Dear Reviewer 2,

The authors greatly appreciate the time you took to review our paper. Your comments and suggestions will greatly improve its overall quality. Please find our responses to each comment below. Reponses are in red. Manuscript changes will be in red italics.

First, the introduction (and other sections) should describe similar efforts – and then focus on the differences in this study. First, the study of Geisler et al. is rather similar to what is presented here (Geissler,J.,Rathmann,L.,&Weiler,M.(2023).Combining daily sensor observations and spatial LiDAR data for mapping snow water equivalent In a sub-alpine forest. Water Resources Research,59(9), e2023WR034460.https://doi.org/10.1029/2023WR034460), yet this paper is not cited or discussed. Geissler et al. develop an approach to fill in the gap between lidar flights using daily depth observations from stations. The scale of the Geissler study and this one are also similar. Second, the methodological framework used in this paper is the same as that presented in Herbert et al. (2025), which used more than 50 lidar acquisitions across 8 basins in Colorado and California.

Thank you for bringing Geissler et al. 2023 to our attention, we apologize for not discussing a clearly related paper.

We will add an additional paragraph in our Intro to describe these two papers. The paragraph will also focus on their previous work, what we are leveraging from their work (i.e., the overlap), and how this work will differ.

Recent work has bridged the temporal gap between lidar acquisitions and daily snow evolution by combining repeated lidar surveys with daily in situ observations to reconstruct basin-scale SWE at sub-alpine forest sites (Geissler et al., 2023). Related studies have framed snow depth estimation as a residual between episodic lidar observations and a continuous temporal driver timeseries, using RF models to learn and reconstruct spatial patterns across multiple mountain basins in the western United States (Herbert et al., 2025). However, these studies do not systematically evaluate how this RF–driver approach performs across different snow–climate regimes and seasons, nor how model skill depends on lidar cadence, acquisition timing, or the choice of driver dataset. Understanding these dependencies is important for designing practical lidar sampling strategies, including how many flights are required, when during the season they are most valuable, and whether reanalysis forcing can replace in-situ measurements.

While Herbert et al. (2025) is cited in limited contexts (e.g., around line 284 when defining the target variable, and around line 650 when discussing the number of lidar acquisitions), the manuscript does not clearly acknowledge that the core problem formulation and modeling approach closely follow that prior work, including (i) casting snow depth estimation as a residual

(relative-depth) between lidar and a temporal driver, and (ii) using a random forest model to reconstruct spatial patterns from those residuals.

We hope that we adequately addressed Herbert's contributions in the above appended paragraph. Two limitations noted in that paper are (1) their reliance on in situ snow depth data, and (2) that they did not test the transferability between basins in different regions and climates. To further demonstrate this difference, we also include the last two sentences in the new paragraph:

However, these studies do not systematically evaluate how this RF–driver approach performs across different snow–climate regimes and seasons, nor how model skill depends on lidar cadence, acquisition timing, or the choice of driver dataset. Understanding these dependencies is important for designing practical lidar sampling strategies, including how many flights are required, when during the season they are most valuable, and whether reanalysis forcing can replace in-situ measurements.

We reiterate our objectives in the last paragraph:

To address these gaps, we analyse two contrasting snowpacks across an extensive multi-year record of airborne snow depth observations at (1) Mores Creek, Idaho, and (2) Hubbard Brook Experimental Forest, New Hampshire. We ask: (i) which terrain and land-cover factors explain spatial variability in prediction error and provide the most predictive value; (ii) how much lidar cadence matters via a drop-date experiment; (iii) how well models transfer across winter phases and which single phase yields the most effective cross-season performance if only one acquisition is feasible; and (iv) whether ERA5-Land can substitute for in-situ forcing for daily prediction. Our aim is to preserve the strengths of full-basin airborne lidar mapping while reducing its operational burden by identifying the minimum flight frequency, optimal seasonal timing, and whether ERA5-Land forcing enables accurate daily mapping in basins lacking in-situ networks.

The contribution of this paper lies not in introducing a new framework, but in extending and testing an existing approach in a different setting. Emphasizing these differences, rather than presenting the method as newly developed, would clarify the contribution and better situate the work within the existing literature.

Agreed. We have revised the Introduction to clearly state that we build on the RF residual framework developed by Herbert et al. (2025) and related lidar station approaches such as Geissler et al. (2023), rather than presenting the method as new. We now explicitly frame the novelty of this study as extending and testing this existing approach in two contrasting snow–climate regimes, quantifying sensitivities to lidar cadence and seasonal timing, and evaluating the use of ERA5-Land as an alternative driver where in situ data are unavailable.

Second, the manuscript includes a fair amount of extraneous text that distracts from the main points. The manuscript would benefit from tightening to better emphasize its core contributions. For example, the first paragraph of the introduction could be shortened to two sentences.

Agreed. We will shorten the first paragraph accordingly:

Seasonal snow cover provides valuable water for billions of people across the Northern Hemisphere. In the American West, snowmelt supplies roughly 53% of annual runoff (Li et al., 2017), and both western and eastern snowpacks are projected to decline under future climate scenarios, with implications for water security, wildfire risk, soil freeze-thaw dynamics, and cold water habitats (Callaghan et al., 2011; Diffenbaugh et al., 2015; Ford et al., 2021; Intergovernmental Panel On Climate Change (Ipcc), 2023). These natural, social, and economic stakes underscore the importance of monitoring snow depth distribution and its seasonal evolution to improve water forecasting, guide resource allocation, and anticipate ecologic and hydrologic impacts. While in situ networks such as snow courses and automated stations (e.g., SNOTEL) provide accurate point-scale snow depth and snow water equivalent (SWE) measurements, they remain sparse, biased toward accessible locations, and poorly capture the spatial heterogeneity of snowpacks across diverse terrain (Dong, 2018; Meromy et al., 2013). This spatial gap has increasingly driven reliance on remote sensing to scale snow observations over larger domains.

The text around line 530 and following could be greatly abridged.

Agreed. We will simplify the text from 530-~550 to:

Cross-phase transfer analyses trained the RF model in one winter phase and tested it in another (Figure 7). At Mores Creek, models trained on all phases perform best, and single-phase experiments show that mid-winter training generalizes well to late season, while early-season training is a workable compromise for predicting both middle and late phases. In contrast, models trained only on late-season data transfer poorly to earlier phases, indicating that late-season acquisitions alone are a weak basis for cross-season prediction. At Hubbard Brook, single-phase models show modest degradation relative to all-phase training, and transfer between early–mid and mid–late phases is approximately symmetric. Overall, these results indicate that mid-season acquisitions are the most transferable at Mores Creek, whereas at Hubbard Brook combining both accumulation and melt phases is more beneficial than optimizing for a single phase, likely because Hubbard Brook's transitional snowpack is more temporally variable.

The discussion includes sections that, while interesting, are not directly tied to the study design or results. For example, lines 604–610 do not relate directly to the results and could be shortened or removed.

Lines 604-610 will be removed.

The conclusion reads more like a short discussion than a set of clear takeaways. I recommend focusing on the text needed to support the main findings and removing or condensing the rest.

Thank you for this comment. Below is a shortened conclusion that focuses on our main findings and shortened to 3 paragraphs:

This study shows that a relative-depth RF model can use a small number of airborne lidar acquisitions and a single daily snow depth driver to produce temporally coherent, basin-scale daily snow depth maps in contrasting snow regimes. The model learns lidar–driver residual patterns on flight days and applies them between flights, yielding accurate, meter-scale fields without a full energy-balance model, provided predictions remain within the observed parameter space. Across both basins, the approach explains roughly 90% of the variance with typical errors of about 8–28 cm and near-zero bias, demonstrating that episodic lidar surveys can be leveraged for daily mapping with modest data requirements.

Spatial diagnostics and feature attributions confirm that terrain and land cover organize both predictions and errors in a regime-dependent way. Elevation, aspect/northness, microtopography, and redistribution proxies exert the strongest influence, with signals most pronounced during transitional periods. In the heterogeneous, wind-affected Mores Creek basin, the model provides substantial gains over using an in situ site alone, particularly in deep-snow zones, whereas in the smaller, more homogeneous Hubbard Brook basin, the main benefits are reducing local biases and refining melt timing rather than dramatically improving basin-mean metrics.

Cadence and forcing experiments translate directly into survey and driver design. Errors decline as additional lidar flights are added, with clear diminishing returns after only a few acquisitions: at Mores Creek, roughly five early-season flights, four mid-winter flights, and five late-season flights recover most of the attainable skill, while at Hubbard Brook a smaller number of surveys (three early–mid and five mid–late flights) is sufficient for similar basin-mean performance. Where only one or two flights are feasible, prioritizing middle season acquisitions captures most of the attainable skill, whereas late-season-only training is least transferable. Using ERA5-Land as the temporal driver produces snow depth predictions that closely match those driven by in situ measurements, with small, mostly modest biases, indicating that reanalysis forcing is a viable substitute in basins lacking station data. Together, these findings outline a practical path from episodic airborne lidar campaigns to cost-aware, daily snow depth products that can

support melt-season operations, access and hazard planning, and short-term water management.

Third, the figures could be improved for clarity and readability. Some appear fuzzy in my version, and several rely on shades of gray where color would improve interpretation.

We will export figures 2, 3, 6, 7, 8, 9, and 10 with dpi = 400.

In Figure 7, the grayscale is difficult to relate to RMSE values. Overall, improved figure clarity would enhance readability.

We will change Figure 7 colorbar from greys to viridis and do the same for figure 8. Figure 10, we changed the lines from grey and black to orange and blue.

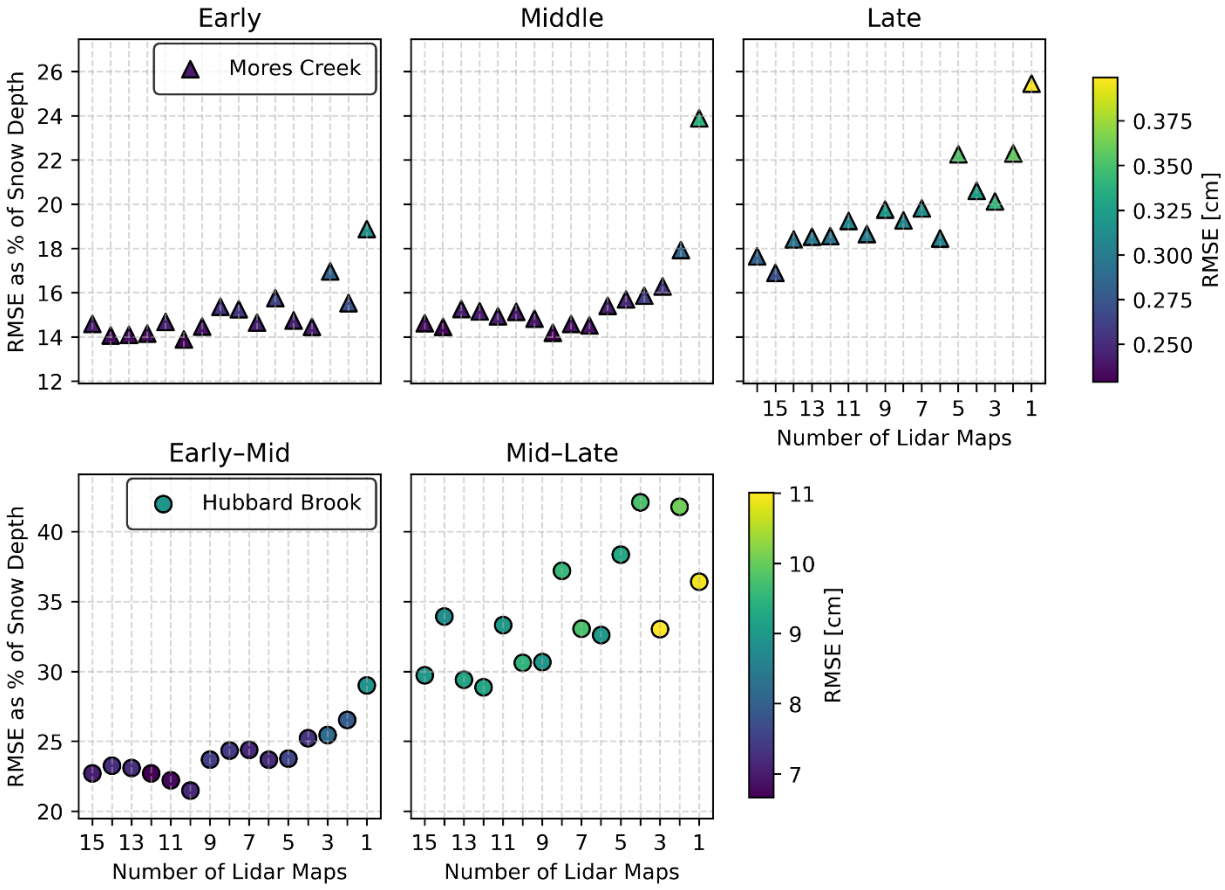


Figure 7. Updated color ramp to viridis and exported at dpi = 400.

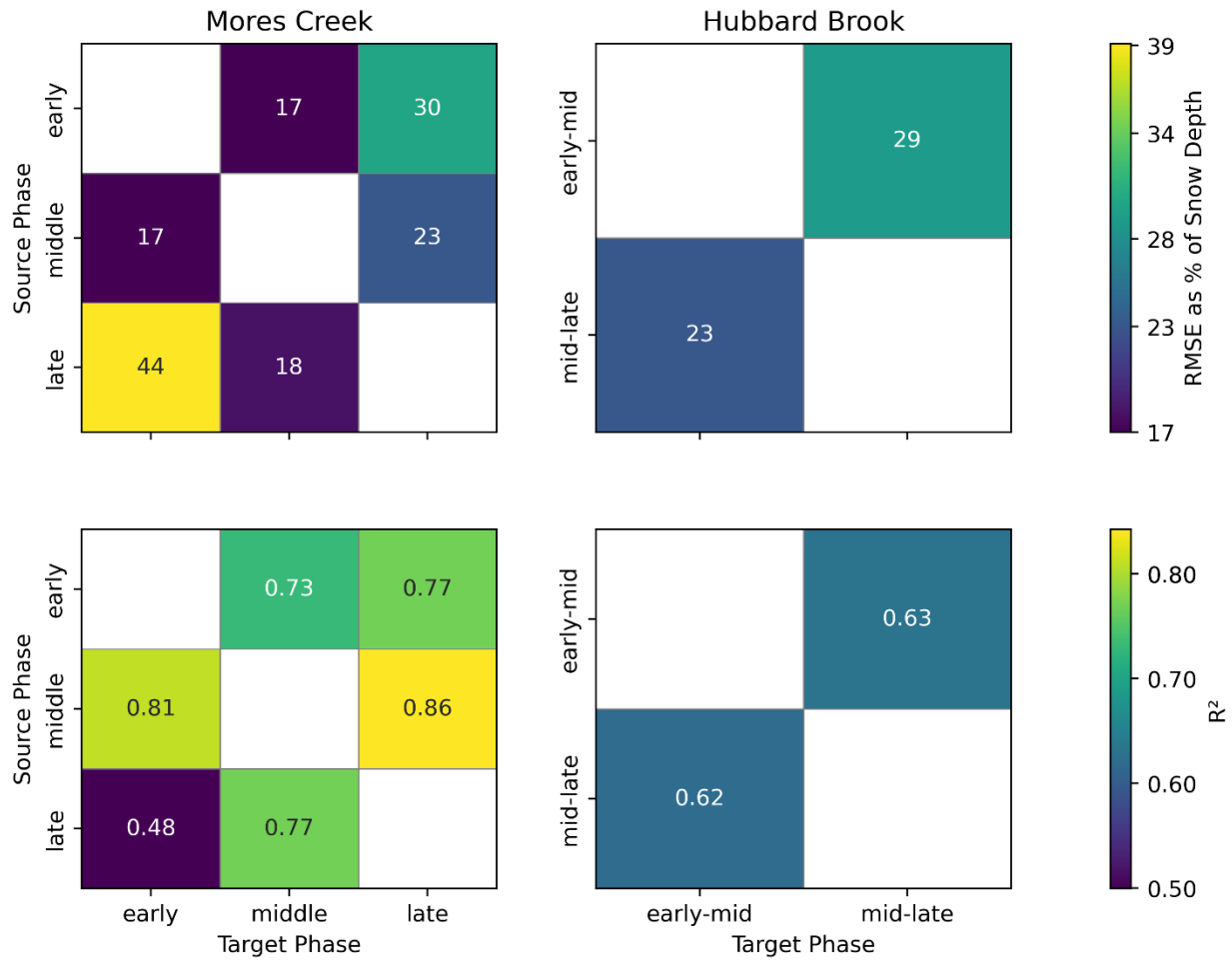


Figure 10. Updated color ramp to viridis and exported at dpi = 400.

In Figure 4c and 4d, the dark histograms appear to show the distribution of differences between each pixel and the in situ site (line 450), which represents basin-wide snow depth referenced to the station, rather than model error as described.

This is an important comment and your interpretation is correct. The dark histograms in Figures 4c and 4d represent the distribution of differences between the spatially distributed snow depth and the in-situ station observations across dates, representing a baseline assumption rather than a model error.

Our intent with this figure was to show the value added by the RF approach. Because the RF model is driven by that same in-situ data, comparing the baseline assumption (shown in black) against the RF model predictions (shown in gray) highlights the improvement achieved by the spatial modeling. Here, baseline assumption implies that the in situ site represents the entire basin. The comparison also demonstrates that assuming a single in situ station is representative

of the surrounding basin is far more consequential at the larger Mores Creek site than at the smaller Hubbard Brook site.

We agree that our previous phrasing inadvertently implied the black histogram was a model result. To prevent confusion, we have revised the text (formerly line 450) and the Figure 4 caption to explicitly state that the dark histograms represent the baseline spatial difference relative to the in situ site, reserving “model error" strictly for the gray RF model distributions

We will update the figure caption to the following:

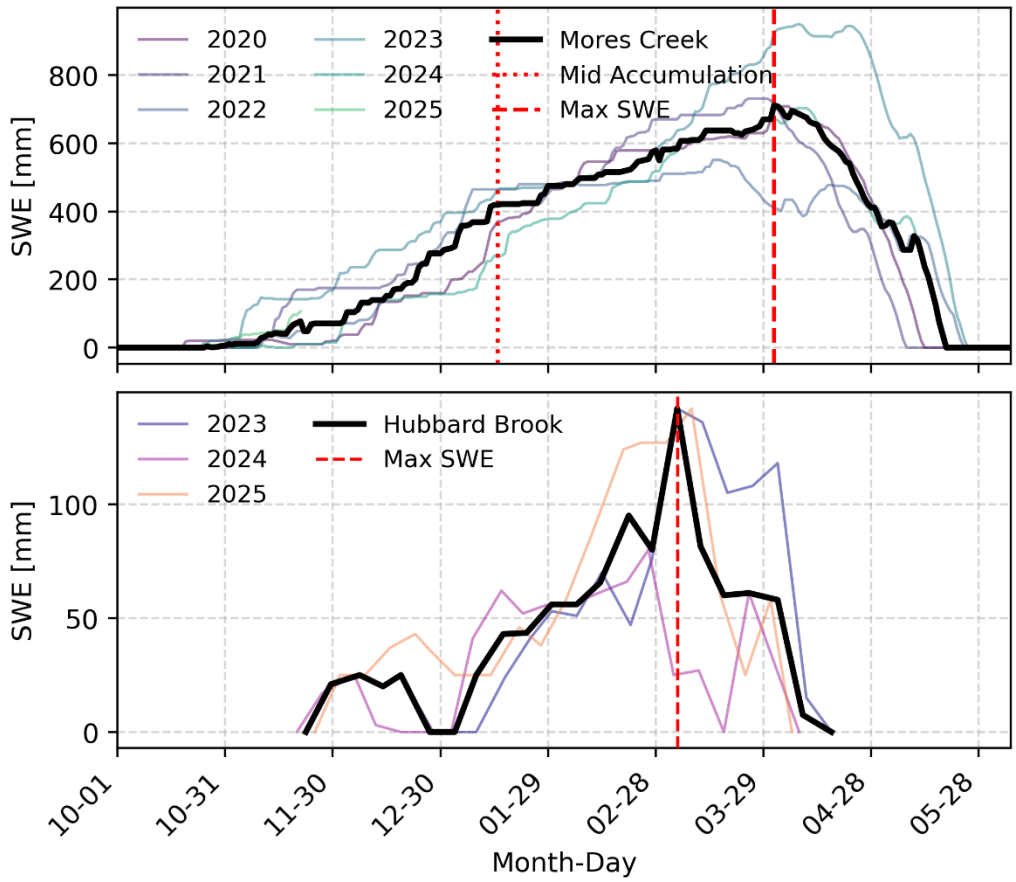
Figure 4: Spatial distribution of averaged model residuals (RF prediction – lidar observation, light grey) for a) Mores Creek and b) Hubbard Brook. Histograms comparing the RF model errors across the basin (light grey) against the distribution of differences when assuming the basin-wide snow depth is equal to the in-situ station observation (dark grey) are shown for c) Mores Creek and d) Hubbard Brook. Arrows and circled areas indicate noted areas of model overprediction (red) and underprediction (blue).

We will also update the paragraph for more clarity:

To isolate the value of the spatial modeling approach, the distribution of mean residuals from the RF model (light grey) was compared against a baseline assumption where the local in-situ snow depth is uniformly applied across all pixels in the basin (dark grey; Figure 4c and 4d). At Mores Creek, this in-situ baseline approach (using SNOTEL daily snow depth) resulted in a basin-wide difference of -13.2 cm, regularly underestimating deep snow areas by more than 50 cm (Figure 4c). At the Hubbard Brook basin, the in-situ baseline (using SC2 daily snow depth) showed a basin-wide difference of -9.4 cm, underestimating snow depth in 91% of pixels. This comparison highlights a key takeaway: the assumption that a single in-situ station is representative of basin-wide snow distribution is significantly more consequential at the larger, more topographically complex Mores Creek site than at the smaller Hubbard Brook site. Still, this assumption can break down at relatively small scales (<1 km2), as shown by local differences exceeding -30 cm at Hubbard Brook (Figure 4d). In contrast to the baseline assumptions, the RF models successfully corrected these discrepancies across both sites, proving generally unbiased (bias within ±2 cm) with an even distribution of positive and negative average residuals (Figure 4d).

In Figure 2, it would be helpful to show individual yearly traces in different colors, similar to a standard SNOTEL plot, rather than a gray band.

In Figure 2, we will add the annual SWE timeseries for both Mores and Hubbard.



Fourth, many of the dynamic variables appear to have limited influence on prediction (Figure 5: "While dynamic predictors provided minimal predictive value…"). The authors should exclude variables with little predictive value, particularly given the large number of predictors used. This would simplify the model and improve interpretability and transferability. Alternatively, if there is a reason to retain these variables, this should be clearly explained and justified.

We agree that many of the dynamic variables have lower SHAP importance than the static terrain and canopy predictors. Our intention in this study was not to derive a minimal predictor set, but rather to assess how dynamic meteorological information contributes alongside terrain and canopy within a single model. We therefore retained the dynamic variables for two reasons: 1) some dynamic features (DOWY, forward-looking air temperature, lagged snow depth) exert important influence on model predictions, consistent with their expected physical roles in snow accumulation and melt (the daily evolution) 2) we want to quantify their contribution relative to terrain and canopy predictors and to diagnose under what conditions temporal drivers matter. For clarity, we will add the following in the section 5.1.4:

Although most dynamic predictors have lower SHAP importance than static terrain variables, we retained them for two reasons. First, some dynamic features (DOWY, forward-looking air

temperature, lagged snow depth) exert seasonally dependent influences on model predictions. Second, part of our goal was to quantify their contribution relative to terrain and canopy predictors within a single model, rather than to derive a pared-down model. Future applications could prune or simplify this dynamic predictor set to improve interpretability and transferability into an operational setting once the relative contributions of these variables are established.

Minor comments:

L63: and perhaps equally important, lidar only provide snow depth, not SWE

Will update the sentence to:

Uncrewed aircraft system (UAS) lidar provides finer-scale, on-demand mapping over smaller domains and is valuable for validation and model training, but it does not by itself solve the cadence gap and only provides SWE.

L134: this could be shortened, at least for the content included. I think it could be helpful to show variability from year-to-year in figure 2 - a series of small subplots perhaps. each with the "average" vertical lines in the text superimposed. This would show how these date compare to the year-to-year evolution of the snowpack. This would be helpful to evaluate the results describing transferability from early to late season, etc.

Thank you for this comment. We updated figure 2 to include the yearly SWE timeseries. We also update the text to be more streamlined and will read as the following:

Winter phases were defined using site-specific SWE climatologies. At Mores Creek, 2019–2025 SNOTEL data identified a mean peak SWE date of April 2 and an accumulation midpoint of January 16, delineating three periods: early (accumulation; N=5 lidar surveys), middle (peak storage; N=8), and late (ablation; N=5). At Hubbard Brook, the transitional snowpack necessitated a two-phase division: early-mid (accumulation; N=8) and mid-late (ablation; N=8), separated by a March 6 peak storage date. This threshold reflects the near-term (2023–2025) average while accommodating the coarse resolution of weekly sampling intervals. To ensure balanced sample sizes across these two phases, two acquisition collected on March 1, 2024 and March 3, 2025, were classified into the mid-late phase despite occurring marginally before the climatological peak.

L230: How was this choice made. Why not reverse the procedure (predictor versus validation) to evaluate robustness of results?

Thank you for this comment. At Mores Creek, we used the SNOTEL station for training and evaluation because it is the only site with a complete record (2020-2025), covering the lidar campaigns. Freeman Station, provides snow depth only for a single winter 2024-2025, lacking the met variables needed to act as the temporal driver. Because this mismatch in records, we

only use Freeman as an independent high elevation validation site. Including Freeman in the analysis allowed us to evaluate the robustness away from the driver location (SNOTEL) without limiting analysis to a single year. We will modify the sentence to the following for clarity:

These measurements are used solely for model validation, providing an independent check at a higher elevation site not used to drive the model.

L512: not monotonically, maybe roughly or generally? it does bounce around a bit, so not strictly monotonically.

Thank you for your comment. We will make the following change:

'…generally increased…'

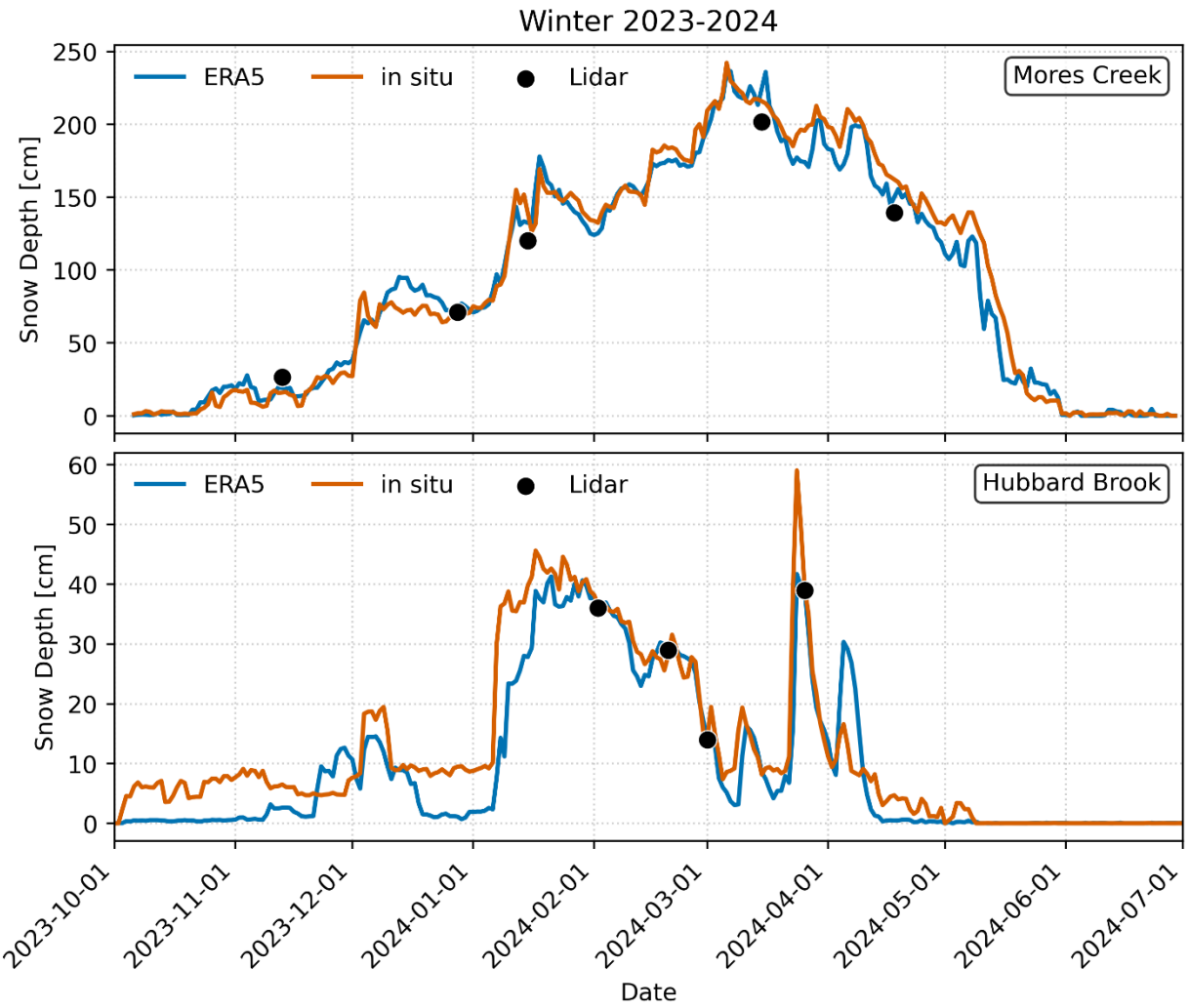
L573: but would it be the same "everywhere", or would you need to make the "on site" comparison, then determine the bias? I.e., is it transferable without local information as a test?

This is an important comment and helpful catch. In our results, the ERA5-driven RF model shows small, nominal biases when compared to in situ observations, and these biases are smaller than those obtained when using another in situ station to drive the model. This pattern is evident in the bias metrics shown in Figures 3 and 9. We will revise the text to emphasize that ERA5-Land forcing does not introduce a strong systematic bias in our basins, rather than suggesting a generic bias correction is required.

Overall, ERA5-Land appears to be a reliable substitute where in-situ forcing data is unavailable.

L584: why not color? the gray line is very difficult to see. the who figure is fuzzy.

Updated with color and dpi = 400.



L614: How is this result supported by the experiments? Rewrite or clarify.

Thank you for this comment. We will update the paragraph for more clarity. We will bring in the work by Herbert et al. 2025 who performed a sensitivity analysis to understand how the RF will predict snow in other parts of the basin once the in situ obs hit 0. We will also add suggestions that might help mitigate RF outputs predicting no snow, too early. The following paragraph will read as follows:

Another limitation arises from the model's reliance on the temporal snow depth forcing from in situ stations used as an input driver. Because the RF effectively predicts relative changes in snow depth through time, its accuracy depends on the fidelity of this forcing. If the forcing time series is biased relative to parts of a basin (e.g., biased low or returning to zero too early), then modelled pixels that historically retain deeper or more persistent snow may lose their dynamic characteristics or be forced toward zero values, despite physically retaining snow. This limitation underscores that the approach relies on representative, unbiased temporal snow depth inputs to

capture the full spatial and temporal variability of the snowpack. In related work, have shown that combining point snow depth with additional dynamic predictors and learned spatial patterns allowed the RF residual framework to maintain persistent snow at higher elevations even after a SNOTEL site had melted out, effectively extending the temporal utility of point measurements (Herbert et al., 2025).

L672: this section seems valuable, given the west-east contrast is central to the study, and unique from previous studies.

Agreed, thank you for this comment.

L746: the conclusions feel a bit redundant with the discussion. The conclusions could be more "conclusions" with the main points from the study.

We shortened the conclusion to 3 paragraphs and focused it more on our main findings. We directly address this in earlier comments.