Answer to R. Essery

We thank R. Essery for his insightful comments. We answered below to all his points. His comments are in bold while our answers appear in normal font. Changes in the manuscript appear in red.

- **Queno et al. present an interesting evaluation of high-resolution snowpack simulations in a mountainous region with a reasonably high density of observations. There is quite a lot of overlap in descriptions of the snow model, NWP model and analysis system with a paper by the same authors cited herein (Vionnet et al. 2015b), which addresses some similar issues in a different region. Some repetition will be inevitable to allow the papers to be read independently, but if both are to be published it is the differences between them that will be most interesting.**

The reviewer is right. There are some similarities between Vionnet et al. (2015b) and our manuscript since they both deal with snowpack modelling issues over mountainous areas using atmospheric forcing from a NWP model. However, the two papers are rather complementary because each one brings a detailed analysis of the spatial variability related to the geographical location of mountains: results over the Alps (discussed in Vionnet et al.) can hardly be generalized to the Pyrenees mountains. Our study focuses on an extended assessment of the quality of snowpack simulations in the Pyrenees, regarding snow depth and SWE point evaluation, snow cover spatial variability, accumulation and ablation processes. On the other hand, Vionnet et al. (2015b) focus firstly on the capabilities of AROME to accurately represent the complex atmospheric variability in the French Alps in wintertime and presents an extended discussion on NWP modelling in complex terrain. Snowpack simulations driven by AROME are then evaluated only against ground-based measurement of snow depth.

Since the snowpack model Crocus and the high resolution NWP model (AROME) are used in both papers, it is quite difficult to avoid redundancies between the two articles which may occur in the description section. We consider that a detailed description of data/models is necessary so that the paper can be read independently. However, we managed to synthesize this section since the atmospheric forecast is not the main focus of this study: the description of AROME physics and data assimilation schemes was deleted, and replaced by a reference to the paper by Seity et al. (2011), which gives a comprehensive description of the AROME model.

--- CHANGES IN MANUSCRIPT (line 152) ---
A detailed description of the physics and data assimilation schemes can be found in Seity et al. (2011). In particular, the precipitation phase is derived from the cloud microphysical scheme.

Moreover and in agreement with referee #2 suggestion, a new section dedicated to a quantitative evaluation of simulated snow cover distribution with respect to MODIS snow cover fraction images has been added. It includes a table synthesizing the mean similarity scores by domain and winter, and a figure representing the evolution of daily similarity scores during the winter 2011/2012. A detailed description of these results is provided bellow.
This new study provides new and relevant evidence on the quality of the snow simulations using AROME-Crocus not yet addressed in other publications.

--- CHANGES IN MANUSCRIPT (line 229) ---

The Jaccard index (J) and the Average Symmetric Surface Distance (ASSD) are two similarity metrics which were used to compare simulated and remotely sensed snow covered areas. They were applied to simulated and observed binary snow covered maps on the same grid. J takes into account every pixel of the surfaces A (simulated snow cover domain) and B (observed snow cover domain), and is thus dependent on the whole snow covered area:

$$J = \frac{|A \cap B|}{|A \cup B|}$$

J ranges from 0 to 1, where 0 235 means no overlap of A and B surfaces, and 1 means A = B. The ASSD is complementary to J since it evaluates a mean distance between the boundaries of the two surfaces. It is based on the Modified Directed Hausdorff Distance between boundaries L_A and L_B, defined by Dubuisson and Jain (1994) as the average distance of the points of L_A to L_B:

$$MDHD(A,B) = \frac{1}{|L_A|} \sum_{a \in L_A} d(a, L_B)$$

where $d(a, L_B)$ is the Euclidean distance between point a and the closest point of boundary L_B:

$$d(a, L_B) = \inf_{b \in L_B} ||a - b||$$

The MDHD is a directed distance, used by Sirguey (2009) for snow patterns matching. The ASSD is its symmetrised version:

$$ASSD(A,B) = \frac{MDHD(A,B) + MDHD(B,A)}{2}$$

It ranges from 0 to +∞, where 0 means $L_A = L_B$. In practice, the maximum value is the highest possible distance between two points of the domain.

Binary maps are built using a 20 mm SWE threshold for simulations and a 50% snow fraction threshold for satellite data. The metrics are calculated only when the cloud fraction on the domain is less than 10% and the snow cover represents at least 10 pixels in MODIS images interpolated on AROME grid (the size of a pixel is 0.025° x 0.025°, i.e. approximately 6.25 km²).

--- CHANGES IN MANUSCRIPT (line 341) ---

4.1.3 Snow cover distribution

The comparison between AROME–Crocus, SAFRAN–Crocus and MODIS snow cover distribution is extended to two entire winters: 2011/2012 (characterized by an average deficit of snow) and 2012/2013 (extremely high amount of snow). Table 3 summarizes two metrics (ASSD and Jaccard index) that evaluate the match of simulated and observed snow covers in different domains. AROME–Crocus scores are better than SAFRAN–Crocus for the whole Pyrenees (higher Jaccard index and lower ASSD for both seasons). This is also true for the Spanish, central and eastern domains, whereas scores are equivalent for France. SAFRAN–Crocus performs better in the western Pyrenees. The seasonal evolution of scores over this domain (not shown) indicates that both models have equivalent skills during the accumulation season, while SAFRAN–Crocus performs better during the melting season. This result is consistent with the results of section 4.1.1: AROME–Crocus strongly overestimates snow quantities in the western Pyrenees, which results in a later presence of snow on the ground in the Springtime.

Table 3. Seasonal means of daily Jaccard index and ASSD for simulated snow cover distribution against MODIS observations in the Pyrenees for winters 2011/2012 and 2012/2013. The best scores are given in bold.

year	domain	N	Jaccard index		ASSD (pix.)	
			AROME	SAFRAN	AROME	SAFRAN
2011-2012	all	57	**0.47**	0.40	**1.34**	1.64
	France	57	0.51	**0.55**	0.91	**0.76**
	Spain	56	**0.42**	0.28	**1.27**	1.88
	West	56	0.45	**0.48**	1.34	**1.04**
	Center	57	**0.51**	0.39	**1.08**	1.64
	East	56	**0.42**	0.31	**1.27**	1.98
2012-2013	all	39	**0.40**	0.36	**1.73**	2.00
	France	39	**0.44**	**0.44**	**1.52**	1.61
	Spain	35	**0.39**	0.32	**1.52**	2.05
	West	37	0.43	**0.45**	1.36	**1.12**
	Center	38	**0.43**	0.37	**1.31**	1.66
	East	26	**0.42**	0.32	**1.37**	1.75

Figure 6 shows the evolution of daily ASSD and Jaccard index for winter 2011/2012 over the whole Pyrenees (within SAFRAN massifs). Both scores attest that AROME–Crocus improves the representation of the spatial snow cover distribution compared to SAFRAN–Crocus until late March. SAFRAN–Crocus shows a slightly better agreement than AROME–Crocus after late March, i.e. at the beginning of the melting season due to the overestimation of snow quantities by AROME–Crocus. On 22 February 2012 (date studied in the previous section, Fig. 4), J = 0.61 and ASSD = 1.22 pixels for AROME–Crocus, while J = 0.40 and ASSD = 2.09 pixels for SAFRAN–Crocus, which quantifies the better agreement seen in Fig. 4.

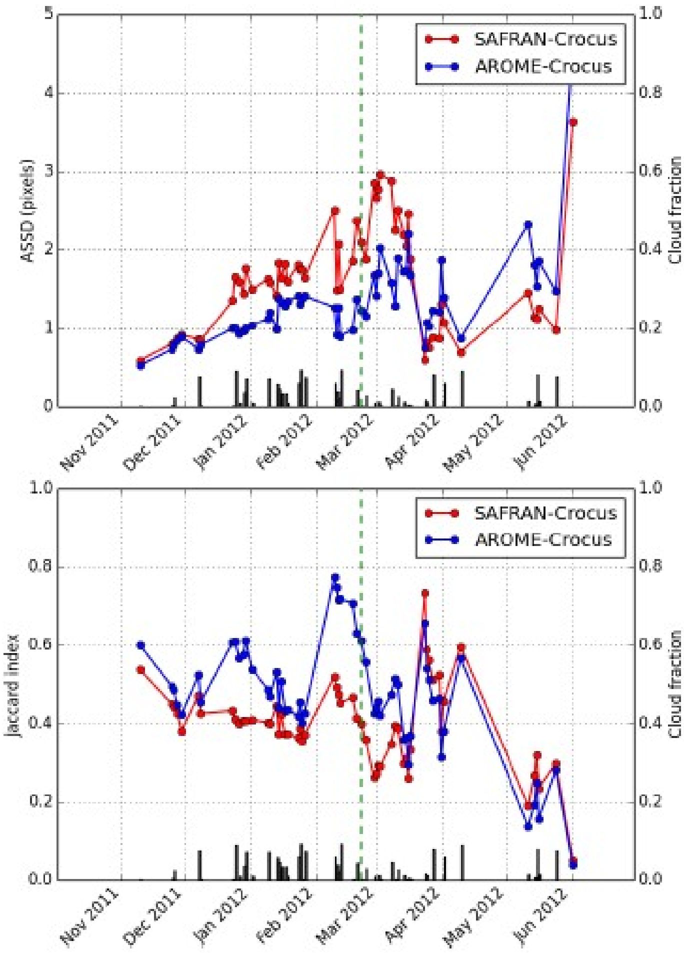

Figure 6. Daily ASSD and Jaccard index, within all massifs, AROME–Crocus vs MODIS (blue) and SAFRAN–Crocus vs MODIS (red), 2011-2012. Smaller ASSD and higher J mean better match with MODIS. The green line indicates 22 February 2012. The cloud fraction is represented by the black bars.

--- CHANGES IN MANUSCRIPT (line 546) ---

AROME–Crocus exhibits a better snow spatial distribution than SAFRAN–Crocus with respect to MODIS images of snow cover fraction. Similarity scores highlighted a better agreement of snow covered areas for AROME–Crocus, for two winters in most domains, except in the western Pyrenees where AROME snowfalls are too large. The added value of AROME–Crocus to represent the spatial variability of the snowpack within each massif was particularly emphasized on winter 2011/2012.

Additionally, the evaluation of precipitation forecast with HSS has been removed, since it did not bring major conclusions to the study.

- **This paper also has a lot of figures, sometimes with quite limited discussion; I think that some consideration could be given to the balance between figures and text.**

We agree with the referee. Figure 7 (cumulated daily SD variations by category) has been removed as suggested in another comment (Figure 6 – 7 after revision – and text were sufficient).

--- CHANGES IN MANUSCRIPT (line 382) ---
In terms of quantities, the categorical sums of ΔSD (not shown) indicate that SAFRAN-Crocus strongly underestimates the high accumulation quantities.

Figure 10 (HSS for precipitation) and the associated paragraph have been removed, as explained previously.

Figure 15 (cumulated daily SWE variations by category) has been removed. Overall, daily SWE variations study did not bring new conclusions, compared to the daily SD variations study. Furthermore, despite the 24h-median smoothing, some noise remained in some time series which increased the uncertainty of the values (compared to daily SD variations).

One figure has been added (new section) since it brings relevant information to our study, so the revised version of the paper has two figures less.

- **line 35. Redistribution of snow by avalanches would also be worth mentioning here.**

Redistribution by avalanches has been added, as suggested.

--- CHANGES IN MANUSCRIPT (line 32) ---
At a smaller scale (less than 100m), processes like wind-induced erosion (Pomeroy and Gray, 1995), avalanches (Schweizer et al., 2003) or preferential deposition of snowfall on the leeward slopes (Lehning et al., 2008), play a decisive role on snow distribution (e.g. Mott et al., 2010).

- **Figure 1. To divide the Pyrenees into western, central and eastern regions, it might seem more obvious to have Haute-Bigorre in the central region and Haute-Ariege and Andorra in the eastern region. Is the division trying to distinguish north-south gradients also?**

Western, central and eastern regions are defined following the climatological study of Maris et al. (2009), and unpublished studies of CNRM/CEN based on SAFRAN reanalyses. As most of the disturbed flows constituting the snowpack in winter are NW/N flows, this division indeed includes north-south gradients.

- **The acronym "SD" is used in the Figure 1 caption but not explained until line 194**

An explanation of the acronyms SD and SWE has been added in the introduction.

--- CHANGES IN MANUSCRIPT (line 91) ---
Section 4 details the results following three main axes: (i) global scores and spatial distribution of snow depth (SD); (ii) daily snow depth variations and winter precipitation; and (iii) comparison to snow water equivalent (SWE) scores and study of bulk snowpack density.

- **216. RMSE is barely mentioned hereafter. Knowing two out of bias, RMSE and STDE, the other one can be determined; what is the point of considering all three?**

The reviewer is right. Only bias and STDE have been kept. RMSE was used shortly in section 4.2.1, it has been replaced by STDE and bias.

--- CHANGES IN MANUSCRIPT (line 207) ---
Two error metrics were used: the bias and the Standard Deviation Error (STDE, which represents the temporal and spatial dispersion around the bias).

--- CHANGES IN MANUSCRIPT (line 364) ---
The STDE of daily ΔSD indicates the ability of the model to forecast (or analyse) the appropriate daily evolution of snow depth. This score was computed for AROME–Crocus and SAFRAN–Crocus. It is equal to 7 cm (and bias equal to 0 cm) for both models, with low spatial variation. The STDE is slightly higher in the most snowy winters (8 cm in 2012/2013 and 2013/2014 against 6 cm in 2010/2011 and 2011/2012).

- **Rainbow colour schemes, as used in Figures 3 and 4, are deprecated.**

The colour schemes of these figures have been changed.

- **Figure 5. The cross section passes close to 2 or 3 precipitation measurement stations on the French side. Could these measurements be shown?**

Precipitation measurement stations have not been included for two reasons: (1) they are located at rather low altitudes and cannot discriminate the precipitation phase; (2) the undercatch issue of precipitation gauges. However, we have chosen three SWE measurement stations located close to the transect, and with a similar exposure to the flows as the modelled topography. Cumulated snowfalls have been derived from cumulated positive daily ΔSWE. These measurements (and their actual altitude) have been added to the cross section in Fig. 5, as well as a short comment of these observations in the text.

--- CHANGES IN MANUSCRIPT (line 323) ---
They are represented in Fig. 5 along a NW/SE cross section, as well as cumulated positive ΔSWE from measurements of three stations close to the transect.

--- CHANGES IN MANUSCRIPT (line 333) ---
AROME simulations are in good agreement with the two Spanish stations, which are located at an altitude close to the model's topography. SAFRAN snowfalls are too low at the station closest to the border, but in good agreement at the second Spanish station. Observations for France are in better agreement with AROME than with SAFRAN, but still higher than both simulations. This may be due to the difference of altitude with the models.

--- CHANGES IN MANUSCRIPT (caption of Fig. 5) ---
Cross section of cumulated snowfall from 1 October 2011 to 22 February 2012 for AROME forecasts (blue) and SAFRAN reanalysis (red), with topography plotted on the right axis in grey. Cumulated positive ΔSWE from measurements of three stations close to the transect are represented with black dots; their actual altitude is represented with black stars. The locations of the transect (red) and stations (blue stars) are given on the upper right map.

- **Figure 7. I am not sure that this figure adds anything that is not already clear from Figure 6 and the text in lines 347 to 361.**

Figure 7 presented the differences in terms of quantities, but the text associated to Figure 6 (Fig. 7 after revision) may be sufficient. Figure 7 was removed, as suggested.

--- CHANGES IN MANUSCRIPT (line 382) ---
In terms of quantities, the categorical sums of ΔSD (not shown) indicate that SAFRAN–Crocus strongly underestimates the high accumulation quantities.

- **350. It is not clear what "mechanically counterbalanced" means here.**

The word "mechanically" was ambiguous and has been removed. A brief explanation has been added.

--- CHANGES IN MANUSCRIPT (line 385) ---
It is counterbalanced by an overestimation of small accumulation quantities, since an underestimated strong accumulation event is counted in the smaller accumulation category.

- **368-376. There is little interpretation of ETS beyond this paragraph. Are Figure 9 and the associated description of ETS essential to the paper?**

We used the ETS in the paper since it provides an overview of models skills. Figure 9 shows that SAFRAN-Crocus and AROME-Crocus have equivalent scores in terms of daily snow depth accumulation, AROME-Crocus being better for accumulations higher than 10 cm/day. The ETS has also been used by Schirmer and Jamieson (2015) for evaluating snow accumulations simulated by equivalent models (GEM-LAM/SNOWPACK). A direct comparison of both models configurations is thus possible. For these two reasons, we decided to keep the ETS in the revised version of the paper.

- **403. Contrasting accumulation error to precipitation error is not straightforward because it also involves modelled snow density (discussed later).**

This issue has been mentioned in the revised version of the manuscript.

--- CHANGES IN MANUSCRIPT (line 437) ---
The difference between accumulation and precipitation errors also involves modelled snow density: this issue is discussed in section 4.3.

- **420. Crocus has its own index of snow drift. Why is it not used? What difference would it make?**

R. Essery points out an interesting way to detect wind erosion. Crocus has indeed a snow drift index derived from surface wind (given by the input forcing) and a mobility index based on the modelled snow surface properties (Guyomarc'h and Mérindol, 1998). Such an index makes it possible to take into account snowpack properties additionally to wind speed in the determination of blowing snow occurrence. However, we have chosen not to use this index for several reasons.

Firstly, if we consider the index coming from our snowpack simulations, it is computed using modelled snow surface properties and 10-m wind as simulated by the atmospheric forcing (e.g. AROME). Simulated wind at 2.5-km grid spacing in mountainous terrain can largely differ from observed wind due to topographic features (ridges, depressions…) not reproduced at 2.5-km grid spacing. Using the AROME wind at the Maupas station would for example give wind erosion detection almost every day in winter, because the simulated wind is too strong.

Then, a solution would be to use wind observed at the automatic stations combined with simulated snow surface properties to derive a new snowdrift index. We computed this index (cumulated for each day) at Maupas station for winter 2012/2013. It is represented on Fig. R1-1 as black bars, together with the BSD detection (green). Both detections are in fair agreement, particularly for the strongest events.

However, we consider that taking into account the modelled surface properties does not necessarily improve the detection of blowing snow events. Indeed, snow surface properties have been modified by previous non-simulated wind erosion events, since there is no ablation by wind transport in Crocus. For instance, a 60-cm decrease of snow depth occurs in mid-December: the snowpack is totally different after this event which is not simulated; the subsequent snowdrift index may be far from reality.

Consequently, in the revised version of the manuscript, the wind erosion index is only based on observed wind and simulated melting.

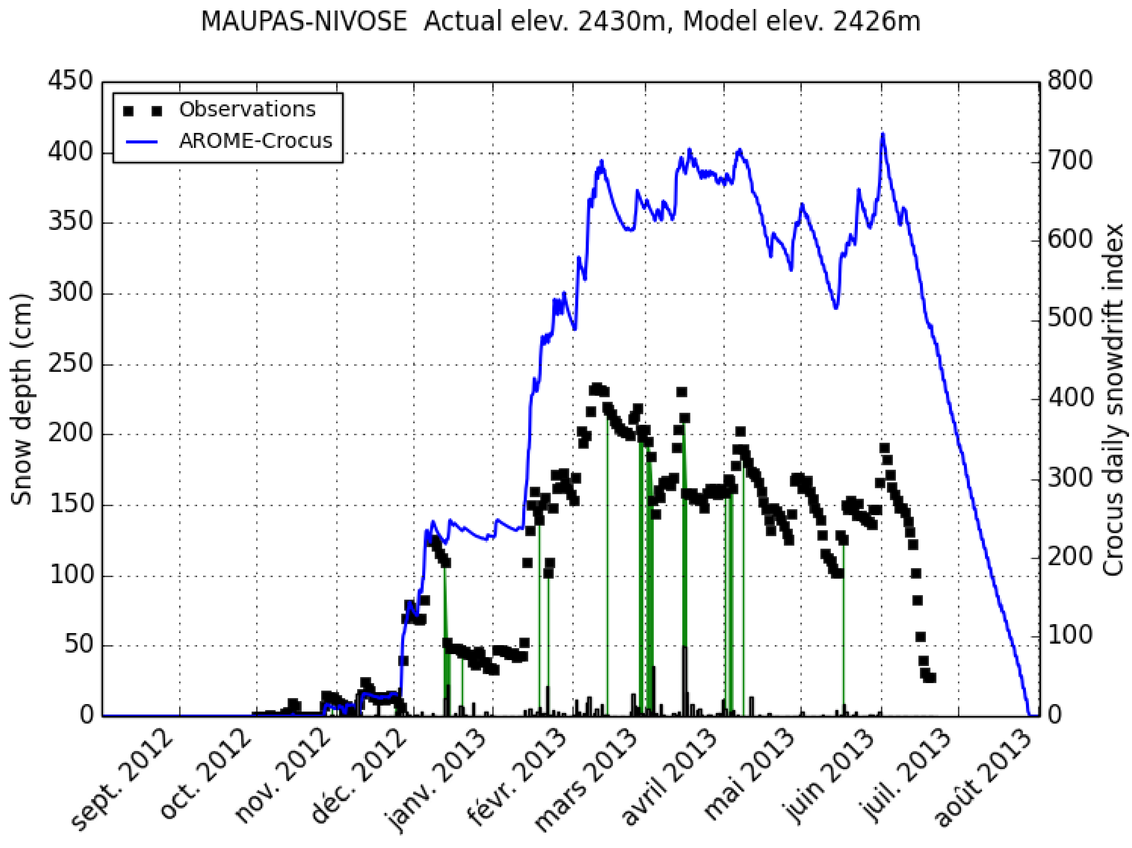

Figure R1-1: Snow depth simulated by AROME-Crocus (blue) and observed (black) at Maupas station, 2012/2013. BSD are identified in green and cumulated daily snowdrift index is represented by black bars.

- **Figure 12 clearly shows that the neglect of wind redistribution of snow in kilometrescale simulations accounts for some errors in comparison with point-scale measurements, but does this really matter? Apart from snow that sublimates in transit (which Crocus can estimate), the snow removed by wind will just end up in a**

drift somewhere else, likely within the same model grid cell. Snow redistribution is of course enormously important for loading on avalanche slopes, but that isn't being discussed here.

We thank R. Essery for this comment. Figure 12 (Fig. 11 after revision) is presented in this paper to illustrate how the computation of SD and SWE scores is affected by the occurrence of wind-induced snow transport at stations measuring SD and SWE. This figure clearly shows that wind redistribution strongly impacts observations used for validating the simulations and this must be kept in mind when discussing model results. We then totally agree with R. Essery concerning the fact that wind-induced snow redistribution cannot be represented on a regular grid at 2.5-km grid spacing. Snow redistribution by wind indeed occurs very likely within each grid cell. In the discussion part of the revised version of the paper, we now mention more clearly these two different points.

--- CHANGES IN MANUSCRIPT (line 584) ---
We first showed that wind-induced erosion of the snowpack constituted the major cause of the underestimation of strong ablations at seven high altitude stations. This small-scale process cannot be captured by a kilometric simulation of the snowpack, since snow redistribution by wind occurs very likely within each grid cell. But the computation of SD and SWE scores is affected by the occurrence of wind-induced snow transport at stations. The impact of blowing snow could not be estimated at all stations. It is probably less significant at lower altitudes.

- **The English is always good enough to be understood but will require editing to be perfect. There are several constructions of the type "allows to capture" and "allows to avoid" often used by French authors writing in English; "allows capturing" and "allows avoidance of" or just "captures" and "avoids" would be better English. "deplored" (212) is a rather strong term; "adequation" (532) and "prescind" (581) are English words but very uncommon ones – there will be better choices.**

The new version of the manuscript has been edited by a native speaker.