Response to the Referees

We thank the Editor and the two Referees for their careful reading of the manuscript and their constructive comments. This document provides point-by-point responses to the Referees and explains how each comment was considered in the revision, indicating also the relevant changes in the manuscript (the lines in the replies refer to the revised marked document). The revised version is accompanied by a new Supplement containing the additional material generated during the revision.

Response to Reviewer 1

General comment

For the evaluation of the land surface datasets against IT-SNOW (for SWE and snow cover duration) or SCIA (for 2-m temperature and precipitation), the authors look into spatial biases between the different products, but do not evaluate the statistical significance of these biases, which make me wonder whether the biases presented are statistically significant or not. As the authors work with a relatively short period (2010-2024) and a large number of grid points, I would suggest assessing the statistical significance of SWE, snow cover duration, temperature and precipitation biases using a two-tailed students T-test combined with a False Discovery Rate (FDR) approach following Wilks (2016).

Reply: We would like to thank the Reviewer for this highly pertinent and constructive observation, which significantly enhances the statistical robustness of our findings. We fully agree that evaluating the statistical significance of the spatial biases is crucial, particularly given the relatively short evaluation period (2010–2024) and the large number of grid points involved in the analysis. Following your valuable suggestion, we have carried out the analysis assessing the statistical significance of the biases for all the variables using the proposed methodology, applying a paired two-tailed Student's t test to the annual values and screening the resulting grid-cell p values with the Benjamini–Hochberg procedure at αFDR = 0.10, and we have updated the results section accordingly. Specifically, rather than overlaying stippling directly onto the main manuscript maps—which would severely compromise their readability given the spatial resolution (9 km), the small pixel size, and the compact format of the figures—we have provided dedicated additional materials. These are included in a new Supplement and consist of expanded versions of the corresponding figures explicitly showing the spatial distribution of the statistical significance (Figs. S1 to S4). The methodology is described in the revised manuscript (ll. 351-360): *"For each product and indicator, the statistical significance of mean annual SWE and SCD biases was assessed at each common-grid cell by applying a two-tailed Student's t-test to the time series of annual product–reference differences, testing whether their mean differed from zero. As discussed by Wilks (2016), interpreting numerous grid-cell tests without accounting for multiple testing can overstate the spatial extent of significant differences because some results are significant by chance alone. The resulting p-values were therefore evaluated using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), which controls the false discovery rate (FDR), defined as the expected proportion of false rejections among all rejected null hypotheses. The procedure was applied separately across the evaluated grid cells for each product and indicator. An FDR control level of 0.10 was adopted following Wilks (2016), who recommends using twice the target global significance level to obtain an approximately 0.05 global test level in fields with moderate to*

strong spatial correlation. Statistical significance is shown in Supplement to preserve the legibility of the main bias maps." and (ll. 391-394): *"In addition, the statistical significance of the mean annual precipitation and temperature biases was assessed using the same paired t test and false discovery rate (FDR) procedure described in Sect. 4.1, with SCIA as the reference. To preserve the legibility of the bias maps in Appendix Figs. B1 and B2, the results of the FDR-screening are shown in the corresponding expanded Supplement (Figs. S3 and S4)."*

The results of this new analysis confirm our initial findings. The statistical significance is generally consistent with the interpretations provided in the original draft; therefore, the core message and the overall conclusions of the Technical Note remain unchanged, but they are now supported by a much more rigorous statistical framework. We sincerely thank the Reviewer again for pointing this out and for helping us improve the quality of our manuscript.

Specific comments (SC)

SC1: L124 and Figure 1: Could the authors add more geographical information to the map of Italy (other than "Alps" and "Apennines") to help readers locate places more easily? For example, could they indicate the location of the Ligurian sector, Sardinia, Sicily, the Strait of Messina, etc.? I know where these places are, but I imagine that not all readers do, so it would be helpful to include a little more geographical background information.

Reply to SC1: We agree that adding more geographical references to Figure 1 helps the international readership better orient themselves and follow the regional differences discussed in the text. In the revised version of the manuscript, we have updated Figure 1 to include additional geographical background information.

SC2: L146-147: Could the authors be a little bit more specific on how the precipitation-phase partitioning is implemented in IT-SNOW?

Reply to SC2: We thank the Reviewer for pointing this out. Phase partitioning in S3M relies on both air temperature and relative humidity, which generally provides more reliable results than using air temperature alone (Zhang et al., 2017). Rain and snow proportions are derived based on the simple method by Froidurot et al. (2014). The separation is performed at an hourly time scale. In the revised manuscript, we have included these additional details (ll. 160-162): *"At each hourly time step, total precipitation is partitioned into rainfall and snowfall based on air temperature and relative humidity, which is more reliable than temperature alone (Zhang et al., 2017), following Froidurot et al. (2014)"*.

SC3: L148: Is the hybrid temperature-index and radiation-driven melt approach in IT-SNOW similar to the enhanced temperature-index approach of Pellicciotti et al., 2005 (https://doi.org/10.3189/172756505781829124)?

Reply to SC3: Yes, the approach in S3M was inspired by the approach of Pellicciotti et al. (2005). The only, significant deviation is the addition of two modulation parameters that reduce snowmelt based on antecedent temperature. This was done to include in the approach some forms of cold content, that is, thermal inertia of

snow. While a comprehensive description is available in Avanzi et al. (2023), we agree that briefly clarifying this connection benefits the readers. Accordingly, we have included these details and the reference to Pellicciotti et al. (2005) in the revised text (ll. 164-167): *"The melt formulation follows Pellicciotti et al. (2005) in separating radiation- and temperature-driven components, but S3M modulates the coefficients of both components using mean air temperature over the preceding 10 d, providing a parsimonious proxy for snowpack cold content and thermal inertia (Avanzi et al., 2022).".*

SC4: Appendix Figures B1-B3: For the middle row figures (representing the temperature biases) I recommend reversing the color bar. In the current state, the colors representing the Tbias are somewhat counterintuitive in my opinion, i.e. warmer is purple and colder is red whereas a positive SWE bias is blue (usually associated with colder temperature conditions) and a negative SWE bias is red (usually associated with warmer conditions).

Reply to SC4: We completely agree that the previous colour combination for the temperature bias might be counterintuitive for the readers. In the revised manuscript, we have updated the colour scale. Specifically, we have reversed the colour bar, using red to indicate temperature overestimation (positive bias).

SC5: L358-359: Another potential explanation for the overestimation of snow depth in the higher parts of the Alps could be that (according to Orsolini et al., 2019; https://doi.org/10.5194/tc-13-2221-2019) data assimilation of snow cover (using IMS – Interactive Multisensor Snow and Ice Mapping System) was discontinued in ERA5 above 1500 m a.s.l. This could explain the snow biases in ERA5 itself, but as snow biases also affect ERA5 temperature (via snow-albedo feedbacks) I can imagine that the negative temperature biases propagated from ERA5 into ERA5-Land also have affected the snow conditions in ERA5-Land, particularly at high Alpine elevations. And what about snow-albedo feedback in ERA5-Land itself? Could these feedbacks also have contributed to the SWE bias in ERA5-Land?

Reply to SC5: We thank the Reviewer for this highly insightful comment and for pointing out the relevance of Orsolini et al. (2019). We agree that these represent very plausible hypotheses that could help explain the observed biases. In the revised manuscript, we have expanded the discussion to explicitly cite Orsolini et al. (2019) and to discuss potential mechanisms (ll. 439-448): *"However, another potential explanation for such biases was given by Orsolini et al. (2019), who identified the discontinuation of snow-cover assimilation from the Interactive Multisensor Snow and Ice Mapping System (IMS) above 1500 m a.s.l. as a possible further contributor to high-elevation snow biases in ERA5 over the Tibetan Plateau, since the simulated snow cover at these elevations is no longer corrected by observed snow cover. The same restriction applies over the Alps, where excessive snow cover in ERA5 could increase surface albedo, artificially lowering near-surface temperatures; these cold biases could then be propagated into ERA5-Land, which relies on ERA5 meteorology as its atmospheric forcing. Moreover, snow–albedo interactions could also reinforce the positive SWE bias within ERA5-Land itself: within CHTESSEL, an excessive snowpack could prolong snow cover and maintain a higher surface albedo, reducing absorbed shortwave radiation and the energy available for melt, thereby further delaying snow depletion."*

SC6: L371-374 and beyond: I guess the negative SWE biases in the Apennines or some regions of the Alps, like the Trentino-Alto Adige region could also be related to differences in how ERA5-Land (via precipitation forcing from ERA5) or CERRA-Land/VR-REA-IT and the S3M model (the basis for IT-SNOW) deal with the precipitation phase, particularly for temperature regimes around 0°C. I don't know how the precipitation-phase partitioning in S3M works, but could it be that precipitation falls as snow in S3M, whereas in ERA5-Land or the other 2 models precipitation falls as rain? To address this question, I recommend adding two more rows to Figures B1-B3 that visualize the seasonal biases in rainfall and snowfall.

Reply to SC6: We fully agree that differences in the precipitation-phase partitioning (solid vs. liquid) around 0°C between the evaluated products and the S3M model are highly likely to be a key driver of the SWE biases observed in such regions. We agree with the rationale behind your suggestion to visualize the seasonal biases in rainfall and snowfall. However, we are unfortunately unable to perform this specific evaluation due to data availability constraints. To ensure a consistent and "symmetrical" evaluation, our analysis relies strictly on the variables that are publicly and openly distributed for all datasets. While ERA5-Land does provide partitioned snowfall and rainfall outputs, these specific fluxes are not available in the public data releases for CERRA-Land and VHR-REA_IT. Also, the reference dataset (IT-SNOW) includes only the final snow variables (SWE, snow depth, bulk snow density and liquid water yield), but not the snowfall and rainfall fields internally computed by the S3M model. Consequently, computing a direct, comparative spatial bias for snowfall against the reference dataset is not feasible. While a quantitative evaluation is not feasible, in the revised manuscript we have briefly mentioned the possibility that differences in the precipitation-phase partitioning approaches could potentially influence the SWE discrepancies, particularly in regions where temperatures frequently fluctuate around the freezing point, together with the reason why this mechanism could not be evaluated directly. (ll. 456-460): *"This apparent inconsistency may reflect factors not resolved by the seasonal precipitation and temperature averages, including differences in precipitation-phase estimation, the limited representation of the high-elevation massifs on the 9-km grid, and differences in the representation of snow processes such as ablation."*

SC7: Also, I am wondering whether there are differences in the density of ultrasonic sensors (used for assimilation of IT-SNOW) among the different regions? Could it be that the density of sensors is higher in the central Apennines and some regions of the Alps, such as Trentino Alto-Adige? I am asking this because the negative SWE biases presented in Figure 3 seem to appear consistently over the same regions, independent from the comparison with the different land surface datasets. Even the SWE biases between ERA5 Land and IT-SNOW show patches of negative SWE biases in the Trentino Alto-Adige region, which seem to appear in the lower-elevated areas of the respective region.

Reply to SC7: Yes, there are indeed regional differences: the bulk of the snow data are in the northern regions, particularly within the Po river basin in the northwest. While a lower sensor density might indeed affect the performance of the model, previous comprehensive assessments of IT-SNOW against a spatially distributed Sentinel-1 snow depth dataset did not reveal any clear spatial bias patterns linked to the sensor network density (Avanzi et al., 2023). So we are inclined to attribute most of these patterns to the evaluated datasets, rather than to IT-SNOW. We appreciate the Reviewer for prompting this clarification. In the revised manuscript, we have

briefly expanded upon this aspect in the text. (ll. 175-177): *"The modelled snowpack is constrained by the daily assimilation of snow depth observations. These come from a network of ~350 ultrasonic sensors that follows the distribution of Italian snow resources and is therefore densest across the northern regions, particularly within the Po river basin."*

Response to Prof. S. A. Margulis

Specific comments (SC)

SC1: Line 59-67. In discussing other observationally-constrained snow datasets, you should consider the mention of Cortés and Margulis (2017), Liu et al. (2021), and Fang et al. (2022), which demonstrate multi-decadal SWE estimation over High Mountain Asia, the Andes, and the Western U.S. respectively.

Reply to SC1: We are grateful to Prof. Margulis for having suggested these three references, which are directly relevant to the paragraph in question, and we have cited them in the revised version of the manuscript. Cortés and Margulis (2017), Liu et al. (2021), and Fang et al. (2022) present multi-decadal, observation-constrained snow reanalyses for the Andes, High Mountain Asia, and the western United States respectively, in which fractional snow-covered area retrieved from the Landsat record is assimilated to produce sub-kilometre SWE estimates. Since the Landsat archive extends back to the mid-1980s, these reanalyses span periods of up to nearly four decades. Their example shows that the temporal coverage of observation-constrained SWE products is determined by the length of the assimilated observational record and by the design of each individual product, rather than being an inherent limitation of the product class, and we have revised the paragraph to make this distinction explicit. (ll. 63-72): *"Prominent examples include those developed for Italy (IT-SNOW, ~500 m; Avanzi et al., 2023), Switzerland (OSHD, up to ~250 m; Mott et al., 2023), the contiguous United States (SNODAS, ~1 km; NOHRSC, 2004), the Po River District (J-snow, ~500 m; Dall'Amico et al., 2025), the extratropical Andes (~180 m; Cortés and Margulis, 2017), High Mountain Asia (~500 m; Liu et al., 2021), and the western United States (~500 m; Fang et al., 2022). Despite their demonstrated reliability, the temporal and geographical coverage of these products varies widely, since it depends on the length of the assimilated observational records and on the design of each individual product. For example, while the Andes and western United States products, which assimilate fractional snow-covered area retrieved from the Landsat record, extend back to the mid-1980s, SNODAS and IT-SNOW begin in 2003 and 2010, respectively, and J-snow covers only the Po River District."*

SC2: Figure 2, panel (b). It may make it easier for the reader to interpret the SWE if you centered the mean annual temperature colorbar on zero degrees Celsius. It would provide visual evidence of regions that are below freezing for a significant fraction of the year and thus more likely to have snowfall than rainfall.

Reply to SC2: We fully agree that centring the temperature colour scale on zero degrees Celsius makes the figure easier to interpret, and we have revised the colour scale accordingly in Fig. 2, panel (b), and in Fig. B2, panels (a) to (d), where the same scale is used. We thank Prof. Margulis for the suggestion.

SC3: Section 3.1. Given that the IT-SNOW is the reference product, a bit more detail on how it works would help the reader. For example, how is the assimilation implemented? Is it in near-real-time? Retrospective? What is the assimilation method used? While these details are in the provided references, having the basic summary would be helpful.

Reply to SC3: We agree that, given the central role of IT-SNOW as the reference product of this study, a basic overview of how it is generated would benefit the reader without requiring consultation of the original publications.

To answer the Reviewer's questions directly, IT-SNOW is a retrospective product. It is generated by rerunning the S3M Italy operational chain over the historical period, and the snow depth observations are assimilated once per day through a Newtonian relaxation (nudging) scheme; the SCA maps act only as a mask on the spatialised snow depth fields prior to assimilation (see our reply to SC4).

In the revised manuscript, we have complemented the description previously provided in Sect. 3.1 with the following text (ll. 148-152): *"IT-SNOW is a high-resolution, observation-constrained snow reanalysis for Italy, available from September 2010 onward, providing daily maps of SWE, snow depth, bulk snow density, and liquid water content at ~500 m spatial resolution (Avanzi et al., 2023). The reanalysis is generated retrospectively by rerunning the operational Snow Multidata Mapping and Modeling (S3M) chain for Italy over the historical period, constraining the estimates by using in situ and satellite snow data available for each date."* and (ll. 175-187): *"The modelled snowpack is constrained by the daily assimilation of snow depth observations. These come from a network of ~350 ultrasonic sensors that follows the distribution of Italian snow resources and is therefore densest across the northern regions, particularly within the Po river basin. The point measurements are spatialised each day through multilinear regressions against elevation, slope, and aspect, fitted separately over 10 subdivisions of the territory that group areas of comparable snow climatology. The resulting fields are then trimmed to the observed snow line, using a blended snow-covered area map that combines Sentinel-2 (Main-Knorn et al., 2017), MODIS/VIIRS, and EUMETSAT H SAF H10 retrievals: pixels observed as snow-free are removed, so that the regressions do not extend snow onto bare ground. Conversion of the masked fields to SWE relies on the modelled snow density, whose spatial variability is considerably lower than that of snow depth (Mizukami and Perica, 2008). The SWE fields obtained in this way are assimilated once per day through spatially distributed Newtonian relaxation, or nudging: at each grid cell the modelled SWE is corrected by a fraction of its departure from the observation-based estimate, and this fraction decreases with distance from the sensors. The procedure is confined to December–April, outside which the ultrasonic sensors may pick up the vegetation growing beneath them after melt-out and introduce spurious signals into the spatialised maps (Avanzi et al., 2023)."*

SC4: Line 151: How informative is the snow covered area in updating SWE? It typically has limited instantaneous relationship with SWE (i.e. often saturates near 100% for much of the accumulation season). Does the assimilation more heavily rely on the snow depth observations as a constraint?

Reply to SC4: We thank Prof. Margulis for pointing this out. The instantaneous relationship between snow-covered area and SWE is indeed weak over a consolidated snowpack, where SCA saturates near 100 % for much

of the accumulation season, and the design of the IT-SNOW assimilation is consistent with this limitation. In S3M Italy, the SCA maps are in fact not assimilated directly: the quantitative constraint on the modelled snowpack comes entirely from the snow depth maps, which are converted to SWE using modelled snow density and assimilated through a spatially distributed Newtonian relaxation scheme (Avanzi et al., 2023). The blended SCA maps are used to remove snow-free pixels from the snow depth maps before assimilation. Their information content therefore enters the system as a binary, snow / no-snow constraint on the extent of the assimilated fields, correcting the regression-based spatialisation where it would otherwise extend snow into areas observed as snow-free, rather than as a constraint on snow mass. The main role of assimilating SCA is, then, to get the snow line right, which is crucial especially during rain-on-snow events (Parajka et al., 2019). We also recognise that the previous wording, which introduced the SCA maps as the first of two assimilated data sources, may have suggested a direct assimilation of SCA. In the revised manuscript, we have rephrased this passage to state explicitly that the SCA maps act as a mask applied to the snow depth fields prior to assimilation, and that the quantitative update of the modelled SWE relies on the snow depth observations. (ll. 179-181): *"The resulting fields are then trimmed to the observed snow line, using a blended snow-covered area map that combines Sentinel-2 (Main-Knorn et al., 2017), MODIS/VIIRS, and EUMETSAT H SAF H10 retrievals: pixels observed as snow-free are removed, so that the regressions do not extend snow onto bare ground."*

SC5: Line 158: It would be helpful to the reader to provide some of the bulk verification statistics from previous work (i.e. define what is meant by "consistent" and "low mean bias" more quantitatively).

Reply to SC5: We agree that the terms "consistent" and "low mean bias" might remain vague without the corresponding figures, and that reporting the main verification statistics gives the readers a more concrete measure of the reliability of the reference dataset. These statistics are reported in Avanzi et al. (2023). In the revised manuscript, we have replaced the qualitative wording with the corresponding quantitative results (ll. 188-193): *"IT-SNOW has been evaluated against Sentinel-1-derived snow depth maps (C-SNOW; Lievens et al., 2019) over the period September 2016–April 2020, and against independent in situ snow depth and SWE measurements (Avanzi et al., 2023). The comparison against C-SNOW showed that IT-SNOW snow depth estimates were consistent with satellite-based snow data, with low mean bias across the evaluated mountain areas (mean bias: −10 mm; mean RMSE: 220 mm), and comparison against ground measurements confirmed that the dataset reconstructs seasonal snow accumulation and melt dynamics (RMSE: 300–600 mm for snow depth; 90–300 mm for SWE)."*

SC6: Figures 3 and 4. I would suggest adding a third row at the bottom showing the bias as a percent of the reference. Given the large difference in mean annual SWE or SCD values, it would be complementary to show the percentage bias in addition to the absolute bias.

Reply to SC6: We agree that a relative measure of the discrepancies can complement the absolute one. We have therefore computed the percentage bias for both mean annual SWE and SCD, dividing the bias by the corresponding IT-SNOW reference value, and added it as a third row of panels to the figures reproducing the current Figs. 3 and 4.

Having examined these figures, we would nonetheless like to comment on what they add and on where they can be best placed.

The analysis domain comprises all grid cells in which the IT-SNOW mean annual SWE exceeded 1 mm in at least one year of the study period, and therefore includes cells where the reference mean annual SWE amounts to a few millimetres and the reference mean annual SCD to a few days. Since the reference values in the denominator are positive, the percentage bias has the same sign as the absolute bias everywhere, and the two rows differ only in the relative weight they give to each cell. Over the Alps and the higher Apennines, where the reference values are large enough for the normalisation to be stable, the patterns of the percentage bias follow those of the absolute bias, and the largest absolute discrepancies correspond to comparatively moderate percentage values. Where the reference values are small, in contrast, differences of little practical consequence in absolute terms correspond to percentage values of several hundred per cent, and the percentage bias saturates the colour scale over much of the lowland part of the domain; the saturation there reflects the magnitude of the reference value rather than that of the discrepancy.

Since the percentage bias adds no information on the sign or the spatial pattern of the discrepancies and is difficult to interpret over the low-elevation part of the domain, and in order to keep two already composite figures readable within the concise format of a Technical Note, we have included these expanded figures in the new Supplement, as Figs. S1 and S2, referring to them explicitly in Sects. 5.1 and 5.2 together with a short note on the interpretation of the percentage bias where the reference values are small. The added text reads (ll. 415-421): *"Supplement Fig. S1 presents an expanded version of this figure, adding percentage-bias panels and marking the cells whose biases are not significant after false discovery rate (FDR) screening. Where reference SWE is substantial, the percentage-bias panels largely mirror the absolute-bias pattern; where it is small, however, even small absolute differences can yield large percentage biases, which should therefore be interpreted with caution (Fig. S1, panels h–j). Across all products, cells with non-significant biases are generally confined to areas with the smallest biases values, whereas large biases are statistically significant (Fig. S1, panels e–g)."*

SC7: Line 359: Follow-on studies to those cited above (i.e., Liu et al. (2022) and Fang et al. (2023) looked at relationships between SWE errors and meteorological forcing (including ERA5 and ERA5-Land). These might be useful for context and supportive of the findings in this manuscript.

Reply to SC7: We thank Prof. Margulis for these further references. Both works address the same question that motivates our analysis, namely how far discrepancies in SWE estimates from large-scale products can be traced back to their meteorological forcing and how far other factors are involved, and they do so for products that partly overlap with those evaluated here, including ERA5-Land, across mountain regions with contrasting snow climates. Their conclusions are consistent with ours in attributing a substantial part of the differences in simulated SWE to the meteorological forcing, while showing that the forcing alone does not account for all of them, which is the same distinction we draw in Sect. 5. They therefore provide a valuable wider context for the interpretation developed there and for the behaviour of ERA5-Land in particular. We have accordingly added these references in the Introduction, in Sect. 5.1 and in the concluding discussion of the transferability of the

evaluation framework. (ll. 86-89): *"Although previous studies have evaluated large-scale SWE products in other regions and confirmed that product performance varies with geographic and hydroclimatic setting (Brown et al., 2018; Broxton et al., 2016; Fang et al., 2023; Liu et al., 2022; Monteiro and Morin, 2023; Mudryk et al., 2025), no systematic evaluation of SWE estimates from large-scale land-surface products has yet been conducted across the whole of Italy."* and (ll. 460-463): *". Ablation, in particular, can be relevant even during the accumulation season: in evaluations of ERA5-Land and other global products in High Mountain Asia (Liu et al., 2022) and in the western United States and the Andes (Fang et al., 2023), accumulation-season snow losses contributed to inter-product differences in SWE estimation."* and (ll. 621-624): *"The complementary precipitation and temperature analysis helps to attribute a substantial part of the SWE differences to their meteorological origins, although in some regions and products the forcing biases alone do not fully account for the SWE discrepancies (Liu et al., 2022; Fang et al., 2023)."*

SC8: I would suggest the authors consider whether the Alps vs. Apennines results should be discussed in terms of seasonal vs. ephemeral snow. Liu et al. (2022) partitioned their analysis of HMA snow into seasonal and ephemeral and found varying errors for the different regimes. From the results shown, it appears the Apennines may be more ephemeral and that may explain some of the errors seen.

Reply to SC8: This is a valuable suggestion, which we have adopted in the revised manuscript, as the contrast between the Alpine and mid- and lower-elevation Apennine results indeed can be framed well in terms of snow regime.

We have classified the study domain into seasonal and ephemeral snow. Rather than the peak SWE criterion of Liu et al. (2022), we adopted a snow persistence criterion, which requires no variable or threshold beyond those already used in the manuscript. The resulting classification is shown in Appendix Fig. A1, and the procedure is now described in Sect. 4.1 of the revised manuscript (ll. 361-367): *"Finally, to obtain further insight into regional differences in product performance, the snow regime of each grid cell was also classified as seasonal or ephemeral according to snow persistence in the IT-SNOW reference. For each cell and snow-year, the longest sequence of consecutive days with SWE of at least 5 mm (the same threshold used for computing SCD) was identified, and its duration was averaged across all the snow-years. Following the 60-day persistence criterion (Sturm et al., 1995; Petersky and Harpold, 2018), cells with a mean duration of at least 60 days were classified as seasonal and the remaining cells as ephemeral. Since the classification serves as an interpretative aid rather than an evaluation result, the resulting regime map is provided in Appendix Fig. A1."*

The resulting classification broadly supports the Reviewer's interpretation. Seasonal snow is largely confined to the higher Alpine sectors, whereas ephemeral snow dominates most of the Apennines. In the Apennines, seasonal snow is restricted to the uppermost altitudes, with the most notable examples being the highest massifs of the central sector, which feature substantial areas characterized by a seasonal snow regime. This is consistent with previous descriptions of seasonal but highly variable snow presence in these high-elevation areas (De Bellis et al., 2010; AINEVA, 2023). This classification therefore provides a useful framework for interpreting the contrasting product performance between the Alps and the Apennines. In the revised manuscript, we have introduced the classification and the supporting references, briefly described the snow conditions of the higher

Apennines in Sect. 2, presented the regime map in Appendix A (Fig. A1), and used this distinction to support the discussion of regional performance differences in Sect. 5.

SC9: Line 393: The fact that the highest resolution (and coupled) model is presumably most computational expensive and yet worst in some error metrics is perhaps worth emphasizing more. Given that meteorological forcings are so important in SWE evolution, but typically relatively coarse, does not guarantee that results will be more accurate at higher resolution. This emphasizes the importance of observations in high-resolution applications.

Reply to SC9: We agree with Prof. Margulis that this point deserves greater emphasis than it previously received, and we have brought it out more clearly in the revised manuscript.

The results are indeed instructive in this respect. VHR-REA_IT combines the finest grid spacing with a fully coupled atmosphere and land configuration, and is presumably also the most computationally demanding among the analysed products, yet it shows the most widespread underestimation of mean annual SWE and SCD and the weakest daily correlations. Higher resolution and a more elaborate representation of land and atmosphere interactions therefore do not in themselves guarantee more accurate SWE estimates. This is probably due to the fact that, among the characteristics distinguishing the three products, VHR-REA_IT is the only one in which no observational constraint on the meteorological variables is applied within the regional domain, with observations entering only through the ERA5 lateral boundary conditions. As discussed in Sect. 5.1, we consider this the most plausible explanation for its performance, although the products differ in several respects at once. We have therefore expanded the discussion and the concluding remarks accordingly, highlighting that the assimilation of meteorological observations remains essential in high-resolution applications. The corresponding additions read (abstract, ll. 25-27): *"VHR-REA_IT shows widespread underestimation and the weakest temporal correspondence with IT-SNOW despite its finer grid spacing, which may be due to its fully coupled atmosphere–land architecture with no assimilation of meteorological observations."* and (ll. 505-507): *"Finer resolution and two-way coupling therefore do not by themselves guarantee more accurate SWE estimates, underscoring the importance of observational constraints in high-resolution applications (Fang et al., 2023)."* and (ll. 644-647): *"This is notable given that VHR-REA_IT combines the finest grid spacing with the most complete land–atmosphere coupling among the evaluated products: higher resolution alone does not guarantee more accurate SWE, and the assimilation of meteorological observations remains essential in high-resolution applications."*

SC10: Section 5.2: Can you provide a bit more detail on how the models parameterize snow covered area? Is it a prognostic variable or is it parameterized as a function of snow depth? Is it binary or fractional?

Reply to SC10: We thank Prof. Margulis for this question, which concerns a model feature that the previous version of the manuscript did not describe.

In all three products, snow cover fraction is a diagnostic rather than a prognostic variable. It is represented as a fraction of the grid cell rather than as a binary indication of snow presence or absence. The prognostic snow variables include SWE and snow density, from which snow depth is diagnosed through the mass–volume

relationship (snow density × snow depth = water density × SWE). Snow cover fraction is then computed as function of the simulated snow amount, expressed either as SWE or as diagnosed snow depth depending on the product, and reaches complete coverage when a product-specific snow amount threshold is reached. The three schemes therefore share the same general structure, although the specific formulation of this diagnostic differs among the products.

To ensure a consistent comparison across the four datasets, we did not use their native snow-cover-fraction fields. Instead, snow cover duration was derived directly from the daily SWE fields by applying the same 5 mm threshold to IT-SNOW and each of the three evaluated products and counting the number of snow-covered days within each snow year. The SCD metric is thus defined consistently across all four datasets and does not depend directly on their native snow-cover-fraction parameterisations.

In the revised manuscript, we have clarified in Sects. 4.1 and 5.2 that the analysed snow cover duration is derived exclusively from daily SWE estimates from each product. (ll. 330-332): *"It is to be underlined that, for consistency across datasets, SCD was derived directly from daily SWE values, for all products, rather than from the snow-cover-fraction fields (that are available in some of the products)."*

SC11: Figure 5: This is a more general comment relevant to some of the other figures too, but Figure 5 provides a useful example. The authors should consider whether using a more discretized colorbar will help communicate the message more clearly. In this case, having a continuous colorbar to demonstrate correlations is not as visually useful as perhaps discretizing the colorbar into a few bins where "poor", "moderate", and "high" correlation would be easier to infer.

Reply to SC11: We appreciate this suggestion and have revised Fig. 5 accordingly, since a discretised colour scale indeed conveys the differences in temporal agreement more clearly than a continuous one.

SC12: Figures A1-B3. Would using consistent colorbars for bias in these plots be easier to interpret or do the authors think using different colorbars for different variables are helpful?

Reply to SC12: Our intention in using a distinct palette for each variable was to make the quantity displayed immediately identifiable across the large number of similarly structured appendix panels. For a given variable, the same palette and the same limits are applied to all products and periods, so the panels remain directly comparable, and all palettes are diverging and centred on zero, so the sign and magnitude of the bias are read in the same way throughout. We would therefore prefer to keep the current colouring, since we believe it helps the reader to recognise immediately that the panels show different variables, while remaining consistent across the multiple figures referring to the different products. The direction of the temperature palette has nevertheless been reversed, following comment SC4 of Referee #1.

Minor suggested edits/corrections (MS)

MS1: Line 112: Is "SCIA" an acronym? If so, define in first instance.

Reply to MS1: Thank you for catching this. SCIA is indeed an acronym, standing for "Sistema nazionale per la raccolta, elaborazione e diffusione di dati Climatologici di Interesse Ambientale", and we have defined it at

its first occurrence. (ll. 120-121): *"The precipitation and temperature fields are derived from SCIA (Sistema nazionale per la raccolta, elaborazione e diffusione di dati Climatologici di Interesse Ambientale; Desiato et al., 2007, 2011), a national gridded observational dataset"*

MS2: Line 352: Use "SWE" instead of "snow water equivalent".

Reply to MS2: Corrected.

References

AINEVA: Appennini, resoconto stagione 2022-2023, Associazione Interregionale Neve e Valanghe, available at: https://aineva.it/wp-content/uploads/Pubblicazioni/resoconto2022-23/appennini2022-23.pdf, 2023.

Avanzi, F., Gabellani, S., Delogu, F., Silvestro, F., Pignone, F., Bruno, G., Pulvirenti, L., Squicciarino, G., Fiori, E., Rossi, L., Puca, S., Toniazzo, A., Giordano, P., Falzacappa, M., Ratto, S., Stevenin, H., Cardillo, A., Fioletti, M., Cazzuli, O., Cremonese, E., Morra di Cella, U., and Ferraris, L.: IT-SNOW: a snow reanalysis for Italy blending modeling, in situ data, and satellite observations (2010–2021), Earth Syst. Sci. Data, 15, 639–660, https://doi.org/10.5194/essd-15-639-2023, 2023.

Cortés, G. and Margulis, S.: Impacts of El Niño and La Niña on interannual snow accumulation in the Andes: results from a high-resolution 31 year reanalysis, Geophys. Res. Lett., 44, 6859–6867, https://doi.org/10.1002/2017GL073826, 2017.

De Bellis, A., Pavan, V., and Levizzani, V.: Climatologia e variabilità interannuale della neve sull'Appennino Emiliano-Romagnolo, Quaderno Tecnico ARPA-SIMC n. 19/2010, ARPA Emilia-Romagna, Bologna, 2010.

Fang, Y., Liu, Y., and Margulis, S. A.: A western United States snow reanalysis dataset over the Landsat era from water years 1985 to 2021, Sci. Data, 9, 677, https://doi.org/10.1038/s41597-022-01768-7, 2022.

Fang, Y., Liu, Y., Li, D., Sun, H., and Margulis, S. A.: Spatiotemporal snow water storage uncertainty in the midlatitude American Cordillera, The Cryosphere, 17, 5175–5195, https://doi.org/10.5194/tc-17-5175-2023, 2023.

Froidurot, S., Zin, I., Hingray, B., Gautheron, A.: Sensitivity of Precipitation Phase over the Swiss Alps to Different Meteorological Variables, J. Hydrometeorol., 15, 685–696, https://doi.org/10.1175/JHM-D-13-073.1, 2014.

Liu, Y., Fang, Y., and Margulis, S. A.: Spatiotemporal distribution of seasonal snow water equivalent in High Mountain Asia from an 18-year Landsat–MODIS era snow reanalysis dataset, The Cryosphere, 15, 5261–5280, https://doi.org/10.5194/tc-15-5261-2021, 2021.

Liu, Y., Fang, Y., Li, D., and Margulis, S. A.: How well do global snow products characterize snow storage in High Mountain Asia?, Geophys. Res. Lett., 49, e2022GL100082, https://doi.org/10.1029/2022GL100082, 2022.

Mizukami, N. and Perica, S.: Spatiotemporal characteristics of snowpack density in the mountainous regions of the western United States, J. Hydrometeorol., 9, 1416–1426, https://doi.org/10.1175/2008JHM981.1, 2008.

Orsolini, Y., Wegmann, M., Dutra, E., Liu, B., Balsamo, G., Yang, K., De Rosnay, P., Zhu, C., Wang, W., Senan, R., Arduini, G.: Evaluation of snow depth and snow cover over the Tibetan Plateau in global reanalyses using in situ and satellite remote sensing observations, The Cryosphere, 13, 2221–2239, https://doi.org/10.5194/tc-13-2221-2019, 2019.

Parajka, J., Bezak, N., Burkhart, J., Hauksson, B., Holko, L., Hundecha, Y., Jenicek, M., Krajčí, P., Mangini, W., Molnar, P., Riboust, P., Rizzi, J., Sensoy, A., Thirel, G., and Viglione, A.: MODIS snowline elevation changes during snowmelt runoff events in Europe, J. Hydrol. Hydromech., 67, 101–109, https://doi.org/10.2478/johh-2018-0011, 2019.

Pellicciotti, F., Brock, B., Strasser, U., Burlando, P., Funk, M., Corripio, J.: An enhanced temperature-index glacier melt model including the shortwave radiation balance: development and testing for Haut Glacier d'Arolla, Switzerland, J. Glaciol., 51, 573–587, https://doi.org/10.3189/172756505781829124, 2005.

Petersky, R. and Harpold, A.: Now you see it, now you don't: a case study of ephemeral snowpacks and soil moisture response in the Great Basin, USA, Hydrol. Earth Syst. Sci., 22, 4891–4906, https://doi.org/10.5194/hess-22-4891-2018, 2018.

Sturm, M., Holmgren, J., and Liston, G. E.: A seasonal snow cover classification system for local to global applications, J. Climate, 8, 1261–1283, https://doi.org/10.1175/1520-0442(1995)008<1261:ASSCCS>2.0.CO;2, 1995.

Wilks, D. S.: "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, B. Am. Meteorol. Soc., 97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.

Zhang, Z., Glaser, S., Bales, R., Conklin, M., Rice, R., Marks, D.: Insights into mountain precipitation and snowpack from a basin-scale wireless-sensor network, Water Resour. Res., 53, 6626–6641, https://doi.org/10.1002/2016WR018825, 2017.