Repsonse to reviewer #1

This study focuses on the seismic detection and identification of the signals generated by snow avalanches in the Davos area in the Swiss Alps during the winter 2010. The authors tested the capability of a machine learning algorithm (hidden Markov models - HMM) to perform this detection and identification from continuous seismic data. They used a reference catalog to evaluate the performance of the algorithm. The first results showed that the algorithm is capable to achieve relatively high positive identification rates of the avalanches in the catalog (70-95% depending on the station that recorded the signals), but also with a high rate of supposedly false detections. This led the authors to propose a post-processing strategy. Three post-processing steps were investigated: (i) analysis of the duration of the signals; (ii) computation of a correlation factor to evaluate the coherence of the signal between each sensor; and (iii) a voting system based on the classification returned by each station for a given event. Using one, or a combination of those proposed post-processing steps, led to a decrease of false alarm rates, but also in most cases to a decrease of the rate of good identification.

The use of seismology to study environmental processes is of growing interest as it allows producing observations with a unique spatio-temporal resolution. This new approach can help to better understand the triggering factors of natural hazards and to mitigate their consequences on our societies. In this context, this study contributes to the continuing effort to develop robust and versatile methods to explore years of continuous data and for the implementation of real-time seismology-based warning systems. Overall, I think the paper is clearly written, and that the Authors have made a good effort to carefully explore the data, explain their approach and discuss their results. Nevertheless I listed below several comments and suggestions that might help to improve this paper.

General comment:

The only major concern I have regarding this work is that the downsides of using the HMM algorithm are not discussed while most of the results presented in this paper suggests that HMM alone, without pre- and post-processing, cannot perform identification of seismic sources with a high success rate. The strengths of the HMM are usually stated to be:

- i) it does not need any pre-detection or picking (STA/LTA, etc.), which should ensure that no event is missed;
- ii) it does not require any inputs from experts.

Yet this paper demonstrates that

- i) a pre-detection can be suitable to remove low-amplitude/noise signals (figure 2);
- ii) post-processing steps with thresholds set by experts (duration, etc.) is needed to achieve a high accuracy.

Moreover, the Authors are building their post-processing strategy based on features that can be incorporated in the identification models constructed with other machine learning algorithms. The post-processing steps the Authors propose seem necessary because the HMM cannot include these features (durations of the signals, coherence between signals recorded at different stations, vote among stations) in the model due to its core design, which is to consider chunk of continuous data and not the entire signal generated by the event. This forces the Authors to manually set thresholds on those features, while with some other algorithms those thresholds are determined through a statistical analysis of the reference data. I think the authors must include a more thorough and objective discussion on the pros (which definitely exist) and the cons of HMM compared to other algorithms/studies in the light of the results of this work.

We agree with the reviewer that we did not sufficiently discuss the shortcomings of HMM models. We now address this in more detail in the Discussion.
We would like to clarify that the signal amplitude threshold value we used should not be considered a pre-detection. We merely applied this threshold to reduce the data volume for the feature computation. Since the goal is to develop this method for operational application, reducing the computational time is of crucial importance. We would also like to point out that the HMM identifies events and returns a duration for the events. Event duration is thus a feature obtained by the model. Furthermore, the coherence between the sensors could also be used as a feature in the HMM. However, due to the high computational time required to calculate the coherence between 21 receiver pairs, we decided to use this feature only during post-processing. Calculating the coherence for a small number of detections is faster than calculating the coherence for the continuous data set. Finally, the reviewer is correct to state that the voting based classification cannot be included in the HMM.

Specific comments:
P3 L6-8: Are those false alarm rates related to the choice of the algorithms or to the choice of the features used to parametrized the signals? The latter might be more important and should be mentioned.

These false alarms rates are related to the choice of the algorithm.

Figure 1: I think a colorscale with more colors would allow to better observe the features of the signals generated by the different sources, especially at frequencies below 50 Hz. This is important as the readers might want to understand what guided your choice of features. Also this figure can be larger.

We chose this colour scale since it is perceptually uniform. We adapted the colour scale and enlarged the figure to more clearly highlight the features of the signals. However, the main goal of this figure was to highlight the ubiquitous environmental noise and not necessarily those of avalanche signals. The features of a typical avalanche signal are more clearly visible in Figure 7.

P8 L25-27: How do you compute the duration?

The duration of the avalanches in the reference catalogue was determined by visual inspection of the seismic data.
The onset was defined as the first appearance of energetic low frequency signals (i.e. between 15 and 25 Hz), while the end of the signal was defined as the time when low frequency signals reverted back to background levels (P5 L4-6).
The duration of any automatically detected event was determined by the HMM (P10 L9).
The minimal duration T_{min} was determined manually and is described in the results section (P14 L12 – P16 L8).

P9 L1-3: How does the voting step in the post-processing would impact the detection of “small” events (especially with a threshold set at 5 stations for a network with 7 sensors)? Are “small” events detected by the whole network? A figure showing the locations of the avalanche corridors and the seismic network would be interesting.

We have added more details on the spacing between the geophones (see P 4 L5-9). A figure of the array can be found in van Herwijnen and Schweizer (2011a) and we refer the reader specifically to their figures (see Figure 3 and 4 in van Herwijnen and Schweizer, 2011a and Figure 2 in van Herwijnen and Schweizer, 2011b).
Due to the short distance between the sensors, the voting step does not neglect small avalanches. Signals of small events are recorded at all stations. The deployment of the sensors resulting in a less desirable SNR impacts more on the detection of small avalanches. Furthermore, most of the confirmed events can be regarded as small avalanches regarding international avalanche classifications.

P11 L4-5: Indeed. How would it have impacted your results if you had chosen another master event? Is this has been investigated in Hammer et al. (2017)? If yes it should be mentioned and referenced.

We used an avalanche event from the 100% group as master event. We compared it with other avalanche events and these all had similar features (Fig 5). We therefore can expect, that the classification results are similar.
We investigated the effect of using another training event and the overall results were very similar. However, since this is a very time consuming endeavour, we did not perform a more in-depth analysis of the influence of the training event.
Hammer et al (2012) investigated the dependence of classification performance on the reference event in detail. They show that the proposed approach is very robust in face of various master events.

P13 L11-12: So the selection of the threshold on the duration is not done by considering the physics of the sources or the distribution of the durations in the reference catalog, but to optimize the POD-FAR ratio? By applying this threshold to the reference data set you lose 40% of the events (P14, last line). How is this threshold choice impacting the detection of “small” avalanches? Can you show a histogram of the duration of the events in the reference catalog? You state on P3 L24 that “For avalanche forecasting information on smaller avalanches is also required”. Hence is the approach you propose suitable to detect the smaller events?

Indeed, we optimized the duration threshold with the POD-FAR ratio. Since signal duration likely relates to avalanche size (van Herwijnen et al., 2013), it is clear that by increasing the signal duration threshold the detection rate of the smallest avalanches decreases. However, overall most of the avalanches in the reference catalogue can be considered to be ‘small’ for the purpose of avalanche forecasting. As such, we believe our approach is still suitable to detect smaller events.

P13 L14-16, P20 L6-7, Table 3: Again, it would be great to have a map of the seismic network to discuss the discrepancies observed at the different sensors. Distance to the sources, traveled paths, attenuation, dispersion, etc., can also be factors impacting the amplitude, the duration and more generally the features of the signals that might in return change the POD at different stations. This could be discussed.

Since the distance between the sensors was rather small (<12 m for the longest distance), we do not believe that local site effects substantially influence the signals. However, since the sensors were inserted in the snow, we are convinced that the main reason for the discrepancies in model performance between the different sensors is due to differences in snow cover properties, as discussed in P22 L5-10. To clarify this point in more detail, we now also mention in Section 2 that sensors 1 and 4 were most deeply covered by snow (P4 L5-8).

P20 L17-18: So in the best case what is your overall accuracy? Considering which range of avalanche sizes? I think it is this information that the readers will seek.

Overall accuracy is difficult to estimate. Although we have a reference catalogue, the actual number of avalanches in the continuous seismic data remains unknown. This is why we implemented the subjective probability classes in the reference data set. While such an approach does not provide a single performance value for the HMM, it shows that for clear avalanche signals (100% class) model performance is reasonable. Furthermore, it highlights the difficulties in obtaining a reliable and independent avalanche catalogue. However, in response to reviewer 2, we now also added information on the detection rate of the avalanches that were confirmed by images from the automatic cameras. These results show that at least 80% of the confirmed events were detected by all sensors.

P20 28-30: Are the results presented in this study supporting this statement or is it based solely on the study by Hammer et al. (2017)?

Our results support this statement since we also only used one training event for the HMM.

P21 L4-5: How can you incorporate localization parameters in the HMM? Is this done directly in the model or during the post-processing?

Both methods are possible. Localization metrics, in particular back azimuth and apparent velocity, of signals above the amplitude threshold could be used as feature for the HMM. On the other hand, it would also be possible to use the localization metrics during post processing by rejecting all detections having clear path.