Response to Referee 1

We thank R1 for this detailed review, which will enable us to significantly improve our article. Enclosed please find a detailed explanation of the revisions we made based on R1's comments. For your convenience, comments are in bold and our response is in Arial italic. Revisions we made in the manuscript are presented in Arial italic with grey background.

General remark

The present work is potentially relevant, representing a method for adjusting RCM output to the conditions in mountainous environments. I particularly like the approach to show the immediate consequences of the method's performance (with respect to the individual meteorological parameters) for energy balance-based land surface models, as done in this study by evaluating snow depth results obtained by driving CROCUS with output by ADAMONT.

Besides, I have several major concerns. The reanalysis data set the authors use as a reference is very specific, as it includes average conditions within different mountain massifs for different altitudes. As such, it is hard to see the relevance of this study in a broader context, for example, when focusing on applications that use more common observation data sets (e.g., local-scale observations, or observations on spatially regular grids). With this regard, the terminology "downscaling method" may also be inappropriate (see my comments below). The authors need to discuss potential implications of using their method for observational data sets other than the SAFRAN. I also have important concerns regarding the evaluation of the results. The authors do not show the performance of the method based on independent data. This affects the entire discussion and the conclusions. Also, the evaluation is not performed at the scale of the application (individual massifs), but at a larger scale (Northern Alps, Southern Alps). However, for the application in energy balance based land surface models we are interested in the skill of the method of reproducing more local-scale conditions. I recommend considering the study for publication in Geoscientific Model Development if the authors perform a major revision.

*We thank the reviewer for this review, please see our specific responses to each point below. Concerning the remark « **As such, it is hard to see the relevance of this study in a broader context, for example, when focusing on applications that use more common observation data sets (e.g., local-scale observations, or observations on spatially regular grids).** », we included a sentence to justify the use of SAFRAN in the introduction (l. 99-104) :*

« However, despite its specificities, SAFRAN is the only reanalysis providing all variables needed to drive energy balance snowpack and land surface models over a long time period (since the 1960s). Moreover, it features a satisfactory altitudinal resolution of 300 m, much more precise than the altitudinal resolution of the RCMs (at a 12.5 km horizontal resolution), which is crucial for assessing the precipitation phase and the altitude variations of snow conditions. »

Moreover, we discussed the possible use of other datasets in the conclusion (l. 714-721):

« Note that beyond the French mountain regions, the method could be applied in France using the SAFRAN-France gridded reanalysis (Vidal et al., 2010). A Spanish version of SAFRAN was also developed recently (Quintana-Seguí et al., 2016). On a broader scale, the method, although detailed and tested here with the SAFRAN reanalysis as pseudo-observation, could be applied to other observational datasets or meteorological reanalyses, such as ERA-Interim surface fields (Dee et al., 2011) or MESCAN (Soci et al., 2016), providing enough data to drive land surface models. Furthermore, it could be used without application to drive a land surface model for more restricted datasets to simply compute atmospheric diagnostics (such as temperature and precipitation). »

Major comments

1. The authors do not perform an evaluation based on independent data. They use different learning periods, but as far as I can follow the validation is based on all the available data (thus, including the data used in the training). I acknowledge the importance of considering different learning periods, but in each case, the validation should be based exclusively on data which have not been involved in the learning process. This point affects the entire discussion section and the conclusions (incl. the abstract) of the submitted manuscript. See also my specific comments below.

Indeed, we used different learning periods, and we based our validation on different periods too (1980-1995, 1995-2010 and the whole 1980-2010 period, e.g. in Fig. 10), and not only on the 1980-2010 period. Please see our response to R1's specific comment below.

2. Grid point selection: The role of the SAFRAN massifs' centroids location and elevation is not clear. Apparently, using one location to represent an entire massif represents a simplification. Also, a potential altitudinal bias between the centroids' altitudes and the altitudes of the RCM grid point is not representative for the elevation differences apparent in reality. Not exactly knowing the SAFRAN reanalysis, and the definition and importance of the centroids, it is hard to understand why RCM grid points should be more realistic if they correspond in altitude and location to the SAFRAN centroids of the massifs, given that SAFRAN represent a simplification per se, assuming horizontally homogenous conditions for the entire massifs (thus areas ranging from 500 to 2000 kmˆ2). Overall, it is not recommended in statistical downscaling to use single grid points by atmospheric numerical models as predictors, because single grid point data are affected by numerical noise (see also: the concept of optimum scale, or effective resolution of an atmospheric numerical model). Also, data by at the RCM surface may be outperformed by the respective data extracted from the relevant pressure levels (see, e.g., Räisänen and Ylhäisi, 2011, Hofer et al, 2012, and references therein).

A remark that is shared by all three reviewers is indeed that we did not give enough details about the SAFRAN reanalysis, which is very specific. It is not a traditional gridded reanalysis, but instead the area of interest (the French Alps in our case) is subdivided into different polygons named massifs inside which the meteorological conditions are assumed to be homogeneous. The centre point of each polygon (centroid) plays no specific role in SAFRAN. It was just the way we chose to

select RCM grid points close to each massif. The use of single grid points can be justified by the fact that it would not be appropriate to mix (average) RCM pixels with different surface elevations, especially in mountainous regions.

We inserted more details about SAFRAN in the manuscript.

In the introduction (l. 95-104) :

« The SAFRAN meteorological analysis has been developed specifically to address the needs of snowpack numerical simulations in mountainous regions, and contains hourly time series of temperature, precipitation, wind speed, humidity, and short- and longwave radiation for so-called massifs (ranging between 500 and 2,000 km² in the French Alps) by elevation steps of 300 m (Durand et al., 2009a, b). However, despite its specificities, SAFRAN is the only reanalysis providing all variables needed to drive energy balance snowpack and land surface models over a long time period (since the 1960s). Moreover, it features a satisfactory altitudinal resolution of 300 m, much more precise than the altitudinal resolution of the RCMs (at a 12.5 km horizontal resolution), which is crucial for assessing the precipitation phase and the altitude variations of snow conditions. »

Section 2.2 has now become Section 2.1 (l. 127-140), and was changed to:

« The SAFRAN system is a regional scale meteorological downscaling and surface analysis system (Durand et al., 1993), which provides hourly data of temperature, precipitation amount and phase, specific humidity, wind speed, and shortwave and longwave radiation for each mountain region (or « massif ») in the French Alps (23 massifs, as illustrated in Fig. 1). Unlike traditional reanalyses, SAFRAN does not operate on a grid, but on French mountain regions subdivided into different polygons known as massifs. Massifs (Durand et al., 1993, 1999) correspond to regions ranging approximately between 500 and 2,000 km² for which meteorological conditions are assumed to be spatially homogeneous but vary with altitude. SAFRAN data are available for elevation bands with a resolution of 300 m. SAFRAN was used by Durand et al. (2009b) to create a meteorological reanalysis over the French Alps by combining the ERA-40 reanalysis (Uppala et al., 2005) with various meteorological observations including in situ mountain stations, radiosondes and satellite data. It was complemented after the end of the ERA-40 reanalysis (2002) by large-scale meteorological fields from the ARPEGE analysis, so that it now spans the period from 1959 to 2016, making it one of the longest meteorological reanalyses available in the French mountain regions. »

3. The term “downscaling method” for the ALADIN procedure is somehow misleading. The RCM has a horizontal grid point distance of 12.5 kmˆ2, thus the individual grid cells cover areas of approx.. 150 kmˆ2, while the SAFRAN assume spatially homogenous conditions within each 300 m altitude band of each massif, which in turn can cover up to 2000 kmˆ2. To my understanding, it is thus not possible to define a horizontal resolution of the SAFRAN reanalysis, in traditional terms. SAFRAN may represent more realistic conditions, in particular regarding the altitudinal differences within each massif, than the RCM. However, “downscaling RCM data to SAFRAN” is certainly not what is intended by the term “downscaling” used for inferring higher-resolution information by coarser-scale atmospheric numerical models in the scientific community. Please clearly discuss the practical differences between fitting a RCM to the SAFRAN reanalysis vs. downscaling a RCM to, e.g., to higher resolution gridded observations. I don’t know any reanalysis data set comparable to the SAFRAN reanalysis. I recommend to discuss the implications of using ADAMONT based on

other observational data sets as are more commonly available. Otherwise this study is important only for a very narrow range of applications (i.e., applications based on the SAFRAN data set).

Indeed, the term « downscaling » is not meant in its traditional sense in our study, because we don't perform any horizontal downscaling of RCMs, as R1 rightly points out. It consists more in some sort of « altitudinal downscaling » in mountainous regions and a statistical adjustment, as we produce adjusted RCM data for every 300m elevation band (and in every SAFRAN massif), starting from only one data value at the surface elevation of each RCM grid cell.

We decided to change the title of our manuscript accordingly to:

« The statistical adjustment method ADAMONT v1.0 for climate projections in mountainous regions applicable to energy balance land surface models »

Furthermore, we removed most references to the word « downscaling » and replaced it by the words « statistical adjustment ».

We inserted some discussion about the differences between our method and traditional downscaling and discussed the implications of using the ADAMONT method based on other datasets such as gridded reanalyses or observations:

l. 99-104 :

« (…) However, despite its specificities, SAFRAN is the only reanalysis providing all variables needed to drive energy balance snowpack and land surface models over a long time period (since the 1960s). Moreover, it features a satisfactory altitudinal resolution of 300 m, much more precise than the altitudinal resolution of the RCMs (at a 12.5 km horizontal resolution), which is crucial for assessing the precipitation phase and the altitude variations of snow conditions. »

and l. 714-721:

« Note that beyond the French mountain regions, the method could be applied in France using the SAFRAN-France gridded reanalysis (Vidal et al., 2010). A Spanish version of SAFRAN was also developed recently (Quintana-Seguí et al., 2016). On a broader scale, the method, although detailed and tested here with the SAFRAN reanalysis as pseudo-observation, could be applied to other observational datasets or meteorological reanalyses, such as ERA-Interim surface fields (Dee et al., 2011) or MESCAN (Soci et al., 2016), providing enough data to drive land surface models. Furthermore, it could be used without application to drive a land surface model for more restricted datasets to simply compute atmospheric diagnostics (such as temperature and precipitation). »

4. The discussion of the results is too lengthy. I see the importance of showing various evaluation criteria. However, the information should be compressed and presented in a more synthesized manner. Also, there are too many figures and the figure fonts are too small. Try to highlight the important points using figures which summarize the results in a more transparent way.

This is a point that was shared by all three reviewers. We decided to remove figures concerning the Northern and Southern Alps, to keep only figures showing results for the Vercors massif as an example (with larger fonts and better quality) + the same figures for every massif in the French

Alps in the Supplement. In the main article, we now have 15 figures instead of 28. Moreover, we decided to include a new synthetic table (Table 3) showing different features (mean values, biases, RMSE values and correlations) for variables of temperature, precipitation and snow depth for every massif in the French Alps + the Northern and Southern Alps, for the « RCM L. 1980-2010 » simulation configuration, at 1200 and 2100 m.

5. The article needs to be proofread by a native English scientist. The language needs improvement.

We sent our article to a professional English translator who helped improve the langage.

Specific remarks

Lines 175-178: This sentence is not clear. At this point, the reader does not know what the authors mean with "adjusted" RCM. Particularly, what do you mean with "there is no risk of introducing any artificial inflation of the simulated series"?

R1 is right. We removed the term « adjusted » in front of RCM, and inserted the information about resolution of the RCM and SAFRAN in the sentence. The end of the sentence ("there is no risk of introducing any artificial inflation of the simulated series") refers to a potential problem of quantile mapping that can occur when it is applied using observations with a much higher resolution than the RCM, as pointed out by Maraun, 2013 : « If, however, the bias correction also attempts to downscale [i.e., if the correction is against station (or very-high-resolution gridded) data], deterministic variance correction and quantile mapping approaches are not feasible. In general, the spatiotemporal variability at the gridbox scale is much smoother than at the local scale. Yet as only the marginals are corrected and no additional local-scale variability is generated, the temporal dependence and the spatial dependence between locations across grid boxes are those of the gridbox scale. Even more, since the correction is a deterministic mapping, within a grid box the spatial dependence between locations is fully deterministic. Hence, in this downscaling setting also deterministic variance correction and quantile mapping rescale the simulated time series in an attempt to explain unexplained small-scale variability. In other words, they inflate the simulated time series. »

We changed our sentence (l. 174-177) to :

« Because the spatial resolution of the RCM (12.5 km) is higher than the resolution of the observations (SAFRAN massifs, 500 to 2,000 km²), no spatial downscaling is attempted, so there is no risk of introducing any artificial variance inflation of the simulated series (Maraun, 2013), and therefore quantile mapping is adapted. »

Lines 185-186: This sentence implies an assumption. The sensitivity of quantile mapping to circulation may change in different climates. Please clearly distinguish the terms "weather regimes" and "climate".

Yes, this is what we imply. With climate change, the frequency of weather regimes may change. Moreover, the model errors are different in different regimes.

We introduced more explanation in this paragraph (l. 180-186):

« Moreover, Driouech et al. (2009) showed that for mid-latitude climates, such as that in Morocco, quantile mapping adjustment can vary for different weather regimes, because model biases vary in

different regimes. Similarly, Addor et al. (2016) demonstrated the sensitivity of quantile mapping adjustment to circulation biases over the Alpine domain. Additionally, the frequency of weather regimes may change in a changing climate (Boé et al., 2006; Cattiaux et al., 2013). To improve the stationarity of our method in a changing climate, weather regimes are thus taken into account in our method. »

Eq. 1: Indicate the value of N you used. How was N determined?

We tested values of N of 50 and 100, and only showed results for N=50, as explained later in the document. This was rather empirical : using N=50 yielded satisfying neighbouring grid points, while N=100 yielded neighbours that were sometimes too far (e.g., more than 80km in the case of Mont-Blanc for high altitudes) from the SAFRAN centroids.

We inserted the values of N we used in this equation (l. 196-198)) :

« (…) and N is referred to as the elevation factor. Values of 50 and 100 were tested, but only results using a value of 50 (N50) will be shown in this study. »

Lines 199-200: Provide more information about the clustering method you applied, since it is a crucial step in the downscaling procedure. For example, why exactly four weather regimes?

More details were provided in step 2 (l. 201-211):

« 2. Four different daily weather regimes were diagnosed from ERA-Interim for each season (DJF, MAM, JJA, SON), based on the geopotential height at 500 hPa, following Michelangeli et al. (1995), similar to the method described in Driouech et al. (2010). In the latter studies, only regimes for the winter season are defined. We chose to apply the same method to determine weather regimes for the other seasons as well. The clustering method used is the dynamic cluster method, whose goal is to "find a partition P of the data points into k clusters C1 , C2 ,..., Ck that minimizes the sum of variances (W(P)) within clusters, [...] (by defining) iterative partitions P (n) for which W(P (n)) decreases with n and eventually converges to a local minimum of W" (Michelangeli et al., 1995). A classifiability and reproducibility analysis in Michelangeli et al. (1995) suggested that 4 weather regimes (k=4) can reasonably be chosen for Europe. This number also ensures a sufficiently large size of the datasets for quantile mapping. »

Four weather regimes were chosen for each season, based on previous analyses by Michelangeli et al., 1995. Moreover, we shouldn't calculate more regimes, as it would result in statistically too small datasets, unsuitable for quantile mapping adjustment.

Line 205: What does the definition of the snow year matter for the downscaling procedure at this point?

R1 is right. In fact, the snow year is not introduced at this point, but at the end of the procedure.

We removed this sentence and introduced a last step (step 9, l. 295-298) to explain the resulting time series we obtain :

« 9. The resulting ADAMONT-adjusted hourly time series for each variable are obtained for each snow year (from the 1 st August to the 31 st July of the following year), matching the format of the SAFRAN dataset. This makes them easy to use as input of an energy balance land surface model such as SURFEX/ISBA-Crocus. »

Lines 208-213: This step is not clear. Is the selection of SAFRAN dates for each RCM date unique? How is this step automatized? Do you consider autocorrelation in the SAFRAN time

series to avoid artificial jumps? Thus this step imply a reordering of the daily RCM time series in order to best correspond to the temporal ordering in the SAFRAN data?

Indeed, this step was not clear enough. We improved its explanation in the new version of the manuscript.

Yes, the selection of SAFRAN dates for each RCM date is unique. For each RCM date, we perform a random draw amongst all available SAFRAN dates, and then browse through the dates chronologically until one meets all the requirements.

No, we don't consider autocorrelation in the SAFRAN time series, but we try as much as possible to use consecutive analogous days to avoid artificial jumps in step 7.

No, this step does not imply a reordering of the daily RCM time series in order to best correspond to the temporal ordering in the SAFRAN data. Analogous dates in the SAFRAN dataset are only used (in step 7) to reconstruct the daily cycle of the RCM based on hourly time series of the SAFRAN analogue in order to obtain final hourly adjusted RCM time series.

This step was re-written (l.229-238) :

« 6. For each day in the RCM dataset, an analogous date is chosen in the SAFRAN dataset, matching the following criteria: the month and the regime must be the same as in the RCM dataset, mean precipitation over the Alps must be consistent between datasets to ensure intermediate-scale (accross the French Alps) climatological consistency (i.e. if precipitation in the adjusted dataset is less than a threshold of 1 kg m −2 day −1 , precipitation in the SAFRAN analogue must also be less than this threshold), and whenever possible, consecutive time slices are chosen in the SAFRAN dataset in order to avoid artificial jumps in the final data linked to the choice of analogues. For each RCM date, a random draw amongst all available SAFRAN dates is performed, then the dates are browsed through until one meets all the requirements. This analogous day is then used in step 7 for all variables. »

Line 214 and below: The differences between the daily and subdaily ALADIN values should be as small as possible, how is the relation between the SAFRAN subdaily and the ALADING subdaily values? You may provide a formula which describes the transfer. The equations provided concern only air temperature, not the other variables, and it is not clear in which way the SAFRAN subdaily values are considered.

We thank R1 for this remark. We have now introduced more details about the procedure, especially for variables other than temperature.

Equation 3 was written at the beginning of the step (which changed the order of eqs 2 to 3) and generalized to any variable to explain the relation between the SAFRAN subdaily and the ALADING subdaily values. More details about the procedure for variables other than temperature were also added (l. 239-277):

« 7. The adjusted RCM dataset is then disaggregated from a daily integration period into an hourly time step, necessary for driving impact models such as the SURFEX/ISBA-Crocus model, by using the hourly SAFRAN data from each analogous date chosen in the previous step to reconstruct the daily cycle of the data. Hourly adjusted RCM value of any variable can be expressed as a function of the (hourly) SAFRAN value for this variable, as:

$$X^h_{RCM}(i) = a \times X^h_{SAF} + b, \quad (2)$$

where $X^h_{RCM}(i)$ is the hourly adjusted RCM value of the variable X and X^h_{SAF} is the hourly SAFRAN

value of the same variable from the chosen analogous date (step 6).Different criteria are chosen to calculate a and b, depending on the variable considered (Table 1). For the disaggregation of RCM adjusted temperature from daily to hourly, (…). Sensitivity tests yielded an optimal value of 2 for α. Following eq. (2), eq (3) transforms into : (…).

This procedure is only applied for temperature, because the use of the maximum and minimum criterion can lead to important jumps between consecutive days, which is not the case for other variables (Table 1). For humidity, eq. (2) is solved using b=0 and a = $X^{d,adj}_{RCM}(i)$ / $X^{h}_{SAF}(24h,i)$, so that the hourly adjusted RCM value and the hourly SAFRAN value at the last time step of day i ($X^{h}_{SAF}(24h,i)$) are equal. For wind speed, the same calculation as for humidity is applied, except if a > 1 (i.e., $X^{d,adj}_{RCM}(i) > X^{h}_{SAF}(24h,i)$) : then b=$X^{d,adj}_{RCM}(i) – X^{h}_{SAF}(24h,i)$ is calculated. For humidity and wind speed, if $X^{h}_{SAF}(24h,i) \leq 10^{-10}$, a=0. For precipitation and radiation, b=0 and a= $X^{d,adj}_{RCM}(i)$ / $X^{h}_{SAF}(mean,i)$, so that the mean hourly adjusted RCM value and the mean hourly SAFRAN value of day i are equal. For solar radiation, if $X^{h}_{SAF}(mean,i) \leq 10^{-10}$, a=0. For precipitation, if this is the case, a=1. »

Line 246: This step is confusing. So you put two quantile mappings on top of each other?

Yes, but the second quantile mapping is only applied to rainfall and snowfall separately, once they are calculated from the hourly total precipitation and temperature and cumulated on a daily basis.

The description of step 8 (l. 278-294) was improved and better justified:

« Finally, total precipitation is separated into rainfall and snowfall based on hourly adjusted temperature (a threshold of 1 °C is used for the transition from snow to rain, consistent with the approach used in SAFRAN). As mentioned above, inter-variable consistency is not guaranteed by quantile mapping. Consistency between temperature and precipitation is the most critical in this study, because we focus on mountain regions where snow plays an important role. As precipitation and temperature were corrected independently from each other (step 5), and because the adjustment can differ for the different precipitation phases, the relationship between temperature and precipitation phase may be modified by quantile mapping, so that the adjusted rain and snow distributions may lose consistency. To avoid this, Olsson et al. (2015) separate their temperature data into wet and dry days before adjustment. In our case an additional quantile mapping against SAFRAN is applied for daily cumulated RCM rainfall and snowfall separately. Hourly adjusted RCM rainfall and snowfall (a_2) are then determined by applying the ratio between daily rainfall or snowfall (taken separately) after quantile mapping (A_2) and daily rainfall or snowfall before quantile mapping (A_1) to the hourly rainfall or snowfall before quantile mapping (a_1):

$$a_2 = a_1 \times A_2/A_1 \quad (5)$$

If $A_1 = 0$ and $A_2 = 0$, then $a_2 = 0$. If $A_1 = 0$ and $A_2 = 0$, then $a_2 = A_2$.»

Line 258: Please clearly define what you evaluate. As far as I can follow, you want to evaluate the output of ADAMONT. However, you repeatedly mix "output of ADAMOMT" with the terms "RCM" and" adjusted RCM". For example, line 286: "ratio of the standard deviations between the RCM time series and SAFRAN". Do you really mean standard deviation of the RCM? Then this criterion is not indicative for the performance of ADAMONT, but for the performance of the selected RCM grid point without any adjustment. Further, whenever you use "adjusted RCM" I am not sure if you mean the ADAMONT output or some intermediate step in the downscaling procedure. If you evaluate the hourly output by ADAMONT applied to ALADIN, please say so.

Indeed, we thank R1 for this remark. « adjusted RCM » is intended as the ADAMONT output.

We tried to be more consistent throughout the manuscript, for example :

l. 300-301 :

« To evaluate our method, outputs by ADAMONT applied to the Météo France ALADIN RCM forced by ERA-Interim (1980-2010) were analysed »

or, l. 333-335:

« the correlation and the ratio of standard deviations between time series of ADAMONT applied to the RCM and time series of SAFRAN for each variable and as a function of the integration window (from 1 day to several years) from 1980 to 2010 ; »

All over the manuscript: don't use brackets inside brackets (e.g., in the references, line 265-266)

This problem was corrected.

Line 270: So the method consists of "downscaling" values at a massif scale (with downscaling not necessarily being the appropriate term), but then the evaluation is not performed at the massifs' scale, but at a much larger scale. The evaluation needs to be applied at the scale of interest, in this case, the individual SAFRAN massifs. Then, the resulting (numerous) scores may be synthesized (e.g., box plots of scores resulting for individual massifs for the Northern alps, box plots of scores resulting for the individual massifs for the Southern alps). The same for the altitudinal ranges. There are various ways how to summarize and appropriately illustrate performance metrics for numerous cases (here, variables, massifs, altitudinal ranges, and variants of the method in terms of learning period, grid point selection, a posteriori corrections, . . .). However it is important that the performance metric is applied to the scale of interest (and: that the reader does not need to interpret too many figures, see also my major comment above).

The evaluation was performed at the scale of the Vercors massif (one individual massif taken as an example). Moreover, supplementary figures display the same evaluation for every single massif in the French Alps (23 massifs in total). Results integrated at the scale of the Northern and Southern Alps are now only provided in Table 3 and not anymore in several plots. Nevertheless, depending on the application, it can be appropriate to operate at the scale of a given massif or at the scale of a wider geographical area such as Northern and Southern Alps. For example, studies addressing water ressources, natural snow abundance, winter tourism, hydrological regimes etc. do not require the same level of geographical resolution. In addition, most of the climate change signal is regional, so that for most regional-scale assessments model results aggregated for the Northern or Southern Alps, repsectively, will be amply sufficient. Concerning the need to reduce the number of figures, we agree with R1, and this was addressed above (General remark n°4).

Line 286: seasonal average time series is not an evaluation criteria per se. Mean annual cycle and mean altitudinal gradient: the same. Please be more specific.

We thank R1 for this remark. The sentence (l. 329) was changed to :

« The following features were analysed for temperature, total precipitation and snow depth »

Line 287: RCM time series or ADAMONT time series?

This was already addressed above (R1 specific remark Line 258)

Line 290: not an evaluation criterion.

Please see our answer above (R1 specific remark Line 286)

Line 317: Is "analysis of different massifs" is limited to the application of scores to the average conditions in the Northern and Southern alps, correctly?

No. We computed scores for the Vercors massif as an example, and for every massifs in the French Alps in the Supplementary Figures.

Line 320: Following the authors description, their evaluation is never independent from the training data. I.e., for the three different learning periods applied, performance metrics are always calculated including the training data. Case 1: Training period: 1980-1994, Case 2: training period 1995-2010, Case 3: training period 1980 to 2010. Evaluation period is always 1980-2010, thus always includes the training data. The performance metrics should be calculated based on split – sample validation (if parametric properties are validated, e.g., biases, distributions, e.g., Cannon), or cross-validation (if the temporal sequencing is validated, e.g., mean squared errors, ect.). Otherwise, the validation has no evidence.

Indeed, we used different learning periods, but we did base our validation on different periods, and not only on the whole 1980-2010 period. For example, Section 3.2 Mean seasonal variations : « Fig. 10 represents the mean annual cycle of temperature, precipitation and snow depth for the different ADAMONT-adjusted RCM simulations vs. the SAFRAN/Crocus reanalysis, for the period 1980-1995 and 1995-2010.».

Line 462: term "neighbor selection" technique is misleading. Use "grid point selection" instead.

We corrected this throughout the manuscript.

Figure 3 is a bit confusing. I see only one SAFRAN centroid being linked with the closest grid points in x, y, and z. Again, I am not sure what exactly you mean with "adjusted RCM". Is this the final output by the ADAMONT procedure based on ALADIN? Please be consistent with the terminology for the output.

Yes, this is because we are looking at one specific massif (the Vercors massif, see our comments above), which has one centroid. In the Supp. Figures, this is done for every massif of the French Alps. Concerning the terminology for the output, we corrected this in the new version of the manuscript.

Figure 16: It is hard to make any conclusions based on a visual inspection of Figure 16. Plotting the deviations from the modelled against the observed cumulative PDFs could help. The same with Figures 17-19.

The conclusion is that the modelled cumulative PDFs are very close to the reanalysis ones, which is reassuring considering the fact that we use a quantile mapping method. We think it is more interesting to show cumulative PDFs than the differences between modelled and observed PDFs. Even if the visual inspection of specific curves may not directly provide estimates of the quantitative deviation between PDFs (there are many quantitative metrics described and used in the manuscript), what is more important here is the fact that the shapes of the PDFs correspond,

rather than the deviation for one or several quantile ranges.

Figures 20 – 22: It is hard to distinguish amongst the different lines. Again, it could help to plot deviations of the model results to the SAFRAN values. Still, there are too many figures and the information content should be compressed (e.g., in terms of summary statistics for each model option, e.g., boxplots of skill scores).

Again, it is hard to distinguish amongst the different lines because the modelled cycle is very close to the observed one, which is a result per se. Showing the shape of the mean annual cycles of temperature, precipitation and snow depth seems again more interesting to us than showing the differences.

Concerning the number of figures and the changes we made, please see our responses above (General remark n°4).