Reviewer Responses - egusphere-2025-6066-RC1

Dear Referee #1,

Thank you for talking the time to review and provide thoughtful, constructive comments for improvement of the study. We agree with all comments and recommendations made by the author, and that revision of the manuscript to include them will improve the study.

Please see below our responses to in blue.

Best,

David Casson, on behalf of all 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. As you highlight the LSE for parameter identification, we reflect that this specific contribution should be better represented in the manuscript abstract, discussion and conclusions.

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 will improve consistency between the Figure 3 and Section 4 titles to allow better cross-referencing for readers, as well adding references to the text.

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.

   In the revised manuscript, we will include evaluation of the Crumley et al, 2020 snow cover metrics, snow cover frequency (SCF) and snow disappearance date (SDD) based on the MOD10A1 snow cover product.

We will also include lidar based SWE available data available over the Tuolumne watershed. Specifically, the processed airborne lidar survey SWE data made available by Plfug and Lundquist (2020).

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 is needed for this specific conclusion, and that the observational SWE data are often unable to constrain and evaluate the melt out timing. The additional metrics computed from spatial snow data will help to clarify this result. Additions will be made to the manuscript based on these findings.

   Magnitude was meant to refer specifically to SWE in the accumulation period, and not melt. SWE magnitude is quantitatively assessed, as shown in Figure 13 (S bias in top plots and RMSE in bottom plots) with as the reviewer suggests, melt timing less effectively. The text requires update for clarity magnitude was meant to refer to peak SWE and not melt. The specific update will be made following the quantitative evaluation.

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
- L. 49-51: You might also note here that the airborne and terrestrial lidar provide high spatial resolution.
  - Agreed
- L. 59: Suggest adding “lumped” to “conceptual degree-day models”.

- Agreed

- L. 207: Add “across study basins” at the end of this sentence.
  - Agreed
- L. 209: Should read “Sun et al. (2019)”.
- L 212-214: Is this the same as ERA5-Land? Please clarify.
  - This is not the exact same product as ERA5-Land. The text will be updated to: *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 will be updated to*: A regional approach that assumes consistent gauge type and configuration 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 Tuolume 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
- L. 379: It would help to have a brief summary of the key model decisions to understand how SUMMA was setup here.
  - Agreed. The text will be added. *With respect to snow modelling, the model decisions select the dynamic snowpack layering scheme adapted from the Community Land Model (Lawerence 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_theta and N_atts variables.
-
    - We thank the reviewer for catching this. We will update the text to clearly describe this as follows: *To this end, we use a matrix of $N_{sites} \cdot N_{i=0}$ 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?
    - The following response and generated figure are provided for the reviewer, and will be updated for clarity in the updated manuscript.

    - For the 162 stations used to train the emulator (in red in Figure 1a), we compared the Taylor Skill Score (TSS; Taylor, 2001) from SUMMA (i.e., simulations vs. observations) and the emulated TSS value using Random Forest (Figure 1b). The emulator comprises 162,000 SUMMA simulations (162 stations × 1,000 parameter sets; density dots in Figure 1b), for step i = 5 of the iterative process. To obtain new parameter sets for each station, we search the emulated response surface using a Genetic Algorithm (GA), yielding 100 new parameter sets per station. SUMMA is then run for these parameter sets to compute TSS, providing an evaluation error for parameter sets unseen during Random Forest training. At step i = 5, the MAE between emulated and SUMMA TSS values for these unseen parameter sets is 0.242 TSS units.

We also apply the trained emulator from each iteration step to search for parameter sets at unseen stations (shown in red in Figure 1c, n = 608), again using a GA to produce 100 candidate parameter sets per station. The error between emulated and SUMMA TSS values is then computed across these 608 stations × 100 parameter sets. As shown in Figure 1d, this error decreases steadily across iteration steps, reflecting both the growing training dataset — which increases from 500 to 1,000 parameter sets per station between i = 0 and i = 5 — and the progressively better-explored parameter space.

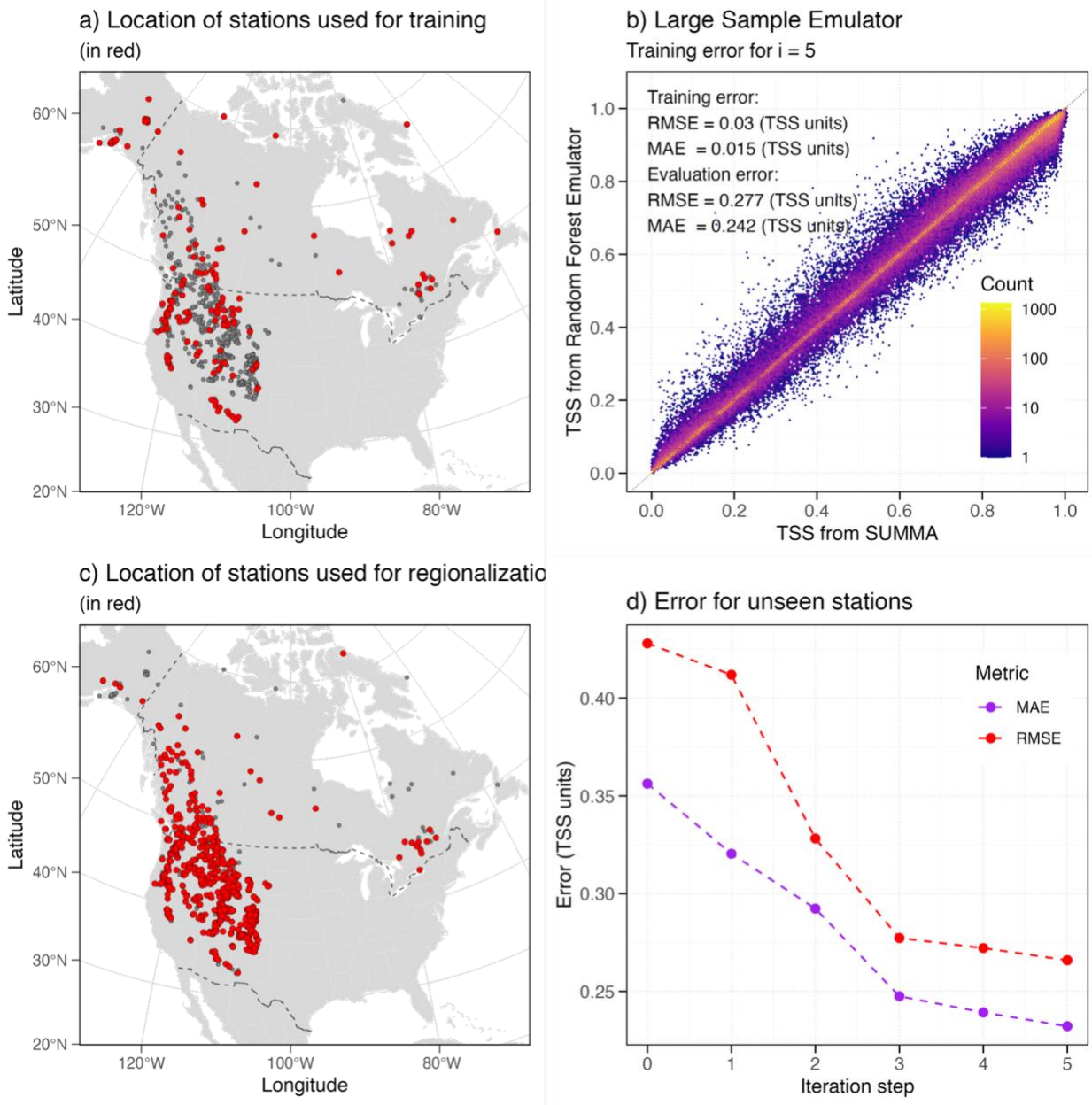

Figure 1: Emulator errors for training and unseen stations. a) location of the basins (in red) used to train the emulator (n = 162). b) Training error for emulator at iteration step i = 5, and evaluation errors for the new parameter sets obtained through the Genetic Algorithm optimization search. c) stations used for parameter regionalization (i.e., unseen stations, n = 608). d) Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) between SUMMA TSS and emulated TSS for the unseen stations and from the emulator trained for each step of the iterative process to refine the emulator.

- *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
- 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
- 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. Text will be updated to: *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 to update and include a percentage
- 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. Information on GRU sizing will be added in data and results.
- L. 562: Suggest moving this sentence to the caption of Fig. 13 as it would be helpful there for interpretation of the figure.
    - Agreed

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. Will review and update color schemes.
- 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 will be 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 and will reference.
- Figure 5: This is not referenced directly in the text. Please do so.
    - Agree and will reference.
- 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 will be updated to clarify this.
- Figure 9: What does the background map represent? Elevation? Please explain in the caption and/or add a colorbar.
    - Agree. The background map represents elevation. This will be updated in the caption for clarity.
- 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 will be 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

Additional References

Pflug, J. M., & Lundquist, J. D. (2020). Inferring Distributed Snow Depth by Leveraging Snow Pattern Repeatability: Investigation Using 47 Lidar Observations in the Tuolumne Watershed, Sierra Nevada, California. Water Resources Research, 56(9), e2020WR027243. https://doi.org/10.1029/2020WR027243

Pflug, J. (2020). Pflug and Lundquist (2020) Data repository, HydroShare, http://www.hydroshare.org/resource/b9b5b667eb074576a932ddc32a5e924