Response to reviewers for "False positives are common in single-station template matching"

Jack B. Muir * [1], Benjamin Fernando [2], and Elizabeth Barret[3]

[1]Department of Earth Sciences, University of Oxford, Oxford, United Kingdom
[2]Department of Physics, University of Oxford, Oxford, United Kingdom
[3]Jet Propulsion Laboratory, National Aeronautics and Space Administration, United States of America

July 13, 2023

1 Editor

Dear Jack Muir, Benjamin Fernando:

I hope this email finds you well. I have reached a decision regarding your submission to Seismica, "False positives are common in single-station template matching". Thank you once again for submitting your work to Seismica.

Based on reviews I have received, your manuscript may be suitable for publication after some revisions.

When you are ready to resubmit the revised version of your manuscript, please upload:

- A "cleaned" version of the revised manuscript, without any markup/changes highlighted.
- A pdf version of the revised manuscript clearly highlighting changes/markup/edits.
- A "response-to-reviewers" letter that shows your response to each of the reviewers' points, together with a summary of the resulting changes made to the manuscript.

Once I have read your revised manuscript and rebuttal, I will then decide whether the manuscript either needs to be sent to reviewers again, requires further minor changes, or can be accepted.

*Corresponding author: jack.muir@earth.ox.ac.uk

If you deem it appropriate, please check that the revised version of your manuscript recognises the work of the reviewers in the Acknowledgements section.

Please note that Seismica does not have any strict deadlines for submitting revisions, but naturally, it is likely to be in your best interest to submit these fairly promptly, and please let me know of any expected delays.

I wish you the best with working on the revisions. Please don't hesitate to contact me with any questions or comments about your submission, or if you have any feedback about your experience with Seismica.

Kind regards,

Yen Joe Tan

Dear Prof. Tan,

Thank you for handling the editorial process for this paper. Please find our responses to reviewer comments in blue beneath each point. In addition to the direct responses to reviewer comments, we have also supplemented the paper with additional discussion from Elizabeth Barrett, who is now a coauthor. We also ultimately decided that having 100 random iterations of the experiment for each configuration was unnecessary for providing adequate statistics, and have reduced the provided data to 32 iterations to save download replication time.

Best regards,

Jack Muir, Benjamin Fernando & Elizabeth Barrett

2 Reviewer A

This MS investigates the rate of false detections for single-station template matching. Authors conclude that false detections are common in template matching based on a simulation of teleseismic application. However, the MS reads more like a report on an issue that everyone knows, and I cannot see any new findings. Thus, at this stage, I think such a paper is not valuable to be published in Seismica. More comments as below:

Thank you for the review comments below. We agree that the rate of false detections is an issue that should be well understood by practitioners of detection by template cross-correlation. We were motivated to write this paper because a) the move to planetary seismology contexts has in some sense regressed the state-of-the-art in seismic detection due to the need to rely on single-station methods, which result in poorer rejection of false events compared to multi-station methods, and b) we are not aware of a detailed simulation based study of the baseline false detection rate for classical single-station methods, and past studies seem to provide overly optimistic estimates of the false-positive rate based on back-of-the-envelope calculations. We hope that this paper serves as a useful resource for cross-correlation practitioners to think about false-positives in there work.

- The MS seems to give a general conclusion for template matching, while it only simulates a teleseismic application with a very low frequency range

(0.1–1.6). In fact, template matching is more frequently used in local seismic detection with a relatively high frequency range (e.g., 1-10). We recognize that Earth-based applications typically use higher frequency ranges (and normally shorter windows, unless searching for exact repeaters); the emphasis of our paper is discussing planetary seismology applications however, for which examples have been more in the teleseismic regime. As we now note in the paper, due to the underlying raw data being white noise, the parameter regimes can be easily rescaled, with the 0.1–1.6 Hz case with a 20 s window being equivalent to a 1–16 Hz case with a 2 s window (with the time taken to reach a particular CC level also being 10 times shorter). Presenting results in terms of windows in samples and filter bands in normalized frequency would be perhaps the most correct, but is quite non-intuitive which motivates our use of physical units in the manuscript.

- Not surprisingly, the low frequency waveforms are simple and thus can result in a high similarity in template matching. Cross-correlation between noise could result in a high time of MAD, what if the template waveform includes a true event? Author may need to add more tests to support your conclusion, especially for a real application of local seismic detection. Thank you for pointing out that some of the motivation of this study is unclear. To the first point, as we mentioned above the frequency range presented in this study is "low", but the window lengths are comparitively "long" for applications other than searching for true repeating earthquakes. We have now made this distinction clearer in the introduction. As mentioned above, we can rescale results as the raw data is white noise and thus invariant under appropriate transformations of the frequency and time scales, and we see that the rescaled ranges are more comparable to local detection contexts. In regards to the addition of template waveforms including a true event; we would presumably expect a higher rate of detections of both the signals and templates include events; the objective of this study is to obtain a *lower bound* on the rate of false detections under typical processing choices, which is why we use pure noise as the raw data, as it contains only false detections.

- Line 97: Please add references for the statement "c=7 is a typical choice" We have changed this line to $c \sim 7$ as being more accurate, and referenced Sun and Tkalčić as a relevant recent example.

- Table 1: the third fmax should be 1.6. Thank you for pointing this out, it has been corrected.

- Lines 138-141: Some common noise can result in false detections indeed, this is also a common issue and can be solved by adopting a multiple-segment cross-correlation strategy (Gao et al., 2020, JGR).Thank you for pointing out this recent study, which contains a useful discussion of methodology used to improve template matching results. We have now

included it in the manuscript — however we note that Gao et al. itself recognizes a tradeoff between their method and traditional CC methods in high-noise situations, and furthermore recognize that the majority of users still imploy these traditional methods. As such, we feel a detailed study of the false-positive rate for traditional processing is warranted.

- Figures 1 and 2: Please use a-i to denote each panel instead of "top left" or "bottom left", it is easy to get confused. E.g., "bottom left" in line 119 does not indicate the case with a longer window and wider passband. Thank you for pointing out this error, it has been corrected and explicitly labeled subfigures have been added and referenced in the text.

3 Reviewer B

3.1 Summary

In this short didactic study, the authors demonstrate that single-station, narrow frequency band tem- plate matching should be used with caution. In particular, detection thresholds used in multi-station template matching may not be adequate for single-station template matching and produce an inac- ceptable number of false detections. The authors make synthetic templates and data from filtered white noise. They investigate the effect of the template length (5s, 10s and 20s) and of the frequency band (0.1-0.4Hz, 0.1-0.8Hz, 0.1-1.6Hz) on the statistics of correlation coefficients. The paper is very well written. However, I have a few suggestions to clarify some points of the manuscript.

3.2 On the use of "white noise"

I found the use of "white noise" a bit misleading. The authors make synthetic signals by first generating white noise and then filtering it. Filtered white noise is, by definition, not white noise anymore. The waveforms of the filtered white noise are very far from looking random (see top panels in Figure 1). Showing some waveforms in the paper would actually be helpful to get an intuition of the effect of filtering on the output of template matching. I suggest that the authors don't use "white noise" alone when talking about the filtered waveforms.

We agree that the terminology could be improved; the point of using white noise for the raw records is to guarantee that the raw data is by definition uncorrelated on average, and then to investigate how that translates to template observations after the application of filters. Using white noise as the initial data also allows the results to be both presented in "intuition friendly" units of Hz for frequency and s for time for a particular choice of windows that are in the "planetary seismology" range, whilst still allowing the results to be easily rescaled for different regimes. As such, it is important to both emphasise that the initial data is white noise while also acknowledging that it has been altered by data processing; we have decided to use "filtered white noise" where appropriate as the simplest response to the above point, and added discussion

throughout the paper as appropriate to emphasise the distinction between the raw data and the processed data used for template matching.

3.3 Why showing the normalized maximum CC instead of the number of samples that exceed the threshold?

To quantify the frequency of false detections, the authors use the value of the maximum correlation coefficient (CC) normalized by the median absolute deviation (MAD) as a function of time. I believe that a more relevant statistic is the number of CCs that exceed the detection threshold as a function of time, which is a way of illustrating the p-value corresponding to the detection threshold (this comment is further developed in the next section). We agree that the most important ultimate statistic is the number of CCs that exceed a certain detection threshold as a function of time; the advantage of the plots shown in the text is that plotting the maximum up to a certain time gives an indication of the rate of false positives at a range of thresholds in one plot. To make this more concrete, we have supplemented these plots with indicative values of false positive rates expected for certain planetary mission timescales for different thresholds.

3.4 Clarifying the meaning of the detection threshold for single-station template matching

Such a didactic paper could benefit from commenting on the p-value of the detection threshold. The p-value is the probability that any CC will exceed the threshold by chance. As the authors say in the manuscript, a common practice in template matching is to define the detection threshold as c $\times$ MAD(CC) where c is adjusted depending on whether the user wants a very conservative catalog or not. MAD is a robust estimator of the standard deviation, for example, in the case of the normal distribution, MAD is about 1.48 times smaller than the standard deviation. Thus, it is easy to convert c $\times$ MAD(CC) into a p-value for an hypothetical gaussian distribution with standard deviation σ = 1.48MAD. This reasoning is usually justified when using several stations and channels because of the central limit theorem (although we know that noise in seismic data can be very exotic and behavein an extremely non-gaussian way). In the authors' experiment, the CC distributions are not gaussian but the p-value can simply be estimated by the histograms. For example, in the case of the 5s template length and 0.1-0.4Hz band (see Figure 1), MAD is about 0.2 and, therefore, for any c $>$ 5 the p-value is 0 because a CC cannot exceed 1. In this scenario, the maximum possible value, 1, is often randomly hit and there is little one can do to distinguish a real detection from a random detection. It really means that the 5s/0.1-0.4Hz is the worst possible scenario for template matching but the manuscript's Figure 2 seems to show that 5s/0.1-0.4Hz is preferrable to 5s/0.1-1.6Hz. Thus, I think that the take-away message can be misinterpreted because the value of the normalized maximum CC as a function of time is not so relevant. Thank you for this excellent suggestion; of course, it is not the intention of the manuscript to argue that a narrower passband is

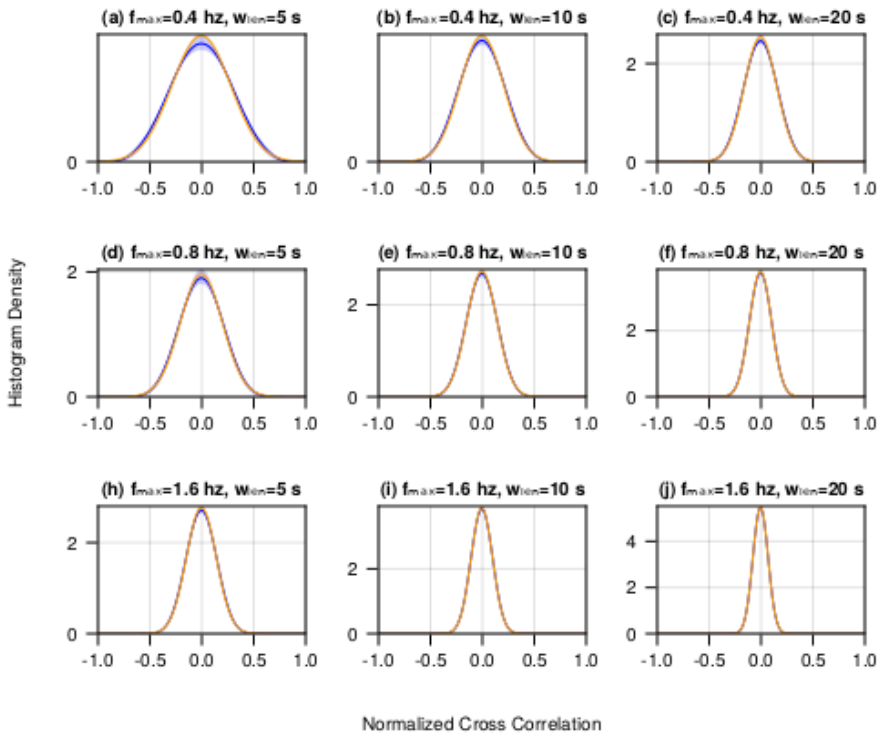

Figure 1: Step-histograms of normalized cross correlation for each configuration (blue), with standard deviations across 32 iterations (light blue) and best fitting Gaussian (orange)

preferable given that you lose all discriminative power, rather that even with wider passbands you perhaps counterintuitively achieve higher MAD ratios and can exceed "common" thresholds surprisingly quickly (the MAD ratio saturation is maybe not counterintuitive to experienced practitioners of template matching but as we have seen studies in the parameter regimes investigated in this paper it is apparent that this point needs to be made more comprehensively). In regards to the p-value consideration, we have included in this response (which is, we believe, publicly accessible in the final version of the paper) a plot in Figure 1 that shows the histograms of NCC results; for the longer windows and passbands they are quite close to normal and so your calculations for p-value would be a good rule of thumb. We think the value of this paper is primarily in the discussion of the relevant timelines to expecting the exceedence of a representative value, so that is what we have kept the focus on in the paper.

3.5 Unimportant comment on the use of GPUs for template matching

A comment to take or leave: Lines 52-55, the authors comment on the computational progress made possible by the use of GPUs to efficiently parallelize the CC computation. Why not citing a study dedicated to the implementation of template matching on GPUs (e.g. Beaucé et al., 2018)? The QTM catalog (Ross et al., 2019) is not the best example of a template matching catalog... We agree that it is appropriate to add the study of Beaucé et al. as a foundational implementation for GPU template matching. Given its high impact in the observational seismology community, we have kept the reference to Ross et al., with the disclaimer "albeit with potential concerns regarding the overall rate of false detections" — this does after all relate to the concerns brought up in this paper!