Reply to the comments by Reviewer#2 on the manuscript "Snow water equivalent in the Alps as seen by gridded datasets, CMIP5 and CORDEX climate models"

We thank the Reviewer for the comments and suggestions regarding the discussion paper. We have addressed all the points he raised, performing supplementary analyses that in some cases were added in the manuscript or in the new Supplementary material. In brief, the new parts cover:

- The clarification of the paper objectives, including the motivation on why we kept the section on future projections.
- The evaluation of the RCMs in the historical period, previously based on ERA-Interim driven runs, and now extended also to the GCM-driven models. A new figure (Figure S03) was added in the Supplementary material.
- The analysis of the SNW distribution for different ranges of elevation, for all datasets (the references, GCMs, and RCMs). This analysis is now included in Fig. S04 of the Supplementary material.

Our point-by-point reply (black) to the suggestions and comments of the Reviewer (gray) is reported below.

Reviewer: "In this paper, the authors assess the snow water content in the Alps as represented in several atmospheric reanalyses, ERA-Interim-driven (and to a lower extent CMIP5-driven) regional atmosphere models (EURO-CORDEX), and numerous CMIP5 models. I appreciate the large amount of datasets analysed in this study, however I have several concerns with the paper in its current form, and I think that a major overhaul is required before publication.

First of all, the aim of the present study is not clearly stated. Does the paper aim to provide projections of snow water equivalent for end users (ecologists, road managers, ski resort), or does it aim to assess the models fidelity in order to point out limitations in our ability to project future snow water equivalent or for any other purpose? The aim should be better explained, and this should also be used to choose and justify which diagnostics are shown in this paper (e.g. why evaluating ERAinterin-driven RCMs in section 4.1 and 4.2 and additionally evaluating CMIP5-driven RCMs only in section 4.3?)."

Reply: The clarification of the aims of this paper was also provided as a response to Reviewer#1 to question #2. Now the objectives of the paper are clearly stated in the introduction (P3 L24 – P4 L6). This paper does not intend to deliver snow water equivalent projections for end users: without a proper *absolute* validation of the accuracy of the model, future projections would be pure speculation. Instead this paper aims to show and point out the strengths and limitations in the current knowledge of snow water equivalent characteristics at regional scale.

In brief the main objectives are:

- to assess the uncertainties in the characterisation of current snow water equivalent in the GAR, from both satellite/reanalyses and climate models.
- to explore how the current model uncertainties project into the future.

For the first objective we need to evaluate ERA-Interim-driven RCMs and, ideally, the AMIP simulations of the CMIP5 experiment, as pointed out by the referee (thank you for the suggestion). Nonetheless, out of

the 6 high resolution GCMs considered in this study, only two, CMCC-CM and MRI-CGCM3, have run AMIP simulations for the CMIP5 experiment (check in March 2017) and none of them is currently available for the download, apparently owing to issues with the servers. As of today, march 29th, we could not retrieve those data. At this stage it is impossible for us to evaluate the 2 AMIP runs.
For the second objective, we need GCM-driven RCMs and fully coupled GCMs. The scope of the manuscript is now better explained in the introduction.

"I also have a concern with the first diagnostics shown in this paper, i.e. the anomalies/biases represented on maps (Figs. 2-4). First, how are the datasets re-gridded prior to compute the difference? Furthermore, as the models (including reanalyses) miss the tail of the elevation distribution (as indicated in Fig.1), it is expected that they cannot account for high snowfalls observed in high-elevation areas. It seems to me that an alternative/complementary diagnostic would be to plot the snow water equivalent distribution per elevation bin. It would indicate whether the models behave well given their grid topography. I guess that the remapping used to build the Taylor diagram in Fig.5c partly addresses this, but it is not sufficient. In my opinion, this could replace sections 4.2 and 4.3 which I don' t find very informative."

Throughout the paper the datasets are regridded using conservative remapping. This remapping allows the conservation of the quantity (SNW) from the original to the output grid.
Remapping methods do not change the original resolution of the datasets, so models and reanalyses that do not represent the tail of the elevation distribution are not expected to represent high snowfalls observed in high-elevation areas. As suggested by the reviewer we produced a plot representing the snow water equivalent distribution per elevation bin (Fig R4). Reanalyses represent elevations up to 2000-2500 m; CMIP5 models generally represent elevations up to 2000 m; RCMs describe high elevation areas up to 3000 m. This plot clarifies what elevation ranges are represented in each dataset, thus it has been included in the text in Figure S04. However this analysis does not show how close the modelled and the reference SNW patterns are, in terms of point-by-point correlation, mean error and variance. This information is instead given in the Taylor diagrams, which in our opinion provide much information in a concise way and, in our opinion, they cannot be replaced by the plot of SNW per elevation ranges alone.

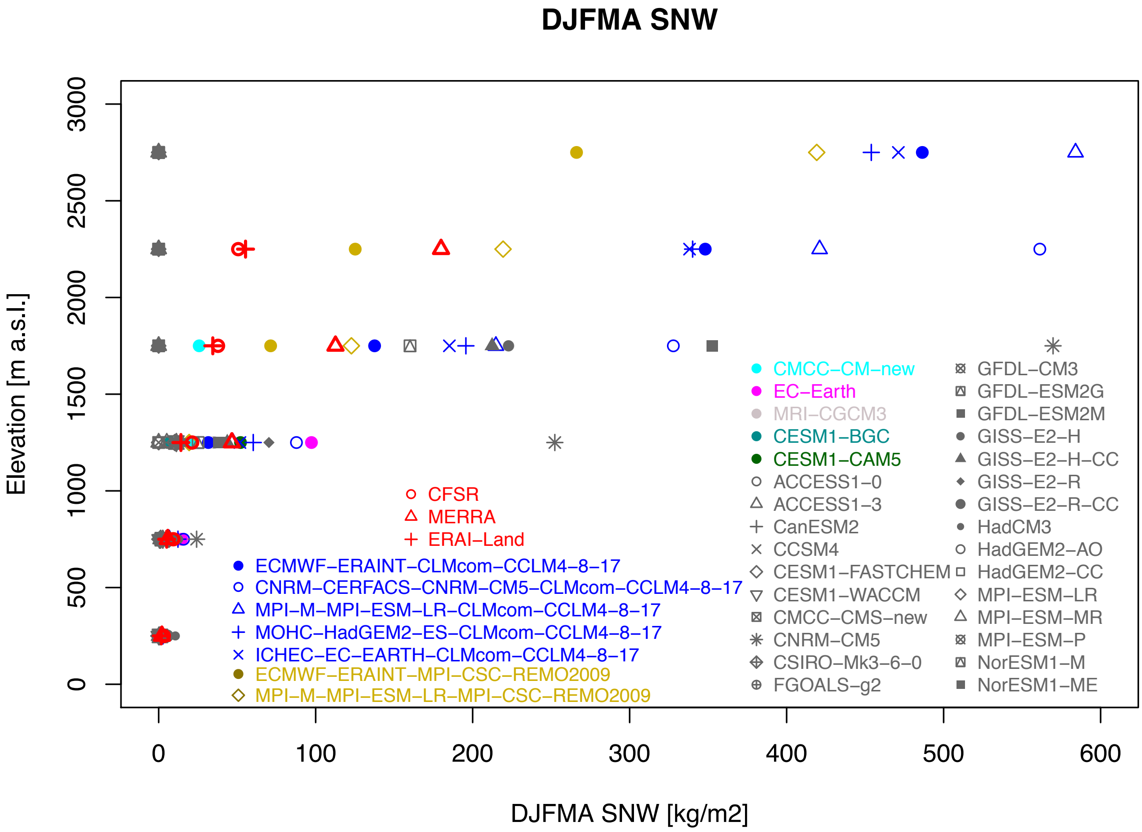

Fig R4. Multiannual mean DJFMA SNW in the Greater Alpine Region spatially averaged over different elevation ranges (500 m wide). The elevation is derived from the topography of each model (reanalysis).

“In addition, despite the limitations mentioned for snow water equivalent derived from passive microwave satellite observations, no attempt is made to discuss the validity of such products. The minimum would be to compare the two datasets described in section 2.1 over their common period. Of course, comparing to more datasets or to in-situ measurements would be even better. Is there any evidence that these satellite datasets are more reliable than the other datasets?”

The two satellite datasets AMSRE and Global SWE have been compared over their common period (2003-2007) and the plot has been integrated in Fig. 2f. We thank the reviewer for this suggestion. Given the relative short period of overlapping (5 years) we did not investigate further the time series, but we reported and discussed previous validation papers (Section 2.1 P5 L10-16)

“I have a possible concern with the choice of the simulations presented in the manuscript. ERAinterin-driven RCMs are more similar to AMIP models (atmospheric- only GCMs driven by observed SSTs) than to CMIP models, so in section 4.1 and 4.2, I think that comparing AMIP GCMS to ERAinterin-driven RCMs would make more sense. Then, in section 4.3 and 4.4 where the CMIP5-driven RCMs are evaluated, it makes more sense to compare to CMIP5 models.“

As stated before AMIP simulations are provided for only 2 HiRes GCMs and they are currently not available for download. However, we added Fig S03 in the supplementary material presenting the biases of GCM-driven RCM, to be compared to fully coupled CMIP5 models

“I have several other specific comments:

- Abstract, l.11: replace “latest” with “fifth” (in a couple of years, latest won't be clear).”

Done, thank you.

- 2nd and 3rd paragraph of the Introduction: there are also concerns related to snow itself (road & airport safety, ski resorts, . . .).

Thank you. We have mentioned in the text the impacts on winter tourism and we added 2 citations (Beniston et al., 2011, Rixen et al., 2011). We preferred not to mention airport and road safety because it is more related to extreme events, i.e. to temporal scales not covered by our analysis.

- Intro, l.25-26: “at relatively high spatial resolution” -> subjective, indicate a typical range.

Done, thank you.

- It would be interesting to discuss the reliability of satellite datasets in the Introduction. Note that the GlobSnow dataset is derived from satellite measurements but uses ground-based weather station data in the SWE retrieval.

Yes, we added:
- a discussion on reliability of satellite datasets in (Section 2.1 P5 L10-16)
- the fact that GlobSnow is based also on “surface measurements” (P3 L17). Thank you

- Section 2.3: Sabin et al. (2013) use LMDz as an atmosphere-only model (i.e. not coupled to an ocean), I don't know how relevant this is to the CMIP models.

Actually also Davini et al 2017 does. At present state of the art, ultra-high resolution simulations are AMIP only.

- Section 2.3, last paragraph, remove “at ISAC-CNR”.
- Tab. 1: there should probably be a line between satellite products and reanalyses.

Done, thank you.

- Given that LMD is mentioned, I'm surprised not to see the IPSL models in the long CMIP5 list, but anyway, there are clearly enough models in this paper.

IPSL models provide the snow depth variable but not snow water equivalent. Being the focus of this study on the latter variable, IPSL models do not appear in the paper.

- Sections 2.3 and 2.4: mention what kind of outputs are used (daily means or monthly means?).

We used monthly means. This detail has been added in the text (P7 L18 and L32)

- Section 2.4: what is a “non-reliable snow accumulation trends”? (and what is a reliable trend?).

Pixels masked as “glaciers” do not reproduce the snowpack evolution (accumulation and melting) but they continuously accumulate snowfall in time (without melting). “Non-reliable trend” refers to this behaviour and it has been clarified in the text (P8 L1-5).

- Section 2.5, about "The ability of climate models to properly reproduce snow water equivalent depends both on the accuracy of their snow schemes and on the reliability of the atmospheric forcings": it actually depends on many kinds on biases in the regional model (e.g. radiation scheme, boundary-layer scheme, etc, all being able to eventually impact snow).

Yes, we see your point. Actually with "reliability of the forcings" we already include all possible biases due to the land-surface and atmospheric schemes. We have rephrased the sentence "The ability of climate models to properly reproduce snow water equivalent depends on the accuracy of their surface snow schemes and on the reliability of the atmospheric fields forcing the snowpack processes."

- Section 2.5: what is the interpolation method for HISTALP and EOBS?

EOBS is kept at its original resolution (0.25° lat-lon regular grid). HISTALP has been conservatively remapped to EOBS grid, as all the other datasets, for the comparison in Figure 2. This has been explained in the corresponding Section 4.1.1

- Section 3 could probably be merged with section 2 into a "datasets and methods" section.

Actually we prefer to keep them separate to make the text more readable.

- Fig.2: why showing the relative precipitation bias (in %) while the temperature and snow biases are shown as absolute errors?

Mainly to be consistent with a previous study by Kotlarsky et al., 2014, presenting the same maps for the same models over the full EURO-CORDEX domain. Here we present a focus on the Alpine region.

- Fig.2: the caption "snow water equivalent in the EOBS observational dataset and the NSIDC Global Monthly EASE-Grid Snow Water Equivalent Climatology respectively" is misleading, it would be clearer at a first read to write that EOBS relates to (a) and (b) while NSIDC relates to (c).

Done, thank you.

- Section 4.1.1, about "In order to facilitate the comparison we present the differences with respect to a given dataset: the NSIDC Global SNW Climatology for SNW, since it is available for a longer period (1980-2005) than the other satellite product AMSR-E (2003-2011)". Ok, but it is a pity not to compare these two products aver the common period, especially given that you have claimed that "we expand the study by Mudryk et al. (2015) by including additional global SNW gridded datasets obtained from remote sensing" in the Introduction.

As previously mentioned, we have compared the two SNW satellite datasets over their common period (2003-2007) and the results are reported in Fig 2f. Given the relative short period of overlapping (5 years) we did not investigate further the time series, but we presented and discussed two papers on the validation of the two satellite products (Section 2.1 P5 L10-16). Thank you for the suggestion.

- Section 4.1.2: I would not say that REMO2009 is much better than the other RCMs, there is a substantial warm bias all over the domain (except maybe just over the mountain range) that could explain the relatively lower bias in SNW compared to other RCMs. Also, I would replace "CCLM4-8-17 and REMO2009 models which present no issues" with "CCLM4-8-17 and REMO2009 models which present weaker biases than other RCMs".

We agree that the performance of REMO2009 are comparable to other RCMs (please note that the plot in Fig 4m has been updated after finding an error in the computation of the DJFMA mean). We have better explained in the text (Sections 2.4 and 4.1.3) the "issues" in ALADIN53, HIRAM5 and RACMO22E models: "in glacier-masked pixels they show continuous snow accumulation and no melting. As this feature hampers the regridding of the models and the calculation of spatial averages over the GAR" we did not consider them for investigating the annual cycle and its projected changes at mountain range scale.

- Section 4.1.3: what period is used for the CMIP5 models, 1980-2005 or 1850-2005?

1980-2005, as clarified at P13 L4-5.

- Section 4.2 and its Taylor diagrams. I don't find the spatial correlation very relevant here, because it mostly relies on correlations between the topographies. Similar comment for RMSE and NSD.

We agree that the correlation coefficient R mainly reflects the *model* topography but we do not this this is a limitation because each model has its own topography, at its own resolution. It would have been meaningless if all models were using the same topography. In our case the objective is to measure the similarity between climate model climatologies (provided at different resolution) and a reference pattern. In such case RMSE, NSD and R provide, in our opinion, a good measure of this similarity.

- Why removing the worst RCMs in section 4.2?

We did not remove the worst RCMs but the models presenting pixels characterized by continuous snow accumulation and no melting, possibly areas masked as glaciers. As this feature hampers the regridding of the models and the calculation of spatial averages over the GAR we retained only two RCMs out of the five for further investigation. This has been explained in the text at P8 L1-5

- I am a bit lost, why using CMIP5-driven RCMs to analyse the seasonal cycle in section 4.3 and not to evaluate the mean spatial patterns in sections 4.1 and 4.2?

Yes, we added the evaluation of the GCM-driven RCMs in section 4.1.3, with one additional plot (Figure S3) in the Supplementary material.

- Section 4.3: I would not call 20CR a "reference dataset", it is a coarse atmospheric GCM only constrained by surface pressure and SSTs, probably more comparable to a coarse AMIP model. . .

Thank you for this suggestion. As already mentioned in the Response to Reviewer 1, in the revised version of the manuscript the 20CR reanalysis is not considered as a "reference" any longer. In fact, we repeated all the analyses excluding the 20CR one from the Multi-Reference-Mean, and consequently figures 5, 6 and 7 have been updated in the main text. The new procedure is explained at lines P12L31-34.