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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.*

  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 $O(N^2)$.

  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

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".

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.

- *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

- *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.

- *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.

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.