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. The relevant adjustments made in the revised manuscript are indicated in red.

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.

Please see the revised introduction.

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.

We have revised Sect. 2 substantially in order to make it more readable and accessible. Fig. 1. illustrates how waveforms traverse the iTrees and why anomalous time windows tend to follow short paths to terminal nodes. The application of DTW is illustrated in Fig. 4. and Fig. 5 using waveform segments taken from the data. We have simplified the mathematical notation, but remark that we have not moved it to the appendix. This is because these expressions are necessary to discuss the maximum depth parameter in the isolation forest as requested by reviewer #3 (see Sect. 2.2.1). As requested by the reviewer we added descriptions of the workflows in Sect. 2.4 and Sect 2.5, including a visual illustration of the semi-supervised exploration workflow used in the Illgraben case study in Fig 6.

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.

In addition to the revisions detailed above, we have moved several of the Figures in the appendices to the main text (see Fig 9. and Figs 14 - 17). We have

included Fig 8. to show how time series of the IF anomaly scores behaves in the Illgraben network over a debris flow event. More waveform-spectrogram plots for segments uncovered at the KARAT stations are given in Figs 11 - 13.

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.

Please see Sect. 3.1.4.

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.

Please see Sect. 2.4 for the adjustments made. We remark that the updated DTW scoring method provided better recall overall and also lead to increased performance at ILL15 over the scoring method used in the preprint. Consequentially we performed one more update of the catalog generated in the preprint (see Sect. 3.1.4).

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.

Please see Table 2 for the test metrics. The dense table were moved to Sect. A6 in Table A5 and A6.

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.

Please see final paragraph in Sect 4.

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.

We have included Table 1 to contain the definitions of all the types of segments used in the manuscript. We hope that this table, alongside Fig 6. and the corresponding explanation of the semi-supervised strategy in Sect. 2.4, will clarify the confusion. These additions meant that the Illgraben case study had to be rewritten and reorganized. We refer in particular to Sect. 3.1.2 - 3.1.4.

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.

  Please see the revised introduction.

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

  This will be clarified in a revised manuscript.

  Please see the first paragraph of Sect. 1.

- 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.

  See Line 65.

- *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.

  See Sect. 2.4.

- *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.

See Table A1 and Table B1.

- *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.

  See Sect. 2.2.2 and Fig. 2.

- *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.

  See Sect. 2.4.

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

  See our response in Section 1.4.

  See Table 2.

- *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.

All these points are discussed in the unsupervised workflow of Sect. 2.5.

- *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.

  See Sect. 2.5. The dendrogram generated from this procedure is given in Fig. B1 of the appendix.

- *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.

  We remark that the IF segment score is defined in line 144. We hope the the final paragraph in Sect. 3.2.1 clarifies any confusion. We remark that in Fig. 10, the segment DTW is now performed between all 194 IF segments flagged at NUUG which is why the Figure is different from that of the preprint.

- *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.

Moved to Fig. B3. in the appendices.

- *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.

  We considered including this paragraph in the introduction, but felt as though this was detrimental to the readability of the paper. In particular, we felt that one needs to read through the methodology to appreciate this discussion.

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.

Author Response to Referee Comment 2

We would like to thank the referee for their careful consideration of our manuscript. Our response is given below, with the comments made by the referee indicated in italics. The relevant adjustments made in the revised manuscript are indicated in red.

My only concern is that some figures that provide a clearer view of the case are located mainly in the appendix (Figures D4, E1, E4), which somewhat underestimates the importance of the study because they are not present in the main body of the paper.

As suggested by the referee, several figures currently presented in the appendix will be moved to the main body to improve clarity and readability. In particular:

- Figure D4 provides a concise example of how the isolation forest does not suffer from event masking as opposed to the STA-LTA trigger, and will be moved to the Illgraben case study in the main body of the paper to illustrate this point.

  Fig. D4 now appears as Fig. 9. in the main body of the paper. We have also included a new Fig 8. to show how time series of the IF anomaly scores behaves in the Illgraben network over a debris flow event.

- To improve readability of the manuscript, the dendrogram currently depicted in Figure 3 of the main body of the paper will be moved to the appendix and replaced with representative example waveforms of the clusters, including Figure E1. In addition, figures such as Figure E4 will be moved to this section to illustrate how the labels were obtained.

  Figs E1 - E4 in the preprint now appear in the main body of the manuscript, see Figs 14-17. Additional waveform-spectrogram plots are given in Figs 11 - 13.

In addition to the above revisions we have included waveforms from the Illgraben data to illustrate the methodolody. Please see Figs 1, 4, 5.

Author Response to Referee Comment 3

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. The relevant adjustments made in the revised manuscript are indicated in red.

1 Major Comments

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

1.1 Introduction

1. *I'm not suggesting that the authors should adopt a different, more recent version of IFs, but it could be helpful to a reader to provide some pointers for future developments (based on recent work). This is done to some extent at the end of the Discussion section, but there it reads a bit like an afterthought.*

   In a revised manuscript we will include a paragraph in the introduction with references to other anomaly detection algorithms, with a focus on more recent isolation-forest inspired methods. This paragraph will contain the references currently in the Discussion section of the manuscript. Additionally, we will cite the review paper [1] and include a discussion of references contained therein that we consider relevant to future developments.

   We considered including this paragraph in the introduction, but felt as though this was detrimental to the readability of the paper. In particular, we felt that one needs to read through the methodology to appreciate this discussion. Instead we included a paragraph starting in Line 470 to explicitly discuss some isolation-based alternative anomaly detection methods and also give references to look for alternative methods in the time series literature. To be consistent, we also included a similar paragraph for alternative applications of DTW starting in Line 486. In these paragraphs we included references to other studies in environmental seismology.

2. *I also want to remark that the motivation for using the proposed signal analysis methods (last paragraph) is not a direct consequence of the*

problem statement(s) laid out in the introduction. The authors state that STA/LTA is not suitable for environmental seismology due to the difficulty of discrimination, which is not solved by IFs or DTW (at least not without an additional clustering/classification step, which could likewise be applied to STA/LTA).

We agree with the referee that, like the STA-LTA trigger, the IF does not natively provide discrimination between events. Neither does DTW, since it provides a quantification of how similar sequences are and not where they came from. In both cases subsequent steps, such as clustering or expert labeling are needed. This will be clarified in a revised manuscript.

We respond to this point, and the one below here. In the revised introduction we motivated the use of the IF primarily from the vantage point of its favorable computational/memory complexity, its strong empirical performance and minimal number of hyper-parameters to tune (see final paragraph of Sect 1.). We did not feel comfortable commenting on the integration of other data sources since we did not perform any related experiments.

3. *In my opinion, the main advantage of IFs is their favourable computational/memory complexity, their enormous flexibility to combine different data sources (multi-sensor detection) and representations (time-series, images, tabular data, ...), and the limited number of free parameters that could affect their performance (for evaluation only "hmin" is a user parameter, which is not used/discussed in this study). Because of these reasons, IFs is a "one-method-fits-all" technique that could become a default choice for seismological data exploration, which would improve consistency across studies and ease of interpretation of their results. Working backward from this notion, the authors could highlight some challenges with multi-sensor, multi-dimensional data analysis (seismometers, GNSS, radar, pluviometry, Insar, DAS, ...) and consistency/interpretability of unsupervised detection and classification methods (STA/LTA, k-NN, t-SNE, auto-encoders, ...).*

   In light of this and the previous comment the introduction will be revised in order to better motivate the use of the IF. In this revision, we will mention

   - the difficulties we encountered in calibrating the STA-LTA trigger.
   - the IF as a powerful alternative given its favourable computational/ memory complexity and limited number of hyper-parameters.

   We will consider incorporating the suggestion made by the referee in relation to other data sources. We remark that while different data sources can be integrated into the IF by appropriately stacking features, the same applies to other methods and one would need to account for potential

outages in data arriving from these sources. At the same time, given the scalability and limited fine tuning required, this task could be easier with the IF. Furthermore, since the IF computes anomaly scores relative to nominal/normal points in the data, it does provide a degree of consistency and generalization when using different combinations of data sources.

see response above.

1.2 Methods

1. *Starting with Section 2.2: for someone who doesn't work with tree-based algorithms on a daily basis, it is not at all obvious that anomalies exhibit short paths in a tree (on average). This is explained in the Liu papers, but I think it would help the reader appreciate this method if the authors dedicate a few lines clarifying some statistical properties of anomalies, and how those are leveraged by IFs.*

   We will include a discussion of why anomalies tend to follow short paths, also taking into consideration the suggestion made by the referee to show how sliding windows traverse a tree.

   See Sect. 2.2 and Fig. 1.

2. *Then in lines 70-74 the authors describe the procedures for training and evaluation lumped together, which could be confusing/misleading to a casual reader.*

   In a revised manuscript, we will make sure to clearly distinguish between the training and evaluation phases, which is correctly described by the referee.

   See Sect. 2.2.1.

3. *It would also be helpful to explicitly state that x is an entire time window, and not just one recorded sample in that window; for a time window of size N, x is therefore a point in N-dimensional space.*

   We will make it clear already in Section 2.2 that $\boldsymbol{x}$ should be interpreted throughout as a vector containing all the data from a sliding window taken from the preprocessed waveforms.

   See Sect. 2.2.

4. *In Section 2.2.2, the authors define the use of "segment" as the collection of time windows for which the anomaly threshold exceeds the onset/offset conditions. This is where things might get a little confusing for someone who reads the paper with less attention; "segment" and "time window" are*

sufficiently generic terms that their specific meaning as used here might get lost as the reader progresses through the manuscript. For example, "Take all sliding windows over a segment" (line 125) is difficult to understand if the realisation has not yet fully set in that a "segment" has a specific meaning in this context. It might be helpful to the reader to systematically refer to "IF segment" in favour of just "segment". This tiny addition may seem insignificant, but it serves as a reminder of what it is that we're talking about, and to snap the brain out of the default mode of thinking about the meaning of "segment".

This suggestion will be incorporated and used consistently throughout the revised manuscript.

There are various definitions of segment types that are needed to describe the methodology contained in the paper. To improve readability we included all of these in Table 1.

5. *Section 2.3 introduces DTW, which is a technique that is best explained visually. It would be most helpful if the authors could create a figure explaining the basic concepts of Section 2.2 (how a time window traverses a tree), Section 2.2.2 (what is a "segment" and "IF segment anomaly score"), and Section 2.3 (how does DTW work, how are template/segment DTW defined).*

   Such visualizations will be provided in a revised manuscript.

   See Fig. 4. and Fig. 5.

1.3 Case Studies

1. *My remarks about the use of "segment" also apply to the word "catalog" as used in Section 3.1. If one takes a pre-existing catalogue and creates a new catalogue with IF, then you end up with two catalogues. But, if I understood correctly, in this manuscript "catalog" exclusively refers to the pre-existing catalogue, so it would make sense to systematically refer to it as "existing catalog" or "WSL catalog".*

   This confusion will be clarified in a revised manuscript, also taking into account a suggestion made by Referee #1 to include visualizations of the procedure. As suggested by the referee, we will refer to the initial catalog as the "WSL catalog". We do remark that the trigger parameters (for both STA-LTA and IF) are calibrated to the WSL catalog and subsequently frozen. In a similar fashion, the DTW scores are computed once in reference to the WSL catalog. It is only the score thresholds and minimum detection length that are re-calibrated when the catalog is updated.

Fig 6. contains a visual representation of the semi-supervised workflow applied to the Illgraben case study. We hope that this figure alongside the corresponding discussion of the workflow in Sect. 2.4 and a description of the various type of segments detailed in Table 1 will clarify the confusion. These additions meant that the Illgraben case study had to be rewritten and reorganized. We refer in particular to Sect. 3.1.2 - 3.1.4.

2. *This confusion is exacerbated by lines 196-197, which are written as if the detections are considered the ground truth, assigning a label to the existing catalogue entries. My suggestion for rewriting this sentence: "A detection is labeled a true positive (TP) if it overlaps with an entry in the existing catalog, otherwise it is labeled a false positive (FP). If no detection coincides with an existing catalog entry, it is labeled a false negative (FN)".*

   We thank the referee for the suggestion, however there is a specific reason why it is written as is in the current manuscript. It is possible that there are multiple detections that overlap with a specific entry in a catalog, and all of these will be labeled as true positives under the suggestion. Since all of these detections refer to the same event, they should be counted once. We will clarify this in a revised manuscript, taking care to be clear which catalog is under consideration.

   See line 235.

2 Minor Comments

- *Lines 41-42: the prominent placement of DTW seems to suggest that this is the cornerstone algorithm of this study, while it is my impression that IF does more of the heavy lifting (it is also the basis for DTW). Perhaps it makes more sense to not mention DTW here and introduce the acronym in line 50.*

  In hindsight we decided to keep the reference here because DTW does play a key role in the analysis of the case study.

  Lines 46-47: the authors could mention here that IFs have an evaluation cost and memory footprint that scale as $O(N)$. These are strong arguments in favour of IFs over pairwise-distance based algorithms that tend to scale as $O(N^2)$.

  We prefer to have this discussion in Sect. 2.2.1 in order to be precise about what this linear complexity means.

  Line 79: it would help the reader to state here that the harmonic number is approximated as $H(n) \approx ln(n) + 0.577...$ for $n >> 1$, or show the Taylor

expansion

See Line 120.

Line 472, "non-linear wave-like pattern": I suppose that the authors describe the shape of the data in analogy of a ocean wave breaking on the beach? Since this paper will be read by seismologists, "wave-like" will invoke a very different mental picture than the pattern of the data in these figures. I would use the term "hook-like".

See paragraph starting at Line 168 and Sect. A1.

These suggestions made will be incorporated in a revised manuscript.

- *Section 2.2: what about the "height limit" ("hlim" in Liu et al. 2018) that is part of the original algorithm proposed by Liu? Is this value set to infinity (= smallest granularity)? Did the authors consider this parameter at all? If all attributes of a data point x follow the same (normal) distribution, then this parameter should have no effect because only a single cluster exist that is centred on the origin.*

  When building IF trees, we stuck to the defaults provided by Scikit-learn and did not experiment with any other configurations. In a revised manuscript we will include a discussion of the Scikit-learn defaults, how this relates to the "hlim" parameter and how these defaults effect the training process.

  We included a detailed discussion of this paramemter in Sect. 2.2.1. We emphasize again that we adhered as far as possible to the Scikit-learn defaults.

- *Figures A1,A2: my PDF reader was struggling a bit to render the large number of data points. Since we can't see any details when 1000 symbols overlap, it would be more rendering-friendly to rasterise these figures, or at least the data points.*

  In a revised manuscript, we will revise Figures A1 and A2 to improve rendering performance while preserving their visual content

  We rasterised these figures to make them more rendering-friendly. We do remark that the plots for ILL11 and ILL14 was moved to Fig. 3 so that the reader can gain an appreciation for how the IF anomaly scores relate to the log standard deviation. The corresponding plots for the remaining stations are given in Fig A1. The rasterisation of the figures also allowed us to expand the plots to contain all points and not just those corresponding to an IF anomaly scores exceeding 0.5.

- *Line 474, "they are fairly highly rank in the log standard deviation bands in which they appear": what does this mean? That ILL11,12,13,18 achieve higher IF scores than the other stations?*

  At stations ILL11,12,13,18 sliding windows that overlap with debris-flow related events can reach relatively high IF anomaly scores compared to the other stations. In contrast to these stations we see, for example, that no debris-flow related sliding window has an IF score of higher than 0.65. However, a large number of non debris-flow related events can be screened away by requiring that the log-standard deviation should be small (e.g. less than around 4.5). This suggests that at "weaker" stations like ILL14, the significance of the IF score associated with debris-flow related sliding windows can be improved by conditioning on the log-standard deviation. This paragraph will be clarified in a revised manuscript.

  See final paragraph of Sect. 2.2.2 .

- *Figures D1-4: in all cases, the IF score seems to increase a lot before the visible onset of the anomaly in the time series. Is this a result of acausal filtering/processing? If so, it would be good to mention that somewhere. If not, that would suggest IF is able to pick up an anomalous signal before it becomes visible in the time-series.*

  While this has not been systematically evaluated, we do not consider the early increase in the IF score to be solely attributable to the acausal filtering. In our experience, acausal-type preprocessing can in some suppress the IF score over debris-flow periods. For example, we experimented previously with preprocessing the entire waveform in a miniseed recording by subtracting the corresponding median and dividing by the mean absolute deviation. If such a recording happened to contain a debris flow, the amplitude in the waveform, and therefore the corresponding anomaly scores, is suppressed over this period because the mean absolute deviation (although robust) is increased by the event. We agree that clarification is needed and will include a brief discussion in the revision.

  See the caption of Fig 8.

References

[1] 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.

Author Response to Referee Comment 4

We would like to thank the referee for their careful consideration of our manuscript. Our response is given below, with the comments made by the referee indicated in italics. The relevant adjustments made in the revised manuscript are indicated in red.

- *Perhaps the authors could consider providing additional information, such as some Python scripts for all these algorithms.*

  All code used to produce the results of the paper is available on github: `https://github.com/FKamper/seismic-isolation-forest`. While this repository does contain python scripts, we will review its content before submitting a revision of the manuscript and make improvements if necessary. Any updates made to the repository will be made public should the manuscript be accepted for publication.

  We have made adjustments to the above repository to facilitate the revision; these are contained inside the branch `https://github.com/FKamper/seismic-isolation-forest/tree/revision_v2`. In this adjustments we reorganized the python scripts to improve readability. As mentioned above, these changes will be merged with the main branch should the manuscript be accepted for publication.

- *The overall findings or main conclusions are notable, and I recommend publishing the article, but taking into account the suggestions of experienced colleagues who have commented previously.*

  Please refer to our responses to the other referee comments.

  Please see the revisions detailed in our responses to the other referees.