Response to Review 1

We thank very much Reviewer 1 for his comments that help improving the manuscript. Please find below our point-by-point replies in blue color.

The authors present a local-scale study aimed at characterizing seasonal snowpack evolution with traditional sampling (snow pits), advanced techniques (SnowMicroPen, IceCube, and Tomography) and model application (SNOWPACK). Applying a multi- scale approach, methods are intermixed to construct a daily time series of vertical variation in snow density and specific surface area. The methods are cross-compared to contribute a recalibration of the Proksch et al. (2015) SMP empirical model and to evaluate SNOWPACK simulations. Analysis of the dataset demonstrates clearly how recent advances in field methodology can support model evaluation at very high vertical resolutions. In particular, the details found in Figures 6 and 9, where SMP derived snow properties are introduce at daily time steps, show ability to track snow events and metamorphosis captured in SNOWPACK simulations. Overall, the paper provides a great summary of the campaign results and demonstrates how future model evaluations can benefit from applying similar seasonal framework.

Prior to publication, the paper would benefit from some restructuring to clarify properties of generated the dataset and promote repeatability. These would be meaningful additions to allow application of this work to other environments:

- Recalibration of the Proksch et al. (2015) model uses collocated SMP profiles and density cutter measurements. No distinction is made between the training and testing data when evaluating Eqns 1 or 2. If the authors felt cross-validation was unnecessary, please include this information so that the reader can determine if the skill estimates may be biased (i.e. Test-Train are identical datasets).

→ The entire cutter and IceCube data have been used to "train" the SMP data and to obtain Eq. (1) and (2). The scatter plots shown in Figure 2 show the quality of these parameterizations for the same dataset, i.e. SMP derived data from Eq(1) and (2) versus cutter and IceCube data. The "Train" and "Test" dataset are thus the same. We aimed here at getting as close as possible to this particular cutter and IceCube dataset from our SMP data, and we did not evaluate the obtained parameterizations with other independent dataset. This is why we wrote page 23 line 10: "We would hope that the parameterization Eq. (1) and (2) are generally applicable to an SMP version 4. However, without an independent validation by measurements under different snowpack conditions, it is not possible to state the range of validity of the parametrizations presented here."

We improved Section 5.1 so that it appears more clearly that the test and train data are the same, p10, L6: "This plot [Figue2] shows the observed density from cutter measurements against the SMP-derived density obtained from Eq. (1) and from Proksch2015 for the 15 days for which both data are available (same dataset as used for the statistical modeling). Similarly, the observed SSA from IceCube measurements are presented against the SMP-derived SSA from Eq. (2) and from Proksch2015 for the 13 days for which both data were available (same dataset as used for the statistical modeling). To do so, and as done for the statistical modeling, SMP-derived properties were averaged over 3 cm resolution and SMP and snow pit profiles of the same day were re-aligned with the snow surface and cropped to the length of the shortest profile."

- I'd like to better understand why realignment resulted in improved correlation between the cutter/IceCube measurements and SMP derived properties in Figure 2 as indicated in text (P9 L23). If alignment with the persistent layer defined in Section 6 resulted in a better vertical matching, why were the better alignments not used for the initial recalibration? Throughout the paper, descriptions of alignment could be improved and are noted in the extended comments below.

→ We thank the reviewer for pointing out this issue in the paper. We agree on the confusion about the alignment. For explanation, we would like to point out the difference between 1/ matching of profiles *of the same day* for statistical analysis, and 2/ matching for visualisation of the data such as the evolution of *profile with time*.

1/ Alignment of co-located, co-temporal profiles can be done by using the snow surface. This is convenient and always applicable (unlike using a specific layer) so it is a suitable method to use when doing a local re-calibration of the SMP parameterizations as in our study.

2/ Alignment of profiles when plotting their evolution with time requires another method of matching since profiles are then not co-temporal and do not share a common height/snow surface. One way is to re-align profiles with the ground. For sites showing a ground that is uneven or bumpy this method can however lead to a mediocre alignment. This was the case of the WFJ (ground is uneven) and we found out that a re-alignment based the crust MF-layer offers a qualitatively better match, when looking at plots of Fig 6 and 9 for example. Hence, we chose this alignment method for to present data in Figure 6, 7, 9 and 10.

→ As pointed out by the Reviewer, the first version of the paper showed an inconsistency related to the choice of the alignment method in Section 5.1. Indeed, method 1 (snow surface alignment) was used to develop the statistical model but method 2 (MF-layer alignment) was used to test the performance of the model (Figure 2).

We fully agree with the reviewer that it is confusing. Thus, we modified so that method 1 (snow surface alignment) is now used for both the statistical model and the analysis of the model performance. Method 2 (layer alignment) is only used later in the paper, in the Result part, for time-series plotting purposes.

Modifications throughout the paper have been done accordingly, especially:

- Figure 2 has been redone, based on data re-aligned using the snow surface
- R2 coefficients associated to Figure 2 have been modified. They are slightly better than the previous version (from layer alignment to snow surface alignment: R2 changes from 0.73 to 0.75 for density and from 0.81 to 0.82 for SSA, using Eq 1 and Eq 2 respectively). This actually makes sense as Eq. 1 and 2 have been developed from data aligned with the snow surface.
- Section 5.1 reads now, p10, L6: The performance of the new parametrizations compared to the original parametrizations of Proksch2015 is presented in Figure 2. This plot shows the observed density from cutter measurements against the SMP-derived density obtained from Eq. (1) and from Proksch2015 for the 15 days for which both data are available. Similarly, the observed SSA from IceCube measurements are presented against the SMP-derived SSA from Eq. (2) and from Proksch2015 for the 13 days for which both data were available. To do so, and as done for the statistical modeling, SMP-derived properties were averaged over 3 cm resolution and SMP and snow pit profiles of the same day were re-aligned with the snow surface and cropped to the length of the shortest profile. "

- In the introduction to the Result part, p12, L3, we included now: "To present the evolution of profile properties with time, vertical profiles presented in the following were re-aligned such as z = 0 cm corresponds to the height of the upper boundary of the MF-layer (i.e. the 20151202-boundary). Choosing this layer as a height reference leads to a qualitatively better match than by simply taking the ground as reference (the field site ground at WFJ is uneven)."

- While the layer tracking analysis is meaningful (Fig 8 and 11), description of the SMP tracking method is difficult (if not impossible) to reproduce. An enhanced description of how transitions in SMP signal were used to define layers would be a helpful addition.
→ Section 5.2 "Layer tracking" has been restructured and some reformulation has been made to improve the description of the method. Layers in SMP data were tracked in the same way as in the cutter and IceCube data, i.e. by a manual identification of boundaries in the snow property profiles. The paragraph now reads: "In the measurements data, the layers of interest were defined by the height of their upper and lower boundaries. Boundaries were manually identified by simply looking at the property profiles, looking for sharp and relevant transitions, and recording heights. This step was performed on all the weekly density profile from the cutter and SSA profile from IceCube, as well as on all the daily representative profile of penetration force resistance obtained from the five daily SMP measurements. The identification of layer boundaries was sometimes challenging for weak stratigraphic transitions, e.g. the transition between a layer of fresh snow that fell onto a soft snow layer. To help in such cases, boundaries could be backtracked in time, starting from a profile where the layer of interest is older and its boundaries more clearly detectable. Also, additional information, such as observed height of new snow, was sometimes used to help delineate boundaries."
Besides, we would like to point out that this method only works when tracking well-pronounced layers, so might be hard to use in a systematic way over entire snowpack profiles. To stress this point, we added p10, L23: "The first step is to define which are the layers of interest, knowing that this method is only possible with layers that contrast well enough with their surrounding, so their boundaries can be identified by a significant and rather sharp transition in the vertical profile of snow properties."

- I can confirm that the revised coefficients presented for SMP density are improved over those Proksch et al. 2015 for Arctic snow and snow on sea ice. However, local calibration with our SMP4 unit resulted in quite different coefficients and better RMSE over the use of global parameters (P23 L11). This may make it important to make clear the calibration methods so that they can be easily repeated for different environments or units(?).
We improved the description of the calibration method in Section 5.1, making sure that each step is clearly described.

General comments

P2 L5 – Suggest removing the 'e.g' and revising as 'data back to 1936 in the case of WFJ'.
→ Modified accordingly

P2 L8 – Please be explicit about which properties are characterized rather than using 'hard hardness . . ..'.

→ We modified accordingly; it reads now "grain size, grain shape, hand hardness, and wetness" (P2, L8).

P2 L9 – Remove the period between the citation and sentence.
→ Modified

P2 L14 – Can you clarify what 'non-empirical snow properties' means? This statement is unclear.
With "non-empirical properties" we refer to properties that are physically/mathematically-defined, such as density and SSA, in contrast to grain shape for instance which has no mathematical definition. We modified the term and use "objectively-defined snow properties" (P2, L15).

P2 L15 –Ideally traditional measurements would be supported with metrics such as SSA but the use of the word 'tends' seems to imply this IS a frequent practice. Could it rephrased with the word 'can' or similar?
→ Modified accordingly. The sentence reads now "Concerning the characterization of snow microstructure, the observer-biased estimate of traditional grain size can be replaced by measurements of specific surface area" (P2, L15).

P2 L19 – Capitalize 'IRIS'. Stands for 'InfraRed Integrating Sphere'.
→ Modified accordingly.

P3 L16 – Should the word 'such' be in this sentence?
→ We modified the sentence as "These examples exploit key advantages of the SMP, namely fast profiling for frequent measurements and high vertical resolution, so that profiles are obtained at a considerably finer scale (mm) than possible with traditional means." (P3, L17).

P3 L21 - It feels a bit discouraging to say that the stated goals are dependent on availability of a large dataset with many tools. As a suggestion, removing the word 'only' might lessen the tone. The wording 'cross-validation' could also be problematic as it refers to a specific statistics method. Later the wording 'cross-comparison' (P4 L8) is used which seems to be a better fit.
→ We agree with the reviewer and modified the sentence accordingly as "In the context raised above, the value of emergent, objective snow properties, their potential to replace traditional means in operational snow monitoring programs, and their requirements on temporal and vertical resolutions for model evaluations can be investigated within a multi-resolution and multi-instrument dataset to facilitate comprehensive cross-comparison analyses."

P4 L12 – Degree symbols should accompany the coordinate units.
→ Modified accordingly

P4 L14 – Consider revising the sentence to mention dry snow conditions only once.
→ Modified as follows "We focused on the period from beginning of December 2015 to end of March 2016 to ensure measurements in dry snow condition as required by some of the used instruments. (P4, L17)

P4 L23 – The second element of the measurement area description is squared. Was this intended?

→ Modified as follows "20 m x 8 m".

P7L7 – If the Zuanon (2013) methods were adopted, were any samples compressed to avoid over penetration of the laser? A sentence on how samples were extracted and prepared would be useful for future comparisons where this has become common practice.

→ The extraction of the sample was performed following the protocol described in Zuanon et al. 2013. In addition, we indeed systematically slightly compressed the extracted sample. We included this information in the paper: p7, L10: "Snow samples were very slightly compressed when inserted into the sample holder and attention was paid to have a flat snow sample surface."

P7 L8 – What about uncertainty with low SSA (i.e. DH or FC)? Standard deviation of the measurements in Figure 10a appears to increase with depth and is quite large relative to tomography.

→ As pointed out in the paper Section 7.3, we report a significant and systematic inter-measurement deviation in the SSA estimates. Although we did not study in details uncertainty of SSA measurements in weak layers, our results do not show that biases are more pronounced for DH or FC layers. We did not observe an evolution of the bias with depth. The paper however stresses that these inter-measurement deviations should be further investigated.

P7 L17 – Would like to see an enhanced description of what goes into the profile quality check. Previous studies have described linear trends while measuring in air while others have provided quantitative methods to apply a noise threshold. Which approach was used to determine drift or accept/reject a profile?

→ We improved the description of the SMP data processing. The paragraph now reads P7, L21: "The quality control of SMP force profiles was done manually by rejecting signals with 1) visible trends either in the air portion of the signal or over the entire depth, 2) high noise levels and unrealistic spikes, and 3) frozen tip problems revealed by a force response that appears to be activated only deeper in the snowpack. Most of these problems are caused by wet conditions. The air-snow and snow-ground interface were detected manually to remove air and ground regions from the signal."

P7 L20 – What were the qualities of the data, snow, or study site that determined the profiles could be matched without an offset correction? In section 6 the opposite seems to be stated that spatial variability required compensation to avoid height mismatches (P11 L13).

→ This seems to be a misunderstanding. We improved the description of the SMP data processing in Section 3.4. By offset correction we mean that the value of the force signal itself was not shifted by a given value as it can be sometimes observed (see previous comment). The force signal in the air was very close to zero (manual check) so we did not correct the force signal. This has no link with the height alignment performed in Section 6 for data visualisation.

P7 L29 – Suggest removing 'Reconstruction followed standard procedure' as it's described in the next sentence.

→ Modified accordingly

P8 L10 – May be helpful to indicate the rate of replacement.
→ During the period shown in this study (no melt out), only missing values of either incoming or outgoing SW or albedo values above 0.95 require a replacement. There are no missing values and the latter amount to at most 0.8%, predominantly at sunrise and sunset.

P9 L7 to 11 – Found this a bit of confusing. Is the single 'median' profile being used to train (1)? Perhaps the alignment sentence could be moved upwards in the paragraph to clarify. As it reads now I was not able to determine if 1 profile per pit is being used or if multiple A-S aligned and cropped profiles are being used.
→ We agree with the reviewer and modified the paragraph to describe more clearly each step of the process. It reads now, P9, L16: "The statistical modeling was applied based on a sub-dataset of data from the days for which both SMP and snow pit measurements were available (15 days for density, 13 days for SSA). From each raw force signals, parameters F and L were computed from the raw penetration force profiles over a sliding window of 1 mm with 50% overlap, yielding profiles of F and L with a vertical resolution of 0.5 mm. Note that Proksch et al. (2015) used a sliding window of 2.5 mm, but tests with different window heights (1, 2.5 and 5 mm) did not show a significant impact. Next, for each day, the five daily profiles of F and L of the same day were aligned by simply using snow surface as common reference and a median operation was applied to get one representative profile of F and L per day, called the median profiles in the following. Next, each median profile was averaged vertically using a 3 cm window to match the vertical resolution of the snow pit measurements. Finally, the median 3cm-averaged profiles F and L and the profiles of rho_cutter and SSA_ic of the same day were aligned by using snow surface again as common reference and cropped to the length of the shortest profile. This way, all profiles of a given day are described on the same vertical scale and values of F, L, rho_cutter and SSA_ic can be paired for the statistical modeling, relying on a total of 590 paired-values for density and 497 for SSA."

P9 L15 – Please provide the number of compared measurements to support of the significance test.
→ The number of compared measurements was 590 for density and 497 for SSA. We included that in the manuscript (see comment above).

P9 L16 – This differs substantially from Proksch et al (2015) where coefficients for SSA were not provided. This new equation requires no estimate of density from the SMP, which arguably is better if SSA is the target (minimizes bias from density coefficients and conversion from d0?). No action to take unless the authors wish to highlight the benefit of avoiding the conversion of L_ex to SSA.
→ We would agree with the reviewer that directly estimating SSA and not correlation length via the density as in Proksch et al. 2015, should lead to a better estimates (less errors). In the paper we simply pointed out this difference in the method by writing "Differing slightly from the one suggested by Proksch et al 2015, a regression of the from [Eq 2] was applied to estimate SSA …".

P9 L23 – An enhanced explanation of why the values in Figure 2 do not reflect the error/skill assessment in this section is needed. Related questions: Why does correlation improve when Eqn. (1) was trained on a different set of comparisons? Why was Eqn (1) was not just trained on this better alignment to begin with?

→ As written in an above comment on the same issue, we agree with the Reviewer and modified Section 5.1. In the revised version, values in Figure 2 (and the associated correlation analysis) are based on the same set of data and same re-alignment with the snow surface than the values taken for the statistical modelling Eq 1 and 2. Besides, our statement that using the MF-layer alignment leads to better correlation of values in Fig 2 was a wrong statement. Slightly better R2 coefficients are indeed found when using the snow surface than using the MF-layer for re-alignment (from layer alignment to snow surface alignment: R2 changes from 0.73 to 0.75 for density and from 0.81 to 0.82 for SSA, using Eq 1 and Eq 2 respectively). This makes sense as Eq 1 and 2 have been developed based on a snow surface re-alignment. This has been corrected in the revision and Section 5.1 is now consistent.

P8L29 – Remove one set of brackets around the Eqn.

→ done

P10 L2 – What was the statistical test that showed the boundary transition to be significant? If untested, consider removing the word 'significant'. See comments in the initiate statement about repeatability as well.

→ Boundaries were detected manually just from looking at the data, so there was no statistical test to identify them as well as to confirm that they are "significant". We deleted the work "significant" and it reads now, P10, L23: "The first step is to define which are the layers of interest, knowing that this method is only possible with layers that contrast well enough with their surrounding, so their boundaries can be easily identified by a rather sharp transition in the vertical profile of snow properties." We modified substantially Section 5.2 "Layer tracking", as described in a related comment above, so the method is better described now and can be repeated.

P10 L3 – Given that the boundaries were identified subjectively, will their heights be provided in the published dataset?

→ Heights of the tracked layers will be provided in the database of this study.

P15 L3 – I agree that the information is really useful to show the formation and evolution of these fine features. However, given that Figure 6b has no minor or major ticks for the initial date (Feb 22) it's fairly difficult to identity the feature. Could a label be provided for easy reference?

→ We prefer to leave the figures as is to avoid an emphasis on a single, annotated feature. Since the location is given exactly in the text and the x-axes of the subfigures are exactly the same, the birth of this layer could be easily taken from the SMP image above.

P23 L11 – I can confirm that the recalibrated density coefficients don't produce a best-possible estimates of snow density with our SMP for Arctic snow. Would be very interesting to combine datasets from multiple units to evaluate this uncertainty.

→ We agree that it would be very interesting to compare different sites to test the re-calibrations presented here.

P25L20 –Citation style should be a paraphrase.

Table 1 - List the number of measurements as a separate column. The large number of measurements is really smoothing to highlight! This will also be helpful in the future to frame comparison.
→ The number of measurements has been included in Table 1 (SMP: 100, Cutter: 15 profiles, IceCube: 13 profiles, Traditional: 11 profiles, Stability tests: 8 tests).

Figure 2 - Add N, Rˆ2 and RMSE be added to these diagrams. Having a quantitative evaluation in the diagram provides a quick reference for the reader.
→ Done

Figure 4 – Please provide a colour legend for the grain type classifications even though they are standardized. Additionally, is it possible to provide sub-hatching for the hand hardness levels? It's challenging to determine the level past the first data.
→ Figure 4 has been modified accordingly.

Figure 6/9 – Has the SMP data been smoothed or aggregated? This does not appear to be mentioned in text but Figure 10 shows variability in SSA absent in Figure 9 at the 1 mm scale.
→ We used the same data with a resolution of 0.5 mm for the seasonal evolution plots as well as for the vertical profile plots (7 and 10).