Author responses to comments of reviewer 1

I sincerely appreciate the efforts that the authors have made to address or rebut my comments and I am satisfied with the revised manuscript, which has improved with regards to the stated novelty, and the presentation and impact of the results. I congratulate the authors for the good and rigorous work. There are still a couple minor considerations that should be addressed, after which I would be happy to see the paper published.

First, we would like to thank the reviewer for the kind words and useful comments, which have helped us to improve the manuscript even further. 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
Follow up on comment 2 about XGBoost: What made the authors choose XGBoost after the comparative study with other models? Was it a better performance, ease of use, computational efficiency?

We took the decision both based on performance and computational efficiency. As mentioned in the revised manuscript, XGBoost is easy and fast to train and tune. The computational time necessary to train one hyperparameter configuration is substantially less compared to the other tree-based algorithms tested, which made XGBoost best suited for the cross-validation experiments. In addition, XGBoost displayed best model performance. We addressed this comment in the revised manuscript as follows:

L235-237 "We chose this ML algorithm after comparative experiments with other tree-based methods (including Random Forest, LightGBM and CatBoost) in which XGBoost showed the best performance and computational efficiency, as well as its use in recent snow studies (e.g., Dunmire et al. (2024) or Goodarzi et al. (2025))."

Comment 2
Note on comment 3: I am happy with the revised sentence and the authors take on my comment about the ML model learning the physics. Indeed, we do not know what the model learns, but what we know is that the relationship between snow depth and climate variables is reproduced.

We are pleased to hear that the revised manuscript is better describing the ability of the ML model to learn the physics behind the processes occurring.

Comment 3
I appreciate the efforts on Comment 4 and the extra analysis/figures, but something is still unclear to me. The authors state in Review Figure 2 and associated text, that the "sample size" refers to the number of sites used. However, the performance values of Table 3 (CombinationML) and Figure 6 (also CombinationML) are the same, and it is clear based on Figure 6 that the sample size is in the order of millions (all measurements used, not per site). Therefore, I suspect that these performance values are calculated based on all the points in the dataset, and not based on average points per site. Please clarify this and if that is the case, then again I would suggest against the use of significant differences based on p-value, for this study.

You are correct to note that the performance metrics discussed in Table 3 and Fig. 6 in the revised manuscript are based on all available in-situ measurements. A total of 593,166 measurements were used in this comparison, of which ∼40% are zero-measured values that were excluded from the computation of the performance metrics.

We agree that using p-values on all measurements could lead to significance attributable to the sample size. Therefore, the revised manuscript specifically describes significant differences only for site-specific performance statistics. As written in L374-375 of the current manuscript, we acknowledge that the overall performance is similar (based on the overall performance metrics), and only refer to significant difference for the MAE, computed per site. As such, for this analysis, the sample size used to compute statistical significant improvements is not based on the 593,166 measurements, but on the site MAE, which are 1,022 values for each configuration. The latter is also most informative, as we hope to see improvements at new unseen sites by incorporating the PolSAR variables. It has also been discussed that while we see a significant improvement in site-MAE for certain configurations and CV frameworks, the improvements are often marginal.

Nonetheless, we acknowledge that L380 of the current manuscript can still be confusing, and we will address this by explicitly stating that all significant differences are based on site-level computed performance metrics. Therefore, we adjusted the following lines in the revised manuscript:

L381-384 "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 further supported by statistically significant gains in site-MAE ($p \ll 0.05$, 95% confidence level) observed in the temporal nested CV framework."

L547-549 "Our results show modest improvements in estimated SD with the inclusion of S1 C-band PolSAR observations, with gains primarily observed in site-level performance under spatio-temporal generalization, whereas no significant improvements are found in these metrics under spatial generalization alone."

Comment 4
I agree with with reviewer 2 that given the length of the Appendix, it should be moved to a Supplement.

We agree that the appendices are quite long and make the manuscript rather long when reading all information. Nonetheless, we think that keeping it as an Appendix would be beneficial. We often refer to the Figures in the Appendices, and the extended analyses provided in the Appendices directly correspond to the main manuscript. Therefore, if accepted by the editor, we would prefer to keep it as an Appendix. Otherwise, we will move the Figures to a supplement.

Comment 5
Data availability: Following open science practices, please provide the links for the freely downloadable snow depth data from the providers across the European Alps. Also indicate which snow depth measurements were provided through personal communication and from who

We agree with the reviewers comment. We already partly addressed this comment by providing the links of the websites for the data collected for Switzerland, France, Italy and Austria. To further address this comment and resolve remaining confusion, we replaced the hyper references with the links to all open source datasets. In addition, we now provide the names and corresponding email addresses for the measurements obtained through personal communication. Note that we removed the TIWAG reference in the manuscript. After careful validation of the data sources, we confirmed that our dataset contains only measurements from sites provided by GeoSphere Austria (open-source). This was a mistake from our side.

References

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.

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.

Author responses to comments of reviewer 2

The authors have effectively and thoughtfully addressed previous comments and improved the manuscript to a more impactful level. The revisions in response to comment have improved the manuscript focus, readability, and clarity of results while maintaining robust detail of the extensive data processing and methodologies. I have suggested below only few, minor revisions to further improve the manuscript, and believe further review not necessary for publication.

First, we would like to thank the reviewer for the useful comments, which have helped us to improve the manuscript even further. 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

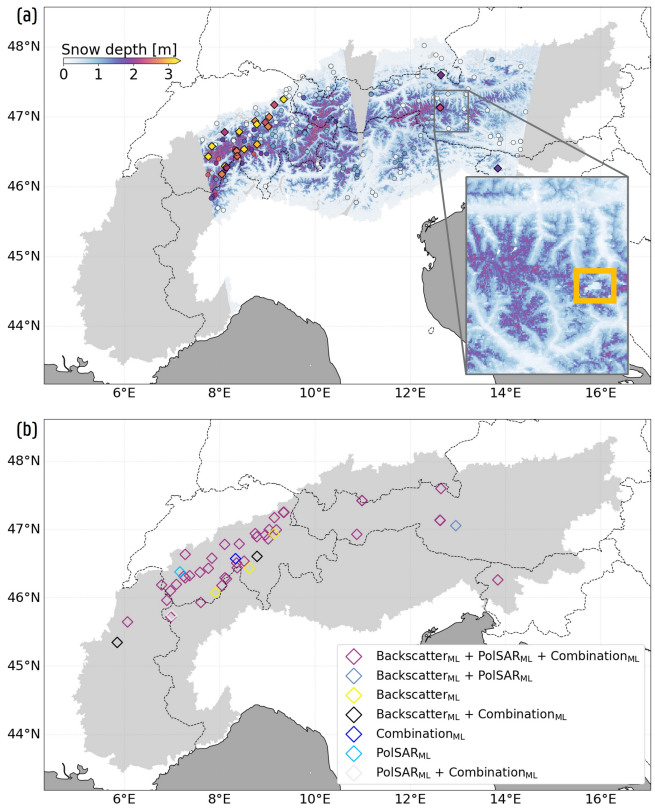
Comment 1
The scale and color schemes of figures (particularly Figure 1, Figure 8, and the inset of Figure 3) make it difficult to see. Suggest increasing the size of Figure 1. Figure 3 it is difficult to see the purpose of the inset and yellow box - the inset can cover a smaller area. FSC errors are discussed, but I don't see how that relates to the figure – is it depicting an area of zero snow depth? Figure 8, the point estimates are difficult to see at this color and scale.

We have adjusted the figures and color schemes to improve the readability of the manuscript. More specifically, we changed the figure size of Fig. 1 and we adjusted both the inset and caption in Fig. 3 to better specify the purpose of the inset (Review Fig. 1). Finally, we increased the size of the point measurements in Fig. 8, increased the image size of sub-figures a), b), c), e), f) and g), and thereby increased the contrast between the plotted measurement locations and the background map.

For Fig. 3 specifically, the inset is provided to highlight the area of SD underestimation error which is not visible in the plot across the whole Alpine region. Because the XGBoost predictions depend heavily on FSC (which can be seen through the feature importance), errors in the FSC product result in errors in the model output. This is clearly the case for the yellow box. The area covered by the box is high-alpine terrain where SD estimates are expected in the deep-blue or purple range of the colorbar. To address this comment, we made L397 of the original manuscript more clear:

L401-402 "FSC is particularly important, which makes the predictions prone to potential errors in this input feature (e.g., the high-alpine area in the yellow box in Fig. 3a, where erroneously low FSC-values led to SD underestimation)."

Comment 2
In response to Reviewer 1, Comment 6, the authors state "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." This could be more explicitly incorporated into the Introduction (line 96-107) to better establish the purpose and goals of the study.



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 to indicate an area of SD underestimation errors (yellow square) attributable to errors in the SCF input feature. (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 and incorporated the main purposes of the manuscript in the revised final paragraph of the introduction:

L96-109 "In this manuscript, we further investigate the potential of ML with different input configurations to accurately estimate SD across the European Alps, essential for accurately quantifying the annual water stored as snow. To this end, we first conduct various experiments comparing the performance of S1 C-band backscatter intensity with PolSAR observations to quantify the added value of PolSAR observations relative to, and in combination with, backscatter intensity. To gain insights in when and where which type of satellite observations contribute to the SD predictions, we use feature importance (FI) analysis under both dry and wet snow conditions. We further evaluate the added value of incorporating meteorological forcing data and process-based snow model SD estimates as features in the ML model, to assess improvements in capturing interannual and site-specific variability, and to determine whether S1 observations remain influential. To validate our approach, we implement a threefold nested cross-validation (nested CV) framework, which masks subsets of the data during train-

ing and predicting (testing). This framework accounts for the spatial, temporal and spatio-temporal dependencies in the data, an essential consideration when validating ML models for spatio-temporal purposes (Meyer et al., 2018). Finally, to address the limitations of relying solely on point-based training data for spatial prediction and representing topographic effects on the estimates, we compare models trained with and without airborne snow survey data, and validate predictions against nine airborne photogrammetry surveys in the Dischma Valley, Switzerland."

Comment 3
Section 3.4: Overall the spatiotemporal CV structure is rigorous. However, the construction of spatial folds raised questions. Fig. 1 shows that the point locations are quite dense in some areas. How many clusters result from the 5 km threshold, and how does the distribution and number of sites vary within these clusters? Since each fold contains a similar amount of data, some folds must contain much denser data with fewer clusters (e.g., the folds containing clusters of sites in NW Italy) than others (e.g., central Austria). Does this impact the CV at all?

The spatial clustering indeed results in clusters that are more dense in case many measurement locations are close to each other. Nonetheless, by imposing a minimal distance of 5 km between the locations, the number of clusters constructed remains high (598 clusters when using a 5 km radius, with maximum 11 nearby sites within the same cluster). In addition, by randomly distributing the clusters into folds to have an equal amount of data and preserving a roughly equal SD distribution, we have a roughly equal distribution of both the static features and target variable (SD). The latter is most important, as a model trained solely on low SD values will not extrapolate well to unseen areas with only high observations and vice versa.

Nevertheless, the CV is impacted by spatial clustering. As written in the original submitted manuscript in L494–496 "Despite these gains, the number of sites with a mean bias ≤ -0.5 m remains similar in both the Weather$_{\mathrm{ML}}$ and *Snowclim*$_{\mathrm{ML}}$ configurations, with most of these sites again located in Switzerland (not shown). Although a smaller clustering radius during fold creation could mitigate this underestimation through more balanced fold SD distributions, the current clustering offers a fair evaluation of model performance, indicating a need for expanded data collection of high observed SD values.", the clustering radius can indeed affect the CV results. A smaller clustering radius results in more clusters (876 clusters when a radius of 1 km is used), such that a single 5 km cluster containing high SD observations may be split into several smaller 1 km clusters, thereby improving the distribution of these high SD observations when randomly assigned to the five folds. In contrast, a larger clustering radius leads to fewer, larger clusters, which can result in less balanced SD distributions across the folds and, consequently, a stronger sensitivity of the CV results to the spatial grouping of the data. The effect of the clustering is currently being tested in a Master's thesis, with preliminary results indicating slight increases in performance metrics for a smaller clustering radius. Future work can focus more on this effect, but is beyond the scope of this manuscript.

Comment 4
Line 456 introduces another XGBoost based model built on the point based data. Is this the "baseline" model referenced in Table 2? This should be introduced earlier, outside of the performance metric section.

In L456 of the current manuscript (Section 4.3 "Added value of meteorological forcings and *Snowclim* SD estimates"), the new introduced model is the Weather$_{\mathrm{ML}}$ configuration (XGBoost models). This indeed refers to the input feature configurations displayed in Table 2. However, we think there is confusion between the use of *baseline* in Table 2 and a *baseline*-model. The use of *baseline* refers to shared input features among the configurations, and not a model 'to improve upon'. We have addressed this confusion by changing *baseline* in Table 2 and related text to *common* features. In addition, we added the following to the revised manuscript:

L460-462 "Figure 6 illustrates the improvements when incorporating either regionally downscaled meteorological forcing data or *Snowclim* SD estimates (Table 2; $\text{Weather}_{\text{ML}}$ and $\textit{Snowclim}_{\text{ML}}$ configurations) when predicting at unseen locations (and time periods; Fig. 6a and b), with the most important improvements for high SD observations ($\geq$ 2.5 m)."

With the following changes and the titles of the subfigures in Fig. 6, no confusion should remain about a *baseline*-model.

Comments by line number

L75: Section 3.3.2 states Ps,c and Uac are summed from September 1st to the prediction date, while SWd7 and MD7 are the 7 preceding day summation. Why is wind speed not at the 7-day interval?

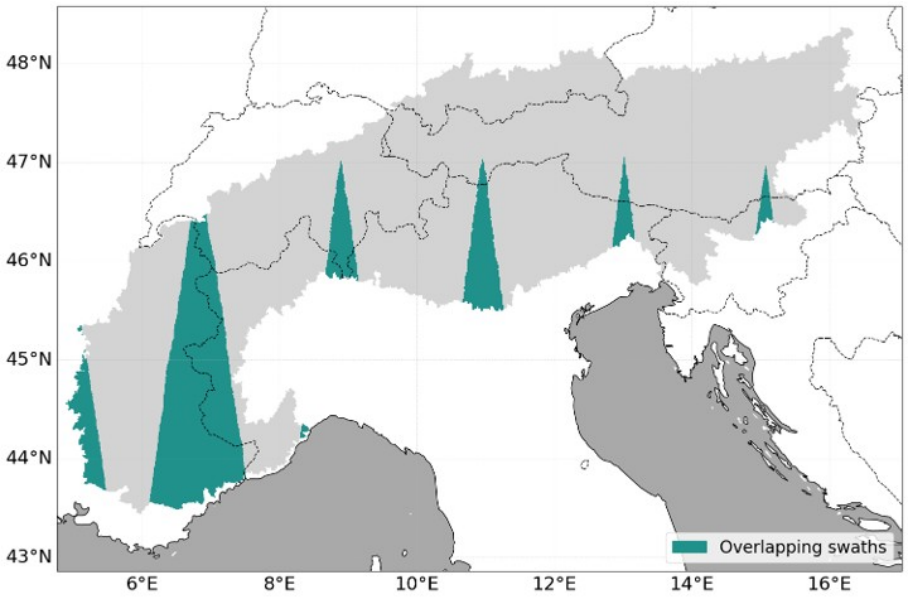
We included a cumulative wind speed feature to provide an indication of wind-prone areas, with higher values expected at locations subject to enhanced snow drift. Rather than using a 7-day sum, a cumulative metric was adopted to allow XGBoost to better distinguish persistently wind-exposed areas, even in the absence of recent wind events. While a short-term aggregate (e.g., the 7-day sum) may also provide useful information, we argue that at a spatial resolution of 500 m, such variability is unlikely to be sufficiently informative. This was supported by SHAP-values on initial experiments with the meteorological forcing data, which showed higher SHAP-value FI for Ua_{c} than for Ua_{7}. In addition, the lack of wind direction data prevents a more explicit characterization of the directionality of snow redistribution.

L289: How many locations have both ascending and descending S1 orbits? Is there any difference between the XGBoost outputs for ascending vs descending observations?

There are several places in the Alps that experience both an ascending and descending S1 observation on the same date, indicated in Review Fig. 2. Across all measurements used for the spatial nested CV framework, we found that $\sim$4% of the instances in the training dataset contained multiple same-date S1 observations.

Referring to the second part of the reviewer's question, there can be a difference in output between A and D observations. However, this difference is not related to the track (ascending vs. descending), rather it is determined by the viewing geometry associated with the orbits belonging to these tracks and the impact of this viewing geometry on the observed S1 variables. More specifically, the local incidence angle (LIA) determines which scattering mechanisms happen at a specific location (see Appendix C: time series of SD and SHAP values). If one location has very different LIAs between the relative orbits, then this can result in distinct SD predictions. This is also visible in Fig. C1b between January and April 2018. Given the differences in observed α^{s}-values between the orbits (and thus tracks; Review Fig. 3), and the positive relationship XGBoost has learned between observed α^{s}-values and SHAP-value contribution, one can see jumps in the predictions as well (Fig. C1b). Although we expected these differences to be (partly) diminished by inclusion of the LIA in the model, it seems from both this Figure and the SHAP-value FI that XGBoost only limitedly learned to account for these viewing geometry differences. Future research should test new methods to improve the scaling of the S1 observations to better account for the viewing geometry differences. This is currently beyond the scope of this manuscript.

To indicate this influence of viewing geometries in the methodology section, we added the following in Subsection 3.3.3 "XGBoost snow depth prediction procedure":



Review Figure 2: Areas in the Alpine region prone to overlapping (same date) S1 observations.

L290-292 " Since S1 observations are influenced by orbit-specific viewing geometries — potentially resulting in different SD estimates — input features from ascending and descending orbits were fed to XGBoost separately."

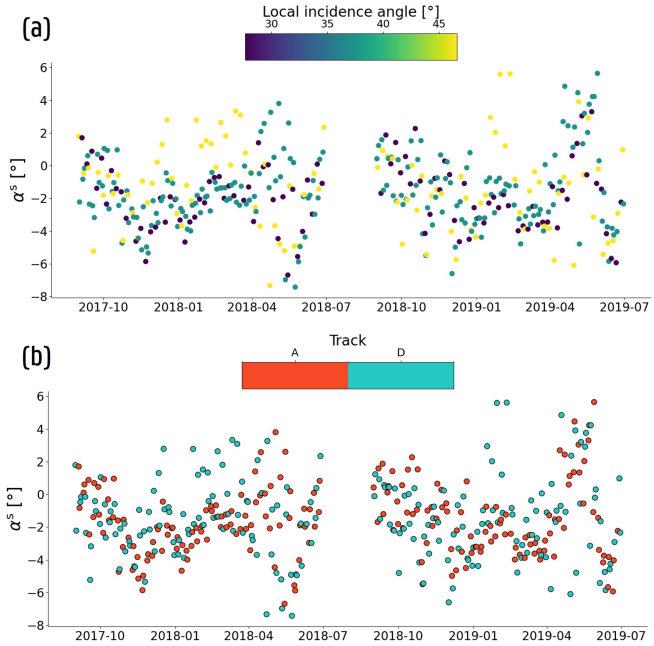
Line 381: there is reference to an increased computation cost, though this is not addressed. Suggest including a computational demand figure or table of each configuration in the appendices.

Processing S1 SLC data is computationally more demanding than processing GRD data. Besides a longer processing graph (two separate processing chains in SNAP to produce either the polarimetric scattering angle or the first Stokes parameter from the C2-matrix; see Section 2.2 of the original manuscript), a raw SLC image takes up 5 Gb compared to 1 Gb for a GRD product. Therefore, also downloading the raw images before processing takes longer for the polarimetric variables compared to the backscatter intensity ones. We did not specifically keep track of the processing time during the S1 preprocessing. In order to resolve the reviewer's comment, we have not added a new Figure or Table, but instead explained the computational cost of the PolSAR data more in detail:

L384-387 "Nonetheless, also here the overall 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. Indeed, besides requiring separate processing chains to retrieve α and S_0 from the C2-matrix (Section 2.2), handling SLC data is more demanding, with raw images ($\sim$5 Gb) substantially larger than GRD products ($\sim$1 Gb)."

Line 440–448: this paragraph discusses correlation between LIA and other input variables, but the cited figures (C1b and 5) do not depict LIA. I think this could be clarified somewhat.

This links to our reply on the second comment-by-line-number. The LIA namely determines how the S1 microwaves penetrate the snowpack. For low LIAs, S1 looks more vertical towards the snowpack, while for higher LIAs, S1 observations penetrate more oblique through the snowpack. In the first case, it has been shown by Jans et al. (2025) that often increases in γ^0_{VV} (and S_0) are observed.



Review Figure 3: Changes in α^s between ascending (A) and descending (D) orbits at a location near Grindelwand: 46.67°N, 8.06°E. The time series displays the snow seasons 2017-2018 and 2018-2019. (a) Time series colored by local incidence angle (LIA). (b) Same as (a), but colored along corresponding track.

One hypothesis is that the signal in co-polarization is thought to to be less depolarized, but instead more sensitive towards surface scattering from snow-layer interfaces. However, this hypothesis remains highly uncertain, and more research is required to unravel the origin of this increase in γ^0_{VV} (and S_0). In addition, increasing γ^0_{VV}- (S_0-) values can be observed at other sites with medium LIAs as well (e.g., Fig C1b), making it location-specific. This is also visible within Fig C1b, especially for the 2017-2018 snow season.

Differently, for S1 observations with medium to high-values, mostly rising α- and/pr γ^0_{CR}-values are observed with increasing SD. As explained in the Introduction, this is thought to be related to signal depolarization when the S1 signal penetrates the snowpack along an oblique path. To address the reviewer's comment, we made the following change:

L449-455 "Instead, increases in $\gamma^{0,s}_{VV}$ or S^s_0 are observed, a pattern shown to be more prevalent at sites with low-LIA satellite overpasses (Jans et al., 2025) where the S1 signal penetrates the snowpack more vertically and decreasing α-values are observed. Nevertheless, these patterns may also occur at locations with medium LIAs (e.g., the site in Fig. C1b that shows increasing $\gamma^{0,s}_{VV}$-values) and further research is required to understand their origin. In addition, these patterns help explain the dual relationship observed for $\gamma^{0,s}_{VV}$ in Fig. 5c: at certain locations, a limited positive contribution is found, whereas at the majority of sites, a strong inverse relationship emerges during wet snow conditions, when a decrease in this S1 variable is typically observed (e.g., Fig. C1a)."

Line 542: "with evident improvements at unseen locations during uncovered time periods, whereas no significant improvements are observed at unseen locations" – this sentence seems to contradict itself. Which time periods is the second half of the sentence referring to? Are uncovered time periods those not included in the training data?

Uncovered time periods refer to those not seen in the training dataset. We rephrased this sentence to follow the lexicon of the CV frameworks. This part is now revised to:

L547-549 "Our results show modest improvements in estimated SD with the inclusion of S1 C-band PolSAR observations, with gains primarily observed in site-level performance under spatio-temporal generalization, whereas no significant improvements are found in these metrics under spatial generalization alone."

Line 552: "XGBoost operates at the pixel level, limiting its ability to incorporate information from neighboring pixels" This is a difficult sentence to introduce without further explanation in the final paragraph of the conclusion. In these feature sets, the incorporation of TPI is incorporating neighboring pixel information, and there are indexing or windowing techniques that may be applied to further spatially inform the model. I suggest removing the sentence.

We agree that spatial information can be incorporated through specific features, such as the TPI. However, in contrast to convolutional deep learning methods, the model cannot inherently learn spatial relationships, and these must therefore be explicitly encoded in the input features. To address the reviewer's comment, we have revised the sentence as follows:

L559-561 "Furthermore, XGBoost operates at the pixel level and cannot inherently capture dependencies between adjacent pixels; instead, spatial context must be explicitly incorporated through engineered features."

Editorial comments

We thank the reviewer for pointing out these textual improvements. We resolved all editorial comments (except from the use of "as such", which we do think can be used within the context), and refer to the diff-file for the changes made.

References

Jans, J.-F., Beernaert, E., De Breuck, M., Brangers, I., Dunmire, D., De Lannoy, G., and Lievens, H.: Sensitivity of Sentinel-1 C-band SAR backscatter, polarimetry and interferometry to snow accumulation in the Alps, Remote Sensing of Environment, 316, 114 477, https://doi.org/10.1016/j.rse.2024.114477, 2025.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation, Environmental Modelling & Software, 101, 1–9, https://doi.org/10.1016/j.envsoft.2017.12.001, 2018.