Response to Referee 2

We thank R2 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 R2'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.

This paper describes a new statistical adjustment method intended to correct the biases in regional climate simulations in order to force land surface models in mountainous regions, and its application over the French Alps. The method is applied to the results of a RCM simulation forced by an atmospheric reanalysis. Precipitation and temperature after correction, and snow cover after land surface modelling with corrected forcing variables are compared to observations. The paper could be an interesting and useful addition to the field. The adjustment method is sound, and the evaluation work is serious. It may be publishable after major revisions. However, the description of the method needs to be much improved and the authors need to totally rethink how they present the results of the evaluation, with much less figures, but that better synthesize the results (see my major comment bellow). Moreover, the authors also need to demonstrate that the novelties of the adjustment method (quantile-quantile mapping that depends on large scale circulation; method used for the temporal downscaling from daily to hourly outputs) are useful. I also think that the English is not very good, and need to be improved.

We thank the reviewer for this review, please see our specific responses to each point below.

General comments

The paper is not particularly well written (despite visible efforts), with long and awkward sentences that sometimes make the paper difficult to understand.

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

Some important methodological aspects of the proposed adjustment method are not well described and sometimes not described at all. For example, the basic quantile-quantile mapping algorithm is not described precisely. In the description of the adjustment method, the authors simply describe the different steps very factually, but don't give the precise objective of the step (which is not always obvious) and very seldom justify the proposed solution (see my specific remarks).

We have improved the description and justification of the method in the new version of the manuscript (please see our specific responses below).

The authors have produced a very large number of figures (28 figures with a very large number of sub-figures. In the end, we have hundred of illustrations and even more in

supplementary materials). The sub-figures are often very small and therefore difficult (and sometimes impossible) to read. I think it is the job of the authors to do an effort to synthesize their results with a limited number of relevant illustrations, and to only show the important results (at least in the main paper): I disagree with the approach that consists in producing as much as possible illustrations and letting the reader finds what is important.

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.

The core of the adjustment method, quantile-quantile mapping, is well known and has been widely used. An originality of the approach proposed in the paper (even if it is not really the first time it is used, as noted by the authors) is to apply the quantile-quantile mapping by regime of large-scale circulation. Unfortunately, the authors do not demonstrate the interest of this approach. Is it really useful to do that? A second originality is the method used to obtain hourly data from the adjusted daily RCM output, which is often a necessary step to be able to force a land surface model. The authors propose a quite sophisticated approach, but do not evaluate its interest compared to simpler approaches (e.g. daily cycle from an analogous day without adjustment, or climatological diurnal cycle), neither directly (using observations with hourly resolution, I'm sure that some are available on the study area) nor indirectly (for example by comparing the simulated snow cover obtained with the different approaches). The authors should demonstrate that the novelties they introduce are really useful. It would significantly reinforce the interest of the paper.

We don't consider our approach to be better than other simpler ones. Moreover, we think it is out of the scope of this paper to perform a comparison of different approaches, which is already done as part of other projects (such as the CORDEX ESD experiment). What was missing throughout the manuscript was clearer justifications for the choices we made regarding the approach, especially concerning the regimes of large-scale circulation. This has now been included :

For the weather regimes, l. 180-186 were modified as follows :

« 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. »

Moreover, we included more details about the weather regimes selection in step 2 (l. 201-211) of Section 2.3 :

« 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. »

It is easy to justify why we did not choose simpler approaches. Concerning the approach of using the daily cycle from an analogous day without adjustment, the main factor for temperature change at a local scale is the large-scale radiation balance, not changes in atmospheric circulation. This main factor cannot be captured by using the analog without adjustment. Moreover, Boé (2007) showed that using analogous days without any adjustment did not enable to represent trends linked to climate change correctly. Furthermore, it is well known that the diurnal cycle of temperature and radiation is highly affected by cloudiness so using a climatological diurnal cycle could definitely not be realistic under many circumstances.

Using direct observations with hourly resolution would not be appropriate, because we only have few stations where all variables are available on the long-term (including radiation), for example the Col de Porte site, with « only » 22 years of hourly observations (Morin et al., 2012). Measurements at these stations could not be extrapolated to the entire French Alps, it is thus better to use a reanalysis such as SAFRAN.

Specific remarks

L67-68. I’m not sure to understand. It depends on how one deals with the distribution tail, I think.

Yes, indeed. In our case, we use a constant correction after the last quantile (99.5%), so this statement remains valid. We completed this sentence (l. 67-71) to take into account R2's remark :

« Moreover, the adjustment is not strictly restricted to the range of observed values in the reference period, which is the case for example for methods based on analog weather patterns (e.g., Déqué, 2007; Themeßl et al., 2011; Rousselot et al., 2012; Dayon et al., 2015), provided that values based on the lowermost and uppermost quantiles are handled appropriately (Gobiet et al., 2015). »

L98-102. Unclear and awkward sentence.

We reformulated the sentence (l. 104-107):

« Using this reanalysis as a pseudo-observation dataset combined with the strong and efficient quantile mapping adjustement methods in order to drive energy balance snowpack and land surface models is a highly desirable goal making full use of the current capabilities of climate impact assessment tools for mountainous regions.»

L102-104. Not clear

This sentence (l. 107-110) was also reformulated :

« In addition, the use of such methods ensures that the chronology of the RCM, which may be affected by climate change through variations of the seasonality of meteorological conditions, will be maintained in the adjusted climate projections.»

L127-128. OK, but in the end, the adjustment method is intended to correct the output of classical RCM projections such as the ones from Euro-Cordex. A smaller domain likely results in smaller biases compared to the biases found in typical RCM projections. Therefore the evaluation shown in this paper does not demonstrate that the adjustment method is able to deal correctly with the larger biases from classical RCM projections. I think it is a limitation of this work that should be pointed (including in the conclusion).

The size of the domain should not impact significantly the way RCM biases are corrected : quantile mapping will correct those regardless of their amplitude. However, if we want to evaluate the method in terms of chronology, it is better to choose a small domain so that it is better constrained. Anyhow, the size of the FRB12 domain (2200 x 2200 km) used in this study is comparable to the size of the commonly employed RCM ensembles such as EURO-CORDEX (5000 x 5000 km), so that we don't expect to experience problems linked to chronology when applying the ADAMONT method to RCMs such as the ones of EURO-CORDEX. We included a map of the EURO-CORDEX domain in Fig. 1.

We moved this paragraph into Section 2.4 « Method evaluation », and included more details (l. 303-309):

« We chose to work on a spatial domain smaller than the one used in EURO-CORDEX (domain covering all of Europe, Fig. 1), in order to evaluate the method and not the output of the RCM itself, especially in terms of chronology. Indeed, the smaller the domain, the more it is constrained (Alexandru et al., 2007). Simulations carried out over a domain centered on France, called FRB12 (Fig. 1) were thus evaluated (Sect. 2.4). The domain size (2,200 x 2,200 km) is on the same order of magnitude as the size of the commonly employed RCM ensembles such as EURO-CORDEX (5,000 x 5,000 km).»

L130-144. The authors need to explain what exactly is SAFRAN, how the values at different elevations are obtained etc. It may help to better understand some of their choices for the adjustment method. They also talk about the "centroids" of SAFRAN massifs. How are the centroids defined, and what do they represent?

Yes, we thank R2 for this advice. 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 of which the meteorological conditions are assumed to be homogeneous. The centre point of each polygon (centroïd) plays no specific role in SAFRAN.

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. »

Part 2.3. It would be good to give the forcing variables, their time step etc. in this section.

The following lines were added (l. 148-150) :

« Similarly to most land surface models, it requires sub-diurnal (ideally hourly) meteorological forcing data including air temperature, humidity, incoming longwave and shortwave radiation, wind speed, as well as rain and snow precipitation. »

L165. "Centroids": see a previous remark.

This was corrected.

L173. Please provide a more precise description about the exact algorithm used for the quantile-quantile mapping. Only a very brief general idea is given for the moment. For example, how many quantiles are used? How does it work for the values between quantiles: is a linear interpolation is used? How does it work for the values greater than the higher quantile? How the fact that the probability of precipitation in the RCM is different than in SAFRAN is dealt with?

This description was inserted later in this section, in step 4 & 5 (l. 218-228) :

« 4. The quantiles (99 percentiles + 0.5 % and 99.5 % quantiles) of the RCM distribution and the SAFRAN distribution are then calculated at each centre point of each massif and for each elevation band, for each variable, each season (DJF, MAM, JJA, SON) and each of the four weather regimes.

5. Quantile mapping is then applied to the entire RCM dataset for the period 1980-2010, taking into account the season and the weather regime. For the values between quantiles, a linear interpolation is used. For RCM values greater than the 99.5 % quantile, a constant adjustment based on the value of this last quantile is applied. For precipitation, it can happen that for low *quantiles, the probability of precipitation is lower in the RCM than in SAFRAN (i.e. several null values in the RCM, which can correspond to different positive values in SAFRAN). In this case, a random draw is performed amongst the SAFRAN values within the same quantile.»*

L174. Plotting? I hope that the authors do not really plot the simulated quantiles versus the observed quantiles.

No. We replaced the term « plotting » by « comparing ».

L179. Most adjustment methods make the same hypothesis. . .

Yes, but it is still a disadvantage of the method worth mentioning.

L187-194. For each massif, the authors use a single RCM grid point, the closest (either horizontally or also taking into account the vertical distance) of the massif centroid. Another solution, maybe better, would be to use all the RCM grid points within a massif, and use, based on their altitude, the most appropriate point for each elevation band within the massif. Another possibility, a priori more logical than the single point approach of the authors, would be to average all the RCM grid points within a massif. Obviously, the statistical properties of the spatial average are not the same than for a single point, but the values from SAFRAN on a massif are already spatial averages, right? I think it could make more sense. In any case, the authors need to justify their approach.

The first alternative solution R2 proposes is exactly what we do when we select the closest grid point also taking into account the vertical distance (N=50). For one given massif, we may have different RCM grid points selected depending on the elevation band, as in Fig. 3 for the example of the Vercors massif. However, we highlighted a clear degradation of scores for high elevations when using this approach, linked to the scarcity of high altitude grid points in ALADIN compared to SAFRAN, resulting in grid points being selected several tens of kilometers from the centre point of most SAFRAN massifs.

The second alternative solution you propose was not considered, because it would mean mixing (averaging) different grid cells with different surface elevation, which we think would be inappropriate.

L201. Hourly to daily what?

Hourly to daily time resolution. This was corrected.

L203-204. What do the authors mean by "each point". Each centroid? Or each elevation band within a massif? If they mean elevation band, using the word "point" is confusing.

Indeed. We meant each elevation band inside of each massif…

We changed the sentence (see our previous response to R2 specific remark L.173).

L204. I’m not sure to understand why this precision is necessary at this point.

Yes, this is something that was noted by R1 also. In fact, the snow year is not introduced at this point, but at the end of the procedure.

We 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. »

L210. I don't really understand how the "analogous dates" work. The authors need to give the general rationale of their approach, justify the choices they made, and better explain the step. Is there just one analogous date used? Is it just one random date among all the dates that match the different criteria? What is the justification for these criteria? In what sense the date is really "analogous"? The authors could search for a real analogous date, with similar temperature and precipitation over the massif for example. The authors need to explain the rationale behind the use of a day "consistent" in terms of precipitation. And why do they look at the average of precipitation over the Alps and not at the average over the massif of interest? Why the consistency is only defined in terms of occurrence of rain? The intensity does not matter?

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

The preponderant criteria in the choice of analogous date are the month and the weather regimes. The occurrence of precipitation was added to limit the domain for the random draw and mostly to benefit from an hourly chronology of precipitation in the analog. The intensity of precipitation could have been used, but this was not tested, and it is not necessarily a large-scale characteristic.

L234. How does the optimal value of alpha is chosen precisely? Is it the same at each point?

Yes, alpha stays the same at each point. The optimal value of alpha (2) was chosen empirically, to obtain the best possible balance between the importance of the minimisation of differences between daily and hourly ALADIN minima and maxima and the minimisation of the jump between two consecutive days.

L255 (point 8 actually). I don't really understand step 8. It seems that, first, total precipitation is adjusted. Then there is a phase separation given temperature and then rainfall and snowfall are readjusted separately (only in a variant it seems later in the paper)? Please improve the clarity of the description of this step (rationale and methodology).

Yes, this is what we do. A method separating rain and snow before adjustment was tested, but it did not yield satisfying results.

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):

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

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

L279. I don’t see in section 2.4 where the different learning periods are introduced.

You're right. They're introduced in Section 2.7 ! The sentence (l. 324-325) was corrected :

« The two RCM grid points neighbour selection techniques and the three different learning periods (1980-1995, 1995-2010 and 1980-2010, see Sect. 2.7) were tested. »

L286 "determined only for each massif". I’m not sure to understand.

We meant we did not calculate the altitudinal gradient for the Northern and Southern Alps (which would make little sense). However, we have now removed figures for the Northern and Southern alps, thus we changed the sentence (l. 332) into :

« – the mean value for each elevation band over 1980-2010 ; »

L349. The "evidenced"? The entire sentence is awkward.

« evidence», this was corrected. The sentence (l. 400-402) was changed to :

« This section provides the evidence needed to assess the ability of the ADAMONT method to reproduce the statistical characteristics of SAFRAN for temperature, precipitation and snow depth from daily RCM outputs. »

L352. "average altitudinal gradient"? I see the averages for each elevation band in this figure: the gradients are not plotted.

This was corrected (l.403-405 and throughout the manuscript):

« Figure 3 presents the location of the Vercors massif and its average temperature, precipitation and snow depth values for each elevation band (...) »

L415. "After 1 month of integration"? This formulation is not very good, I think.

This was corrected (l. 473 and throughout the manuscript):

« For integration windows of one month or more, (...) »

L472. It is really useful to plot hundred of time series (in the main document)? I think some integrated scores would be much better. Temporal averages in addition to the correlations shown later would be largely sufficient, I think.

Please see our previous response concerning figures (3rd R2 General comment).

L504-505. I don't think that the good scores are mainly due to the adjustment method. The small size of the RCM domain is likely the main responsible for the good correlations. With a small domain, the RCM results are very constrained by the boundary conditions as noted by the authors in a different context. The affirmation is therefore misleading (and references would be needed in any case).

Please see our response to a previous comment on this specific point (R2 Specific comment L127-128). Indeed, the correlations we evaluate are due to the whole system, not only to the ADAMONT method.

L550-551. Why? The authors do not explain how they deal with extremes values in their algorithm (there are many possibilities. . .). It is therefore difficult for the reader to understand this affirmation.

Yes, R2 is right. We have now included more details about how we deal with the tail of our ditributions in Sect. 2.3 step 5 (see our previous comment to R2 Specific remark L173).

L564. The temporal transferability is only very partially tested. To my opinion, it is not a major problem that the mean state changes with the learning period. What really matters is whether the trends or the differences between two periods change with the reference period. This is not assessed in the paper, and I think this point should be made.

R2 is right. However, trends in SAFRAN are too uncertain and affected by the heterogeneity of the assimilated data to be evaluated in details (Vidal et al., 2010). At least, the differences between the periods can be deduced from the figures.

L630. As noted previously, it does not really make sense to compare the results of different adjustment methods applied to different domains (and RCMs). The differences of performances are more likely to result from the differences of models and domains than from the adjustment methods…

In our study, the domain is larger than the ones used in Lafaysse (2011) and Lafaysse et al. (2014), and we obtain similar or even better results. So, if we consider your previous remark that larger domains give larger biases, it means that our method is at least as performant as the ones of Lafaysse (2011) and Lafaysse et al. (2014). Moreover, the latter methods use only reanalyses and statistical downscaling (without using RCMs), thus current RCMs and the ADAMONT method perform at least as well, regardless of the recent evolution of RCM perfomance.

L652-656. As the sentence is written, one may think that the authors want to apply the adjustment method over the entire Europe. Is it really the case? (which data-set would be used instead of SAFRAN in this case?). Or, they simply want to use RCM simulations with a larger domain, as I suspect?

Yes, the last proposition is correct, we simply want to use RCM simulations with a larger domain (i.e. EURO-CORDEX domain) and apply it to the French mountainous regions (Alps, Pyrenees, Massif Central, Jura, Vosges, Corsica). The sentence (l. 727-730) was re-written :

« In the framework of EURO-CORDEX, as we will be working with RCMs driven by GCMs, the objective, on the contrary, will be to focus on a larger RCM domain covering all of Europe, in order to analyse results over the French Alps depending less on the biases of GCMs. »

L657. "RCM model" : The M of RCM stands for model.

Yes, R2 is right. We removed the term « model ».