Reviewer 1

The article demonstrates that mechanical snow profiles can be statistically evaluated and stratigraphic features reconstructed. Traditional snow profiles are limited in number due to limitation of field work. Using 921 snowmicropen profiles from different expeditions, the authors build on the work of Kaltenborn et al, who demonstrated the use of machine-learning methods to evaluate highly resolved force profiles in snow. Based on their parameterisation and algorithm, significant differences between first-year and multi-year ice floes are quantified.
The statistical quantification of snow properties is an important step towards synthetic modelling of the snowpack on Antarctic ice floes. The data are also essential to improve detailed snow pack models and develop stochastic modelling tools for these stratigraphically complex snowpacks. Concerning remote sensing, the data help to understand variation in albedo and microwave brightness temperatures. However, the article lacks standard statistical tests between observables.

On these lines, the paper is a major step towards a much more quantitative and statistically testable snowpack.
The paper is well structured. Propositions to improvements are detailed in the "Detailed comments". Figures should be adapted to the final format of TC, and the layout unified.

We thank Reviewer 1 for the careful assessment of our manuscript and for the detailed and constructive review. The comments helped us to improve the presentation and consistency of the figures, clarify methodological choices, strengthen the statistical comparison between first-year and multi-year ice, better define the reliability limits of the spatial autocorrelation analysis, and expand the discussion of methodological limitations. Below, we provide a point-by-point response. Reviewer comments are shown in black text, our responses in blue text. Manuscript text is quoted and manuscript changes are set in bold. Line numbers refer to the preprint manuscript.

We have revised the figures to better follow The Cryosphere formatting requirements and to ensure a consistent layout, typography, panel labeling, and units throughout the manuscript.

Detailed comments
Title: could be more specific, eg.
Different spatial scales and snow stratigraphy between first- and multiyear ice floes in the Weddell Sea, Antarctica
We agree that the original title could be more specific and should make clear that the study focuses on internal snow properties and their spatial variability. We have therefore revised the title.
The revised title reads: **"Variability of internal snow properties over Antarctic late summer sea ice on different spatial scales"**

Introduction: compare here the difference to the Arctic sea ice, cite the work of Sturm et al.

line 22: a short introductory paragraph on the work in the Arctic on sea ice (traditional: Sturm et al, SMP the pioneering work of King, then the MOSAiC) that forms in many ways the base to improved methods for the Antarctic.
(This is to some degree covered in lines 68 ff, but it much more in detail)
We agree that the studies by Sturm et al., King at al., and the SMP work conducted during MOSAiC have contributed substantially to our understanding of snow on Arctic Sea ice and provide an important methodological basis for subsequent sea-ice snow studies. However, because the focus of the present manuscript is on Antarctic snow rather than a direct hemispheric comparison, we have kept the Arctic backgorund focused on the studies most relevant to our methods and objectives, which are presented from line 68 onwards.

Lines 112-116: are the data of these other measurements available? Then cite data location, references. Otherwise delete

We agree that the data sources for the accompanying sea-ice physics and biological measurements should be identified more clearly. We have therefore added the relevant references to the manuscript and clarified which observations are used in the present study:

"At all ice stations, a comprehensive sea ice physics and biological program was carried out, comprising measurements of sea ice thickness and snow depth distributions across individual ice floes, extensive ice coring for subsequent analyses of internal sea ice structure and biogeochemical composition, vertical snow characterization, and, at selected stations, the deployment of autonomous ice-tethered platforms **(Arndt et al., 2018, Arndt et al., 2021, Haas et al., 2021). The present study uses the snow pit, SnowMicroPen (SMP), snow depth and sea ice thickness observations described below.**"

Figure 1: The background sea ice concentration is only for 2018. Likely it was different for 2019 and 2021 - relevant to the interpretation of the data? The geographic separation between PS111/PS124 (southeastern Weddell) and PS118 (northwestern Weddell) confounds the FYI/MYI comparison, since regime and region covary almost completely.

We thank the reviewer for highlighting both issues. We agree that the sea ice conditions differed among the three campaigns and have therefore revised Figure 1 to show the mean sea ice extent for February 2018, February 2019 and February 2021 rather than a single sea ice concentration.

The covariation between sampling region and ice regime is addressed separately as a methodological limitation in Sect. 5.3 (line 495 onwards; see our response to the corresponding Discussion comment below).

line 155: calculate a histogram of transect length and present. The calculation of autocorrelation are only statistically significant to about half of the total transect length (see Cliff & Ord 1981 on Moran's I confidence intervals).

We agree that the transect-length distribution and the decreasing statistical support at larger lag distances should be reported more explicitly. The 26 transects included in the correlation analysis range from 19 to 64 m. 17 are located on FYI and 9 on MYI, with mean lengths of 49 and 40 m, respectively. Because this relatively small distribution can be summarized compactly, we report the length ranges and means in the text rather than adding a separate histogram. We also restrict the physical interpretation of Moran's I to lag distances of approximately 35 m.

We have added a description of the present transect lengths to line 165 in Section 2.3: **"For correlation length analysis, only transects with 10 profiles or more were considered. Out of those 26 longer transects, of which 17 transects were on FYI and 9 on MYI, 10 transects between 36 and 50 m and 6 transects were 19-31 m long. Transects lengths on FYI were on average 49 m and on MYI 40 m."**

We also revised line 178 accordingly. The final sentence of Section 2.5, now reads: **"Because the number of transects and profile pairs decrease strongly for larger lag distances, we restrict the physical interpretation to lag distances of approx. 35 m. Values at greater distances are not interpreted as robust spatial signals."**

line 163: why was the unmodified (?) Calonne parameters used? This model is calibrated to alpine snow at rather high snow temperature (average approx -5° C). Explicitly justify this choice and discuss potential biases.

We agree that the choice of parameterization and its potential biases should be justified more explicitly. King et al. (2020) provide sea-ice-specific density coefficients but no corresponding SSA retrieval, whereas our analysis requires a consistent framework for both properties. Snow-pit temperatures in our dataset ranged mainly from 0 to −7 °C, with a mean of −3.3 °C and a mode of −1.7 °C, and therefore fall within the temperature range relevant to the Calonne et al. (2020) parameterization. In addition, this parameterization showed the best agreement with the co-located snow-pit density profiles and stratigraphic observations. Potential differences in absolute values arising from alternative parameterizations are additionally discussed in Sect. 5.3.

The revised text from line 163 onwards reads: "With appropriate assumptions and calibration, penetration resistance data **can be** converted to snow density and SSA following the approach of Calonne et al. (2020). **This parameterization was selected because it produced density profiles consistent with the co-located snow pit density and stratigraphy observations and provides both density and SSA, which are required for the subsequent analysis. Although sea-ice-specific SMP density coefficients have been proposed by King et al. (2020), no corresponding SSA retrieval is available.**"

line 214: "weighted Cross-Entropy Loss function": could you add a brief justification for the weighted version?

We agree that the purpose of the weighting should be stated more explicitly. The inverse-frequency class weights compensate for the unequal occurrence of the snow classes, preventing the more frequent classes from dominating the loss and reducing the risk that good overall performance masks poor classification of less frequent classes. We have restructured the text to connect the class weights directly to the weighted cross-entropy loss.

The revised text reads in line 210 and following:
"To address **the** imbalance between **snow class occurrence**, class weights inversely proportional to their frequency of occurrence were applied in the calculation of the **Cross-Entropy** loss function. **Without weighting, the more frequent snow classes like rounded and faceted would contribute more strongly to the loss function, so that the model could achieve good overall performance while performing poorly on the less frequent classes."**

And in line 214:
"We therefore set up the experiments as a supervised learning, multiclass problem, and used the weighted Cross-Entropy Loss function **accounting for class imbalance** to guide the optimization process.

line 246 -248: repeated statement: "Rare ..." (cf line 209)

We thank the reviewer for pointing out the apparent repetition.
The two passages refer to different aspects: the first identifies the rare or weakly expressed classes excluded from model training, whereas the second explains the possible consequence of this exclusion when the model is applied to the full dataset. We have revised the latter sentence to make this distinction explicit.

Line 246 ff now reads: **"Because** rare or weakly expressed snow types, such as new snow, surface hoar, or thin ice lenses, were excluded from model training due to insufficient sample sizes, **their occurrence in the full dataset** may be locally classified as the most similar major snow type."

line 268: what do you mean by "and the finite vertical support of the SMP signal"?
We agree that the phrase “finite vertical support of the SMP signal” was unclear. We have therefore removed it.

The sentence now reads: **“Such effects reflect the non-discrete nature of snow stratigraphy.”**

line 274: "by a single operator": The authors acknowledge single-operator labelling as a source of subjectivity but do not discuss whether this represents the dominant source of uncertainty relative to the ambiguity of the SMP signal itself. A brief discussion of this hierarchy of uncertainties would strengthen Section 3.3.
We thank the reviewer for raising this important point. We have expanded Sect. 3.3 to clarify that operator-dependent uncertainty and ambiguity in the SMP signal are closely linked and cannot be quantified independently with the available annotations. We also note that repeated labelling by independent operators would be required to assess their relative contributions.

The revised manuscript now reads in line 274 ff: ” ...where expert inter-
pretation of SMP signal characteristics is ambiguous. **Manual labelling by a single experienced operator provided a consistent interpretation across all profiles and avoided additional inter-operator variability. The remaining labelling uncertainty is closely linked to inherent ambiguity in the SMP signal and is expected to be greatest around transitions and where different snow classes produce similar responses. Since operator-dependent uncertainty arises primarily in these ambiguous intervals, the two sources cannot be considered fully independently. Repeated annotations by independent operators were not available, so their relative contributions could not be quantified, but such an inter-operator comparison would provide a useful direction for future work.** Also, while manual stratigraphic analysis may distinguish multiple adjacent layers…”

line 293-297: paragraph repeated from the methods section (cf lines 120 ff)
We agree that the original paragraph repeated information already presented in the Field Data section. We have shortened it to retain only the information needed to define the dataset used in the Results. In response to Reviewer 2, we have also clarified the thermal state of the snow during sampling.

The revised paragraph now reads:
“The results presented here are based on the full set of processed SMP data, including both manually labeled and automatically classified profiles, as described above, excluding the snow pit data. **Although sampling took place during the austral summer, observed snowpack temperatures ranged from -7°C to 0°C with a mean temperature of -3.3°C throughout all campaigns, indicating that the snowpack was not undergoing active surface melt at the time of measurement.** For the analysis, the dataset is separated into first-year ice (FYI) and multi-year ice (MYI) classes […]”

Table 3 between FYI and MYI (density, layer thickness, SSA, snow type fractions) lacks any formal significance testing throughout the paper. A Welch's t-test or Wilcoxon rank-sum test would be straightforward additions.
We agree that the original manuscript relied primarily on descriptive comparisons between FYI and MYI. We have therefore applied Welch’s t-test, which does not assume equal variances, to the comparisons summarized in Table 3, using $p < 0.05$ as the significance threshold. For snow-type-specific density and SSA, the SMP-derived values were first averaged within each individual layer so that vertically correlated measurements from the same layer were not treated as independent

observations. Table 3 now marks statistically significant differences in means in bold, and Figures 4 and 5 have been revised to show the corresponding distributions of layer-mean SSA and density. The relevant Results and Discussion sections have also been updated to distinguish significant from non-significant differences.

The following text was added to the methods in Sect. 2.4:
“For snow-type specific density and SSA, values were averaged over each contiguous classified snow layer. Probability density distributions, and statistical comparisons of snow-type density and SSA were calculated from the resulting layer-mean values. To test for differences in the mean values of the analysed parameters between FYI and MYI, Welch’s t-test was applied, which does not assume equal variances, with a significance threshold of $p < 0.05$.”

line 320 "is closely related to the snow grain size": this is only the case in theory: see the paper by Painter, Thomas H., Noah P. Molotch, Maureen Cassidy, Mark Flanner, and Konrad Steffen. “Contact Spectroscopy for Determination of Stratigraphy of Snow Optical Grain Size.” Journal of Glaciology 53, no. 180 (2007): 121–27. https://doi.org/10.3189/172756507781833947.
We thank the reviewer for this clarification. We agree that a simple one-to-one relationship between SSA and grain size is not generally valid. We have therefore removed the statement and retained only the direct physical definition and relevance of SSA.

The revised manuscript now reads in line 320:
”The specific surface area (SSA) of snow quantifies the interfacial area between the ice crystals and the pore air enclosed within the snowpack and influences key physical snow properties such as radiative transfer and albedo.”

line 347: replace "in the methods in Section 2.5" by "(see Section 2.5, page xx)"
We have revised the cross-reference.

The sentence now reads:
” The spatial autocorrelation of fractional snowpack composition is computed using Moran’s I (**see Section 2.5)**.”
We have omitted the page number because it may change during typesetting.

Figure 7: very interesting data. I guess however that all values and interpretation beyond 30 m are spurious (see comment to line 155).
We agree that Moran’s I values at the largest lag distances are not sufficiently supported for physical interpretation. Based on the decline in the number of contributing transects and profile pairs, we restrict interpretation to lag distances of approximately 35 m and have stated this explicitly in line 535 in Sect. 4.4.

The revised sentence now reads: **“As the number of profile pairs and transects decrease for larger lag distances, Moran's I values at lag distances larger than 35 m cannot be interpreted as robust signals.”**

We have also added this restriction to the length scale discussion in Sect. 5.2
The revised text from line 453 ff now reads: ”… whereas FYI snowpack correlations decayed rapidly over the first tens of meters from 0.4 - 0.6 down to below 0.2 and stayed low **thereafter. However, it needs to be kept in mind, that signals past 35 m cannot be considered statistically robust. Nevertheless,** this contrast indicates …”

line 355 ff: Are this differences statistically significant? Perform a test.

We agree. This point is addressed by the Welch's t-tests described in our response to the Table 3 comment above. The corresponding Results text has been revised to state explicitly which FYI-MYI differences are statistically significant.

Discussion

I wonder why you did not apply standard statistical tests too all your comparisons between FYI and MYI. Are the differences significant? Which differences are significant?
We agree that formal tests are needed to distinguish robust FYI-MYI differences from descriptive contrasts. As described above, we have applied Welch's t-tests to the comparisons summarized in Table 3 and revised the Results and Discussion to distinguish significant from non-significant differences.

The dataset has 19 FYI floes versus only 9 MYI floes, and the MYI floes are almost entirely from a single expedition (PS118). This is a notable limitation, as it could bias the spatial autocorrelation comparisons.
We agree that the unequal number and regional distribution of FYI and MYI floes limit the attribution of the observed differences solely to ice regime. Most MYI floes were sampled during PS118 in the northwestern Weddell Sea, whereas most FYI floes were sampled during PS111 and PS124 in the southeastern Weddell Sea. We have therefore added the coupling between ice regime and sampling region to the methodological limitations in Sect. 5.3 and now make clear that the FYI-MYI contrasts may also reflect regional differences.

The revised discussion from line 495 on now reads:
"The interpretation of Antarctic snow variability derived from SMP measurements and machine-learning-based snow-type classification is subject to several methodological limitations. These primarily affect the absolute quantification of snow properties and the resolution of basal and fine-scale stratigraphy, with implications for interpreting variability across spatial scales.
The sampling design limits the attribution of differences specifically to a single ice-regime because ice-type is partly coupled to the sampling region. Most MYI floes were sampled during PS118 in the northwestern Weddell Sea, whereas most FYI floes were sampled during PS111 and PS124 in the southeastern Weddell Sea. The FYI-MYI contrasts therefore also reflect regional differences. However, because the southeastern Weddell Sea is typically dominated by FYI and the northerwestern Weddel Sea by MYI (Gupta et al., 2025), the coupling between ice regime and sampling region reflects, at least in part, the natural regional distribution of these ice types. The observed contrasts should therefore be interpreted as differences between the sampled FYI- and MYI-dominated environments rather than as region-independent effects of ice regime alone."

Practical recommendation for sampling in future campaigns?
We agree that the practical implications for future sampling should be stated clearly. The discussion therefore emphasizes two recommendations: sampling multiple floes rather than concentrating measurements on a single floe (line 465 onwards), and using a higher within-floe sampling density for layer-resolved properties than for bulk properties (line 480 onwards).

Code availability: TC data policy request that the code must be made public. Without code results are irreproducible, as the machine-learning part depends on subjective choices. This is strong request for the next round of reviews.

We agree that public availability of the classification code is necessary to ensure reproducibility, particularly because the machine-learning workflow includes methodological choices that should be transparent.
We are currently finalizing the repository and will make the code available on GitHub. The repository link will be added to the Code Availability statement.