Reviewer #1:

Manuscript entitled „Interpreting the cause of bound earthquakes at underground injection experiments" by Ryan Schultz, Linus Villiger, Valentin Gischig, and Stefan Wiemer is well designed with clearly presented scope and reasoning. It deals with important topic of the determination of maximum magnitude for injection experiments, which pose important insight into physics of the earthquakes induced by fluid injection and further any other seismicity related to fluid-rock interactions. Methods are clearly described as well as data used for the estimations. Reasoning is documented well with the former works of various authors. I have one major critical comment related with methodology and some minor comments related with the literature review and technical.

We thank the reviewer for their comments and positive outlook on our paper! We are more than happy to address the major/minor critiques of the reviewer.

Major issue, which may need some explanation is sensitivity of the CAP test to magnitude range. There is 2-3 magnitude unit span between the smallest and the largest events and even smaller when we consider completeness. I would like to see any discussion about the magnitude range on the CAP tests efficiency in cases used here. Authors only discuss the role of the events number in datasets suitable for the tests.

For this point, we haven't provided much for sensitivity tests in this paper, since this concept has already been extensively tested on synthetic data. The largest factor contributing to the resolvability of M_{MAX} *is the degree-of-truncation. We define this term and point to its importance on Lines 169-186.*

We note that the span between smallest/largest magnitudes is not a significant factor for determining M_{MAX}*. Sometimes, having more small events can actually be detrimental to resolving* M_{MAX}*, since these events don't sense the GR-MFD roll-off very well.*

That said, we have added a citation to a newer paper [Schultz, 2026] that also demonstrates these points methodologically – to try and be clearer for future readers who have similar questions.

In the Introduction authors refer to different maximum magnitude estimation methods, however not mentioning any Bayesian methods (Kijko, 2025) or methods dealing with small catalogs or

incomplete catalogs (eg. Kijko et al., 2021, Vermuelen and Kijko (2017)). I think, that taking into account above works may be informative for reader interested in dealing with seismic catalogs with narrow magnitude range and/or small event number.

We thank the reviewer to pointing us towards these works! We have now added a citation to the 2025 paper on Line 72.

Minor technical remarks:

Line 78 and below: Acronyms such as CAP, KS, MLE and EW should be explained as they are introduced.

Similar to the comment from Reviewer #2, we have defined the KS, MLE, & EW acronyms at this point in the paper now.

Line 868: All the symbols from equation should be explained here again. Some are introduced earlier (but not all), and it may be hard to follow for the reader.

Fair point, we have added a sentence after these equations (Lines 893-895) that reminds the reader of all these terms.

References:

Kijko, A., Vermeulen, PJ., Smit, A. (2021) Estimation Techniques for Seismic Recurrence Parameters for Incomplete Catalogues SURVEYS IN GEOPHYSICS Vol.43 Issue 2 pp. 597-617, DOI:10.1007/s10712-021-09672-2

Kijko A., (2025) Bayesian Assessment of the Maximum Possible Earthquake Magnitude mmax. JOURNAL OF THE GEOLOGICAL SOCIETY OF INDIA. Volume 101 Issue 6 Page764-769 DOI: 10.17491/jgsi/2025/174157

Vermuelen, P., Kijko, A. (2017) More statistical tools for maximum possible earthquake magnitude estimation. Acta Geophysica 65(4), pp.579-587. DOI10.1007/s11600-017-0048-3

We would like the thank the reviewer again for their comments. They are greatly appreciated.

Reviewer #2:

In this manuscript, the authors extend their study on the statistical characterization of induced seismic sequences with respect to the expected maximum magnitude from field-scale observations (Schultz et al, 2025) to underground laboratory experiments. A reliable estimation of the maximum expected magnitude during EGS and other subsurface operations, which goes beyond a static traffic light system, is paramount operations operations to be conducted safely.

Mine-scale experiments have the advantage that their environment is much more controlled and findings from field-scale (Schultz et al, 2025) can be tested with higher confidence due to improved monitoring setups and other comprehensive datasets. That said, the extension of their research goes beyond the mere application of the existing method to new test cases.

The work is clearly structured from simple (HF-dominated) to complex (mixed-mode) cases.

While the authors clearly draw the difference between bound and unbound in regard to limited fracture length and the reactivation of preexisting faults, the authors also discuss when this simple binary reasoning might be wrong. Their generalization of the V^n model and the Vn-EW test provides a novel approach to the study of underlying processes in seismic sequences beyond the analysis of induced cases alone.

We would like to thank Peter for his review. It is clear that the reviewer has put significant care and thought into their critique, which is very much appreciated. Also, having the reviewer's prior experience/insights for the Äspö stimulations is an asset for this review!

Given these thoughtful critiques, we have endeavoured to handle them with a similar level of care.

The authors concisely describe their statistical tests and the workflow, allowing it to be applied to other sequences by other researchers. I have to assume that the github link with data and code will be available upon publication of the article, as they did for the previous paper introducing the CAP test.

The reviewer is correct here. Once the paper is accepted for publication, we will be happy to add the codes to GitHub so that other researchers can replicate this work or apply it to their cases.

General minor remarks/questions/suggestions

1. I appreciate that the authors state that they are careful in estimating Mc and set it in a rather conservative way, even when it reduces the number of events to the point of hindering the unambiguous classification of bound sequences. That said, I suggest adding Mc in Table 1-3 for completeness and for the reproducibility of the subset of events used in each analysis.

We agree with the reviewer on this point. We have added the optimal Mc choice for all the cases to Tables 1-3 now.

Note that the GitHub repository will provide all the information/data required to fully reproduce the results. This includes all the tables and figures in this paper. Having access to the optimal Mc choices will also be a part of the code release.

2. 246f: "Testing on both synthetic and real datasets suggests that the MLE-test is sensitive to quantifying MMAX within a hundredth of a magnitude unit when MLRG-MMAX discrepancies are -0.5 M or better." Are magnitude uncertainties considered in these tests? For synthetic datasets, the precision might be within a hundredth of a magnitude, but for real cases, inherent magnitude uncertainties will not allow for such precision.

This is an astute point raised by the reviewer. Yes, we have considered magnitude uncertainties via perturbations/dithering of the magnitudes during the MLE-tests (and other tests).

We perform 50 bootstrapped MLE tests, where each of the catalogues have their magnitudes perturbed by ±0.1. In each of these 50 bootstrap trials, the catalogue is also reshuffled 100 times. This point is mentioned fairly extensively in the manuscript, so we haven't modified the text in response here.

As a side point, for this study, we largely use this error as a sign that the MLE-test is confident in its result, rather than actually believing the fitted value. Essentially, we agree with the reviewer that real cases will certainly be trickier than synthetic cases!

3. Despite being discussed in the methods section, no sequence is assigned to the 'unresolved' class in the tables. Did all sequences pass the resolution checks?

Thanks to the reviewer for raising this point. Admittedly, we could have been clearer here.

The tables provide the mechanism that is favoured, while the colour indicates the confidence. The CAP-tests may indicate a bound process, but this can be low confidence.

We have revised the tables to be more in line with this interpretation (i.e., explicitly stating indeterminate instead of bound/unbound). We have also amended a sentence on Line 666 to be explicit on this point.

4. The generalization of Vn models and the integration into the EW analysis is a very interesting approach to further characterize bound sequences. However, I see some discrepancies between the Mmax model choice in the result tables and the Vn-EW analysis that might be worth discussing.

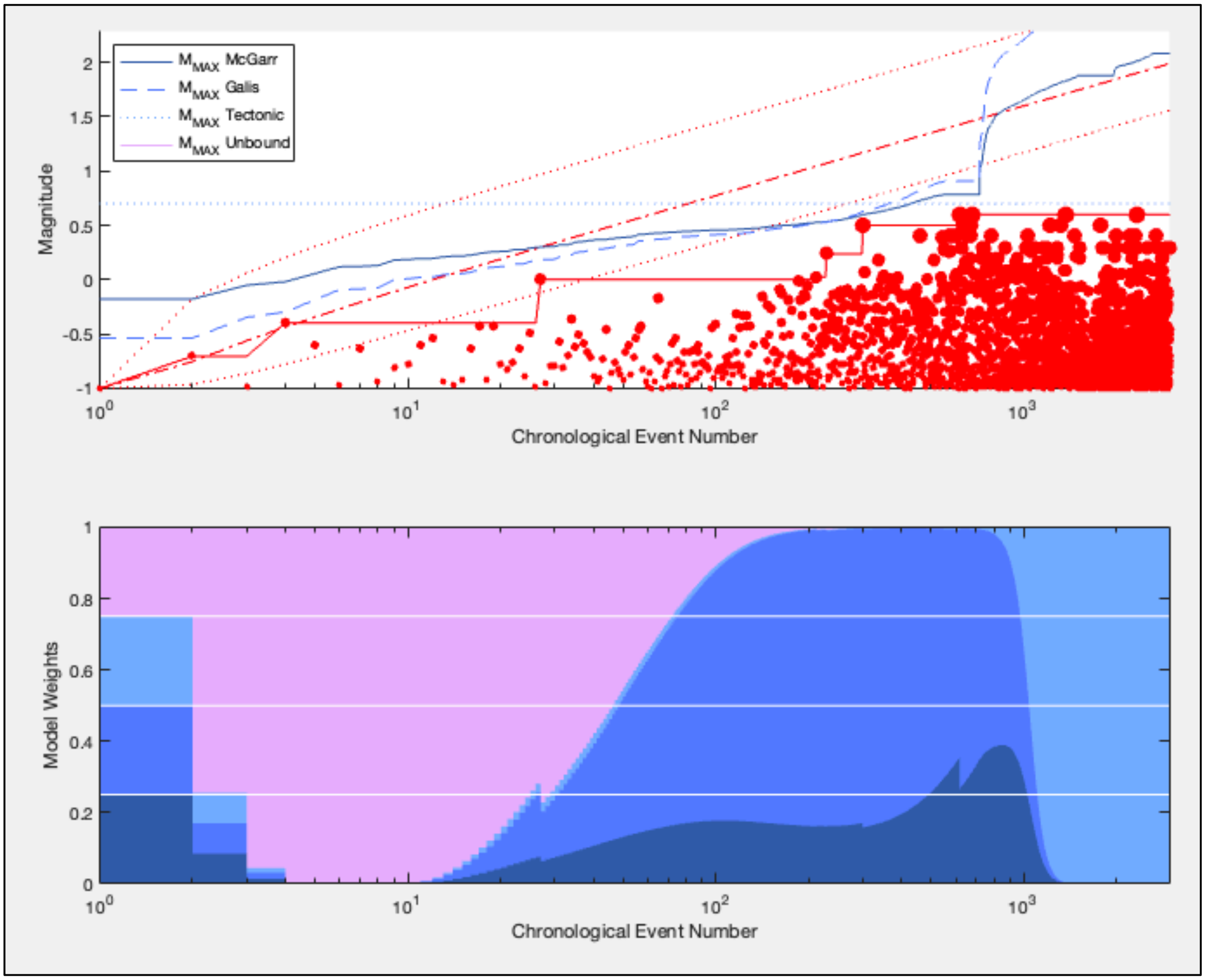
We thank the reviewer for identifying these discussion-worthy points. See our responses to the two related sub-points below.

4a. Aspo HF3 and SURF N164 fall into the unknown (X-like) category in the vn analysis. Obviously, these two cases were in a different model class before, since X-like is not part of the models in the standard EW test. While Aspo HF3 ended up in the next closest category, SURF N164 was found to be in between McGarr and Tectonic (the models indicated in the result table). Both make perfect sense.

Agreed! We have added two sentences on Lines 854-857 that mentions this point now.

4b. However, PNRC 2w (Schultz et al 2025) is assigned to McGarr or Galis in the result table and should therefore fall into one of those categories in the Vn-EW analysis, yet it peaked in tectonic. Is this a different subset of data? How do the authors explain the difference?

The reviewer is correct, for this case it shows up as tectonic bound here rather than McGarr/Galis-like. Partly this is because in the previous paper, we did not test for a tectonic upper limit – in this sense, there was a better fitting model that was missing from the ensemble. See below for the EW-test applied to this case. The Galis/McGarr-like models are initially favoured, until later when the tectonic model wins out.



The reasons for this aren't perfectly clear at the moment. On the one hand, this could be accounted for by an unaccounted for a pre-existing fracture network. This mean that the Mmax curve would be flatter, with growth really only appearing later in the sequence. Effectively that observational bias that is discussed on Lines 847-854, just restated in a slightly different way.

My personal opinions on this are that it can still be explained by this observational bias, although I'm not completely certain quite yet. Disentangling this effect is going to take dedicated efforts to better understand the EW-test and its sensitivity to various forms of V^n relationships.

We now mention much of these points on Lines 854-857.

I am looking forward to seeing Vn-EW tests for other seismic sequences in future research, as these may shed light onto the processes driving them.

We thank the reviewer for their kind words here!

Some remarks on HF6 of the Aspo dataset:

The damping coefficient used in the magnitude calculation was estimated as a uniform value across the entire rock volume. As correctly stated by the authors, most events were in fact induced during stages HF1 and HF2. The larger number of events in this domain could have biased the uniform estimation of the damping coefficient. As a consequence, the deviating rock type around HF6 might not be well represented by the uniform coefficient, which may lead to biased magnitudes that influence the statistical analysis.

Unlike all other stages in which the volumes of substages were ramped up over time or kept stable, the first substage of injection in HF6 was the substage with the largest injected volume. The volume was reduced afterwards, resulting in an early occurrence of Mmax even during the injection. For all other stages the largest magnitude was induced after the shut-in.

Additionally, the last three substages of HF6 were pumped on another day. This might have allowed some relaxation and flow back over hours after the largest substage, HF6-1, altering the response of the reservoir. Unfortunately, the flowback for HF4-HF6 was not rigorously monitored.

This is fantastic information! We now share these details with the reader on Lines 452-455.

Additional minor points:

84f: The abbreviations of the statistical tests should be introduced, here, when first mentioned.

This has been added.

107: "This is significant"

I have added the "is" back into the sentence now.

330: "catalogues between 10^2-10^3 events (above Mc)"

Corrected.

374: "HF1, HF2, HF4 & HF6"

Very nice catch! This has been corrected.

727f: Repeating that the "analogous stages" were not aiming for new fracture creation would help to emphasize the difference between the experiments.

We've added a small.

798: What are sd and k in the time-varying fracture radii equations for shear and tensile cracks?

These are coefficients related to the propagation of these fractures. We now explicitly mention this point, right after these equations are introduced.

822: "stages 1-3"

Corrected.

824: "for cluster 1 (stages 1-2) and cluster 2 (stages 3–6) as defined in Schultz et al, 2025"

Modified, as suggested.

880: "or , the extent of asperities"

Added the missing 'the' into the sentence.

866ff: I suggest to add the categories in front of the equations for clarity

V^n: …

Tensile: …

Shear: …

Agreed. These categories have been added now.

868ff: c, sd and k not formally introduced, see also line 798

These terms are formally introduced on Lines 815-816 now. Similar to Reviewer #1's comment, they are also re-mentioned following the equations.

881: “between seismic asperities”

This has been corrected.

Figure 1: I suggest adding the resulting classification at the end of the arrows (bound, unbound, not resolved).

Agreed! The bound/unbound labels have been added.

Figure2 : Increasing the unbound/bound labels on top of the figure would help to quickly spot which is which.

The font size has been increased here.

Figure 4:

1. **The labels got mixed up. HF3 instead of HF1 for the green cluster; Last label in Fig. 4a should be HF6. There was no microseismic activity during HF5.**

We thank the reviewer for catching these typos. They have been fixed.

2. **The injection rate is mostly hidden behind the seismic events, see also Fig. 7 and 10. Since the stages are analyzed individually in this study, the time period between the stages is not important, so each stage could get its own subplot to improve the visibility.**
3. **The map is distorted. Unequal scale of easting and northing (possibly a layout choice?), see Fig. 7**

For these points, we would prefer to keep the plots simpler and fewer. They are largely serving to give readers a very brief overview of the experiments, rather than give full details.

Our reasons are two-fold. First, the provided references (and online data) can give greater detail as to the specifics of the stimulations/catalogues. Second, the paper is already quite long, the figures are large. We would prefer to not increase the number of figures/panels in this paper.

We hope that this is agreeable/understandable.

Figure 7: The discussion regarding GTS HS4 and GTS HF2 is based on geometric considerations of the microseismicity. I suggest including subfigures that zoom into these two stages due to their complexity.

On this point, similar to the ones above, we have respectfully have declined the reviewer's request. These stages are examined in greater detail in prior papers [Villiger et al., 2020; 2021], which we cite. Readers who are curious about the details of those two sequences can find them there.

Furthermore, this plot can be recreated with our "F10.m" script on GitHub, so those readers will also be able to examine the details of these sequences there.

Figure 10: Flow rate axis label missing
Thanks for this attention to detail. The flow rates have been added to the y-axis now.

Figure 11: Labels a and b should be added in the caption
We now mention that these two panels refer to either hydroshearing or hydrofracturing experiments.

We appreciate the reviewer being so diligent in their review. Thanks again!