Author Response to Reviews of

SeisMIC - an Open Source Python Toolset to Compute Velocity Changes from Ambient Seismic Noise

Peter Makus, Christoph Sens-Schönfelder
Seismica, doi:-

RC: ***Reviewer Comment***, AR: *Author Response*, ☐ Manuscript text

Dear Dr. Koelemeijer,

Thank you for giving us the opportunity to respond to the reviewers' comments. We thank them for their comments and implemented the requested changes in the revised version of the manuscript. Below, we list the reviewers' comments and our respective responses. We also show the modifications we applied to the text. With regard to your request to discuss data representation standards, we added a minor subsection 2.2.1 to the main manuscript. In addition, we made some minor language corrections.

Please find a a marked-up and a cleaned version of the revised manuscript attached to this response.

Thank you for your efforts in handling our manuscript, and we hope that you can now accept our manuscript.

Sincerely yours,

Peter Makus and Christoph Sens-Schönfelder

1. Reviewer #1: Laura Ermert

1.1. Comments on the Code and Tutorial

RC: ***Tests ran fine (output attached below)***

AR: *Excellent.*

RC: ***jupyter is not included in the environment.yml file; it may be worth to make a brief note somewhere to encourage users to install it.***

AR: *We thank the reviewer for the comment. This is indeed a good point. We added this information to (1) the introduction of both jupyter notebooks, (2) SeisMIC's README, and (3) SeisMIC's documentation. We do not make jupyter a dependency in order to leave the package as "light-weight" as possible since it is strictly not required to use the code.*

RC: ***the link at https://petermakus.github.io/SeisMIC/modules/modules/tutorials does not work***

AR: *We could not find any occurrence of this link with an automated search through the manuscripts, the code, the documentation, and the tutorials. So we do not know where the reviewer found this link.*
The correct link would be `https://petermakus.github.io/SeisMIC/modules/tutorials`*, which works as expected.*

RC: ***I installed the conda environment specified by the environment.yml and jupyter separately through conda.***

In this setup, I could not launch mpi processes through the jupyter notebooks and had to return to the command line (where it worked fine). This is not a real issue, but it may be worth to point out to the users that they could directly launch the mpi processes from command line if needed.

AR: *We thank the reviewer for pointing this out to us. We added this information to the appropriate fields in the jupyter notebook.*

RC: ***It may also be worth to introduce a "channel" variable in the tutorial alongside "station" and "network", so that users have more flexibility to adapt the example.***

AR: *Currently, the code and tutorial allow to limit the download to specific channels. For the workflow, correlations are then computed for the specified combination method (i.e., autocorrelations, self-correlations, or cross-correlations) for all available channels.*
However, we agree that such a feature would be useful and will be available in future versions. As common procedure, it will first be implemented into the developer's version before arriving on the main branch to ensure the code's stability.

RC: ***Other than that, the tutorial ran very well and within a short time and with quite minimal effort produced a dv/v curve showing velocity drops in Mexico City coincident with earthquakes in Chiapas and Puebla (2017)***

AR: *We greatly appreciate the effort of the reviewer. We are excited to see that it was easy to reproduce the results mentioned above.*

1.2. Comments on the text

RC: ***Lines 42 ff:***

> ***Still today, many scientists use simple "home-grown" scripts to produce results for later publication and interpretation. Often, these scripts lack the required efficiency and make reproducing and adapting analyses cumbersome since there are no agreed-upon standardisations.***

I am one of the scientist presumably using a "simple home-grown collection of scripts" (which has been relying on mpi4py and hdf5 for a long time) and reading this sentence did not put me in the mildest mood towards the rest of the manuscript. Formulating things in this way may be detrimental to getting groups with already-established workflows on board. What does home-grown refer to? It should be mentioned explicitly what sets the presented code apart from many other science codes that have been developing over the years in the community (this questions is to some degree answered in the next section). Since the performance comparison shown later includes MSNoise and SeisMIC, but no other codes, the statement that other groups' collections of scripts lack efficiency seems based on hearsay. I believe that there is a diverse collection of codes in the international community that all have their particular strengths,specialties and weaknesses (and I believe that some of them are very efficient). At the moment, which code users rely on seems to depend mostly on the institute where they start their career in ambient noise studies. I think a community benchmarking exercise would be needed to come to the above conclusion and to try and establish a clearer picture of all the tools out there. There are indeed no agreed-upon standardisations. However, the present manuscript does not explicitly propose any as far as I can see (e.g. a new data format standard for correlation data). So, the only way the manuscript could improve the state of affairs is by becoming the standard. Maybe this point should be taken up in the discussion section, and the authors might propose to adopt the SeisMIC standard or to initiate a community discussion on such standards

(especially on data formats).

AR: *We admit that our phrasing could lead an unintended irritation of some readers. So we are grateful for the reviewer's remark. Our use of* home-grown *was intended to describe unpublished (according to FAIR principles, doi:s41597-022-01710-x) code. Our implicit assumption was that such codes are often not publicly accessible, hard-to-read and using hard-to-adapt scripts which slows down our progress as a community. But of course there are exceptions to these assumptions and such codes could be published in a FAIR way to foster exchange, bench marking and reproducible research.*
To avoid misunderstandings, we have replaced the two sentences in question by:

> *Still today, many authors use unpublished codes to produce results for later publication and interpretation making it difficult for fellow researchers to reproduce or adapt the analyses. Using community codes published in the spirit of the FAIR principles (Barker et al., 2022) can facilitate the reproducibility of research, exchange in the community, and progress in science.*

Note that we also modified some of the introductory part giving a more accurate oversight of how SeisMIC compares to MSnoise and NoisePy following a comment by reviewer #2.

RC: ***Along similar lines, Line 126***

> ***To our knowledge, SeisMIC is currently the only software that supports spatial inversion of velocity change time series.***

Since there are a number publications showing this type of inversion, are you referring to publicly available codes?

AR: *Yes, here, we are referring to publicly available codes. For clarification, we adapted the text as follows:*

> *To our knowledge, SeisMIC is currently the only publicly available software that supports spatial inversion of velocity change time series*

RC: ***In the performance comparison, instrument response removal is not included, and it can be one of the surprisingly cumbersome steps. Does including it affect the comparison with MSNoise?***

AR: *We agree with the reviewer that the response removal can be a surprisingly expensive processing step, which is mainly due to long time series that need correction. Both MSNoise and SeisMIC use the obspy implementation for the response removal and do consequently need the same amount of time to execute the response removal. Speaking in absolute terms, the comparison would thus not differ. We added this information to the manuscript:*

> *We do not remove the instrument response. Note, however, that both MSNoise and SeisMIC perform the response removal using ObsPy (Beyreuther et al., 2010) and will therefore take the same amount of compute time and resources.*

RC: ***Line 93 "imaginary number" → imaginary unit?***

AR: *We changed the formulation to:*

> *...where* $i = \sqrt{-1}$ ~~*is the imaginary number*~~.

RC: ***Line 138 "message parsing interface" → message passing interface***

AR: *We modified the text accordingly.*

RC: ***Line 202***

> ***For conciseness, we restrict this example to 11 days of data from 25 January to 5 February 2016.***

It would strengthen the case for presenting this code as highly efficient and suitable for HPC if a test on big data was run; e.g. Clements & Denolle tested Seisnoise.jl on over 1.6 million cross-correlation pairs. I am not suggesting to run such a test because it would cost significant effort; but if larger tests are already available from previous studies, why not include some key figures on performance on a larger dataset?

AR: *We agree that such tests and figures are highly relevant. However, the line quoted by the reviewer refers to the exemplary result. We believe that the reader benefits from a simple and easy-to-interpret example, which is used to illustrate the functionalities and methods available in SeisMIC. For the benchmarking section we employed a significantly larger dataset. For the comparison with MSNoise, we compute more than 16,000 cross-correlation pairs. We added the following to the cited section to avoid ambiguities:*

> *For conciseness, we restrict this example to 11 days of data from 25 January to 5 February 2016. In section 2.3, we show how SeisMIC performs when confronted to much larger datasets.*

RC: ***Figure 5, the component (pair) should be specified. This figure, which can be reproduced in the tutorial, shows an autocorrelation. However, the trace is not symmetric with respect to lag time 0. This suggests that the code uses causal filters in the background. Users should be made aware of this to avoid any possible problems caused by the asymmetry with respect to zero lag. How is this handled, e.g., during windowing the correlation waveforms?***

AR: *We thank the reviewer for bringing this to our attention. The reason that the correlation function was asymmetric is that we indeed used a causal filter in the tutorial notebook. We changed this line in the tutorial so that a zero-phase filter is used. We replaced Figure 5 accordingly. We modified the figure caption to indicate that the east-component was used to create the figure:*

> *Hourly autocorrelations of ambient noise recorded ~~at X9.IR1~~ by the east component of X9.IR1.*

RC: ***More information in the implementation; This paper introduces the software, so I think it needs to provide short (not necessarily technical) descriptions on the parallelization strategy and memory management or whether it includes any knobs that the users can turn to influence memory management.***

AR: *We agree with the reviewer that such information is useful and have therefore added the following to section 2.3:*

In SeisMIC, the computationally most expensive parts of the workflow described in section 2.2 are the calculation of correlation functions, the associated preprocessing and the estimation of the final velocity change time series. Therefore, an effective parallelisation scheme matters the most in these steps. For users, it is also important to understand how memory requirements scale. For the computation of CFs and the preprocessing of raw data, each core reads different raw data in chunks of equal length (see Listing 3 for details). Subsequently, the same core performs the preprocessing. For the cross-correlation operation, each core is responsible for a different component combination. This implementation makes the RAM usage practically independent of the number of cores used. Thus, RAM usage will mainly depend on the length of the raw data chunks read in each step (i.e., a smaller read length will lead to lower memory usage) and its sampling rate (i.e., a lower sampling rate will lead to lower memory usage). Resulting CFs are written to h5 files immediately after correlation or stacking and the memory is freed. In contrast, SeisMIC computes the final dv/v estimate with "1-core per component combination". Here, a single core loads all available CFs for one component combination and executes the stretching algorithm and the associated processing. Therefore, for the final dv/v calculation, the memory requirement scales with the number of employed cores.

2. Reviewer #2

2.1. Import Issues

RC: ***After following the guidelines on GitHub for the "Installation of SeisMIC from Source Code" some errors occurred during the execution of the provided Jupyter notebooks. First of all, executing the first cell in the Jupyter notebook tutorial.ipynb resulted in:***

```
ModuleNotFoundError: No module named 'seismic'.
```

AR: *It is difficult to provide debug instructions in the given frame. The provided error claims that SeisMIC has not been installed. This can either mean that the installation step with the command*

```
pip install -e .
```

has been omitted before the installation or that, in a conda installation, the wrong environment was active.

RC: ***After fixing this error, the next problem occurred in section 2.2 Start correlation. Executing the cell***

```
from seismic.db.corr_hdf5 import CorrelationDataBase

with CorrelationDataBase(f'data/corr/{network}-{network}.{station}-{station}.HHE-HHE.h5

# find the available labels

print(list(cdb.keys()))
```

resulted in:

```
from seismic.db.corr_hdf5 import CorrelationDataBase
AttributeError: module 'numpy' has no attribute 'typeDict'.
```

AR: *Numpy typeDicts have been deprecated with numpy version 1.21 (see numpy changelog). This error occurs, when an older version of h5py is combined with a newer numpy version. To address this issue, we now set a hard version-requirement for h5py 3.9.0.*

RC: ***In the provided Jupyter notebook spatial.ipynb an error occurred already in the first cell***

from seismic.monitor.spatial import DVGrid

TypeError: unsupported operand type(s) for |: 'type' and 'type'

It seems like there´s a problem with some dependencies in the provided environment.yml file. This should be carefully checked and possibly fixed.

AR: *The "|" for 'type' was implemented with python 3.10, which is one of the reasons why SeisMIC requires python 3.10 or 3.11. This requirement is defined in all relevant installation files (i.e., setup.cfg, environment.yaml, and requirements.txt). After following the installation instructions as given in SeisMIC's documentation, these requirements should automatically be fulfilled. From this errors and the described import issues it seems like one or several steps in the installation instructions might have been omitted. We could not reproduce this error on any of the machines used for testing.*

2.2. Minor Comments

RC: ***Page 2 line 47:***

> ***However, as we will show and discuss here, the existing software still leaves a niche to fill.***

In the manuscript some differences to MSNoise and NoisePy are mentioned, such as computation times and preprocessing options. However, the manuscript might benefit from a more extensive comparison between the three softwares, including advantages/disadvantages of each software, functionalities,… I would recommend adding a few sentences concerning this into the Introduction.

AR: *We agree that such a comparison would certainly be useful. However, a detailed comparison poses several challenges. Many of the reasons that researchers prefer a certain code are due to small details and subtleties in features, performance, syntax, documentation, available support, and countless others. Thus, they are very difficult to capture even in a detailed comparison and require an extensive amount of testing as many of these details are not or cannot be documented and would justify a separate publication, as for example done for earthquake phase picking by Münchmeyer et al. (2022). Often, preferences have to do with subjective perceptions, most of the time resulting from knowledge of prior codes, which is why we point out that SeisMIC has a similar syntax to ObsPy. Moreover, software is under constant development and such a comparison would be outdated after only a short time.*
In its current state, we believe that our manuscript outlines the most important differences in features (e.g., SeisMIC offers a spatial inversion algorithm, but currently only the stretching method for dv/v estimation) and performance (in the benchmark section). In addition, we provide insights on the syntax of SeisMIC to emphasise the code's simpleness.
Nevertheless, we agree with the reviewer that there should be an additional statements in the introduction. Here, we choose to formulate this statement focusing on the intended purpose and speciality of each of the three mentioned solutions:

> *However, as we will show and discuss here, the existing software still leaves a niche to fill. For example, MSNoise is more specialised for end-to-end workflows and automated monitoring solutions, lending it more towards applications in observatories, whereas, recently, NoisePy has undergone development towards cloud computing. To fill the remaining gap, we introduce SeisMIC (Seismological Monitoring using Interferometric Concepts) (Makus & Sens-Schönfelder, 2022), a fast, robust, flexible, and easily adapted Python tool to compute, process, and analyse dv/v. Due to these attributes, SeisMIC especially excels at the analysis of campaign data, where both ease of use and flexibility are crucial.*

RC: ***Page 2 line 65: possibly missing word in sentence "…we follow the FAIR principles after Hong et al. (2022)…"***

AR: *We modified the sentence so that it reads:*

> *...we follow the FAIR principles (Hong et al., ~~(~~2022~~)~~)*

RC: ***Listing 3 line 254: Specify what the parameter corr_inc stands for.***

AR: *We thank the reviewer for pointing this out. We modified the manuscript to:*

> *Most fundamentally, we must set the correlation length, corr_len, (i.e., the duration of the time windows to be correlated), the increment between these time windows, corr_inc,...*

RC: ***Page 11 line 295: missing word. … or might be due to lunar …***

AR: *We implemented the change as suggested.*

RC: ***Figure 8: the font size of the axis labels (Northing, Easting, dv/v) should be increased.***

AR: *We changed the font sizes as suggested.*

RC: ***Supplement Figure 1 c): the numbers in the color bar are overlapping.***

AR: *We fixed the format of the colourbars.*

3. Reviewer #3

3.1. Comments on Content and Figures (Main Text)

RC: ***In Section 2.2: Please explain for each equation what each letter/variable means: e.g., letters m, n, k, n, in equation 1, even if you assume that people know it intuitively. Letter o in equation 2 is explained in line 106, which is too late, for the reader it would be easier to follow if the letters were explained in the text right after the equations.***

AR: *We thank the reviewer for the comment and implemented the following changes:*
l. 89:

> *Suppose we want to calculate all available correlations from a dataset of M waveforms, of which each has N samples (indices m and n, respectively).*

l. 93:

> *where $i = \sqrt{-1}$ and k is the sample index of the signal in the frequency domain.*

l. 97:

> *where the bar indicates the complex conjugate and o indexes the station pair.*

RC: ***Velocity changes/coherence estimates are jointly inverted/calculated for causal and acausal sides of cross-correlations. Is there the option look at causal and acausal parts separately, for example if the signal-to-noise ratio is much better on one side of the cross-correlations? Addressing this briefly in 1 or 2 sentences would be helpful.***

AR: *This feature is indeed implemented in SeisMIC and we added the following sentence to lines 123-125:*

> *In SeisMIC, dv/v can either be jointly inverted from causal (right) and acausal (left) side or estimated from either side, which might be desirable for active source experiments or if one side of the CF exhibits a superior signal-to-noise-ratio.*

RC: ***Section 3.5: is there a bias in the results caused by the station distribution for the spatially resolved velocity changes and if yes, does SeisMIC address this? (like the checkerboard test in tomography studies).***

AR: *We hope we do not misunderstand this question. In our understanding, the synthetic test presented in the manuscript does function exactly like a checkerboard test. The user would only have to alter the station arrangement to the actual configuration of a given dataset.*

RC: ***Figure 4: Please add the date of the earthquake you are referring to in the caption as well and not only in the main text (or indicate it otherwise in the figure).***

AR: *We added the following clause to the figure caption:*

> *...earthquake, which occurred on 28 January 2016,...*

RC: ***Figure 5: I believe the time axis is in Figure 5a) is slightly off. Like this the autocorrelations are not symmetric around 0 lag time. If this is a specific plotting style it would be good to indicate it in the figure caption. Also, why are amplitudes of autocorrelations diffrent on causal and acausal sides? The label for the colorbar 'correlation coefficient' is confusing, maybe put simply 'Amplitude' since you speak about correlation coefficients in a different context in the paper.***

AR: *We have addressed this issue according to the comment by reviewer #1. It was fixed by replacing the bandpass filter in the tutorial by a zero-phase filter. As for the colourbar label, we tend to disagree since mathematically we are indeed showing correlation coefficients, whereas 'amplitude' is a term that we would associate with raw seismograms.*

RC: ***Figure 8: Please explain in the text/caption what you mean with 'Correlation length' in this case. Previously you define it as the duration of the time window to be correlated (line 226).***

AR: *We agree, this should have been explained in the the text body. Therefore, we added the ll. 344, 345:*

> *This inversion relies on two damping parameters, the correlation length λ determining how strongly related neighbouring grid cells are and the model variance σ_m, that the model may assume.*

3.2. Suggestions on Presentation/Language (Main Text)

RC: ***L21 and L40: make -> makes***

AR: *We implemented the suggested change for l. 40. For l. 21 however, using the plural would be grammatically correct as there are two subjects.*

RC: ***L54 – 56: unclear formulation starting at 'or down to…'. Maybe rewrite sentence slightly?***

AR: *We thank the reviewer for the comment and changed the sentence as follows:*

> *As opposed to working with a black box, users work close to the source code, making it easy to develop individualised workflows. M~~or even use m~~odules, submodules, or even single objects and functions ~~separately.~~ of the code can also be used individually.*

RC: ***L134 – 137: long sentence and comma missing between (HPC) and compatible?***

AR: *We split the sentence in two to make it a little easier to read:*

> *We address the arising computational and storage challenges with efficient and high-performance computing (HPC) compatible code design. ~~enabling~~SeisMIC enables parallel computing of correlations, velocity change estimates and spatial inversions, where the computation of CFs is the most expensive operation by a large margin.*

RC: ***L 179 - 180: This sentence is a bit unclear – in particular I don't understand what 'more evenly writing operations' and the 'slightly improved scaling' (of what?) means***

AR: *To clarify this sentence, we added some more specific descriptions:*

> *MSNoise creates one miniseed file per CF, resulting in less complex ~~and more evenly distributed~~ writing operations, which are more evenly distributed across the cores. For this benchmark, this translates to a slightly better scaling between the number of cores and the computational time but also in a high number of files, which can be undesirable for large datasets*

RC: ***L301 – 303: Sentence is a bit complicated, I had to read it twice to understand it. Shortening it would help.***

AR: *We modified the sentence as follows:*

> ~~*The spatially extended sampling of coda waves increases the sensitivity to distributed weak velocity changes and the detectability of localised changes but prevents a simple inference of the affected location along a ray path or Fresnel volume.*~~ *Coda waves, as used in PII, sample the medium at a high spatial extent. While this allows to detect distributed weak velocity changes or changes located away from the path of direct waves, it prevents a simple inference of the affected location along a ray path or Fresnel volume.*

3.3. Supporting Information

RC: ***L33: Please discuss a bit more in detail why your approach is better for averaging, smoothing, etc. An example would help.***

AR: *For clarity, we modified the text as follows:*

> *This definition is especially useful for smoothing, averaging, or otherwise manipulating velocity change time series as, in contrast to the common definition, it ensures linearity (see below).*

RC: ***Equations: like in the main text, please make sure that you explain each variable of the equations in the text.***

AR: *After re-reading the text several times, we could not find an undefined variable. I hope we are not missing the point here.*

RC: ***Figures 1-5) The ticks for the colorbars need to be fixed.***

AR: *We fixed the colourbar for Figure S1 (see comment by reviewer #2 above). For all other figures, they seemed to be in order.*