Author responses to reviewer 1 comments on:

"SPASS – new gridded climatological snow datasets for Switzerland: Potential and limitations"

by Marty et al. in *the Cryosphere*

We thank the reviewer for the time to assess our work and for the valuable feedback and suggestions. We respond to each point of the reviews below. The reviewer comments are highlighted in blue while our responses and comments are kept in black.

Marty et al. present an evaluation of spatially gridded datasets of snow cover over Switzerland, with high spatial resolution (1km) and long duration (60+ years). They evaluate different datasets, with and without assimilation using ground observations using different metrics across elevation, and also compare long-term trends. The manuscript falls well within the topic of TC. It is well written and results are discussed critically, for which I compliment the authors (especially sec 3.4 on limitations is great). However, there are a few issues that remain unclear.

Major points:

The novel contribution of this study with respect to previous studies is not completely clear – from the literature review in the intro it seems that a comparison of the "new" dataset has already been performed (Scherrer et al 2024) and the dataset itself has been created and validated in Michel et al. 2024. Please highlight the differences between the past studies and the current one more clearly, as well as the novelty of this particular study.

We agree to have been biased in this regard (this point was also mentioned by reviewer 2): Michel et al. (2024) introduced the quantile mapping method, which was used to create the new SWE datasets. They also provided some first validation of these datasets. Moreover, Scherrer et al. (2024) compared preliminary versions of these SWE datasets to other gridded SWE datasets. The main novelty of our study is the creation of the corresponding gridded snow depth datasets and thus possible comparisons to station-based measurements based on elevation and time aggregation dependent analysis as well as elevation dependent trends. This procedure allows to provide quantitative information on time aggregation- and elevation-dependent uncertainties, which was missing so far. We will change the manuscript text accordingly.

I assume one novelty is the elevational analysis and temporal aggregation unit analysis. While the first one is evident and well explained, the second seems very minor to me after reading the paper. First, all of the figures are in the supplement, and often the results are presented as similar across temporal units. Moreover, the motivation behind the different temporal units is not evident. And also why daily was not a temporal unit.

You are right, there is less emphasis on temporal aggregation than on elevational analysis. The main reason is the fact that elevation dependent differences are larger than temporal aggregation dependent differences and that the main message is always the same: "lower performance for smaller time aggregations". Moreover, when describing Fig: S2, S3, we also emphasized the differences due the shorter time aggregations in absolute and relative terms. We agree that the corresponding description of Fig. S4 is so far is only brief and could easily be extended, which will be done in the revised manuscript.
The motivation behind the used temporal units was given by the following facts: Climatological analysis are often provided by yearly or monthly reports and we wanted to assess the uncertainty of the new snow products with the goal to include them in future such reports. Moreover, knowing about the need for timely (near real time) public information about possible current extraordinary situations, we also assessed the weekly aggregation. Daily aggregations were by purpose not assessed as quantile mapping is not expected to do more than a climatological bias correction, meaning biases at short time scales, like on a single weeks or days are not necessarily corrected. This is demonstrated by the fact that already the analysis weekly values revealed relative errors of more than 100 % at low elevations. Additionally, model results of the current day are only available after 15:00 local time, which is too late for timely public information. We will add this information to the revised manuscript.

On the other hand, another important factor is seasonality. Did you consider how error metrics vary across the snow season? E.g, if they are constant or increase/decrease towards the end of the season?

Yes, we did specifically analyze how the error metrics vary across season. The beginning and end of the snow-covered season has generally a lower performance than mid-winter also at higher elevations because the situation is similar as at low elevations during the entire winter. This implies the transition seasons between no-snow and snow also at higher elevations have the same potential problems as at low elevations during the entire winter. These problems involve among others high spatial variability and no information on the soil temperature, which is decisive for the survival of potential snow fall. But since our focus was between November and April this seasonality issue does only affect the 1000 and 1500 m elevation band.

While the authors have a great choice of evaluation metrics, including MAAPE, which seems very interesting, there needs to be some consideration of whether the metrics (and the associated figures and statistics) refer to spatial, temporal, or their combined spatiotemporal

variability. For example, L176-180 is unclear (and also not really relevant to know computational details like how your array looks like). It would be great if you could identify what the metrics and variability refer to, i.e., where you average over space, time (years or other), or where you show variability across space or time or both. Also the order of calculation matters, so if you first do metrics, then average (e.g., over weeks, or over gridcells), or first average and then do metrics. Less for bias, but significantly for all other metrics. The ordering of calculations is not completely clear from the manuscript.

This aspect indeed has potential for improvement. We always first averaged over time for each elevation band. This means that the boxplot shows the variability across space for each temporal aggregation in each elevation band. This means in the case of the model-to-model intercomparison (Fig.4) the boxplots were created based on the number of grid points per elevation band (as listed in Table S2). In the case of model-to-station intercomparison (Fig. 5), the boxplots were created based on the number of stations per elevation band (as listed in Table S2). We will add this missing information to the revised manuscript.

Minor points:

Abstract is sometimes confusing. It mentions two datasets, named old and new, which remains unclear. My suggestion is to try to make the abstract as self-contained as possible. Also, some numbers would make it less vague.

We will reformulate the Abstract in the revised manuscript to make it more coherent and also add some numbers to make it more concrete. BTW, we do not find "old" anywhere in the abstract.

L40 "many applications", please provide some examples.

We will extend the sentence with something like: "…limit their usefulness for all climatological applications which go beyond station-based analysis, i.e. which are dependent on temporally and spatially consistent long-term regional datasets to calculate trends or elevation dependent anomalies.

Besides the general intro to snow, the introduction focuses exclusively on the history of gridded snow datasets in Switzerland. Since the topic of spatially gridded snow cover datasets is not trivial and can be tackled (in theory) in multiple ways (stations, remote sensing, modelling), maybe a broader introduction into gridded climate datasets and gridded snow in particular might be useful for readers.

We understand the argument and will add some sentences, which give a broader introduction to gridded (snow) data sets.

Introduction and Methods are somewhat mixed, since the used models/datasets are presented in the introduction, but then in the method the models/datasets are not described further. I guess there are other studies presenting this in detail, but for completeness, I suggest including a brief summary of the key model characteristics and meteorological input for the different datasets.

We believe that the used models/datasets are also shown in Fig. 1 in the Methods section, but we will streamline the Method section and add a brief summary of the meteorological input and the models used in the revised manuscript.

L114 some reference would be useful

We assume that the reviewer refers to "The data of these stations have been carefully quality-controlled and gap-filled in separate steps." The details of the applied methods have changed over time. In general, these are physical threshold checks, as well as temporal und spatial consistency checks (we will add this information to the text) among others involving the relationship between depth of snowfall and snow depth. According to MeteoSwiss and SLF experts unfortunately there is no publication available documenting the various QC methods, but we added a reference for the gap-filling.

L135 unclear if for the climatological analysis the reference period was 1991-2020 or 1999-2023.

We agree that this paragraph is confusing and will change the last sentence to: "When investigating performance differences between OSHD-CLQM and OSHD-EKF, as well as to have enough in-situ data (Fig. 2, Table S2) available in the different elevation bands (mean per elevation band is 20 stations, minimum 14 stations, maximum 34 stations), the evaluation is based on the period 1999-2023."

L151 so relative trend is based on the Theil-Sen slope, but relative to what? Theil-Sen intercept, mean over the whole period, something else?

This sentence will be deleted in the revised manuscript as we do not show any relative trend results.

Sec 2.4 did you compare the difference in trends between CLQM and Comb?

Yes, see section 3.3.1: There we write: “The OSHD-Comb trend magnitude is marginally weaker than the OSHD-CLQM trend magnitude and thus closer to the station-based trend magnitude for all investigated elevations with the exception of the 2000 m band.”

Related, has the meteo input (the temp and precip grids) been tested for homogeneity? Otherwise, I guess the snow trends could reflect input dishomogeneities as well…(ok this comes around L405...)

No, we did not check the meteo-inputs for homogeneity, as they are potentially inhomogeneous due to their creation method, as described in section 3.4.

Sec. 2.5 Since you use relative errors, I guess relative bias would also be interesting? While absolute bias increases with elevation, relative one should decrease, no?

Yes, this is true for SWE, but not necessarily for HS. We did not specifically analyze relative bias, but when combining the results of Fig. 6 with Fig. 8, we can calculate that the relative bias for HS is decreasing with elevation only until about 1500 m, i.e. if the absolute bias is small (see Fig. 5).

L190 “because HS has been derived from SWE” but this is true for all elevations.

Yes, this is true for all elevations, but the used SWE2HS algorithm sometimes seems to underestimate HS at higher elevations (L 247).

L210 “boxplots consisting of the 25 yearly values” but there more points than this in the boxplots?

Thanks for this tip, this information was wrong. Since the boxplots show the spatial variability (see our answer to your last major point), they were created based on the number of grid points per elevation band (as listed in Table S2) in the case of the model-to-model intercomparison (Fig.4). In the case of model-to-station intercomparison (Fig. 5), they were created based on the number of stations per elevation band (as listed in Table S2). We will correct this information in the revised manuscript.

Fig 4: very unusual choice for the whiskers to go from 5th to 95th percentile. Why not the standard boxplot variant with 1.5*IQR from the box edges (up to the largest value, if within range)? Also because your choice highlights a lot of "outliers", which are not really outliers, but continuous variability, in my opinion. One could also do the other standard whiskers that go to min and max.

Thanks for this statistical help. We used the 5th to 95th percentiles as we believe that this is easier to communicate to practitioners. However, we agree with your argument about the outliers and now recalculated all boxplots with standard 1.5 IQR whiskers. This did reduce the number of outliers. We will change the text and exchange the figures in the revised manuscript.

Fig4 and 5: for bias a line at y=0 would be useful. Or some light background grid lines in all panels.

Good idea, which will be included in the revised manuscript.

Fig5: if the focus is on the comparison between CLQM and EKF, it would be useful to show the boxes side-by-side (e.g., with different fill or line colours) and not in separate panels. If the focus is on comparing by elevation, it's fine like this.

Yes, we chose this representation because our focus is the comparison by elevation band.

Fig. 6: a polynomial of first degree should be a straight line…

Thanks a lot for the hint, this indeed wrong. It is a polynomial fit of second degree. We will change this in the revised manuscript.

Besides Fig6, is it possible to produce the same figure as Fig5 also for the non-assimilated stations, or put them side-by-side to compare the performance metrics between assimilated and non-assimilated stations?

Technically this would be possible, but it would not be comparable to Fig. 5 as the non-assimilated stations cover any time period between 1999-2023 (median 12 years, minimum 5 years) instead of 25 years as in Fig. 5.

L278-284 this paragraph feels a bit off in the current section. Or what does it refer to? To all stations assimilated and not? It also contains something on climatology and trends…

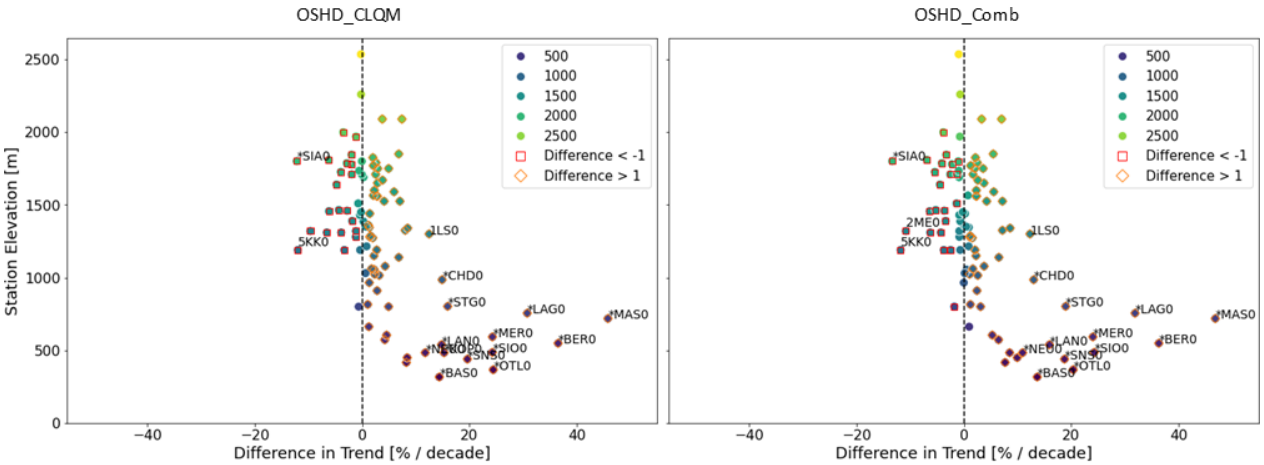
We agree. This paragraph rather belongs after the explanations to Fig. 5. Therefore, we will move and rephrase it in the revised manuscript.

Fig7 could you please increase resolution or use vector graphics? It's not possible to zoom in easily.

You are right: We will increase the resolution of Fig. 7 in the revised manuscript.

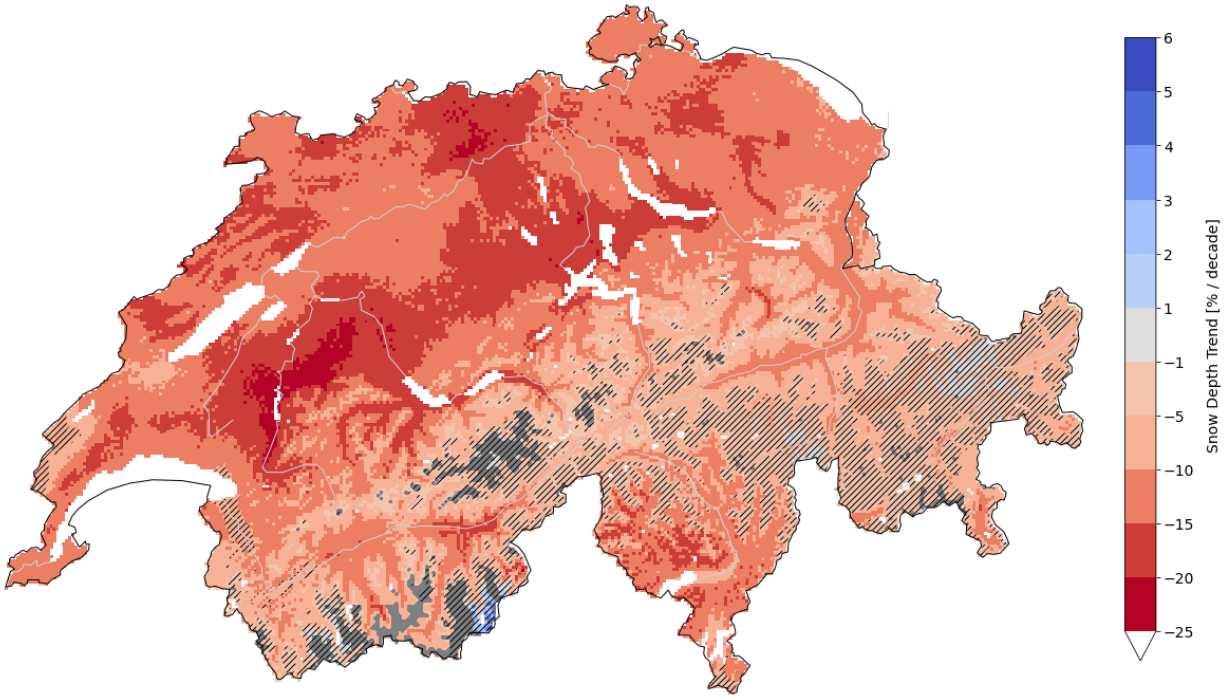
Fig9, by chance, do you also have a relative trend figure of this?

We are not sure what you exactly mean. In our opinion the most meaningful way to show Fig. 9 in relative terms is the difference with respect to the mean snow depth at each station (see below figure, where stations that show a difference greater than ± 10%/decade are labeled). As expected from Fig. 5, the relative difference shows large values at low elevation. On the other hand, above 1000 m only 4 stations show differences larger 10 %.



Similarly, I guess a relative trend map (Fig10) could also be useful?

As shown below we additionally produced a relative trend map based on Fig. 10. Similar as above, the largest relative trends are shown at low elevation (mainly Swiss plateau), which reach almost -25 % per decade. In the Alps typical relative trends are between -5 and -10 % per decade.



L427 unclear, might be resolved with a more detailed method section description

This refers to the fact that the daily degree factor of the temperature index model is seasonally dependent, but the same for all pixels, i.e. derived from the mean of all station measurements in the last 25 years but applied to the entire 60 years. We agree that this was unclear and will add this information in the methods and rephrase this section in the revised manuscript.

L429 Why not use tmin and tmax instead of tmean? It's also much more stable over time, considering the 60year period.

We agree, but gridded data of tmin and tmax is only available back to 1971 in Switzerland.

L461 please repeat the reasoning from Michel et al. 2024 here, shortly.

Generally, quantile mapping is known to not necessarily correctly preserve extreme events. The applied quantile mapping method can by definition not really capture extreme values, as they are corrected according to the correction of the 99th quantile. We will add this information to the revised manuscript.

Conclusion could be a bit more general. E.g., what are the implications of the results? Can the fairly simple degree-day model be trusted or is the station data maybe more accurate? What are use cases of such a dataset in climatology, hydrology, ...?

The question if station or SPASS data can be more trusted depends on the application and region in focus. Generally, as shown in our study, at low elevation station data is more thrustful. At higher elevations SPASS data from larger regions and longer time aggregations have the advantage of being less locally dependent and available in pre- and after season periods (early autumn and late spring). Anyway, we will add some more implication related sentences to the Conclusion in the revised manuscript.

Not necessarily for this study but have you considered evaluating the grids with remote sensing? Like with optical-derived snow presence?

The SWE dataset has already been compared with remote sensing data (doi.org/10.3929/ethz-b-000652501).

PS: Thanks for the review invitation, I had SPASS reading the paper, or Spaß, as we spell it here :)