



From Sensor-Wise Accuracy to Episode-Wise Evidence: Auditing DAS Slope Monitoring at Hollin Hill and Brienzen

Gunjan Kumar Mishra¹ and Bijaya Ghimire¹

¹Department of Electronics and Computer Engineering, Institute of Engineering, Pashchimanchal Campus, Tribhuvan University, P.O. Box 46, Lamachaur, Pokhara, Nepal

Correspondence: Gunjan Kumar Mishra (mgunjan67@gmail.com)

Abstract. Dense distributed acoustic sensing (DAS) records can contain thousands of sensor rows but only one hazard trajectory, allowing row-wise evaluation to overstate warning evidence. We present a finite-record framework linking each claim to its valid unit, split, target provenance, background support, metric, and interpretation boundary. Hollin Hill contains one rainfall-triggered event. In a controlled section-split stress test, every one of 844 evaluation timestamps recurs in fitting, and clock probes reconstruct the shared chronology; purges from 30 min to 24 h reduce usable class support but cannot create independent events. A causal retrospective replay therefore reports alert burden, anchor-relative declaration delay, and rupture-horizon occupancy separately. Brienzen provides a second, clustered collapse sequence. Across ten grouped partitions, the balanced-accuracy increase from full sibling-view exposure is positive in every partition, while the one-sibling matched contrast is small. Forty-three of 97 radar intervals match 43 of 101 DAS episodes; a 5,000-draw within-day circular-shift null yields a median of 14 matches (95% range 8–20; empirical $p < 0.001$), with the association persisting after excluding 15 June. Sensor-dense data are not event-dense, and chronology, augmentation, episode construction, and incomplete reference catalogs must be represented before DAS results support engineering-warning claims.

1 Introduction

Distributed acoustic sensing (DAS) converts an optical fibre into a dense array of axial-strain observations and now supports applications from seismic imaging to near-surface monitoring (Jousset et al., 2018; Ajo-Franklin et al., 2019; Lindsey and Martin, 2021). A single cable can sample deformation along a slope more densely than conventional point instruments (Rossi et al., 2022; Wang et al., 2017; Schenato, 2017). At Hollin Hill, low-frequency DAS revealed head-scarp straining, rupture-zone development, retrogression, and lobe motion at nanostrain-rate sensitivity (Ouellet et al., 2024a). That observational richness makes the record valuable for asking how DAS evidence should enter an engineering-warning evaluation.

The statistical unit, however, remains the event rather than the channel or cable section. When one rainfall-triggered sequence is represented by many sensor rows, splitting by sensor can place the same timestamps, rainfall history, and global target in both model fitting and evaluation. The section-specific DAS features are not literal duplicates; the dependence arises because their chronology, target, and global covariates are shared. Classification scores can then answer whether another section reconstructs a known event phase while appearing to answer whether a future landslide is detected. This distinction is consequential for



engineering decisions because an operational warning assessment must jointly describe detection timing, exceedance burden, missed events, non-event monitoring, and target provenance (Intrieri et al., 2012; Piciullo et al., 2018; Stahli et al., 2015).

This paper asks a bounded question: *what can be inferred from finite, multi-sensor DAS records, and which evaluation choices create misleading apparent skill?* It contributes (1) a claim–unit–split protocol for finite-record warning evaluation; (2) a Hollin section-split stress test and retrospective burden analysis with declaration-time persistence and uncertain onset scenarios; and (3) a Brienz grouped-view experiment and partial-catalog episode audit. Hollin is not presented as operational prediction, and Brienz does not validate Hollin timing. The physical-grid ablation is confined to the Supplement.

2 Background

2.1 DAS slope monitoring and warning validation

DAS response depends on gauge length, cable direction, coupling, and the physical relation between axial strain and the incident wave or deformation field (Daley et al., 2013; Mateeva et al., 2014; Dean et al., 2016; Fichtner et al., 2022). Hollin Hill concerns slowly evolving low-frequency strain rate (Ouellet et al., 2024a, b); Brienz concerns short rock-slope signals amid vehicle, random, and narrow-band noise (Kang et al., 2024a, b). A 2026 typhoon case reports public 1 Hz DAS and meteorological data at two confirmed landslide locations, illustrating the emerging opportunity for multi-location evaluation while also showing that weather-correlated disturbance is not automatically landslide-specific (Zhu et al., 2026). These process domains differ, but all require explicit grouping, event definitions, and background burden.

Rainfall thresholds and local or territorial landslide early-warning systems have long required explicit treatment of threshold exceedance, missed events, false alarms, warning duration, and decision context (Caine, 1980; Guzzetti et al., 2007; Aleotti, 2004; Baum and Godt, 2009; Crosta and Frattini, 2003; Segoni et al., 2018). Frameworks for local and regional warning systems emphasize that a warning event is not interchangeable with a row-wise classification (Intrieri et al., 2012; Calvello and Piciullo, 2016; Piciullo et al., 2016, 2018, 2020). Displacement-based time-of-failure methods similarly depend on how alarm thresholds are set and on whether apparently accelerating motion is sustained (Rose and Hungr, 2007; Dick et al., 2015; Carla et al., 2016; Intrieri et al., 2019; Carla et al., 2019). These studies motivate our decision to report anchor-relative declaration delay and alert occupancy separately, without turning a pre-anchor alarm into a detection success.

2.2 Dependent samples, grouped evaluation, and leakage

Pseudoreplication arises when multiple measurements are analysed as if they were independent experimental units (Hurlbert, 1984; Lazic, 2010). In predictive modelling, leakage occurs whenever information unavailable for the intended target setting enters fitting, selection, or evaluation (Kaufman et al., 2012; Kapoor and Narayanan, 2023). Grouped, temporal, and spatial data require splits that respect the dependence structure (Roberts et al., 2017; Bergmeir et al., 2018; Cerqueira et al., 2020; Tashman, 2000). Time-series anomaly research has further exposed run-to-failure bias, flawed benchmark composition, and point-adjusted or event scoring that can reward trivial detectors (Wu and Keogh, 2023; Kim et al., 2022; Liu and Paparrizos,



2024). A section split can be useful for studying same-time spatial phase reconstruction, but it cannot substitute for future-event or cross-storm validation.

2.3 Event metrics and uncertain targets

Precision–recall, ROC, and proper scoring rules summarize row-wise discrimination or probability quality, but they do not by themselves define an alert episode (Brier, 1950; Fawcett, 2006; Davis and Goadrich, 2006; Gneiting and Raftery, 2007; Hand and Till, 2001; Saito and Rehmsmeier, 2015). Event-based scoring must define episode merging, tolerance, latency, and burden (Molchan, 1990; Tatbul et al., 2018). When onset is known only between observations or from approximate published anchors, interval-censoring principles require the interval to remain visible rather than being converted into artificial temporal precision (Turnbull, 1976; Finkelstein, 1986). We accordingly report deterministic scenario ranges, not confidence intervals. Empirical day resampling for Brienz is explicitly conditional on the observed monitored days and is not interpreted as population uncertainty. More generally, stating the target setting before fitting is an estimand problem, not just a model-choice problem (Hernan and Robins, 2016).

3 Finite-record evaluation framework

We distinguish six conceptual claim levels: in-event description; same-time cross-section reconstruction; future-time performance within a storm; cross-storm performance at one site; cross-event performance; and cross-site prospective performance. Table 1 consolidates them into four broader validation classes and adds partial-catalog agreement as a separate catalog estimand. Before fitting, an evaluation should state its intended claim, independent unit, permissible split, target provenance, non-event support, metric, and prohibited interpretation. The present audit applies this prospective template retrospectively; it does not imply that the case-study analyses were preregistered.

The framework treats sensor replication, same-time target exposure, circular target provenance, overlapping windows, chronology confounding, threshold selection, and domain ambiguity as distinct failure modes. Each generated result carries an estimand label. Table 1 condenses the framework; its “do not claim” column is part of the protocol rather than a repeated narrative disclaimer.

4 Study records

4.1 Hollin Hill

The public Hollin Hill release contains six processed 1 Hz strain-rate NetCDF files, 30 min rainfall observations, ShapeArray records, and channel coordinates in OSGB36 (Ouellet et al., 2024a, b). The record spans 12 January 2021 11:38:10 to 15 January 14:39:59 UTC and contains one rainfall-triggered slope event. A controlled audit construction created here assigns 5,064 section rows to 844 common sequence endpoints. Each trailing 30 min sequence contains six 5 min frames; 201 endpoints belong to the published head-scarp, rupture-triggering, or peak-deformation phases. This section-split stress test represents



Table 1. A prospective evaluation template applied retrospectively in the present audit.

Intended claim	Valid unit and split	Evidence requirements	Report / do not claim
Same-record reconstruction	Timestamp or label group; group all replicated views	Disclose sensor-derived targets and shared chronology	Group AP/BAcc and recall / future-event skill
Retrospective event description	One slope timestamp and declared episode; forward replay	Sequential pre-event estimation; interval-valued targets	Burden, occupancy, scenario-relative delay / preregistration or false-alarm rate
Future-period event detection	Independent episode; forward date split with grouped views	External or blind intervals and representative background	Event match, declaration delay, unmatched burden / window precision as event skill
Cross-event or cross-site warning	Independent events, storms, or sites; prospective holdout	Multiple failures and non-failures with independent adjudication	Detection, misses, burden, calibration / rows, channels, seeds, or days as event replicates
Partial-catalog agreement	Catalog and DAS episodes; one-to-one assignment	Explicit monitored duration, tolerance, and catalog limitations	Conditional agreement fractions / false positives, F1, or complete-ground-truth precision

a realistic failure mode whenever spatial sensor rows share one chronology and target trajectory. It was not proposed as a benchmark by the field-study authors.

The separate retrospective replay contains 8,718 slope-level timestamps (72.65 monitored h) after validity filtering and seven acquisition gaps. Rainfall begins at 13 January 12:00 UTC. Approximate published DAS anchors, deterministic timing intervals, and the ShapeArray-derived movement proxy are described with their provenance in Methods. Figure 1 shows the public cable geometry and the evaluation sequence used in this audit.

4.2 Brienzen

The Brienzen release supplies 1,491 manually labelled groups with 15 augmented views per group, four-class probability outputs, radar intervals, and noise summaries (Kang et al., 2024a, b). The classes are C0 vehicle noise, C1 random noise, C2 slope failure, and C3 narrow-band noise. The published protocol fits on 1–14 June 2023 and evaluates from 15 June onward. Radar informed the original label development and is incomplete as a failure catalog; the 97 future-date intervals are therefore a partial sensor reference, not complete ground truth. Table 2 distinguishes rows, grouped units, events, and monitored duration for both records.

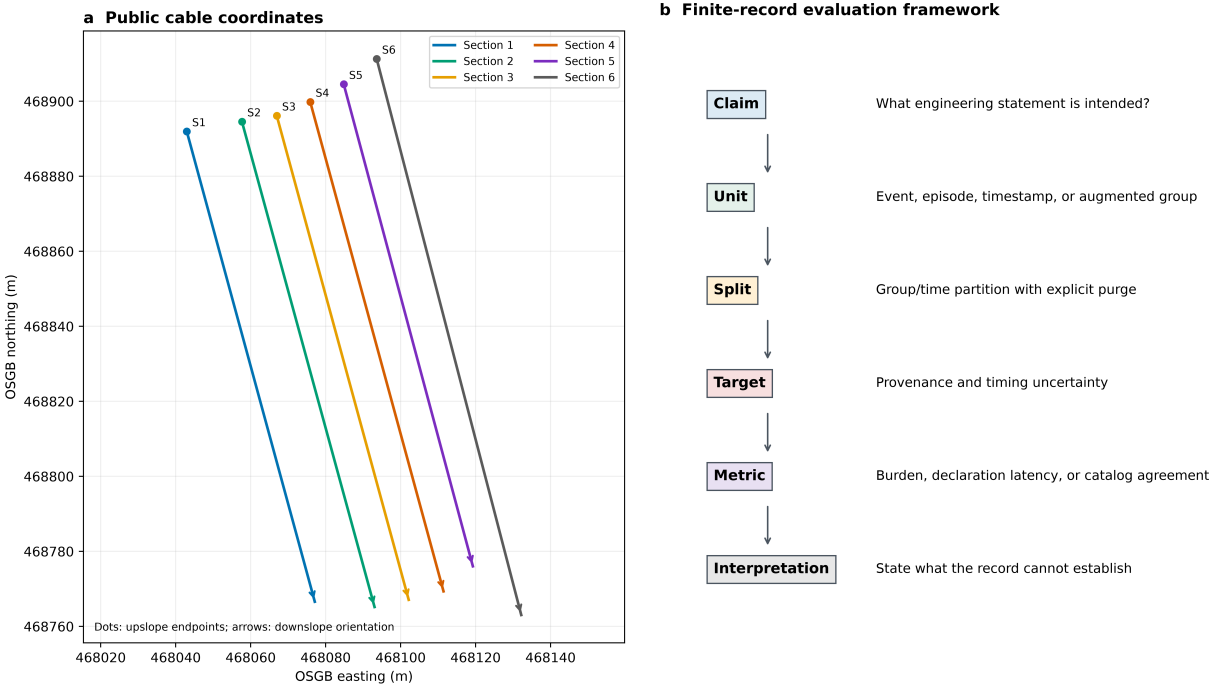


Figure 1. Site geometry and evaluation framework. (a) Public Hollin Hill cable coordinates; points mark upslope endpoints and arrows the common downslope orientation. (b) The claim–unit–split–target–metric sequence used to determine the allowed interpretation before fitting.

Table 2. Observation, grouping, event support, and monitored duration. Counts describe data layout rather than independent scientific replicates.

Component	Records / grouped unit	Event support	Duration
Hollin section stress test	5,064 rows / 844 endpoints	One target trajectory; one storm/event	75.03 h
Hollin retrospective replay	8,718 slope timestamps	Same Hollin storm/event	72.65 h after gaps
Brienz labels	22,365 views / 1,491 groups	381 future-date groups	1–30 June protocol
Brienz event replay	64,093 windows / 101 episodes	97 partial-catalog intervals over 16 days	364.6 h

5 Methods

5.1 Hollin section-split stress test and clock probes

We first reproduce the section split. Sections 1–4 form the nominal fitting rows and section 6 the nominal evaluation rows; because global rainfall and labels are copied, all 844 evaluation timestamps also occur in fitting. We then collapse to one row per timestamp and use 6 h calendar blocks with 30 min, 6 h, and 24 h purges. The 24 h setting also removes overlap in the longest rainfall-history feature. These blocked analyses are support diagnostics, not event-generalization estimates, because the positive state is contiguous and most test blocks contain one class.



Clock probes use elapsed hours only: a linear logistic model, a cubic spline of degree 3 with six knots, a depth-4 tree, 1-nearest-neighbour time, and an exact timestamp lookup oracle. A flexible clock can represent a bounded phase interval, unlike a monotone linear score. Perfect oracle reconstruction is expected whenever the same timestamp occurs on both sides of the section split.

110 The reported DAS-only and DAS-plus-rain stress-test models use six consecutive 5 min frames per endpoint. Each frame is a nearest-neighbour-resampled 96×96 time-channel array, robust-normalized by section over 12 January 2021 11:38:10 through 13 January 05:38:10 UTC (inclusive) and clipped to ± 8 . A three-block CNN (8/16/32 channels) produces 48-dimensional frame embeddings; a one-layer 56-unit GRU, 24-unit attention layer, and 32-unit dropout head produce the binary score. The rain variant adds endpoint rainfall over 30 min, 1 h, 6 h, and 24 h through a 16-unit encoder. Sections 1–4 fit the model,
115 section 5 selects the best of six AdamW epochs by average precision, and section 6 is evaluated; batch size is 32, learning rate 8×10^{-4} , weight decay 10^{-4} , and seed 42. The rainfall logistic uses the same four rain features, training-set median imputation and standardization, class balancing, and 5,000 maximum iterations on the same section split. A paired sensitivity refits both neural variants for seeds 42, 61, and 101 using either section-specific pre-event parameters or one center and scale pooled only from sections 1–4; the latter never uses sections 5 or 6 to fit normalization. Supplement Sect. S2.1 gives the exact model,
120 interval, and fitting specification.

5.2 Trailing-window slope score

For section s , component $k \in \{\text{RMS}, p95\}$, and score time t , let $x_{s,k}(t)$ use only samples in $[t - 59\text{s}, t]$. The $p95$ component is the 95th percentile of absolute strain rate over the section-window cells. Values below 10^{-30} are floored before $\log_{10}$, and any zero or extremely small MAD is floored at 10^{-12} . The normalization-fit segment $\mathcal{N}$ defines

$$125 \quad z_{s,k}(t) = \frac{\log_{10} x_{s,k}(t) - \text{median}_{\mathcal{N}}(\log_{10} x_{s,k})}{1.4826 \text{ MAD}_{\mathcal{N}}(\log_{10} x_{s,k})}, \quad (1)$$

where $\mathcal{N} = [12 \text{ January } 2021 \text{ } 11:38:10, 13 \text{ January } 06:00:00) \text{ UTC}$. Threshold calibration uses $[13 \text{ January } 06:00, 10:00) \text{ UTC}$, the untouched background holdout uses $[10:00, 12:00) \text{ UTC}$, and rainfall begins at 12:00 UTC. The section score is $a_s(t) = \max_k z_{s,k}(t)$. The slope-level score is the mean of the two largest available section scores,

$$A(t) = \frac{1}{2} \{a_{(1)}(t) + a_{(2)}(t)\}, \quad (2)$$

130 so each timestamp is one slope observation rather than six nominal samples. Requiring two elevated sections reduces sensitivity to an isolated section spike, while retaining a distributed response; it may miss a genuinely local precursor and is therefore treated as a fixed transparent design choice, not an optimized rule. Because top-two order statistics depend on the number of available sections, the generated records include the valid section identities at every timestamp, section-availability summaries around each target and horizon, and sensitivity rows for a fixed complete-section subset, a five-of-six support rule, and
135 conditional-by-support-count calibration.

The primary displayed threshold range is 0.95–0.995, crossed with persistence lengths of 1, 2, 5, 10, and 20 steps. The 0.999 quantile is retained in the complete 25-setting archive only as an exploratory empirical-resolution sensitivity. The illustrative



lower and upper levels are the 0.990 and 0.995 quantiles of 13 January 2021 06:00–10:00 UTC (4.715 and 4.934 robust-score units). NumPy linear quantiles (Hyndman–Fan type 7) are used. With $n = 480$, the 0.995 value interpolates between the fourth- and third-largest scores. The calibration diagnostics therefore report raw exceedance rows and contiguous exceedance clusters; calibration-length and boundary sensitivities support the tail assessment.

Score-series dependence is summarized separately. Let x_{ci} be score i in contiguous segment c , and let $\bar{x}$ be the pooled mean. The biased autocorrelation estimate is

$$\hat{\rho}_k = \frac{\sum_c \sum_{i=1}^{n_c-k} (x_{ci} - \bar{x})(x_{c,i+k} - \bar{x})}{\sum_c \sum_{i=1}^{n_c} (x_{ci} - \bar{x})^2}, \quad (3)$$

so no lag pair crosses an acquisition gap. Lags extend to $\min(240, \lfloor n/2 \rfloor)$. Following Geyer’s initial-positive-sequence rule (Geyer, 1992), define $\hat{P}_0 = \hat{\rho}_0 + \hat{\rho}_1 = 1 + \hat{\rho}_1$ and $\hat{P}_j = \hat{\rho}_{2j} + \hat{\rho}_{2j+1}$ for $j \geq 1$. Let J be the last consecutive pair before the first non-positive $\hat{P}_j$. Then

$$n_{\text{eff}} = \frac{n}{-1 + 2 \sum_{j=0}^J \hat{P}_j}. \quad (4)$$

The calibration segment has no acquisition gap and yields $n_{\text{eff}} \approx 14.7$. This is a dependence diagnostic for the score series, not an effective sample size for an extreme quantile.

For persistence m and 30 s stride Δt , t_{run} is the first threshold exceedance and the causal declaration time is

$$t_{\text{declare}} = t_{\text{run}} + (m - 1)\Delta t. \quad (5)$$

Thus five-step and 20-step declarations occur 120 and 570 s after run onset. Anchor-relative declaration delay and active-state burden use t_{declare} ; run onset is retained only for retrospective signal description. State terminates below level or at an acquisition gap. Parameters are estimated from sequential pre-rain segments and remain fixed during replay; the overall design is retrospective. The 0.999 quantile is an empirical-resolution diagnostic, and the short background holdout does not estimate a false-alarm rate.

5.3 Anchor-relative timing, horizons, and label provenance

Anchor-relative timing uses the first *new declared* exceedance episode whose declaration is at or after the target anchor. An episode declared earlier remains pre-anchor burden even if it crosses the anchor. For every possible 30 s anchor throughout each analyst-defined interval, we report active-at-anchor state, preceding burden, and delay to the next declaration. Fixed 1, 3, 6, and 12 h pre-rupture occupancy is reported separately and is not a true-positive count. Structured output fields retain the earlier `onset_detection` and `latency` names for backward traceability; the article interprets them only as analyst-anchor-relative quantities.

The public analysis notebook is available at <https://github.com/smouellet/hhdas>. The inspected commit was 2b6179395531a3f07964cbb2c535e5e2d53bfff50. The notebook plots the event record but contains no recoverable sequence-box coordinates. We therefore use deterministic ± 60 min intervals around the approximate published



Table 3. Target and reference ontology. Independence is stated relative to the evaluated DAS score, not as a claim that sensors measure identical physical constructs.

Target or reference	Source and definition	Timing / provenance	Independence from evaluated DAS score
Hollin head-scarp and rupture anchors	DAS; approximate interpretations with analyst ± 60 min scenarios	Interval-valued; analyst scenario	No; DAS-derived
Hollin ShapeArray proxy	ShapeArray; two samples above pre-rain median +3 MAD	Observation interval	Separate sensor and score; different construct
Brienz C2 labels	Manual slope-failure class groups	Window labels	Not blind; radar informed development
Brienz radar intervals	Ground-based radar start/end intervals	Interval-valued partial catalog	Separate sensor; incomplete reference
Article confirmation stage	Unreleased adjudication; 54 aggregate confirmations	Individual timings unavailable	Cannot be independently audited

anchors: head-scarp onset, 13 January 2021 19:00–21:00 UTC; rupture-zone triggering, 14 January 2021 02:00–04:00 UTC; and peak strain rate, 14 January 2021 11:00–13:00 UTC. The one-hour half-width is an analyst scenario choice for approximate clock descriptions; it is not estimated from the data and is not a confidence interval. Their provenance remains circular with DAS to the degree stated in Table 3.

The separate ShapeArray-derived movement proxy is reconstructed sensor-first. For each lobe, sensors with complete XYZ support form a fixed set. Vector velocity is calculated per sensor before aggregation; the primary statistic is the 90th percentile of sensor speed. Onset requires two consecutive post-rain statistics above the pre-rain median plus 3 unscaled MAD, with 2/3/4-MAD and leave-one-sensor-out sensitivities. ShapeArray is not treated as rupture truth.

Physical-grid construction and the ten-seed DAS-only preprocessing ablation are documented in the Supplement. They check geometry, masking, and normalization but do not supply independent event evidence or select the replay score.

5.4 Brienz grouped-view experiment

The released Brienz table has 1,491 label groups, 15 views per group, and 14,336 features per view. This audit deterministically compacts each view to 448 values: mean and RMS within 16 channel blocks for each of 14 frequency rows. Released view 0 is the first stored view; the release does not establish that it is an unaugmented raw observation. The split is applied only after this released representation exists, so the controlled estimand is downstream sibling-view exposure within the released table. Raw-signal feature construction and any upstream global exposure remain unresolved; Supplement Sect. S5 provides the information-flow ledger.

All 15 released views remain attached to their manual label group. The broad naive-view, grouped-label, and published-date splits are reported as different estimands. The original controlled experiment fixes one class-stratified partition (seed



20260712) with 373 held-out groups for ten ExtraTrees algorithmic repetitions. A robustness audit instead uses ten independently generated class-stratified group partitions, three independent class-matched removal draws per partition, and fixed model randomness. Evaluation always uses released view 0. The grouped baseline excludes all sibling views. Full exposure includes the other 14 siblings and more fitting rows. Class-matched controls expose 1, 5, or 14 siblings per test group while removing the same number of baseline rows without replacement, preserving fitting-row and class counts. The model uses 160 extremely randomized trees, square-root feature sampling, minimum leaf size two, and class balancing.

5.5 Brienz protocol reconciliation and episode matching

The reconciliation ladder separates the released notebook’s automatic stage from the article’s unreleased manual-confirmation stage, then applies the future-date restriction and a 0.75 audit threshold. A qualifying output has argmax class C2 and $P(C2) \geq 0.75$. Let u_i be qualifying window-start times on the 20.48 s grid. Adjacent windows merge when $u_i - u_{i-1} < 60$ s; this is start-to-start separation, not silent time. An episode support is $[u_{\text{first}}, u_{\text{last}} + 25.6\text{s}]$, and its first classifier decision is available at $u_{\text{first}} + 25.6$ s.

For radar interval $R = [r_s, r_e]$, DAS episode $D = [d_s, d_e]$ is eligible when $d_e \geq r_s - 60\text{s}$ and $d_s \leq r_e + 60\text{s}$. Thus a long episode can be eligible through any support overlap. Pairwise boundary distance is

$$b(R, D) = \begin{cases} r_s - d_e, & d_e < r_s, \\ d_s - r_e, & d_s > r_e, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Chronologically sorted eligible pairs receive a one-to-one assignment that first maximizes cardinality, then minimizes total b , then total absolute support-start offset; exact residual ties follow chronological indices. The support-start offset $d_s - r_s$ describes signal support. The first-window-completion offset is $(d_s + 25.6\text{s}) - r_s$; it assumes zero acquisition, computation, and telemetry delay and is neither measured operational latency nor warning lead time. Negative support offsets therefore need not imply an available early decision.

The 108-setting surface is 3 probability thresholds $\times$ 3 start gaps $\times$ 3 tolerances $\times$ 2 assignment modes $\times$ 2 June-15 policies. The primary 43/97 quantity is maximum-cardinality agreement conditional on the selected episode definition and ± 60 s tolerance. We also report episode confirmation, unmatched episodes per monitored hour, unmatched duration, size/noise strata, daily counts, leave-one-day-out values, and 2,000 calendar-day resamples. The resamples describe the observed days rather than population uncertainty. A separate 5,000-draw temporal null circularly shifts each day’s complete episode sequence within monitored support, preserving episode number, durations, within-day relative spacing, radar intervals, and the one-to-one matching rule. It tests whether temporal alignment exceeds that expected from the observed clustering; it does not create independent events or estimate out-of-sample detection probability.



215 6 Results

6.1 Hollin timestamp reuse and clock probes

The 5,064 Hollin rows are six section-indexed observations at each of 844 endpoints: DAS features differ, while timestamp, target, and global covariates are shared. All 844 nominal section-6 timestamps occur in the section 1–4 fitting rows. The actual DAS-only temporal model reaches AP 0.514 and the DAS-plus-rain model AP 0.982 on section 6, but these remain same-time cross-section estimates. The rainfall logistic reaches AP 0.998. The spline reaches AP 1.000 and the tree, 1-nearest-neighbour, and timestamp oracle reach 1.00. These probes demonstrate vulnerability to chronology reconstruction; they do not establish the fitted neural model’s decision mechanism.

Purging removes endpoint overlap, but it does not manufacture event support: only 2 of 13 six-hour test blocks contain both target classes at each of the 30 min, 6 h, and 24 h purges. At 24 h, only 7 of 13 folds also retain both classes in fitting. Paired normalization transfer is likewise model- and seed-dependent: pooled training-section parameters change DAS-only BAcc by -0.034–0.097 and DAS-plus-rain BAcc by 0.001–0.209 across three paired seeds. Neither sensitivity supplies independent-event evidence. Table 4 reports the intervention-specific estimands, and Fig. 2 shows the complete ladder.

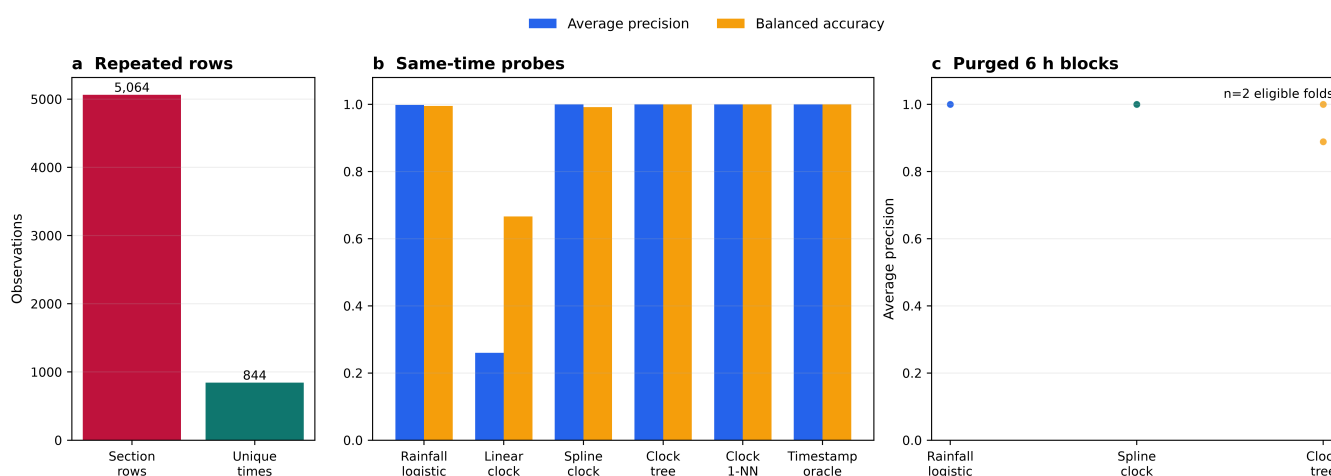


Figure 2. Leakage ladder for Hollin Hill. Replicated section rows reduce to unique timestamps (left). Under the section split, global rainfall and flexible clock probes reconstruct the repeated target trajectory (centre). Purged time blocks remove exact endpoint overlap, but only two test folds contain both classes at every purge duration; the 24 h purge further reduces fitting support (right). Empty-class folds are not pooled into a generalization score.



Table 4. Leakage interventions and conditional results. AP is average precision; BAcc is balanced accuracy; Δ is exposed minus grouped baseline on identical evaluation rows.

Construction	Split / intervention and support	Result	Interpretation
Hollin section split	Rainfall logistic; 844/844 times reused	AP 0.998; BAcc 0.995	Shared rainfall and target chronology
Hollin section split	Cubic-spline clock; same times reused	AP 1.000; BAcc 0.991	Flexible time reconstructs phase
Hollin section split	Timestamp oracle; same times reused	AP/BAcc 1.000	Exact same-time target exposure
Hollin blocked time	6 h blocks; 30 min–24 h purges; 0 endpoint overlaps	2/13 mixed test folds	One event leaves inadequate class support
Brienz paired control	Full sibling exposure; ten grouped partitions	median Δ BAcc 0.042 (0.018–0.063)	Positive in 10/10 partitions
Brienz paired control	One sibling; size/class matched	median Δ BAcc 0.012 (-0.003–0.022)	Positive in 9/10 partitions
Brienz published dates	Group all 15 views; 0 group overlap	BAcc 0.817	Future-date group estimand only

6.2 Hollin burden and anchor-state results

The replay has 8,718 unique score times and 72.65 monitored h, including 48.3 h after rainfall. The background holdout contains 0 min of lower and 0 min of upper exceedance; this is only short-record behavior. Post-rain valid sections range from 6 to 6, with complete support for 100.0% of times. The 0.95–0.995 threshold–persistence surface is primary; 0.999 remains an exploratory archived sensitivity. The illustrative 0.990/0.995 traces occupy 16.1/15.7 h (33.2/32.4%) after rainfall in 10/7 episodes. Timing is not summarized as a precise post-rupture delay: Fig. 4 evaluates every possible anchor throughout the rupture interval and shows the discontinuity when an episode changes from a future declaration to active-at-anchor burden.

At the rupture point anchor, both illustrative exceedance states have zero occupancy in fixed 1, 3, 6, and 12 h horizons. Across the full 14 January 2021 02:00–04:00 UTC anchor interval, however, early scenarios await a new episode whereas late scenarios begin inside an existing episode (Table 5). The latter state is burden/coverage, not detection. Guaranteed occupancy is zero, while non-zero occupancy is possible for 16.6% of lower-level and 15.4% of upper-level anchor scenarios; the complete horizon matrix is in the Supplement. The sensor-first 3-MAD ShapeArray proxy is 14 Jan 11:00–14 Jan 13:00 UTC and follows the DAS-derived peak interval rather than establishing an independent pre-rupture warning. The score is strongly associated with rainfall and cannot separate movement from hydrological strain, temperature, coupling, or common-mode environmental response because no rainfall-without-failure period exists. Section-localization summaries are therefore descriptive, not attribution of mechanism.

Fig. 3 aligns the finite record without assigning causal meaning to the score. Fig. 4 shows burden trade-offs and continuous onset-scenario timing; it does not select a setting against the event.



Table 5. Rupture-anchor state segments on 14 January 2021 (UTC). Run onset is the first threshold exceedance and is available only retrospectively; declaration follows the fifth consecutive exceedance, 120 s later. State and anchor-relative delay use declaration time. Scenario segments are exhaustive deterministic timings, not probability or confidence intervals.

Level	Anchor scenario	State at anchor	Run onset / declaration	Quantity
Lower (0.990)	02:00:00–03:40:00	Awaiting declaration	03:38:09 / 03:40:09	0.1–100.2 min delay
Lower (0.990)	03:40:30–04:00:00	Declared state active	03:38:09 / 03:40:09	0.3–19.8 min prior burden
Upper (0.995)	02:00:00–03:41:30	Awaiting declaration	03:39:39 / 03:41:39	0.1–101.7 min delay
Upper (0.995)	03:42:00–04:00:00	Declared state active	03:39:39 / 03:41:39	0.3–18.3 min prior burden

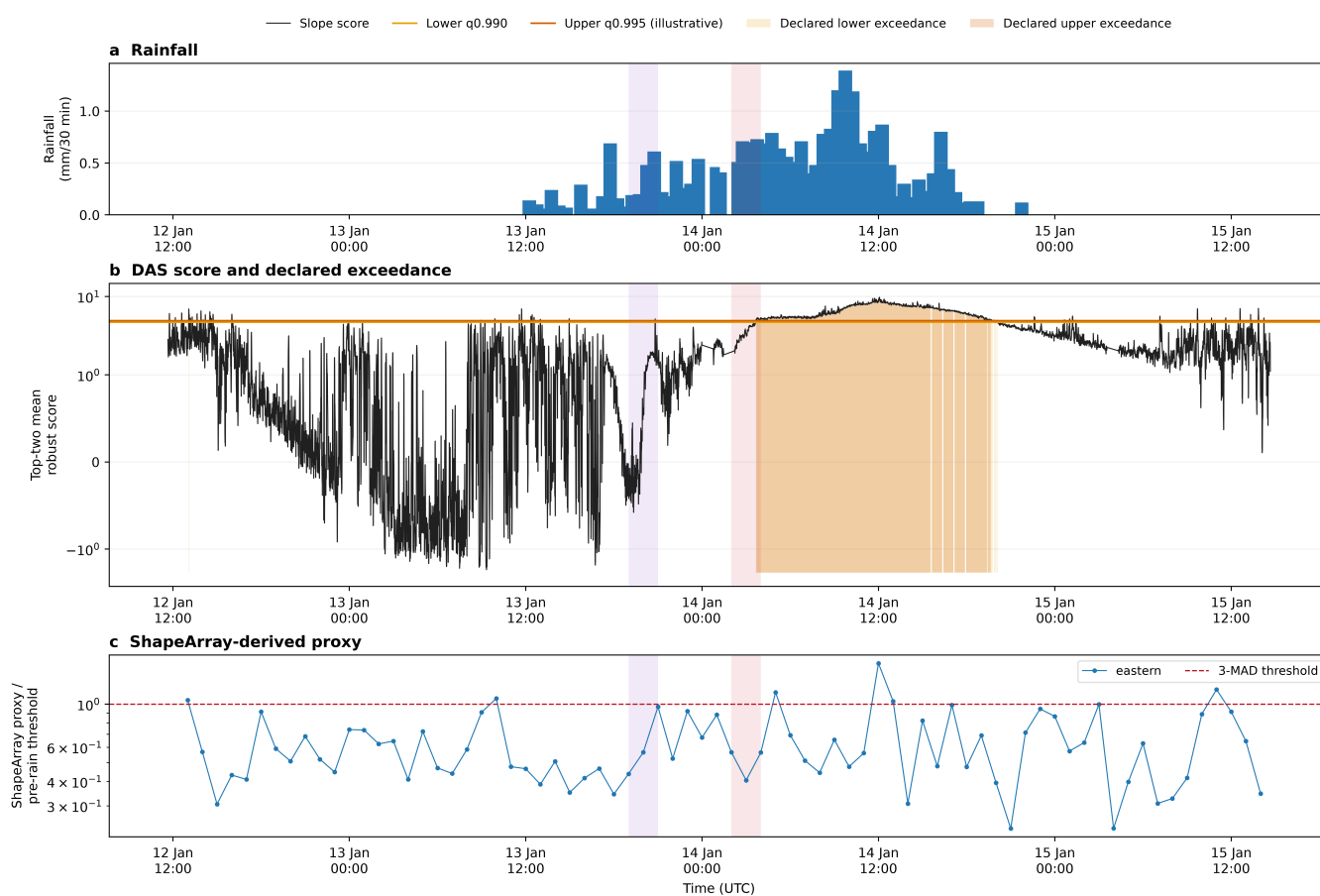


Figure 3. Retrospective Hollin replay with rainfall, the slope-level DAS score, and a separate ShapeArray-derived movement proxy. DAS windows are strictly trailing, and normalization and thresholds are estimated from sequential pre-rain segments; the overall protocol was designed retrospectively. Markers show declaration after five consecutive exceedances, not backdated run onset. Shading denotes deterministic anchor scenarios. Acquisition gaps remain unmonitored and reset persistence.

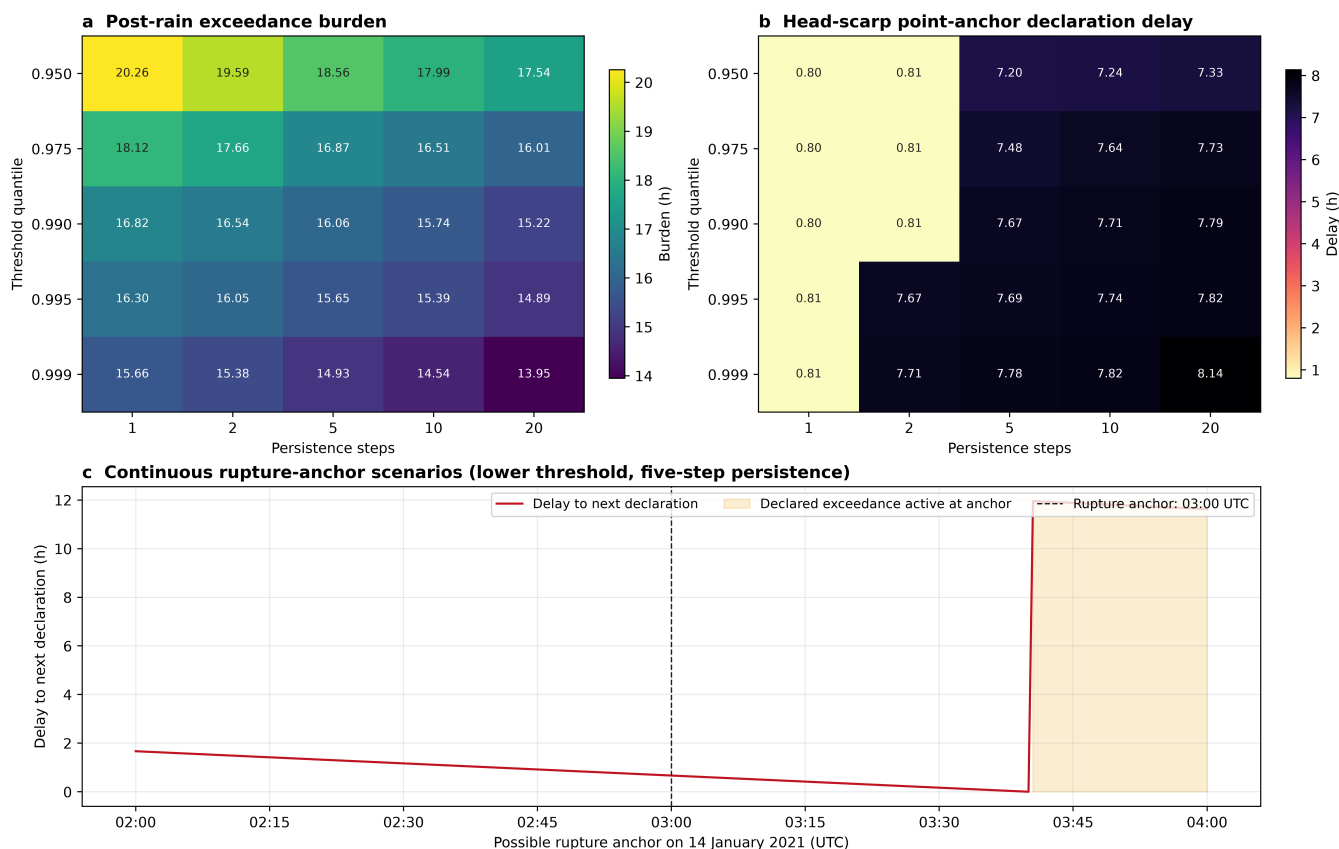


Figure 4. Burden and continuous anchor-scenario timing. Panel (a) shows post-rain exceedance burden for the full archived quantile–persistence surface; 0.999 is exploratory. Panel (b) shows declaration delay from the exact 13 January 2021 20:00 UTC head-scarp point anchor. Panel (c) evaluates every possible anchor in the 14 January 2021 02:00–04:00 UTC rupture scenario; an already declared episode is pre-anchor burden, not a new declaration.

6.3 Brienz grouped-view and episode results

The broad Brienz splits change both leakage and test composition, so their BAcc values (0.943, 0.900, and 0.817) are different estimands. In the ten-partition robustness audit, the full 14-sibling contrast has median ΔBAcc 0.042 (range 0.018–0.063) and is positive in 10/10 partitions. Size/class-matched medians are 0.012, 0.032, and 0.045 for 1, 5, and 14 exposed siblings; the one-sibling range is -0.003–0.022 and is positive in 9/10 partitions. The original fixed-partition ten-seed means (0.046 for full exposure and 0.009 for one matched sibling) remain archived as optimization sensitivity, not scientific replication. The full contrast combines sibling exposure with additional fitting rows, and all inference is confined to the released feature table because upstream representation exposure is unresolved.

The released notebook’s automatic stage rerun contains 727 radar events, 993 C2 windows, 295 matched radar events (0.406), and 896 radar/day-rule window hits (0.902). The article adds manual confirmation of 54 of the remaining 97 windows, giv-



ing 950/993 (0.957); those individual adjudications are not released, so we do not claim to reproduce that manual stage. After future-date restriction and the primary episode protocol, 101 episodes are compared with 97 radar-catalog intervals. Maximum-cardinality matching gives 43/97 conditional agreements (catalog fraction 0.443) and pairs 43/101 DAS episodes (radar-confirmation fraction 0.426). The temporal null gives a median of 14 matches (95% range 8–20) versus 43 observed, with empirical $p = 0.0002$. Excluding 15 June gives 9 observed versus a null median of 0 (95% range 0–1; $p = 0.0002$), showing alignment beyond the preserved within-day clustering without implying event independence. Median support-start offset is -10.3 s, whereas the median first-window-completion offset is 15.3 s after adding the 25.6 s window duration; the latter excludes acquisition, processing, and telemetry delay. There are 58 unconfirmed episodes, or 0.159 per monitored hour; they are not labelled false detections. Table 6 reports the conditional episode results, and Fig. 5 shows their temporal and stratified context. The complete reconciliation ladder and 108-setting sensitivity surface are supplied.

Table 6. Brienz episode-level partial-catalog results overall and by radar-event size. Agreement is matched/catalog events (fraction). Negative support offsets begin before the radar interval. The first-window-completion offset adds the 25.6 s window but excludes acquisition, processing, and telemetry delay. Timing values are median [IQR] in seconds.

Stratum	Catalog agreement	Support-start offset	First-window completion offset
All future-date events	43/97 (0.443)	-10.3 [-23.8, 8.4]	15.3 [1.8, 34.0]
Size: small	11/42 (0.262)	-7.6 [-13.3, 10.4]	18.0 [12.3, 36.0]
Size: medium	12/30 (0.400)	-11.2 [-24.7, -8.3]	14.4 [0.9, 17.3]
Size: large	20/25 (0.800)	-12.1 [-36.6, 12.4]	13.5 [-11.0, 38.0]

Overall DAS-episode radar-confirmation fraction: 0.426. The 58 unconfirmed episodes contribute 0.002 h of unmatched episode state per monitored hour. Overall episode duration is 46.1 s (90th percentile 128.0 s). The 2,000 day-resampling range is an empirical sensitivity, not a population confidence interval.

7 Discussion

7.1 What the Hollin benchmark estimates

The central result is not that rainfall or a clock is an excellent predictor. It is that the section split exposes the same target trajectory at the same timestamps. The flexible probes and oracle make that estimand visible. A high score under this protocol measures within-event chronology or spatial phase reconstruction. It cannot estimate future-event detection probability, cross-storm calibration, or an operational false-alarm rate. Collapsing evaluation rows after fitting is also insufficient because the repeated timestamp has already entered fitting.

The forward-only retrospective replay answers a narrower question: after sequential pre-rain fitting, calibration, and holdout, when would this statistic have exceeded fixed levels during the observed record, and at what burden? It is not causal inference and its design is retrospective. Continuous anchor sensitivity prevents an episode that crosses an uncertain interval from being relabelled as a precise detection success.

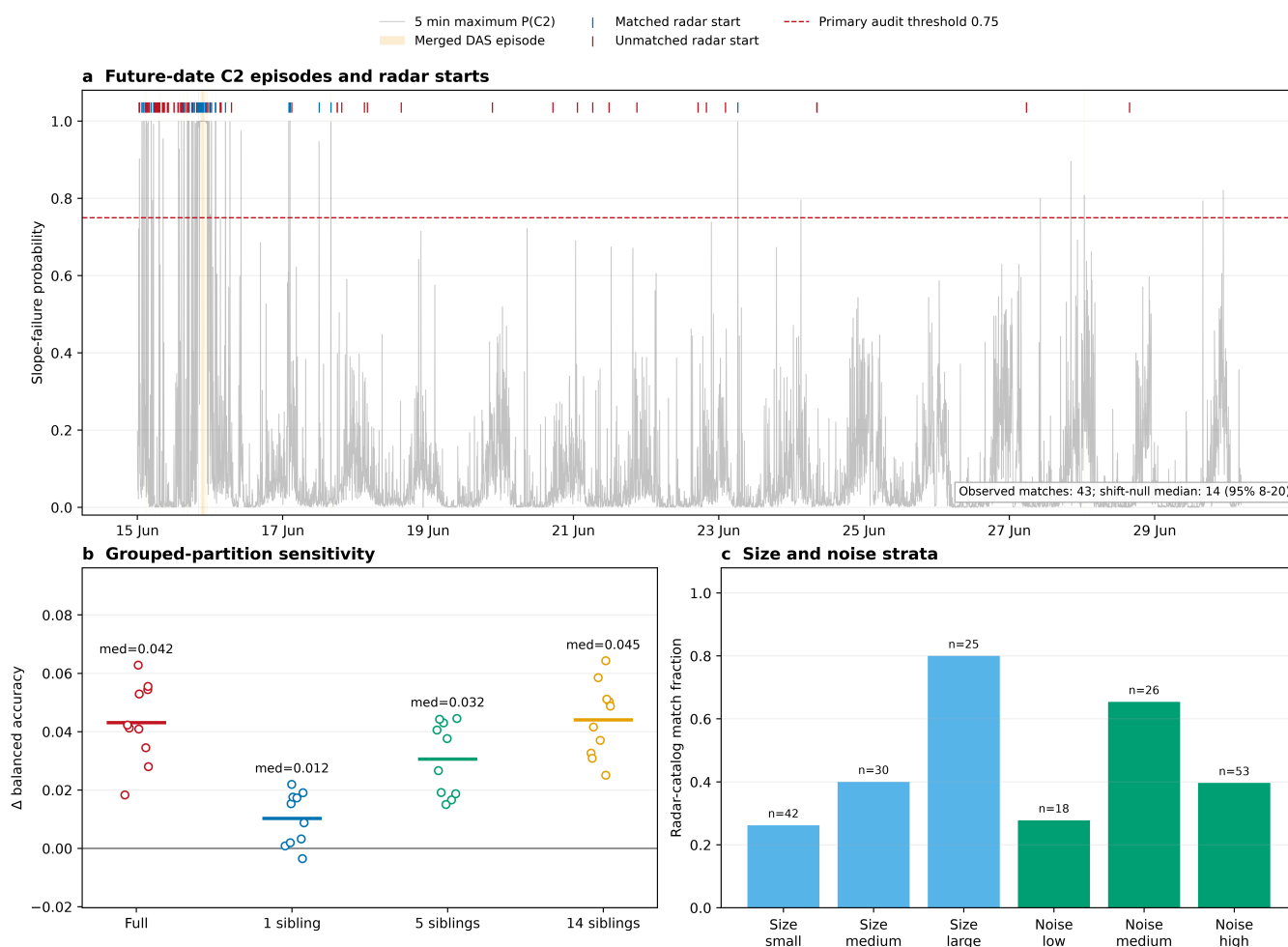


Figure 5. Brienz second-dataset triangulation. (a) Future-date C2 probabilities with merged DAS episodes, radar starts, and the timing-preserving null summary. (b) BAcc differences across ten class-stratified label-group partitions; each point is one partition after averaging its three removal draws. (c) Conditional radar-catalog agreement by event-size and pre-event RMS-noise strata.

7.2 Target provenance and physical interpretation

The head-scarp and rupture anchors remain partly circular because published DAS evidence helped define them. Their intervals are analyst scenarios, not confidence bounds. Sensor-first ShapeArray velocity moves the 3-MAD proxy to 14 Jan 11:00–14 Jan 13:00; this change from the earlier maximum-first calculation demonstrates sensitivity to proxy construction. Its hourly sampling, fixed-sensor choice, p90 aggregation, MAD floor, and leave-one-sensor-out results remain visible. Stronger studies should preregister targets from blind field observations, radar, GNSS, pore pressure, or independent expert annotation.



7.3 Second-dataset triangulation and evaluation implications

Brienzi shows that the evaluation critique is not confined to the Hollin construction. Partition-level controls quantify sibling-view exposure, while the date split asks a separate temporal-shift question. The temporal null shows that the observed episode–radar alignment is much stronger than expected after preserving the observed within-day episode clustering and monitored support. It remains an association within one collapse sequence: the 97 radar intervals are not 97 independent geotechnical validation events. Partial-catalog terminology prevents unmatched DAS episodes from being treated as known false positives.

Day sensitivity is dominated by 15 June. Resampling 16 consecutive dates cannot create new collapse mechanisms, background regimes, or sites; its percentile range is therefore not inferential uncertainty. Engineering decisions require multiple failure and non-failure periods, explicit costs for missed and prolonged exceedances, and thresholds frozen before deployment.

7.4 Limitations and evidence needed next

Hollin Hill supplies one rainfall episode and one interpreted process sequence. There are no independent non-failure storms, and the background record is short. The transparent score is not claimed optimal. Its top-two aggregation can suppress a genuinely local one-section precursor, while reducing sensitivity to isolated section noise; the availability records show the section support under which that order statistic is evaluated. The ShapeArray rule is an audit target with hourly resolution.

Brienzi concerns a different rock-slope process and signal band. Radar intervals are an external sensing reference but informed the original labels, so this is an episode-level replication of evaluation effects rather than sensor-blind discovery. The released feature table predates every split available to this audit; view 0 is merely the first released view, and raw-signal feature construction cannot be reconstructed from the release. One-to-one matching prevents a single DAS episode from crediting multiple radar events. Operational claims would additionally require multiple failures and non-failures, independent targets, long background monitoring, and fixed event-level partitions.

8 Conclusions

Sensor-dense records can look large while containing one target trajectory. At Hollin Hill, a section split repeats every evaluated timestamp in fitting, and flexible clock or exact-time probes reconstruct the labels; longer purges and transferred normalization do not create event independence. In the pre-rain-fixed replay, point-anchor rupture-horizon occupancy is zero and is not guaranteed across the anchor interval; late scenarios instead begin within pre-existing burden. At Brienzi, sibling-view effects persist across grouped partitions and episode–radar alignment exceeds a timing-preserving null, but both findings remain conditional on one released feature table and one clustered collapse sequence.

The practical lesson is procedural: count events and unique times before windows; separate detection from horizon occupancy; preserve interval-valued provenance; freeze parameters before replay; report the complete burden surface; group augmented observations; and compare episodes with qualified external references. These safeguards prevent one event from being mistaken for a validation cohort.



Code and data availability. Hollin Hill data are available from Zenodo (<https://doi.org/10.5281/zenodo.8356348>), and Brienzen data from
315 EnviDat (<https://doi.org/10.16904/envodat.541>). Version-controlled analysis source and tracked result tables are publicly available at <https://github.com/mgunjan67/das-himalayan-geohazard-ai>; the evaluated revision is commit 2832a4a. The repository contains the analysis configuration, commands, source, tests, environment locks, and generated result tables. The reproducible tensor cache is represented by metadata and a checksum. Analyses use NumPy, SciPy, scikit-learn, PyTorch, and xarray (Harris et al., 2020; Virtanen et al., 2020; Pedregosa et al., 2011; Paszke et al., 2019; Hoyer and Hamman, 2017).

320 *Author contributions.* GKM conceived and led the study, developed the methodology and software, curated and formally analysed the data, performed validation and visualization, and wrote and finalized the manuscript. BG independently reviewed the methodology, results, and scientific interpretation and contributed to validation and manuscript revision. Both authors approved the submitted manuscript and accept responsibility for their contributions.

Competing interests. The authors declare no conflict of interest.

325 *Acknowledgements.* The authors thank the Hollin Hill and Brienzen investigators for making their data and analysis materials publicly available. No external funding was received. OpenAI Codex assisted with code refactoring, reproducibility checks, manuscript organization, and language editing. Both authors verified all outputs and accept responsibility for the publication.



References

- Ajo-Franklin, J. B., Dou, S., Lindsey, N. J., Monga, I., Tracy, C., Robertson, M., Rodriguez Tribaldos, V., Ulrich, C., Freifeld, B., Daley, T.,
 330 and Li, X.: Distributed Acoustic Sensing Using Dark Fiber for Near-Surface Characterization and Broadband Seismic Event Detection,
 Scientific Reports, 9, 1328, <https://doi.org/10.1038/s41598-018-36675-8>, 2019.
- Aleotti, P.: A warning system for rainfall-induced shallow failures, Engineering Geology, 73, 247–265,
<https://doi.org/10.1016/j.enggeo.2004.01.007>, 2004.
- Baum, R. L. and Godt, J. W.: Early warning of rainfall-induced shallow landslides and debris flows in the USA, Landslides, 7, 259–272,
 335 <https://doi.org/10.1007/s10346-009-0177-0>, 2009.
- Bergmeir, C., Hyndman, R. J., and Koo, B.: A note on the validity of cross-validation for evaluating autoregressive time series prediction,
 Computational Statistics & Data Analysis, 120, 70–83, <https://doi.org/10.1016/j.csda.2017.11.003>, 2018.
- Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 78, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2), 1950.
- 340 Caine, N.: The Rainfall Intensity: Duration Control of Shallow Landslides and Debris Flows, Geografiska Annaler. Series A, Physical
 Geography, 62, 23, <https://doi.org/10.2307/520449>, 1980.
- Calvello, M. and Piciullo, L.: Assessing the performance of regional landslide early warning models: the EDuMaP method, Natural Hazards
 and Earth System Sciences, 16, 103–122, <https://doi.org/10.5194/nhess-16-103-2016>, 2016.
- Carla, T., Intrieri, E., Di Traglia, F., Nolesini, T., Gigli, G., and Casagli, N.: Guidelines on the use of inverse velocity method as a tool for
 345 setting alarm thresholds and forecasting landslides and structure collapses, Landslides, 14, 517–534, <https://doi.org/10.1007/s10346-016-0731-5>, 2016.
- Carla, T., Intrieri, E., Raspini, F., Bardi, F., Farina, P., Ferretti, A., Colombo, D., Novali, F., and Casagli, N.: Perspectives on the prediction
 of catastrophic slope failures from satellite InSAR, Scientific Reports, 9, 14137, <https://doi.org/10.1038/s41598-019-50792-y>, 2019.
- Cerqueira, V., Torgo, L., and Mozetic, I.: Evaluating time series forecasting models: an empirical study on performance estimation methods,
 350 Machine Learning, 109, 1997–2028, <https://doi.org/10.1007/s10994-020-05910-7>, 2020.
- Crosta, G. B. and Frattini, P.: Distributed modelling of shallow landslides triggered by intense rainfall, Natural Hazards and Earth System
 Sciences, 3, 81–93, <https://doi.org/10.5194/nhess-3-81-2003>, 2003.
- Daley, T. M., Freifeld, B. M., Ajo-Franklin, J., Dou, S., Pevzner, R., Shulakova, V., Kashikar, S., Miller, D. E., Goetz, J., Henniges, J.,
 and Lueth, S.: Field testing of fiber-optic distributed acoustic sensing (DAS) for subsurface seismic monitoring, The Leading Edge, 32,
 355 699–706, <https://doi.org/10.1190/tle32060699.1>, 2013.
- Davis, J. and Goadrich, M.: The relationship between Precision-Recall and ROC curves, Proceedings of the 23rd international conference on
 Machine learning - ICML '06, pp. 233–240, <https://doi.org/10.1145/1143844.1143874>, 2006.
- Dean, T., Cuny, T., and Hartog, A. H.: The effect of gauge length on axially incident P-waves measured using fibre optic distributed vibration
 sensing, Geophysical Prospecting, 65, 184–193, <https://doi.org/10.1111/1365-2478.12419>, 2016.
- 360 Dick, G. J., Eberhardt, E., Cabrejo-Lievano, A. G., Stead, D., and Rose, N. D.: Development of an early-warning time-of-failure analysis
 methodology for open-pit mine slopes utilizing ground-based slope stability radar monitoring data, Canadian Geotechnical Journal, 52,
 515–529, <https://doi.org/10.1139/cgj-2014-0028>, 2015.
- Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874, <https://doi.org/10.1016/j.patrec.2005.10.010>, 2006.



- Fichtner, A., Bogris, A., Nikas, T., Bowden, D., Lentas, K., Melis, N. S., Simos, C., Simos, I., and Smolinski, K.: Theory of phase transmission fibre-optic deformation sensing, *Geophysical Journal International*, 231, 1031–1039, <https://doi.org/10.1093/gji/ggac237>, 2022.
- Finkelstein, D. M.: A Proportional Hazards Model for Interval-Censored Failure Time Data, *Biometrics*, 42, 845, <https://doi.org/10.2307/2530698>, 1986.
- Geyer, C. J.: Practical Markov Chain Monte Carlo, *Statistical Science*, 7, 473–483, <https://doi.org/10.1214/ss/1177011137>, 1992.
- Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, 102, 359–378, <https://doi.org/10.1198/016214506000001437>, 2007.
- Guzzetti, F., Peruccacci, S., Rossi, M., and Stark, C. P.: The rainfall intensity-duration control of shallow landslides and debris flows: an update, *Landslides*, 5, 3–17, <https://doi.org/10.1007/s10346-007-0112-1>, 2007.
- Hand, D. J. and Till, R. J.: A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, *Machine Learning*, 45, 171–186, <https://doi.org/10.1023/A:1010920819831>, 2001.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Rio, J. F., Wiebe, M., Peterson, P., Gerard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, *Nature*, 585, 357–362, <https://doi.org/10.1038/s41586-020-2649-2>, 2020.
- Hernan, M. A. and Robins, J. M.: Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1., *American Journal of Epidemiology*, 183, 758–764, <https://doi.org/10.1093/aje/kwv254>, 2016.
- Hoyer, S. and Hamman, J.: xarray: N-D labeled arrays and datasets in Python, *Journal of Open Research Software*, 5, 10, <https://doi.org/10.5334/jors.148>, 2017.
- Hurlbert, S. H.: Pseudoreplication and the Design of Ecological Field Experiments, *Ecological Monographs*, 54, 187–211, <https://doi.org/10.2307/1942661>, 1984.
- Intrieri, E., Gigli, G., Mugnai, F., Fanti, R., and Casagli, N.: Design and implementation of a landslide early warning system, *Engineering Geology*, 147–148, 124–136, <https://doi.org/10.1016/j.enggeo.2012.07.017>, 2012.
- Intrieri, E., Carla, T., and Gigli, G.: Forecasting the time of failure of landslides at slope-scale: A literature review, *Earth-Science Reviews*, 193, 333–349, <https://doi.org/10.1016/j.earscirev.2019.03.019>, 2019.
- Jousset, P., Reinsch, T., Ryberg, T., Blanck, H., Clarke, A., Aghayev, R., Hersir, G. P., Henningses, J., Weber, M., and Krawczyk, C. M.: Dynamic strain determination using fibre-optic cables allows imaging of seismological and structural features, *Nature Communications*, 9, 2509, <https://doi.org/10.1038/s41467-018-04860-y>, 2018.
- Kang, J., Walter, F., Paitz, P., Aichele, J., Edme, P., Meier, L., and Fichtner, A.: Automatic Monitoring of Rock-Slope Failures Using Distributed Acoustic Sensing and Semi-Supervised Learning, *Geophysical Research Letters*, 51, e2024GL110672, <https://doi.org/10.1029/2024GL110672>, 2024a.
- Kang, J., Walter, F., Paitz, P., Aichele, J., Edme, P., Meier, L., and Fichtner, A.: Distributed Acoustic Sensing Brienz, *EnviDat* dataset, <https://doi.org/10.16904/envidat.541>, 2024b.
- Kapoor, S. and Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science, *Patterns*, 4, 100804, <https://doi.org/10.1016/j.patter.2023.100804>, 2023.
- Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O.: Leakage in data mining, *ACM Transactions on Knowledge Discovery from Data*, 6, 1–21, <https://doi.org/10.1145/2382577.2382579>, 2012.



- Kim, S., Choi, K., Choi, H.-S., Lee, B., and Yoon, S.: Towards a Rigorous Evaluation of Time-Series Anomaly Detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 7194–7201, <https://doi.org/10.1609/aaai.v36i7.20680>, 2022.
- Lazic, S. E.: The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?, *BMC Neuroscience*, 11, 5, <https://doi.org/10.1186/1471-2202-11-5>, 2010.
- 405 Lindsey, N. J. and Martin, E. R.: Fiber-Optic Seismology, *Annual Review of Earth and Planetary Sciences*, 49, 309–336, <https://doi.org/10.1146/annurev-earth-072420-065213>, 2021.
- Liu, Q. and Paparrizos, J.: The Elephant in the Room: Towards a Reliable Time-Series Anomaly Detection Benchmark, in: *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- Mateeva, A., Lopez, J., Potters, H., Mestayer, J., Cox, B., Kiyashchenko, D., Wills, P., Grandi, S., Hornman, K., Kuvshinov, B., Berlang, W.,
 410 Yang, Z., and Detomo, R.: Distributed acoustic sensing for reservoir monitoring with vertical seismic profiling, *Geophysical Prospecting*, 62, 679–692, <https://doi.org/10.1111/1365-2478.12116>, 2014.
- Molchan, G.: Strategies in strong earthquake prediction, *Physics of the Earth and Planetary Interiors*, 61, 84–98, [https://doi.org/10.1016/0031-9201\(90\)90097-H](https://doi.org/10.1016/0031-9201(90)90097-H), 1990.
- Ouellet, S. M., Dettmer, J., Lato, M. J., Cole, S., Hutchinson, D. J., Karrenbach, M., Dashwood, B., Chambers, J. E., and Crickmore, R.:
 415 Previously hidden landslide processes revealed using distributed acoustic sensing with nanostrain-rate sensitivity, *Nature Communications*, 15, 6239, <https://doi.org/10.1038/s41467-024-50604-6>, 2024a.
- Ouellet, S. M., Dettmer, J., Lato, M. J., Cole, S., Hutchinson, D. J., Karrenbach, M., Dashwood, B., Chambers, J. E., and Crickmore, R.: Dataset for Previously hidden landslide processes revealed using distributed acoustic sensing with nanostrain-rate sensitivity, *Zenodo dataset*, <https://doi.org/10.5281/zenodo.8356348>, 2024b.
- 420 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: *Advances in Neural Information Processing Systems*, vol. 32, pp. 8024–8035, 2019.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal*
 425 *of Machine Learning Research*, 12, 2825–2830, 2011.
- Piciullo, L., Gariano, S. L., Melillo, M., Brunetti, M. T., Peruccacci, S., Guzzetti, F., and Calvello, M.: Definition and performance of a threshold-based regional early warning model for rainfall-induced landslides, *Landslides*, 14, 995–1008, <https://doi.org/10.1007/s10346-016-0750-2>, 2016.
- Piciullo, L., Calvello, M., and Cepeda, J. M.: Territorial early warning systems for rainfall-induced landslides, *Earth-Science Reviews*, 179,
 430 228–247, <https://doi.org/10.1016/j.earscirev.2018.02.013>, 2018.
- Piciullo, L., Tiranti, D., Pecoraro, G., Cepeda, J. M., and Calvello, M.: Standards for the performance assessment of territorial landslide early warning systems, *Landslides*, 17, 2533–2546, <https://doi.org/10.1007/s10346-020-01486-4>, 2020.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical,
 435 or phylogenetic structure, *Ecography*, 40, 913–929, <https://doi.org/10.1111/ecog.02881>, 2017.
- Rose, N. and Hungr, O.: Forecasting potential rock slope failure in open pit mines using the inverse-velocity method, *International Journal of Rock Mechanics and Mining Sciences*, 44, 308–320, <https://doi.org/10.1016/j.ijrmms.2006.07.014>, 2007.



- Rossi, M., Wisen, R., Vignoli, G., and Coni, M.: Assessment of Distributed Acoustic Sensing (DAS) performance for geotechnical applications, *Engineering Geology*, 306, 106729, <https://doi.org/10.1016/j.enggeo.2022.106729>, 2022.
- 440 Saito, T. and Rehmsmeier, M.: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, *PLOS ONE*, 10, e0118432, <https://doi.org/10.1371/journal.pone.0118432>, 2015.
- Schenato, L.: A Review of Distributed Fibre Optic Sensors for Geo-Hydrological Applications, *Applied Sciences*, 7, 896, <https://doi.org/10.3390/app7090896>, 2017.
- Segoni, S., Piciullo, L., and Gariano, S. L.: A review of the recent literature on rainfall thresholds for landslide occurrence, *Landslides*, 15, 1483–1501, <https://doi.org/10.1007/s10346-018-0966-4>, 2018.
- 445 Stahli, M., Sattelle, M., Huggel, C., McArdell, B. W., Lehmann, P., Van Herwijnen, A., Berne, A., Schleiss, M., Ferrari, A., Kos, A., Or, D., and Springman, S. M.: Monitoring and prediction in early warning systems for rapid mass movements, *Natural Hazards and Earth System Sciences*, 15, 905–917, <https://doi.org/10.5194/nhess-15-905-2015>, 2015.
- Tashman, L. J.: Out-of-sample tests of forecasting accuracy: an analysis and review, *International Journal of Forecasting*, 16, 437–450, [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0), 2000.
- 450 Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., and Gottschlich, J.: Precision and Recall for Time Series, in: *Advances in Neural Information Processing Systems*, vol. 31, pp. 1920–1930, 2018.
- Turnbull, B. W.: The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 38, 290–295, <https://doi.org/10.1111/j.2517-6161.1976.tb01597.x>, 1976.
- 455 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Haggstrom, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavic, J., Nothman, J., Buchner, J., Kulick, J., Schonberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodriguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kummerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., and Vazquez-Baeza, Y.: SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nature Methods*, 17, 261–272, <https://doi.org/10.1038/s41592-019-0686-2>, 2020.
- 465 Wang, X., Shi, B., Wei, G., and Chen, S.: An Energy Demodulation Based Fiber Optic Sensing System for Landslide Early-Warning, *Applied Sciences*, 7, 699, <https://doi.org/10.3390/app7070699>, 2017.
- 470 Wu, R. and Keogh, E. J.: Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, *IEEE Transactions on Knowledge and Data Engineering*, <https://doi.org/10.1109/TKDE.2021.3112126>, 2023.
- Zhu, C., Yang, Y., Yang, K., Zhu, W., and Yang, Q.: Monitoring landslide disturbances using distributed acoustic sensing under extreme weather conditions, *npj Natural Hazards*, 3, 23, <https://doi.org/10.1038/s44304-026-00182-y>, 2026.