Referee Comment #1

In this study, the authors take advantage of a temporary, dense and relatively large seismic array (Swath-D Network) to build a new earthquake catalog and improve the current knowledge of the seismicity in the eastern and southern Alps. The authors present the results of a thorough analysis mixing automated and manual methods with which they detected and located 4 times more events than existing catalogs. The manuscript is well organized and written. I enjoyed reading about sophisticated methodology that doesn't involve machine learning. I only have minor comments about the methodology and some surprising results.

Comments

- Did the INGV and SED agencies use the same data set when building their catalogs? I find quite surprising that after the energy-based detection there are only 286 new events.

  Yes, the seismic networks that are operated by the local agencies are publicly available. The agencies therefore have access to the same data and also state on their websites that they use data from external networks. Combining the local networks, the agencies have quite impressive detection capabilities in the area. This explains the relatively low number of additional energy-based detections. Note that this is only one of the first steps in our workflow, and the majority of events is detected by template matching.

- Lines 104 and 296, you mention the correlation coefficient but don't give details about how many stations are involved in its computation. I think this information is important to understand what the value means.

  The cross-correlation coefficient referred (line 104, 296) refers to the median of the three largest three values, using a maximum of 15 stations closest to the event, depending on availability. This has been clarified in the manuscript at the locations mentioned above, as well as in the figure captions of Figures A5 and A7.

- Line 140 "Additionally, the number of I/O operations is reduced to a minimum by loading each continuous data trace only once." I don't really understand the meaning of this statement because I don't see why one would load the data several times?

  Because our machine, especially the GPU, has limited memory, only a small number of templates and a small amount of continuous data can be loaded at once. In our code, the set of continuous data traces are held in memory while the templates are cycled (reloaded for each new set of continuous data traces). This is cheaper than cycling through the continuous data, but has some limitations. For example, if one were to use a summed CC approach, it would be necessary to load the data template-wise (i.e. load traces from different stations simultaneously), or to write the continuous cross-correlation functions to disk (which would take up a huge amount of space). Both of these options would require much more reading and writing.

- The detection threshold is set in a very arbitrary way: How much is 0.5 in terms of standard deviations? CC detection thresholds are usually defined upon the root mean square or median absolute deviation of the CCs so that they correspond to a given p-value of false detections assuming a gaussian distribution. Based on my personal experience, 0.5 can also be very conservative. What is the motivation for not summing the CCs and applying the threshold on the summed CCs? Array techniques are all based on the idea that signal-to-noise ratio of a sum increases as N where N is the number of traces. But by applying the detection threshold on single stations, one loses the full benefits of network information.

  We have experimented with using Median Absolute Deviation (MAD), as well as a summed CC approach. Both of these options reduced the stability of our results. In general, our CC threshold of 0.5 corresponds to an MAD of about 7-9. However, the CC obviously has a maximum of 1.0 when the waveforms are exactly the same, whereas the maximum of the MAD varies greatly over time and from station to station, making it difficult to set a universal threshold. Our first attempts at applying template matching to our dataset revealed that an enormous amount of triggers were due to station noise. The best way to stabilise the results was to use three independent station measurements instead of a combined network value. This greatly reduced the number of false triggers. It also has the advantage of being much faster, because we don't need to store all the continuous cross-correlation functions (see answer to previous question).

- Lines 166-167: For phase picking, the authors use higher frequencies for S waves than for P waves. What motivated this decision? It is unexpected since P waves carry higher frequency energy (the P-wave corner frequency is about 50% higher than that of the S wave).

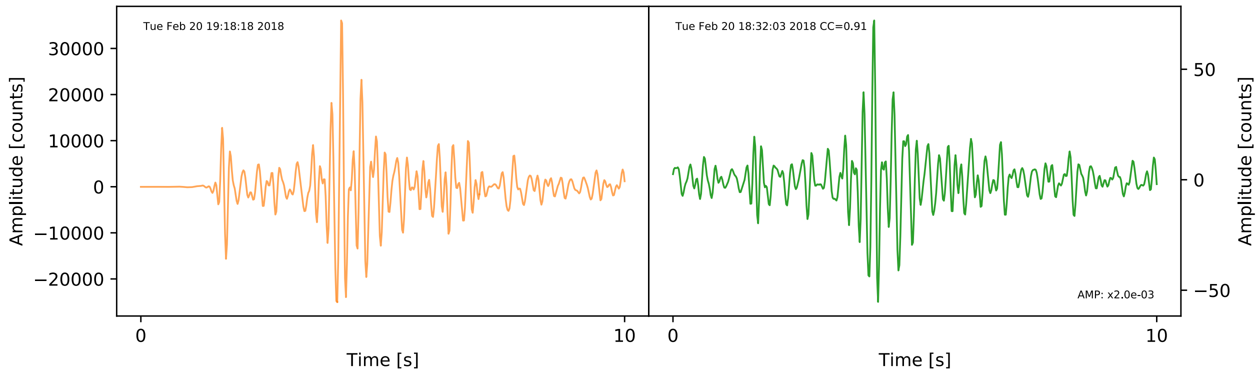
  This was a typing error. We use the frequency band from 1-20 Hz, whereas for S-waves, frequencies between 1-12 Hz are used. Thanks for pointing this out.

- Lines 178: I got a bit lost in these explanations. We start reading about picking on the template events, then on the newly detected events, and in this paragraph the authors go back to hand picks, which I thought were only for template events? I found this paragraph hard to understand so perhaps it is worth editing it.

  Thanks for this suggestion. The paragraph was rewritten to make it clearer, and the previous two paragraphs were also revised slightly to improve the structure of the section. As for your question: only the master events are hand-picked. This is the first step. In the second step, these handpicks are used to find picks for the detected events by cross-correlation. However, changing station availability might therefore limit the number of picks for detected events. We therefore apply a third step, where a STA/LTA trigger is applied to traces that might be available for the detected event only (and not for the master event).

- Magnitude estimation. Figure A5 shows some surprising observations. The lack of knowledge of what the CC means exactly (see comment above) might partly feed this comment. For event pairs with CC > 0.9, even within the reduced frequency band of 2-8 Hz, I don't see how the magnitude difference can be up to 2 units. Could it be caused by errors in the magnitude computation?

  An explanation of the definition of the CC we used has been included here and in the figure caption (Fig. A5). We do not attribute the variation in magnitude to errors in the calculation. The frequency band used for cross-correlation (2-8 Hz) is well below the corner frequency of the events in this magnitude range (~ 20 Hz for M_W 2). The cross-correlations are therefore performed on the flat part of the frequency spectrum. On the recommendation of the second referee, we have changed the method for magnitude calculation to a standard local magnitude approach. This actually increases the magnitude difference slightly, because the magnitude now scales directly to the log amplitude. An example of this is shown in the figure below, where the left panel shows a trace from a master event with Ml 1.72, and the right panel shows a detected event with Ml -1.01. The two traces have a maximum cross-correlation coefficient of 0.91 although the maximum amplitude differs by a factor of 500, confirming the magnitude difference of log(500) = 2.7. This particular event pair has four stations with a maximum cross-correlation coefficient greater than 0.9, all of which show a similar difference in maximum amplitude.



- Line 291: A mean location uncertainty of only 300m? Most earthquakes are located outside the network (Figure 4) so I would expect a much higher number.

  Although it is true that some events are located slightly outside of the network, this is only a small fraction of the event catalogue. The density of the events within the network area is much higher, something that can not be directly derived from Figure 4, as there are very many overlapping points in the central parts of the map. On request of the second referee, we included a figure with the original NonLinLoc errors to support this claim.

- Line 297: "upper bound" on location uncertainty. I would avoid making a quantitative statement from a hand wavy argument. Is the idea that events with CC>0.9 are more-or-less exactly co-located? In this case, shouldn't it be a lower bound (true uncertainty = within-family dispersion + NLLoc uncertainties)?

  If we assume that events with CC>=0.9 are exactly colocated, we can attribute the dispersion to location uncertainty entirely. In reality, part of the dispersion may be natural (e.g. events are not exactly colocated), and in this case the contribution of location uncertainty would be smaller. Therefore we call it the upper bound.

- It would have been nice to have the catalog as a supplementary file. The authors do say that they are willing to share the file upon request, but this seems incompatible with anonymity.

  Yes, the final manuscript will contain a link to the event catalog. The statement referred to was intended for the referees during the review process.