Answer to Referee #2

We thank the referee 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.

1 General comments

1.1 Summary of the manuscript

Queno et al. are evaluating an analysis model and a numerical weather prediction (NWP), respectively, to be able to force a snow cover model at up to 130 stations in the Pyrenees. A time period of four winter seasons was analysed covering very different winters. Their goal was to compare the quality of a 2.5 km resolution NWP model with a coarser analysis system in terms of spatial variability, timing and amount of precipitation, ablation processes and settlement, and amounts of snow depth (HS) and Snow Water Equivalent (SWE). They concluded that the NWP and analysis system produced a positive bias in snow depth, which resulted from an underestimation of accumulation and a larger underestimation of ablation fluxes. Especially large fluxes were particularly underestimated. For decrease of HS they addressed issues causes by melt, wind and settling separately and concluded that wind erosion was responsible for the largest error during ablation. In general the fine resolution NWP model was found to be better in many analysed aspect.

1.2 Overview of the review

Queno et al. addressed an interesting topic for mountain snow hydrology or avalanche research. Reliable input of solid precipitation and resulting SWE or HS states, is crucial for applications in mountainous terrain, and studies covering long time periods are rare. Also, ablation and densifications processes also interesting to evaluate. The study addressed different error sources, i.e. the meteorological forcing, the snow cover modelling, not included processes, and observation errors. While timing and amount of fluxes and amount of state variables were quantitatively analysed, the conclusion of a better spatial representation by the finer resolution NWP model was analysed qualitatively at one single (but interesting) point in time. The better spatial representation of the NWP forcing is one major conclusion and thus this analysis needs to be enhanced.

We thank the referee for his suggestions. We have chosen to increase the impact of the manuscript with a more extensive study of the snow cover spatial distribution. This study is detailed in the answer to the specific comment 2.1.

The impact of the manuscript can be enhanced using different NWP models as meteorological forcing to additional snow cover models with different settlement or melt implementations, which will allow users of those kind of models to choose accordingly. Another increase in impact could be achieved with addressing reasons for errors during melt and settling.

This study focuses on the use of kilometric resolution NWP models as meteorological forcing to a snowpack model. Over our domain of study (the Pyrenees), only the AROME model is available.

Concerning the snowpack model, we work with Crocus, because our study is performed with a view to operational avalanche forecasting. The aim of this study is not to discuss the quality of results depending on the complexity of the snowpack model. Furthermore, addressing the reasons of the errors during melting and settling requires an extensive study of Crocus physics formulations, which would go beyond the scope of this paper.

Meteorological variables responsible for errors provided by the NWP model or the analysis system could be evaluated, as solar radiation or air temperature.

An extensive evaluation of meteorological variables forecasted by AROME in alpine terrain has been already performed by Vionnet et al. (2015b). We have decided to keep the focus of the manuscript on snowpack modelling, through the assessment of snowpack-related variables only. In the manuscript, we also refer to the uncertainties due to the meteorological forcing (precipitations for snow accumulation, incoming radiations for melting...).

Snow model runs with meteorological weather stations instead of modelled input data would be a solution to discriminate error sources between meteorological forcing or subsequent snow cover modelling.

We thank the referee for this comment. Using meteorological weather stations as input to the snowpack model is indeed an interesting way to discriminate error sources. However, there is no station in the Pyrenees providing all the measurements necessary for the atmospheric forcing of Crocus. The only station suitable for such a study in the French mountains is the Col de Porte located in the French Alps.

These suggestions would also decrease the similarity to a cited non-published study including many of the authors of this manuscript (Vionnet et al., 2015b). I think this manuscript is worth publishing despite these similarities after addressing the comments mentioned below.

There are language and spelling issues, so I suggest an accurate editing by a native speaker.

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

2. Specific comments

2.1 Spatial variability of snow depth

The spatial variability of snow depth was evaluated with Figure 4 in comparison with snow cover fraction at a single point in time. This is indeed an interesting situation, but a more quantitative comparison is needed to conclude that AROME delivers a more realistic spatial variability. First, station observations can be used for this situation of large differences between South and North, pooled in two groups, for example. This would decrease the problem that only snow depth variability and snow cover fraction is

compared. Second, one more year can be easily be included. Third, depletion curves can be derived between observed and modeled snow cover fraction.
So far, the authors only discussed precipitation amount differences between SAFRAN and AROME for differences in spatial variability of snow depth. The authors may also comment on differences in precipitation phase or in melt processes, which probably happened repeatedly at lower elevations on the Spanish side.

We thank the referee for this very relevant comment. The section dealing with the spatial variability of snow cover has been updated as suggested by the referee. We have completed the study which only described initially a single date of winter 2011/2012. In the revised version of the paper, we present a more quantitative study of AROME-Crocus and SAFRAN-Crocus representation of the snow distribution, through comparisons to MODIS snow cover images during two winters (2011/2012 and 2012/2013). Two new scores have been used to evaluate how simulated snow cover agrees with the MODIS satellite images: the Average Symmetric Surface Distance (average distance from one snow line to the other) and the Jaccard index (evaluating surfaces matching). Both are presented in the manuscript because the ASSD describes more the correspondence of snow lines while the Jaccard index is more representative of the total areas. We get the same results with the two metrics: AROME-Crocus better represents the snow cover distribution than SAFRAN-Crocus. The new section includes a table synthesizing the mean similarity scores by domain and winter, and two figures representing the evolution of daily similarity scores during winter 2011/2012.

--- 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.

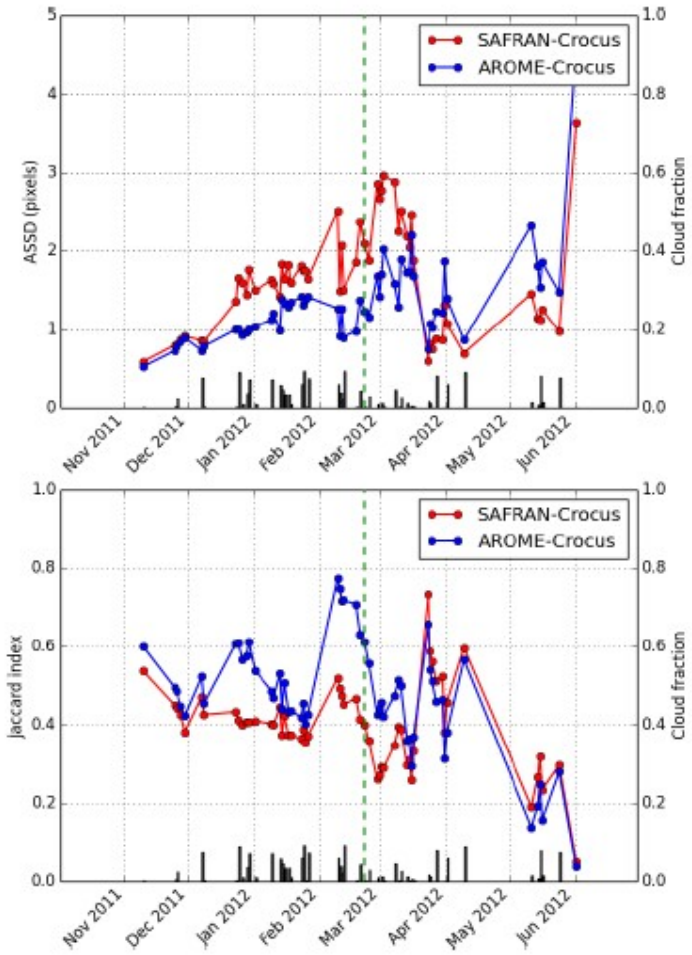

Figure 6. Daily ASSD and Jaccard index, within all massifs, AROME–Crocus vs MODIS (blue) and SAFRAN–Crocus vs MODIS (red), 2011-2012. Smaller ASSD and higher *J* mean better match with MODIS. The green line indicates 22 February 2012. The cloud fraction is represented by the black bars.

--- CHANGES IN MANUSCRIPT (line 547) ---

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.

2.2 Wind erosion major cause for underestimating ablation or decrease in HS

To my opinion it is not clear that wind erosion is the major cause with presented results (line 552). The authors need to be more precise when discussing the data to draw this conclusion. One concern in this regard is that in Figure 13 only a small subset of stations are used, for which wind effects are anticipated. This makes it difficult to compare errors caused by melt or wind erosion.

In order to better highlight the contribution of wind erosion to strong decreases of snow depth observed at these seven stations, we have added a quantitative discussion of the results exposed in Fig. 13 (Fig. 12 after revision). We show that wind erosion constitutes 71% of high decreasing rates. There is no overlap of blowing snow days (BSD) with melting snow days (MSD), which means melting is part of the 29% remaining.

For the sake of clarity, we have plotted BSD (instead of all days excluding BSD) on Fig. 13 (Fig. 12 after revision). The same representation has been chosen for MSD in Fig. 14 (Fig. 13 after revision). Similarly to the BSD study, we have shown that MSD represented 42% of high decreasing rates at all stations.

Concerning the smaller subset of stations used for wind erosion study, this is due to the fact that only automatic stations measure wind speed. We have shown that wind erosion is the major cause for underestimating strong ablations for these seven stations located at high altitude (mean altitude: 2203 m.a.s.l). The referee is right that we don't have enough data to conclude that it is the major cause at all stations. Indeed, the contribution of blowing snow may be less significant at lower altitudes. We have qualified this assertion in the discussion.

--- CHANGES IN MANUSCRIPT (line 457) ---

To quantify the impact of wind-blown snow events on the performance of models, the cumulated ΔSD for AROME–Crocus and observations are plotted in Fig. 12, for BSD and all days, with a finer categorization of SD decreases. This study is restricted to seven automatic stations measuring wind speed and SD (mean altitude: 2203 m.a.s.l). For observations, BSD contribute to all decreasing rates, in the strongest proportion for high decreasing rates (less than -20 cm). For AROME–Crocus, BSD do not contribute to the strong ablation categories but to small ablation and accumulation categories in the same proportions. Cumulated ΔSD for high decreasing rates is equal to -1106 cm for all observations, and equal to -781 cm for BSD only (excluding MSD), while it is equal to 0 cm for AROME–Crocus in both cases. It means that wind-blown snow is the main contributor (71%) to this category, the remaining contribution coming from MSD or other processes.

Similarly, the cumulated ΔSD is plotted in Fig. 13 for MSD and all days, at all SD stations. Very strong melting (more than 20 cm.day^{-1}) is seldom observed, but never predicted. Strong melting (between -20 20 cm.day^{-1} and -10 20 cm.day^{-1}) is much under-represented by models, while melting of less than 10 20 cm.day^{-1} is over-represented. Cumulated ΔSD for high decreasing rates (more than 20 cm.day^{-1}) is equal to -7741 cm for all observations, and equal to -3215 cm for MSD only, while it is equal to -41 cm for AROME–Crocus in both cases. Melting snow represents 42% of this category, the remaining contribution coming from BSD or other processes. The behaviour of SAFRAN–Crocus is similar to AROME–Crocus for BSD and MSD (not shown). The simple diagnostics of BSD and MSD may miss some blowing-snow or melting events.

Consequently, the underestimation of strong decreasing rates comes mainly from ablation processes: on the one hand, from wind-blown snow events which are not represented by models, as they are small scale processes; and on the other hand, from an underestimation of strong snowpack melting (more than 10 cm.day^{-1}). Other reasons for very high decreasing rates can be the strong settling after an intense snowfall or a rain-on-snow event, but it probably constitutes a limited part of this category.

--- 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.

2.3 Similarity to Vionnet et al. (2015b)

The same model setup was evaluated not in the in the Pyrenees but in the French Alps by Vionnet et al. (2015b). They also evaluated spatial distribution of snowfall similarly to Figure 4 and 5 in this manuscript. They also assessed categorical scores of daily precipitation. Additional aspects of this manuscript are analysed processes of ablation and settling. This manuscript also uses SWE and HS measurements, additionally to precipitation gauges, to evaluate accumulation. After enhancing the spatial variability part I suggest that this study is publishable additionally to Vionnet et al. (2015b). Other strategies to enhance the impact of this manuscript (see section 1.2) will further discriminate the both studies.

The reviewer is right. R. Essery in his review pointed out the same aspect and we reproduce below the answer that we gave to R. Essery.

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.

Like the focus on accumulation and ablation processes, the study of snow cover spatial distribution of simulations vs MODIS images (Fig. 4) is specific to our paper. Cross sections of simulated snowfalls (Fig. 5) are also presented by Vionnet et al. (2015b), but it is used in the present paper as an interpretation of the differences of snow cover distribution between AROME-Crocus, SAFRAN-Crocus and MODIS.

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

Following the referee's suggestion, a new quantitative study of spatial variability has been added (described previously).

3. Technical comments

Figure 15: Number of observations are missing.

Following a comment of R. Essery, Figure 15 has been removed to compensate the increasing number of figures due to the new section about spatial variability and to improve the balance between figures and text in this paper. 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).

How is the precipitation phase determined? As output from the NWP model and analysis system or with the by the snow model?

Snowfall and rainfall are distinguished as outputs of the NWP model (from the cloud microphysical scheme) and the analysis system (threshold T_{2m}=1°C). A mention of this issue has been added in the description of models.

--- CHANGES IN MANUSCRIPT (line 153) ---
In particular, the precipitation phase is derived from the cloud microphysical scheme.

--- CHANGES IN MANUSCRIPT (line 177) ---
The precipitation phase is derived from a simple threshold of 1°C air temperature at 2 m above the ground.

The problematic observations of precipitation gauges can be better defined and speculations of the precipitation phase could be reduced if the analysis of Figure 11 would also be performed only for days when snowfall is likely (cold, or dependent on NWP model output).

A brief mention of the effect of wind on snowflakes trajectories has been added, in supplement to the reference to literature, which seems sufficient for further details.

--- CHANGES IN MANUSCRIPT (line 432) ---
The undercatch of solid precipitations by gauges, mainly due to wind effects on falling snowflakes trajectories, is well known and very variable. This issue is investigated by the WMO Solid Precipitation InterComparison Experiment (e.g. Wolff et al., 2015).

As suggested by the referee, the analysis of Figure 11 (Fig. 10 after revision) has been restricted to "cold" days (i.e. daily maximal 2m-temperature lower than 2°C), when snowfall is more likely (rainfall now represents only 6% of total AROME precipitation). With this criterion, the study period has been extended from October to June (instead of December-April).

--- CHANGES IN MANUSCRIPT (line 413) ---
A complementary information on winter precipitation comes from the network of gauges in the French Pyrenees (red dots in Fig. 1). Daily accumulations of precipitation (rainfall plus snowfall, cumulated from 6UTC to 6UTC) from the forcing models are then directly compared to precipitation gauges measurements, for days with a maximum temperature of 2°C in order to reduce the proportion of rainfall amongst precipitation. Most of these observations are assimilated in SAFRAN reanalyses, while they are not taken into account in AROME forecasts. Figure 10 shows cumulated precipitation by category for both models and observations (right) compared to cumulated ΔSD at the same stations (left). Contrary to ΔSD, AROME overestimates precipitation measured by gauges (+ 73 %). The optimal interpolation basis of the SAFRAN analysis system should mathematically not be biased on the assimilated observations over a long period. The slightly positive bias obtained in this study (+ 17 %) may be linked to the fact that some assimilated observations are not included in our evaluation dataset and/or to differences between the climatological guess and the mean precipitation amount of the 4 years under study. The strong overestimation of AROME is particularly notable for the largest amounts. The different distribution of precipitation and ΔSD for AROME, with a higher proportion of strong precipitation than of strong snow accumulations, may be due to settling effects: the stronger the snowfall, the stronger the snowpack settles under its own mass, which shifts the distribution to the left.

Why do the authors use the HSS for the evaluation with precipitation gauges and the ETS for snow depth sensors? This reduces the direct comparison between the both evaluation measures.

We agree with the referee that using two different scores could bring some confusion. The HSS was used for precipitation evaluation in order to facilitate comparisons with other NWP precipitation evaluations, and particularly with AROME precipitation evaluation in the French Alps by Vionnet et al. (2015b). The ETS was used for daily snow depth variations evaluation in order to allow direct comparison with the categorical study of snow accumulations by Schirmer and Jamieson (2015). This comparison is particularly relevant since equivalent NWP and snowpack models were used (GEM-LAM and SNOWPACK).

Following a comment of R. Essery concerning the balance between figures and text in the paper, the evaluation of precipitation through the HSS has been removed since it did not bring major conclusions to the article.

Line 255: The authors could later provide a summary for reasons causing this high standard deviation error.

The STDE represents the temporal (within a season and between seasons) and spatial (between stations) dispersion around the bias. The underestimation of the intensity of daily snow depth variations may explain a high STDE: daily variations are not well reproduced which implies a daily variation of the bias, and thus a higher dispersion. This issue has been mentioned in the discussion as suggested by the referee.

--- CHANGES IN MANUSCRIPT (line 601) ---
Consequently, all processes contributing to the decrease of the snow depth are underestimated, in a stronger proportion than for accumulations, which leads to a global overestimation of snow depths, through a smoothing of extreme variations. These opposite biases artificially imply a smaller bias for SAFRAN--Crocus than for AROME--Crocus. The underestimation of the intensity of daily variations also implies daily variations of the bias, hence a high dispersion around the mean bias, which partly explains a high STDE.

Two many figures. I would suggest to delete Figure 12 since there is no additional value shown, and either Figure 16 or 17.

We agree with the reviewer. Figure 7 (cumulated daily SD variations by category) has been removed as suggested by R. Essery (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 has been removed as explained previously.

One figure has been added (new section), so the revised version of the paper has two figures less.

Line 577 and in References: Gruenewald and Lehning (2015) must be 2014.

The article was first published online in 2014, but the actual date of publication is 2015: http://onlinelibrary.wiley.com/doi/10.1002/hyp.10295/abstract