Reviewer B:
This manuscript summarises the development of a labelled dataset of seismic waveforms from regional earthquakes (defined by the authors as being recorded at distances of 1-20 degrees from the source). This dataset is intended to be used to aid the development of machine learning algorithms focused on regional earthquakes. Most recent applications of machine learning to seismic waveforms have been focused on local earthquakes in part due to the more comprehensive suite of labelled data for local seismicity to train these models on. I expect that this new dataset will be a useful and well-used aid for development of the next generation of machine learning models. The authors have explained their curation process well, and thanks to this curation process, this dataset is likely to be useful. I have several minor comments that I hope will clarify the manuscript, and suggestions for additional metadata that I would like to see addressed before final acceptance.
Comments:

1. The change in central longitude between Figures 2 and 3 was somewhat confusing. I suggest you chose one and stick to it.

Figure 2 is more focused on instrumentation and monitoring, where the land masses are emphasized. On the other hand, Figure 3 presents the distribution of hypocenters in the CREW dataset, of which the Pacific ring of fire is the most remarkable feature. In our opinion, having different central longitudes clarifies the presentation.

(similar comment by reviewer A)

Line 126: You note that you only retrieved earthquake catalogs, but in your discussion you note that this will be a useful dataset of test ban treaty monitoring. Did you explicitly exclude blasts from this dataset? If you did, it might help to clarify how you see this dataset helping with treaty monitoring. If you did not and there are explosions in this dataset please include that in the metadata (e.g. something like QuakeML's eventtype attribute).

There is an attribute in the data and metadata files called 'event_type', and the counts are:

'earthquake': 1567159, 'rock burst': 2592, "mining explosion': 21347,

'explosion': 4040, 'nuclear explosion': 3, 'induced or triggered event': 3898,

'experimental explosion': 234, 'collapse': 29, 'not existing': 19, 'not reported': 2}

The intention with the phrase is that the models trained with this dataset will be most useful at monitoring seismicity in places far from instrumentation, which is typically the case for explosions and nuclear explosions in certain regions of the world.

2. I was a little lost in the description of the manual intervention (lines 184-218), and the following points reflect that confusion, perhaps a couple of clarifications would help:

   This section has been extensively reworked to address the concerns of the reviewer, and to simplify the workflow. Instead of using a model that detects the location of the earthquake signal, we trained a phase picker for P and S waves. Initially we trained the model on the whole pool of data and then filtered the dataset based on the fit between the labeled picks and those predicted by our model.

   1. Line 190: Did you remove 28% of the random 10k examples, or 28% of the whole dataset. I assume you are referring to the random sample. Was it based on this manual analysis that you decided to implement the Nueral Network to remove bad data?

      The 28% is from the original pool of data, started at 3.3 M and ended at ~2.2M. It was based on the manual analysis that I decided a massive cleanup was needed and due to the scale of the data, ML models were best suited

   2. Line 194: Is this the initial data pool after you removed 28% of the 10k manually checked? This reads like this hasn't had any cleaning yet. If this is really the initial data pool it isn't clear where you removed the initial 28% from.

      From the initial 3.3M pool of data

   3. Line 199: Did you look at all 100,000 waveforms after each training step to remove faulty examples, and if so why did you have to look at the same data multiple times? Please clarify what waveforms you reviewed.

      No, we inspected a random sample each time, not the same data was reviewed every time, due to some example being removed after each pruning stage.

   4. Line 209: How did you determine this threshold? From your example in figure 8, this threshold clearly threw away good data as you point out and corrrect later.

      The threshold was empirically set, the BCE of 0.1 was the one that seemed to separate good data from bad data, better than other tried thresholds, both higher and lower.

   5. Line 218: What size of random sample was checked?

3. You note in the text that the y-axis of the middle panel of figure 6 is cut. Please put this note in the caption and state the frequency of the 1-2degree bar that is truncated.

   The y-axis has been replaced by a logarithmic scale and is not cut anymore.

4. Line 266: Can you specify what depth range your earthquakes fall in: I would be surprised (but happy to be wrong) if you have earthquakes spanning "all" depths (which I read as surface to core at least every few km in depth). Also line 146.

   min_depth=-4.9, max_depth=700 km. We mean all seismogenic depths here.

5. In figure 8, it looks like you include the location code in your metadata text to the right of the panel, but you don't include the location code in table 1. You should include location code to provide the full seed id of the recording instrument. There are multiple instances that I know of of a surface and borehole sensor attached to the same site, with the same channel code, but different location codes.

   Thanks for noting this point. However in the ISC catalog no location code is provided with the picks, so it is unavailable.

6. Similarly, it would be useful to include channel depth in your metadata alongside station elevation.

   Station elevation is in the metadata and data attributes as 'station_elevation_m'.

7. It would be useful to include reporting agency in your event information metadata so that people can find the event and any additional metadata they require for an event.

The latest version of the data includes only ID's from the ISC, as noted in the updated manuscript, with only one, we don't note it in every entry.

8. It would be helpful to include a pick method-id (as used in QuakeML under the pick.method_id attribute) unless all your picks are manual.

   This one is not provided with the ISC catalog, so it is not available for us to provide.

9. There are some incomplete references in the bibliography:
   1. Assran et al.
   2. Ng et al.
   3. Northcutt et al.
   4. Tan et al.

5. Zha et al.

Thanks for finding these. We have now fixed/added them.

Note that the Ng et al., is a website that links to a web competition.

10. It was disappointing not to be able to check the code (line 295). Although I do not feel I need to check now, please ensure that the url to the github repo is included. Ideally please also deposit the code in a version controlled repository site (e.g. zenodo) to ensure that the code that was initially used remains available and obvious.

The dataset will be hosted in Stanford University's Datafarm https://redivis.com/Stanford, where it will get a permanent URL and doi. While we finalize the license with the University and the provider (Redivis), reviewers can find part of the dataset here https://redivis.com/datasets/601a-5ygtwrrzb and example queries and notebook here https://redivis.com/projects/k4k5-1hdesm2pc

The scripts used for assembling the dataset are available in github, alongside example jupyter notebooks.

https://github.com/albertleonardo/CREW

11. Similarly, I'm not sure whether google-drive supports versioning, and the url was not provided. I would strongly encourage the authors to version their dataset and to link from the versioned dataset to the correct versioned code and vice-versa. I expect that CREW might require updating in the future and it may become important to know what version of CREW was used for a specific model.

The google drive version was an ad-hoc storage space so the reviewers had access to the dataset, improvised at the time. The dataset is now hosted in Stanford University's DataFarm, with versioning. In this space, users can make custom queries and download part or all of the dataset. (see point above)

In summary, after some minor changes I look forward to seeing the updated version of this likely useful dataset and manuscript accepted.

Reviewer A:

The manuscript “Curated Regional Earthquake Waveforms (CREW)” by Aguilar and Beroza introduces a novel training dataset for machine learning applications with a focus on regional distances (1 to 20 degrees). With the focus on this distance range, the authors identified a type of waveforms currently underrepresented in ML datasets. At the same time, the authors make a clear case why this distance range is important for seismic monitoring. The compilation and publication of the CREW dataset is, therefore, a valuable contribution for the community.

Overall, the paper is well-written even though some modifications I detail below would greatly improve the manuscript. Below I outline the larger points that should be addressed.

Regarding the data, I have two concerns. First, while the hdf5 format chosen is generally accessible and should provide good reading performance, the particular format has not been used before. With several benchmark datasets or frameworks published (STEAD, INSTANCE, PNW, MLAAPDE, SeisBench, QuakeLabeller), I think a new dataset should not introduce a novel format without good reason. In this case, I think the data could easily be presented in the format of a previously published dataset. Having community-agreed standard formats is essential to allow easy reuse and interchangeability of data sets. Second, the currently indicated sharing strategy is through Google Drive. This is not a permanent storage and, in particular, the data is not issued a persistent identifier. I'd urge the authors to deposit the data in a research data repository with DOI, for example, Zenodo.

Data attributes are now using mostly Seisbench notation. For example we differ from the two character channel identifier of Seisbench 'trace_channel' ,e.g HH, as we provide a list of the channels, e.g. HH1,HH2,HHZ. That are distinguishable from HHE,HHN,HHZ.

The google drive was an adhoc way of sharing the data with the reviewers while we sorted out permanent storage. The dataset will be hosted in Stanford University's Datafarm https://redivis.com/Stanford, where it will get a permanent URL and doi. While we finalize the license with the University and the provider (Redivis), reviewers can find part of the dataset here https://redivis.com/datasets/601a-5ygtwrrzb and example queries and notebook here https://redivis.com/projects/k4k5-1hdesm2pc

(See similar comment as reviewer B above)

One thing that should be clarified is the pick labelling strategy. The dataset contains P, Pg and Pn picks (and the same for S waves). What phases can the pick labeled P represent? In particular, looking at the dataset there are cases in which P, Pg and Pn are all labeled, without a visually separate arrival at the P label (2 plots attached). In the manuscript, there is also a mention of P and Pn labels at the same sample but labelled separately. In the interest of consistency, I think there should be a clear description what the P label means, e.g., always the first arriving P phase. Consequently, I'd suggest to either label all first arrivals with a P or none. Clearly, it will not be possible to identify the type of P phase in all examples. Upon inspection of the data, I found several example where there are P and Pn/Pg annotated with very close temporal succession. In these cases, I could not visually identify distinct arrivals for both phases. Therefore, I think, some form of label deduplication or at least a discussion of the issue is necessary.

We ensured the accuracy of the earliest arrival with our workflow. We leave the details of those secondary and/or duplicated arrivals for future research. Users might find it useful so we decided to keep them. We added a note of clarification about this in the manuscript with an example that shows this issue.

The introduction/background section of the manuscript is fairly long. Maybe it could benefit from additional sub-headings to give it more structure. In addition, the long introduction makes it difficult to extract the key motivation. From what I understood, the key argument for the need for CREW is that there is a lack of regional-distance training datasets and that the performance of models trained on local distances will likely degraded when applied to regional distances. I find this claim not fully convincing. Even though it is likely, the claim is not backed with references or tests. In particular the paragraph starting at line 48, discussing distances over 100 km as potentially degraded is exaggerating as such distances are already contained in INSTANCE and STEAD.

You extracted the gist of the introduction as we intended, so the intro is doing its job. I have only anecdotal evidence of the performance degradation as the source to receiver distance increases. Once the S-P time is close to the window length of 1 minute that most available models work with, the wavetrend does not fit, which is where our dataset and subsequent models will fill in.

Nonetheless, I don’t think that the lacking quantification of performance degradation with distance is an issue to the publication of this manuscript of CREW itself. Without datasets such as CREW it will not be possible to measure the impact of this data gap and the gap is undeniable. I’d just suggest the authors to rephrase the introduction in this respect in a more careful wording.

Thanks

In the current version of the manuscript, references to the underlying raw data are missing. Please reference all used seismic networks using the DOI and all FDSN webservices queried. This is essential to give credit to the data collecting agencies and in many cases mandated by their license agreements.

All seismic networks are now referenced in the manuscript, those that have a citation and a doi are cited, and those for which no doi exists are only mentioned (same as reviewer B)

Given the points listed above, and the smaller points listed below, I think the manuscript requires revision before publication in Seismica. Some of my comments below refer to my observations from inspecting an example chunk of the data. I believe this manuscript and the CREW dataset will be a useful contribution to the development of ML models for observational seismology.

Minor points:

- Line 34: The Pacific Northwest Dataset contains noise examples too.

  Noted and clarified.

- In the list of seismic benchmark datasets, the datasets published with the SeisBench library (Woollam et al. 2022) are not included. I think, in particular, the GEOFON dataset should be mentioned, as it covers teleseismic distances.

  A short comment has been added regarding this dataset.

- Please include the recently published MLAAPDE dataset in the introduction (Cole et al 2023, https://doi.org/10.1785/0220230021). This dataset (partially) covers a similar distance range as CREW.

  Added the citation and the comment, it was not published at the time of writing of the manuscript.

- Line 68ff: I think this description should be in the figure caption rather than the main text.
  We believe this paragraph helps us guide the reader through the structure and characteristics of the dataset, thus we would like to retain it, while also adding some of this information into the figure caption.

- Figures 2 and 3: I find it difficult identifying the land-sea boundary (and therefore orienting myself) on the map. I'd recommend to remove topography from the map and instead show the coastlines.

  We increased the darkness and thickness of the coastlines, and made the topography lighter. We would like to retain the topography.

- Figures 2 and 3: Please use the same central longitude for both plots

The purpose of figure 2 is towards showing monitoring, which is central on the continents and the central longitude is chosen such that the contours are not cut, but Figure 3 displays the seismicity mainly, and since most of it is around the Pacific the central longitude is chosen such that the seismicity is displayed as continuously.

(Same point as reviewer B)

- Figures 2 and 3: Please replace the rainbow colormap with a perceptually uniform colormap.

  (See https://www.nature.com/articles/s41467-020-19160-7 for details)

  Replaced the colormap with the viridis colormap, which is perceptually uniform.

- Figure 3: I'd recommend increasing the figure size to that of Figure 2.

  The figure shows the global distribution of seismicity and stations, there is nothing novel here. Also, the size of this Figure is such that it fits in one row of Seismica's publication format.

Line 125: Please list all queried data centers explicitly. Obspy can query any data center with an FDSN implementation.

All the seismic networks used in the dataset have been cited. The ones for which a doi exists are included, the ones for which it doesn't are mentioned.

(Similar comment by same reviewer addressed above)

- Line 126: "We only retain [...] S, Sg, Sn." → I'm not sure how to interpret this statement. Please specify if this relates to the catalogs as a whole (at least on of these arrivals in the whole catalog) or every single event-station pair.

  This refers to catalogs as a whole. For instance, the NCEDC catalog only reports P arrivals, no S arrivals are reported. Thus, this datacenter is not used.

- Line 140: “Ultimately [...]” → Does this mean that the majority of the data comes from the ISC catalog and from the IRIS DMC? I’d suggest a clarification here.

  In the updated version we retained only data from the ISC catalog. It has been explicitly noted in the manuscript.

- Line 149: “We display [...]” should in my view be part of the figure caption instead.

  It would be too long of a caption and we feel it keeps the walk through the characteristics of the dataset clear.

- Line 192: Please provide details for the model you used. This should include at least model architecture, train/test split, hyperparameters, and evaluation metrics. Possibly this could be part of a supplementary material, as it’s important for the validity of the results but not integral to the main story.

  The model architecture is now mentioned, but the details of this model and other models trained on CREW will be presented in a separate, forthcoming paper.

- Line 208: For assessing the feasibility of the semi-supervised approach, there should be a quantitative assessment of the classifiers performance on a hand-labeled test set.

  We agree that a quantitative assessment would be required, but that is beyond the scope of this paper.

- The figures are not ordered in the same order they are referenced in the paper (Figure 8 is referenced early). I’d suggest reordering the Figures.

  Thanks for catching this, we corrected it.

- Line 221: I find the observation that the model performs substantially better on the bandpass filtered data surprising, as the model could in principle learn the bandpass filter. Can you add a hypothesis what this is related to? My personal guess would be either limited training data or model architecture.

With the right choice of filters, the signal to noise ratio increases dramatically when bandpassing. This is a standard technique for making phase picks in noisy data because it makes it easier to see the arrivals, and it makes it correspondingly easier for the models to learn to pick. Given enough data, and enough training, it should also be possible for the model to learn to do that, as the reviewer suggests. In the updated version, this section has been replaced by a workflow based on phase picking rather than the detection of the location of the earthquake signal.

- Figure 5, panel D: I think this panel should be rescaled. The spiky signals in the data make it impossible to visually check if there a P and S arrivals at the indicated locations or not.

  We agree that the scale doesn't allow for proper checking of the presence of the arrivals, but in our experience, when the scale of the spikes is orders of magnitude larger than that of the earthquake signal, these examples lead to problems in training, so we decided to remove them.

- Figure 6: I think it would be helpful to indicate how many waveforms are contained at distances not already available in other benchmark datasets, maybe above 6 degrees, the approximate maximum distance in INSTANCE.

  We want to avoid a long section of comparing datasets, but for example: Instance contains 21,362 labeled S arrivals over 1 degree of distance, and only 1257 over 2 degrees. For comparison, CREW contains 608,079 examples with distances over 2 degrees, each one with both P and S arrivals, and 103,318 examples with distances larger than 6 degrees with both P and S labeled arrivals. CREW is the only dataset in the regional distance range that ensures that every example has both P and S labels. A line was added to the text that states this. "CREW contains more examples with both P and S arrival information than other datasets that cover the same distance range."

- Line 265: I'd recommend deleting this paragraph. It feels like an early conclusion that is directly followed by the actual conclusion.

We believe this section of the paper needs a short wrap up paragraph, otherwise it feels the end of the section is too abrupt. We would like to retain this paragraph as it keeps the flow of the reading.

- Line 277: The word “propel” seems ambitious and I would recommend exchanging it for a less strong word.

We believe this dataset will propel developments in the field, but we agree to replace the wording, for the word enable.

- Currently the paper contains both the spelling “dataset” and “data set”. I’d recommend a consistent choice.

Thanks for the catch and the suggestion.

- Missing samples in the data are currently indicated as 0 in the “*_arrival_sample” attributes. This should be changed to NaN to avoid downstream issues.

Good point! Noted and modified, missing picks are now NaN.

- For the metadata, I think it is suboptimal to only store them as attributes of the hdf5 data groups. This means that any filtering operation, e.g., by magnitude or distance, requires parsing all entries of the hdf5. Instead, I’d recommend an additional metadata table.

Metadata is available as a separate csv file. In the drive folder there was a file CREW_metadata.csv that contains the metadata only. In the permanent storage space on redivis users can query the metadata and then collect the resulting data.

- Judging by the data, the waveforms have been normalized using the peak amplitude. However, I did not find a description of this process in the paper and would ask to add this.

Yes, normalized by the peak amplitude, and we have noted that in the manuscript now.

- Please specify if the instrument response has been restituted.

  No instrument response has been removed, as indicated in the manuscript.

- The reporting authority of the picks should be noted in the metadata.

  All the picks come from the ISC as noted in the manuscript now.

- Please specify the total amount of data in terms of storage needs.

  1.1 TB data, 578 MB metadata, which is now noted in the text.

- (nitpick, feel free to ignore) The terms “machine learning” and “deep learning” are currently capitalised in the manuscript. I’d argue they can’t be considered proper nouns anymore and would recommend not capitalising them. Clearly a matter of taste though.

  They are capitalized to emphasize that the acronyms ML and DL are used in the text, and indicate their correspondence. Thanks for noting this.

In the revised version of the manuscript “Curated Regional Earthquake Waveforms (CREW) Dataset”, the authors addressed many of the critical comments raised by the reviewers. This has improved the quality of the manuscript.
Nonetheless, I think it is necessary to reiterate one of my main points of criticism from the initial review round. While the authors have modified the metadata naming scheme to resemble the one of SeisBench, the format of the data is still not compatible with any other dataset, in a sense that software written for other datasets will not work with the CREW dataset as is. However, such compatibility is integral for the aspect of interoperability within the FAIR principles for open data. In particular, for benchmark datasets such interoperability is desirable as it allows users to easily compare there models on multiple datasets. I admit that this is some additional work for the authors but in my perception comes at a substantial added value to the community.
Asides this point, I only have a few minor points listed below, that should be addressed before acceptance. The CREW dataset will be a valuable contribution for the community.

We agree with the reviewer that making access to CREW easier will help users create implementation faster. However, with the years of work we have built on the dataset it is impractical for us to change it right now, given the amount of code that has been developed on top of it. We provide examples on how to convert the dataset to Seisbench. Although easy to do, we do not have the space to host that version at the moment, but will work to make it in the future.
In the meantime, we provide a notebook with details about converting the format here:
https://github.com/albertleonardo/CREW/blob/main/CREW_convert_to_Seisbench.ipynb

Minor comments:

- Some waveforms (e.g., examples 0, 5, 7 in 000_CREW.hdf5) seem to be zero-padded at the end and then postprocessed, leading to low amplitudes after the actual signal. This makes it difficult to remove this padding again in application. Maybe this padding could be removed in favour of mixed-length traces. Alternatively, the number of padding samples could be specified in the metadata. Otherwise padding might lead to artifacts in models trained on the CREW dataset.

Yes, some waveforms were zero padded to complete the 300 s duration. The postprocessing consists of detrending, demeaning), resampling, padding when necessary and then normalizing. Lines 142-147.

We kept the padded version over mixed length traces because the data is read and fed into algorithms that require all traces to be the same length in order to be cast into tensors in popular frameworks like keras and pytorch.

We have added a metadata entry to the metadata files: -trace_completeness- (following seisbench convention), which goes between zero and one, reflecting the fraction of samples in the trace that are not from padding.

Given the quality control process, we have eliminated most artifacts if they occurred. For instance Figure 6E shows an example for which a lot of zero padding occurred, such that the Sn arrival is not present. Our quality control process eliminated that example.

- The sampling rate of the traces is not specified in the metadata. I'd recommend adding it as a column to be easily available for users.

We added the attribute 'trace_sampling_rate_hz" (seisbench notation) to the metadata, which is 100 Hz for all of the dataset.

- The clarification that the PNW dataset contains noise examples seems to have not been added to the manuscript, contrary to the statement in the response letter.

We moved one line of text down to include the PNW dataset among the ones that contain noise samples. Line 36 now reads: 'These four datasets also contain noise waveforms.' This statement includes the PNW dataset. Since noise is not central to the review the paragraph intends, we do not elaborate on it.

Not to editor about Figure 3:

In addition I would recommend adjusting figure 3 to include both the Pacific centered and African centered views & to make the figure width larger; one of the advantages of Seismica is that we will refine the layout to best match the requirements of your paper (within the page margin limit).

We chose the size of this figure based on its fit into the seismica overleaf template (a version of the manuscript on that template has been also uploaded). We believe this size as a page wide figure suits the figure very well and makes it look elegant.