The abstract is quite lengthly and maybe not very focused on the core content of the paper itself. For example the part about improving meteorological models, volcanic explosion detection is too long and would better suit in the introduction than in the abstract. Same about the fact that the IMS is supposed to detect 1 KT explosions, information about waveguides, etc. I would suggest to move all this in the Introduction. The key message in the abstract should probably be that a high-quality dataset is made available to the scientific community. Then the authors could briefly describe the dataset and then list all the possible/foreseen applications of this dataset.

Line 12. "The IMS is supposed to detect any explosion of at least 1 kt of TNT equivalent underground, underwater, and in the atmosphere." -> There is no 1 kT minimum
requirement within the CTBT. Maybe the sentence could be rephrase as “The IMS was initially designed to be able to detect any [...]” This would also be more in line with the “design goals” that are mentioned Line 121.

Line 40. I would remove the “respectively” as several technologies can detect a test in a same environment. Especially as later (Line 65), the authors provide examples of infrasound detection produced by underground explosions.

Line 41. I am not sure about the wording “radionuclide detector” vs. “stations” for waveform technologies. I would also suggest to use station for radionuclide technologies as a single station can include several “detectors”. Or maybe use “facility”.

Line 41. I would maybe add “16 radionuclide laboratories” as those only apply to this technology.

Line 59. "a flat response from 0.02 Hz to 4 Hz”. The response is flat but +/-3 dB over this passband. This does not need to be added in the introduction but I did not find any information in the rest of the paper whether or not the amplitudes were corrected in the dataset from the response of the sensors used at the stations. This should mainly impact the lowest frequency bands. Something should be said about this in the paper.

Line 130. IMS Stations are all referred to as “ISXX” in the paper. This has often been used in the past and is a minor comment only. Another option would be to use the official station names (as defined by ISC or CTBT) such as I01AR, etc. But that would probably be too much work as all data files are already named this way. Just a recommendation for the future.
Station upgrades also lead to lacks of data since these often require a station to be revalidated. The revalidation process in most cases does not affect data availability. The upgrade process can (power off, etc.) but not the revalidation process. Station could be taken out of processing during the revalidation process (although not common anymore) but data availability would stay high.

The more sensors are progressively incorporated (generally from the inside to the outside of an array), the more potential aliasing is limited", I am not convinced that the limited aliasing is the main factor that allow improving the computation of wave parameters when the number of sensors increases for the selected processing technique (PMCC). Could the authors add some explanation here.

Pixels adjacent to others in terms of time, back azimuth, and apparent velocity are grouped into detection families if at least 10 pixels contribute” -> not frequency ?

I would recommend to add what is the source of each of these artifact categories and how the applied criteria help filtering each of these artifacts. No explanation is given neither on how the thresholds such as the family size of 40 or 50 were chosen. The word “obvious” is used, but it is not obvious when reading the paper that a family size of 39 would be an artifact but 41 would definitely not be. So the chosen thresholds were probably defined based on statical analysis (ROC curves ?) and probably do not set a 100% clear line between “obvious artifacts” and real events. This also makes the sentence “We post-process the detection lists to discard obvious artefacts” quite strong statement as real events might have been filtered our as well.

Effectively raising the lower family size threshold ensures the global
comparability of the stations’ detection lists and the derived products”: Maybe not so clear for the reader. What derived products?

No comments are made in Section 2 and 3 about the choice of the filter bank vs. the shape of the infrasound spectra. Were the spectra flattened before applying the PMCC processing? If not, do we expect the frequency of the detection to be shifted towards higher detection because of leakage effects. Was some testing performed to compare the processing results with flatten spectra vs. raw data. Are the filters sharp enough?

Line 310: “The white vertical lines near the center frequencies in (a) result from cleaning the detection list of ringing artefacts;

with the newer version and configuration, the cleaning is easier to narrow down to the respective center frequencies.” I think this is an interesting comment that only appears in the figure caption. This should probably be added to the text and a link made with the discussion at the end of Section 2.2.

Section 3.3 (general comment as well): According to the authors, this dataset is mainly made available because the raw data is not available to the scientific community. But I think these dataset is also very valuable for those who have access to IMS data including the CTBTO. It allows identifying station performance issues that could be reported to the CTBTO and provide a very valuable dataset to all NDCs. Not all NDCs have the resources to compute such dataset and the fact that the authors made this available is great for everybody (not only those who don’t have access). This is a positive point that should be emphasized more in the paper I think. I read it in the conclusion afterwards but it should probably be highlighted earlier in the paper.

Line 369-376: There should probably be more explanation about why the microbarom band was divided into 2 categories. Because for the reader, it does not really appear very clearly in Figure 3 for example that there is 2 distinct categories in this frequency band.
Maybe a some explanation could be added about the different use cases of these 2 datasets (and because this allows to have 2 different values instead of an averaged one over the entire frequency band).

Fig 8(b) the purple square (0.02-0.07 Hz) is shifted in time by about 20 min compared to the detection (does not align with the detection). This is probably because of the time is defined as the middle of a predefined time windows but something should probably be said about it.

Figure 9. The legend about what the color-coded squares are is missing.

Line 681-684: This is true and should be included in the paper, but I am not sure it is one of the main highlights of the paper that should be included in a conclusion. As for the abstract itself, the conclusion could be slightly re-written to better summarize the main points of the paper with possible opening (future work).