Review Reports

Reviewer A Comments
For author and editor
In this paper, the authors expand on their previous published work on magnitude clustering in two directions, namely by (i) analyzing a high-quality catalog from Northern California, and (ii) applying additional methods to establish the temporal and spatial footprint of magnitude clustering across various seismic catalogs including lab experiments. I believe the results are publishable in Seismica, but there are a number of points that need to be addressed first.

I) Methodology
a) While the authors perform a spatial and temporal analysis of magnitude clustering separately, I am missing a joint spatio-temporal analysis. This would be particularly interesting for a comparison with the known spatio-temporal properties of aftershocks. I suggest performing such an analysis or discussing why it is not feasible.
b) Temporal analysis: Why fixing n<=100 first and then use an upper bound of 150 hours? What happens if you just use 150 hours, independent of n?
c) l.278-280: By not using proper time lags in your autocorrelation approach but rather allowing mixing across real time scales by considering the discrete numbering of events and then trying to fix it by using the average interevent time difference leads to significant biases. This is because the interevent times for a fixed discrete lag h will certainly follow a broad, non-Gaussian distribution (see, e.g., J. Geophys. Res., 114, B01316, doi:10.1029/2008JB005870). This can explain the observed behavior of a higher "base level" in Fig. 6 B,D.
d) l.307-308: You can't really compare the two temporal measures and say which one indicates "higher" clustering unless they are normalized in a similar way.
e) For me, the main approach is the ECDF method. Yet, I am missing a significance analysis related to the ECDF method. The presented comparison between the original catalogs and individual randomized ones only goes that far. Can the authors assign significance levels to the findings highlighted by Figs. 3-6 and 8, for example?
f) For the randomization, I am missing a clear statement whether excluding interevent times of less than 2 minutes to limit short-term aftershock incompleteness is implemented. If it is, then how? For example, is the randomization done first over all events and then certain pairs are excluded or is it done by performing the randomization AFTER the pair removal?
g) l. 220: Is there a typo in the formula? Seems to be always negative since N>N_m. Please explain clearly.
h) I appreciate the effort by the authors to limit the effect of catalog incompleteness and short-term aftershock incompleteness. However, it might be advisable to take an even more cautious approach. In terms of the magnitude of completeness, it typically varies in space and time due to changes in the seismic monitoring network as documented for Southern California (e.g., Bull. Seism. Soc. Am., 98, 2103–2217) and many other seismic networks including Japan, Switzerland, and Italy (e.g., Geophysical Journal International, Volume 181, Issue 3, June 2010, Pages 1713–1724, Bulletin of the Seismological Society of America (2010) 100: 3261–3268). The magnitude of completeness estimated by the two methods used in the paper at hand can only be

considered as a lower bound of the true magnitude of completeness for the overall study area and period considered. This will particularly affect the observed magnitude clustering involving smaller events, which is the second strongest group in the presented analysis. Similarly, excluding interevent times of less than 2 minutes to limit short-term aftershock incompleteness is only a crude approximation (see, e.g., Seismological Research Letters, 87, pp. 337—344; Bulletin of the Seismological Society of America (2022) 112 (1): 494–507; Phys. Rev. E 78, 041115; Bull. Seismol. Soc. Am., 96, 90–106.). In particular, one potential side effect of the chosen approach by the authors (removing event pairs that are more than 2 min apart) could be the differences between the spatial and temporal analysis (Fig. 6A,C vs 7A,B) mentioned on l. 360-361. Most importantly, from the current presentation it is not absolutely clear that this removal is used in Figs. 4-6, for example. I assume it is but it needs to be clearly stated in the manuscript.
Finally, I am missing a clear description how the authors deal with catalog incompleteness and short-term aftershock incompleteness in the lab experiments (e.g. Fig. 8). It might be described in earlier papers but at least a clear summary needs to be given in the paper at hand.

II) Lab experiments and Universality
a) I found some of the statements too strong given the limited experimental support. For example, it is my understanding that magnitude clustering in lab experiments occurred in TWO marble samples. This does not seem to warrant statements like "was shown ... universally in laboratory catalogs under a shear stress condition" (l.85) and a similar one in l. 374-375. Similarly, it is also important to note that there are cases involving shear without any magnitude clustering (e.g., Phys. Rev. Lett. 131, 154101; Physical Review Letters 108, 038501) that might challenge the proposed universality.
b) It is also quite odd that the cited literature on the universal scaling between lab and field scales is exclusively focused on work done by some of the authors (l.399) and ignores the large body of literature by others in that area.
c) As mentioned above, I am missing a clear description how the authors deal with catalog incompleteness and short-term aftershock incompleteness in the lab experiments. In addition, while the spatial uncertainty of the AE events is given (l.387), the temporal uncertainty including the "dead times" associated with the detection an AE event as well as the limitations of the AE sensors (clipping of large events, etc.) needs to be mentioned.

III) Seismic catalogs
a) l.152-153 (Fig. 2C,D): This seems to be an unfair comparison, since the used m_th is different between the two catalogs. Note also that for increasing m_threshold, the magnitude correlations decrease (Fig. 2A,B) and become insignificant at the 3 sigma level for the highest m_threshold (see also SI) such that the term "universality" is not clearly warranted here. Do the authors have an explanation why the magnitude correlations decrease with increasing m_threshold?
b) Why is the maximum magnitude capped at 5.9 (SC) and 6.0 (NC) in Fig 3?
c) Fig. 10 and l.478-479, 482: The corresponding statements need to be more careful since no significance levels have been established and log-scales do not allow to clearly talk about "near zero".

IV) Aftershocks (l.340-341, l.344-345, l.513-515)
The cited observations regarding the spatial and temporal footprint of aftershocks are a bit outdated. In particular, the statement that "The persistence of the magnitude clustering signature beyond 50 km indicates that magnitude clustering is not driven solely by processes like Omori aftershock rate decay or repeated rupture of an identical fault patch" is not generally true since aftershocks can occur at even longer distances in Southern California (see, e.g., J. Geophys. Res. Solid Earth, 119, 5518–5535, Geophys. Res. Lett., 41, 8818–8824). Related to that, the authors should investigate how the peak at short distances in Fig. 7a depends on the lower magnitude threshold of the analyzed catalog. It is possible that the peak is controlled by the rupture length associated with that smallest magnitude, which might also explain why the peak is absent in Fig. 7b (which has a lower m_c). Similarly, the longer range in Fig. 7A might be a consequence of the larger magnitude events in Southern California (Landers, Hector Mine), which are absent in Northern California.

V) ETAS model
To properly mimic field data, the background rate in the ETAS should not be homogeneous in space and a more realistic spatial kernel (l. 419) would also be desirable (see again, e.g., J. Geophys. Res. Solid Earth, 119, 5518–5535, Geophys. Res. Lett., 41, 8818–8824).

Reviewer B Comments
For author and editor
The topic is interesting, and the manuscript reads well. However, there is a lack of discussion of physical models/interpretations of the statistical observations, there are unclear or unexplained methods, and there are unclear results. Below are some major and minor suggestions I offer to improve the manuscript in these aspects.

<u>Major Comments</u>

Lack of discussion of physical models that can be linked to the statistical observations.

Throughout this manuscript, I asked myself: What are the other models and processes that the authors' continue to refer to that are responsible for their observations/results ? For instance, in Line 38-39 of the abstract, it's stated that there are other processes that control magnitude clustering, but no suggestions for those processes are offered, nor is a new model proposed. It is the same for Line 77, but still the new model has not been mentioned. In Lines 244-247, it's again proposed that there are possible relationships for the observations: "controlling earthquake size", but there is no physical model offered to explain the results. And even later at Lines 295-296 and Line 347, it is mentioned that the linear regression helps to "narrow down the physical mechanism" or there is another physical model but then that physical mechanism/model is not proposed or discussed. In Line 362-363, a physical explanation/interpretation of the result would be useful. For example, is it a combined factor of static and dynamic stress? Finally, in Lines 514-515, Omori aftershock rate is stated as a 'physical' explanation, but it is a statistical view of a physical process (presumed to be related to

relaxation after an earthquake), but there is no discussion of the physical processes in the context of the results of this study. Overall, the entire manuscript could be improved if more physical explanations of the statistical tools used and observations made are included.

Unclear or unexplained methods

In Line 504, it is stated that the "magnitude clustering signature remains significant" based on a comparison with results from a randomized catalog (Lines 138 and 398). As the results are only compared to a single randomized catalog, it is difficult to understand this significance. To truly justify this claim, bootstrapping should be applied to the randomized catalog to see if the true observations fall within the 1-sigma error bounds of the randomized versions. Bootstrapped results should be updated in Figures 5, 6, 7, 8, and 9 where comparisons are made with the true observations.

In Lines 368-369, "specifying time windows that are too narrow leads to an insufficient number of events to conduct reliable statistical analyses"… This seems to be related to the fact that only the next 100 events are considered? It is suggested if more events are included/considered, then perhaps there would be a sufficient number of events to see a stronger magnitude clustering signature. This is a difficult problem, but perhaps the event selection could be dynamic such that the number of events could be scaled to magnitude, based on the b-value and magnitude of completeness analysis that was done? Physically, this could make sense because larger magnitude earthquakes should 'trigger' more earthquakes than smaller magnitude earthquakes (based on the Gutenberg-Richter relation), but it may affect the statistics (or stacking, if done).

In Line 131, it's not explained why there is a 2-minute limit. Is it because there is a M7 event in the catalogs, and this limit is equivalent to the duration of a M7 event? I suspect useful events may be excluded with this time limit. Have you considered scaling the time limit with magnitude, i.e. earthquake duration (M2 = 80 seconds, M7 = 120 seconds). In that way, more events would be included, especially since there is only 1 M7 event in the catalogs, but there are 10,000 or more M2 events in the catalogs. However, I think in Figure 2 a filtered and unfiltered version is shown where this 2-minute limit and STAI time window is not applied? It's unclear.

With such a large catalog extent (North or South California), why is only timing between events considered and not spatial proximity (e.g. Lines 141-144)? It appears to be only 'timing' even though in Lines 239-241, spatial distance windows are mentioned, but the spatial distance windows are not discussed in the approach. Unless this is what you are describing in Line 212, but based on your description this is only catalog distance (index), not spatial distance, I presume? Even in Lines 202 and 226, there is a presumed 'sequence', but there does not appear to be any consideration for the spatial relationship between the events. Unfortunately, there is a chance that successive events that are separated by 100s of km are being compared (see your Figure 1); this is something that was observed in Aiken and Obara (2021). In their study, unrelated events were clustered together – the events occurred close in time but far in space

(see their supplemental – cluster identification). Was there ever a case of an event occurring more than 100 km away in a similar time frame of the selected 100 events (Lines 324-326)? Or is the smallest distance accepted 100 km and also includes events that occur further away? Also, in Lines 228-229: What is distance decay in the ECDF? This is shown in the Table S1 but not presented in the main text. Is this associated with a spatial clustering feature? An explanation of the how the spatial feature is accounted for in this study is needed. Furthermore, in Line 204, it is stated that the result is 'remarkable', but it seems that the results indicate that at any time we can have similar magnitude events occurring over large distances. It's not clear how this tells us something new. The feature that is missing in this analysis is spatial relationship to magnitude clustering.

Aiken, C. and Obara, K. (2021), Data-driven clustering reveals more than 900 small magnitude slow earthquakes and their characteristics, https://doi.org/10.1029/2020GL091764.

Unclear or unexplained results

It seems the results in Figure 6 are for D and H shown in Figure 4? It's hard to understand what this figure represents exactly following your previous analysis. A percent difference was calculated for each mi and mi+100 along the diagonal shown in Figure 4 (D&H; I presume). So there should be 5 percent differences lines, but only one set along that diagonal is shown? Were the results stacked or something else was done? Why not show all?

In Line 336, it is stated that there is a "linear trend" after 20 km for Southern California and after 5 km for Northern California, but the fast drop in short distance (in Figure 7) is not explained.

In Lines 340-341: "linear density of aftershocks…" I suspect that this is true for short distances and larger magnitudes (where static stress is important) but linear at longer distances (where dynamic stress is more important). Perhaps more discussion of this physical aspect can be added and linked to the statistic observations? This links to a previous comment about the manuscript needing more physical interpretation.

In Line 469: "logarithmic decay". There seems to be a contradiction between what is shown in Figure 10 A & B and what is said about the relationship between PD and time. For example, in Lines 286-287, the relationship is described as a rapid decrease at short time intervals and that it is more gradual at longer time intervals. Indeed, in Figure 6, it is mostly linear and "gradually decreases" outside of the short time intervals. Even in Figure 10, the plot is log-log, in which the data does appear to be *mostly* linear, but it is clear that there are tails deviating from the linear relationship for the 0-20% and 20-80% groups. So there seems to be some contradiction unless I am missing something?

<u>Minor comments</u>

Line 57: It is not clear to me what seismic modeling is.

Line 67-69: Magnitude clustering is presented as the topic of this article, but it has not been defined what exactly magnitude clustering is. Is it events of similar magnitude occurring near each other or events of different magnitude occurring near each other or something else? How close in magnitude must they be? Exactly the same, difference of .1 or something else?

Line 77: This is very general; too general for the topic. It would be helpful to the reader if there is an expansion on what is not understood about seismogenesis and fault interactions. For instance, we still don't know how slow slip is related to seismogenesis, but in this study, I think it is more about aftershock sequence interactions?

Line 88 and 97: Xiong et al. missing year (ignore if this is an acceptable style of reference).

Line 124: “summarizing the steps of the analysis”; this phrase is awkward.

Line 126-128: Need to specify that the catalogs are processed separately and that this observation is for the North California catalog.

Line 176-179: This sentence reads like it belongs in a figure caption and not the main text.

Line 241: I think you are trying to reference # 5, but it looks like km^5

Line 386-387: It's scaling issue?

Line 433-434: Missing some words to make this a complete sentence.

Line 498: It’s unclear how the method is “novel.” It appears that this work is just an extension of a previous work (Xiong et al.).

Figure feedback

Figure 2. I don't understand what the difference colors in A and B mean. It looks like magnitudes but magnitudes of what? the subsequent event or the triggering event? Or is it different magnitudes of completeness? Or differenced magnitude bins?

Are the filters the 2-3minute differencing and the STAI factors?

Response to Reviewers

Reviewer A:

In this paper, the authors expand on their previous published work on magnitude clustering in two directions, namely by (i) analyzing a high-quality catalog from Northern California, and (ii) applying additional methods to establish the temporal and spatial footprint of magnitude clustering across various seismic catalogs including lab experiments. I believe the results are publishable in Seismica, but there are a number of points that need to be addressed first.

I) Methodology

a) While the authors perform a spatial and temporal analysis of magnitude clustering separately, I am missing a joint spatio-temporal analysis. This would be particularly interesting for a comparison with the known spatio-temporal properties of aftershocks. I suggest performing such an analysis or discussing why it is not feasible.

- We considered how one might perform the proposed joint analysis, but it was unclear how to combine the information of space and time, and it was further unclear how to interpret the resulting decay in magnitude clustering with increasing combined distance-time. For example, we found that higher weighting of interevent distance information resulted in a more linear decay and higher weighting of interevent time resulted in a more logarithmic decay. This did not provide more insight than what we gleaned from the separate space and time analyses. However, we did find that we could better represent the relative decay of magnitude clustering with varying space and time differences using a 3-dimensional plot. We believe this new figure helps the reader to see how the decay patterns persist across a range of different time and space constraints. We have added a short section to explain this analysis of combined spatial and temporal patterns.

b) Temporal analysis: Why fixing n<=100 first and then use an upper bound of 150 hours? What happens if you just use 150 hours, independent of n?

- This was an important suggestion and we adjusted the time and distance decay analysis to remove the interevent number (n) restriction. Instead, we analyzed the decay relationships using only upper bounds of 150 hours interevent time and 100 km interevent distance. Not only does this change better preserve the decay relationship trends, it increases the strength of the magnitude clustering signature seen. The trends in the distance decay plots are similar to the original plots in terms of the observed signature relative to the randomized version. The time decay, however, remains well above random variation throughout the 150 hour period with the ECDF approach, whereas the original analysis with n<=100 decayed down to non-significant levels within the 150 hour period. We also observed a less steep decline in the signature for the smaller interevent times in the Southern California catalog.

c) l.278-280: By not using proper time lags in your autocorrelation approach but rather allowing

mixing across real time scales by considering the discrete numbering of events and then trying to fix it by using the average interevent time difference leads to significant biases. This is because the interevent times for a fixed discrete lag h will certainly follow a broad, non-Gaussian distribution (see, e.g., J. Geophys. Res., 114, B01316, doi:10.1029/2008JB005870). This can explain the observed behavior of a higher "base level" in Fig. 6 B,D.

- The autocorrelation plots have been adjusted to use interevent times from 0-150 hours for the lag rather than the event number. Again, the catalogs are restricted to upper bounds of 150 hours and 100km to be consistent with the ECDF method. The results show a similar increase in the clustering signature (represented by the autocorrelation coefficient in this case) as the new ECDF time decay plots, and the trend in the decay is now more similar between the ECDF and autocorrelation methods.

d) l.307-308: You can't really compare the two temporal measures and say which one indicates "higher" clustering unless they are normalized in a similar way.

- We appreciate the reviewer's point here, and have removed language regarding which indicates "higher" clustering signature from the text. To further compare the similarities between the decay trends across the two methodologies, we performed a min-max normalization on the two datasets and directly compared the rescaled patterns in a new supplementary material plot (Figure S4).

e) For me, the main approach is the ECDF method. Yet, I am missing a significance analysis related to the ECDF method. The presented comparison between the original catalogs and individual randomized ones only goes that far. Can the authors assign significance levels to the findings highlighted by Figs. 3-6 and 8, for example?

- Bootstrapping of randomly removing 10% of the data over 100 cycles was completed on both the real and randomized data to estimate the uncertainty and enable a more statistically relevant estimation of significance. The data points and error bars in the time and distance decay plots now represent the mean and standard deviation calculated with this bootstrapping (Figs. 5-7, 9-10).

f) For the randomization, I am missing a clear statement whether excluding interevent times of less than 2 minutes to limit short-term aftershock incompleteness is implemented. If it is, then how? For example, is the randomization done first over all events and then certain pairs are excluded or is it done by performing the randomization AFTER the pair removal?

- The catalog is filtered for completeness and STAI, and then the randomization is completed over these filtered versions of the catalog. We have updated the explanation of our catalog filtering and processing to address this.

g) l. 220: Is there a typo in the formula? Seems to be always negative since N>N_m. Please explain clearly.

- We thank the reviewer for pointing out this typo, the N and N_m values have been switched around and the formula is correct now.

h) I appreciate the effort by the authors to limit the effect of catalog incompleteness and short-term aftershock incompleteness. However, it might be advisable to take an even more cautious approach. In terms of the magnitude of completeness, it typically varies in space and time due to changes in the seismic monitoring network as documented for Southern California (e.g., Bull. Seism. Soc. Am., 98, 2103–2217) and many other seismic networks including Japan, Switzerland, and Italy (e.g., Geophysical Journal International, Volume 181, Issue 3, June 2010, Pages 1713–1724, Bulletin of the Seismological Society of America (2010) 100: 3261–3268). The magnitude of completeness estimated by the two methods used in the paper at hand can only be considered as a lower bound of the true magnitude of completeness for the overall study area and period considered. This will particularly affect the observed magnitude clustering involving smaller events, which is the second strongest group in the presented analysis. Similarly, excluding interevent times of less than 2 minutes to limit short-term aftershock incompleteness is only a crude approximation (see, e.g., Seismological Research Letters, 87, pp. 337—344; Bulletin of the Seismological Society of America (2022) 112 (1): 494–507; Phys. Rev. E 78, 041115; Bull. Seismol. Soc. Am., 96, 90–106.). In particular, one potential side effect of the chosen approach by the authors (removing event pairs that are more than 2 min apart) could be the differences between the spatial and temporal analysis (Fig. 6A,C vs 7A,B) mentioned on l. 360-361. Most importantly, from the current presentation it is not absolutely clear that this removal is used in Figs. 4-6, for example. I assume it is but it needs to be clearly stated in the manuscript.

- We thank the reviewer for their helpful suggestions and references. We took the suggestion to implement a version of filtering using a rate-dependent magnitude of completeness, as outlined in [Hainzl, 2016]. However, this method actually kept more events in the catalog, and in fact raised the amount of magnitude clustering observed. However, to be even more conservative in our demonstration that magnitude clustering is not due to artifacts of incompleteness or STAI, we have chosen to stay with the more conservative 2-minute window approach.
- To further address the question of incompleteness, we implemented a version of the ETAS catalog with incompleteness added by removing an increasing number of small magnitude events from the catalog (Supplementary Figure S6). Adding this incompleteness does not artificially introduce magnitude clustering in the ETAS catalog, providing further evidence that incompleteness is not driving the magnitude clustering we observed in the real catalogs.
- By imposing the interevent time (150 hr) and distance (100 km) restrictions described in an earlier comment, we also sought to restrict the potential influences of temporal and spatial variability of incompleteness. The amount of magnitude clustering seen in the

time and distance decay plots is now more similar. The manuscript has been updated to reflect this.
- We updated the manuscript to clarify that the catalogs used to make the original Figures 4-6 have undergone the described filtering process.

Finally, I am missing a clear description how the authors deal with catalog incompleteness and short-term aftershock incompleteness in the lab experiments (e.g. Fig. 8). It might be described in earlier papers but at least a clear summary needs to be given in the paper at hand.

- Thanks for the reviewer's inquiry. The description on how to catalog incompleteness was dealt with in the lab catalogs has been added into the main text. The explanation to the STAI has been clarified in both the main text and our response to the point II) c) of the reviewer's comment below.

II) Lab experiments and Universality

a) I found some of the statements too strong given the limited experimental support. For example, it is my understanding that magnitude clustering in lab experiments occurred in TWO marble samples. This does not seem to warrant statements like "was shown ... universally in laboratory catalogs under a shear stress condition" (l.85) and a similar one in l. 374-375. Similarly, it is also important to note that there are cases involving shear without any magnitude clustering (e.g., Phys. Rev. Lett. 131, 154101; Physical Review Letters 108, 038501) that might challenge the proposed universality.

- Thanks for the reviewer's comment. Indeed, the observation of clustering under shear stress condition was verified by a cumulation of catalogs compiled from the rock mechanics tests conducted at different institutes, by different experimentalists, and acquired by different sensor and data acquisition assemblies. For each institute we have the acoustic emissions under the loading conditions of shear and tensile, separately. Among all the tests, magnitude clustering can be observed under shear stress condition and is absent once the stress condition is dominantly tensile. Such observation is consistent regardless if the tests (for both shear and tensile) were conducted at different institutes by different experimentalists using different data acquisition systems. The corresponding clarification has been added to the main text. To illustrate the fact that extensive laboratory catalogs have been included, we attached the references to the corresponding place. That is where the references 21, 26-32 come from.
- The references from Physical Review Letters provide interesting and more inclusive discussions about the magnitude clustering phenomena. However, the focus of the second reference (Physical Review Letters 108, 038501) was the energy releases after the mine blasting, and the scale was way larger than the laboratory-scale, i.e., the scale for the 3550m depth Mponeng gold mine in South Africa. We acknowledge that our laboratory rock mechanics tests have not included the blasting tests. Such damage tests, if conducted in the laboratory, are difficult for acoustic emission data acquisition at laboratory scale. However, such observations remind us that the stress after blasting would be undergoing a relaxation process, which we suspect would have some types of

physical similarities with the tensile stress condition in laboratory rock fracture processes. The first reference (Phys. Rev. Lett. 131, 154101) has conducted a similar field-scale study on a catalog from Italy. For its laboratory-scale catalog however, it was provided from a numerical model. It is difficult to compare the catalog from numerical simulation with that from actual rock mechanics tests.

b) It is also quite odd that the cited literature on the universal scaling between lab and field scales is exclusively focused on work done by some of the authors (l.399) and ignores the large body of literature by others in that area.

- Thanks for the reviewer's comment. The citations are for illustrating that extensive laboratory catalogs had been investigated for reaching the conclusion. We have removed them from here and put them into the place where such illustration of extensiveness is needed. Also, we have added the citation by other authors for emphasizing the scaling between lab and field.

c) As mentioned above, I am missing a clear description how the authors deal with catalog incompleteness and short-term aftershock incompleteness in the lab experiments. In addition, while the spatial uncertainty of the AE events is given (l.387), the temporal uncertainty including the "dead times" associated with the detection an AE event as well as the limitations of the AE sensors (clipping of large events, etc.) needs to be mentioned.

- Thanks for the reviewer's inquiry. The issues with incompleteness in laboratory rock mechanics tests are different with the field-catalog. In the laboratory, the loading rate can be controlled while the tectonic loading cannot be controlled. The commensurate control on seismicity rates minimizes the issues from short-term shaking that prevents recording of small subsequent events. As such, the incompleteness in the lab studies is primarily the incompleteness due to the limitation of detecting low magnitude events, which incorporates into the deviation of power-law in the frequency-magnitude distribution for the laboratory catalogs.
- The deadtime for the acoustic emission data acquisition system was at the order of the data sample rate, and the maximum signal/hit rate for this test was far below the saturation level. As a result, the STAI is not a concern for the laboratory catalog being analyzed in this study. This has been added to the manuscript. Regarding the clipping of large events in the sensor limitations, the processing uses relative magnitudes based on the p-wave first arrival peak for recording the AE events rather than the overall signal peak, so clipping of the signal is not a concern.

III) Seismic catalogs

a) l.152-153 (Fig. 2C,D): This seems to be an unfair comparison, since the used m_th is different between the two catalogs. Note also that for increasing m_threshold, the magnitude correlations decrease (Fig. 2A,B) and become insignificant at the 3 sigma level for the highest

m_threshold (see also SI) such that the term "universality" is not clearly warranted here. Do the authors have an explanation why the magnitude correlations decrease with increasing m_threshold?

- We have adjusted the wording of this sentence to remove the term universality and to acknowledge the difference in magnitude of completeness. We have clarified that we do not see that the decreasing magnitude correlations with m_threshold is due to incompleteness and thus we interpret it as a real feature of the magnitude clustering process. Our tentative hypothesis is that larger magnitude events generate larger stress changes that can trigger a wider range of magnitude earthquakes which reduces magnitude clustering particularly when only the next event is considered. However, justifying this hypothesis would take some considerable additional investigation that we believe is beyond the scope of this paper considering the current focus on spatial and temporal patterns. We have adjusted the text to indicate investigating this variation with magnitudes should be the focus of future work.

b) Why is the maximum magnitude capped at 5.9 (SC) and 6.0 (NC) in Fig 3?

- The way that our catalog was filtered for STAI removed events based on the magnitude of the event compared to the mainshock magnitude and the time between the two events, based on mainshock of magnitude greater than or equal to 6, (Helmstetter, 2006, referenced in the manuscript). Our implementation compared mainshocks to themselves, so our STAI filtering was also filtering out the mainshocks. We have revised our ECDF analysis to include these events, which adds 13 events to each catalog. This number of events is too small to affect the amount of magnitude clustering observed in each of analyses.

c) Fig. 10 and l.478-479, 482: The corresponding statements need to be more careful since no significance levels have been established and log-scales do not allow to clearly talk about "near zero".

- The referenced lines were discussing the patterns in order to compare the two different patterns of decay (time vs. distance) and relate them to possible physical interpretations for the difference in decay pattern. Since the main purpose of this section is to show that the respective decay patterns still hold for all magnitude ranges in the catalog, we felt that these particular points were ultimately not necessary, especially since we have updated the manuscript to focus less overall on physical processes and models. We have therefore removed these lines from the manuscript.

IV) Aftershocks (l.340-341, l.344-345, l.513-515)

The cited observations regarding the spatial and temporal footprint of aftershocks are a bit outdated. In particular, the statement that "The persistence of the magnitude clustering signature beyond 50 km indicates that magnitude clustering is not driven solely by processes

like Omori aftershock rate decay or repeated rupture of an identical fault patch" is not generally true since aftershocks can occur at even longer distances in Southern California (see, e.g., J. Geophys. Res. Solid Earth, 119, 5518–5535, Geophys. Res. Lett., 41, 8818–8824). Related to that, the authors should investigate how the peak at short distances in Fig. 7a depends on the lower magnitude threshold of the analyzed catalog. It is possible that the peak is controlled by the rupture length associated with that smallest magnitude, which might also explain why the peak is absent in Fig. 7b (which has a lower m_c). Similarly, the longer range in Fig. 7A might be a consequence of the larger magnitude events in Southern California (Landers, Hector Mine), which are absent in Northern California.

- We have revised this specific sentence to focus on the points that a repeated fault patch is not a plausible model. We were attempting to draw attention to the observation that Omori aftershock temporal decay rates are not observed beyond 10 km (Richards-Dinger et al., 2010) because we felt this implies that the physical processes associated with causing Omori aftershock decay rates are likely not responsible for the magnitude clustering patterns at significantly larger distances. However, the literature pointed out by the reviewer implies that some form of aftershock driving mechanism is still present at larger distances, so we have decided to remove the reference to Omori aftershock processes.
- Regarding the peak in the distance decay for Southern California, after rerunning the plots imposing time and distance restrictions, the peak is less pronounced, and it is not yet clear whether it is simply a statistical variation in the data or if it has a physical significance considering the peak does not occur in both catalogs. At this point in time we are hesitant to speculate on physical models, associated with rupture length or otherwise, but have noted that this should be a focus of future work. We did check versions of these plots with a higher magnitude threshold of 2 and 2.5 for both catalogs, and there was no systematic change in where a peak occurs, and in some cases it does not occur at all (for example, when increasing the magnitude threshold to 2.5 for the Southern California catalog the peak does not occur), which we would expect if it were due to physical processes associated with rupture length. Therefore it is plausible that this peak is simply due to statistical variability in the datasets.

V) ETAS model

To properly mimic field data, the background rate in the ETAS should not be homogeneous in space and a more realistic spatial kernel (l. 419) would also be desirable (see again, e.g., J. Geophys. Res. Solid Earth, 119, 5518–5535, Geophys. Res. Lett., 41, 8818–8824).

- We appreciate the reviewer’s concern, and we attempted to find publicly available code incorporating nonhomogeneous background rates in the ETAS model. However, such code is not readily available, and implementing this would take considerable time and effort. We do not see a theoretical reason that incorporating spatial heterogeneity in the background rate of the ETAS model would introduce magnitude clustering into the ETAS model, so we believe incorporating this aspect of ETAS modeling is out of the scope of this particular study.

Reviewer B:

The topic is interesting, and the manuscript reads well. However, there is a lack of discussion of physical models/interpretations of the statistical observations, there are unclear or unexplained methods, and there are unclear results. Below are some major and minor suggestions I offer to improve the manuscript in these aspects.

Major Comments

Lack of discussion of physical models that can be linked to the statistical observations.

Throughout this manuscript, I asked myself: What are the other models and processes that the authors' continue to refer to that are responsible for their observations/results ? For instance, in Line 38-39 of the abstract, it's stated that there are other processes that control magnitude clustering, but no suggestions for those processes are offered, nor is a new model proposed. It is the same for Line 77, but still the new model has not been mentioned. In Lines 244-247, it's again proposed that there are possible relationships for the observations: "controlling earthquake size", but there is no physical model offered to explain the results. And even later at Lines 295-296 and Line 347, it is mentioned that the linear regression helps to "narrow down the physical mechanism" or there is another physical model but then that physical mechanism/model is not proposed or discussed. In Line 362-363, a physical explanation/interpretation of the result would be useful. For example, is it a combined factor of static and dynamic stress? Finally, in Lines 514-515, Omori aftershock rate is stated as a 'physical' explanation, but it is a statistical view of a physical process (presumed to be related to relaxation after an earthquake), but there is no discussion of the physical processes in the context of the results of this study. Overall, the entire manuscript could be improved if more physical explanations of the statistical tools used and observations made are included.

- We understand the reviewer's concern regarding the lack of specificity regarding the connection between magnitude clustering and physical models of earthquake rupture/interactions. While we do believe that deciphering the patterns of magnitude clustering can help us gain insight into physical mechanisms, and this is one of the main long-term goals of this research, at this point in time we do not have a definitive physical model to present that matches the spatial and temporal patterns we observed. There is still much debate regarding physical models for static and dynamic stress interactions and their role in earthquake rupture. Understanding what leads to clustering of earthquake magnitudes can ultimately add to the understanding of these physical processes, and this will be a main focus of our future work on this topic, exploring different simulation strategies for generating magnitude clustering. Yet for this manuscript, we have decided to place less emphasis on the discussion of physical

models, and instead focus on a more straightforward presentation of the statistical patterns, along with discussion on the usefulness of including magnitude clustering in forecasting models.

Unclear or unexplained methods

In Line 504, it is stated that the “magnitude clustering signature remains significant” based on a comparison with results from a randomized catalog (Lines 138 and 398). As the results are only compared to a single randomized catalog, it is difficult to understand this significance. To truly justify this claim, bootstrapping should be applied to the randomized catalog to see if the true observations fall within the 1-sigma error bounds of the randomized versions. Bootstrapped results should be updated in Figures 5, 6, 7, 8, and 9 where comparisons are made with the true observations.

- As discussed in a prior response, we have now performed bootstrapping on both the randomized and real data, and we have represented the bootstrap-estimated standard deviations in our plots.

In Lines 368-369, “specifying time windows that are too narrow leads to an insufficient number of events to conduct reliable statistical analyses”… This seems to be related to the fact that only the next 100 events are considered? It is suggested if more events are included/considered, then perhaps there would be a sufficient number of events to see a stronger magnitude clustering signature. This is a difficult problem, but perhaps the event selection could be dynamic such that the number of events could be scaled to magnitude, based on the b-value and magnitude of completeness analysis that was done? Physically, this could make sense because larger magnitude earthquakes should ‘trigger’ more earthquakes than smaller magnitude earthquakes (based on the Gutenberg-Richter relation), but it may affect the statistics (or stacking, if done).

- As discussed in an earlier response, the analyses have been revised to remove the limitation on subsequent event number, and instead consider all subsequent events with an upper interevent time and distance bound of 150 hours and 100 km. The results show that without this interevent number restriction, the strength of the clustering signature is indeed higher in the time decay plots, and more similar to the strength of signature observed in the distance decay.

In Line 131, it’s not explained why there is a 2-minute limit. Is it because there is a M7 event in the catalogs, and this limit is equivalent to the duration of a M7 event? I suspect useful events may be excluded with this time limit. Have you considered scaling the time limit with magnitude, i.e. earthquake duration (M2 = 80 seconds, M7 = 120 seconds). In that way, more events would be included, especially since there is only 1 M7 event in the catalogs, but there are 10,000 or more M2 events in the catalogs. However, I think in Figure 2 a filtered and unfiltered version is shown where this 2-minute limit and STAI time window is not applied? It’s unclear.

- The 2-minute time filter is based on analysis of coda waves from larger events in the catalog. Overlapping of coda waves after the larger events can mask the detection of

smaller events, leading to spurious correlations of magnitudes due to short term aftershock incompleteness (STAI). We did implement a test using a rate-dependent incompleteness filter using the method of [Hainzl, 2016]. This method of filtering did keep more events in the catalog, and actually raised the amount of observed magnitude clustering. However, to be even more conservative in our demonstration that magnitude clustering is not due to artifacts of incompleteness or STAI, we have chosen to use the 2-minute window. The difference in figures 2A,2C and 2B,2D is that B and D show the magnitude clustering probability deviations after the filtering for catalog incompleteness and the 2-minute STAI filter have been implemented. This shows that significant deviations still occur after this filtering. Explanation for why the 2-minute filter was chosen and the difference between the filtered and unfiltered figures have been added to the manuscript text.

With such a large catalog extent (North or South California), why is only timing between events considered and not spatial proximity (e.g. Lines 141-144)? It appears to be only 'timing' even though in Lines 239-241, spatial distance windows are mentioned, but the spatial distance windows are not discussed in the approach. Unless this is what you are describing in Line 212, but based on your description this is only catalog distance (index), not spatial distance, I presume? Even in Lines 202 and 226, there is a presumed ‘sequence’, but there does not appear to be any consideration for the spatial relationship between the events. Unfortunately, there is a chance that successive events that are separated by 100s of km are being compared (see your Figure 1); this is something that was observed in Aiken and Obara (2021). In their study, unrelated events were clustered together – the events occurred close in time but far in space (see their supplemental – cluster identification). Was there ever a case of an event occurring more than 100 km away in a similar time frame of the selected 100 events (Lines 324-326)? Or is the smallest distance accepted 100 km and also includes events that occur further away? Also, in Lines 228-229: What is distance decay in the ECDF? This is shown in the Table S1 but not presented in the main text. Is this associated with a spatial clustering feature? An explanation of the how the spatial feature is accounted for in this study is needed. Furthermore, in Line 204, it is stated that the result is ‘remarkable’, but it seems that the results indicate that at any time we can have similar magnitude events occurring over large distances. It’s not clear how this tells us something new. The feature that is missing in this analysis is spatial relationship to magnitude clustering.

Aiken, C. and Obara, K. (2021), Data-driven clustering reveals more than 900 small magnitude slow earthquakes and their characteristics, https://doi.org/10.1029/2020GL091764.

- Regarding lines 141-144, the process being described here refers to the process in the previous Xiong et al. paper, which was itself based on work by the Davidsen et al. paper mentioned in that section. These works examined statistically significant deviations in magnitude differences when comparing an event to only the next subsequent event in the catalog. We agree that examining clustering relationships with event pairs considered over a large area such as the in the California catalogs could be affected by events that are separated by large distances, which is part of what this paper aims to address by looking at not just the subsequent event in a large catalog, but actually

examining event pairs and their magnitude correlations in terms of space and time. The discussion in lines 202, 212, and 226 is introducing the idea that with increasing event separation in a sequence of events in a catalog (which, considering the accepted spatiotemporal relationships mainshocks and their aftershocks, would also generally mean increasing time and distance between event pairs), there seems to be a decrease in the amount of magnitude clustering behavior. When only examining event pairs based on their position within a sequence over a large area such as this, it is true that there could be some cases of very similar magnitudes in subsequent events that don't have any actual relation to each other. However, the fact that we observe a significant decay pattern at all signals that magnitude clustering is somehow linked to spatiotemporal patterns of earthquakes (and by extension presumably related to physical mechanisms in the rupture process itself) and we are not just observing similar magnitudes in a sequence by random chance.

Lines 239-241 are mentioned to provide a foundation of previous literature that would naturally lead us to the question of how this change in magnitude clustering behavior occurs with changes in the actual time and distance between events, not just their positioning in a sequence. The sections on time and distance decay address this question, showing that indeed there is a decay in the amount of magnitude correlations seen as you examine two events that are further separated from each other in both time and distance when treated separately. However, we still observed significant correlations at longer time and distance separations than has been shown in previous studies.

We do appreciate the concern that our analysis is actually considering event pairs that are legitimately related to one another spatiotemporally, and recognize that the spatial areas of the catalogs we were allowing were quite large. This is additional justification for the previously mentioned interevent spatial restriction when examining the time decay of magnitude clustering, excluding any event pairs separated by more than 100km. We also implemented an interevent time restriction when examining the distance decay, excluding event pairs separated by more than 150 hours. In both cases, the overall decay patterns that were previously observed remain or are increased. Additionally, we added a section to the manuscript that examines the spatiotemporal decay of magnitude clustering when looking at interevent time and distance together on the same plot to further enable readers to consider the influence of space and time on magnitude clustering.

Regarding lines 228-229, the supplementary table S1 with distance decay included was mentioned in this section before the distance decay analysis was introduced in the main text. Thank you for pointing this out, and discussion of table S1 has been removed from this section. It is mentioned later when discussing the differences in the spatial and temporal decay trends in the distance decay section.

Unclear or unexplained results

It seems the results in Figure 6 are for D and H shown in Figure 4? It's hard to understand what this figure represents exactly following your previous analysis. A percent difference was calculated for each mi and mi+100 along the diagonal shown in Figure 4 (D&H; I presume). So there should be 5 percent differences lines, but only one set along that diagonal is shown? Were the results stacked or something else was done? Why not show all?

- Figure 4 D and H are looking at the percent difference in event pairs that display magnitude clustering (which is given by the event pairs that fall along the diagonal line in the ECDF heat map) when only looking at pairs of events in the catalog that were separated by 100 events within the time-ordered sequence. In figure 6, we have calculated interevent separation times for each event pair, and we examine event pairs separated by specific interevent time intervals, regardless of how separated those events are in a sequence in terms of event number. Note that for each time interval (for example, 0-3 hours interevent separation, or perhaps 130-133 hours separation, etc..) we are considering all event pairs that are separated by that interval of time, not just from event n to event n+1 or n to n+100. This way we can directly examine how the strength of magnitude correlations evolves with increasing time between two earthquake events. We do the same for interevent distance (in 5 km intervals) so we can examine how it evolves with increasing distance between two events. Prior studies had mainly focused on looking at comparisons between one event and the next event (n+1), so our study expanded the range of subsequent events any given event could be compared to.
- We chose to average the percent difference for each of the 5 bins that fall along the line of magnitude clustering, to plot one value representing the strength of magnitude clustering across the full range of magnitudes in the catalog, rather than plotting magnitude ranges separately. We believe this is more representative of the full catalog by incorporating comparisons at each of the 5 magnitude tiers, but we also sought to address the varying amounts of magnitude clustering seen in those different bins in the “Variations in Time and Distance Decay Patterns for Different Magnitude Ranges” section at the end of the paper. That section shows that the decay patterns are present regardless of the magnitude range you specify, but the strength of the magnitude clustering signature does vary within the different magnitude ranges.

In Line 336, it is stated that there is a “linear trend” after 20 km for Southern California and after 5 km for Northern California, but the fast drop in short distance (in Figure 7) is not explained.

- We have reviewed the literature seeking an explanation for what could explain a faster drop in the magnitude clustering at short distances, mainly whether static vs. dynamic stress factors could play a role in this, but found no consensus. Moreover, after the updates to the spatial decay plots based on removing the n<=100 restriction, the decay difference in the shorter distances compared to the rest of the distance ranges is less pronounced, and it seems it does not occur at all in the Southern California catalog. The overall trend of the spatial decay for both catalogs appears to be linear, and at this point in time we don’t think there are definite variations from this trend that could not be attributed to statistical variability in the dataset.

In Lines 340-341: “linear density of aftershocks…” I suspect that this is true for short distances and larger magnitudes (where static stress is important) but linear at longer distances (where dynamic stress is more important). Perhaps more discussion of this physical aspect can be added and linked to the statistic observations? This links to a previous comment about the manuscript needing more physical interpretation.

- After further review of the literature regarding the relative importance of static and dynamic stress at different distances, we have concluded that there is not a consensus on the topic, and there are many studies with compelling arguments for the importance of one over the other. There does seem to be a consensus in the literature, however, that power-law decay of aftershocks is a main feature of the spatial clustering of seismicity. The wording of this section has been updated to provide more clarity on our main points.

In Line 469: “logarithmic decay”. There seems to be a contradiction between what is shown in Figure 10 A & B and what is said about the relationship between PD and time. For example, in Lines 286-287, the relationship is described as a rapid decrease at short time intervals and that it is more gradual at longer time intervals. Indeed, in Figure 6, it is mostly linear and “gradually decreases” outside of the short time intervals. Even in Figure 10, the plot is log-log, in which the data does appear to be *mostly* linear, but it is clear that there are tails deviating from the linear relationship for the 0-20% and 20-80% groups. So there seems to be some contradiction unless I am missing something?

- We appreciate the reviewer's concern regarding the deviations of the data points in these plots. After rerunning these plots with our updated datasets imposing interevent time and distance restrictions instead of n<=100, the resulting decay curves are a better fit to the respective decay curve types, and the issue with a deviating tail in the data is much less apparent. We have also added a least squares regression fit to the data with a 95% confidence interval.

Minor comments

Line 57: It is not clear to me what seismic modeling is.

- We changed the text from “using seismic modeling” to “by comparing with popular statistical models of seismicity”.

Line 67-69: Magnitude clustering is presented as the topic of this article, but it has not been defined what exactly magnitude clustering is. Is it events of similar magnitude occurring near each other or events of different magnitude occurring near each other or something else? How close in magnitude must they be? Exactly the same, difference of .1 or something else?

- We added the following statement to the introduction: “We define magnitude clustering as statistically significant correlations between magnitudes of earthquakes in a given region and time period, beyond random occurrence and other spatiotemporal relationships such as the Gutenberg-Richter frequency-magnitude distribution”

Line 77: This is very general; too general for the topic. It would be helpful to the reader if there is an expansion on what is not understood about seismogenesis and fault interactions. For instance, we still don't know how slow slip is related to seismogenesis, but in this study, I think it is more about aftershock sequence interactions?

- This sentence has been removed for reasons discussed in our answer above about limiting discussion about physical models.

Line 88 and 97: Xiong et al. missing year (ignore if this is an acceptable style of reference).

- The year has been added to the reference.

Line 124: “summarizing the steps of the analysis”; this phrase is awkward.

- This has been deleted.

Line 126-128: Need to specify that the catalogs are processed separately and that this observation is for the North California catalog.

- We added a statement specifying “After performing this processing on the Northern California Catalog,”

Line 176-179: This sentence reads like it belongs in a figure caption and not the main text.

- We have improved the wording in the main text.

Line 241: I think you are trying to reference # 5, but it looks like km^5

- We switched our reference format in the main text from superscripts to this format required by Seismica for the final manuscript: [Xiong et al., 2023].

Line 386-387: It's scaling issue?

- We updated plots with a 2mm interevent separation interval appear similar to the field distance decay plots. The 1mm event separation was too small given the detection limits of the lab experimental setup.

Line 433-434: Missing some words to make this a complete sentence.

- We changed the sentence to “Extending the emissions recording period to hours, or days, after the experiment ends is necessary for future laboratory analysis of magnitude clustering trends.”

Line 498: It’s unclear how the method is “novel.” It appears that this work is just an extension of a previous work (Xiong et al.).

- We changed the text to simply say “We examined seismic magnitude clustering…”

Figure feedback

Figure 2. I don't understand what the difference colors in A and B mean. It looks like magnitudes but magnitudes of what? the subsequent event or the triggering event? Or is it different magnitudes of completeness? Or differenced magnitude bins?

- The different colors correspond to different magnitude of completeness thresholds. This is to show that even when raising the completeness threshold above what we have determined to be the actual magnitude of completeness for the specific catalog, statistically significant magnitude clustering is still observed. Raising this threshold too much leads to an insufficient number of events that can be used for a reliable statistical analysis, so our largest completeness threshold examined is 2.3, which is still considerably larger than the determined magnitude of completeness of 1.4 for the Northern California catalog.

Are the filters the 2-3minute differencing and the STAI factors?

- Yes, and we have updated the text to make this clearer.

Review Reports, Round 2

Reviewer A Comments

For author and editor

The authors have taken great care addressing the points raised in my previous report. From my perspective, there are two issues that need to be addressed before accepting the paper.

1) l.240: I am still not sure how the averages are exactly calculated. It also seems that the right hand side of the equation can only become zero if the average N equals N (and N is stated to be the TOTAL number of events!). Why would that correspond to the "uncorrelated" case? I do not see it. Given that, I am also currently not able to assess the correctness of the bootstrapping used to establish uncertainties.
Since this equation is fundamental to the majority of the presented analysis (the vast majority of figures build on it) it is crucial to explain in detail how the different quantities in that equation are calculated, even if that means to be a bit pedestrian.

2) ETAS catalog with incompleteness added by artificially removing smaller magnitude events from the catalog (Supplementary Figure S6): I am very surprised that this did not give rise to any magnitude correlations. If one uses the Helmstetter equation (l. 138) to remove events from an ETAS catalog that contains large events, an analysis similar to what is shown in Fig. 2A (without using any filters to correct for STAI etc.) DOES give rise to magnitude correlations. I suggest the authors clarify exactly what they did.

Minor:
i) Since the authors have tried the Hainzl (2016) approach, I would suggest adding the corresponding analysis to the supplemental material.

Reviewer C Comments

For author and editor

This study aims to investigate and quantify possible clustering in earthquake magnitude, considering the distance in space and time between events in two dense seismic catalogues. The results are publishable in Seismica if the authors can address the following recommendations:

Major and minor revisions:

1. I appreciate the details and clear response of the authors regarding previous comments from reviewer A and B. A common issue, which I also identified is related to the methodology. In the current version of the manuscript, the problem in understanding the method remains.

L187-L189 referenced a novel method introduced in the Xiong et al (2023) paper; the ECFD method is not really detailed in the referenced paper so I suggest considering to include in a supplementary material a detailed description of the methodology.

2. I tried running the code to better understand the method and the calculations made. After reading the README file and carefully following the instructions, I can say that I was not successful in running all scripts. It would have been helpful if the authors added the catalogue/original input data in the repository so I could confirm the results. After creating my own catalogue to match the input data, I also write the"toepoch.csh" script (is missing from the archive). Another issue was the "fmd.csh" script where errors appear (under Ubuntu22.04) I think the authors should revisit the scripts and make sure that their analysis can be replicated by a regular user of Unix system.

3. As one reviewer pointed out, the definition of magnitude clustering is not clear. As I read the definition in the updated manuscript (L69-72), it is still rather vague what magnitude clustering actually represents. As the authors mentioned in the introduction, there is a debate regarding magnitude clustering. A clearer definition is required for the readers to understand the proposed correlations.

4. L102-105 For me it is not clear if the mentioned model is proposed by the authors or by the Xiong study or it comes from different studies? A reference would be required in this case.

5. Following the methodology proposed in Davidsen and Green (2011), the authors find significant deviations from 0 in the magnitude differences for successive events. This result contradicts the conclusion of the above-mentioned study. Thus, I consider that the interpretation provided at L164-L167 is not substantial and needs further discussion to be understood by the reader.

6. L197 – L203 As I could not replicate the analysis, the meaning of Figure 3 and Figure 4, which represent the results of this study, are hard to understand for me. What exactly is the significance of the difference relative to the expected mean and why this is divided by 25?

7. In Figure 6 two methods are used to investigate the correlations in magnitudes, considering the time difference between event pairs occurring during a 150 h window. The results are consistent between the two methods. The ECDF method is used to further analyse the distance decay. L 341-344 refers to direct comparison between the two field catalogues. I wonder why such a

comparison was not made for the time decay? Is the magnitude correlation in respect to time decay the same for both regions? Figure 8 could include an inset showing the differences (since the grid is the same for both catalogues).

8. Figure 10 shows for the x-axis either a Time Difference or a Time or Distance. I don't understand how the distinction is made between Time Difference and Time, respectively Distance and Distance difference?

9. L526-528 In Figure 11 C. and D. we observe a clear variation in slope of the spacial decay for the three categories of magnitude ranges. I consider that a discussion is need to understand the implications of this variations.

Dear Dr. Llenos,

Thank you for another round of helpful reviews for our manuscript to Seismic entitled "The Pattern of Earthquake Magnitude Clustering Based on Interevent Distance and Time".

We have made additional revisions to the manuscript and code to help ensure a better experience for readers. We have detailed our revisions in the point-by-point response to reviewer comments below.

Best regards,
Derreck Gossett
Mike Brudzinski
Qiquan Xiong
Jesse Hampton

Dear Derreck Gossett, Dr. Michael R. Brudzinski, Dr. Qiquan Xiong, Dr. Jesse C. Hampton:

I hope this email finds you well. I have reached a decision regarding your submission to Seismica, "The Pattern of Earthquake Magnitude Clustering Based on Interevent Distance and Time". Thank you once again for submitting your work to Seismica.

I have received two more peer-review reports for your manuscripts. Both reviewers still agree that the manuscript describes an interesting and worthwhile topic, but request a few more details before it can be acceptable for publication. In their comments, both reviewers note places where the method(s) and interpretation could be clarified and supported with a bit more detail. Reviewer C also notes that they had some difficulty in running the provided code, and so that should also be checked. Based on the reviews I have received, your manuscript may be suitable for publication after some revisions.

The comments submitted in the webform by Reviewers A and C can be found below.

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.

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,
Andrea Llenos
U.S. Geological Survey
andrea.llenos@seismica.org

Reviewer A:

The authors have taken great care addressing the points raised in my previous report. From my perspective, there are two issues that need to be addressed before accepting the paper.

1) l.240: I am still not sure how the averages are exactly calculated. It also seems that the right hand side of the equation can only become zero if the average N equals N (and N is stated to be the TOTAL number of events!). Why would that correspond to the "uncorrelated" case? I do not see it. Given that, I am also currently not able to assess the correctness of the bootstrapping used to establish uncertainties.
Since this equation is fundamental to the majority of the presented analysis (the vast majority of figures build on it) it is crucial to explain in detail how the different quantities in that equation are calculated, even if that means to be a bit pedestrian.

Thank you for the suggestion, we agree that the equation could be confusing to readers in its current form, and have altered the equation and accompanying term descriptions to make them more clear to the reader. The equation itself is simply a formula for percent difference, and we believe this new version of the equation makes that a lot clearer and easier to understand:

$$\overline{PD}_{similar} = \frac{\overline{N}_{similar} - \overline{N}_{all\ bins}}{\overline{N}_{all\ bins}}$$

We have changed the subscript names to be simpler for the reader to quickly understand. The $\overline{N}_{similar}$ term is the average number of event pairs that fall into the bins of similar magnitude comparisons along the diagonal line in the ECDF plots. The $\overline{N}_{all\ bins}$ term is simply the average number of event pairs in a bin based on consideration of all bins in the plot. Therefore, the

$\overline{PD}_{similar}$ term is the percent difference between these two averages. The higher the number of the $\overline{PD}_{similar}$ term, the more event pairs there are that are correlated in magnitude compared to the total event pair comparisons.

2) ETAS catalog with incompleteness added by artificially removing smaller magnitude events from the catalog (Supplementary Figure S6): I am very surprised that this did not give rise to any magnitude correlations. If one uses the Helmstetter equation (l. 138) to remove events from an ETAS catalog that contains large events, an analysis similar to what is shown in Fig. 2A (without using any filters to correct for STAI etc.) DOES give rise to magnitude correlations. I suggest the authors clarify exactly what they did.

Our original manuscript created incompleteness by artificially removing events from the catalog at random times based on the proportion that we have sought to remove across the different magnitudes. This was designed to replicate overall detection threshold catalog incompleteness, which is different than trying to force STAI into an ETAS catalog. In this revision, we have sought to employ the reviewer's suggestion by using the Helmsetter equation to identify a time-varying level of completeness and then we remove events from the ETAS catalog with a progressively higher percentage across a 1.0 magnitude level below the Helmsetter calculated completeness value. In contrast to the reviewers expectation, employing this incompleteness to the ETAS catalog does not produce magnitude clustering, as shown in supplementary Figure S9. Our regular correction for STAI would remove all events below the Helmsetter value, but employing the reviewer's suggestion demonstrates that incompleteness below the threshold does not artificially produce magnitude clustering. We have clarified in the main text that multiple forms of incompleteness have been investigated and added more extensive text and a figure to the supplement regarding the STAI version of ETAS incompleteness.

Minor:
i) Since the authors have tried the Hainzl (2016) approach, I would suggest adding the corresponding analysis to the supplemental material.

We have added a description of the Hainzl method and the results of our ECDF analysis when filtering the catalog based on this method to the supplementary material. The associated figures show that after filtering out events that are incomplete based on Hainzl's definition, the amount of magnitude clustering observed using our ECDF method remains unchanged.

Reviewer C:

This study aims to investigate and quantify possible clustering in earthquake magnitude, considering the distance in space and time between events in two dense seismic catalogues. The results are publishable in Seismica if the authors can address the following recommendations:

Major and minor revisions:

1. I appreciate the details and clear response of the authors regarding previous comments from reviewer A and B. A common issue, which I also identified is related to the methodology. In the current version of the manuscript, the problem in understanding the method remains. L187-L189 referenced a novel method introduced in the Xiong et al (2023) paper; the ECFD method is not really detailed in the referenced paper so I suggest considering to include in a supplementary material a detailed description of the methodology.

We thank the reviewer for their suggestion, and have included a detailed description of the ECDF method in the supplementary material.

2. I tried running the code to better understand the method and the calculations made. After reading the README file and carefully following the instructions, I can say that I was not successful in running all scripts. It would have been helpful if the authors added the catalogue/original input data in the repository so I could confirm the results. After creating my own catalogue to match the input data, I also write the“toepoch.csh” script (is missing from the archive). Another issue was the “fmd.csh” script where errors appear (under Ubuntu22.04) I think the authors should revisit the scripts and make sure that their analysis can be replicated by a regular user of Unix system.

We have updated the scripts in the Zenodo repository. These updated scripts are based on the successful testing of a colleague with coding knowledge but no prior experience with our particular code or research.

3. As one reviewer pointed out, the definition of magnitude clustering is not clear. As I read the definition in the updated manuscript (L69-72), it is still rather vague what magnitude clustering actually represents. As the authors mentioned in the introduction, there is a debate regarding magnitude clustering. A clearer definition is required for the readers to understand the proposed correlations.

To make the definition more clear for the reader, we have added more specific text to the introduction section where we define magnitude clustering, clarifying that the magnitude difference between two earthquakes is smaller than would be expected from the Gutenberg-Richter frequency-magnitude distribution, based on a large number of event comparisons.

4. L102-105 For me it is not clear if the mentioned model is proposed by the authors or by the Xiong study or it comes from different studies? A reference would be required in this case.

This model was proposed by the authors in the Xiong (2023) study, and we have clarified this in the text.

5. Following the methodology proposed in Davidsen and Green (2011), the authors find significant deviations from 0 in the magnitude differences for successive events. This result contradicts the conclusion of the above-mentioned study. Thus, I consider that the interpretation provided at L164-L167 is not substantial and needs further discussion to be understood by the reader.

We understand the reviewer's concern and have elaborated on the contradictions between our findings and those of Davidsen and Green (2011). The main points of our findings in their rebuttal of Davidsen's findings are as follows:

- We determined a lower magnitude of completeness than Davidsen, taking care to use multiple well-established methods of determining the Mc value based on the frequency-magnitude distribution (Maximum curvature and b-value stability methods). We chose the higher value found among these methods (b-value stability) to ensure we were being conservative in our choice of Mc. After filtering for STAI in the same way as Davidsen, we find that magnitude clustering deviations are still observed. One disadvantage of having too high of a completeness value is that it doesn't leave the catalog with enough event comparisons to establish a reliable statistical relationship at the accepted levels of standard deviation, so keeping more events in our catalog increases the accuracy of these event comparison statistics.
- Furthermore, using a different method (Hainzl, 2016, now added to supplementary material) that establishes a varying Mc value based on the earthquake rate at different times instead of a static Mc value for the length of the catalog, we still observed magnitude clustering in our ECDF analysis after filtering for incompleteness and STAI in this way.
- Finally, our ECDF method has a distinct advantage over the cumulative distribution method used by Davidsen and by us earlier in our paper, in that it is able to show there are significant magnitude correlations across the full range of magnitudes in a catalog, not just for the smaller event comparisons which would be those most likely to be affected by catalog incompleteness.

We have added text to lines 169-177 and lines 222-225 that discuss these differences in our results.

6. L197 – L203 As I could not replicate the analysis, the meaning of Figure 3 and Figure 4, which represent the results of this study, are hard to understand for me. What exactly is the significance of the difference relative to the expected mean and why this is divided by 25?

We believe that the detailed description of the ECDF that is now in the supplementary material will help with the understanding of these figures, and we have also updated and clarified the wording in the figures themselves.

7. In Figure 6 two methods are used to investigate the correlations in magnitudes, considering the time difference between event pairs occurring during a 150 h window. The results are consistent between the two methods. The ECDF method is used to further analyse the distance decay. L 341-344 refers to direct comparison between the two field catalogues. I wonder why such a comparison was not made for the time decay? Is the magnitude correlation in respect to time decay the same for both regions? Figure 8 could include an inset showing the differences (since the grid is the same for both catalogues).

The time decay of magnitude clustering is very similar for both catalogs across both time decay methods examined, with some slight differences in both the amount of clustering and the pattern of decay. We have added text to the Time Decay section that discusses the similarities and differences of the two catalogs in the time decay represented by both the ECDF and autocorrelation methodologies.

8. Figure 10 shows for the x-axis either a Time Difference or a Time or Distance. I don't understand how the distinction is made between Time Difference and Time, respectively Distance and Distance difference?

This was simply an inconsistency in labeling between the ECDF and autocorrelation plots. The plots have been updated to remain consistent with each other using the labels "Time Difference" and "Distance Difference".

9. L526-528 In Figure 11 C. and D. we observe a clear variation in slope of the spacial decay for the three categories of magnitude ranges. I consider that a discussion is need to understand the implications of this variations.

The difference in the steepness of the distance decay slopes can likely be explained mathematically by observing the initial strength of the magnitude clustering value for each of the different magnitude ranges. There is a clearly similar linear decay for each of the magnitude ranges. However, since the initial strength of the magnitude clustering value in the highest magnitude range is much larger compared to the other two ranges, this would naturally lead to a steeper decay slope mathematically if it decays in a similar fashion to other two ranges. To decay to a similar value of magnitude clustering observations at a distance where we believe the distance is likely too large for magnitudes to significantly cluster, it naturally must decay at a steeper slope from its initially higher value. The time decay slopes don't show this difference due to being presented in a log-log space.

This explanation has also been added to this section of the manuscript.