Dear Editor,

Please find attached the “response letter” in which we addressed all the points raised by the reviewers. First of all, we wish to thank the reviewers for their constructive comments. We think that following their indications improved the clarity of the description of our novel approach and made the paper more readable for a wide audience. Mainly, we modified the manuscript as follows: (a) we improved the description of the methodology, re-organising the manuscript’s structure to avoid misinterpretation of the covariance matrix and the computation of the differential travel-time; (b) we extended the discussion including the reviewers’ comments on the cross-correlation, seismicity rate and changhepoints relevance; and (c) we included the synthetic tests in an appendix. We also improved in writing, incorporating all the comments/corrections annotated by Reviewer #1 (Cliff Thurber).

In the following, reviewers’ main comments are reported together with our replies (in *italics*). Line numbers refer to the current version of the manuscript. The original reviews are attached at the end of the letter for reference.

– –

REVIEWER #1 - Cliff Thurber

[R1.1] My first suggestion is that the rather vague title be modified to better reflect this key theme of the manuscript.

We have an alternative title: "Preparing double-difference data for seismic tomography: an application of a transdimensional algorithm for data-space exploration”. However, focusing on seismic tomography only in our opinion would put aside the application on DD data for other purposes, such as refined hypocenter location. We therefore prefer keeping the original title, but we are open to more discussion

[R1.2] The authors need to be more careful about the use of the term's error (errors) versus uncertainty (uncertainties). I believe that for the majority of their uses of error (errors), uncertainty (uncertainties) is the correct word to use.

We changed errors with uncertainties in the relevant places.

[R1.3] I have difficulty understanding their constructions of the covariance matrices Ce, Ce*, and Ce,w. They refer to Piana Agostinetti and Malinvervo (2018) for the explanation of Ce,w, but in my opinion, this needs to be explained here. Where Ce comes from is not at all clear, and then Ce* is the final version of Ce, which comes from Figure 4b, but what is Figure 4b based on? Further confusion is added because Figure 4b is not the least bit diagonal in appearance, but Ce* is assumed to be diagonal. The entire development section needs to be made crystal clear and thereby be reproducible.

We agree with the reviewer about the potential mis-interpretation and we explicitly separated (and changed the names) the two Covariance matrices: the one used to compute the double-difference data (Ce,w) and the ones used in the exploration of the data space (Ce and Ce). We re-organized the manuscript according to point [R2.2] and we think that the Method section is much clearer now than in the previous version. The preparation of the DD data is now presented in details in an Appendix (Appendix B) (Lines 418-424), LL 418-424 hereinafter).*

[R1.4] I also found the presentation of the weights on page 11 to be overly confusing. One would expect that for a DD datum for which the events cross a change point, the weight would be small. However, the wij factor that enters the exponent in Equation 9 is a sum of “the weights associated

to (sic) each changepoint", giving the impression that the corresponding weight Wij(m) can be large. It confuses things even more by having the variable m be used for two different things in Equation 10.

We corrected Equation 10. To avoid confusion, we explicitly presented the weights (LL 207-209) as generally expected (i.e. lower weights for DD data crossing many change-points). This is basically a simple change in the presentation style, while it does not influence the algorithm itself. Due to the fact that we are sampling a uniform prior for the logarithmic value of the weights, we plotted the inverse on a logarithmic scale (negative values).

[R1.5] The discussion of the MCMC sampling could be improved. The four "moves" have specified probabilities (0.4, 0.4, 0.1, and 0.1, respectively). How is the move to be made actually chosen? Is it the case that some random number between 0 and 1 is drawn, and if it is between 0 and 0.4, move 1 is selected, etc.? Is model fitness used as a factor, as it usually is in MCMC? If so, how? Clarity would be appreciated.

We extended the Method section and the description of the McMC (LL 185-196 and 230-232) including a new flow-chart (Figure 2). We do not enter too many details about McMC because the algorithm follows a standard McMC sampling structure. What is really important to specify (because the moves are unique to the application) is the “recipes" which is described in detail.

[R1.6] I also recommend that the synthetic tests referred to on page 12 be presented to provide confidence in the claimed parsimonious character of the method.

We included the synthetic tests in Appendix A.

[R1.7] A recent paper by Roecker and coworkers (Double Differencing by Demeaning: Applications to Hypocenter Location and Wavespeed Tomography, BSSA, 2021) shows that a demeaning approach performs as well as, and is mathematically equivalent to, the DD approach. This paper should be cited and the potential application of the new method presented here to their demeaning approach should be addressed, if possible.

We added the reference and commented our results in light of the demeaning approach (LL 354-356).

[R1.8] In general, a methods paper is more valuable if it is actually applied to something tangible. Here, the authors stop at determining changepoints, which by themselves don’t really teach us anything about Katla volcano. Isn't there something worthwhile that can be done with these data and the new method?

We do not think that including a full application of the results obtained here in terms of, e.g. a tomography, is a good idea. We think that such addendum might shift the focus of the paper from the novel methodology.

Reviewer #2 - Jiaqi Li

[R2.1] The authors give a clear basic background introduction to 'double-difference data' and? 'Bayes theorem'. However, besides introducing the background, this part should also serve the following sections or the whole paper. For example, one of the main focuses of the paper is to construct the covariance matrix of the error, which is only simply mentioned in the introduction part. If the introduction part can be more specific, say focus on the methods used in the paper, the readers are more clear how these methods fit in the current Bayes theorem, and the following parts can also be much easier to be understood.

We now specify in detail the role of the Covariance matrix at LL 139-145 and 164-171. See also answer to point [R1.3]

[R2.2] Following #1, I also found that the Data and Method parts seem to be somehow mixed. Maybe first clarify the methods and then talk about their application would be better, especially for a method-focused paper.

We re-arranged the manuscript following the reviewer's suggestion. Now the Method section come first (LL 146 ss)

[R2.3] As for the inconsistency when compared with the cross-correlation (CC), my understanding is that this paper's method can reflect the similarities/differences between any traces (because the covariance matrix is used), however, the CC is calculated according to one major event. I wonder, what if you calculate the CC within each window, instead of with the largest event? Whether the comparison will be more consistent or not?

We plotted a figure with the comparison proposed by the reviewer (see the figure attached at the end of the letter). We think that the comparison between our change points and the CC analysis has not changed much and we kept the old figure in the manuscript. We think that plotting all CC values for all event pairs makes the figure a bit confusing (for example, we should attribute each CC value to a certain time. In the presented figure we decided to plot it at the center of the time interval between the two events, but alternative choices could be made). For some windows, the mean of the CC is increased with respect to the average CC over the entire period, but not for all. Thus, we can say the our time-windows are not the one that "optimise" (in some sense) the CC.

[R2.4] As for the inconsistency when compared with the seismicity rate, I think it might be dangerous to draw the conclusion that 'the time-history of the seismicity rate should be carefully evaluated', based on the current comparison and evidence. From my understanding, the data the authors used are the individual traces, and the time-related information, also in the data, is not considered (e.g., what is the time interval between two traces). If this is the case, in my opinion, the authors are looking at 'different parts' of the data from the people who study seismicity rates. For example, assuming a synthetic case (two time periods) with a constant subsurface elasticity but with a different stress/strain field. The waveforms in these two time periods should be exactly the same (therefore no changepoint is defined from this paper's method), but the seismicity rate can be different. Therefore, the discussion on seismicity rate needs more considerations.

We extended the discussion about the seismicity rate at LL 312-314. The reviewer is correct, and we modified the discussion to smooth our conclusion. Nevertheless, we still point out the differences in the time-evolution of the two indicators (DD data and seismicity rate) and suggest that an integrated analysis could be necessary.

[R2.5] A minor question about the terminology. From my understanding, when there is one event and one station, and we compare the differences between the observed data and the synthetic data, this can be valued as 'single-difference'. And when there is a stationor event-pair, this is called 'double-difference'. But in this paper, only the observed data (observed travel time) is mentioned. I wonder if it is still suitable to call it doubledifference data when the synthetic data part is not involved in?

We kept the original wording to make the paper easy to follow for a general audience. We agree with the reviewer that our terminology is not totally correct, but it makes easier for the reader to put our algorithm for DD data preparation in the correct context rather than using the word "single-difference", which is quite unusual.

[R2.6] I think a major revision, especially on the structure (how to better present the method), is needed to make it a better paper. The method should be easier to be understood, rather than

harder, through the authors' re-construction. Then, the following-up applications and the scientific findings behand have the chance to be addressed, as it is cleared mentioned at the beginning.

See answer to point [R3.1]

Reviewer #3

[R3.1] The Data, Method, and Result sections are juxtaposed in the current form of this manuscript. This may be a good way to summarize / discuss with someone who is already familiar with this work but personally, it was difficult to follow. I think this manuscript would be much clearer if authors delineated the main contents (before the Discussion section) more strictly so that each section can be fully understood on its own.

We focused the paper more on the Method section (Data section is now a subsection of the Method section, see comment [R2.2], and part of the "Data preparation" is now Appendix B, see comment [R1.3]). We summarised in more detail the content of each subsection of the Method section at the end of the initial paragraph of the Method section itself where we inserted a new flowchart (LL 185-196)

[R3.2] Furthermore, it would be great to include a flow chart that summarizes and describes the overall methodology and the corresponding sub-steps.

We inserted a new flow-chart (Figure 2) in the method section

[R3.3] Why not start from a simple synthetic dataset for more straightforward assessment to understand whether the number and time-length of their analysis windows accurately capture elastic changes of the subsurface? It sounds like the authors already have tested their method with some synthetics (e.g., L251-253) so why not share them more extensively in the manuscript?

We added the synthetic tests as Appendix. See answer to point [R1.6]

[R3.4] How significant is the number of changepoints inferred from the analysis? For example, if the max. number of changepoints is fixed to a certain value, would the prediction still be able to explain the data equally-well (the largest probable changepoint, #3 seems pretty good in terms of clustering the data into ones that have stable CCs vs. ones that have variable CCs while the rest of changepoints don't seem to correlate well with the CCs)?

The relevance of each chagepoint is directly measured as its probability of occurrence in time. Having the full Posterior Probability Distribution (PPD) we can easily "count" in how many models a particular changepoint is present or not. We make a test with a fixed number N=4 of changepoints, to show, for example, which are the four most relevant changepoints. Obviously this does not mean that the other changepoints can be excluded, but they are definitely less relevant. We include the figure with the results of the test at the end of the letter, but not in the manuscript to keep it easier to read. From the figure, it is evident that Changepoints #1, #3, #6 and #8-9 (following the manuscript order) are the most relevant. It is worth noting that such changepoints are the ones with the highest values found by our algorithm for the data-space exploration (Figure 5d).

[R3.5] In addition, the evolution of the number of changepoints in Fig. 5 generally shows that there are mainly 3-5 clusters / families of models which are stable across the model ensemble. What are the differences between those models? Do all of them show similar behavior shown in Fig. 7-8?

In Figure 5b (now Figure 4b), the number of change points for the sampled models is represented by the blue crosses. It is hard to say that families with different and stable number of change points

exist. If the reviewer means the red symbols, there are 3-5 clusters of models which display similar values of Nd/Nf. This could mean that the position of the changepoints, more than the number, could be clustered for some changepoints (i.e. there are more relevant positions in the sampled models). However, this is exactly what happens with the "fixed N" example (see point [3.4]). Our idea is that the clusters of models with higher Nf/Nd are the ones that does not include less relevant positions.

[R3.6] if you were to compare the resulting clusters to any other widely used clustering algorithms (perhaps with an optimization-based approach that would be much faster to converge) would you expect similar segmentation of the data?

Could be. However, it would remain the main issue of defining the number of "significant" clusters, which is similar to the problem of selecting the number of change points. Our probabilistic approach can help in defining the most probable changepoints and the less ones.

[R3.7] it appears to me that the time-occurrence of the resulting changepoints is only discussed based on the timings of the event cluster (e.g., L190-193). Could you comment on any observable spatial patterns (e.g., Fig. 2b)?

Spatial patterns seem to be rather chaotic and not relevant to our analysis. We did not comment on the spatial patterns in the seismicity as it has been already done in two different papers (Sgattoni et al, 2016ab)

[R3.8] To test the significance of your optimal parameters, can you evaluate a "null hypothesis" model in which you randomly assign those two parameters within the distribution of the actual data?

We do not really understand the comment. Are we talking about how a randomly-picked 2-changepoints model would be able to separate the data?

ORIGINAL REVIEWS

REVIEWER #1 - Cliff Thurber

The manuscript by Piana Agostinetti and Sgattoni presents a novel scheme for identifying times in an earthquake arrival time data set in which a change in the pattern of differential times occurred (presumably due to a change in the velocity structure). This can be useful for the purpose of 4-D tomography (3-D plus temporal change) when data from different, relatively stable time periods should be separated. My first suggestion is that the rather vague title be modified to better reflect this key theme of the manuscript.

The authors need to be more careful about the use of the term's error (errors) versus uncertainty (uncertainties). I believe that for the majority of their uses of error (errors), uncertainty (uncertainties) is the correct word to use.

I have difficulty understanding their constructions of the covariance matrices Ce, Ce*, and Ce,w. They refer to Piana Agostinetti and Malinvervo (2018) for the explanation of Ce,w, but in my opinion, this needs to be explained here. Where Ce comes from is not at all clear, and then Ce* is the final version of Ce, which comes from Figure 4b, but what is Figure 4b based on? Further confusion is added because Figure 4b is not the least bit diagonal in appearance, but Ce* is assumed to be diagonal. The entire development section needs to be made crystal clear and thereby be reproducible.

I also found the presentation of the weights on page 11 to be overly confusing. One would expect that for a DD datum for which the events cross a change point, the weight would be small. However, the wij factor that enters the exponent in Equation 9 is a sum of “the weights associated to (sic) each changepoint", giving the impression that the corresponding weight Wij(m) can be large. It confuses things even more by having the variable m be used for two different things in Equation 10.

The discussion of the MCMC sampling could be improved. The four "moves" have specified probabilities (0.4, 0.4, 0.1, and 0.1, respectively). How is the move to be made actually chosen? Is it the case that some random number between 0 and 1 is drawn, and if it is between 0 and 0.4, move 1 is selected, etc.? Is model fitness used as a factor, as it usually is in MCMC? If so, how? Clarity would be appreciated. I also recommend that the synthetic tests referred to on page 12 be presented to provide confidence in the claimed parsimonious character of the method.

A recent paper by Roecker and coworkers (Double Differencing by Demeaning: Applications to Hypocenter Location and Wavespeed Tomography, BSSA, 2021) shows that a demeaning approach performs as well as, and is mathematically equivalent to, the DD approach. This paper should be cited and the potential application of the new method presented here to their demeaning approach should be addressed, if possible.

In general, a methods paper is more valuable if it is actually applied to something tangible. Here, the authors stop at determining changepoints, which by themselves don’t really teach us anything about Katla volcano. Isn't there something worthwhile that can be done with these data and the new method?

My final point is that the manuscript needs major improvements in the writing. I have tried my best to provide suggestions for such improvements in the annotated scan of the paper.

Reviewer #2 - Jiaqi Li

The authors provided an automatical and data-driven approach to pre-process seismic data and applied their method to the Katla volcano data to define the changepoints. The aim is clear and the method is suitable. My major suggestion is that the way how they present their method and its application is hard to follow and can be re-organized so that the idea can be clearly conveyed and understood. I have the following suggestions and comments:

1. The authors give a clear basic background introduction to 'double-difference data' and? 'Bayes theorem'. However, besides introducing the background, this part should also serve the following sections or the whole paper. For example, one of the main focuses of the paper is to construct the covariance matrix of the error, which is only simply mentioned in the introduction part. If the introduction part can be more specific, say focus on the methods used in the paper, the readers are more clear how these methods fit in the current Bayes theorem, and the following parts can also be much easier to be understood.

2. Following #1, I also found that the Data and Method parts seem to be somehow mixed. Maybe first clarify the methods and then talk about their application would be better, especially for a method-focused paper.

3. As for the inconsistency when compared with the cross-correlation (CC), my understanding is that this paper's method can reflect the similarities/differences between any traces (because the covariance matrix is used), however, the CC is calculated according to one major event. I wonder, what if you calculate the CC within each window, instead of with the largest event? Whether the comparison will be more consistent or not?

4. As for the inconsistency when compared with the seismicity rate, I think it might be dangerous to draw the conclusion that 'the time-history of the seismicity rate should be carefully evaluated', based on the current comparison and evidence. From my understanding, the data the authors used are the individual traces, and the time-related information, also in the data, is not considered (e.g., what is the time interval between
two traces). If this is the case, in my opinion, the authors are looking at 'different parts' of the data from the people who study seismicity rates. For example, assuming a synthetic case (two time periods) with a constant subsurface elasticity but with a different stress/strain field. The waveforms in these two time periods should be exactly the same (therefore no changepoint is defined from this paper's method), but the seismicity rate can be different. Therefore, the discussion on seismicity rate needs more considerations.

5. A minor question about the terminology. From my understanding, when there is one event and one station, and we compare the differences between the observed data and the synthetic data, this can be valued as 'single-difference'. And when there is a stationor event-pair, this is called 'double-difference'. But in this paper, only the observed data (observed travel time) is mentioned. I wonder if it is still suitable to call it doubledifference data when the synthetic data part is not involved in?

6. Some of the typos:
Line 100: Such solution have (has)
Line 108: where p(mld) represent (represents)
Line 173: we compute compute (delete one of the compute)

To conclude, I think the aim of this paper is clear: how to find changepoints of the data, and their method should be suitable. However, the current structure prevents readers from fully understanding what is and how to exactly use their method. I think a major revision, especially on the structure (how to better present the method), is needed to make it a better paper. The method should be easier to be understood, rather than harder, through the authors' re-construction. Then, the following-up applications and the scientific findings behand have the chance to be addressed, as it is cleared mentioned at the beginning.

Reviewer #3

Piana Agostinetti and Giulia Sgattoni present a study that focuses on partitioning / clustering double difference seismic travel time measurements using a transdimensional Bayesian-based method that they have devised. The authors apply their new method to a sample dataset which comprises an earthquake cluster near Katla volcano in Iceland. Given the null hypothesis of all 1141 events within the earthquake cluster being co-located beneath invariant subsurface elastic field, the authors test how their new method can quantitatively partition the data into different groups and explain the waveform similarities within groups as well as the seismicity rate in the region.

I find this manuscript interesting and think that it can potentially be an important contribution to more objectively selecting key parameters (e.g., the number and length of the time-windows) in double difference data. This can eventually enhance the robustness of subsequent seismological analysis such as seismic tomography or earthquake relocation. I have a few points worth further clarification before the manuscript is suitable for publication.

Structure of the manuscript:
The Data, Method, and Result sections are juxtaposed in the current form of this manuscript. This may be a good way to summarize / discuss with someone who is already familiar with this work but personally, it was difficult to follow. I think this manuscript would be much clearer if authors delineated the main contents (before the Discussion section) more strictly so that each section can be fully understood on its own. Furthermore, it would be great to include a flow chart that summarizes and describes the overall methodology and the corresponding sub-steps (as well as descriptions of the priors).

Application to a complex dataset before validating the method:
For a methodology-oriented research paper, I think it is logical to first test the new method on a very simplistic case (e.g., synthetic dataset such as Fig. 1) prior to its application to more complicated datasets. As the authors state in the introduction, the selection of the key parameters in double difference data is nonunique and largely qualitative. Therefore, even if the authors find some plausible posterior distributions of the hyperparameters with their new Bayesian approach, a minimum level of manual supervision is expected to draw a final set of the optimal parameters. That said, this final step can be very difficult to assess when the dataset that is being analyzed itself is complex. The dataset that the authors used in the manuscript is a rather complex one, at least for validation purposes. As the authors pointed out in the main text, their dataset includes some complexities: 1) changes in station network configuration that resulted in loss of data and degraded the quality of earthquake locations as well as origin times; 2) other temporal changes in waveforms that were identified through previous work. Why not start from a simple synthetic dataset for more straightforward assessment to understand whether the number and time-length of their analysis windows accurately capture elastic changes of the subsurface? It sounds like the authors already have tested their method with some synthetics (e.g., L251-253) so why not share them more extensively in the manuscript?

Sensitivity of the analysis on data collected near Katla volcano:

-How significant is the number of changepoints inferred from the analysis? For example, if the max. number of changepoints is fixed to a certain value, would the prediction still be able to explain the data equally-well (the largest probable changepoint, #3 seems pretty good in terms of clustering the data into ones that have stable CCs vs. ones that have variable CCs while the rest of changepoints don't seem to correlate well with the CCs)? In addition, the evolution of the number of changepoints in Fig. 5 generally shows that there are mainly 3-5 clusters / families of models which are stable across the model ensemble. What are the differences between those models? Do all of them show similar behavior shown in Fig. 7-8?

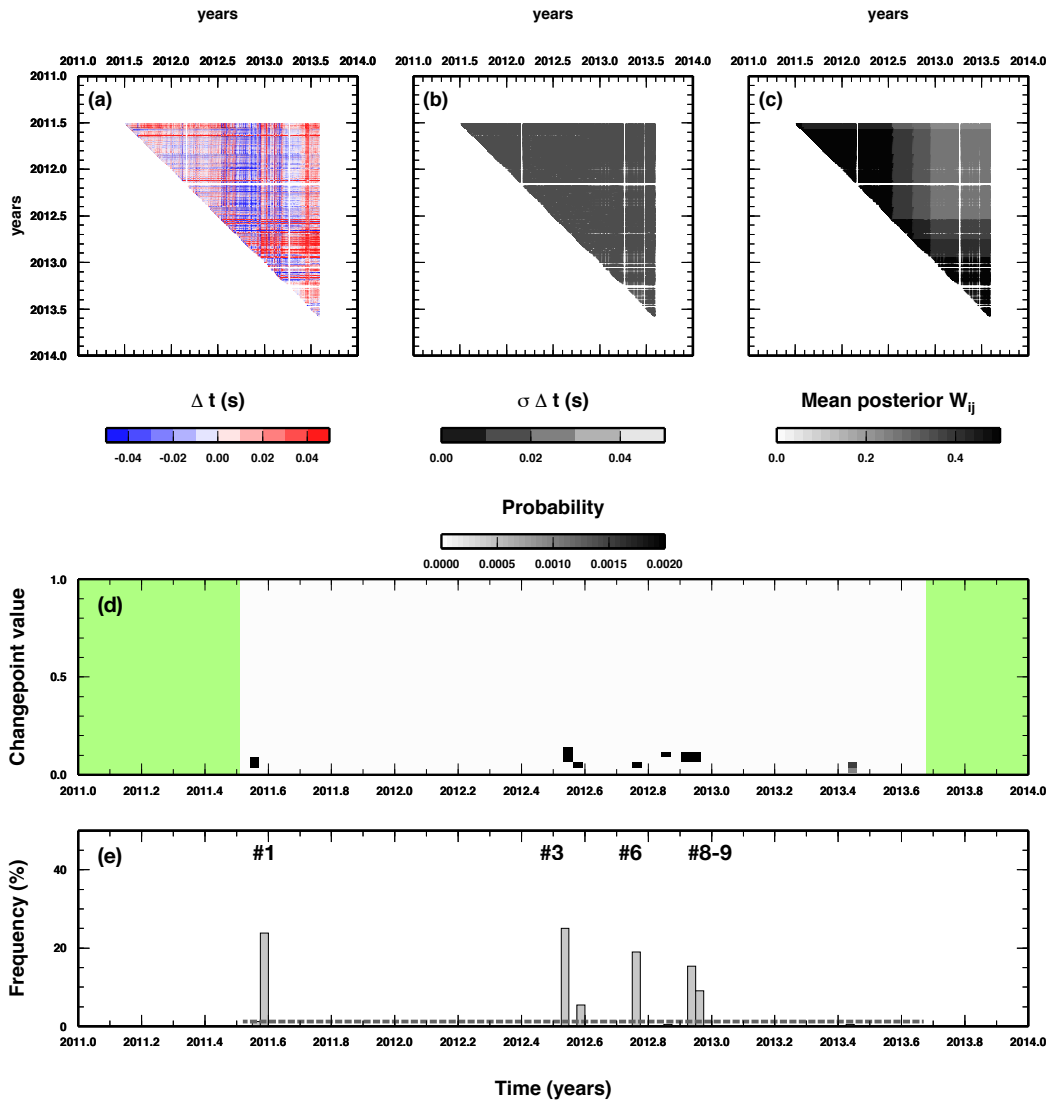
-Another related note here is how significant are those less-probable changepoints (other than changepoint #3); 1) would they correlate with other variables that are not currently discussed, e.g., variances of the CCs within the data delineated with gray lines in Fig. 7 or changes in seismicity rate in the region (e.g., in addition to what you state L322-323)? 2) if you were to compare the resulting clusters to any other widely used clustering algorithms (perhaps with an optimization-based approach that would be much faster to converge) would you expect similar segmentation of the data?

-It appears to me that the time-occurrence of the resulting changepoints is only discussed based on the timings of the event cluster (e.g., L190-193). Could you comment on any observable spatial patterns (e.g., Fig. 2b)?

-To test the significance of your optimal parameters, can you evaluate a “null hypothesis” model in which you randomly assign those two parameters within the distribution of the actual data?

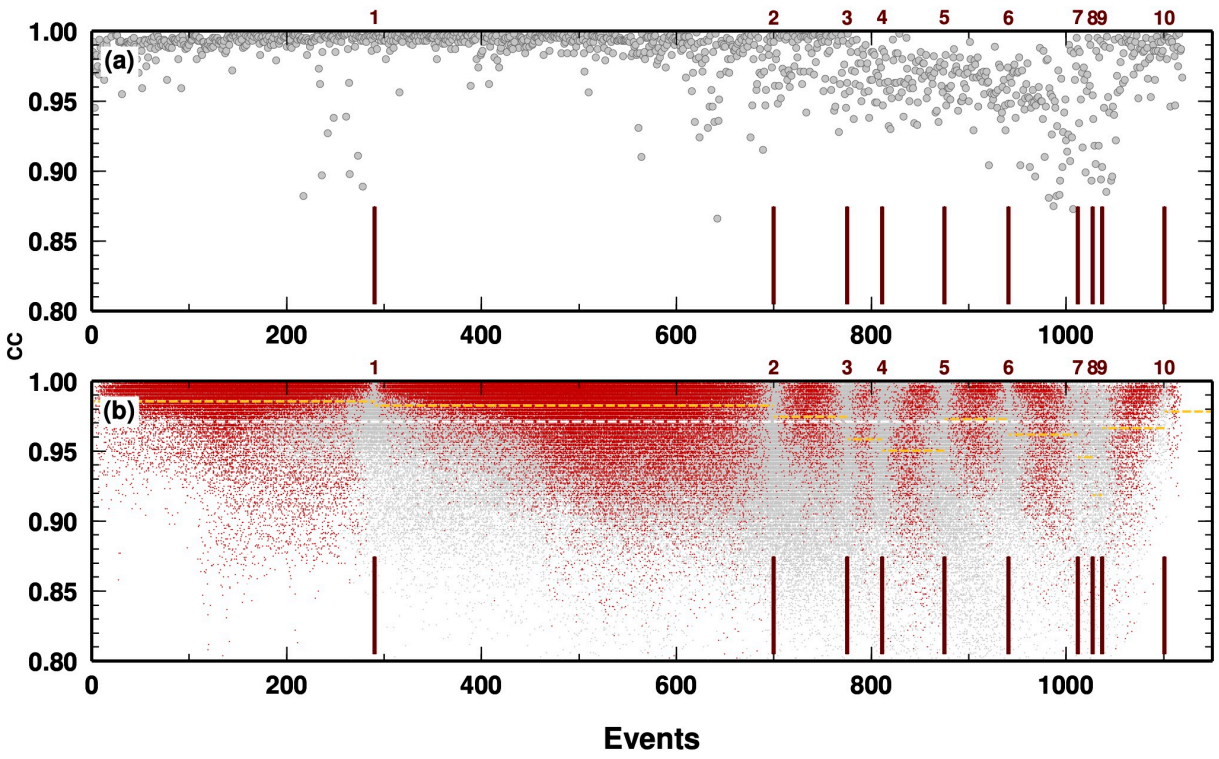
Despite my comments / questions and reservations, I am confident that the new method presented in this manuscript will be a useful tool by providing a complementary solution for choosing the hyperparameters associated with double difference data. This would allow the community to move beyond more common, largely manual, hard-partitioning approaches. I hope the authors find my comments constructive and helpful as they make revisions.

RESULTS FOR A TEST USING A FIXED (N=4) NUMBER OF CHANGEPOINTS



Results for the test with a fixed number of change points (N=4). Symbols as in Figure 5 in the main manuscript. The changepoint numbers refer to the numbers in the main manuscript.

ANALYSIS OF CROSS-CORRELATION VALUES



Comparison of cross-correlation (CC) values for the Master event and between all event pairs. (a) Grey circles represent the CC value computed between the MASTER event (event 435) and all other events. The vertical lines indicate the occurrences of the changepoints found in the analysis (with their respective numbers). The figure is slightly different form the figure in the main text, due to the different length of the time-window used to compute CC value. Circles are plotted along X-axis at the occurrence of the event, i.e. they are equally spaced along X-axis. (b) Grey dots represent all CC values between all events pairs. Each dot is plotted along the X-axis at the mid point of the time-interval between the two events for which the CC value is computed. Red dots report the CC values for the event pairs NOT separated by one or more changepoints. Horizontal white dashed line indicates the average CC value for all event pairs (i.e. the average of the grey dots). Horizontal yellow dashed lines indicate the average CC value for all event pairs between two changepoints (i.e. the average of the red dots within two change points). Vertical lines indicate changepoints found in the previous analysis. It is worth noting that: (1) for some time windows, the average CC value is higher analysing only event pairs not separated by any change point, but not for all. (2) Looking at the first two time-windows, we observe that the CC values are not stable not stable within each whole window, with dominant higher values in the first half with respect to the second half. This seems to indicate that the algorithm is not "optimising" with respect to the CC value, in the sense of separating time-widows with homogeneous CC values.