Dear Dr. Radiguet,

We sincerely thank you and the two reviewers for the thorough, professional reviews and consideration of our paper. Our responses to the reviewer comments are shown below in blue. Changes to the main text are also indicated in blue.

Best regards,

Amanda

Reviewer D:

This paper proposes a method for denoising 3-component high resolution GNSS time series in presence of an earthquake, using a U-net deep learning model derived from the DeepDenoiser developed for seismic waveforms (Zhu and Beroza, 2019). For this, it combines real HR-GNSS time series in absence of earthquakes (noise), with a simulation of 3-component displacement time series due to a earthquake using the FakeQuakes output (signal). This Fourier transform of this combined simulated time series is the input of a deep learning model, which in output gives the signal mask of the frequency content. 3 variations of this model are proposed and compared. The main difference with the DeepDenoiser is that the 3 components are used simultaneously, as 3 channels, because GNSS data is noiser and the components are highly linked together, which will help the network to distinguish the signal. The proposed method is tested on simulated noisy data, and then on real HR-GNSS on the Ridgecrest earthquakes, and is compared with strong-motion sites (assimilated as a ground truth). The results are promising.

The paper is well written and interesting, the method is well designed, different architectures are tested and the results are showing interesting properties, both on simulated and real data. The problem is of interest, and this kind of method could indeed be very useful for the community. I think that this paper is worth publishing, yet there are few points that need to be addressed first:

1) I think there is a misunderstanding of what is a sample in machine learning, which makes the paper unclear. 'To create an algorithm capable of separating signal from noise we need many thousands of samples of both signal (i.e. HR-GNSS displacement time series from real earthquakes) and noise', 'Overall our testing dataset includes 1,458,606 three component records split evenly between signal and noise', These sentences are clearly indicating to the reader that there are 2 types of input samples, the signal and the noise, as we would do for a classification task. But later on I think that what the authors would like to say is that they created samples by combining (adding) real noise and synthetic signal waveforms, i.e. the number of samples are 1,458,606/2. The problem is a multi-output regression, not a classification. This needs to be clarified, as it looks like the authors did not really understand what the machine learning model is doing. In particular, a workflow figure would be needed in order to be clear about it.

We did not refer to denoising as a classification problem. The addition of signal and noise to generate network input on is described in the section titled “2.3: Inputs, Outputs, and Model variations”. We used the term sample(s) to refer to signal recordings, noise recordings, the combined signal and noise recordings that are input to the model, the sampling rate of or number of samples in a timeseries. We agree this can be confusing. To obviate this confusion we’ve changed any referral to a signal or noise recordings to waveform. This distinguishes the purely signal or noise waveforms from their combination, which is the input to the CNN.

Also, there are not 1,458,606/2 samples. The generator augments the data by combining different signal and noise waveforms and shifting them in time, therefore there are far more possible combinations of signal, noise, and shift. To make this clear we explicitly state the number of signal and number of noise waveforms we used in L183-4. We also added the text “The combinations of signal and noise and the time-shifting are data augmentation strategies that significantly increase the size of the training dataset.” to section 2.3. We keep the term sampling rate since this is common nomenclature.

2) Why is there no comparison with traditional denoising techniques? I find it very strange, and it is usually a good habit to compare to a baseline. I am pretty sure that your model would be best, but this is clearly missing in the paper.

Great question. Traditional denoising techniques are not applicable or can actually worsen SNR at the frequencies present in the short time windows used in this study. For example, the most common type of denoising, sidereal filtering, can increase noise amplitude in the high-frequency band. We expanded the background on denoising techniques in the second paragraph to make this clear.

For example, sidereal filtering leverages repeating satellite-receiver geometries to correct for noise resulting from multipath errors (i.e. when a transmitted signal arrives at a receiver via an indirect path). Simply stated this technique involves taking displacements recorded during the previous orbital repeat period (the time since the satellite constellation was last in the same configuration), applying a low pass filter (e.g. 11 s corner frequency), and subtracting the filtered displacement record from the displacement recorded at the present time (e.g. Choi et al., 2004). Spatial filtering targets common-mode noise that’s highly correlated across GNSS stations in close spatial proximity (Wdowinski et al. 1997). This technique simply averages detrended records on many nearby stations and the resulting average is subtracted from each station. Principal component analysis (e.g. Dong et al. 2006, He et al. 2015) has been employed to remove long period noise (0.2-0.1 cycles/year) and various match-filtering approaches have also been employed to reduce noise levels (Frank, 2016; Rousset et al., 2019). With respect to HR-GNSS, there are a dearth of denoising techniques that are both applicable to high rate data and efficient. For example, sidereal filtering is one of the most commonly employed techniques but, because of the low-pass filtering, it does not apply to frequencies higher than the chosen corner frequency. Additionally, Geng et al. (2017) noted that sidereal filtering can also increase noise levels for periods

between 20 and 33 s (those authors used a 10 s corner frequency). While data-driven denoising strategies are capable of mitigating high-frequency noise (e.g. Li et al. 2018), as proposed they involve multiple decompositions using techniques such as empirical mode decomposition, which is known to be computationally time consuming.

3) Did you test on only-noise samples? I guess that in a real case scenario, you might want to use it on a continuous recording and in that case having waveforms without earthquakes would be important in order to not 'create' unexisting signals. Even if you will test it only on earthquakes, maybe sometimes the sensor will be too far and in that case you could end up in a 'signal free' sample.

Yes, the data augmentation strategy we employed allows for some or all of the windows to be noise in both training and the testing data. To test this explicitly we downloaded another 10000 noise waveforms and applied the model to them. We've added this figure to the supplement along with the caption.

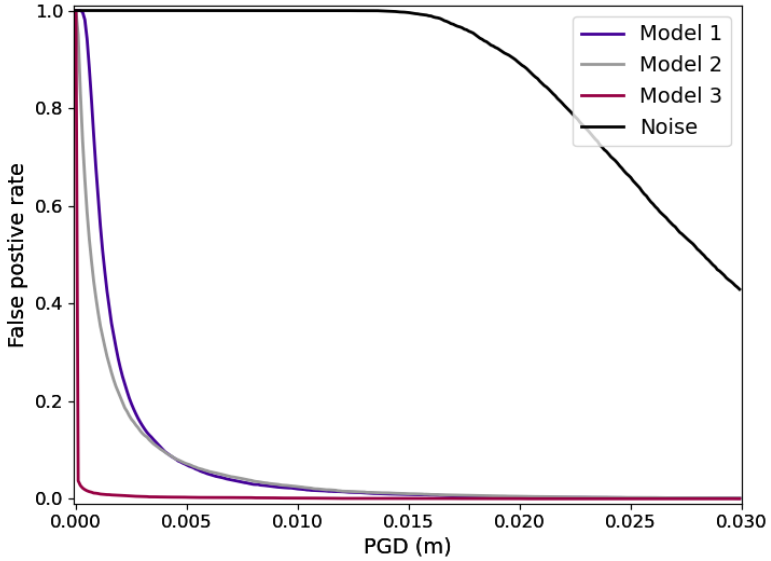


Figure S13. The false positive rate of Models 1, 2, and 3 applied to unseen noise data (N=10000) plotted against a PGD decision threshold. For comparison we also show the PGD of the noise waveforms the models are applied to. The false positive rate at a decision threshold of 1 cm PGD is 2.05%, 2.51%, and 0.16% for Models 1, 2, and 3 respectively. The false positive rate at a decision threshold of21 cm PGD is 0.4%, 0.53%, and 0.06% for Models 1, 2, and 3 respectively.

We also added the following texts to the end of the results section.

Finally, as denoising may be applied to HR-GNSS in real time, we explored the false positive rate of models 1, 2, and 3 on additional 10,000 noise waveforms. The results of this exercise is shown in Figure S13. The false positive rate at a decision threshold of 1 cm PGD is 2.05%, 2.51%, and 0.16% for Models 1, 2, and 3 respectively. The false positive rate at a decision threshold of 2 cm PGD is 0.4%, 0.53%, and 0.06%

for Models 1, 2, and 3 respectively. These values are far smaller than the PGD of the noise waveforms themselves. Additionally, they could be further reduced by considering waveform character at multiple stations as all earthquake early warning algorithms do.

4) It is not clear at first that your 3.1 and 3.2 sections are the results on synthetic data. If I understood correctly, these are the results on the test synthetic dataset (10% of your simulated samples).

Both the training and testing dataset contain synthetic signal and real noise records. Sections 3.1 and 3.2 do refer to the testing dataset. We've added "from the testing dataset" to line 343. Section 3.2 is titled "model performance on the testing data" and it is explicitly stated that we are evaluating performance on the testing dataset in the first sentence in this section.

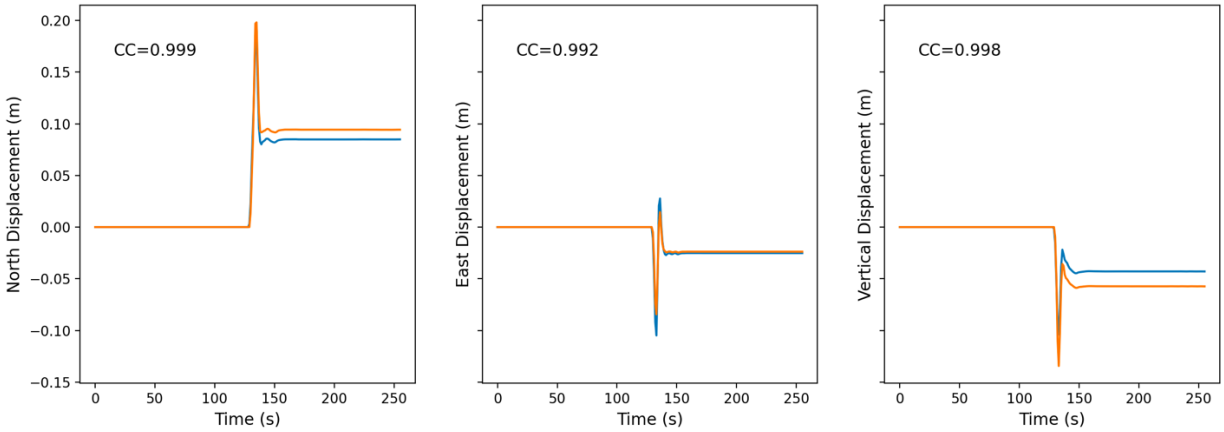
Usually, we use a 3rd set, called sometimes validation set, to select the hyperparameters of the model, to determine the best epoch to stop the training, etc. This is in order to avoid overfitting on the test set, and to show the results on completely unseen data. Did you use a validation data set?

Yes. Generating the synthetic earthquake waveforms for denoising took several iterations and our workflow was to generate a "small" (~10,000) waveform dataset to develop the model. We used this initial suite of synthetic waveforms to develop the CNN architectures mainly adjusting inputs and outputs but also evaluating optimal values of parameters such as the learning rate and number of epochs.

Also, did you pay attention to group all samples from the same earthquake event (registered by the different stations) on either train or test (otherwise it would be too easy, as nearby stations might have very similar signals)?

No, this is unnecessary. Depending on the kinematics of the earthquake there are sometimes similar waveforms on nearby stations but they are similar, not identical. The figure below shows an example of waveforms on nearby stations for the same earthquake. The waveforms have similar character but ultimately they have different amplitudes and static offsets so they constitute independent examples of earthquake waveforms. Beyond this, synthetic signal waveforms are augmented by adding independent noise samples so even if identical signal waveforms were included in training and testing data, the noise added to each would be different, as would the time

shift. We want an algorithm that can learn these differences.



There is also a precedent of not omitting similar waveforms from testing datasets in applications of machine learning in seismology. For example, one of the most highly cited papers on earthquake phase detection is Zhu and Beroza (2019). In that paper they obtain earthquake records on NCEDC stations. They do not omit waveforms because of waveform similarity but undoubtedly these exist in the dataset since there are many nearby or even repeating earthquakes in northern California.

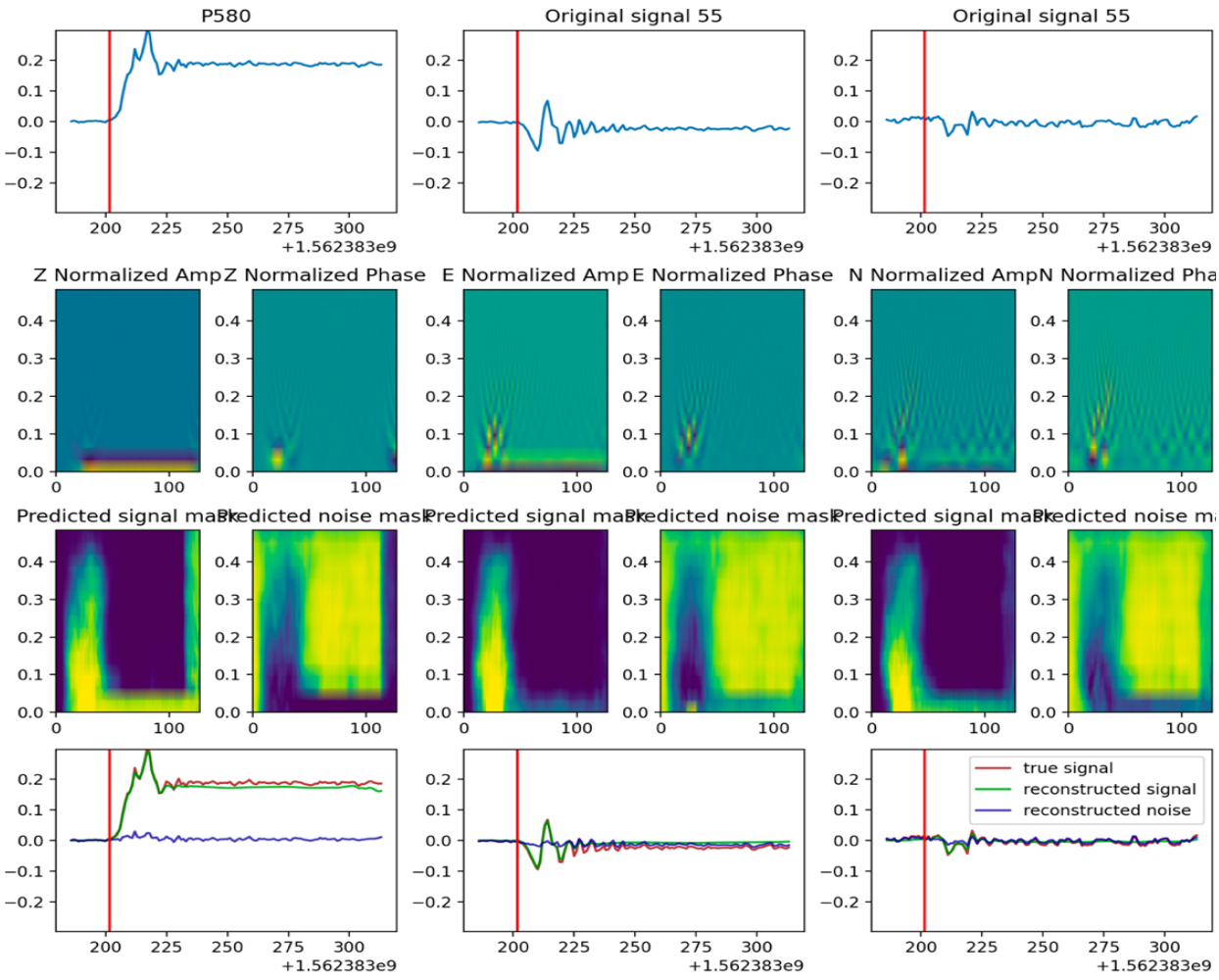
Smaller remarks:

- 'Finally, we test the models on HR-GNSS records from the Ridgecrest earthquakes recorded at stations that have nearly colocated strong-motion sites used ground-truth the denoising results.' --> word missing?

Indeed. We changed this to “Finally, we test the models on HR-GNSS records from the Ridgecrest earthquakes recorded at stations that have nearly colocated strong-motion sites **that can be** used ground-truth the denoising results.“

- l 370 : it would be interesting to see the complex valued masks predicted by models 2 and 3,

Original versions of Figures 6-11 had the phases and amplitudes of the complex masks. We opted for the simpler three panel figure because this ultimately results in an 18 panel figure and this seems a bit excessive for a typical reader (see below). If accepted for publication we will make all plotting codes available to the reader via Github and Zenodo and anyone interested can explore these masks at will.



and it would be interesting to see the differences between the 3 models on some common examples.

Figures S1-S10 show the performance of all three models on some common examples.

- Figure 13 L2 distances: did you look at the relative error?

No. You can design or utilize any number of performance metrics. We felt that the two we explored in the manuscript were sufficient.

- frequency content and amplitude seems to differ between simulated and real signals, as well as simulated and real noise. Have you looked at it?

The Ridgecrest waveforms in Figures 14 and 15 are one particular example of an earthquake. We did not tune our kinematic rupture simulations to reproduce the Ridgecrest events so given that any given kinematic rupture can have different magnitude, duration, rupture direction, speed, etc. there is no reason for the simulated waveforms to match the Ridgecrest waveforms identically. The main difference we noticed during the preparation of this manuscript was that the Ridgecrest events had

more “ringing” (i.e. long period coda) than the waveforms from the kinematic ruptures. This prompted us to go back and include soft layers in the Greens functions.

- 545 while we have demonstrated here the models --> word missing

Changed to “While we have demonstrated that the models we developed”

- This HR-GNSS: what about standard GNSS? Do you think your model would work on GNSS? If not, what would be needed to be changed?

Standard GNSS has much smaller errors than HR-GNSS so denoising is less critical there (though clearly still interesting). Beyond this, the goal of the paragraph on L564 is to say that, while we tried this single station algorithm, this isn’t really the way to do this problem. Noise is spatially and temporally correlated on GNSS stations hence using an algorithm that exploits that information is a better method for this problem. So I don’t want to add discussion on how to adapt this model to regular GNSS because I do not think that is a useful next step.

- Model 2: please add a figure to better show the model(s). Why does a linear activation make it possible to predict a complex value?

We’ve added “the real and imaginary parts” to line 274 to make this clear.

are all layers in your model for complex values (there are some papers specifically on CNN for complex values)?

No.

Recommendation: Revisions Required

Reviewer E:

Dear Editor and authors,
I find your work interesting and there is potential to be a good machine learning based denoising method for HR-GNSS data. However, there are several issues that I addressed below that are required some attention. I listed them as below:

Major/Moderate comments:

My main question about the paper is what about the ‘traditional’ denoising methods? Is any of these three models are better at denoising signals with respect to other machine learning models or non machine learning approaches? One can clearly see that the models are working. But we cannot see if they can actually provide any improvement on the satellite data denoising topic in general. For instance Zhu et al. 2019, compare the results with other methods. Is there a way to carry out a similar study?

Great question. Traditional denoising techniques are not applicable or can actually worsen SNR at the frequencies present in the short time windows used in this study. For example, the most common type of denoising, sidereal filtering, can increase noise amplitude in the high-frequency band. We expanded the background on denoising techniques in the second paragraph to make this clear.

"For example, sidereal filtering leverages repeating satellite-receiver geometries to correct for noise resulting from multipath errors (i.e. when a transmitted signal arrives at a receiver via an indirect path). Simply stated this technique involves taking displacements recorded during the previous orbital repeat period (the time since the satellite constellation was last in the same configuration), applying a low pass filter (e.g. 11 s corner frequency), and subtracting the filtered displacement record from the displacement recorded at the present time (e.g. Choi et al., 2004). Spatial filtering targets common-mode noise that's highly correlated across GNSS stations in close spatial proximity (Wdowinski et al. 1997). This technique simply averages detrended records on many nearby stations and the resulting average is subtracted from each station. Principal component analysis (e.g. Dong et al. 2006, He et al. 2015) has been employed to remove long period noise (0.2-0.1 cycles/year) and various match-filtering approaches have also been employed to reduce noise levels (Frank, 2016; Rousset et al., 2019). With respect to HR-GNSS, there are a dearth of denoising techniques that are both applicable to high rate data and efficient. For example, sidereal filtering is one of the most commonly employed techniques but, because of the low-pass filtering, it does not apply to frequencies higher than the chosen corner frequency. Additionally, Geng et al. (2017) noted that sidereal filtering can also increase noise levels for periods between 20 and 33 s (those authors used a 10 s corner frequency). While data-driven denoising strategies are capable of mitigating high-frequency noise (e.g. Li et al. 2018), as proposed they involve multiple decompositions using techniques such as empirical mode decomposition, which is known to be computationally time consuming."

The second doubt about the study is about the model prediction of the low SNR data from Ridgecrest earthquake. As given in supplementary material, the outcomes are zeros. However, in Figures like 10C, there are some improvement. Other figures like 9C, some others examples get almost zeros. In fact in line 503, it is stated that the biggest improvement are SNR between 1 to 3. However, for the performance analysis of the models (Figure 12 and Lines 417-423) very low SNR values are also included. I believe it is more explicitly stated in the conclusion that in very low SNR cases, the models fail.

We're not sure we fully understand the reviewers concern here. It would be difficult for any denoising method, machine learning or not, to recover signals with very low SNR. The stations shown in the supplement were the results of the Models applied to the waveforms of the smaller M6.4 Ridgecrest earthquake on stations that are 67 and 88 km away. The SNR on all channels is below 1 and the Models do not detect the earthquake. We stated this clearly in the Discussion on L550-553.

My last question is which model should we use in the end? In the paper, 3 models are compared with each other and their results are discussed. But in the end, we need to decide a model to denoise the data we have. In this study, it is not clearly stated which models should be preferred. The preferred model can be case specific and/or data specific which is fine. Model 2 shows better performance in the metrics that have been used. I would assume that Model 2 should be the final product of the study for the future studies. But it is not written in the conclusion. Some final thoughts about the model performance can be written in the conclusion.

In the discussion on line 536 we state "Models 2 and 3 were motivated by the amplitude distortion inherent in Model 1 which was originally developed for seismic data (Zhu et al. 2019). After assessing the performance of all models we find that Model 2, simple direct amplitude prediction, generally performs better than Models 1 and 3. It does not suffer from the amplitude distortion and is better at predicting both ringing and static offsets than the other models. By both performance metrics utilized here, it performs better at all SNRs than Model 1 and Model 3."

My other moderate to major comments are in below:
1. In the 2nd paragraph of Introduction, noise sources and noise reducing studies are not really matching. Giving some insight about the noise and studies that are dedicated to reduce them would be a better approach for the paragraph.

The second paragraph has been significantly modified. We hope these modifications address this concern.

2. In the 3rd paragraph of Introduction, other recent studies about denoising are not mentioned. Several examples can be seen below:
Tibi et al. 2021:
https://doi.org/10.1785/0120200292
Even though Tibi et al. 2021 is not cited, it is in the references. However, Tibi et al. 2019 which is cited in Line 2018 is not in the references.
Zhang et al. 2021:
https://doi.org/10.1093/gji/ggab099
Novoselov et al. 2022:
https://doi.org/10.1029/2021JB023183

We have corrected the Tibi citation.

There are many, many studies on seismic denoising. We chose to cite Zhu et al. and Tibi et al. because they use a similar ML model to the one applied to GNSS data in this study. The only commonality between the studies mentioned above and this manuscript is that they use ML (but different approaches on seismic data) so they do not seem particularly relevant to the present work and hence we opt not to cite them.

3. In Line 110-110, it is said that the time range where no earthquake with magnitude larger than M4.3 are selected. Is there any reason for the magnitude threshold? Is there

any possibility of contamination in noise database which may reduce the performance of the models?

We've added "This magnitude is a conservative lower bound for the size of earthquake you may expect to generate an observable signal in GNSS data." To L127.

4. In section 2.3, it'd be better to see the architecture of the models. Even though they are similar, there is no information about what do they look like. It would be nice to see how output parameters looks like. It doesn't necessarily be in the paper too. You may provide it in the supplementary, if you believe it will cover unnecessary space in the paper.

I will add the figures of the model architecture to the Zenodo/Github repositories. They are so large they cannot fit on a single page and are better viewed digitally (see below). We made note that these are available in the supplemental material. " Figures showing the architectures of Model 1 (v1_plot.h5.svg), Model 2 (v2_plot.h5.svg), and Model 3 (v3_plot.h5.svg) are available in the Github and Zenodo repositories that accompany the manuscript."

5. Line 292- What is the definition of prior to the first arrival? In the end of the paragraph it is written that the signal and noise samples be at least ten seconds in duration but it the noise is more than 10 seconds, did you use the 10 seconds before the P wave arrival or all noise samples before P arrival?

SNR is calculated using all samples prior to the P wave arrival. We require that this time window is a minimum of 10 seconds long so that $2 * \sigma_{noise}$ can be robustly estimated. This minimum window duration was determined empirically.

We changed this to "on the time window prior to the P-wave arrival" to make this clear. We don't want to use the word sample because of the first comment of reviewer #1.

6. Line 315-320 – Stations are not defined on any map. It would be nice to see them on map. It can be given in the supplementary material as well.

We added these to figure 1A.

7. Line 330 – Types of noises are not presented in the Data section. It may be necessary to give insight about the types of noises in the data section (2.1).

We added content on types of noise in the introduction.

8. In Section 3.2, SNR when you talk about the SNR, there is also ΔSNR which is mentioned in the figures but not in the text. ΔSNR plays an important role on the denoised signals and it needs to be presented in the text.

This is described on L320.

9. CC parameters varies between 0 and 1 which means that it is not susceptible on polarity changes. Even though in Figure 3L, CC has a negative value, in Figure 12 it varies between 0 and 1. Can you better explain how this process has been carried out?

CC is the normalized cross correlation coefficient that is commonly employed in GNSS and seismic studies. It varies between -1 and 1 so it is susceptible to polarity changes. We've adjusted the y-axis of figure 12 to make it clear that it can be negative.

We've added "First, we employ the normalized cross-correlation coefficient to measure the similarity between the signal and the model predicted signal, denoted as CC. CC varies between 1 when two signals are exactly correlated and -1 when they are exactly anticorrelated." After the definition of CC.

10. Line 573 – "these models" needs to be better explained (briefly) in the conclusion.

We've changed "these models" to "three machine learning models"

11. Line 577-579 – there are several exceptions in this part most of which are reducing the quality of the models. Maybe you can write a sentence about the limitations or weak sides of the models.

We don't understand the reviewers concern here. This sentence basically says that using three component data is beneficial which is true and one of the things this study did that previous studies did not. Most of the discussion (last three paragraphs) is dedicated to describing when the model doesn't perform and how it could be improved and alternative methods for GNSS denoising that may be better suited to the problem.

Minor Comments:
1. Figure is sometimes written with capital F and sometimes not. Same applies for "Model", "U-Net". Figures are presented as "Figure X" or "X" without the "Figure". In Line 358, "Figure" is not written at all.

We capitalized Figure X in all instances, we added a Figure in line 358, we capitalized the word model in all instances when we refer to Model X.

2. HR-GNSS sometimes referred as GNSS as in Line 224. It would be better to stick with the same terminology.

GNSS is different than HR-GNSS. We do use HR-GNSS when that is the type of data we are referring to but we also use GNSS generally if it's more appropriate.

3. In section 2.4, training and testing dataset are divided with 90% to 10% proportions. Is there any specific reason to use this proportions? Zhu et al. 2019 also used similar proportion without justification. Did you follow the similar proportion by following previous studies or do you have any other selection criteria?

This is a design choice; both 80/20 and 90/10 are common.

4. Two name articles are sometime cited as "A and B" and other times "A & B".

We have changed all &s to ands.

My line by line minor comments are in below:
Line 39 – instead of >= 1 samples per second, it can be written as sampling rate of >= 1Hz. Done.
Line 43 – there can be a citation for the long period recordings. Added Melgar et al. 2015.
Line 50 – citation to a previous study that used HR-GNSS would be good. Done.
Line 55 – Thomas et al. 2016 is not cited. Corrected.
Line 70 – He et al. 2015 is not in the references. Corrected
Line 104 – instead of "we need many thousand of samples of both …" "we need extensive amount of both …" can be used. Extensive is not specific – this could mean

10s or hundreds. We will opt to keep the thousands.
Line 141 – SCEC should be written “Southern California Earthquake Center” in the first time it is mentioned. Corrected.
Line 143 – (GFs) and citation to Zhu & Rivera can be merged into single parenthesis. Corrected.
Line 145 – Trained model is mentioned before providing any information about what ‘training’ is. We changed this to “After comparing our kinematic rupture waveforms with HR-GNSS data from the Ridgecrest earthquakes we noticed that many waveforms from the Ridgecrest earthquakes had significant “ringing” or long duration coda that was not present in the synthetic waveforms.”
Line 151 – An example of these signal can be given in supplementary material. Done.
Line 182 – Citation to ReLu would be good (http://citebay.com/how-to-cite/relu/). Done.
Line 237 – Explicit information about the time and frequency resolution of the data would help reader to understand the input data more clearly. We stated the sample rate and duration, the parameters for computing the STFT.
Line 309 – CA is not defined (even though it is obviously refers to California, it would be better to write it). Changed to California
Line 314 - Melgar et al., 2013 is not cited. Done.
Line 373 – “Despite this …” can be re-written like “Despite low SNR …”. Done.
Line 381 – It’s hard to say which example is “this example” since in the text several figures are referred. Changed to “in this same example (Figure 6C),”
Line 414 – black line doesn’t refer anything. In the following sentence Figure 12 is mentioned and the black line is there but before that sentence black line doesn’t tell anything. We’ve removed this reference to the black line.
Line 467 – “below” is unnecessary. Removed.
Line 476 – There is no Figure S3 in the supplementary material. Figure S1 is also not defined in the text. We have heavily modified the supplement and these numbering issues are now fixed.
Line 612 – Last access date for the github link may be necessary. Zenodo archives a snapshot of the repo so a last access date is not necessary and is FAIR compliant.
Line 648 – Lay 2018 is cited as Lay et al. (2018) in Line 53. Corrected.
Line 706 – Wdowinski is misspelled as Widowinski in line 69. Corrected.
Line 720 – Zhu and Beroza 2019b is defined as 2019. It also needs to be defined as Zhu et al. 2019. Corrected.
Figure 1 – Panels are not really separate figures both panel A is partially on top of panel B. Hence, I think Figure 1 can be referred without specifying as A and B in the text. Moreover, another scale would help to understand magnitudes of the events more clearly. Even though events are scaled with magnitude, they all look same.
We will opt to keep the figure as is since there is value in showing the regional scale from which the noise data was taken and the local scale on which the GNSS waveforms were computed. We don’t feel there is a need to add another panel to highlight the different event magnitudes since only the M6 and M7 events are used in this study – the goal was to show the fault geometries which can be determined from B.
Figure 12 – It is hard to understand percentiles for different model dataset. What is light pink and dark pink (or purple) stand for? We’ve added “and have the same color coding as the median CC”

Figure 14 – Legend can be rearranged to see the waveform more clearly. Done.
Figure 16 – Label of X axes can be “before and after the origin time”. Negative seconds after the origin time is seconds before the origin time. We don’t see this as confusing so opt to leave it as is.