Author Response to Referee Comment 1

We would like to thank the referee for their careful consideration of our manuscript. We divide our response to the referee under sections for the major- and minor comments made, with the comments made by the referee indicated in italics.

1 Major Comments

Our response to the major comments made by the referee is detailed below.

1.1 Literature Review

A first and major concern lies in the perplexing incompleteness of the bibliography. The reference list remains very narrow and omits several first-order contributions to both mass-wasting and debris-flow seismology, as well as recent methodological developments in anomaly detection and machine learning applied to seismology (for mass-movement, but also for glaciers, volcanoes, earthquakes etc). Without these references, the manuscript, which is, first and foremost, a methodological paper, does not convincingly situate its contribution in the broader research landscape.

In a revised manuscript we will expand the literature review by including more recent references to applications of machine learning in seismology and anomaly detection including, but not limited to, [1, 2]. As suggested by Referee #3, we will aim to prioritize studies that are most directly relevant to the content of the manuscript, rather than providing an exhaustive overview.

1.2 Methodology

Beyond this, the exposition of the methodology is at times poorly balanced. Sections presenting detailed mathematical derivations of the methodology, such as the full formalism of the isolation forest, could be more appropriately relegated to appendices. In the main text, a more didactic explanation would be far more valuable. It would make the methodological contribution both clearer and more accessible to the readership. At present, the paper tends to emphasize equations at the expense of interpretability and understandability.

This section will be revised to make it more readable and accessible. Less emphasis will be placed on mathematical detail and more effort will be invested to illustrate the principles underlying the methods. For example, we will include visualizations on why anomalies tend to isolate into terminal nodes at shallow depths, how waveforms traverse trees, and how DTW is performed between waveforms.

1.3 Figures and Visualizations

In addition the manuscript currently places an excessive amount of important content in the appendices, including crucial figures (e.g., D2 or D3, E1, E3) showing seismic signals and examples of detected events. These visual results are central to a seismological study and should appear in the main body of the paper.

Figures and visualization more broadly need to be improved. Beyond the introductory material, the reader is given few direct visualizations of the detection process or of anomaly scores. Examples of time series with IF anomaly scores, accompanied by a schematic representation of the full workflow, would make the study more intuitive and strengthen its appeal for a seismological audience.

In a revised manuscript, examples of time series plots such as those displayed in Figs D2 - D4, waveform-spectrogram plots (Figs E1, E3) will be moved to the main body of the paper. As suggested by the referee, clusters of segments will be illustrated by representative examples, rather than dendrograms.

In addition, the semi-supervised and supervised exploration procedures of Sections 3.1 and 3.2 will be moved to the methodology section, and workflows will be illustrated with diagrams as far as possible.

1.4 Role of Dynamic Time Warping

Similarly, the role of DTW, while potentially promising, is not convincingly established. At some stations DTW improves detection, but in other contexts its added value is marginal. A more explicit discussion of the specific conditions under which DTW enhances performance would strengthen the manuscript considerably.

We remark that since DTW is applied to segments extracted by the IF trigger, there is little room for improvement at stations where the IF tends to assign relatively high anomaly scores to segments associated with debris flows. This is why the improvement made by IF-DTW over IF is marginal at stations ILL11-ILL13 and ILL18. While this is discussed in Section 3.1.5, we will phrase this more clearly in a revised manuscript.

In addition, we will attempt to further improve the performance of IF-DTW by (a) exchanging Template DTW with Segment DTW and (b) experimenting with an alternative scoring strategy over the simple mean. Any changes made to the methodology will be clearly described in the methodology section and visually illustrated as far as possible.

1.5 Evaluation of Results

The evaluation of the results, although rigorous, would benefit from a clearer and more accessible presentation. Metrics such as precision, recall, and IoU are appropriate, but their distribution across dense tables makes comparisons difficult to follow. Averaged summary values or graphical representations would make the performance differences between methods easier to grasp.

To better illustrate the results in Section 3.1.5 of the paper, the dense Tables 1 and 2 will be moved to the appendix and replaced with a table of aggregated metrics over the test period only, with the best performing method for each metric bolded. Following our response to the previous comment, we will report the metrics for each station set {ILL11-ILL13, ILL18}, {ILL14-ILL17} and {ILL11-ILL18}. This will be done so that the stations where IF-DTW provides improvement are clearly indicated.

1.6 Generalization and Scalability

The generalization and scalability of the approach also deserve further elaboration. The manuscript focuses on two case studies, but it would be important to reflect on the applicability of the methodology to larger seismic networks, to other types of gravitational mass movements, and to real-time operational monitoring. A presentation and a discussion of all the hyper-parameters used and their values is mandatory.

We will include such a discussion in the revision, although we remark that a discussion of extending the methodology to real-time monitoring is given in the final paragraph on page 15, and that the hyper-parameters for the Illgraben case study are given in Appendix B. As mentioned by Referee #3, the IF has favourable computational and memory complexity, meaning that it is highly scalable. On the other hand, DTW is more challenging computationally, which is why it is appealing to be able to sparsify the waveforms into smaller segments. As long as other gravitational mass movements manifest as anomalies, it is not unreasonable to assume that the methodology would extend to such cases as well. We will discuss this accordingly in a revised manuscript.

1.7 Terminology

Related to this, the terminology is sometimes confusing. The distinction between "trigger segments," "detections," and "catalog entries" is central but not always presented with sufficient clarity. A clear diagram of the entire processing chain would help avoid such ambiguities.

We agree that the readability of Section 3.1 needs to be improved. The distinction between the terms mentioned by the referee will be clearly stated in a revision, also taking into account the suggestions made by Referee #3. All procedures used will be discussed with the aid of a diagram, as far as is reasonable.

2 Minor Comments

- *Larose et al. (2015) focuses exclusively on seismic noise monitoring. There are many other references that would more accurately illustrate the point being made here.*

  Bahavar et al. (2019) and Collins et al. (2022) represent significant contributions, but they are not the only efforts (particularly regarding machine learning) which is directly relevant to the present study.

  L28: Replace "see for example" with "e.g.," followed by citations. More exhaustive referencing is needed to 1) provide the correct context for the study and 2) guide readers to other relevant works.

  L34–36: The bibliography on background noise monitoring is more complete than that on machine learning in environmental seismology, even though the latter is central to this paper...

  Please see our response in Section 1.1.

- *L21–25: STA/LTA is a detector, not a discriminator. The current phrasing is misleading.*

  This will be clarified in a revised manuscript.

- L47: Clarify what "vanilla" refers to. In seismology or in machine learning? Many algorithms now exist that combine anomaly detection and classification (e.g., VAEs, contrastive learning).

  "vanilla" refers to unsupervised. This will be clarified in a revised manuscript.

- *Sections 3.1.3/3.1.4: These are methodological and should not appear in the results section.*

  These will be moved to the Methodology section in a revised manuscript.

- *The datasets description is insufficient (number of samples, classes distribution, training/validation/test splits).*

  We will include a table with detailed information in a revised manuscript.

- *L272: Provide justification for onset/offset thresholds; where do these "rule-of-thumb" values come from?*

  In a revised manuscript we will motivate where the rule-of-thumb suggestions for the IF trigger came from. While such recommendations are inherently heuristic in nature, we will motivate these using theory from the isolation forest combined with what we observed in the case studies.

- *L217: Clarify how grid search is performed in an unsupervised context. Does this not undermine the intended advantage of IF as a parameter-free exploratory tool? And the use of IF for true unsupervised exploration?*

  We remark that the mining strategies of this section operate in a semi-supervised context (see line 168). In this case we have a catalog of events that allows us to calibrate the thresholds of the IF trigger. In the unsupervised context one has to resort to rule-of-thumb thresholds. This will be clarified by discussing the semi- and unsupervised procedures used in the case studies separately in the methodology section.

- *Tables: Highlight best-performing results in bold to facilitate interpretation.*

  See our response in Section 1.4.

- *L272–273: Why the "top 50" segments? What if more than 50 are of interest? This seems to be central to your approach, and should be thoroughly discussed*

  Also specify the inconsistency threshold used in agglomerative clustering.

  L295: The description of four clusters "in increasing order of diversity" reflects a subjective choice. Justify why the dendrogram splits were regrouped this way and acknowledge the subjectivity involved.

  In a revised manuscript, the remaining segments (i.e. those outside the top-50) flagged by the IF trigger at KARAT will be clustered. However, under the current workflow, performing DTW between all the pairs of segments is not feasible computationally. Instead, we will use the approach by Wu et al. (2018), a reference already contained in the manuscript. In this approach, we take one of the remaining segments and perform DTW between this segment and each of the top-50 identified by the IF-trigger. These 50 distances will then serve as the features associated with the relevant segment and subsequently used in a clustering procedure.

  We remark that specifying an inconsistency threshold is not the only way to obtain a clustering from a dendrogram.

We will provide clarity on how any clustering was obtained.

- *L283–286: The explanation is unclear. A diagram of the complete processing chain would help.*

  Section 3.2.3: Replace dendrograms with examples of seismic signals in clusters (A, B, C and D). For a seismological audience, the waveforms themselves are far more informative.

  See our response in Section 1.3.

- *L289–290: Define the metric by which segments are "most anomalous." Provide values. Clarify what is meant by "further emphasized by the agglomerative clustering."*

  The most anomalous segment is the one with the largest IF segment anomaly score, which was defined in Section 2.2.2 of the paper. The detection associated with the rock avalanche was the one which merged with an existing cluster of detections in the dendrogram at the highest height, indicating that this segment is highly unusual according to the DTW distance, even among the highly anomalous IF segments. These points will be clarified in a revised manuscript and corresponding values given.

- *Figure 4 is not useful for the discussion and could be removed or moved to the appendices.*

  Figure 4 will be moved to the appendix.

- *L346–355: This discussion belongs in the introduction, not in the conclusion.*

  This section does contain a discussion of how the methodology could be extended to real-time monitoring, which relates to the comment discussed in Section 1.6. Here the referee is asking us to "reflect" suggesting that such a discussion should be included here and not in the introduction. We will, at the very least, move the literature related to the IF to the introduction in the revision, and consider where the remainder best fits in the flow of the manuscript.

References

[1] Jack Woollam, Jannes Münchmeyer, Frederik Tilmann, Andreas Rietbrock, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Carlo Giunchi, Florian Haslinger, Dario Jozinović, et al. Seisbench—a toolbox for machine learning in seismology. *Seismological Society of America*, 93(3):1695–1709, 2022.

[2] Yang Cao, Haolong Xiang, Hang Zhang, Ye Zhu, and Kai Ming Ting. Anomaly detection based on isolation mechanisms: A survey. *Machine Intelligence Research*, 22(5):849–865, 2025.