survey locations is more variable, and in some Bow GRUs the deterministic simulations more closely match the observed SWE magnitude than the RF ensemble mean. These differences reflect the challenge of evaluating grid-scale simulations using observations that may not fully represent the spatial variability of snow within a GRU. Consequently, the in situ SWE observations primarily constrain snow accumulation and peak SWE, whereas snow disappearance date and snow cover duration are evaluated independently using MODIS-derived products (Figure~\ref{fig:SCF_and_SDD}). Complete SWE time series, including snow pillow evaluations, are provided in Appendix~\ref{app:swe_timeseries}.

The following text on limitations has also been added to Section 6.3 Potential Application and Limitations

A key limitation of this framework is the representation of snow processes at the GRU scale, which aggregates substantial sub-grid heterogeneity.

Figure~\ref{fig:mountain_observatory} illustrates this limitation, showing that SWE measurements collected under contrasting local conditions within a single GRU can differ markedly. Preferential accumulation areas, including wind-sheltered slopes, leeward deposition zones, and forest openings, often sustain deeper and more persistent snowpacks than surrounding terrain, yet are not explicitly resolved. These locations are also underrepresented in meteorological and snow observation networks, compounding uncertainty in both forcing and model evaluation particularly in melt periods.

Specific Comments

- L. 38: Consider also citing Günther et al. (2019) here, as they compared multiple sources of uncertainty for a physical snow model.
 - Agreed. The updated text : *These biases propagate through snowpack simulations, making meteorological forcing a dominant source of model uncertainty (Raleigh et al., 2015; Günther et al., 2019).*
- L. 49-51: You might also note here that the airborne and terrestrial lidar provide high spatial resolution.
 - Agreed. The updated text: *Airborne and terrestrial lidar, along with satellite remote sensing, have expanded the ability to map high resolution snow depth,*

cover, and albedo across larger domains (Deems et al., 2013; Painter et al., 2016; Gascoin et al., 2024).

- L. 59: Suggest adding “lumped” to “conceptual degree-day models”.
 - Agreed. The updated text: *Approaches range from lumped conceptual degree-day models to physically based, high-resolution systems that solve processes such as resolving energy balance, blowing snow, canopy interception, and avalanching (Hrachowitz and Clark, 2017; Magnusson et al., 2015)*
- L. 207: Add “across study basins” at the end of this sentence.
 - This sentence describing the limits of study to in-situ snow products has been removed, as we now include remote sensing and lidar products.
- L. 209: Should read “Sun et al. (2019)”.
 - Agreed and updated.
- L 212-214: Is this the same as ERA5-Land? Please clarify.
 - This is not the exact same product as ERA5-Land. Updated text: *The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) is a state-of-the-art global reanalysis dataset that provides consistent and high-quality estimates of atmospheric, land, and oceanic variables from 1950 to the present (Hersbach et al., 2020). In this study, we use ERA5 atmospheric (meteorological) forcing variables from the full reanalysis product, rather than the higher-resolution ERA5-Land dataset.*
- L. 246-251: I think this text should state more directly the types of gauges and corrections used for the study basins. This is implicit in Figure 4, but seems like it should be referenced more explicitly.
 - Agreed. The text has been updated to: *A regional approach that assumes consistent gauge type and configuration within each basin is taken.. Specifically in the Bow Basin WMO SPICE UTF (Kochendorfer et al., 2017a, b) was applied for assumed unshielded Geonor and Pluvio gauges, and in the Chena and Tuolumne the WMO (1998) standard rain gauge correction (Goodison et al., 1998; Yang et al., 1998) is applied for assumed unshielded US standard 8” gauge.*
- L. 334: This seems like a logical place to reference Appendix C.
 - Agreed and updated
- L. 379: It would help to have a brief summary of the key model decisions to understand how SUMMA was setup here.

- Agreed. Updated text: *With respect to snow modelling, the model decisions select the dynamic snowpack layering scheme adapted from the Community Land Model (Lawrence et al, 2011). Model decisions also inform snowpack boundary conditions, such as the throughfall of snow based on new snow density, itself formulated as an empirical function of temperature Hedstrom and Pomeroy (1988). Snow density is also used to calculate thermal conductivity (Jordan, 1991) The albedo decay rate is held constant in time (Verseghy, 1991). For greater detail on the snow model implementation and model decisions, the reader is referred to Clark et al, 2015a)*
- L. 380-385: What is a single GRU and how is this different than an HRU? I read this part multiple times and the differences remain unclear to me.
 - To avoid confusion, use of HRUs is removed from the text. It was only described in this one paragraph, and not relevant to the study approach.
 - In this model framework distinction is perhaps best articulated in the original Clark 2015: *A GRU is composed of one or more HRUs. The key characteristics of the HRUs are: (i) similar to the GRUs, the HRUs can be of any shape and size, but there is no longer the restriction that HRUs are spatially contiguous (e.g., an HRU can lump together hydrologically similar areas from different parts of the landscape); (ii) the meteorological forcing data can vary across the HRUs, as opposed to the approach in Kouwen et al. [1993] where all HRUs within a given GRU receive the same meteorological input; and (iii) we include the option for lateral subsurface flow among HRUs*
- L. 432: Need to define the N_{θ} and N_{atts} variables.
 - We thank the reviewer for catching this. Updated text: *To this end, we use a matrix of $N_{sites} \times N_{i=0}$ rows and $N_{\theta} + N_{atts}$ columns as x , where N_{θ} is the number of parameters to calibrate, and N_{atts} the number of site attributes.*
- L. 423-442: I find the large sample emulator to be an innovative and appealing approach but I wonder about how accurate it is in predicting SUMMA skill in this multi-dimensional parameter space. Have the authors checked and verified that the predicted skill (from the emulator) matches evaluations with actual SUMMA runs with the final/optimal parameter sets?

A detailed response and figure have added to the manuscript in response to this comment in a new Appendix D. Please refer to Appendix D in the manuscript for the complete explanation and figure.

Updated Text in Section 4.52 Snow Model Calibration: *To train the emulator, we consider snow gauges across Canada and the US, with four water years of measurements, each with at least 20 measurements per year taken from NorSWE dataset and quality-controlled SNOTEL stations (770 stations in 465 total; the location of the stations and the emulator error are shown in Figure D1 in Appendix D; Sun et al. 2019; Mortimer and Vionnet 2025).*

- L. 443-452: I think some clarification is needed here in terms of the model calibration. Specifically, were the parameters derived by SUMMA simulations driven just by station observations? If not, how where did the additional meteorological data come from?

The emulator is trained using SUMMA parameter sets, attributes, and performance metrics from different stations ($n = 162$) and uses an iterative approach to further refine regions of the parameter space with higher performance metric values. The performance metrics are computed for each parameter set by contrasting SUMMA simulations and observations. The forcing data used to run SUMMA simulations in training is ERA5 meteorological data interpolated at each site.

To clarify that each station used to train the emulator uses ERA5-derived forcing data, we have modified the text in section 4.5.3: *"Note that for each of these stations, we obtain the required forcing data to run SUMMA from the ERA5 reanalysis, as well as the SWE measurements to compute performance metrics."*

- L. 480: Replace "This" with "These".
 - Agreed and updated
- L. 503-504 and Fig. 7: It might help to add markers (asterisks) above the predictors in Fig. 7 to denote which were retained.
 - Agreed. This update has been made across all predictor screening figures, and the caption updated *"Asterisk indicates the predictor retained in sequential forward screening."*
- L. 510-518: Given the limited number of study basins, variables, and models, I wonder about how much can really be generalized here. You might rephrase the language to not use words like "tends".
 - Agree with this comment. Updated text in Section 5.3 Predictor Combination Selection and Regression Performance:
 - *Across the study basins, elevation or its smoothed derivatives are the primary controls on temperature estimation, with only a small number of additional*

predictors contributing. For LWR, elevation alone explains most of the variability, while adding a single dynamic predictor (such as reanalysis air temperature or shortwave radiation) provides modest but consistent improvements. RF models for temperature use combinations of elevation and one or two additional topographic descriptors (e.g., aspect or south–north derivative), with further gains when reanalysis air temperature or humidity is included. For precipitation, both static and dynamic predictors contribute. In the Bow basin, LWR identifies aspect and latitude as key predictors, with air temperature as the main dynamic input, while RF additionally includes longitude and surface pressure. In the Chena and Tuolumne basins, RF uses elevation and horizontal derivatives together with reanalysis precipitation rate and surface pressure, indicating that interactions between orography and large-scale meteorology play an important role in daily precipitation patterns within these basins.

- L. 526: Suggest being more quantitative to clarify what “Most stations” means (e.g., include a percentage).
 - Agree. This text has been updated to: *Stations achieve $\mathrm{KGE} > 0$ for precipitation and $\mathrm{KGE} > 0.4$ for temperature, with the proportion of stations exhibiting lower skill not exceeding 3% for any catchment+variable combination and being 0% in most cases, likely reflecting localized measurement issues or unresolved sub-grid variability.*
- L. 548-560: Here and elsewhere, it would be helpful to have a sense of how large the GRUs are in terms of area for the comparisons against the SWE observations.
 - Agreed. The GRU size is included across all figures showing observation comparisons.
- L. 562: Suggest moving this sentence to the caption of Fig. 13 as it would be helpful there for interpretation of the figure.
 - Agreed and updated.

Figures and Tables

- Multiple figures (e.g., Fig. 5, 8, 10) include colors that are not accessible for someone with red-green color deficiency. Please check all figures and fix the color schemes.
 - Agree. We have reviewed and updated the color schemes across these Figures and checked the remaining with the Cryosphere tool.
- Figure 2: I am confused by the legend for the Tuolumne Snow Stations, which indicate they are either NOAA-NOHRSC or US NRCS. My understanding is that

many/most of the snow stations in the Tuolumne are part of the CA Department of Water Resources cooperative snow network. Please double check and correct as necessary.

- From Mortimer and Vionnet (2025), which describes the NorSWE dataset: *NRCS data compile observations from state-level data collection offices across the western US and Alaska. Some states, such as California (<https://cdec.water.ca.gov/snow.html>, last access: March 2025), also provide these observations through their own data portals. To ensure broad consistency across the region and to avoid introducing duplicate records, we chose to draw only from the NRCS database.*
- The Figure 2 caption has been updated for clarity: *The NRCS datasource compile observations from state level collection offices, including California Department of Water Resources.*
- Figure 4: This is not referenced directly in the text. Please do so.
 - Agree. This is now referenced.
- Figure 5: This is not referenced directly in the text. Please do so.
 - Agree. This is now referenced.
- Figure 6: It is unclear what the boxplots and area time series are summarizing. Is this GRUs or year or both? Please clarify.
 - Agree. The boxplots and area time series are summarizing calculations at station locations that are grouped per basin. The Figure caption has been updated with: *Both the boxplots and area time series summarize calculations at station locations grouped by basin*
- Figure 9: What does the background map represent? Elevation? Please explain in the caption and/or add a colorbar.
 - Agree. The caption is updated to read: *The background map represents elevation.*
- Figure 14: Please state/clarify the study basin. Also, provide the area of the GRU, as this is important context.
 - Agree. This is the Bow Study Basin, this information and the area of the GRU has been added.

References

Günther, D., Marke, T., Essery, R., and Strasser, U.: Uncertainties in Snowpack Simulations—Assessing the Impact of Model Structure, Parameter Choice, and Forcing

Data Error on Point-Scale Energy Balance Snow Model Performance, *Water Resources Research*, 55, 2779–2800, <https://doi.org/10.1029/2018WR023403>, 2019.

Citation: <https://doi.org/10.5194/egusphere-2025-6066-RC1>

RC2: 'Comment on egusphere-2025-6066', Anonymous Referee #2, 14 Mar 2026

Review of *Propagating Meteorological Uncertainty in Physically Based Mountain Snow Simulations* by Casson and co-authors for EGUSphere

Summary

The authors present a framework to quantify and propagate uncertainty in meteorological forcing through physically based snow modeling in mountainous terrain. It generates probabilistic meteorological forcing ensembles by combining station observations, topographic predictors, and atmospheric reanalysis with both regression and machine-learning methods to estimate spatial fields and their uncertainty. Precipitation observations are corrected for wind-induced gauge undercatch, and spatially correlated random fields are used to produce

ensembles of meteorological inputs. These ensembles drive the energy-balance snow model SUMMA in three contrasting basins: the Chena (Alaska), Bow (Canadian Rockies), and Tuolumne (Sierra Nevada); basins that are generally well known in the North American snow hydrology community.

Results show that precip undercatch correction improves winter precipitation estimates, dynamic predictors such as reanalysis enhance regression skill, and Random Forest methods outperform regression in generating meteo fields. Ensemble verification shows positive probabilistic skill. When forced by the ensemble, snow simulations reproduce observed SWE in realistic ways: timing, magnitude, and uncertainty ranges seem reasonable. Overall, the study demonstrates a reproducible and transferable approach for explicitly representing meteorological forcing uncertainty in mountain snow simulations to support hydrologic modeling and water-resource decision making.

I commend the authors on a compelling demonstration of best practices in hydrological modeling, reproducibility, and uncertainty characterization. I believe this paper will serve as a guide for many early-career hydrologists seeking practical examples of rigorous modeling workflows and code development. The use of FAIR principles and Snakemake-based workflows is particularly notable and valuable for the community. The paper is clearly within the scope of the journal and will likely be well cited in the hydrological modeling literature, and potentially beyond in other environmental modeling disciplines.

[Thank you for this positive assessment and the time taken to review our manuscript.](#)

Main Comments

Interpretation of Figure 13 – I found these results particularly interesting. Could the authors expand on the implications of the results shown in Figure 13? Specifically, what do the cross-site differences in S-bias & S-pattern relationships reveal about the relative strengths and limitations of the different forcing datasets?

[We thank the reviewer for highlighting Figure 13, which could benefit from further interpretation, including what cross site differences reveal about the site climatology and relative forcing datasets.](#)

[Across all study basin sites, the lower S-pattern values for the RF based ensemble indicate better temporal evolution of snowpack, including accumulation and melt processes. The ensemble forcing appears to better captures event-scale meteorological variability and temperature dynamics compared to deterministic](#)

reanalysis products, which tend to smooth these processes and may miss specific events captured in gauge records. |

|

The cross-site comparison of S-bias is indicative of how forcing datasets contribute to SWE magnitude more specifically. Deterministic reanalysis products exhibit more consistent but often biased behavior across sites, with a pronounced negative S-bias in the Bow basin, likely related to precipitation undercatch and unresolved orographic enhancement in precipitation and temperature. The RF-based ensemble spans both positive and negative S-bias values, indicating that it captures a wider range of plausible accumulation magnitudes. This spread reflects uncertainty in precipitation inputs and phase partitioning, while also demonstrating that many ensemble members reduce magnitude bias relative to deterministic forcings. |

|

The Taylor diagrams demonstrate similar findings, but displayed with different statistical metrics. Ensemble members generally achieve comparable or higher correlation with observations than deterministic forcings, though with a wider spread. |

Deterministic products occupy a narrower region of the Taylor space, reflecting more consistent but often biased representations of variability, whereas the ensemble provides a distribution of outcomes that includes both improved and degraded simulations. Across basins, these differences highlight how forcing uncertainty propagates differently depending on hydroclimatic regime. The Bow basin shows systematic magnitude bias linked to precipitation underestimation, whereas the Chena exhibits greater spread that may be associated with sensitivity to precipitation phase and accumulation variability. The Tuolumne shows tighter clustering, reflecting stronger topographic control and more consistent storm dynamics. |

|

The results depicted in Figure 13 demonstrate the strength of an ensemble framework to provide a more complete representation of uncertainty and process variability, while also containing realizations that improve upon deterministic forcing, rather than uniformly outperforming them.

|

The following additions will be made to the manuscript, to improve the discussion and interpretation of the results. |

Addition to Section 5.5.1 Evaluation with In Situ Measurements |

The RF-based ensemble generally exhibits lower S pattern values than deterministic forcings, which indicates better modeling of the temporal evolution of snow accumulation and melt. This suggests that ensemble forcing better captures the temperature driven processes and event variability that are often smoothed in reanalysis products. In contrast, differences in S bias highlight systematic magnitude errors. Deterministic forcings show more consistent but

often biased behavior, with pronounced negative SWE bias in the Bow basin, likely linked to precipitation underestimation and unresolved orographic effects. The ensemble spans both positive and negative bias, reflecting uncertainty in precipitation magnitude and phase partitioning.

Addition to Discussion Section 6.1 Ensemble Mountain Meteorology for Snow Modelling

The snow simulation results show that the ensemble framework improves representation of snowpack temporal dynamics while explicitly capturing uncertainty in forcing. Rather than uniformly outperforming deterministic products, the ensemble provides a physically meaningful range of outcomes that reflects sensitivity to precipitation magnitude, phase, and variability. This reveals both improved and degraded simulations across sites, highlighting where deterministic forcings are biased and where snow processes are most sensitive to meteorological uncertainty.

Treatment of precipitation phase: There was surprisingly little discussion of how precipitation phase is treated in the model and how the calibration and ensemble framework may influence the impacts of precipitation phase errors on snow simulations. Could the authors expand on this aspect?

I would be interested to know whether there is a metric that could evaluate the relative uncertainty of the ensemble during precipitation events when air temperatures fall within the model's specified snow-rain transition range. In other words, is there a way to quantify how the different forcing-generation methods influence the model's treatment of precipitation phase and the resulting uncertainty in snow accumulation?

We agree that precipitation phase deserves clearer treatment in the manuscript, as it is so implicit to both precipitation and temperature forcing uncertainty translation to snow.

We have added in explicit study of precipitation phase to address this reviewer comment, please refer to:

- Methods Section: 4.6.3 Evaluation of Precipitation Phase Uncertainty
- Results Section 5.5.3 Precipitation Phase Uncertainty
- Appendix D: Spatially Distributed Snow Measurement and Precipitation Phase Results

These sections provide a description of precipitation phase handling within SUMMA, as well as calculation of the uncertainty in precipitation phase (quantified as snowfall

fraction in mixed-phase events). This is evaluated for each basin over an annual climatological cycle. A spatial assessment also is presented, showing the standard deviation of snowfall accumulation normalized over the mean water year accumulation.

Representation of preferential snow accumulation areas: While the study focuses on GRUs and GRU-averaged snowpack metrics, it would be helpful if the authors could comment on the potential importance of resolving landscape units that exhibit preferential snow accumulation and harbor deep, persistent snowpack. These areas may provide critical surface and groundwater inputs and support important ecological functions. Such regions may also be the least likely to have in situ observations for either forcing or validation.

These landscape units may disproportionately contribute to basin water inputs and may either be more resilient to warming (snowpack refugia) or particularly susceptible to change. The manuscript clearly highlights the need to better resolve slope-scale wind effects from the perspective of gauge collection efficiency, but what about the potential to improve resolution of slope-scale orographic dynamics (e.g., the improvement of high-resolution reanalysis products that can capture precipitation banding from orography and atmospheric rivers, e.g.) in locations where precipitation gauges are absent?

We agree that preferential snow accumulation areas and sub-GRU heterogeneity are critical components of mountain hydrology that are not explicitly resolved in this study. Our use of GRU-averaged representation prioritizes consistency across large domains, but necessarily smooths processes such as wind redistribution, preferential deposition and areas resilient to warming.

We also agree that improving representation of slope-scale orographic dynamics is an important direction for future work. While our framework partially accounts for this through dynamic predictors and spatially correlated uncertainty, it does not explicitly resolve fine-scale precipitation variability such as orographic banding or atmospheric river structure. Emerging high-resolution reanalysis and regional atmospheric models offer a promising pathway to better represent these processes, particularly in data-sparse regions. We have added discussion of these limitations and opportunities in the revised manuscript.

The following text has been added to the Discussion Section 6.3: Potential Application and Limitations

A key limitation of this framework is the representation of snow processes at the GRU scale,

which aggregates substantial sub-grid heterogeneity. Figure~\ref{fig:mountain_observatory} illustrates this limitation, showing that SWE measurements collected under contrasting local conditions within a single GRU can differ markedly. Preferential accumulation areas, including wind-sheltered slopes, leeward deposition zones, and forest openings, often sustain deeper and more persistent snowpacks than surrounding terrain, yet are not explicitly resolved. These locations are also underrepresented in meteorological and snow observation networks, compounding uncertainty in both forcing and model evaluation particularly in melt periods. Although the ensemble framework partially accounts for unresolved variability through spatially correlated perturbations, it does not explicitly represent slope-scale processes such as wind redistribution and preferential deposition. Addressing these limitations will require finer spatial discretization and higher-resolution meteorological forcing, with emerging regional reanalyses and atmospheric models offering improved representation of orographic precipitation and other terrain-driven processes.

Detailed Comments

Lines 253-254: Typo: "where data were co-located measurements"

- Agreed, and updated

Line 410: Replace with "Meteorology"?

- Agreed, and updated.

Lines 458-459: It may be helpful to mention that snow pillows also record conditions in the absence of overhead forest canopy, which may influence the representativeness of comparisons with spatially averaged model outputs.

- Agreed and updated.

Citation: <https://doi.org/10.5194/egusphere-2025-6066-RC2>

Additional References:

Crumley, R. L., Palomaki, R. T., Nolin, A. W., Sproles, E. A., & Mar, E. J. (2020). SnowCloudMetrics: Snow Information for Everyone. *Remote Sensing*, 12(20), 3341. <https://doi.org/10.3390/rs12203341>

Mortimer, C., & Vionnet, V. (2025). Northern Hemisphere in situ snow water equivalent dataset (NorSWE, 1979–2021). *Earth System Science Data*, 17(7), 3619–3640. <https://doi.org/10.5194/essd-17-3619-2025>

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., Marks, D., Mattmann, C., McGurk, B., Ramirez, P., Richardson, M., Skiles, S. M., Seidel, F. C., & Winstral, A. (2016). The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sensing of Environment*, 184, 139–152.
<https://doi.org/10.1016/j.rse.2016.06.018>