Reply to reviewer 1

We thank the reviewer for the constructive comments. Below we reply in detail.

P2, 5: what does "rather poor" mean? Can you quantify it or specify? Is this their conclusion or your interpretation?
Bessason et al. obtained a propability of detection around 65%. Seismic signals generated by rockfalls and debris flows were wrongly classified as avalanches and vice versa. Information about the false alarm rate was not given directly, but seems to have been high. (P.2 L.3)

Bessason et al. 2007 used a nearest neighbor approach to classify newly recorded events. Using this approach, they were able to detect 65% of all confirmed avalanches.

P2, 30: I think it is unclear what these arrays are. e.g. the one to locate avalanches and the one at 14 km distance and then your are talking about one in Dischma Valley and one at Wannengrat array. Are these the same arrays or different ones? Maybe the names or location of arrays should be introduced earlier and a link to figure 1 should be added?
Thanks for pointing this out. We rephrased this sentence and added a reference to Figure 1. (P.2 L.34)

We performed the classification and localization of the events with the data recorded at the seismic array located in the Dischma Valley above Davos, Switzerland during the winter season 2017 (yellow square in Figure 1). These results were then combined with data obtained at the Wannengrat array, which is located 14 km to the northwest of the Dischma field site (red square in Figure 1).

P2, 32: is this the winter season 2016/17?
Yes, we referred to the winter season 2016-2017. We changed the number in the text.

P3, 10: two ")" too much
Removed.

P3, 15: ")" too much
Removed.

P3, 15: where were these cameras and weather stations located?
We included a detailed view of both field sites including all different types of weather stations and automatic cameras.

P4, 2: In this sentence you describe that the cameras helped to identify avalanches in the winter of 2016/17. But then you cite a publication from 2011? Clearly this publication does not describe the winter 2016/17? Maybe rephrase.
We rephrased this sentence to clarify that van Herwijnen et al. performed a similar analysis in 2011 in the surroundings of Davos. (P.4 L.5)

As already shown for avalanche activity periods in the winter season 2009-2010 by van Herwijnen et al. 2011b, those images can help to identify and confirm seismic events produced by avalanches.

P4, 6: Is the amplitude in noise that stable in time, that it ok to use a fixed threshold like you do or did you change it in time?

As mentioned in the text, we use a threshold value that is 5 times higher than the daily mean amplitude, i.e. the noise amplitude threshold changes each day.

P4, 6: given a sampling rate of 500 Hz your time window is only 2 seconds long when selected. This sounds pretty short to me when looking for avalanches.

In this step we are looking at the seismic energy recorded at the sensors for a 2 second window. If the energy for this window is higher than the threshold, we look at the following windows. If the energy for these windows also reaches the threshold value, we concatenate the windows. As soon as the energy decreased below the threshold value, we cut an additional 60 s before and after the total length of the event to not dismiss the onset and the coda of the event.

We clarified this in the text. (P.4 L.13)

For a window i with a length of 1024 samples a mean absolute amplitude A_i *was determined. When* $A_i <= 5 A^*$, *with* A^* *the daily mean amplitude, the data within the window were cut. If the amplitude threshold for the following window was also reached, data were concatenated. Furthermore, a section of t=60 s was cut before and after the window to ensure that the onset and coda of each event was incorporated. Doing so, data were reduced by 80% to several data windows of various lengths.*

P5, 4 "Using these properties, a widespread background model can be learned from the general properties" I think this sentence sounds odd. Are you trying to built a model from information you derive from the general properties?

We rephrased the sentence. (P. 5 L.14)

From these properties a widespread background model can be learned.

P5, 4-7: I cannot follow how your method works in detail. Maybe the text should be rewritten with more reference to figure 3?

We rephrased the section of the brief introduction of the Hidden Markov Models. However, this topic is rather complex and we provided references. (P.5 L.4)

To automatically identify avalanches in the continuous seismic data we used hidden Markov models (HMMs). These statistical classifiers use a sequence of multivariate Gaussian probability distributions to model observations (e.g. seismic time series). To determine the characteristics of the distributions (i.e. mean and covariance) a large number of training sets of known events, so called pre-labeled training sets, are required. For each different type of observation (e.g. avalanche, airplane or earthquake in the seismic data) a separate HMM is trained. By combining all HMMs the whole classification system with several classes is constructed. This classical approach, which relies on a large number of well-known pre-labeled training sets, was successfully used to automatic identify seismic events in continuous seismic data (Orhnberger 2001, Beyreuther et al. 2012).

Avalanches, however, are rare events and it is nearly impossible and too time consuming to obtain a large training set.
To circumvent this, we performed the classification based on an approach developed by Hammer et al. (2012) exploiting the abundance of data containing mainly background signals to obtain general wave-field properties. From these properties a widespread background model can be learned. A new event model (e.g. representing avalanches) is obtained by using the background model to adjust the event model description by using only one training event. In contrast to the classical HMM approach, the classification system of this approach consists of a background model and one event model for each implemented event [or observation]? class.
The classification process itself calculates the likelihood that an unknown data stream was generated by a specific event [or observation]? class for each individual HMM class (Hammer et al. 2012, Hammer et al. 2013).

P5, 14: are assuming that two events are separated by at least 24 hours? And if two events have a closer spacing in time they cannot be picked/ located?
We do not assume that two events are separated by at least 24 hours. We just use the data within a 24-hour window to construct our background model and a one hour window is then classified. Once the classification is performed, we shift both windows one hour.
We added additional information to clarify this. (P.6 L.8)

This means, that our background model is always determined by the pre-processed data of a 24-hour window. By choosing a length of 1 h for the window t_{class}, we were able to classify the pre-processed continuous seismic data of one hour during one step of the operational classification. Once one classification step, which is the classification of the window t_{class} is finished, both windows are shifted by one hour and the classification was executed for the shifted windows.

P5, 15: you state that the t_class window is 1 h long, but on p4, 6-8 you state that the chosen event is only 122 s long. Is there an error somewhere?
t_class is the length of the window we want to classify, which means that we want to identify all events within this window and define their origin using the HMM. The length of the identified events is independent of t_class.

P5, 25: What is an instantaneous frequency?
The instantaneous frequency is the time dependent change of phase.
Taner, M., F. Koehler, and R. Sheriff (1979). Complex seismic trace analysis, Geophysics 44, no. 6, 1041–1063. We now include this reference in the text.

P5, 27: maybe you should explain cepstral coefficients?
We added a reference for the cepstral coefficients. (P.7 L.7)

P5, 30: what do you mean with “the first half-octave band has [. . .] a total number of 6 bands”?
In total, we calculated 6 half-octave bands. The first half-octave band had a central frequency of 3.9 Hz. Depending on this frequency, the central frequency for the remaining half-octave bands are already determined. We rephrased this sentence. (P.7 L.14)

We used in total 6 half-octave bands for the classification and the first half-octave band had a central frequency of 3.9 Hz.

P6, 6: what if the event model is very unlike the avalanche signal you try to detect?
Then we would not be able to identify avalanches in the winter season. Therefore, it is crucial to use a training event which is representative of avalanches at this specific field site. We changed the wording in the text to emphasize the importance of the event model and included a reference to Heck et al. (2018). (P.7 L.24)

The event model HMM_{Event} was learned using only one training event that is representative of avalanches at a specific field site (Heck et al. 2018a). It was determined once and then applied for the entire winter season.

P6, 10: “Each classified event having a duration shorter than 12 s was dismissed” replace with “Each classified event shorter than 12 s in duration was dismissed”
We rephrased this sentence.

p7, 6: I am a bit surprised that you state that your second array at 14 km distance does not record the avalanche any more. After all you mentioned this array in the introduction that could detect avalanches up to a distance of 30 km (p2, 10).
The detection radius strongly depends on the size of the avalanche, and to a lesser degree on the sensors used to record the signals. It is possible to identify avalanche within a range of 30 km, however, only by using broadband seismometer and only for very large events (runout distances > 2000 m). These events are very uncommon and were not observed during our study period. The avalanches we monitored were substantially smaller avalanches and we used less sensitive geophones. For these avalanches, we recently showed that the detection radius is about 3 km (Heck et al., 2018b).

p7, 8: “12 km away” replace with “at 12 km distance”
We rephrased this sentence.

p7, 10: rephrase the heading as I find it pretty unspecific
We rephrased the heading. (P.9 L.1)

Localization results to confirm avalanches

p7, 12: what MUSIC code did you use? Where is it available?
The MUSIC code we used was developed by Dr. Manuel Hobiger from the Swiss Seismological Service. We will supply the scripts and data on an open access storage.

p7, 27: does this approach not exclude avalanches along other potentially longer or more curved paths?
We analyzed one event which approached the array to less than 100 m and therefore exhibited large changes in the back-azimuth. These changes were still well within the limits we used here. For more distant avalanches, changes in back-azimuth will be much smaller, even for more curved paths.

p7, 33: what is the “used array”?
The used array is also the Dischma array. We rephrased this sentence. (P.11 L.4)
… performed at the Dischma array.

p7, 34: “through further analysis” instead of “by further analysis”?
Changed as suggested.

p8, 2: “to speed up the calculation time”: you “reduce the calculation time” or “speed up the calculation”
We rephrased the sentence.

p8, 2: so if I understand this correctly for a 2 minute long window it takes 6 minutes to process? So in order to do this in real time you need to skip time windows e.g. of “noise”
That is correct. This is also one reason to pre-process the data.

p8, 15: figure 4a p8, 16: figure 4b p9, figure 4: maybe remove the legend in figure 4b as the information is already there as label of the y axis. Could you limit the yaxis at 110 or so in order to make the low numbers of avalanches in February more visible?
We changed the figures as suggested.

p9, 2: On p7, 30 you state that you minimum event length is 20s whereas here you state it is 12 s.
These are two different post-processing steps. In the first step, we applied a duration threshold to the events obtained from the HMM classification, as suggested by Heck et al. (2018). In the second part, however, we applied a duration threshold to the median back-azimuth path obtained by MUSIC. This median path was calculated using a sliding window of 8 seconds. To obtain reliable results, it was thus necessary to use a second duration threshold value. We clarified this better in the text. (P.9 L.20)

Heck et al. 2018a suggested that a detected event should have a minimum duration of 12 s to be considered as an avalanche. For the localization step, however, it was necessary to increase this duration because the window length used for the median smoothing filter was already 8 s long (Heck et al. 2018b).
To cover enough data points to use the minimal event duration as a reliable classification criterion, we therefore required a minimum length of 20 s for the back-azimuth path.

p9, 6: What do you mean with “classes with 5 and 6 votes” what votes?
This term refers to the events that are detected by 5 sensors or 6 sensors of the array. We rephrased this sentence and removed the terms votes and classes~~,~~ to avoid confusion. (P.11 L.15)

A quarter of the events were detected by 5 sensors, a quarter by 6 and about half by 7.

p9, 10: It that a good thing or a bad thing that you detect avalanches that are not listed in figure 4? E.g. does this mean that there are avalanches missing in figure 4 that should have been listed or are there completely different avalanches recorded in different areas and the only common thing is the huge amount of snow in that time period?

In principle, it is entirely possible that we detect avalanches with our monitoring system that were not recorded by visual observations in the area of Davos. However, at this point in the analysis, it remains unclear if the events automatically identified with the HMM correspond to avalanches or not. We therefore manually went through all the 117 detected events to evaluate if the signal characteristics correspond to those typically seen for avalanches.

p10, figure 5: move the sentence “the red area...” up to the description of subfigure a
We changed the caption. (P.11)
Avalanche released on 9 March 2017 at 06:47 used as training event for the classifier. a) time series for the 7 sensors. The red area indicates the part of the time series used as training event. b) corresponding spectrogram of the seismic time series.

p10, figure 6: what do you mean with vote in the legend? What is a vote in the context of avalanches?
As already mentioned earlier, this refers the number of sensors the signal was detected at. We changed this term, since it is confusing here and throughout the text and figures.

p11, 2: two “)” too much
We changed it.

p11, 21: I keep wondering why you detect the avalanched only up to 4 km distance and not 30 km distance as mentioned in the introduction.
This is due to the instrumentation used and the size of the avalanches that were monitored. We added some references in the introduction considering this small range of detection. (P.2 L.26)

In the both studies of Lacroix et al. (2012) and Heck et al. (2018) less sensitive vertical component geophones were used for the seismic monitoring resulting in an avalanche detection of approximately 3 km.

p11, 31: “except for detections at the beginning of April” consider replacing it with: “except for detections at the beginning of April and no detections at the beginning of February” Any ideas as of why these avalanches in February could not be detected?
During the winter 2016 – 2017 we had slightly different meteorological conditions at the Dischma field site compared to the surroundings of Davos. In particular, there was less snow in the back of the Dischma valley resulting in fewer avalanches early in the season. Based on field observations and the images from the automatic cameras, we determined that in early February there were no medium-sized to large avalanches in the vicinity of our array.

P12, figure 7: Does the gray area in figure 7 have the same meaning as the red area in figure 5? If yes I suggest to use the same color.
The red area in Figure 5 is used as the training event, so we manually defined this area. The gray area in Figure 7, however, shows the part of the time series, which was identified as an event by the hidden Markov model. Therefore these two areas have a different meaning.

P12, figure 7: I am surprised about the low frequency content of the airplane and the lack of overtones. How did you classify this as an airplane?
We agree with the reviewer that typically signals generated by airplanes either have clear overtones or at least a clear Doppler effect in the signal. However, based on our experience with seismic data from both arrays above Davos, we are confident that the signals shown in Figure 7 are generated by airplanes. We do not know yet why airplanes generate such signals, and we have not identified a clear pattern, which explains their presence. Perhaps it is related to the altitude or distance of the airplane, or perhaps specific atmospheric conditions are

responsible. Nevertheless, we have seen multiple signals like these recorded at both arrays and we are confident that these signals are generated by airplanes.

P13, figure 8: What is the unit of the normalised time and how is it calculated? Do the events have the same length or did you just stretch/ squeeze them to fit in between 0 and 1?
In Figure 8, time is normalized by the duration of each event. The normalized time therefore has no -units. 0 corresponds to the start of the event and 1 to the end of the event. The duration of the events can therefore not be determined from this plot as each event has a different duration. We chose this representation to better highlight similarities in the temporal behavior of the features for different events.

p13, 6: one ")" too much
We changed it.

p13, 8: one ")" too much
We changed it.

p13, 11: Do you know what these 37 other avalanche like events might be? Maybe these are just avalanches along an unexpected path or longer paths?
Based on a visual inspection of the seismic time series and spectrograms we assumed that we have several sources for these signals. For one part, some airplane signals were still present in the classification results after the combined array classification and signals produced during periods of strong winds.

p13, 16: It sounds to me a bit like you remove events until you end up with back azimuths or locations you would like to get.
We did not filter specific back-azimuth angles. We only require the derivative of the median back-azimuth path to fall within certain threshold values (i.e. the source of the signal should not vary too much within a certain period of time), independent of the values of the back-azimuth angle. The fact that the remaining events all had a mean back-azimuth pointing towards nearby avalanche slopes therefore confirms that this method is suited to identify signals likely generated by avalanches.

p14, figure 9: How do you know that these are airplanes?
During the last years we analyzed the continuous seismic data recorded during the last winter periods. These types of signals were often recorded almost simultaneously at both field sites at any time of the winter season, even in the beginning, when there was no snow at the field sites. We also compared the times of some signals with flight information and were able to identify the airplanes. See also comment above.

p15, discussion: I find the discussion a bit repetitive with respect to the rest of the manuscript. Many points seem to have been made already in the rest of the text. Also my impression is that they barely refer to work of others in the discussion i.e. papers that are not lead by "Heck" or "Hammer".
We agree with the reviewer that some parts of the discussion were redundant. We therefore partly rewrote the discussion to avoid this. However, since we are not aware of many studies which tackle the automatic detection of mass movements (avalanches or other) in continuous seismic data, there are not many other studies which are relevant. (P.20 L.5)

Apart from HMMs, several other machine learning techniques are suited to classify signals in seismic data. It is possible to use a convolutional neural network for earthquake detection and location (Perol et al., 2018) or to pick the P-wave arrival of seismic wave fields (Ross et al., 2018). Comparable to the classical HMM approach, these studies rely on large pre-labelled training data sets. Another approach is the so-called Random Forest classifier, which can be used to discriminate seismic waves (Li et al., 2018). Automatic classification approaches are also suitable to differentiate between earthquakes and quarry blasts (Hammer et al., 2013) or to characterize larger rockfalls (Dammeier et al., 2016). Further mass movements, such as landslides, can also be identified in the seismic data based on automatic classification approaches (Esposito et al., 2006; Hibert et al., 2014; Provost et al., 2016). The automatic classification of avalanches yet remains a difficult task. Rubin et al. (2012) used several machine learning algorithms to identify avalanches in seismic data and compared the results obtained with the different approaches. With all methods a high probability of detection was achieved, but the number of false alarms was too high. A recent study by Heck et al. (2018a) showed that HMMs are a suitable tool to detect avalanches, but there is still a need for additional post-processing steps. In the work presented here we confirm that HMMs in combination with further post-processing steps provide reliable classification results.

p15, figure 10: change to that the legend is not overlapping the bar any more
We changed the legend.

p16, figure 11: “for avalanche event” replace with “for an avalanche event”
We rephrased it.

Figure 11a: I don’t understand to what part of the figure you refer to with “solid part”. Beneath what threshold?
The solid part refers to the solid line in the polar plot.

Figure 11b: is this really the derivative of the angle (y axis label) or derivative of the back-azimuth path (caption)? To me this figure seems to show the “angle” or “back azimuth” during, before and after the avalanche event with very stable back azimuths during the event and larger scatter afterwards.
This plot shows the approximate derivative of the angle (due to the small amount of data points). It was calculated using the angdiff function of matlab and the step size dt between the data points angdiff/dt. We changed the label to clarify this. (P.17)

p17, figure 12: so there are 100 visually observed avalanches in Davos but you could detect only 20? Were you too far away or was this recorded but not classified as event? Move the legend so that it does not overlap with the bars
Indeed the differences in the number of avalanches relate to the size of the area that is monitored. The visual observations are made for an area of about 175 km^2, while with the seismic system only avalanches within a radius of 3 km can be monitored (~30 km^2). We changed the legend of the figure as suggested. (P.18)

p17, 1: “closer” replace with “closer to”?
We changed it.

p17, 8-10: First you say that you could confirm no avalanche visually, but in the next sentence you state that "another 12" events were identified. Were they identified in a different way i.e. not visually or is there an error in the sentence?
The events were not visually identified, but by inspecting the seismic time series and analyzing the events in more detail. Based on the results, these 12 events were identified as avalanches.
We clarified this point in the manuscript. (P.20 L.3)

However, Heck et al. 2018B manually identified 13 avalanches during 9 and 10 March 2017, 12 of which were automatically identified with the approach presented here.

P18, figure 13: How do you know to what distance the duration of the event corresponds to?
The duration and back-azimuth of the event are automatically determined with our method. We do not have any information on the distance of the source. Perhaps the reviewer misunderstood the results shown in Figure 13 and we therefore changed this figure. In the polar plot the direction of the lines indicates the direction of the back-azimuth, the thickness of the lines indicates the duration of the event. Thin lines correspond to short events, thick lines to long events. The thicker the line, the longer the event. The color code shows the release time of the avalanche. (P.19)

P18, 4: number of votes: in my opinion it would be better to replace "vote" with something like "detections on sensors" or similar.
As already mentioned above, we changed this term, since it is confusing.

p18, 12: the overall feature behavior from distance airplanes... "was" not "were"
We changed it.

p19, 9: remove "really". Based on the 5 events that were possible to locate, it is apparently possible to detect some avalanches on both arrays.
We changed it.

p19, 9: I am not sure I fully agree. It is not possible to record an avalanche at 14 km distance if it couples to the ground sufficiently or is large enough?
It might be possible to record avalanches at distances > 14 km. However, as previously mentioned, this is only for large and catastrophic avalanches which were not observed during our study period.

P19, 10: "since distance" replace with "since the distance"
We changed it.

p19, 10: I am not sure where installing two arrays at 2-3 km distance would help. They would then pick up the same avalanches, and hence "events recorded at both arrays" are then not a valid criteria any more to find falsely classified earthquakes or airplanes.
We agree. Two arrays will mainly improve the localization.

p19, 11: "improving" replace with "improve"
We changed it.

p19, 22-24: Can you not locate airplanes and earthquakes with the array because the frequency content is different? So if the MUSIC method is perfectly suitable of detecting

avalanches, why should one go through the hassle of finding a exemplary event, the need of having two arrays and then removing a lot of false detections? Rather than using the output from the array method to detect events?
We performed several test during this study considering the calculation time of the MUSIC method. Depending on the frequency range of the signals, the duration of the calculation varied. However, based on the actual code, it was not possible to apply the MUSIC method in near real-time. If we skipped the classification process at all, we would have to analyze all data remaining after the pre-processing step. It would take too much time to calculate the MUSIC results for all these events, even if we reduced the frequency range to a minimum. The classification process, however, has proven to be faster and allowing us to perform a near real-time classification.

p19, 26: typo in “theses”
We changes it.

p19, 30: typo “form”
We changed it.

p19, 32: “avalanches were released” instead of “avalanches released”?
We changed it.

p20, 5: Why is it that costly? Can the processing be sped up?
At the moment the algorithm is written in MATLAB. In this algorithm, the covariance matrix of small time windows is calculated. It might be possible to speed up the calculation process by using computers with more RAM, since matrix calculations depend on the size of RAM. Unfortunately, we are not computer engineers and our knowledge is too limited to answer the question.

p20, 14: “be still needed” replace with “still be needed”
We changed it.

p21, references. There are 11! referrals in the text to a not published paper (Heck et al. 2018b). Can the authors provide the manuscript in order to cross-check e.g. the content?
The paper is now accepted and accessible online: https://doi.org/10.1093/gji/ggy394