We thank the reviewer# 1 for the review and the constructive comments! All reviewer comments (in italics) are addressed below.

General comments

My largest concern is what portion of the data set was used to calibrate the model (Section 4.3) vs evaluate the model in Section 4.4.1. The description of the methods in Section 4.3 wasn't clear (see further points made below) and seems to suggest that 80% of the data was used to calibrate the model and then the results in Table 1 and Figure 10 were obtained using 100% of the data (or was it the remaining 20%?). Please clarify these points. If this is indeed the case, I don't think it's appropriate to refer to the results in Section 5.4.1 as an "evaluation."

We first splitted the data set in 25 sub pools for all available domain sizes L by using $L \geq$200 m, $L \geq$240 m etc. We then took 500 random samples from each of the 25 sub data pools but each sample comprised only 80 % of the sub pools data. Fitting each sample separately revealed the variance among the samples increasing with L (colored lines in Figure 9). Instead of fitting over 100 % of the data with $L \geq 200$ m, we fitted over the ensemble median of all scale-dependent parameters derived from each of the random sample. This procedure should ensure sufficient robustness of the derived scale-dependent parameters $c(L), d(L)$ (Eq. 3).

It is however true, that we did not compile an additional independent data pool for an independent evaluation and that we also did not split our compiled data set in development and validation subset. Given that we compiled and preprocessed a vast snow depth data pool covering large variability in snow climates and topographic characteristics, we were however able to perform a scale- as well as geographic region-dependent evaluation to reveal the spread in errors we can expect when applying the parameterizations on an individual independent data set (Figure 11).

We have rewritten Section 4.3 also considering the reviewers specific comment and suggestions below and further adapted Section 5.3.

Since the data set wasn't split for calibration and validation (line 364), I think the evaluation by region is particularly important and it will provide a better sense of how differently new data is likely to perform compared to pooled performance of the calibrated model. You state this explicitly on line 389, but it could be emphasized more. It also indicates that climate models would benefit from observations across as broad a selection of alpine regions as possible.

We agree that the scale- as well as region-dependent evaluation allowed the assessment of performances we can expect when applying the parameterizations on independent data sets. We now additionally mention this in the abstract. We also extended the available discussion on that in line 364. We did not add it to the conclusions, since it was already mentioned in the conclusions of the manuscript (line 437-441 of first submission).

While this study uses data from across a greater number of regions than your previous work, it may still not sample enough regions to be applicable to global mountain snow. Is it possible to provide a sense to the reader (perhaps in the discussion) of how well the distributions shown in Figure 5c represent values from global snow-covered mountain ranges? What about how much variability is there in ξ across your pooled data vs globally? Would interannual variability in snow in a particular mountain region affect the values calculated for σ_{HS}? I'm guessing it wouldn't if the snow and terrain is deep but perhaps in particularly low water years it might.

The variability in summer terrain characteristics in our pooled data is rather large. Spatial average slope angles range from 4° to 60° (μ from 0.05 to 1.22), terrain correlation lengths ξ from 6 m to 775 m and L/ξ-ratios from 3 to 40. Thus, typical summer terrain characteristics captured by coarse climate model grid cells are well represented. We now discuss this in Section 3.1.

While interannual variability in snow depth in a particular mountain region might affect the values calculated for σ_{HS}, when normalizing σ_{HS} with the corresponding spatial mean HS the regions order very closely to each other (cf. Figure 6a). This was also a finding of [1] who therefore introduced spatial mean snow depth in the parameterization for σ_{HS} as a climate indicator which we followed here.

While you cite Helbig et al., 2015 and Essery and Pomeroy, 2004 in the introduction there is a lot of context from those two papers which is essential to understand this current study, and I think this manuscript would benefit from including more thorough explanations and context from them. I've tried to specify several examples in the comments below which I think would help but there may be others.

Thanks for pointing that out. We address all specific comments below.

Specific comments

L100-106: I think the terminology "peak of winter fSCA" parametrization causes some confusion here. I recognize that σ_{HS} is being calibrated by mid-winter (March) data, but from your previous work (Helbig et al., 2015) you are expecting that the parametrization will apply during accumulation and melt as well. Furthermore the fSCA you describe in eq 2 is calculated from Helbig et al. (2015) assuming melt events resulting in a SCD curve. This paper never mentions the accumulation season or melt until the very end at line 445. Unless you have changed your opinion on whether or not the formulation used here and in Helbig et al., 2015 can be extrapolated outside of the peak season, I think you should mention this at some point in the introduction. If you are truly concerned that it can't be extrapolated outside of the peak-snow season I think you need to justify its potential use.

We agree this description has been confusing. The terminology "peak of winter" actually only refers to the σ_{HS} parameterization. We went over the manuscript to correct any further inconsistencies.

By tracking snow depth over the season, a seasonal $fSCA$ model implementation of Eq. (1) and (2) is possible. We will present this in a different article, which will be submitted soon. We rephrased the last paragraph of the conclusions to improve the seasonal $fSCA$ outlook.

A visual representation of L, ξ, dx,dy would be helpful. E.g. representational lines on Figure 3, another panel in Figure 3, or at least explicitly refer the reader to previous work (e.g. Fig 2 of Helbig et al., 2009).

In Section 3.3 we introduce the summer terrain parameters. For a visual representation we now refer to Figure 2 of Helbig et al. (2009).

Equation 1: I think it would be helpful to state that equation 1 has been shown to reasonably parametrize fSCA for both nonmoutainous and mountainous regions, while the relationship in Eq 2 is derived using only mountain data (at approx. seasonal peak, if you'd like). And/or state this distinction in the introduction (e.g. at line72, "While the standard deviation of snow depth introduced by Essery and Pomeroy did not depend on subgrid terrain characteristics, the formulation shown in Equation 2 was introduced by Helbig et al. (2015) in order to better model Equation 1 in mountainous terrain.

Thanks. We now point this out in Section 3.3 as well as in the introduction.

You refer the reader to [1] at the start of Section 3, but it's not clear if this is to describe the domain sampling procedure, or even if the same method used in the 2015 paper is used in this manuscript. The 2015 reference specifies 12 domain sizes between 50 and 3000m were randomly sampled. In this manuscript there are 20 bins shown on Fig 4. Please provide additional information on how each data set/scene is decomposed into domain sizes.

Thanks for pointing this out. We added more details to Section 3.1 and now show the numbers for the full range of the 41 domain sizes in Figure 4 instead of in bins. The geographic site sampling procedure used here is in regular grids (per various domain size) per geographic site.

L207: The symbol HS is being used to represent both the domain-average snow depth and the high-resolution observed snow depths at fine scale resolution (e.g. figures 5a, 6a, lines 73-100). I suggest you distinguish these uses.

Though high-resolution observed snow depths at fine scale resolution are actually also spatial mean values over a much finer sampling, we agree the symbol usage might have been a little confusing and rephrased or clarified the usage of the symbol HS where applicable.

L240: Do you sample the autocovariance in each domain 40 times? Why do you single out L=3km and then say you find inflection points for each domain size L?

We derived a total of 40 autocovariances from the available 3 km domains. Unfortunately, the description of the results on spatial autocorrelation caused some confusion. We improved Section 4.1 to make it more clear.

L253: This is the first time you use sqS, and σ_{sqS}. Again for context it would be good to mention that you are repeating previous analysis that established μ and ξ/L as the most important correlates, and you are examining these two variables to compare to results from Skaugen and Melvold, which you do in the Discussion section.

All candidates for terrain parameters are introduced in the Methods Section 3.4. We extended this description to make it more clear that these parameters are later evaluated.

L267-280: While I understand the results shown in Figure 9, I couldn't understand your description of the methods used to produce them. I suggest removing/reordering the first 4 sentences from this paragraph. The discussion of domain size dependent fitting only confuses things when you then discuss the fit to the entire pooled data set. I suggest beginning the paragraph with "Fit parameters were first calibrated for the entire data pool yielding c = 0.6589 (±0.0037) and d = 0.5638 (±0.0043) with the 90 % confidence interval... ... ... larger than the previously derived constants a, b in Eq. (2) (cf. Figure 9). For each step-wise domain size between 200 m to 5 km scale-dependent parameter values are also fit from the data (cf. individual colored lines in Figure 9)." At this point please provide a more complete description of the subsampling used to derive c, d for each step-wise domain size. What does 80% mean? Are the parameter values fitted from all the data within a randomly chosen domain of the appropriate size and this process is repeated 500 times? For domain sizes above 1km there are ¡500 domains total so are the same values just replotted? After this description, you can continue on with the discussion of how the parameters increase with L and the subsequent fitting of c(L) and d(L)

We have rewritten Section 4.3 considering the suggestions and also adapted Section 5.3. We split the entire data pool in 25 sub pools for any available domain size between 200 m to 5 km (cf. Figure 4). Thereby, each sub data pool included all domains larger or equal to the corresponding domain size, i.e. $L \geq$200 m, $L \geq$240 m etc. ... From each of the 25 created sub data pools, we randomly took 500 sub samples where each sub sample comprised 80 % data of the sub data pool. Each of the sub sample per sub data pool was unique. Scale-dependent parameter values were derived for each of the 500 sub samples drawn from each of the 25 sub data pools (cf. individual colored lines in Figure 9). ... By fitting the ensemble median of all scale-dependent fit parameters (dark blue dots in Figure 9) across all domain sizes between 200 m to 5 km, we obtained scale-dependent parameters $c(L)$ and $d(L)$.

Fig 9: Please use a different description on the legend in place of 'f(L) – Eq. (3)' which can read as 'f(L) minus Equation (3)'.

Figure 9 is changed.

Section 4.4.4: Does the different choice of domain aspect ratio (square vs rectangular) affect the differences described in this section?

The domain aspect ratio is not important. The mean domain size L is derived from $L_x = 500$

m and L_y = 1000 m resulting in L = 750 m. The parameterization of [2] was developed for mean domain sizes L of 750 m. This means that, unlike the parameterization of [1], the parameterization of [2] wasn't developed across spatial scales. We added this to the discussion in Section 5.4.4.

L335-337: Please rephrase or clarify: "at these scale lengths." I think you are saying something like "Above scale-lengths of 200m all three effects (precip/wind/radiation-interactions) come into play, while we think there are different physical effects which establish the breaks at 20 and 60m," but please confirm. Also consider rephrasing "scale-independent parameterization", since the parametrization incorporates scale information from the sub-domain terrain parameters as well as in the constants (c(L), d(L)). Perhaps something like "The results presented here indicate that the model described by (eqs. 1 and 2) is a reasonable fSCA parametrization in mountainous terrain for spatial scales between 200m to 5km."

Yes, you are right and we rephrased the paragraph. While we got rid of "scale-independent parameterization" here, we kept it elsewhere when required to make clear that the empirical parameterization is not developed on fixed scale lengths but rather for a broad range in spatial scales.

Given that you are aiming to have this used as a fSCA parametrization in climate models which can still use grid scales as high as 50-100km please comment on the extrapolation of your results substantially beyond 5km.

You are right. We now comment on the applicability of the parameterization for larger grid sizes than 5 km in Section 5.1.

"The results presented here indicate that the model described by Eq. (1) and (3) is reliably parameterizing the spatial snow distribution shaped by the longer range precipitation, wind and radiation interactions with topography for spatial scales between 200 m and 5 km. Above the detected scale range of around 200 m not only the spatial autocorrelations approach zero (Figure 7), but normalized σ_{HS} clearly start levelling out as well as the normalized variability of σ_{HS} among similar sized L (Figure 6). Thus, even though we could not verify the $fSCA$ parameterization for length scales larger than 5 km, we believe that as long as grid cell mean slope angles are larger than zero, Eq. (1) and (3) might also hold for larger grid cell sizes than the 5 km."

L343: Do you mean "for spatial scales between 0.5km and 1km"?

[2] investigated correlations for spatial snow depth distributions in grid cells of 0.5 km x 1 km. We rephrased this.

L357: "Furthermore, larger (about 17% and 45%, respectively) but overall consistent constant fit parameters were obtained compared to those from [1] based on a more limited number of data sets and just two geographic regions (cf. a, b...

Rephrased.

L411-413: I'd suggest that the appropriate standard for how different parametrizations perform is the range of MPE seen among different regions, not the difference between your previous calibration and the current one.

We completely agree, which is why we presented our model performances (by MPE) for the different regions (Figure 11). However, we additionally evaluated the empirical parameterization of [1] on the large evaluation data pool of the present study to investigate if the underlying functional form of our parameterization is reliable and if the parameterization works in independent geographic regions. Only one out of the 11 data sets was used for developing the empirical parameterization of [1].

Technical comments

L185: 3m to 5km Corrected.

Discussion, several places: "origin" as a verb - "originate" Thanks, corrected.

L379: I'd suggest splitting this sentence in two. Done, thanks.

L400: rephrase Rephrased.

L395: "decrease from 80cm..." Done.

L409: "sensitivity"? Corrected.

*References

[1] N. Helbig, A. van Herwijnen, J. Magnusson, and T. Jonas. Fractional snow-covered area parameterization over complex topography. *Hydrol. Earth Syst. Sci.*, 19:1339–1351, 2015.

[2] T. Skaugen and K. Melvold. Modeling the snow depth variability with a high-resolution lidar data set and nonlinear terrain dependency. *Water Resour. Res.*, 55:9689–9704, 2019.