Letter to Reviewers:

Dear Reviewers,

Thank you for your thorough review of our manuscript. Your insightful comments have significantly improved its quality. We have addressed all applicable suggestions and provided responses to the remaining points in the following pages.

Best regards,

Sina Sabermahani and Andrew Frederiksen

Reviewer A

1. It is not entirely clear to me why this model is supposedly applicable outside of the region used for training/testing, while the similar previous approach (Gong et al. 2022) was not. Can the authors offer any insight into this beyond the fact that U-Net was a different and potentially better deep learning framework, or is that just it?

   While there is not a certain answer to this question, there some factors that cause our model to perform better. First, U-net is designed for feature extraction, and it excels at it and in our case, we need a model to recognize the features of good quality receiver functions. Another factor is that we used data augmentation techniques that not only increase the size of the dataset, but also make the dataset richer in terms of diversity. The third factor is that we validated the model from stations apart from what they are trained on, and it makes our model more general, as the model is not specific to data collected from certain tectonic regions.

2. Relatedly, lines 340-341 imply that DeepRFQC is meant to be applied to Canadian datasets. So, which is it supposed to be for – global or regional datasets?

   We trained the model based on regional data (mostly Churchill province) but tested the model on Canada wide stations and it performs well enough. We predict it works well on at least the North American continent. However, we are optimistic that it will work at any other continental seismic station, though this requires more investigation.

3. A definition of what you mean by "quality control" actually seems like it would be important to include (lines 69-70). Does it mean removing type 2 error (i.e. traces where a simple SNR screen on waveforms would still keep some things that produce "bad" RFs")? Is "low quality" always reducible to a visual criterion in the end, or is there some quantitative basis?

   Your suggestion is applied.

   Yes, when we talk about "low quality", we talk about signals that we can visually recognize to be noisy. However, we had criteria for visual inspection of RFs (Lines 183-194).

4. In the description of augmentation, I was a little confused by the 'rolling' method – it is not actually listed in the cited reference (Iwana and Uchida, 2022). Are all components of the waveform shifted equally, or are Z/R/T shifted different amounts for one event/station pair? Is the phase pick adjusted accordingly? Basically, it's not clear to me how this method works and why it should be effective.

   This technique wrapping around the signal by several steps like turning A-B-C-D to B-C-D-A by one step. It is added to the text. We applied the technique exclusively to the radial component of the seismic data. As demonstrated in Table 2, the efficacy of this augmentation method is substantial. When this technique is omitted from the process, we observe a significant increase in training loss. The underlying rationale for this approach is that P-wave arrivals do not consistently occur at their predicted times. By

implementing small temporal shifts in the data, we can simulate this natural variability in arrival times, thereby enhancing the robustness of our model.

The citation was mistaken, and the rolling method is not introduced in Iwana and Uchida, 2022. Although it is a widely used method with different names including but limited to "rolling", "time shifting", and "translation". In the corrected citation, they used "translation" as the terminology.

5. I would like to see a bit more discussion of limitations of DeepRFQC. I know this isn't explainable ML, but some sense of what the model is picking up on and how that might or might not apply to other cases would be helpful. In particular, I'm thinking about choice of frequency bands, and about targeting deeper phases vs crustal/lithospheric – what range of other datasets might this actually apply to directly? What kinds of testing must a user do to ensure that it would work for their dataset?

   We tested different frequency bands to create receiver functions and then manually label them. The problem of systematically considering different frequency bands is that we must create different datasets using different frequency bands and then label them and then train the model based on them and it takes too much time.

   For a user to be make sure if the results are reliable, it is on them to inspect their results one-by-one, but the advantage of using a model like the one we introduced is that a user does not need to inspect 10,000 or even more waveforms to find out which one is useful. The main feature of this model is saving time, but we tested the model on different seismic stations' data to show that the results are quite reliable.

6. The section on H-k stacking (5.1) needs a few things: first, it is never actually stated that the H-k stacking results are similar for the "manual" vs DeepRFQC datasets, though I assume that is the point being made by figures 5 and 6. Second, what actually was the "manual" QC? Was it trace-by-trace visual inspection, or was it using the labeling program described at ~line 185?

   A paragraph added to explain we are not claiming that this model outperforms human performance as it is built upon manual quality control.

   It was a trace-by-trace visual inspection, and that mentioned program is a tool that we created to streamline the inspection process.

7. For the training/validation vs testing datasets, were there any significant differences in X5? Were the sensors the same type as the other networks? Was everything about the networks the same except for when they were deployed and exact station locations?

   The network X5 (HuBLE project) did not have any special characteristics. We had four networks in the study area, and we wanted to isolate one of them to test our model on it. We followed the same procedure for the PO network, and the results were quite similar. To keep the paper concise, we did not mention that as it is not adding anything informative. To enhance the model's reliability, in addition to testing it on the isolated dataset, we tested the model using data from three stations outside the study area. This ensured that the model was not specifically trained on data from that particular area.

8. Lines 64-65: water level deconvolution is computationally cheap, I suppose, and relatively easy to use, but I'd argue that there are better methods (especially multitaper) that are now widely available. Maybe providing some simple justification here for using water level would be helpful, and I would like to know also if the deconvolution method matters much for the application of DeepRFQC. Iterative decon, for example, tends to produce RFs that "look" a little different from other methods in terms of the frequency content/noise. If DeepRFQC is only intended for water level, that's fine, but it should be clearly stated.

   We add a figure in supplementary materials (Figure S.7) to show the effect of deconvolution techniques on final product. As it can be seen the model works well on multitatper deconvolved signals; however, more waveforms were extracted by water level, and it is not surprising as the model is trained on that.

9. In the first paragraph of the introduction (lines 53-60), I understand that there is no possible way to cite every study related to the broad topics mentioned. However, I am a bit puzzled by the choice of papers to cite. In particular, why Hansen and Schmandt 2017 for receiver function sensitivity instead of an earlier paper on the development of the RF method like Vinnik 1977 (doi: 10.1016/0031-9201(77)90008-5)? That's just one example, there are several from around that time, and more from the 1990s/2000s that significantly furthered method development. Similarly, the few references given for "surface wave analysis" and "understanding tectonic processes" are having to do a *lot* of work as representatives of a massive literature. At the very least, those parenthetical citations should have "e.g." prepended to recognize that these are only some examples.

   Three additional references (Vinnik, 1997; Vinnik et al., 2004; Bensen et al., 2007) *are cited to acknowledge the pioneering contributions that laid the groundwork for the development of these methodologies.*

10. Line 206 says 84.8 second traces were used. That seems very long for RFs, particularly for water level. Why? Are there particular aspects of the coda that need to be captured that long after each event? How much does trace length influence the model?

    Half of this time is before P-wave arrival and that means we deemed only 42.4 seconds after P arrival. Having a longer waveform contributes to a more accurate model (highly probably) but it costs more processing time. In this research we followed the time length recommended by RfPy, the package we used for receiver function deconvolution (Audet 2020 and Audet et al., 2020).

11. Line 124 paragraph, and the "Open research statement": please list all the networks you used and provide full citations with DOIs if available. If none of that, at least include URLs for FDSN and whatever the NRC seismic data portal is. Several if not all the software packages listed should also have full citations (I know that RFPy, Numpy, and GMT have citations at minimum).

    Citations are added.

12. The Ronneberger et al paper cited as the source for U-Net is not in the reference list. I have not cross-checked all the other references (I suggest that the authors do so) but this one seems particularly important.

    Fixed.

13. The paragraph starting at line 344 makes me wish there was a comparison provided between DeepRFQC and other non-ML QC methods, especially the labeling scheme described on line 185 (though I can't really tell how hands-on that was based on the paper). How well does it out-perform SNR-based methods?

    We have added Figure S.8, to show the application of SNR as a tool for quality control and we see by higher SNR, the is quality getting closer to the results of DeepRFQC, but it sacrifices the number of waveforms.

14. Line 392 is cut off and ends with an incomplete sentence.

    The sentence should be removed. Fixed.

15. Line 200: I believe this should be Vp/Vs rather than Vs/Vp.

    Fixed.

16. Lines 395-6: I do not know what is meant by a "reliable means of preserving data integrity in seismological studies." Please clarify?

    We rewrote the corresponding paragraph to clarify.

17. Line 399 references "identified anomalies at specific stations" and I am not sure what that refers to. My best guess is section 6.1, and the differences in RFs seen there between different tectonic settings; if so, though, I don't think that "anomalies" is the best term to use here.

    Your points are correct, and we replaced it with "low quality waveforms".

Reviewer B

Review for “DeepRFQC: automating quality control for P-wave receiver function analysis using a U-net inspired network” The methodology presented in “DeepRFQC: automating quality control for P-wave receiver function analysis using a U-net inspired network” is innovative and valuable - it is exciting to see a manuscript addressing the process of culling RFs with advanced ML methods. The writing and figures are generally effective but need to be refined. The ML process/presentation also could be refined. Below I provide detailed suggestions, but here is a summary of the most important points:

1. Stations in the training/test/validation data sets should be random (not by network) and clearly described.

   Test set should be selected in a way that could simulate the real-world problem, which is not necessarily random.

   In our work, we had two main categories of dataset: “train” and “test and validation”. In each epoch, data for training and validation were selected randomly from “train” and “test and validation” set, respectively. To see the performance of the model, we finally applied the model to the test set. However, our testing was not limited to the test set and we tested the model on four different stations in Canada: PGC, ULM, EDM and SJNN.

2. Based on the writing, it appears that an automated (non-ML) process is used to label the data, when the primary point in the abstract/intro is that Ps RFs require tedious manual labeling. The authors should briefly describe the automated (non-ML) process and explain why the ML approach of this study is superior, otherwise this portion of the methods contradicts the study’s motivation.
   The data labeling was performed manually through tedious waveform-by-waveform quality control, not via an automated non-ML process. This laborious manual effort motivated exploring an ML approach to automate quality control, as described initially. The ML model was trained on the human-labeled dataset to replicate expert judgments efficiently, without manual inspection of every waveform.

3. A comparison with non-ML automated Ps RF culling methods would strengthen the study’s motivation. One simple comparison that should be added is to show how the results compare with simply using a minimum SNR criterion.
   We added Figure S.8 to address your valuable comment. In that figure, it is obvious that the higher SNR, the lower number of waveforms extracted from dataset. By SNR equal to 15, we only have 15 events and as we can see the quality of the dataset is lower than what we predicted using DeepRFQC.

Finally, further analysis of the actual receiver function results in context of the geologic setting and prior work would create a very well-rounded paper, although this addition is not necessary.

General comments

Additional comparison of the crustal structure observed in these results with local tectonic history and in comparison, to previous studies (either in the supplement or main text) could provide a more interesting paper. Other methodological papers often do this (e.g. Eilon et al., 2018).

Thompson et al. (2010) provided an excellent comparative example, as they applied H-k stacking to the same area. While results from other techniques are interesting to compare, such comparisons fall outside the scope of our research, which focuses specifically on quality control of receiver functions. Regarding Eilon et al., 2018, their stations are far away from the stations we used and there is no overlap.

26/40/115/Figure 1: The abstract emphasizes the THO as the focus of the study, but in Figure 1 it appears that many (most) stations used are not in the THO. General comments in the abstract, non-technical summary, and introduction about the THO should be generalized to something such as: Archean and Paleoproterozoic regions of northern Canada.

Fixed.

Main Text

Abstract:

28 - 1,508,449 features may be difficult to grasp for some readers. If possible, listing a few examples of features that are intuitive to a seismologist may make the abstract more effective.

These are elements of matrix in the U-net network and there is no meaning in them. Their combination makes the network. All values are available on the model file uploaded on GitHub.

32: No need for a hyphen in receiver function.

Fixed.

1 Introduction:

36: Ideally define the acronym for receiver functions (RFs) in this paragraph instead of waiting until methodology.

Fixed.

57-60: Please provide some mention of the magnitude of velocity change at the interface as well. Such as: “The arrival time of P-to-S converted phases after the direct P-wave corresponds to the depth of the interface at which the conversion occurred and the amplitude of this phase indicates the velocity contrast at the interface.”

Fixed.

64-65: It seems unnecessary to provide only a single example of a deconvolution approach. I would suggest either removing this sentence or providing a list that includes a few other deconvolution approaches that includes at least one time-domain approach.
Done.

85-86 / 91-92: There are some studies that use unsupervised ML techniques for semi-automated RF QC. One that comes to mind is Krueger et al. (2021).
Done.

98: Suggested rewording for clarity:
"In this study we focus on the Trans-Hudson Orogen in northern Canada. We initially used the model of Gong et al. (2022) to quality control Ps RFs but found this method to be insufficient. Consequently, …."
Fixed.

2 Data collection & preparation:
126-128: This sentence is unnecessary. (The data…temporal parameters.)
Removed.

133: Earlier, water-level deconvolution is cited with Ammon. Is this citation missing here?
Cited.

138: Some work has criticized the use of frequency domain deconvolution because of its tendency to introduce processing artifacts (e.g. Kind et al. 2020). Please acknowledge this here or at some later point in this paper.
Done.

141-146: This block of text (Upon these…was assembled) could be summarized: "We successfully performed deconvolution on 12,860 waveforms (Figure 2)."
Done.

147: Please provide the number of waveforms in each data set. For example: "…into three distinct categories: training data (# in set), validation data (# in set), and test data (2,226)."
Fixed. We changed the text a bit as validation and test set and subset of a dataset, data from X5 network.

148-149: Why is the X5 network notable?
We just wanted to keep a network unseen by the model to see how good the model predicts the unseen data. The size of this network was enough to be considered as test set. The other candidate was network CN.

148: It would be more thorough to have the networks evenly split up between the training/ validation/test sets. Single networks often have similar data/issues as they often use the same

seismometers, installation methods, and were often installed by the same people. These networks are also generally separated by region (Figure 1). I strongly recommend randomizing the stations between the sets.
Thank you for raising that point. In our study, our hypothesis was to train a model using data from a limited number of stations and then evaluate its performance on stations that were not included in the training process. This allowed us to assess the model's ability to generalize to unseen data.

Regarding your concern about overfitting, we thoroughly tested the trained model on stations located in completely different tectonic settings from those used for training. The model maintained its high performance, suggesting that overfitting was not a significant issue.

Additionally, it's worth noting that the distribution of stations in the X5 region is relatively random within the area and among other stations. This randomness in station locations further supports the robustness of our approach and the generalizability of the trained model.

We appreciate your feedback and are always open to constructive discussions to improve our methodology and enhance the reliability of our findings.

150-155: Often "synthetic data" in RFs refers to forward modeled RFs given a certain Earth structure. Here is a suggestion to make the meaning of the "synthetic" data in this study clearer & more concise: "To supplement the original dataset, we add white noise with a peak amplitude of 0.01 to the original set of RFs (Chang et al., 2022). By adding this additional synthetically noisy data to our dataset, we double the quantity of waveform data used in the analysis. After normalizing…"
Fixed.

155: The way the paragraph is ordered makes it appear that normalization of waveforms occurs after adding noise, but the following sentence contradicts this. If the normalization occurs before adding synthetic noise, could the set of normalization be described in the previous paragraph?
Those augmentation techniques are distinct from one another and operate independently, without any direct influence on each other's functionality.

158-165: Please describe the methodological steps in chronological order by moving this description to above the description of the synthetic noise process.
As mentioned in the previous comment, they are independent and there is no certain order to apply them.

159 & 164: Are the "high" and "good" quality datasets different? Are these high/good/bad equivalents to the acceptable/unacceptable labels used later? If so, please be consistent in language.
Fixed.

162: Scaling process: Does this mean that all of the data is no longer normalized to have amplitude +/- 1? If so, is the comment made about noise being 1% of amplitude at line 157 still true? Furthermore, after this augmentation, are the only data with max amplitudes exceeding +/-1 classified as “acceptable” - could this add an artificial feature that the NN could recognize?
Yes, this set of data is not normalized, however, each waveform is scaled randomly and individually.
This makes the dataset more diverse. Although comparing the other two augmentations, this technique is the less effective (Table 3).

165: After this augmentation, please provide an updated list of the total number of traces in the training/validation/test datasets?
Done.

3 Methodology
168: Are any of these features intuitive to a seismologist? If so, could a few examples of the intuitive ones be provided?
These are just weights (numbers) and not meaningful on their own.

177: This statement needs a reference.
Fixed.

180: The discussion of the “Panda”/”baby panda” approach may be opaque to some readers. Please describe the process of training the model more explicitly without this jargon. For example, describing training the model sequentially, individually adjusting hyperparameters (this is my understanding of “Pandas”).
Done.

183: If the acronym RF is used in this paper, it is best to define it earlier.
Done.

184-197: Please lay out the process of data labeling as more of a step-by-step list and explicitly state the role of the seismologist in the data labeling process (i.e. what is automated vs. manually done). More specific comments related to this:
Already fixed and explained.

185: As it reads now, this process seems to contradict the point made initially in this paper that the culling process for Ps RFs is done manually and very time consuming. If the labeling process (equivalent to a culling process) is automated in python, why not always use this process instead of going to the effort of building a NN? When reworking how to present the data labeling process, explicitly addressing this potential criticism will strengthen the paper.
Already fixed and explained.

192: By “scrutinized” does it meant that this is done manually?

We changed the word to "inspected" in order to make it clearer.

197: Was the labeling of the validation set done using a different method from the other data? If so, why? If this labeling is automated, why not use this instead of the NN?
We did the same way labeling for this stations and then using the trained model, we also labelled the same dataset to see the performance of the model (how similar are the results to human performance).

196: As noted at line 148, it would be more thorough to have the networks evenly split up between the training/validation/test sets. Please revisit the assignments of training/validation/test sets or describe the reasoning for keeping networks together.
In the area, we have 4 networks, and our hypothesis was to keep one of them unseen by the training process and let the rest contribute to training the model. When the model was trained, we supposed this model could predict the acceptable and unacceptable waveforms in the other network, and we saw that it worked. To make sure that the tectonic setting was not dominating the prediction results, we tested the model to predict other stations located in different tectonic settings, and we found that the results were acceptable.
We have trained the model using other networks, but the results were identical. The point is that different ways of assignment of training/validation/test sets are important when we see a problem in prediction like overfitting. When the model is tuned and tested on an unseen dataset and shows sufficient accuracy, changing the order of assignment may not bring any advantage to the matter. In practice, it makes no difference if we randomly choose waveforms for training from the entire dataset, but we thought it would be more informative for readers to know that we chose a few networks for the training set and the remaining ones for validation or testing.

4 Experimental Setup
Overall, this section may be unnecessary.
(1) The tested hyperparameters could be included as a table.
(2) Remaining information could be combined with Section 3 Methodology.
Please also add a discussion of training and validation loss. This could also include a figure of training and validation loss vs. training epoch.
Table 3 wraps up all hyperparameters and configurations tested but in the section we meant to briefly explain parameters involving in the developing the model.
To create the final model we used the transfer learning concept meaning that after producing many models with different configurations and hyperparameters, we select the best model and train the model the model several times to make the accuracy higher, so there is no meaningful plot for this section.

4.1 Training model parameters and setting
205 - In sections 2, 3, and/or 4 please add explicit descriptions on the size of the train-validation-test split of the labeled dataset. Additionally, Figure 2 needs to be adjusted to show this more clearly.

This information is present in section 2. Sections 3 and 4 are about the methodology and model configuration that are independent of data size.
In figure 2, our purpose was to show a comparison between data labeled as 0 and 1. We add the information in the caption of figure 2.

206 - What is the meaning of the number of samples for each waveform? It would be clearer to state the time range the waveforms are cropped to and the time step they are sampled at (i.e. each Ps RF is limited to -40 to 40s around the estimated parent P-arrival sampled every 0.25s). Furthermore, how is the data window picked? Was software, such as TauP used to estimate the P-arrival time? If so, please describe this process.
Fixed.

218 - This sentence needs additional clarification.
Fixed.

4.2 Fine-Tuned Training Strategies
Suggestion: Even if Section 4 is kept separate from 3, Sections 4.1 and 4.2 could be combined.
Fixed.

218: Is 96 in %?
Yes. It is a percentage, and we fixed the text.

5 Results
232-236: Is "training and test accuracy" meant in 232? Otherwise, the accuracy given in 236 (93%) for the validation contradicts the 97% for validation given in 234?
%93 is with respect to the whole test set but during training we have batch processing and the model randomly selects data based on the batch size, so it is a smaller set and the value is different.

240: Earlier it was stated that the entire X5 network was the validation set - this does not seem "random". Please explain.
X5 is used as test and validation set. When it was being used as validation set, in every epoch, only a certain number of waveforms (based on batch size) were randomly chosen.

6 Discussion
357: This may be out of the scope of this study, but adding a third label that represents "Somewhat acceptable" may be a nice addition. These waveforms could be weighted as 1/2 or 3/4 instead of fully in final stacks.
It's an interesting idea, and we had something similar in mind, but labeling a huge number of waveforms one by one and deciding which category each belongs to would take an enormous amount of time, making it unfeasible currently. In further studies, we'll try to do so, as it would make the network more versatile.

6.1 Additional Test

377: What automated quality control is used? Is it the same pre-processing as with RFs used earlier?

Whenever we say automated quality control, it is the method we developed. We edited the text to make it clearer.

380: Expand on “complex structure” with a sentence or two.

Fixed.

381: “delayed”?

Fixed.

382: Is “station” meant since the text only discusses the sedimentary basin station (EDM)?

Yes. Fixed.

383: What is meant by “expectations given the sedimentary basin’s influence on station positioning”? Could this statement be replaced with a sentence that is more specific?

Fixed.

386: Add the word “quantity” after “Data” - “Data quantity is the most important…”

Fixed.

7 Conclusion

393: Sentence not finished.

Fixed.

Open Research

Links for the data from each network used in this study should be provided here.

Fixed.

Main Text Figures

Figure 1:

Is it possible to provide dashed lines of estimates of where the tectonic boundaries extend toward the northeast?

It would be helpful to indicate which stations belong to the training/test/validation sets – possibly by making them different shapes in addition to the colors that indicate network.

The boundaries are digitized from Whitmeyer and Karlstrom, 2007 and this is what we have from that study.

We add a sentence to the figure caption emphasizing the color related to test dataset.

Figure 2:

What does the 0/1 represent in this figure?

Unacceptable and acceptable.
Fixed.

Validation dataset information is missing. Could this be included?
Validation is the test set but randomly selected based on batch size in each epoch.

The colors in the graph are slightly different from those in the legend.

Legend: All labels could be made more meaningful. For example:
count —> Number of Ps RFs
label —> Culling of Ps RF (labels for ML algorithm)
0/1 —> Keep/Remove or Unacceptable/Acceptable
train_label —> Training data
test_label —> Test data
Caption: Please more clearly describe the content of the figure.
Period is missing at the end of the caption sentence.
Fixed.

Figure 7:
This figure very effectively shows the improvement of this algorithm over the Gong algorithm for this data. Please add an additional column to this figure that includes a comparison of what simply using SNR to cull the data (SNR >= 3 or so) would look like. This would more effectively make the point as often the alternative to culling RF data by hand is simply applying an SNR criterion. This should also be discussed briefly in the text.
We add another plot in supplementary materials to show the performance SNR test on a new station, St. John's, Newfoundland (SJNN), to expand the testing part.

Tabel 2:
It may be more effective to plot this data for the main text (and reserve the table for the supplement).
Fixed.

Supplementary Material
Please include a summary of the supplement contents at the beginning.
Missing "return" after "Supplementary Materials:"
Fixed.