Author responses to comments of reviewer 1

This manuscript presents an extensive analysis of machine learning capabilities for snow depth estimation. The authors compare a variety of machine learning (XGBoost) model configurations and apply a threefold nested cross validation to evaluate their approach. The inputs to the machine learning model are remote sensing data from Sentinel-1 (of which PolSAR had not been used before to estimate snow depth), downscaled meteorological forcing data and physically-based model simulations. The authors then evaluate the importance of features in the machine learning model, and the spatial predictions of snow depth at unseen locations by the model.

The aims and findings of this study are interesting, with the main novelty being the inclusion of PolSAR variables as well as meteorological forcings or a physically-based model forced with those to predict snow depth at high resolution (100 m) over the Alps. I am impressed with the amount of data processing and careful methodological procedures that the authors went through, which seems very robust.

However, I think the manuscript should more clearly state the novelty of this study in comparison with Dunmire et al. (2024). While the authors claim that the snow depth estimations are improved with the inclusion of PolSAR and meteorological forcings, it is hard to see any significant improvement when comparing similar figures between the manuscripts. Even within this manuscript, it is often claimed that a method improves snow depth estimates without this clearly seen in the figures. Furthermore, I have several concerns regarding the presentation of results, some of them are not very clear and there are many instances of "results not shown". I believe the authors need to improve the manuscript before it can be published, and I hope my comments below will help.

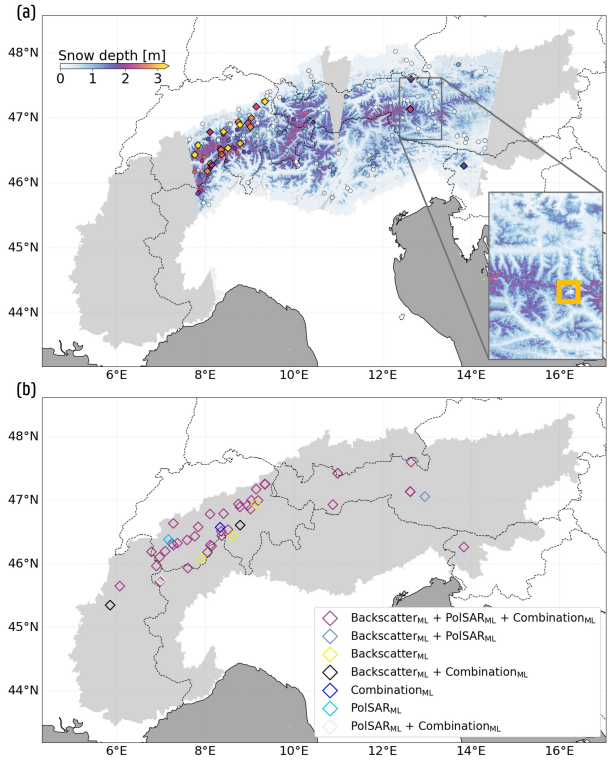
First, we would like to thank the reviewer for the useful comments, which have helped us to improve the manuscript. Below, we list our replies to the comments (red text). The original comments are included in black, new text that will be added to the manuscript is shown in blue. The indicated line numbers refer to the revised manuscript.

Main Comments

Comment 1
The title claims snow depth estimation over the European Alps, but there is no map of estimated snow depth over the European Alps, and no map of the predicted snow depth validation over the entire mountain range (as there is in Figure 2, and Figure 7, in Dunmire et al 2024).

We have included a Figure of predicted SD on 18 January 2018 in the revised manuscript (Review Fig. 1), that displays how the predictions for a specific date would look like for the Combination$_{\mathrm{ML}}$ configuration. Note that not the whole Alps are covered as in Fig.7 in Dunmire et al. (2024), as the area covered by the predictions is limited by the S1 swaths available on a specific day. The Figure also displays the measured SD at the stationary sites as validation, and indicates those sites that display a site bias $\leq$0.5 m within the spatial nested CV framework.

Comment 2
About XGBoost: Besides referring the reader to Chen and Guestrin (2016), I think there should be at least a few lines description of what this ML model is and its characteristics, and why the authors (or previous authors) chose this model.



Review Figure 1: Spatial SD estimate and underestimation of the Combination$_{\mathrm{ML}}$ configuration. (a) Prediction on 18 January 2018 encompassing both an ascending and descending S1 observation, derived as the mean SD from the five XGBoost models trained within the spatial nested CV framework. Observations are indicated by dots and squares, with squares representing sites exhibiting a site bias ≤ -0.5 m for this framework (displayed in (b)). Estimates over Austria are highlighted, with the yellow square marking an area of underestimation attributable to errors in the SCF product used. (b) Measurement stations exhibiting a mean site bias ≤ -0.5 m within the spatial nested CV framework (zero-measured SD excluded). Sites with fewer than 10 observations are excluded. Colors indicate the configurations for which this underestimation occurs. Across the displayed sites, 23% of the observed SD measurements ≥ 2.5 m, compared to only 2% across the full training dataset.

We agree with your comment. The revised manuscript includes a small new Subsection "3.1 XGBoost model selection", which describes the model briefly, and clarifies the usage of XGBoost in our manuscript. In early research stages we did a small model selection experiment, in which we compared different tree-based traditional ML models, including Random Forest, LightGBM, CatBoost and XGBoost. This new Section (L229-235 in the revised manuscript) is as follows:

L229-235 "Within this study, we deployed XGBoost to estimate SD across the European Alps. XGBoost is a tree-based traditional ML algorithm that constructs multiple decision trees in a sequential order, to minimize a differentiable loss function (Chen and Guestrin, 2016). Thereby, new trees are trained to fit the residual errors of the previously fitted trees, while incorporating regularization terms to reduce model complexity, and including additional mechanisms that contribute to its computational efficiency (Chen and Guestrin, 2016). We chose this model after comparative experiments with other tree-based models, including Random Forest, LightGBM and CatBoost; and because of its use in recent snow studies (e.g., Dunmire et al. (2024) or Goodarzi et al. (2025))."

Comment 3
The inclusion in the ML model of physically-based model simulations with meteorological forcing yields a comparable accuracy than using the meteorological forcings directly as input to the ML model. As I understand it, there is therefore no advantage of using physically-based model simulations, as this adds unnecessary complexity. It seems the ML models learns the physics already with the meteorological forcing. I think this is an interesting finding and should be better discussed.

This is indeed an interesting finding, which is now better discussed in the revised manuscript. With respect to the nested CV section (Section "4.3: Added value of meteorological forcings and Snowclim SD estimates"), we updated the discussion as follows:

L466-478 "Overall, the *Snowclim*$_{\mathrm{ML}}$ configuration yields the best performance (Fig. 6a and b). Under the spatio-temporal nested CV framework, site MAE during snow-covered periods decreases by at least 2.5 cm at 55% of sites, while only 16% exhibit an increase of 2.5 cm or more, relative to the Combination$_{\mathrm{ML}}$ configuration. Compared to the Weather$_{\mathrm{ML}}$ configuration, the difference in site MAE is minimal, and only significant for the spatio-temporal framework. However, the *Snowclim* SD estimates appear to more accurately represent conditions at low-elevation sites ($<$ 1000 m) during the early snow season (Fig. 6c), highlighting the importance of reliable FSC information to correctly indicate snow-free conditions when relying solely on meteorological forcing data. In addition, while the Weather$_{\mathrm{ML}}$ configuration displays the lowest bias of the three configurations at medium-elevation (1000-2500 m) sites (Fig.6c), the *Snowclim*$_{\mathrm{ML}}$ configuration still performs best at high-elevation sites ($>$ 2500 m) during peak snow depth and the ablation period. Nonetheless, both the Weather$_{\mathrm{ML}}$ and *Snowclim*$_{\mathrm{ML}}$ configurations seem to underestimate SD near March-May, with the differences being less pronounced for the 1500-2500 m elevation class. As such, the results indicate that directly using meteorological forcings can achieve comparable predictive performance within the applied XGBoost setup. This would eliminate the need to run a snow model, which reduces both computational cost and model complexity."

Next, we also further discussed this interesting finding within section "4.4 Spatial snow depth prediction" as follows:

L506-512 "Across the nine snow surveys, the inclusion of spatially distributed SD training data primarily improves predictions for lower observed SD values (Fig. 7 and E3e–g), whereas differences for higher SD observations ($\geq$2.5 m) remain less pronounced, albeit best captured by the *Snowclim*$_{\mathrm{ML}}$ configuration. Despite a persistent positive bias across all configurations, the Weather$_{\mathrm{ML}}$ configuration exhibits a bias reduction of approximately 10 cm relative to the *Snowclim*$_{\mathrm{ML}}$ configuration, suggesting that explicitly running *Snowclim* may not be required when the primary objective is to obtain accurate SD estimates across the European Alps. Indeed, XGBoost effectively learns the relationship between the meteorological forcing data and observed SD, without explicitly representing the physical processes governing snowpack evolution."

Nonetheless, *Snowclim* remains a lightweight snow model with a small computational cost. In addition, we do not fully agree that XGBoost learns the physics. Rather, the ML models can learn the relationship between the meteorological forcing data and observed SD well, which is most of interest when the purpose is to deliver the SD estimates as a product. However, the ML models do not consider nor learn physical processes occurring in the snowpack. This we also added in the revised text (see final sentence in the blue text above).

Comment 4
Section 4.1 states a couple of times that differences are significant because p $<<$ 0.05 (e.g. differences due to Table C1). However, the improvements are quite marginal (R2 0.88 vs R2 0.89; MAE 0.3m vs MAE 0.29 m). I think the authors should discuss the significance based on the ab-

solute improvement, which is very little, and not the statistical significance, which in this case is clearly just due to the large sample size. See https://www.nature.com/articles/s41598-021-00199-5 and https://linkinghub.elsevier.com/retrieve/pii/S026151771730078X . With this in mind, the authors should revise this section carefully.

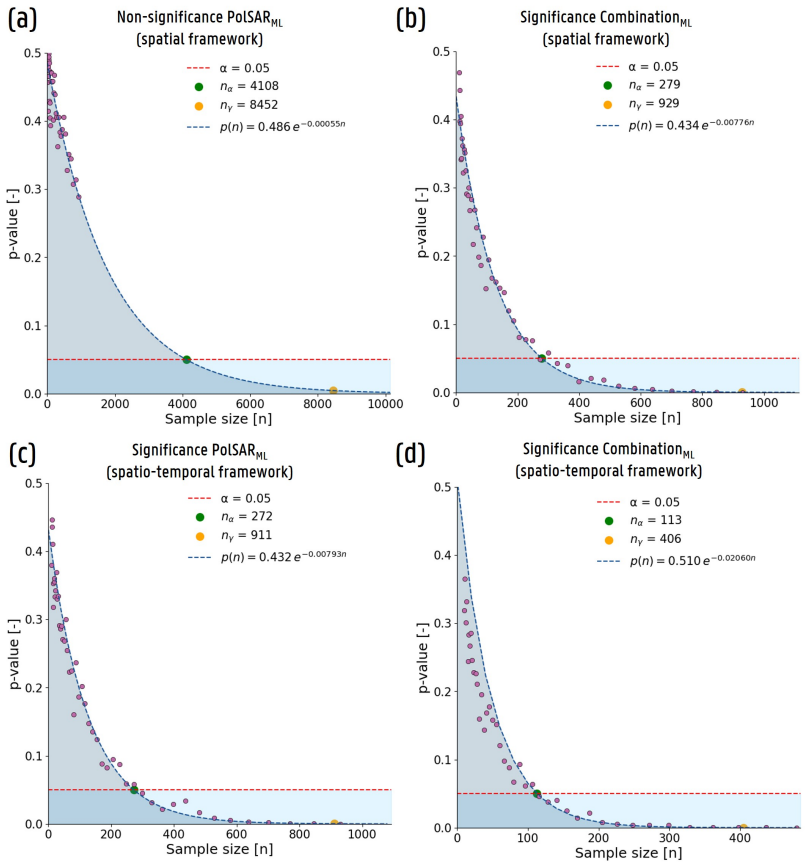
We acknowledge that the improvements are at a rather small magnitude. This is now more stressed within the manuscript as well, whereas before it was stated mainly in the abstract and conclusion. As an example, we rewrote part of section "4.1 Performance of ML configurations with S1 PolSAR and backscatter intensity variables" as follows:

L370-382 "The absolute improvements of using PolSAR observations, as a replacement for (or in combination with) backscatter intensities, are small within all nested CV frameworks, or even minimal within some CV frameworks. Within the spatial framework, the $\text{PolSAR}_{\text{ML}}$ and $\text{Backscatter}_{\text{ML}}$ configurations display marginal differences in performance metrics (Table 3), which likely stems from the similar relationship of the variables with SD, and the spatial noise present in both types of S1 variables. The $\text{Combination}_{\text{ML}}$ configuration also shows similar overall performance compared to the $\text{Backscatter}_{\text{ML}}$ configuration, yet the improvements in MAE, computed per site, are significant ($p \ll 0.05$, 95% confidence level), resulting in an overall MAE of 35 cm, a mean site MAE of 29 cm, and a mean site bias of 11 mm (excluding zero-measured SD). PolSAR observations moreover improve model performance for the spatio-temporal framework, with significant improvements in site MAE for both the $\text{PolSAR}_{\text{ML}}$ and $\text{Combination}_{\text{ML}}$ configurations ($p \ll 0.05$, 95% confidence interval). This improvement suggests that PolSAR observations, alone or in combination with backscatter intensities, more effectively capture the seasonal evolution of SD, a conclusion that is also supported by statistically significant gains ($p \ll 0.05$, 95% confidence level) observed in the temporal nested CV framework. Nonetheless, also here the absolute improvements remain limited, and the increased computational cost of processing S1 SLC data into PolSAR variables relative to deriving backscatter intensity from GRD data must be considered."

Nevertheless, we still think it is useful to include the statistical difference. To investigate the effect of sample size, we consulted the paper of Gómez-de Mariscal et al. (2021), and calculated the exponential decay of the p-values for our experiments (Review Fig. 2). The values n_α and n_γ in Review Fig. 2 indicate the sample sizes where the significance level is reached and where 'the p-value does not change anymore', respectively. However, n_γ is based on an arbitrary chosen p-value γ, and this γ strongly influences the binary θ value used by Gómez-de Mariscal et al. (2021) to decide whether the differences are truly significant, independent of the sample size. What is evident from Review Fig. 2 (with $\gamma = 2.5e^{-6}$) though, is the following: For the spatial nested CV framework comparing the $\text{PolSAR}_{\text{ML}}$ with $\text{Backscatter}_{\text{ML}}$ configuration, n_α is $\gg$ the amount of sites used within the comparison (1022; Review Fig. 2a). This supports the non-significance reported in the manuscript. When comparing the $\text{Combination}_{\text{ML}}$ with the $\text{Backscatter}_{\text{ML}}$ configuration, however, n_γ is 929, smaller than our sampling size, but n_α is 272, indicating that the statistical significance would have been reached if only 30% of the sample size (i.e., the amount of measurement sites) would have been used (Review Fig. 2b). Similar results are obtained for the spatio-temporal framework (Review Fig. 2c and d), supporting our conclusion.

Comment 5
There are several instances of "results not shown" in the paper. I think they should all be included as they seem relevant (lines 396, 411, 465, 474, 484, 494, 501, 526, 550). There are also instances where a result is discussed but not seen on any Figure (lines 397-399, 431-433, 497-498, 511-512, 534-537, 586)

This issue has been addressed in the revised manuscript, and all instances of "not shown" have been removed either by the inclusion of additional Figures, discussion, or text revisions. As such, several Figures have been revised or added to the Appendix (e.g., Fig. 6 and Fig. 3 of the revised



Review Figure 2: Exponential p-value decay with sample size, indicating non-significance between the PolSAR$_{\mathrm{ML}}$ and Backscatter$_{\mathrm{ML}}$ configurations within the spatial nested CV framework, and significance between the Combination$_{\mathrm{ML}}$ and Backscatter$_{\mathrm{ML}}$ configurations for both the spatial and spatio-temporal nested CV frameworks. (a) and (b) Comparing the PolSAR$_{\mathrm{ML}}$ and Combination$_{\mathrm{ML}}$ configurations with the Backscatter$_{\mathrm{ML}}$ configuration within the spatial nested CV framework, respectively. (c) and (d) Same as (a) and (b), but for the spatio-temporal framework. A value of $\gamma = 2.5e^{-6}$ is used to determine n_γ, while 100 samples are used within the bootstrap procedure to fit the p(n)-curve. The mean p-values obtained from the bootstrap procedure for the different sample sizes are shown in purple.

manuscript to backup lines 497-501, 510-512 or 586 of the original manuscript), while more detailed discussions, previously not supported by Figures, have been relocated to the Appendix, where they are now supported by new Figures (e.g., Fig. C2, Fig. E2, Fig. D2a, and Fig. D1 to backup lines 484-485, 494, 535-537, and 522-524).

In addition, some parts of the text have been revised to provide a more direct link with the corresponding Figure. As an example, we revised line 397-399 to:

L416-420 "Excluding the non-valid areas increases the overall and dry-period FI for both α^{s} and $\gamma^{0,\mathrm{s}}_{\mathrm{CR}}$ within the PolSAR$_{\mathrm{ML}}$ and Backscatter$_{\mathrm{ML}}$ configuration, respectively. As these non-valid areas include low-elevation areas (i.e., valleys) and densely-forested regions, locations where typically low SD values are observed, these results suggests that XGBoost places relatively more importance on these S1 variables at bare, higher-elevation locations that are often characterized by deeper snowpacks."

Comment 6
About the novelty with respect to Dunmire et al (2024). Line 344 even states that the errors are slightly higher in this study than in Dunmire, and another example in line 434 shows very similar results. There should be a more open discussion about the little improvement, despite the novelty of this paper.

The intention of this paper was not to compare our results directly to Dunmire et al. (2024). Rather, the purpose of this manuscript was to assess different input feature configurations within an XGBoost ML setup, to see 1) if PolSAR variables improve performance compared to backscatter intensity ones, 2) how much improvement can be achieved by including either meteorological forcing data or *Snowclim* SD estimates (with smaller side questions as to how the latter two compare and whether S1 observations are still be influential), and 3) to determine whether training XGBoost with spatially distributed SD training data is necessary for a good representation of topographical impacts.

In the original manuscript, a comparison with Dunmire et al. (2024) was included to validate the results with previous studies. It is true that the results within our manuscript perform slightly worse, but this now has been addressed in Section "4.4. Spatial snow depth prediction" of the revised manuscript:

L494-497 "Despite the improvements, SD remains overestimated for the 16 March 2017 snow survey across all configurations. This overestimation is most pronounced for the Combination$_{\mathrm{ML}}$ configuration and exceeds the bias reported by Dunmire et al. (2024). Although it is not the objective of this study to perform a direct comparison, the persistence of the Combination$_{\mathrm{ML}}$ bias indicates potential limitations in the current feature selection or training dataset."

However, improvements are made upon the results of Dunmire et al. (2024) with the Weather$_{\mathrm{ML}}$ and *Snowclim*$_{\mathrm{ML}}$ configurations when comparing mean SD for this survey date. This is now also added as follows:

L497-503 "Nonetheless, both the Weather$_{\mathrm{ML}}$ and *Snowclim*$_{\mathrm{ML}}$ configurations outperform the results reported by Dunmire et al. (2024). For the 9 March 2016 and 16 March 2017 snow surveys used by Dunmire et al. (2024) for model validation, the Weather$_{\mathrm{ML}}$ and *Snowclim*$_{\mathrm{ML}}$ configurations achieve higher correlation coefficients (0.68 and 0.64) and lower mean absolute errors (31 and 36 cm, respectively) than those reported by Dunmire et al. (2024) (0.56 and 41 cm). For the 16 March 2017 survey specifically, Dunmire et al. (2024) reported a mean SD difference (predicted minus observed) of 16 cm, whereas the Weather$_{\mathrm{ML}}$ and *Snowclim*$_{\mathrm{ML}}$ configurations exhibit differences of -5 and 14 cm, respectively."

Referring to line 434, the original manuscript states that within forests, the contribution of S1 is limited, as was found by Dunmire et al. (2024). We acknowledge this reviewer's concerns about the marginal improvement made in our paper. However, referring to this line, not much more information can be learned within any ML model, as the relationship between S1 and SD over forest-dominated areas remains poor. For these areas, improvements are expected with the inclusion of either the meteorological forcings or *Snowclim* model simulations. As an example, we included the SD estimates from the Weather$_{\mathrm{ML}}$ configuration for the time series of the measurement site in Sobrio, Switzerland (Fig. C2d in the revised manuscript).

Comment 7
Sometimes in the manuscript, it seems that meteorological forcing data AND physically-based model simulations are used simultaneously in the ML model, but that is not the case. I suggest to change the following to OR (not in the title, as that is a list of all the inputs). (Line 8, 568)

We made sure there is no confusion in the revised manuscript.

Comments by line number

L34: a reference for "essential climate variables" is missing

The reference for essential climate variables is given by Gascoin et al. (2024). We changed the sentence as follows:

L34-37 "As a result of its importance, key snowpack properties such as SD and snow water equivalent (SWE), the latter relating to SD through snow bulk density, have been designated as essential climate variables (Gascoin et al., 2024; Global Climate Observing System), and various scientific institutions and international organizations have prioritized their enhanced observation and monitoring (Dozier et al., 2016; World Meteorological Organization, 2023)."

L40: I suggest to add example datasets as example for local point measurements

We added Fontrodona-Bach et al. (2023) and Matiu et al. (2021) as two reference datasets of point measurements. We adjusted the text as follows:

L41-44 "Manual and automated point measurements offer frequent data at many locations globally (e.g., Fontrodona-Bach et al. (2023) or Matiu et al. (2021)), but fall short in capturing the snowpack's spatial variability due to their relatively sparse and uneven spatial distribution within alpine regions (Miller et al., 2022; López-Moreno et al., 2015). "

L53: Does "this work" refer to the one in this manuscript? Not clear if it refers to the previous references.

No, "this work" refers to the studies of Lievens et al. (2019) and Brangers et al. (2024), not to this manuscript itself. We changed "this work" to "These previous studies", making the link more clear as follows:

L65-68 "This algorithm has been developed based on the results of Lievens et al. (2019) and Brangers et al. (2024), who demonstrated the sensitivity of active microwave observations to snow at C-band. These previous studies have focused on C-band backscatter intensity..."

L54: "an increasing snowpack DEPTH"?

We added 'depth' to make it more clear.

L55: I recommend against the use of etc. Either complete the list or simply state the examples.

We removed 'etc.'.

Lines 62-64 and 75-77 seem to be a repetition of each other regarding the current gap in knowledge.

We removed this sentence in the rephrasing of the introduction to avoid repetition.

L74: perhaps: "snow depth retrieval"?

We prefer to retain the original phrasing of this sentence, as the cited studies encompass both SD and SWE retrieval. Nonetheless, the sentence has been relocated within the introduction to improve flow and readability.

L83: "compared to in-situ measurements." This needs references.

In order to support our sentence that 'uncertainties in the forcing data and/or limitations in model representation can cause the model results to differ from in situ measurements', we added the following three references (L61 in revised manuscript):
Terzago et al. (2020) (DOI: https://doi.org/10.5194/hess-24-4061-2020),
Largeron et al. (2020) (DOI: https://doi.org/10.3389/feart.2020.00325), and
Ryken et al. (2020) (DOI: https://doi.org/10.1016/j.advwatres.2019.103473).

L85: such as instead of e.g.

Addressed.

L88: This needs a reference at the end of the sentence.

We added an example of an emulator that models mountain SD and SWE (Charbonneau et al. (2025)). This paper uses physics-constrained ML to estimate SD and SWE at SNOTEL sites, and holds potential in climate modeling.

L91: perhaps: "contribute to improving SD predictions"?

As the satellite variables do not always improve SD estimates, we will not take this suggestion into account.

L103: coarse instead of course.

Addressed.

L116: The GHC needs a reference (and isn't It GHCN?). Does the end of the sentence mean only Germany and Slovenia are taken from this dataset?

You are correct that the proper abbreviation of the Global Historical Climatology Network should be GHCNd instead of GHC. We changed the abbreviation after consulting the paper of Fontrodona-Bach et al. (2023) and the paper of Menne et al. (2012). The end of the sentence does not indicate that only in Germany and Slovenia sites from these providers appear. However, in the other countries the amount of sites/measurements originating from Synops or GHCnd is small or even minimal. To make it more clear, we revised the text as follows:

L122-125 "In addition, we augmented our SD dataset with measurement data sourced from the Synoptic Data platform and the National Oceanic and the Atmospheric Administration's Global Historical Climatology Network-Daily database (GHCNd, Menne et al. (2012)). With the addition of these data, our dataset also contained sites within Germany and Slovenia."

L145: Why does rescaling matter for interannual start of season differences? It is unclear what this means

Our original manuscript outlines a two-step rescaling procedure of the S1 PolSAR and backscatter intensity observations. First, we applied a scaling of the first and second-order moment to reduce the

inter-orbital differences. Second, for each location and snow year, we subtract the mean summer value from the corrected time series. This step aligns the baseline conditions at the onset of the snow season across sites and years. Without this adjustment, the S1 variables may exhibit multi-year trends (e.g. a gradual increase or decrease across the years), which would constrain the use of these non-corrected values in our experiments.

We originally wrote these summer-adjustments as being part of the scaling procedure. We revised line 144-148 of the original manuscript to make it more clear:

L153-160 "Finally, both the PolSAR and backscatter intensity observations were rescaled and summer-corrected, to reduce inter-orbital and interannual start-of-season differences, respectively. To this end, we first reprojected the satellite observations onto the 100 m WGS84 grid using linear averaging, and resampled to the 500 m WGS84 grid by taking the mean. Then, we performed a first and second-order moment scaling (to correct for differences in the mean and variance) between orbits. Finally, from each scaled time series for each snow season (September-June), we subtracted the mean summer value (July-September) of the summer preceding that snow season, thereby reducing interannual start-of-season differences that are not related to snow conditions (e.g., changing vegetation). To indicate that these variables have been rescaled, we further append a superscript "s" (^{s}) to the satellite variable notation."

L158: How many are these remaining gaps? How many were filled?

To answer this question, we split up the necessary analysis between the dataset used for cross-validation, and the FSC data across the whole Alps and time period. For the data used within the cross-validation (i.e., the measurement data from the sites in Fig. 1 of the original manuscript), 5.1% is gap-filled with IMS binary snow cover information. Across the whole study area and time period (2015-2024), 3.95% is IMS gap-filled.

To address this comment, we added a small statement in the revised text referring to the percentage of data that has been filled in such manner.

L169-171 "Finally, remaining gaps (3.95% of the data points across the whole Alps and study period) were filled using IMS' binary snow cover information, reprojected to the same 500 m WGS84 grid using majority resampling... "

L166: a quick definition of majority resampling would be useful.

This comment has been addressed in the revised text as follows:

L176-179 "The data, available at 60 m resolution from September 2016 onward, were first reprojected onto the 100 m WGS84 grid using majority resampling, whereby the most frequently occurring SWS value within each 100 m target pixel was selected while excluding any no-data values..."

L178: What other downscaling techniques?

In L178 of the original manuscript, 'with other downscaling techniques' refers to the downscaling techniques that are explained in the following sentences (L179-193). In order to avoid confusion, we have changed this part of the sentence to:

L190-191 "Subsequently, we applied bilinear interpolation in combination with different downscaling techniques, explained below, to account for local terrain features."

Equation 1: Where does this downscaling equation come from? A reference or explanation is needed.

Equation 1 of the original manuscript is an adaptation from Equation 2 in the study of Huss et al. (2013), who used meteorological forcing data to model accumulation and melt for two glaciers in Switzerland. Within the study, the authors corrected observed (in situ) precipitation data for gauge undercatch errors using a scaling factor c_{prep} and an altitudinal gradient. Within our study, however, we did not address gauge undercatch. Instead, we adapted equation 2 of Huss et al. (2013) to downscale coarse-scale precipitation data to account for orographic precipitation. To this end, we modified this equation to 1) focus on areas with major elevation differences (by introducing D_{dif} and limiting D to 1, so that Equation 1 within mountainous areas becomes $P_{\text{coarse}_{x,t}} \cdot \left[0.75 + 0.5\frac{z_x - z_{\min}}{z_{\max} - z_{\min}}\right]$) and 2) by limiting the precipitation adjustments to 75-125% of the non-downscaled values. The original manuscript has been adapted as follows:

L192-201 "First, coarse-scale precipitation (P_{coarse}; [mm $3h^{-1}$]) was corrected as a function of elevation to account for orographic effects, using a rescaling function adapted from Huss et al. (2013), who corrected in situ precipitation data for gauge undercatch in a glacier mass-balance study. Thus, for each location x within the 500 m grid at time step t, $P_{\text{coarse}_{x,t}}$ was downscaled using the following equation:

$$P_{x,t} = (1 - D) \cdot P_{\text{coarse}_{x,t}} + D \cdot P_{\text{coarse}_{x,t}} \cdot \left[0.75 + 0.5\frac{z_x - z_{\min}}{z_{\max} - z_{\min}}\right] \text{ with } D = \frac{z_{\max} - z_{\min}}{D_{\text{dif}}} \quad 0 \leq D \leq 1 \tag{1}$$

with z_x the elevation [m] of the 500 m Copernicus GLO-30 DEM, $z_{\min}$ and $z_{\max}$ the minimum and maximum elevation [m] within an interpolation window — centered on the location x and spanning an area roughly matching the original 0.1° grid size — and D_{dif} a user defined difference in elevation, set to 250 m. The user defined difference was introduced to focus the corrections on the study area, with minor adjustments for areas with small elevation differences. Different from Huss et al. (2013), Equation 1 limits the downscaled precipitation values between 75 and 125% of the original $P_{\text{coarse}_{x,t}}$ values."

L223: With a rather long paper and a lot of specific nomenclature, it is sometimes easy to forget what LIA, or TPI mean, especially for unspecialised readers. I suggest to include a table or list in the Appendix with all abbreviations used (or expand Table B1)

We did not include an additional table in the Appendix, but adjusted Table 2 of the original manuscript that now includes a description of both the abbreviations and full names of the different features.

L239: Perhaps it is useful to remind the reader that here the input of meteorological data or physical model simulations is still not assessed. I thought there should be 5 configurations otherwise.

We adressed this by making the following change:

L263-265 "To assess the performance of ML setups using S1 PolSAR versus backscatter intensity observations, XGBoost was trained and validated under three distinct configurations (Table 2), without including meteorological forcing data or *Snowclim* SD estimates in these experiments."

L241: Why "next" and not together with the previous?

In line 240 and 241, it has been written that the three conducted experiments to compare S1 PolSAR vs. backscatter intensity observations include the same time-independent auxiliary data. We

use "next" in the following sentence to make the distinction with FSC, DOY and LIA data, that do depend on time and location.

L243: I suggest "The second configuration, focusing..." instead of "Conversely, the configuration..."

We took this suggestion into account in the revised manuscript.

Table 2 caption: I suggest "within this study."

Addressed in the revisions.

Table 2: some of the features are presented in the text after the table is presented, therefore the reader does not know what all these variables are. I suggest to spell them out in the caption, or in the acronym table I suggested above.

Thank you for pointing this out. As discussed above, we changed the caption of Table 2, in which abbreviations are spelled out.

3.2.3 Snow depth prediction: I do not understand how this title links to the paragraph. It seems that the paragraph is about standardization of features.

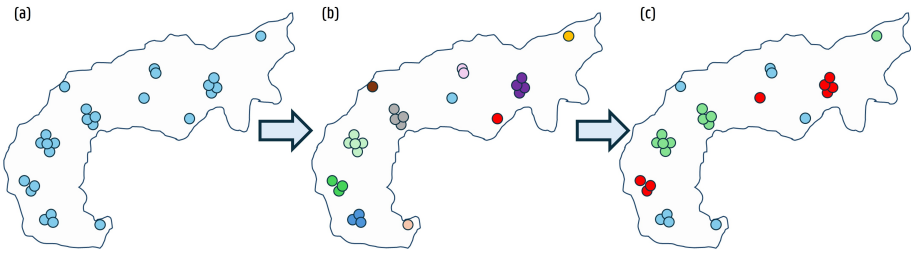
We chose the title *Snow depth prediction* as it describes how the features are applied within XGBoost to output snow depths. Indeed, the first two sentences describe the standardization procedure applied to the features, before inputting in the model. The last two sentences however, give important information to the reader about the post-processing of the predictions. We explicitly explain that ascending and descending orbits are treated separately, and that the average of the SD outputs is taken in case a location has both an ascending and descending S1 observation on the same date.

We will adjust the title to "XGBoost snow depth prediction procedure", to provide a clearer link with the applied steps.

L276: If all folds contain at least one station from each of all the boxes of stations within 5 km of each other, how is this a blind validation? I am possibly understanding this wrong, please clarify this.

Review Fig. 3 gives a schematic overview on the fold creation procedure for a dummy set of measurement sites across the European Alps, to better understand the blind validation and testing. The Figure explains the two-step procedure, in which first sites close to each other are added to clusters, followed by a shuffling of the clusters (not the sites) within folds. As a result, each fold has a subset of the measurement sites (within the spatial and spatiotemporal setups at least), which allows the nested CV to assess model evaluation at unseen data points in space. In nested CV, namely, 1 fold is always excluded as test set, while the remaining ones are used to train and tune the ML model. Through looping over the folds, every site will have been evaluated at the end, but always with a model that has not seen the site during model training. We revised L275-280 of the original manuscript to:

L300-307 "The spatial folds were constructed using a two-step approach: first, sites located within five km of one another were grouped into clusters, thereby preventing nearby sites with similar (climatic) characteristics and SD patterns from being split across training and test sets. Subsequently, these clusters were randomly assigned to five unique folds, ensuring that each fold contained a comparable amount of data and preserved a similar SD distribution. For the temporal folds, sites were not clustered; instead, all observations from a given snow season (September–June) and from across the study area were grouped into nine blocks (corresponding with the number of snow seasons for which SD data is available), which were then partitioned into five folds, again ensuring comparable fold sizes and a



Review Figure 3: Fold creation procedure for a dummy dataset of measurement sites across the European Alps. (a) original geographical distribution of a set of manual and/or automated measurement sites. (b) creation of clusters of sites within 5 km of each other. Note that clusters may consist out of individual sites. The clusters are displayed with different colors (c) Final fold distribution (three folds colored by blue, green and red), in which the sites contained within the same cluster, are added to the same fold.

consistent SD distribution. Finally, by combining both fold-creation techniques, we constructed 25 unique spatio-temporal folds."

L278: The procedure for the temporal fold is also not entire clear to me. What does it mean that sites were kept separate, but grouped and divided in 5 folds?

With this sentence, we imply that there is no creation of clusters that are shuffled into folds. Instead, the different snow seasons are split into nine blocks (as nine snow seasons are included within the study period 2015-2024), and those blocks are partitioned over 5 unique folds. See the revised text as our answer to the previous comment.

Figure 2: Please a add a legend for the colours and textures.

Included in revised Figure.

L308: Does this mean that for these sites, the snow season is less than 10 days?

Not necessarily. L308 of the original manuscript states that if for a certain measurement site and year, there are no 9 days with a SD $\geq$ 1 cm following a snow event with $>$ 10 cm, the first snow event with $>$ 10 cm of snow marks the beginning of the snow season. Note that for marking the end of the snow season of such sites, we loop back in time to find the latest '10-day 0 cm SD period', to not miss the actual time with snowpack presence. As such, for these sites it is possible that one big snowfall event was followed by a longer period with no snow, and then again by snow covered periods. But the no-snow instances in the early winter season months, as well as the (late) spring and summer months, are mostly filtered out using this procedure.

L314: The bias, although discussed, is not always shown in Figures or tables. Please include it (e.g. Table 3, Table C1).

We have included the bias metric in the revised Tables and Figures.

L332: Why is Table C1 not together with Table 3? As it seems quite important and thoroughly discussed.

Within the manuscript, the main focus is on the spatial and spatio-temporal frameworks, as these

give a more accurate representation of model performance for the XGBoost usage within the study: predicting SD across the Alps at many unseen sites for the model. Predicting at known locations within a different snow year is easier, as the model can rely on recurrent snow patterns, and can learn the mapping of the input features to the observed SD. This mapping will be similar in unseen years. As an illustrative example, consider a site at 2000 m elevation with a TPI of 10, an aspect of 187°, and a slope of 10°, where peak SD values of approximately 3 m are consistently associated with α^s values of about 10° across multiple years. In such a case, XGBoost can readily predict an SD of ∼3 m for an unseen time period when α^s is again close to 10°. However, if at an unseen test site with similar topographic characteristics but different SD dynamics an α^s value of ∼10° corresponds to an SD of ∼4 m, XGBoost may underestimate SD. Likewise, even when seasonal SD patterns and most topographic attributes are comparable, subtle differences such as a slightly more south-facing aspect may induce earlier surface wetting and associated signal attenuation at the test site. In this case, reduced α^s values at the test site, not yet observed at the training site during similar time periods, may again lead to SD underestimation.

As such, the temporal nested CV gives an overly optimistic view, as we will not deploy XGBoost to just predict at the ∼1000 measurement sites used within this study. Rather, the goal of our study is to indicate how good XGBoost will perform if someone uses our ML setup to predict SD at unseen locations across the whole Alps. Therefore, we prefer to have the table in the Appendix as additional information on how the model would perform in this setting.

L336: "the temporal framework overestimates model performance" what does this mean? Can model performance be overestimated or is snow depth overestimated?

The sentence explains that temporal nested cross-validation gives an overly optimistic view on how XGBoost would perform in predicting SD across the European Alps, given the training dataset that is available within our study (see comment above). Model performance (i.e. R, MAE, RMSE, bias...) is thus too high on unseen data. As explained above, predicting SD in unobserved snow seasons is easier for the model, as snow patterns generally recur, especially at higher elevations. This is why the main focus is on the two other types of nested CV experiments, as this is the setting in which our model will be deployed in reality as well. The revised manuscript gives a better description as follows:

L359-369 "In general, all configurations (i.e., $\text{PolSAR}_{\text{ML}}$, $\text{Backscatter}_{\text{ML}}$ and $\text{Combination}_{\text{ML}}$) achieve best results within the temporal nested CV framework (Table E1), and performance progressively deteriorates in the spatial and spatio-temporal frameworks (Table 3. This behavior is expected, as SD patterns within the same area (i.e., at the same sites) tend to recur across different years, such that SD prediction for an unseen snow season (i.e., the temporal framework) is inherently more favorable for XGBoost than prediction at unobserved locations (i.e., the spatial framework). Predicting at unobserved locations and during unseen snow seasons (i.e., the spatio-temporal framework), including times outside the study period is even more challenging, as XGBoost cannot exploit season-specific information from the training data nor rely on recurrent site-specific SD patterns. Consequently, because the intended application of XGBoost is to predict SD at previously unseen locations across the European Alps within the training dataset time period (e.g., Fig. 3a), reliance on the temporal framework would lead to an overly optimistic evaluation of model performance, whereas the spatio-temporal framework would provide a more conservative assessment."

L339: This says that the spatio-temporal framework provide a more realistic evaluation of model performance for this study, but lines 331-332 say that performance is highest for the temporal framework and progressively deteriorates in the spatial and spatio-temporal frameworks. These two statements contradict each other.

The framework yielding the highest model performance does not necessarily provide the most

realistic evaluation, as this depends on the intended application. We refer to L359-369 of the revised manuscript (see added text above), from which it should be clear that the statements do not contradict each other. The spatio-temporal framework, however, is less representative of the intended deployment scenario of our ML setup. As discussed above, the spatial framework provides the best match.

L349: in the Figure c1b caption it says observed-predicted, so a negative bias would mean an overestimation of snow depth. Please standardise this.

Thank you for notifying this mistake. As written in L315 of the original manuscript, bias within our study is defined as predicted minus observed. We will check this for the other sections as well, to have it standardized.

L353-355: How is a deterioration of model performance seen as an accurate representation of model performance? This sentence is unclear.

This again relates to our replies to the reviewer's comments above. The sentence however does not link the deterioration specifically with accurate model representation. The sentence wants to make clear that if you would apply XGBoost at unseen locations, outside the time period for which training data is available, the results for the spatio-temporal nested CV framework give you a more accurate indication of model performance, rather than looking at the model performance for the spatial or temporal frameworks. This comment has been addressed in the revised text in L359-369 (see our reply to the reviewer's comment on L336 of the original manuscript).

L355: why "also"? Which other improvements were there?

We used "also" to indicate that PolSAR observations improved the model outcomes for the spatio-temporal framework as well. The sentence is changed in the revised section in the new manuscript.

L355-365: As stated in a general comment, I don't see this little improvement as a significant improvement, I think it is the effect of the sample size.

To answer this question, we refer to our answer on main comment 5.

L368: FSC instead of fSC.

Thank you for noticing! Addressed.

Figure 3: It is difficult to see differences between configurations, perhaps a table in the supplement or Appendix would be useful.

The differences between the configurations are sometimes small, but we think the Figure does allow to see the relative order of the input features well.

L397-399: How are these results seen in Fig. 4a?

Line 397-399 state that "the SHAP value FI suggests that at bare, higher-elevation locations with deeper snowpacks, XGBoost places relatively more importance on α^{s}, a pattern also observed for $\gamma^{0,s}_{CR}$ in the Backscatter_{ML} configuration.". This can be seen in Fig. 4a of the original manuscript, by the decrease in SHAP FI α^{s} and $\gamma^{0,s}_{CR}$ when including non-masked areas (i.e., comparing the fully colored bars with the gray line). As the SHAP value FI is computed by taking the mean of absolute SHAP values, this indicates that at the low-elevation and/or densely-forested areas, the SHAP contributions from S1 are smaller, in favor of other features. To make it more clear, we revised the text to:

L414-421 "Figure 5a presents the FI for both the full set of predictions (fully colored bars), and the subset remaining after excluding the non-valid areas masked within the wet snow product (dark gray lines). Excluding the non-valid areas increases the overall and dry-period FI for both α^s and $\gamma^{0,s}_{CR}$ within the PolSAR$_{ML}$ and Backscatter$_{ML}$ configuration, respectively. As these non-valid areas include low-elevation areas (i.e., valleys) and densely-forested regions, locations where typically low SD values are observed, these results suggests that XGBoost places relatively more importance on these S1 variables at bare, higher-elevation locations that are often characterized by deeper snowpacks. Conversely, the decrease in SHAP value FI when including these non-valid areas indicates the limited added value of the S1 variables to predict SD at low-elevation and/or densely-forested sites."

Figure 5: The predicted snow depth time series show a clear flat long period in the middle of the accumulation period (especial at 5a and 5b), which does not match observations very well, suggesting that snowfalls and increasing snow depth in mid-winter are not well captured? This should be discussed. Also please state which sites are these (name, location, source of the data).

Addressing the first part of the remark, this indeed seems to suggest that snowfall and increasing SD in mid-winter are not captured well. However, we think this does not necessary indicate that it is not captured. Instead, we believe that this phenomenon arises from the very limited amount of high SD observations ($\geq$ 2.5 m) within the training dataset (see Fig. E1a in the revised manuscript). We included your remark in the revised discussion as follows:

L631-636 "Unlike the Arolla site, however, α^s is not only mainly informative during snow accumulation, but remains relatively elevated throughout February and March for both snow seasons, when the snowpack exceeds 2 m of snow. Nonetheless, the SD predictions appear to level off with only a limited response to new snowfall events, which is observed at the Arolla site during the 2017-2018 snow season as well. One explanation lies in the limited number of high SD observations used to train XGBoost (Fig. E1a), which constrains the model's ability to represent higher SD values and results in a flattening of the predicted SD range (Fig. 6a)."

Next, within the original manuscript it is stated that 'Fig. 5 displays the results for two snow seasons at three distinct sites in Switzerland, all operated by SLF'. We added this information to the caption of the revised Figure, as well as the coordinates and location names for each site of the displayed time series.

L505-507: Linking to one of my main comments, I think the results underscore the potential of using meteorological forcing data alone, as input to ML models (as the improvement of Snowclim is minimal). I think this should be included here.

See our responses to main comment 3.

Figure 6. I suggest adding the title of each configuration on each row, to make the Figure more easily readable.

We revised Figure 6 of the original manuscript. While doing so, we included this reviewer's advice and added the names of the configurations.

L531: again, the improvement seems quite minimal.

Overall, yes the improvement of *Snowclim* SD estimates might seem minimal, but for high observed SD, there is a clear improvement. Across the nine surveys, the *Snowclim*$_{ML}$ configuration displays a 20 cm less underestimation (-64 vs. -84 cm) for measured SD $\geq$2.5 m.

542-543: the potential inability to correctly predict snow density is a key limitation for further refining this method to predict daily time series in the future. This could be discussed.

We agree and added the following in the revised manuscript:

L517-518 "Further research should attempt to include factors or use ML models that govern variations in snow density, especially to enhance the potential of predicting short-term variations in SD."

L539: The authors say weatherML and snowclimML overestimate snow depth, based on the biases. However, when comparing Figure 7b with the measured Figure 7d, it seems the opposity. It seems that 7b (weatherML) shows much lower snow depths than measured. In fact, the scatter plot suggests that weatherML outperforms snowclimML, but the snowclimML snow depth map resembles the observations more. This discrepancy should be clarified.

Referring to the first comment: "weatherML and snowclimML overestimate snow depth", especially the $Snowclim_{\mathrm{ML}}$ configuration displays a positive bias for the 16 March 2016 snow survey. The $Snowclim_{\mathrm{ML}}$ configuration overestimates SD for all elevation bins > approximately 1500 m, and this is most evident for the areas in the (south)western part of the Dischma valley. As a result, the $\mathrm{Weather}_{\mathrm{ML}}$ configuration shows a near-zero bias between 1500-2500 m, differently from the $Snowclim_{\mathrm{ML}}$ configuration that displays a consistent positive bias (Fig. D2a in the revised manuscript). Comparing the results in terms of the observed SD values (i.e., not looking at the spatial distribution of the observations nor estimates), the $Snowclim_{\mathrm{ML}}$ configuration displays less underestimation (bias of -1.16 m for observations $\geq$2.5 m, compared to -1.34 for the $\mathrm{Weather}_{\mathrm{ML}}$ configuration), but these high SD values occur less often in space, resulting in more overestimation in space (see Fig.D2a). Therefore, we do not fully agree with the reviewer's remark. Instead, we argue that the $\mathrm{Weather}_{\mathrm{ML}}$ resembles the measurements more, except for the high SD estimations (for which both display underestimation). The detailed explanation given above is added to the revised manuscript within a new Appendix discussing the 16 March 2017 snow survey (Appendix D).

It is important to note that the scatterplot shown in Figure 7e,f and g display the results for all the nine snow surveys collected within our study. Therefore, it displays different results compared to the shown SD maps for the survey of 16 March 2017. In the revised text, we addressed this possible confusion in the revised manuscript.

Figure 7. Why do maps have different MAE than their respective scatter plots? Why do the scatter plots have a low density of points when approaching 0 m snow depth?

The scatter plots display the results for the nine collected surveys (as explained in the data section). As stated above, we have now addressed this confusion.

Referring to your second question, the scatter plots indeed display a low amount of zero-measured SD. This results from the time period the surveys are conducted (around peak SD; often near the end of March or the beginning of April), and the limited amount of collected data points in the valley (which mainly situates above an altitude of 1500 m). As such, the majority of the measured (and predicted) SD values are between approximately 0.5 and 2 m.

Figure 8. It would be better to show the maps of snow depth with survey data, without survey data, the difference, and the measured maps, to enable a better comparison.

In the revised Figure, we have added both the maps with and without the inclusion of spatially distributed SD training data. However, it is not possible to include any measured map, as no such data is available over these areas. Nonetheless, we have added the available point measurements taken

on this date as reference data.

L593-595: Compare these results to estimates from other studies, such as the results from Dunmire et al 2024.

We decided to leave out this sentence in the conclusion. To address the reviewer's comment, however, we included a comparison with Dunmire et al. (2024) in Section "4.4 Spatial snow depth prediction". More specifically, we downloaded the predictions made with the final trained XGBoost model of Dunmire et al. (2024) and compared their outcomes with the $\text{Weather}_{\text{ML}}$ and $\textit{Snowclim}_{\text{ML}}$ configurations for the snow surveys conducted on 9 March 2016 and 16 March 2017. We refer to our answer on main comment 6 for the included comparison.

Equation A1: Can Wsat not become infinite if any weight is zero? Revise or clarify.

The equation is formulated such that the denominator in Equation A1 cannot become zero. In Eq. A2, the values range between 1 (QA=0) and 2 (QA=2). In Eq. A3, the square of CP+1 is taken, with values of CP between 0 and infinity. As such, also this equation never becomes zero. Finally, the same applies for Eq. A4. We adjusted the original manuscript to make clear the denominator never becomes zero:

L578-583 "... in which W_{QA} is the weight associated with the quality flag (QA, that ranges between 0 and 2 with 0 indicating the highest quality) of the satellite data; W_{CP} is the weight based on the number of days since the last cloud-free observation (CP; values between 0 and ∞); and W_{DD} is the weight accounting for the difference in days between the most recent cloud-free Terra and Aqua observations (DD; values between 0 and ∞, where 0 corresponds to the satellite with the most recent cloud-free data). The individual weights were computed using the following formulations, which were designed to ensure that none of the weights attain zero:..."

L633-635: what downscaling techniques and what parameters?

The downscaling techniques refer to the type of downscaling for the different meteorological forcings (i.e., different approaches to downscale precipitation, temperature, relative humidity...). Differently, the standard parameters refer to the *Snowclim* parameters (described in Table B1 of the original manuscript) used by Lute et al. (2022) for their full model run (Table 2, superscript [c] in Lute et al. (2022)) across the Western USA. We have addressed this comment in the revised mauniscript as follows:

L588-591 "*Snowclim* model parameters were calibrated with a two-step approach, involving a subset of the snow dataset. First, we identified the best working downscaling techniques — multiple options were assessed to downscale temperature, precipitation, relative humidity, and downward shortwave radiation — by comparing model performance with measured SD using the standard parameter set used by Lute et al. (2022) (Table 2, superscript [c]) in their full model run."

Figure B1: It would be interesting to see different scatter plots for the snow surveys and the point measurements.

Addressed in the revised text.

Figure C1. Why not just a map with the bias per station, and compare it with the one from Dunmire et al. (2024) in their Figure 2?

We chose to not display the bias per site individually to focus the attention to only those sites that display serious underestimation. Comparison with Figure 2 of Dunmire et al. (2024) can be made, as

there seems to similarity for the sites that display serious underestimation. This comment is addressed in the revised text as follows:

L383-392 "Additionally, all three configurations exhibit difficulties in predicting high SD values at unobserved sites (Fig. 3), which likely arises from the scarcity of high SD observations in the training dataset (Fig. E1a). For the spatial nested CV framework, the $\text{PolSAR}_{\text{ML}}$, $\text{Backscatter}_{\text{ML}}$ and $\text{Combination}_{\text{ML}}$ configurations exhibit a bias of -1.46, -1.51 and -1.48 m for observed SD $\geq$ 2.5 m, which further deteriorates in the spatio-temporal framework. This is also visible in Fig. 3a, when comparing the observations (squares) with the SD prediction, and in Fig. 3b, that displays sites with a bias $\leq$ -0.5 m for the different configurations within the spatial nested CV framework (zero-measured SD excluded). Interestingly, many of the sites displayed in Fig. 3b also appear to have strong negative biases in Fig. 2a of Dunmire et al. (2024). Consequently, the reduced ability of XGBoost models, trained with S1 variables to explain interannual and site-specific variability, to predict high SD values must be taken into account when deploying the model across the European Alps, particularly when applying it beyond the temporal range covered by the training data."

References

Brangers, I., Marshall, H.-P., De Lannoy, G., Dunmire, D., Mätzler, C., and Lievens, H.: Tower-based C-band radar measurements of an alpine snowpack, The Cryosphere, 18, 3177–3193, https://doi.org/10.5194/tc-18-3177-2024, 2024.

Charbonneau, A., Deck, K., and Schneider, T.: A Physics-Constrained Neural Differential Equation Framework for Data-Driven Snowpack Simulation, Artificial Intelligence for the Earth Systems, 4, https://doi.org/10.1175/aies-d-24-0040.1, 2025.

Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785–794, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2939672.2939785, 2016.

Dozier, J., Bair, E. H., and Davis, R. E.: Estimating the spatial distribution of snow water equivalent in the world's mountains, Wiley Interdisciplinary Reviews: Water, 3, 461–474, `https://doi.org/10.1002/wat2.1140`, 2016.

Dunmire, D., Lievens, H., Boeykens, L., and De Lannoy, G. J.: A machine learning approach for estimating snow depth across the European Alps from Sentinel-1 imagery, Remote Sensing of Environment, 314, 114 369, https://doi.org/10.1016/j.rse.2024.114369, 2024.

Fontrodona-Bach, A., Schaefli, B., Woods, R., Teuling, A. J., and Larsen, J. R.: NH-SWE: Northern Hemisphere Snow Water Equivalent dataset based on in situ snow depth time series, Earth System Science Data, 15, 2577–2599, https://doi.org/10.5194/essd-15-2577-2023, 2023.

Gascoin, S., Luojus, K., Nagler, T., Lievens, H., Masiokas, M., Jonas, T., Zheng, Z., and De Rosnay, P.: Remote sensing of mountain snow from space: Status and recommendations, Frontiers in Earth Science, 12, 1381 323, https://doi.org/10.3389/feart.2024.1381323, 2024.

Global Climate Observing System: Snow — Essential Climate Variables, `https://gcos.wmo.int/site/global-climate-observing-system-gcos/essential-climate-variables/snow`, accessed: 2026-01-21.

Gómez-de Mariscal, E., Guerrero, V., Sneider, A., Jayatilaka, H., Phillip, J. M., Wirtz, D., and Muñoz-Barrutia, A.: Use of the p-values as a size-dependent function to address practical differences when

analyzing large datasets, Scientific Reports, 11, 20 942, https://doi.org/10.1038/s41598-021-00199-5, 2021.

Goodarzi, M. R., Barzkar, A., Sabaghzadeh, M., Ghanbari, M., and Fathollahzadeh Attar, N.: Uncertainty Analysis of Machine Learning Methods To Estimate Snow Water Equivalent Using Meteorological and Remote Sensing Data, Water Resources Management, 39, 4471–4491, https://doi.org/10.1007/s11269-025-04164-z, 2025.

Huss, M., Sold, L., Hoelzle, M., Stokvis, M., Salzmann, N., Farinotti, D., and Zemp, M.: Towards remote monitoring of sub-seasonal glacier mass balance, Annals of Glaciology, 54, 75–83, https://doi.org/10.3189/2013AoG63A427, 2013.

Lievens, H., Demuzere, M., Marshall, H.-P., Reichle, R. H., Brucker, L., Brangers, I., de Rosnay, P., Dumont, M., Girotto, M., Immerzeel, W. W., Jonas, T., Kim, E. J., Koch, I., Marty, C., Saloranta, T., Schöber, J., and De Lannoy, G. J. M.: Snow depth variability in the Northern Hemisphere mountains observed from space, Nature Communications, 10, 1–12, https://doi.org/10.1038/s41467-019-12566-y, 2019.

Lute, A. C., Abatzoglou, J., and Link, T.: SnowClim v1.0: High-resolution snow model and data for the western United States, Geoscientific Model Development, 15, 5045–5071, https://doi.org/10.5194/gmd-15-5045-2022, 2022.

López-Moreno, J. I., Revuelto, J., Fassnacht, S. R., Azorín-Molina, C., Vicente-Serrano, S. M., Morán-Tejeda, E., and Sexstone, G. A.: Snowpack variability across various spatio-temporal resolutions, Hydrological Processes, 29, 1213–1224, https://doi.org/https://doi.org/10.1002/hyp.10245, 2015.

Matiu, M., Crespi, A., Bertoldi, G., Carmagnola, C. M., Marty, C., Morin, S., Schöner, W., Cat Berro, D., Chiogna, G., De Gregorio, L., Kotlarski, S., Majone, B., Resch, G., Terzago, S., Valt, M., Beozzo, W., Cianfarra, P., Gouttevin, I., Marcolini, G., Notarnicola, C., Petitta, M., Scherrer, S. C., Strasser, U., Winkler, M., Zebisch, M., Cicogna, A., Cremonini, R., Debernardi, A., Faletto, M., Gaddo, M., Giovannini, L., Mercalli, L., Soubeyroux, J., Sušnik, A., Trenti, A., Urbani, S., and Weilguni, V.: Observed snow depth trends in the European Alps: 1971 to 2019, The Cryosphere, 15, 1343–1382, https://doi.org/10.5194/tc-15-1343-2021, 2021.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, Journal of Atmospheric and Oceanic Technology, 29, 897 – 910, https://doi.org/10.1175/JTECH-D-11-00103.1, 2012.

Miller, Z. S., Peitzsch, E. H., Sproles, E. A., Birkeland, K. W., and Palomaki, R. T.: Assessing the seasonal evolution of snow depth spatial variability and scaling in complex mountain terrain, The Cryosphere, 16, 4907–4930, https://doi.org/10.5194/tc-16-4907-2022, 2022.

World Meteorological Organization: Rapid changes in cryosphere demand urgent, coordinated action, https://wmo.int/news/media-centre/rapid-changes-cryosphere-demand-urgent-coordinated-action, 2023.

Author responses to comments of reviewer 2

Quantifying the volume of frozen water stored as seasonal mountain snow is a fundamental topic in hydrology and water resources management. The authors address the research need by modeling snow depth (SD) at in situ locations and for spatially distributed applications, using an XGBoost machine learning algorithms and forcing it with Sentinel-1 dual polarized synthetic aperture radar (PolSAR) data products and traditional inputs (e.g., precipitation , elevation, slope, aspect). The novelty of the work is using the PolSAR data in the ML workflow, and rigorously assessing the information gained by comparing to existing ML SD modeling workflows through model performance metrics, feature importance, and SHAP analysis-for a total of 5 ML SD models. The results indicate that the PolSAR features improved model performance at in situ and spatially distributed SD modeling locations, often a key model input in the feature importance assessments but behind meteorological and fraction snow cover area products. A key element of the manuscript is the three-fold nested cross-validation technique used for model evaluation. It is rigorous and establishes a standard for others to follow as the method supports an independent assessment of the model's spatial, temporal, and spatiotemporal prediction skill.

While the manuscript provides a publishable contribution to the hydrologic and cryosphere modelling community, the manuscript is quite long and could be much more focused. Primarily, this is the results and discussion section that is 9 full pages. While the information is useful, many of the case-by-case examples could be placed in a supplementary information section and the manuscript could communicate the key findings of the study and discuss. I suggest a minor revision, emphasizing a restructuring of the Results/Discussion section to have it more effectively communicate the key results and discuss their relevance to the snow modeling community.

First, we would like to thank the reviewer for the useful comments, which helped us improve the manuscript. Below, we listed our replies to the comments in red. The original reviewer comments are included in black, new text in the manuscript is shown in blue, with the indicated line numbers referring to the revised manuscript. As response to the main comment on the manuscript length, we have shifted some of the case-by-case examples to the Appendix. As a result, this section now comprises around 2 pages less text and is more focused towards the key findings.

Main Comments

Comment 1
The introduction has great information but could improve in flow. Key research gaps are shared throughout the section, rather than the gaps being shared at the conclusion. Revising the introduction to have each paragraph build on the prior and clearly state the research gap(s) identified in the literature review would improve the section's flow and support reader comprehension.

The introduction has been revised. Necessary missing references have been added, and the paragraphs have been restructured to improve flow and readability.

Comment 2
Additionally, I suggest additional commentary in the figure captions, highlighting the key takeaways in addition to the description of the labels. Having the authors highlight the key findings in the figure captions will support a more detailed understanding of the work for a broader audience. Lastly, I request that figure 8 includes the No Survey mapped product. Adding this plot to the figure would improve the understanding of the spatial performance of the in situ trained model for spatial prediction.

We modified the captions in the revised manuscript to highlight key findings. Next, Figure 8 of the revised manuscript includes the "No Survey" mapped product. The revised manuscript also includes a map of SD estimates across the whole study area, generated with the models obtained from the spatial nested cross-validation approach.

Comment 3
Much of the manuscript results/discussion section would be appropriate for a supplementary material section so that the main document could be clearer and concise, aiding in interpretation to a broader audience.

Many of the case-by-case examples have been moved to the Appendix. For example, we moved Section "4.2.2 Time series of SD and SHAP values" entirely to the appendix. Next, we moved the more detailed discussion of the 16 March 2017 Dischma valley snow survey to a new Appendix "Appendix D: 16 March 2017 Dischma valley snow survey SD estimates", where the results are discussed based on two new figures to resolve many of the "not shown" instances (in response to concerns from Reviewer 1).

Comment 4
The abstract could highlight the impacts more, rather than saying what is done.

We have modified the abstract in the revised manuscript, highlighting more results and how these results relate to the last sentence of the original abstract. Therefore, we changed the abstract to:

L1-18 "Seasonal mountain snow is an indispensable resource, but accurate estimates of this water storage remain limited, even in the European Alps, where there is a dense network of in situ monitoring stations. In this study, we address Alpine snow depth estimation at a 100 m spatial resolution and sub-weekly temporal resolution over the 2015–2024 period using multiple input configurations within an extreme gradient boosting (XGBoost) machine learning (ML) model. We explore the potential of Sentinel-1 C-band dual-polarized synthetic aperture radar polarimetry (PolSAR) observations, and include either regionally downscaled meteorological forcing data or modeled snow depth as additional inputs to further explain interannual and spatial variability. A threefold nested cross-validation scheme is used to account for the spatio-temporal dependencies present in the snow depth data. XGBoost's internal booster and Shapley additive explanation (SHAP) values are used to relate the input features with the predictions for both dry and wet snow conditions. Our results indicate that the inclusion of PolSAR observations leads to modest improvements over a backscatter-intensity-based configuration, whereas the SHAP-based feature attribution reveals a high reliance of XGBoost on the polarimetric scattering angle and co-polarized (VV) backscatter intensity. Next, incorporating either meteorological forcing data or modeled snow depth substantially enhances predictive performance, particularly when spatially distributed training data, proven to be essential for capturing topographic controls on snow depth variability, are included. When supplemented with spatial training data and either meteorological forcing data or modeled snow depth estimates, XGBoost shows good agreement with nine snow surveys conducted in the Dischma valley (Switzerland), achieving correlation coefficients (R) of 0.76 and 0.78 and mean biases of 0.07 and 0.17 m, respectively. When applied to unseen locations across the Alps, the performance remains high, with $R = 0.80$ and biases of -0.04 and -0.03 m, respectively."

Comment 5
The abstract shares a complete summary, but covers too much. It could be more concise with respect to the main experiments, results, and conclusions.

Within the revised abstract, we put more focus on the key takeaways. We still opted to include the main experiments conducted, to give the reader a complete overview of what to expect when consulting our manuscript.

Comment 6
The manuscript would substantially benefit from being more focused and highlighting key conclusions. Additional information could be referenced and placed in a supplemental information section.

Addressed by moving some of the detailed discussions to the appendix.

Comment 7
I suggest in Table 2 that the authors provide a definition of the variables (part of the ML configuration). This would support a broader audience's understanding of the work and provide a clear reference for variables mentioned throughout the manuscript.

Addressed.

Comment 8
In the conclusion (line 582-583), the authors state "These results showcase the potential of using PolSAR observations within non-machine learning applications, e.g., within a conceptual model." This is new information for the reader, and I do not see where the results support this conclusion.

We agree that this suggestion is not previously discussed throughout the manuscript and not necessarily relevant for the work presented in our manuscript. We have therefore chosen to remove this sentence.

Comments by line number

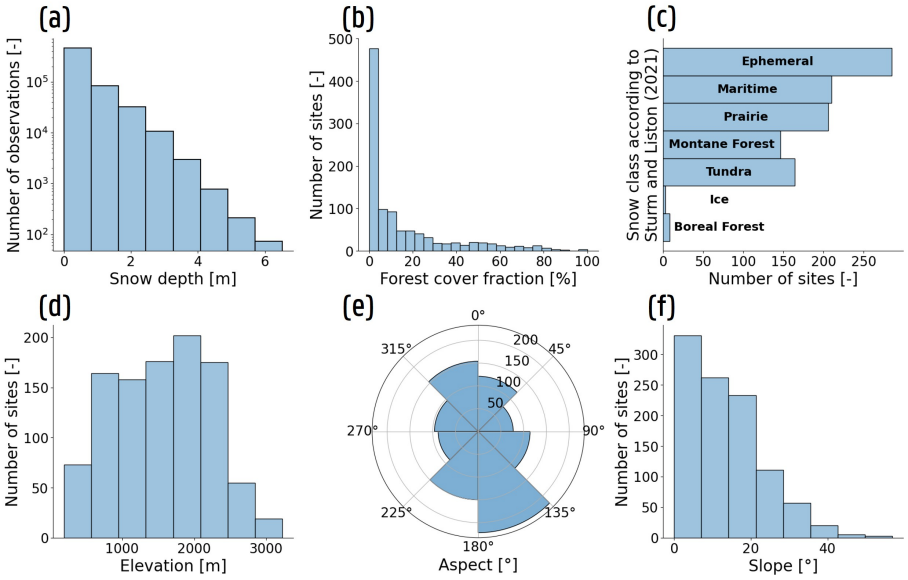
107-108, 118: How do the in situ and point-based measurements represent the variability of SD throughout the study domain (European Alps). What is the range in snow environments (e.g., Sturm snow classification), elevation, aspect, slope, etc observed in the data. Sharing the distribution of training data would support more nuanced conclusions related to the model output, such as providing more context to model skill at locations other than in situ locations.

The point-based in situ measurement sites shown in Fig. 1 of the original manuscript represent the snow classes defined in Sturm and Liston (2021) well, with the exception of the Ice and Boreal Forest classes (Review Fig. 1). However those two classes are not often observed within the European Alps and, as outlined on line 216 of the original manuscript, we excluded sites located in glacial areas following the Randolph Glacier Inventory version 7.0. Nonetheless, some sites that did not appear in the Randolph Glacier Inventory are classified as Ice by Sturm and Liston (2021). Next, the measurement sites represent most elevations found within the European Alps, with the exception of elevations above 3000 m, and the lowest elevations. The training dataset contains fewer east and west facing slopes, and has roughly the same amount of north and south facing slopes. Finally, the amount of sites with steep slopes is limited, as was described within the original manuscript (lines 524-526), and only 2% of the measured snow depths is $\geq$ 2.5 m.

In order to give a better overview of the distribution of the training data, we added Review Fig. 1 of this response file to the Appendix of the revised manuscript.

180-181: Can the authors provide a citation for their precipitation downscaling method?

Equation 1 of the original manuscript is an adaptation from Equation 2 in the study of Huss et al. (2013), who used meteorological forcing data to model accumulation and melt for two glaciers in Switzerland. Within the study, the authors corrected observed (in situ) precipitation data for gauge



Review Figure 1: Distributions of measured SD and static features of the in situ measurement sites used to train, validate and test XGBoost. (a) Distribution of measured SD. (b) Distribution of forest cover fraction of the unique measurement sites. (c), (d) (e) and (f), same as (b), but for the snow classes described in Sturm and Liston (2021), elevation, aspect and slope, respectively.

undercatch errors using a scaling factor c_{prep} and an altitudinal gradient. Within our study, however, we did not address gauge undercatch. Instead, we adapted equation 2 of Huss et al. (2013) to downscale coarse-scale precipitation data to account for orographic precipitation. To this end, we modified this equation to 1) focus on areas with major elevation differences (by introducing D_{dif} and limiting D to 1, so that Equation 1 within mountainous areas becomes $P_{\text{coarse}_{x,t}} \cdot \left[0.75 + 0.5\frac{z_x - z_{\min}}{z_{\max} - z_{\min}}\right]$) and 2) by limiting the precipitation adjustments to 75-125% of the non-downscaled values. The original manuscript has been adapted as follows:

L192-201 "First, coarse-scale precipitation (P_{coarse}; [mm 3h^{-1}]) was corrected as a function of elevation to account for orographic effects, using a rescaling function adapted from Huss et al. (2013), who corrected in situ precipitation data for gauge undercatch in a glacier mass-balance study. Thus, for each location x within the 500 m grid at time step t, $P_{\text{coarse}_{x,t}}$ was downscaled using the following equation:

$$P_{x,t} = (1 - D) \cdot P_{\text{coarse}_{x,t}} + D \cdot P_{\text{coarse}_{x,t}} \cdot \left[0.75 + 0.5\frac{z_x - z_{\min}}{z_{\max} - z_{\min}}\right] \text{ with } D = \frac{z_{\max} - z_{\min}}{D_{\text{dif}}} \quad 0 \leq D \leq 1 \tag{1}$$

with z_x the elevation [m] of the 500 m Copernicus GLO-30 DEM, $z_{\min}$ and $z_{\max}$ the minimum and maximum elevation [m] within an interpolation window — centered on the location x and spanning an area roughly matching the original 0.1° grid size — and D_{dif} a user defined difference in elevation, set to 250 m. The user defined difference was introduced to focus the corrections on the study area, with minor adjustments for areas with small elevation differences. Different from Huss et al. (2013), Equation 1 limits the downscaled precipitation values between 75 and 125% of the original $P_{\text{coarse}_{x,t}}$ values."

259: Table 2. Can the authors include a description of the model variable here? The would provide

a reference for the reader as they move through the manuscript and want to know what variables the authors are related to model skill and feature importance.

We adjusted the caption of Table 2 in the revised manuscript, including a description of the variables.

362/493: Table 3. The authors mention Bias as a model evaluation metric, but it is not in any table or figure. Can the authors add Bias throughout?

We have included bias in the revised Figures and Tables.

508: "SHAP value FI", I think this is referring to Figure 7. Can the authors clarify what FI figure to look at, or include it if it is not present?

The SHAP value FI referred to in line 508 does not refer to Fig. 7 within the original manuscript. We admit that this was not clear in the original manuscript. As part of other comments, we revised Fig. 6 of the original manuscript. The Figure now displays SHAP values for the three configurations for both the spatial and spatio-temporal nested CV schemes. From this revised Figure, our description in lines 508-511 of the original manuscript should be clear. Thereby, we omitted line 512 to focus the text more to how the S1 features are used during SD prediction.

534-536: The elevation bands do not provide meaningful information without connecting to a snow environment (e.g., alpine, tundra, sub-alpine...) as, depending on the location, this could be alpine or forest. I suggest coupling the elevation band ranges with a snow environment type. Lastly, 1000m is substantial and could cover several snow environments. A more representative snow environment communication other than elevation will be more impactful to the readers.

Thank you for this comment. As we have the snow class information from Sturm and Liston (2021) available, we have adapted this part of the discussion in the revised text. This discussion has been moved to the new Appendix "Appendix D: 16 March 2017 Dischma valley snow survey SD estimates" (See main comment 3).

542: Figure 8: I suggest adding the No Survey Data spatial map so the readers can visualize the snow depth prediction across the landscape. Based on the difference map (b,e) it appears that the No survey data map product may estimate the same snow depth for all pixels, not acknowledging terrain impacts on mountain snow depth distribution. Adding two more maps would show that this is not true, or if it is, highlight the need for spatial data to be included in models. This would also help support the claim of spatial SD leading to improved spatial patterns (line 556)

The model output from the "No survey Data" does not produce the same snow depth for all pixels, as it still includes topographic features during model training. The main difference with the "Survey data" models (nine in total), is that the latter are trained on a more versatile dataset with respect to topographical features. The 'original' training dataset, not including the snow-survey SD maps, contains more low-slope values, and more locations with low TPI-values. As such, an ML model trained solely on the data of the stationary sites may already perform well. Nonetheless, including spatially distributed data within the training procedure improves the representation of topographical impacts on the predictions, which is evident from Fig. 8 in the revised manuscript.

582-583: I do not know how the authors came to this conclusion and do not recall (nor can I find) supporting information within the manuscript to come to this conclusion. I suggest either removing or ensure supporting findings are mentions earlier in the document.

Thank you for pointing this out. You are right and this has now been removed in the revised manuscript.

References

Huss, M., Sold, L., Hoelzle, M., Stokvis, M., Salzmann, N., Farinotti, D., and Zemp, M.: Towards remote monitoring of sub-seasonal glacier mass balance, Annals of Glaciology, 54, 75–83, https://doi.org/10.3189/2013AoG63A427, 2013.

Sturm, M. and Liston, G. E.: Revisiting the Global Seasonal Snow Classification: An Updated Dataset for Earth System Applications, Journal of Hydrometeorology, 22, 2917 – 2938, https://doi.org/10.1175/JHM-D-21-0070.1, 2021.