

Interactive comment on “Random Noise Attenuation of Sparker Seismic Oceanography Data with Machine Learning” by Hyunggu Jun et al.

Anonymous Referee #1

Received and published: 22 April 2020

This study applies machine learning method to extract the signal from noisy seismic signal from the water column collected using sparker source. This is an alternative means of collecting data (to air guns) offering lower costs despite larger noise in the signal. Authors offer a method that can successfully clean the data and help gain insight on physical oceanography. For the method, the training data are important, and the authors present cases using synthetic and field data.

Overall the study is technical, but can be of interest to the Ocean Science readers. Authors demonstrate convincingly that the random noise attenuation with machine learning satisfactorily cleans the data and reveals the oceanographic structure in the seismic

C1

signal.

I have a list of detailed comments below. These are mainly clarifications and suggestion for better presentation. The presentation of the material in figures is poor and needs improvement. The analysis is solid and does not need any revisions in my opinion (non-expert on this technique). However, I am surprised the authors do not give a more thorough analysis of the error and skill (e.g., by histograms using all 3072 data patches instead of only 6).

The context in an oceanographic application comes finally in the last figure and the last page of the text (line 364 and on). I would strongly encourage the authors to improve the “Ocean Science” part by adding more insight using the clean seismic sections. Can you make some oceanographic inferences, interpretations (or better quantification) using the cleaned Lines 1 and 2?

Finally, it would serve the community much better if the authors made available some code for noise attenuation using machine learning. They offer the code through communication with authors, but the impact would be far larger if they make the code available as a supplement.

Detailed comments:

Li 6-7. The opening sentence of the abstract is a bit confusing. SO exploits water column reflections to interpret the oceanic features (fronts, eddies, water mass boundaries) as well as ocean fine structure (internal waves etc.). Furthermore, “compensating for the drawbacks of conventional PO equipment” is a very strong (and erroneous) statement. Perhaps “supplements the conventional PO observations”. We should also be very careful when interpreting the seismic images to describe PO quantitatively.

Li 8. The low / high frequency band introduction is not very helpful unless you relate it to spatial resolution.

Li 10. Reword “To solve the problem”? For example, “To extract reliable signal from the

C2

low S/N . . .”

Li 23: “measurements [from cruises] are performed. . . . observation [stations].”

Li 27: mention how the sea water characteristics can be estimated (through the acoustic impedance contrasts and expand a bit more to inform the reader)

Li 29-32: Fine, but please do not oversell. Perhaps mention “qualitative images” and then move to “quantitative information after careful analysis where temperature/salinity contrasts produce well-defined horizons of seismic reflections” or similar. Also SO is not “widely used”.

Li 32: reword “determine the behavior of turbulence and internal waves” to, for example, “quantify the internal wave spectral distribution and infer turbulence”

Li 33: clarify what central frequency is (since the source covers a range of frequencies)

Li 43: vertical resolution of 1.5 m is not much superior to the vertical resolution of “several meters” stated in line 34. Perhaps specify the latter as 5-10 m?

Li 67-68: If not using MLP and AE (and any other acronym), no need to introduce them. It is difficult to read the text.

Li 64-75: If there’s a possibility to thin out various methods introduced (and refer to a few key references and citations therein), it can be easier for the reader to follow.

Li 77: East Sea appears very abruptly here, out of context.

Li 135: delete “On the other hand,”?

Sec 2.1 and 2.2: can any of these descriptions refer to Fig 1? (I only see a reference in the end, at li 141, and it is not very instructive.)

Li 147: This is actually one line, but two repeats (in different travelling directions). Please mention the date of data collection, vessel speed during data collection. Transect duration etc.

C3

Li 163-164: there’re CTD /XCTD profiles, but the authors shown only 2xtemperature profiles from XBTs. It would be nice to increase the oceanographic context in the paper.

Li 167: please describe what a reflection coefficient is.

Li 184-185: what do you mean by “thus, the subsurface seismic data have a better S/N ratio than the SO data.”? Is subsurface seismic data not SO data? I suspect you mean beneath seabed by subsurface. Please clarify.

Li 187: It is confusing: “We used the interval from 0.2 to 0.6 s of the original data where the noise level is relatively low”. Earlier you mentioned that part was just noise!

Li 190: Reword “the data are field data recorded with the same equipment.” as “the data are collected by the same equipment”

Li 204: what is g/cc? Please use SI units.

Li 249: bottom right (instead of right bottom)

Li 249-250: The sentence is confusing: “. . .using training dataset 1 has one problem. The ground truth of test data 5 contains noise in the right bottom part, and training dataset 1 also contains noise in some parts of the ground truth”. Dataset 1 has 6 test data. With the last reference to dataset 1 do you mean test data 1 or the entire dataset 1? Perhaps cut out the entire last part after the comma. Overall, I would appreciate a more distinct wording for test data. For example subset 1 to 6, or patch (you use it in line 280)?

Li 279-280: 20th and 30th traces from the last patch: which epoch is this? Are the traces from the 50x50 patch? Can you please mention for the reader: “. . .traces out of the 50x50 size patch 6 of the test data”.

Li 310-311: can be cut out; simply cross reference Fig 13 after 25 epochs. Overall there are repetitions throughout the authors could try to simplify.

Li 325: perhaps specify, “is the number of test data patch (3072)”

C4

Li 327-328: too many significant digits at RMS errors? (perhaps enough with 6.37 and 6.34). For which epoch are these values? (Also the normalized values in line 331 could be 0.27 and 0.15)

Li 332: delete "than that of the D1 model"

Eq 1, is a division by nmode missing?

Li 364: "The data slope spectrum is the slope spectrum..." this is all very confusing. The data slope spectrum is first referred to in line 276-277 (again without explanation). Please introduce what the data slope is. For example, "the slope spectrum is the horizontal wavenumber, k_x , spectrum of the horizontal gradient of the vertical displacement of a digitized horizon. The data slope spectrum is ...?" (or a similar explanation. Note my interpretation of the slope spectrum can be in error.)

Li 367: replace "we calculated the data slope spectrum ... and compared the data slope spectra" with "we calculated and compared the data slope spectra using the outcome of the D1 and D2 models..."

Li 376: "slope" is missing before "at wavenumbers"

Li 377-378: I cannot quite follow the subranges and the mentioned slopes in this panel. Perhaps mark on the figure?

Li 390: Here again mention why sparker SO data may be desirable

Fig 1. Please offer some more explanation in the caption. If not possible, defer reader to the main text.

Fig 2. Elevation is grayed out for $>0\text{m}$, so the colorbar can stop at 0. It would be useful to add a few isobaths. I would call Line 1 and Line 2, Repeat 1 and Repeat 2.

Figs 4 and 5 can be combined into 1 figure. I suggest two panels, T profiles in one panel with different color. Reflection coeff in the second panel with different colors and one offset by 0.0001 unit. Does the coefficient have a unit?

C5

Fig 6 can be removed. It is already shown in Fig 3 and with the statement $\text{time} > 0.28$ s. You can mark the region by a rectangle in Fig 3.

Fig 10 (and Fig 13)- this is the average PSNR and SSIM for the 6 subsets of dataset 1? Would it not be better to show all 6 lines, or the average with one standard deviation? Actually the number of test data is 3072 (6 is an arbitrary pick), why not show the mean and std over all 3072? And also show a histogram?

Fig 11 caption: in the end the cross reference should be to Fig 9.

Fig 14 caption: in the end the cross reference should be to Fig 12. In the text the model is referred to as D2 (but here D1).

Fig 16. This figure is not needed either. It is simply the upper 0.28s of Fig 3. However, I appreciate that it is zoomed in and compared to the cleaned sections. See suggestion below Fig 17 comment.

Fig 17. Please consider removing xaxis labels from panels c to d, and placing panel labels in the upper left corner of panels (it's grayed out anyway), so that you can have a more condensed 4-panel, 1 page figure with minimum white space vertically between panels.

I think a reorganized version of Figs 16- 18 will be much better for the reader. I suggest 2 figures, each with 5 panels (with identical x-axis limits and width, and minimum white space between them). New Fig 16: Line 1 results. New Fig 17 Line 2 results with corresponding 5 panels for each figure:

- a) data (as in Fig 16a)
- b) clean after D1
- c) clean after D2
- d) noise after D1

C6

e) noise after D2.

It is most impactful, if you can fit all 5panels in one page please.

Interactive comment on Ocean Sci. Discuss., <https://doi.org/10.5194/os-2020-13>, 2020.