Repsonse to reviewer #2

General comments:

The paper discusses the performance of an automatic seismic detection system of snow avalanches using hidden Markov models (HMMs). The study is based on a 107- day continuous dataset acquired during the winter 2010 by a small seismic array consisting of 7 vertical geophones deployed above Davos, Switzerland, and surrounded by several avalanches starting zones. The HMMs system is tested against a reference avalanche catalogue that is based on the work by van Herwijnen and Schweizer (2011a). The reference avalanche catalogue gathers 283 events distributed into 7 probability classes following an independent re-evaluation of the avalanche signals of van Herwijnen and Schweizer (2011a) by three of the authors. Among the 283 events, only 20 (7%) are finally classified by the aforementioned three authors as being with 100% probability a snow avalanche. For the HMMs detector, the authors apply the approach developed by Hammer et al. (2012) that requires a background noise model and only one training event to learn the system. Because important diurnal and seasonal variations are observed in the seismic feature of the background noise over the season, a background noise model is recalculated for each day. Since the authors observe that the change from dry to wet avalanches through the winter does not have a strong impact on the selected seismic features used for the HMMs, only one avalanche signal is selected as a training event for a single avalanche class HMMs detector. The first HMMs avalanche detector results in high probability of detection (POD 70-95%) in the 100% probability class of the reference avalanche catalogue for all 7 stations. However, a large number (up to 2091) of additional events (unassigned detections) are detected by the HMMs for each sensors. In order to reduce the number of unassigned detections, the authors decide to introduce post-processing steps to the HMMs: (i) a duration based classification and (ii) an array based (multiple-station detection/coherence of signal among station pairs) classification. The postprocessing steps reduce the number of unassigned detections by the HMMs; however, the POD in the highest probability class of the reference avalanche catalogue is also reduced. A manual re-evaluation of the unassigned detections after post-processing shows that part of these detections correspond to avalanche missed during the manual detection procedure.

The paper is well structured and clearly written. Despite the noisy character of the data, the authors show HMMs as developed by Hammer et al. (2012) to be a promising approach for operational avalanche forecasting; however, with fine tuning. Strategies will be needed for a regular update of the background noise model and post-processing must be implemented. Much attention should be paid to the stations deployment. The author recommend to install the sensors 30-50 cm in the ground below a homogenous snow cover in order to reduce environmental noise and increase coherence of the data among the network and thus, stabilize the array based post-processing. In addition, a larger sensor inter-spacing would allow the system to apply array-processing techniques and enable to incorporate localization
parameters of avalanche source area in the automatic system. In my opinion, the implementation of the post-processing step is of great interest, as it helps to deal with the noise contamination that is found at any surface site data. The main concern I have is that the authors didn’t make use of the 33 confirmed avalanches of van Herwijnen and Schweizer (2011a) and used instead a re-processed

probabilistic reference catalogue where only 20 avalanches remain considered as certain. In the following, a number of specific comments are related to this concern. I leave it to the authors to decide to which point they want to integrate these suggestions. It would be interesting to know how the 33 confirmed events match the post-processing steps: do they have durations above 12 seconds, are they detected by 5 sensors or more, what are the inter-station correlation indices for these events. In addition, if the location of the 33 confirmed events is known, how important are the variations in the spectral feature of the signals as a function of station-source distance? Finally, what is the performance of the HMMs system including the three post-processing steps against this original 385 events/33 confirmed events reference catalogue?

We agree with the reviewer that adding the model performance for the confirmed avalanches would be insightful. We therefore updated figures (Figures 3 and 9) and tables (Table 1,2 and 3) to include these results and discuss them throughout the paper (e.g. P10 L10-11). Note that the number of confirmed avalanches originally mentioned in the manuscript was wrong and should have been 25. We believe that these results better highlight the model performance and we thank the reviewer for this valuable suggestion.

Specific comments:

P1 L9-12. Please rephrase and clarify this part of the abstract (according to remarks about P4-7 Section 3).

We rewrote this sentence (P1 L10).

P4 L5-7. Although the station inter-spacing is specified in van Herwijnen and Schweizer (2011b), you should mention it here, so the reader has an idea of the size of the network.

We now added this information in the text (P4 L6).

P4-7. I think the section 3 needs some clarification so the reader can get a better idea of the reference data:

1) What are the exact data that were analyzed visually by van Herwijnen and Schweizer (2011a) (385 detection, 33 confirmed events): (1) 107-day of continuous data without pre-processing of section 3.1 at sensor number X or (2) XX% of pre-processed data (section 3.1) at sensor number X? In case (1), I would place P4 L28-30 before section 3.1 for clarity.

We now clarify these points (P5 L1) but prefer to keep the sentence where it is.

2) Section 3.1: The pre-processing of the data is obscure to me. Was it apply only to reduce the amount of data to (re-)process manually? Did van Herwijnen and Schweizer (2011a)/Heck et al. AND the HMMs perform only on the 20% of high

amplitude data (which would impact on the paper title as it would be no more continuous data)? If a local network is intentionally deployed to detect small avalanches, why place an energy threshold in the investigated data series? See for example avalanche number 5 (Av 5) in Figure 7 of van Herwijnen and Schweizer (2011b) which has poor SNR. Furthermore, HMMs are not dependent on energy thresholds, please clarify.

The manual detection of avalanches was performed on the complete data set. However, calculating the features for all seven sensors and the entire data set is extremely time consuming. We therefore applied an amplitude threshold to reduce the amount of data by only taking those parts of the time series when some energy is arriving at the sensors. The amplitude threshold we used was very conservative and all the confirmed avalanches were still in the pre-processed data (Table 1 or Figure 3). Since the goal is to develop this method for operational application, reducing the computational cost is of crucial importance. As such, we do not believe that the pre-processing step we applied warrants a change of the title.

3) Section 3.2: 20 merged avalanches represent 7% of the 283 reference events. In these data, where are the 33 confirmed avalanches of P4 L30? You could have used 33 confirmed events (9%) against 352 uncertain events of the original 385 events of van Herwijnen and Schweizer?

We now include these confirmed events in Figure 3 and Table 1,2 and 3. Most of these events belong to the 100% or 83% class of the re-evaluated reference data set. For sensor 1, 20 of these events were detected and 5 missed.
For the voting based detection without using a minimal duration, also 20 were detected and 5 missed. By applying a minimal duration, the number of detections reduced to 18 hits and 7 missed.

4) P6 L11-15: I think it would be interesting to discuss here in a few words the 283 events reference avalanche catalogue in more details. For example, do you find trends in the probability classes, are the 20 certain events confirmed detections, long duration signals, high energy events detected by all stations? Are the lower confidence event simply lower quality events (low SNR, only recorded at a few stations and not consistent among the stations)? These low probability events represent in my opinion the real challenge of environmental seismology case studies. And I expect the subjectivity of the analyst to have more influence on low quality data (how would/did you rate avalanche Av 5 in Figure 7 of van Herwijnen and Schweizer (2011b)) than on high SNR data? This is where automatic, quantitative systems should help :)

We clarified these points by adding model performance statistics for the confirmed avalanche events throughout the text (Figures 9 and Tables 2 and 3). Furthermore, we added a figure (Figure 4 in the new manuscript) which shows the distribution of signal duration for each probability class. This figure clearly shows that the number of short duration events decreases with increasing probability class.

5) P7 Figure 3. How does the original catalogue with 385 events, 33 confirmed avalanches plot? Are the two periods of higher activity well represented? I personally would insert a plot of the 33 confirmed events at the bottom of Figure 3.

We now show the confirmed avalanches in Figure 4. Indeed, most of these avalanches also released during the high activity periods in March and April.

P10 L8-20. Does this very sophisticated approach to find a threshold value bring more than an approach using common sense (for which arguments are well described in P8 L24-29 – P9 L1-20)? The minimum duration can be evaluated (or decided) as a function of the expected distance/size of the target events; the number of stations that must detect the signal can be selected as a reasonable value knowing the network/stations specific performances; the inter-station correlation threshold can be evaluated by investigating the 33 confirmed events. Please comment.

We used a systematic approach to find threshold values by optimizing model performance. Since we had a reference avalanche catalogue this was possible. However, in the absence of such a reference data set, for instance when installing a system at a new location, we agree that to some extent it should be possible to select these thresholds based on some a priori available knowledge on the local topography and the array geometry. However, this requires some assumptions on the minimum avalanche size that can be detected as well as sensor performance.

P11 L11-13. I agree that in Figure 5a the feature distribution is very similar for the 4 selected events. However, in Figure 5b, I find events of 21 Jan/21Feb dissimilar to events of 22Mar/24Apr, especially at the events onset where one group goes up while the other one goes down. Please comment on that and see remark on P15 Figure 8.

Due to the different length of the signals, we decided to use a normalized time. As the events in March and April were quite long, the decrease for cepstral coefficient is longer. In January and March the decrease is shorter, thats why it seems as it only increases. This is also visible for the central frequency, where the frequency decreases slower for the March and April events.
Furthermore, for the HMM all 4 events are similar, since the feature behaviour (the decrease) is important. The absolute appearance of the features does not matter. This explains why the detection of events with different durations is possible.

P12 Figure 5. In (a) and (b) vertical dotted lines at normalized time 0 and 1 would help to visualize start and end time of event.

Figure 5 changed as suggested.

P15 Figure 8. For both sensor 1 and 7, the HMMs missed 4 events at the end of January:
1) I don’t find this 4-event spike in Figure 3?
2) Are any of the 33 confirmed avalanches found at this date? I see there is one 100% probability class event around this date in Figure 3.
3) Could this speak for an influence of the HMMs system by the avalanche type (dry-
wet) selected to train the system?

The small spike in avalanche activity at the end of January was also present in Figure 3, marked as 33% avalanches. However, the spike is more obvious in Figure 8 since there we also applied the signal duration threshold to the reference data set. None of the confirmed avalanche events occurred on this day, as can now be seen in Figure 3 where we added the confirmed events.

P18 L1-6. Clarification for the single sensor results:
1) Did you investigate the signals of the unassigned detections individually at station 1 and 6? Or, 2) Did you look at the signals recorded by all stations at the time of the unassigned detections of sensor 1 and 6?
I think these results should be commented. What could explain the higher false alarm rate of the array based approach against the single sensor approach?

We investigated the results for each sensor individually. We clarified the sentence (P20 L1-2).
The higher false alarm rate is due to the higher signal-to-noise ratio for the other 5 sensors, as stated in the Discussion section (P22 L5-10).

P19 L8-10. Related to comment on section 3. Please clarify. At P4 L28-30, the 33 events are part of the 385 van Herwijnen and Schweizer (2011a) events. (sorry to insist ;))

The 25 confirmed events were part of the 385 original events and also remained in the 283 events after applying the amplitude threshold during pre-processing.