August 11, 2026

To: The Editor-in-Chief, The Associate Editor, And the Reviewers
Natural Hazards and Earth System Sciences

Title: Reducing False Alarms in Crackmeter-Based Early Warning of Small-Scale Slope Deformation via Deep Learning-Driven Asynchronous Displacement–Rainfall Data Fusion

Authors: Shubing Ouyang, Daichao Li*, Minjiang Liu, Fengjian Ge, Xia Zheng, Yuan Li, and Sheng Wu

Dear Timothy Tiggeloven and Reviewers:

Thank you very much for your response. We thank the editor and reviewers for the time and effort that they put into reviewing the previous version of the manuscript. Following the reviewers' comments, we have carefully modified and improved our manuscript. We hope these revisions now resolve the problems and uncertainties pointed out by the reviewers.

A point-by-point response to the reviewers' comments is provided in the following pages.

For clarity, we note at the outset that the numerical results（Table 7） in the revised manuscript differ from those reported previously because the experimental protocol was substantially revised. Specifically, (1) the dataset was repartitioned to enforce strict sensor-level isolation among the training, validation, and test sets, including within the five-fold cross-validation; (2) all normalization parameters were estimated exclusively from the corresponding training subset and then applied unchanged to the validation and test sets; (3) the positive-class weight was recalculated for each fold based on its training-set class distribution; and (4) MVM was removed as an independent input channel and retained only for fused-feature construction. All models were retrained under this revised, leakage-controlled protocol, and the corresponding results, tables, figures, and conclusions were updated accordingly. Therefore, all conclusions in the revised manuscript are based on the newly generated results rather than those reported in the previous version.

We sincerely hope that our manuscript is now suitable for publication in Natural Hazards and Earth System Sciences.

Thank you for your assistance and we look forward to hearing from you.

Yours sincerely,
Daichao Li and Shubing Ouyang
The Academy of Digital China (Fujian)
Fuzhou University,
Fuzhou, China
Email: lidc@fzu.edu.cn; 394666@fzu.edu.cn

Table 7. Accuracy results and hyperparameter search spaces of the evaluated models

Model	Input data/structure	Precision of true early warning	Recall of true early warning	F_2-score of true early warning	Overall accuracy	AUPR
Bayesian I–D Threshold	Rainfall data	18.62±0.74	**95.94±0.35**	52.39±1.16	54.73±2.31	29.67±2.36
Logistic regression	Handcrafted predictors	60.71 ± 0.00	79.69 ± 0.00	75.00 ± 0.00	92.33 ± 0.00	71.21 ± 0.00
Decision tree	Handcrafted predictors	75.64 ± 0.66	75.16 ± 1.40	75.25 ± 1.02	94.76 ± 0.07	81.34 ± 1.05
Random forest	Handcrafted predictors	87.93 ± 0.51	64.84 ± 0.00	68.44 ± 0.06	95.30 ± 0.05	88.53 ± 0.08
TCN	Dual-branch	84.72±2.17	78.44±2.25	79.6±1.73	96.18±0.25	84.73±1.08
LSTM	Dual-branch	80.09±1.87	82.34±3.06	81.87±2.67	95.93±0.45	88.89±1.24
TimesNet	Dual-branch	85.85±1.81	84.69±3.81	84.87±2.78	96.86±0.17	92.17±1.39
DLinear	Dual-branch	67.74±1.13	78.28±3.55	75.88±2.55	93.69±0.19	73.2±1.58
PatchTST	Dual-branch	90.14±2.16	82.81±1.91	84.18±1.9	97.2±0.4	92.87±1.44
iTransformer	Dual-branch	80.03±5.36	84.22±5.7	83.16±3.54	96±0.47	86.1±1.8
PatchDC	Dual-branch	**90.32±2.64**	89.38±1.8	**89.54±1.09**	**97.83±0.22**	**94.78±0.56**

The hyperparameter search space

Bayesian I–D Threshold: IETD ∈ {6, 12, 24, 32} h; $I = \alpha D^{-\beta}$; $1/\alpha \sim U(0.001, 10)$, $\beta \sim U(0.1, 2.0)$; selected by mean validation F_2. Search sampling: 500 draws + 500 tuning steps, 2 chains; final refit: 2000 draws + 1000 tuning steps, 4 chains.

Logistic regression: RobustScaler; class_weight = balanced; solver = liblinear; max_iter = 5000. Grid: C ∈ {0.001, 0.01, 0.1 , 1, 10, 100}; penalty ∈ {L1, L2}

Decision tree: class_weight = balanced. Grid: max_depth ∈ {2, 3, 5, None}; min_samples_leaf ∈ {5, 10, 20}

Random forest: n_estimators = 500; max_features = sqrt; class_weight = balanced_subsample; bootstrap = True. Grid: max_depth ∈ {4, 8, None}; min_samples_leaf ∈ {2, 5, 10}

TCN: num_levels ∈ {8, 16, 32, 64}; dropout ∈ {0.1, 0.3, 0.5}; branch_channels ∈ {1×, 2×, 4× num_levels}.

LSTM: hidden_size ∈ {32, 64, 128}; num_layers ∈ {1, 2, 3}; dropout ∈ {0.1, 0.3, 0.5}.

TimesNet: top_k ∈ {2, 3, 4, 5}; d_ff ∈ {8, 16, 32}; num_kernels ∈ {3, 5, 7}; d_model ∈ {32, 64, 128}; e_layers ∈ {1, 2, 3}; dropout ∈ {0.1, 0.3, 0.5}.

DLinear: moving_avg_window ∈ {15, 21, 27, 33, 39, 45}; pred_len ∈ {1, 12, 24, 48}; dropout ∈ {0.1, 0.3, 0.5}.

PatchTST: hidden_size ∈ {32, 64, 128}; encoder_layers ∈ {1, 2, 3}; n_heads ∈ {2, 4, 8}; patch_len ∈ {8, 12, 16, 20, 24}; dropout ∈ {0.1, 0.3, 0.5}; stride ∈ {3, 5, 7}.

iTransformer: d_model ∈ {32, 64, 128}; e_layers ∈ {1, 2, 3}; n_heads ∈ {2, 4, 8}; dropout ∈ {0.1, 0.3, 0.5}.

PatchDC: hidden_size ∈ {32, 64, 128}; encoder_layers ∈ {1, 2, 3}; n_heads ∈ {2, 4, 8}; patch_len ∈ {8, 12, 16, 20, 24}; dropout ∈ {0.1, 0.3, 0.5}; stride ∈ {3, 5, 7}.

PART II. REVIEWER 2

This manuscript proposes a dual-branch Transformer model (PatchDC) that reframes crackmeter alert validation as a binary classification problem ("true early warning" vs. "false alarm"), fusing asynchronous displacement and rainfall data. The operational motivation is well documented (a reported false-alarm rate below 0.03% over five years), and the proposed architecture — causal attention respecting the physical precedence of rainfall over displacement, and a hybrid fusion mechanism for correlated variables — is reasonably justified and convincingly validated through a well-designed ablation study. The interpretability analysis, and especially the dedicated failure-case section (Section 6.4), reflect a level of scientific transparency that is uncommon in this type of applied study.

However, the manuscript does not allow a full assessment of the internal validity of some of the reported results. The main concern relates to the construction of the positive dataset: expanding the 93 field-confirmed events to 255 samples by adding temporally adjacent points raises a circularity issue that is not discussed. A related concern involves the risk of information leakage through several normalisation steps computed "across the dataset", without the exact scope (training set only, or training+test combined) being specified. Both issues bear directly on the reliability of the performance metrics highlighted in the abstract and should be clarified before publication.

The manuscript is scientifically sound and likely publishable after revision. The comments below are intended to help the authors strengthen the internal validity of their results and the clarity of their evaluation protocol.

Major Comments

Comment 1 — Risk of circularity in the construction of the positive dataset. The initial number of field-confirmed true early warnings was only 93 (l. 199-200). To reach 255 samples, the authors added points temporally adjacent to these confirmed events, validated through a triple-blind review by three experts (l. 204-207). This procedure raises an important question that is not addressed: to what extent are these additional points independent of the original events in terms of signal (RDS, HMD)? If they share the same underlying displacement dynamics — which is plausible, since they were selected for their temporal proximity — the model may be learning an "episode signature" rather than generic early-detection features. Sensor-level isolation between training and test sets (l. 208-211) limits direct leakage but does not resolve this intra-episode homogeneity risk. If feasible, a sensitivity analysis training/evaluating the model on the original 93 events only would help demonstrate robustness.

Response:

Thank you for raising this important and technically insightful concern. We also apologize for an error in the number of originally confirmed true-warning events reported in the previous manuscript. The initially reported number included several field-confirmed warning events for which corresponding rain-gauge observations were unavailable. After excluding these events according to the data-availability criteria used in this study, 82 originally field-confirmed true-warning events remained in the final dataset, of which 42 were assigned to the training–

validation set and 40 to the independent test set under the sensor-disjoint partitioning strategy. Thus, in the sensitivity analysis described below, model training relied on only 42 originally confirmed positive events.

The added text of sensitivity analysis is as follows:

Page 24：6.2 Sensitivity analysis using only the originally confirmed true warnings

To assess whether the temporally adjacent samples introduced intra-episode homogeneity and caused the model to learn episode-specific signatures, a sensitivity analysis was performed using only the 82 originally field-confirmed true-warning samples as the positive-sample pool for both training and testing. These samples comprised 42 originally confirmed events from the augmented training–validation set and 40 from the augmented test set, with no overlap between the two subsets (Table 4). All temporally adjacent, expert-validated samples were excluded from this analysis. The sensor-disjoint partitioning strategy, model architecture, optimization procedure, and evaluation protocol were identical to those used in the main experiment. For every training split, the normalization parameters were estimated exclusively from the corresponding training set and were subsequently applied, without refitting, to the validation and test sets. Similarly, the positive-to-negative sample weighting in the loss function was recalculated according to the class ratio of the corresponding training set. This ensured that the sensitivity analysis was self-contained and that no information associated with the augmented dataset entered the confirmed-only training process.

As shown in Table 9, compared with the model trained on the augmented dataset and evaluated on the same confirmed-warning test samples, the confirmed-only model showed a precision higher by 1.74 percentage points, whereas its recall, F_2-score, and AUPR were lower by 7.50, 5.15, and 3.96 percentage points, respectively. Overall accuracy was essentially unchanged, with the confirmed-only model showing a marginal decrease of 0.07 percentage points. These results show that the model retained substantial early-warning detection capability when the temporally adjacent samples were completely removed, indicating that its predictive performance was not solely dependent on signatures introduced by the augmentation procedure. The higher recall, F_2-score, and AUPR obtained with augmented training suggest that the additional samples mainly improved positive-class coverage and sensitivity. The augmented-training configuration also produced smaller standard deviations for recall, F_2-score, overall accuracy, and AUPR, consistent with improved model stability resulting from the larger positive training set.

Table 9. Performance on originally field-confirmed true warnings under confirmed-only and augmented training

Train dataset (positive samples)	Test dataset (positive samples)	Precision of true early warning	Recall of true early warning	F_2-score of true early warning	Overall accuracy	AUPR
originally confirmed true warnings	originally confirmed true warnings	76.87 ± 5.11	84.00 ± 6.52	82.39 ± 5.55	98.51 ± 0.35	88.25 ± 2.79
augmented data	originally confirmed true warnings	75.13 ± 5.20	91.50 ± 3.79	87.54 ± 1.79	98.58 ± 0.26	92.21 ± 1.44

Comment 2 — Ambiguous scope of global normalisation statistics. Several preprocessing steps rely on values computed "across the dataset": the four derived statistical features (IQR, ADM) are normalised by their "collective maximum value across the dataset" (l. 244-245), and RMR uses a global maximum across all stations as well as a regional five-year maximum (l. 265, 270). The text does not specify whether these maxima are computed on the training set alone or on the combined training+test set. In the latter case, this would constitute a minor but real information leak into the test set. Explicitly clarifying this point, and reformulating the pipeline to ensure these statistics are computed exclusively from training data (per cross-validation fold), would meaningfully strengthen the credibility of the results.

Response:

Thank you for your professional comment regarding the normalisation statistics. We confirm that the original normalisation statistics for IQR, ADM, and RMR were calculated using the complete dataset, which introduced unintended information leakage. We have corrected this issue by calculating all normalisation parameters exclusively from the training subset of each cross-validation fold. The resulting parameters are then applied unchanged to the corresponding validation and test data. **As a result, we re-ran all the experiments.** We have updated the manuscript and results accordingly.

We have revised the text as follows:

Page 11 (3.2 Dataset Construction): Each of the four derived statistical features was subsequently normalized using its corresponding maximum value, calculated exclusively from the training subset of each cross-validation fold.

Page 12 (3.2 Dataset Construction): D. The preliminary relative rainfall index was normalized using its maximum value calculated from the training subset of each cross-validation fold. The core formula is defined as follows:

$$\mathrm{rain}^{\mathrm{i}}_{statistic} = \frac{\text{Single} - \text{hour rainfall\_max\_i/ Annual rainfall\_i}}{max_{j \in Train}(\text{Single} - \text{hour rainfall\_max\_j/ Annual rainfall\_j})} \quad (1)$$

Following the aforementioned calculation, Relative Maximum Rainfall Intensity was defined as a dimensionless index characterizing the extremity of the maximum hourly rainfall relative to the local historical rainfall background. The maximum hourly rainfall within each observation window was scaled by the corresponding regional average annual rainfall, thereby emphasizing rainfall anomalies relative to local climatic conditions rather than absolute rainfall magnitude.

Comment 3 — Inconsistency between the two modes of reporting results. Section 5 selects the median run by recall among the five repetitions as the "representative" result (precision 90.91%, recall 93.53%, F2 92.99%), whereas Table 5 reports the mean ± standard deviation over 25 runs (precision 89.24±2.43%, recall 93.38±0.95%, F2 92.50±0.49%). These two sets of figures are close but not identical, and the use of two different reporting conventions in two places in the manuscript is not justified. The authors should harmonise their presentation, or otherwise clearly explain why each metric is used where it appears.

Response:

Thank you for pointing out this inconsistency. We apologize for the inconsistent and

redundant reporting in the previous manuscript. We agree that the presentation should be harmonized. Therefore, **we have removed** the “representative” result selected as the median run by recall in Section 5. In the revised manuscript, all performance results are reported consistently as the mean ± standard deviation over 5 repetitions.

Comment 4 — Sampling uncertainty likely underestimated for the positive class. The test set contains only 139 "true early warning" samples (Table 4; confirmed by the confusion matrix: 130 TP + 9 FN). At this scale, a single additional misclassified sample shifts recall by roughly 0.7 percentage points. The inter-run standard deviation currently reported (Table 5) captures training-related variance, but not the sampling uncertainty inherent to a test set of this size. A bootstrap confidence interval on the reported metrics would give a more complete picture of their statistical reliability.

Comment 5 — No significance testing between competing models. The gaps between PatchDC and the best competitor (PatchTST) are sometimes small relative to the reported standard deviations — particularly for recall (93.38±0.95% vs. 92.95±1.24%, a gap of 0.43 points for standard deviations of comparable magnitude, Table 5). Without a paired statistical test (e.g., a t-test or Wilcoxon test across the 25 runs), this recall advantage could well be training noise rather than a genuine improvement, even though the precision gap (3.62 points) looks more robust.

Response to Comments 4 & 5 (Combined):

Thank you for these important comments. To assess both test-sensor sampling uncertainty and training variability, we added a paired sensor bootstrap, paired t-tests on the five seed-level probability ensembles, and paired t-tests on the 25 matched seed–fold results between PatchDC and PatchTST. Holm correction was applied to the t-tests, and statistical significance was evaluated using adjusted p-values and paired 95% bootstrap confidence intervals.

We also note that the numerical results in the revised manuscript differ from those reported in the previous version because several methodological corrections and refinements were implemented during revision. First, the dataset was repartitioned to enforce **sensor-level isolation among the training, validation, and test sets**, and the same sensor-isolation constraint was maintained within the five cross-validation folds. Second, the normalization procedure was revised so that all normalization parameters were estimated **exclusively from the corresponding training subset** and then applied unchanged to the validation and test data, thereby preventing information leakage. Third, instead of using the previously fixed 10:1 class-weighting ratio, the positive-class weight in the loss function was recalculated separately for each fold according to the actual negative-to-positive ratio in that fold's training subset. Finally, the long-term branch of PatchDC was refined by removing MVM as an independent input channel; MVM is now used only within the Hybrid Feature Processing module for fused-feature construction. Accordingly, **all conclusions in the revised manuscript are based on the newly generated results under the revised and leakage-controlled experimental protocol**, rather than on the numerical values reported in the previous version.

The added text of sensitivity analysis is as follows:

Page 22-23 (5.1 Comparative Model Accuracy Analysis): To evaluate whether the performance differences between PatchDC and PatchTST were robust to uncertainty in test-sensor composition and model training, three complementary analyses were conducted (Table 8). First, a

fixed-prediction paired sensor bootstrap was used to quantify sampling uncertainty associated with the independent test sensors. Second, paired t-tests were applied to the five seed-level final probability ensembles to assess the influence of random initialization on the final prediction results reported in Table 7. Third, paired t-tests were applied to the 25 matched seed–fold results as a sensitivity analysis of individual fold-model variability. Precision, recall, F_2-score, overall accuracy, and AUPR were evaluated separately. Holm correction was applied across the five metric-specific t-tests in the seed-level and fold-level analyses, and Holm-adjusted p-values below 0.05 were considered statistically significant. The bootstrap analysis was evaluated using the paired 95% confidence interval of the performance difference; an interval excluding zero indicated a consistent difference under resampling of the test sensors.

First, a fixed-prediction paired sensor bootstrap was used to quantify uncertainty associated with the independent test set. Sensors, rather than individual warning samples, were resampled with replacement because samples originating from the same sensor were not independent. PatchDC and PatchTST were evaluated using the same resampled sensors in each iteration. PatchDC increased recall by 6.562 percentage points, F_2-score by 5.357 percentage points, overall accuracy by 0.634 percentage points, and AUPR by 1.906 percentage points (Table 8). Because all four intervals excluded zero, these improvements were maintained when the composition of the test sensors was varied. In contrast, the precision difference was only +0.172 percentage points, and its 95% CI included zero, indicating no reliable precision difference under sensor resampling. Therefore, the bootstrap analysis further demonstrated that, although PatchDC and PatchTST showed comparable performance in distinguishing false alarms, PatchDC was more effective in correctly identifying true warnings.

Second, paired t-tests were conducted using the final probabilities averaged across the five folds for each of the five random seeds. This analysis corresponded directly to the prediction strategy and results reported in Table 5. Compared with PatchTST, PatchDC increased recall from 82.81% to 89.38%, an improvement of 6.562 percentage points, with a Holm-adjusted p-value of 0.0041. It also increased the F_2-score from 84.18% to 89.54%, an improvement of 5.357 percentage points, with a Holm-adjusted p-value of 0.0025. The Holm-adjusted p-values for both metrics were below 0.05, making recall and F_2-score the only metrics with statistically significant differences after Holm correction. Overall accuracy and AUPR increased by 0.634 and 1.906 percentage points, respectively, but neither improvement remained significant after correction because both Holm-adjusted p-values exceeded 0.05. Mean precision was also similar between PatchDC and PatchTST, at 90.32% and 90.14%, respectively, and the difference was not statistically significant after Holm correction. Therefore, the paired t-test results based on the five random seeds showed that the significant advantages of PatchDC were primarily reflected in recall and F_2-score, indicating that PatchDC was more effective than PatchTST in correctly identifying true warning cases.

Finally, paired t-tests were conducted on the 25 matched fold-level results obtained from five random seeds and five folds as a sensitivity analysis of variability among the individual fold models. Because all Holm-adjusted p-values exceeded 0.05, none of the fold-level differences was statistically significant after correction. Nevertheless, the Holm-adjusted p-value for recall was 0.0641461, which was below 0.1, while that for the F_2-score was 0.102878, which was close to 0.1. These results indicated directional improvements in recall and F_2-score, although they did not demonstrate consistent superiority across the individual fold models. Importantly, the fold-level

analysis was not equivalent to the final ensemble evaluation because the final metrics were recalculated after averaging the predicted probabilities across the five folds. The directional advantages observed at the fold level may therefore help explain the statistically significant improvements in recall and F_2-score achieved by the final five-fold probability ensemble.

Taken together, the three complementary analyses showed that the most robust advantage of PatchDC over PatchTST lay in its ability to identify true warnings. The improvements in recall and F_2-score were maintained under sensor resampling and remained statistically significant across the seed-level final ensembles after Holm correction, whereas precision was consistently comparable between the two models. The evidence for improvements in overall accuracy and AUPR was less consistent across the different analyses. Although the fold-level paired t-tests did not identify statistically significant differences after Holm correction, recall and F_2-score still showed directional improvements. These findings suggest that five-fold probability averaging helped reduce the influence of individual fold-model variability and produced a more stable advantage for PatchDC in true-warning detection, thereby supporting the robustness of the performance gains reported in Table 7.

Table 8 Complementary paired comparisons between PatchDC and PatchTST.

Metric	Fixed-prediction paired sensor bootstrap	Five seed-level five-fold ensembles: paired t-test	Twenty-five fold-level pairs: paired t-test
Precision	95% CI [−2.450, +2.914] Not significant	Δ=+0.172 pp 95% CI [−4.433, +4.776] raw p=0.922579 Holm p=0.922579	Δ=−1.984 pp 95% CI [−4.590, +0.623] raw p=0.129328 Holm p=0.387983
Recall	95% CI [+2.667, +10.845] PatchDC significantly higher	Δ=+6.562 pp 95% CI [+4.437, +8.688] raw p=0.00101666 Holm p=0.00406665	Δ=+6.406 pp 95% CI [+1.489, +11.323] raw p=0.0128292 Holm p=0.0641461
F_2-score	95% CI [+1.967, +9.227] PatchDC significantly higher	Δ=+5.357 pp 95% CI [+3.909, +6.805] raw p=0.000506377 Holm p=0.00253189	Δ=+4.893 pp 95% CI [+0.646, +9.140] raw p=0.0257194 Holm p=0.102878
Accuracy	95% CI [+0.098, +1.261] PatchDC significantly higher	Δ=+0.634 pp 95% CI [+0.173, +1.095] raw p=0.0187921 Holm p=0.0563763	Δ=+0.264 pp 95% CI [−0.319, +0.846] raw p=0.359515 Holm p=0.461722
AUPR	95% CI [+0.236, +4.339] PatchDC significantly higher	Δ=+1.906 pp 95% CI [+0.469, +3.342] raw p=0.0211363 Holm p=0.0563763	Δ=+1.872 pp 95% CI [−1.271, +5.016] raw p=0.230861 Holm p=0.461722

Note: Δ denotes PatchDC minus PatchTST in percentage points. The seed-level analysis uses each seed's final five-fold probability ensemble and is the primary inferential analysis. The sensor bootstrap quantifies uncertainty associated with the sampled test sensors.

Comment 6 — Geographic scope and generalisability insufficiently foregrounded. The authors

acknowledge in the conclusion (Section 7) that the model has only been validated in a homogeneous geological and climatic context (Fujian Province, rainfall-triggered events exclusively). This is honestly stated, but it only surfaces at the very end of the manuscript; it should be flagged as early as the abstract and introduction so readers don't over-generalise, especially given that only 52 of the 429 sensors actually provided the true alerts used for training — a fairly narrow base of geological sites.

Response:

Thank you for pointing this out. *This point has now been explicitly addressed in both the Abstract and the Introduction of the revised manuscript, as follows:*

(1) The Abstract: However, the present validation is limited to the Fujian rainfall-triggered monitoring context, and application to other geological or climatic settings requires further external validation and local adaptation.

(2) the Introduction:

Page 4: It should be noted that the operational monitoring network comprised 429 crackmeters, whereas field-confirmed true-warning samples were obtained from 52 sensors. Accordingly, the present validation is limited to the rainfall-triggered monitoring context in Fujian Province, and the resulting model should not be assumed to generalize directly to other geological, climatic, or monitoring settings.

Minor Comments

Comment 7 — Typographical error. l. 642: "moder ate" should read "moderate".

Response:

Thank you for pointing this out. The word error has been corrected.

Comment 8 — Striking figure insufficiently detailed. The 0.03% rate cited in the introduction (l. 132-134) is striking but presented without methodological detail on the exact period covered or the operational definition of "validated". A supplementary table would strengthen the credibility of this figure.

Response:

Thank you for your suggestion. *Both the table and the text have been added:*

Page 5 (2 The Engineering Problem: Early Warning Characteristics and System Limitations): From 2020 to 2024, the Fujian Provincial Geological Hazard Early Warning Platform received approximately 2.05 million threshold-triggered device alerts. Following expert assessment and field verification, only slightly more than 500 alerts, including slightly more than 100 crackmeter alerts, were validated as true warnings, corresponding to an overall effective warning rate of less than 0.03% (Table 1).

Table 1. Summary of device alerts received by the Fujian Provincial Geological Hazard Early Warning Platform

Period	Total device alerts	Validated true warnings	Effective warning rate
2020–2024	Approximately 2.05 million	More than 500, including slightly More than 100 crackmeter alerts	<0.03%

Comment 9 — Missing precision-recall curve. AUPR is presented as an important metric (Section 4.3), yet no precision-recall curve is shown in the main text. A figure would usefully illustrate the model's stability across decision thresholds, consistent with the Section 6.5 discussion of lowering the threshold to 0.4.

Response:

Thank you for your suggestion. *Both the figure and the text have been added:*

Page 35 (6.5 Monitoring Process Improvement Based on Model Limitations): The second strategy is to lower the classification threshold from 0.5 to a lower value, such as 0.4 (Figure 13). At a threshold of 0.4, the mean recall reaches 92.5%, which is 3.12 percentage points higher than the recall of 89.38% obtained at a threshold of 0.5. This will increase the recall (reduce missed alarms) but will also increase the number of false alarms, shifting the operating point on the PR curve.

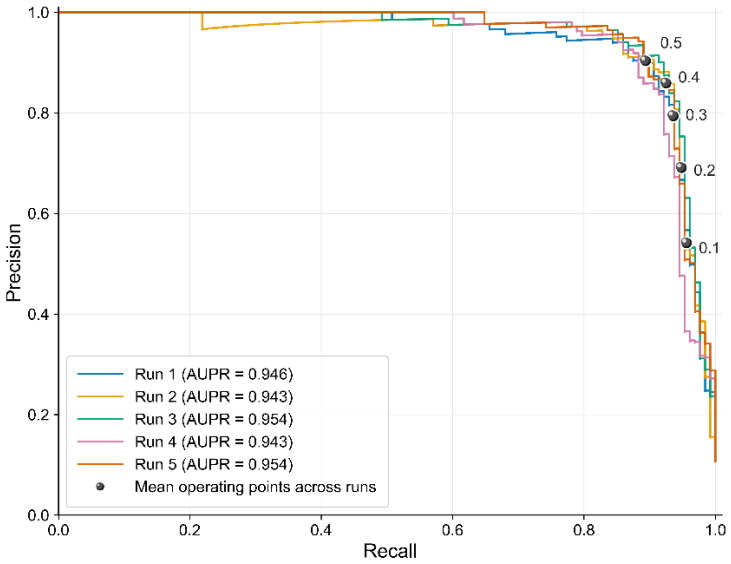

Figure 13 Precision–Recall Curves Across Five Independent Runs with Mean Operating Points at Selected Decision Thresholds

Comment 10 — Proposed but untested ensemble strategy. The "one-vote positive" strategy discussed in Section 6.5 as a future improvement is not empirically evaluated in this article, even though the five models required (one per fold) are already available. Adding this evaluation would strengthen the practical relevance of the discussion.

Response:

Thank you for your suggestion. *The text has been added:*

Page 35(6.5 Monitoring Process Improvement Based on Model Limitations): The one-vote-positive mechanism increased sensitivity to true-warning samples. Across five random-seed runs, it achieved an accuracy of 91.843 ± 2.13%, precision of 57.715 ± 6.86 %, recall of 95.312 ± 0.78 %, F_2-score of 84.128 ± 3.02 %, and AUPR of 91.216 ± 0.77 %. Compared with the mean recall of 89.38% obtained by five-fold probability averaging mechanism, the “one-vote-positive” mechanism increased recall by 5.932 percentage points, substantially reducing missed detections, albeit at the cost of lower precision and more false alarms.

Comment 11 — Imprecise wording regarding the weighting ratio. The loss weighting ratio (10:1) is described as "strictly aligning with the actual class distribution ratio" (l. 367), whereas the actual ratio in the training set is closer to 9.3:1 (1082/116). A more cautious phrasing would be preferable.

Response: Thank you for pointing out this imprecise wording. We agree that describing the fixed 10:1 loss-weighting ratio as strictly matching the class distribution was inaccurate, because the overall negative-to-positive ratio was approximately 9.3:1 and varied slightly among cross-validation folds. In the revised experiments, we removed the fixed 10:1 weighting scheme and recalculated the positive-class weight separately for each fold using only the class distribution of that fold's training subset. All models were subsequently retrained using these fold-specific weights.

The corresponding description in the manuscript has been revised accordingly:

Page 17 (4.2 Training Strategy, Environment and Parameters): To address class imbalance, the positive-class weight in the loss function was calculated separately for each cross-validation fold as the ratio of negative to positive samples in that fold's training subset.

Comment 12 — Sensor-level isolation within cross-validation. The manuscript states that the overall train/test split strictly respects sensor-level isolation (l. 208-211), but does not specify whether the same constraint applies to the 5 cross-validation folds within the training set. Clarifying this would remove any ambiguity regarding the hyperparameter optimisation procedure.

Response:

Thank you for highlighting this ambiguity. We have now enforced strict sensor-level isolation across the training, validation, and test sets and rerun all models accordingly.

We have clarified this procedure in the revised manuscript as follows:

Page 17-18 (4.2 Training Strategy, Environment and Parameters): All samples, including both true and false alerts, originating from the same sensor were assigned exclusively to either the training–validation set or the test set. Sensors, rather than individual alarm samples or events, were randomly assigned to the two sets, thereby preventing sensor-level data leakage. The same sensor-level isolation constraint was applied to the five-fold cross-validation procedure within the training–validation set, ensuring that, in each fold, no sensor appeared in both the training and validation subsets. Detailed splitting results, including counts per set, are provided in Table 4.

Table 4 Division of Training, Validation, and Test Sets

Dataset	Alert events of True warnings (sensors)	Alert events of False warnings (sensors)	Alert events of originally confirmed true warnings
Training–Validation Set	127 (22)	1,071 (195)	42
Fold 1 Training Set	101 (16)	857 (156)	35
Fold 1 Validation Set	26 (6)	214 (39)	7
Fold 2 Training Set	101 (18)	857 (156)	32
Fold 2 Validation Set	26 (4)	214 (39)	10

Fold 3 Training Set	102 (18)	856 (156)	35
Fold 3 Validation Set	25 (4)	215 (39)	7
Fold 4 Training Set	102 (18)	857 (155)	34
Fold 4 Validation Set	25 (4)	214 (40)	8
Fold 5 Training Set	102 (18)	857 (157)	32
Fold 5 Validation Set	25 (4)	214 (38)	10
Testing Set	128 (30)	1,071 (185)	40

Recommendation

Minor Revision

This manuscript makes a useful methodological and operational contribution to a concrete false-alarm problem in geotechnical monitoring. The proposed architecture is well motivated and validated through a solid ablation study, and the failure-case analysis section reflects notable scientific rigour. However, the internal validity of some reported results depends on clarifications regarding the construction of the positive dataset and the exact scope of the normalisation steps — points that should be addressed prior to acceptance. Once these concerns are resolved, the manuscript has clear potential to make a valuable contribution to the field.