https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

SeismicX-Cont: a coverage-aware continuous seismic data product for operational earthquake monitoring

Ziye Yu[1], Xin Liu[2, 3], Yuqi Cai[4], Yanru An[5], Fajun Miao[6], Tengchao Dong[7], and Tingting Li[6]

[1]Institute of Geophysics, China Earthquake Administration, Beijing 100081, China
[2]Laboratory of Seismology and Physics of Earth's Interior, School of Earth and Space Sciences, University of Science and Technology of China, Hefei 230026, China
[3]Institute of Advanced Technology, University of Science and Technology of China, Hefei 230088, China
[4]University of Chinese Academy of Sciences, Beijing 100029, China
[5]China Earthquake Networks Center, Beijing 100081, China
[6]Jiangsu Earthquake Administration, China Earthquake Administration, Nanjing 210014, China
[7]Shandong Earthquake Agency, China Earthquake Administration, Jinan 250014, China

Correspondence: Xin Liu (xinliu_geo@outlook.com)

Abstract. Many seismic datasets are organized around individual earthquakes or phase arrivals and support model training and evaluation on selected waveform windows. Benchmark datasets for testing earthquake monitoring workflows on real continuous records remain limited. Such evaluation must account for background noise, overlapping arrivals, waveform gaps, and variable station availability. SeismicX-Cont provides a fixed continuous-waveform benchmark that links observations with machine-readable station–time coverage, reference annotations, and instrument-response metadata. The release contains 14 d of waveforms from 979 unique network–station identifiers across the Southern California Seismic Network, Berkeley Digital Seismic Network, and USGS Northern California Seismic Network. It covers 1–7 July 2019, including the onset of the Ridgecrest sequence, and 8–14 November 2021, a lower-seismicity period. Fourteen daily Hierarchical Data Format version 5 (HDF5) files preserve native sampling rates and waveform gaps and are linked to annotations for 7519 catalogued events and 290 329 P- and S-wave source readings. Exact coverage checks and reference deduplication yield 182 442 unique manual references with valid waveform samples at the arrival times. These references define the primary recall denominator, separating algorithmic misses from unavailable input; stricter input requirements can be applied using the linked coverage information. The benchmark also supports two complementary proxies for phase-pick precision, defined as positive predictive value: catalogue agreement within reference windows over a declared processing domain, and corroboration by eligible external outputs with the assessed output excluded. Higher eligibility-normalized support is associated with greater agreement with point-covered manual references across both periods and phase classes. These measures support evaluation under incomplete reference catalogues without classifying every unmatched pick as a false positive. Validation scripts, a quick-start notebook, and a two-hour mini dataset facilitate reproducible testing of phase-picking, association, and location workflows under continuous monitoring conditions. The full release is available from Hugging Face at https://doi.org/10.57967/hf/10283 (Yu et al., 2026a).

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

1 Introduction

Automated earthquake monitoring increasingly combines artificial intelligence with established seismic processing methods (Kong et al., 2019; Mousavi and Beroza, 2022, 2023). A typical workflow detects seismic phases and picks arrival times, associates picks across stations, locates earthquakes, and constructs an event catalogue. Representative detection and picking methods include Generalized Phase Detection (Ross et al., 2018), PhaseNet (Zhu and Beroza, 2019), and EQTransformer (Mousavi et al., 2020), while REAL (Zhang et al., 2019), GaMMA (Zhu et al., 2022), and PyOcto (Münchmeyer, 2024) support phase association. Other approaches combine processing stages: ConvNetQuake performs direct earthquake detection and location (Perol et al., 2018), and QuakeFlow integrates multiple monitoring components within a scalable workflow (Zhu et al., 2023). At archive scale, Ni et al. (2025) demonstrate continuous-waveform processing through a global database of automatic phase picks.

These developments have benefited from a growing collection of labelled seismic datasets. STEAD (Mousavi et al., 2019), INSTANCE (Michelini et al., 2021), DiTing (Zhao et al., 2023), LEN-DB (Magrini et al., 2020), and TXED (Chen et al., 2024) provide curated waveform examples for model training and evaluation. PNW-ML (Ni et al., 2023) and the California Earthquake Event Dataset (CEED; Zhu et al., 2025) expand regional data resources, while SeisBench (Woollam et al., 2022) standardizes access to datasets and models. Many widely used training and evaluation datasets are organized around individual earthquakes or phase arrivals, with waveform excerpts selected relative to known events. This organization facilitates supervised learning and controlled testing, but does not preserve the full temporal context of continuous monitoring.

The resulting difference in evaluation conditions affects the interpretation of both phase-level and event-level performance. Consider a 60 s waveform window selected around an earthquake and containing one target P arrival and one target S arrival. During a quiet monitoring day, the same station might record only a few earthquake phases over 86 400 s. Event-centred sampling therefore changes the proportion of earthquake signals to background data and excludes much of the time over which spurious picks can accumulate. Precision measured within selected windows consequently need not represent positive predictive value over the full continuous stream. Recall measured on those windows is also conditional on the selected labelled arrivals and cannot establish recovery of all phases in the original record. Moreover, earthquake detection depends on whether picks can be successfully associated and located across stations, so phase-picking scores alone do not determine event-level recall. Continuous-data experiments have demonstrated corresponding differences between performance on selected windows and operational phase-pick and event outputs (Cai et al., 2025).

To bridge the gap between event-centred evaluation and testing under real monitoring conditions, we introduce SeismicX-Cont, a fixed continuous-waveform benchmark for earthquake-processing workflows. The release links 14 daily waveform files to a coverage index, CEED-derived annotations, station and event geometry, instrument responses, and validation software. Two seven-day case studies—the 2019 Ridgecrest sequence and a lower-seismicity California period in 2021—preserve contrasting monitoring loads, including dense arrivals and extended background intervals. The fixed 14 d scope makes repeated continuous inference and association computationally tractable, providing a common basis for workflow evaluation, stress testing, and reproducibility.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

The core contribution is the versioned alignment of continuous observations, exact station–time coverage, and annotation provenance, with attribution retained to the source data centres and networks. Coverage-qualified canonical manual references define the primary phase-recall denominator, preventing unavailable input from being counted as an algorithmic miss. Two complementary phase-pick precision proxies support assessment under incomplete reference catalogues: one-to-one catalogue agreement within reference windows over a declared processing domain, and corroboration by eligible external outputs after exclusion of the assessed output. The linked annotations and geometry also support evaluation of association and location, with event-level recall calculated against a declared observable reference-event set. Following the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (Wilkinson et al., 2016), documented schemas, validation software, licensing, and reuse interfaces make these evaluations reproducible. Optional multi-output records support interoperability checks and precision-proxy calculation.

2 Source data and selected monitoring periods

2.1 Network sources

The waveforms come from three permanent regional networks: the Southern California Seismic Network (CI; California Institute of Technology and United States Geological Survey Pasadena, 1926), the Berkeley Digital Seismic Network (BK; BDSN, 2014), and the USGS Northern California Seismic Network (NC; USGS Menlo Park, 1966). CI records are archived at the Southern California Earthquake Data Center (SCEDC; Southern California Earthquake Data Center, 2013). BK and NC records are archived at the Northern California Earthquake Data Center (NCEDC; Northern California Earthquake Data Center, 2014). Together, these networks provide dense regional coverage across California and adjacent parts of Nevada using broadband, short-period, and strong-motion instruments. The released Hierarchical Data Format version 5 (HDF5) archive contains 350 CI, 113 BK, and 516 NC network–station keys, each represented by at least one indexed waveform segment.

The data product preserves the original network, station, location, and channel (NSLC) identifiers from the source archives. The HH, BH, EH, and HN channel families provide the principal waveform inputs considered for phase-picking evaluation in this release. Together, they comprise 54 400 indexed segments, corresponding to approximately 52 875 component-days. Additional waveform channel families and low-rate state-of-health channels are retained in the full HDF5 inventory, maintaining traceability to the source MiniSEED files.

2.2 Ridgecrest and lower-seismicity monitoring periods

Candidate periods were drawn from 2010–2023 using the CEED earthquake catalogue (Zhu et al., 2025). They were defined as calendar weeks, from Monday 00:00 to the following Monday 00:00 in Coordinated Universal Time (UTC), and ranked by catalogue event count. From the five weeks with the highest event counts, 1–7 July 2019 was selected as a representative high-seismicity case encompassing the onset of the Ridgecrest earthquake sequence. For the lower-seismicity case, one week was randomly selected from the five weeks with the lowest event counts, yielding 8–14 November 2021 (Fig. 1). The two periods

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



provide contrasting monitoring loads for continuous-stream evaluation. Their exact UTC intervals are [2019-07-01 00:00:00, 2019-07-08 00:00:00) and [2021-11-08 00:00:00, 2021-11-15 00:00:00), with inclusive start times and exclusive end times.

The first period spans the onset of the Ridgecrest earthquake sequence, including the 4 July M_{w} 6.4 foreshock and the 6 July M_{w} 7.1 mainshock (Ross et al., 2019; Yue et al., 2021). It contains 7253 catalogued events and 281 471 P- and S-wave source readings. Daily event counts increase from fewer than 110 on each of the first three days to 2629 on 5 July and 2273 on 6 July. This interval provides a high-load test case with closely spaced and frequently overlapping arrivals.

The second period, 8–14 November 2021, contains 266 catalogued events and 8858 P- and S-wave source readings. Its longer inter-event intervals provide complementary conditions for evaluating continuous pick output and association behaviour. The continuous records preserve local seismicity, background noise, signals outside the reference catalogue, and time-varying station conditions. Differences in network availability and channel composition also contribute to the contrast between the periods, which should therefore be interpreted as operational case studies rather than a controlled experiment isolating seismicity alone.

3 Data-product concept

SeismicX-Cont provides a fixed continuous-waveform benchmark for evaluating earthquake-monitoring workflows, from phase detection and arrival-time picking through phase association, event location, and catalogue construction. The evaluation domain comprises continuous station records across selected monitoring periods, including earthquake signals, inter-event background, overlapping arrivals, and interruptions in waveform availability. This setting extends evaluation beyond event- and phase-centred excerpts to the recovery of earthquakes and the volume of automatic outputs under observed monitoring conditions. The linked data layers support both end-to-end workflow evaluation and separate assessment of individual processing stages.

3.1 Core data objects and relations

Five linked layers underpin the data-product concept: continuous waveforms, exact waveform coverage, reference annotations and geometry, instrument responses, and multi-output phase-picking records with processing metadata. The first four provide the observational and reference basis of the core release. The multi-output evaluation extension supplies automatic phase records and cross-output evidence for precision assessment and downstream workflow reuse.

Daily HDF5 files preserve waveform segments with their native sampling rates, amplitudes, and gaps. A SQLite index maps exact network–station–location–channel (NSLC) intervals to their HDF5 paths, supporting waveform retrieval and input-availability checks. Reference annotations and geometry provide event identifiers, origin times, hypocentres, P- and S-wave source readings, annotation provenance, and station information. Channel-epoch instrument responses are stored in JavaScript Object Notation (JSON) and linked to waveform channels through NSLC identifiers and UTC times.

These relationships connect evaluation at the arrival and event levels. Each reference arrival can be checked against the waveform input required by a declared processing configuration, while its source-event membership links the arrival to a reference earthquake. Source annotations remain traceable when the corresponding waveform input is unavailable. Exact re-

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

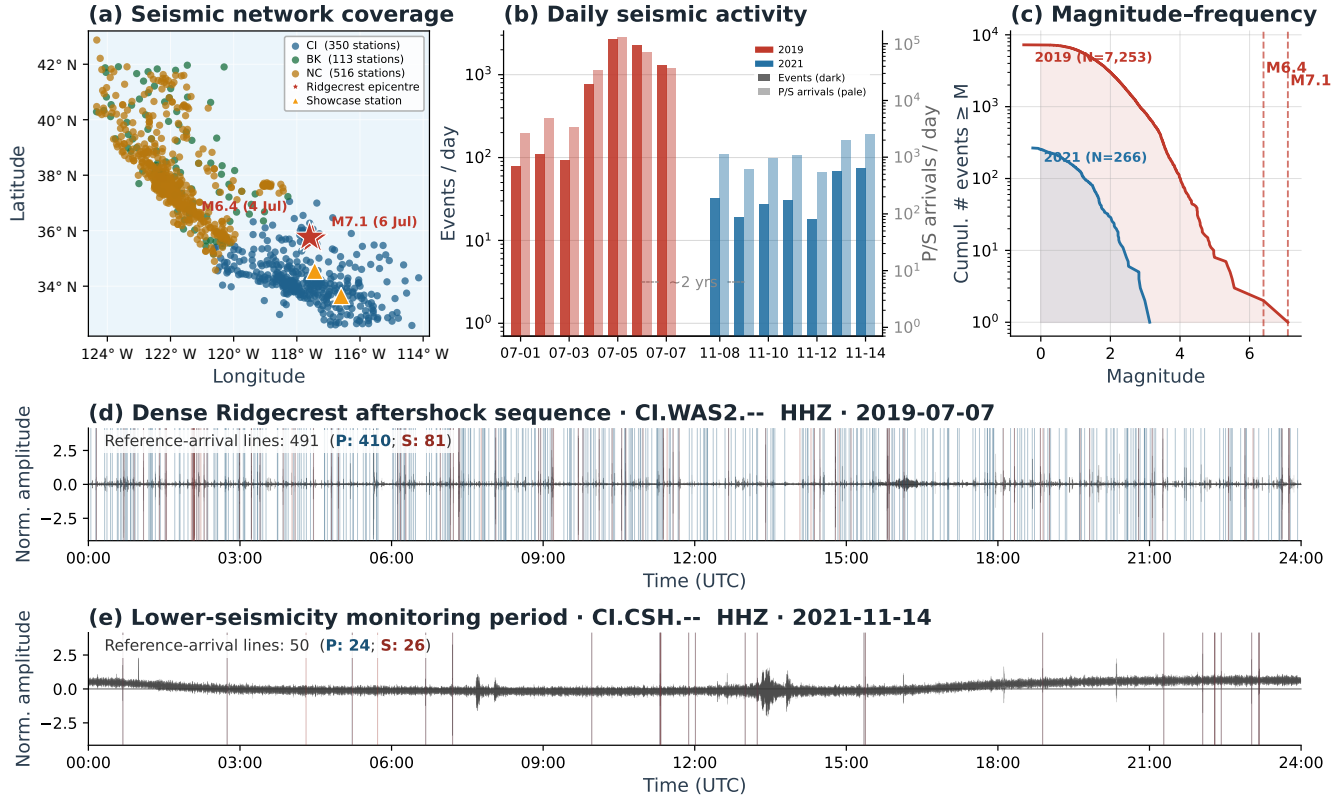

Figure 1. Inventory and monitoring-regime overview of SeismicX-Cont. (a) Regional network–station geometry; (b) daily event counts (dark bars, left axis) and P- and S-wave source-reading counts (pale bars, right axis); (c) cumulative event counts at or above each source-reported magnitude, retaining the original magnitude types; and (d, e) examples of continuous-waveform amplitude ranges with reference P and S arrivals. Panel (d) shows CI.WAS2 on 7 July 2019, selected as the available 2019 station-day with the largest source-reading count within the 50–500-reading display criterion. Panel (e) shows CI.CSH on 14 November 2021. The examples were selected separately to illustrate continuous-stream structure. Red and blue in (b, c) distinguish the two periods. Grayscale envelopes in (d, e) show minimum and maximum amplitudes within each 1 s display interval after subtraction of the trace median. Each panel is scaled by the 99.5th percentile of its per-interval absolute peak amplitudes and clipped at ±4 for display. Full-height blue and dark-red lines mark P and S arrivals, respectively.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



peated station–location–phase–time readings are collapsed into canonical references while retaining their source memberships. Coverage-qualified canonical manual P/S references define the primary phase-recall reference set; operational automatic P references remain available in a provenance-distinct expanded setting. Event identifiers, origin times, and hypocentres provide the reference information needed to evaluate earthquake recovery after association and location.

The multi-output layer contains automatic phase picks from nine example workflows for the selected monitoring periods. A consolidated phase-candidate catalogue groups temporally corresponding picks by station, location, and normalized phase class while preserving each member's output identity, original phase name, arrival time, and reported probability. Single-output candidates retain picks without cross-output agreement, while candidates supported by multiple outputs provide corroborating evidence. An output registry records documented model and checkpoint information, and a processing-eligibility database identifies the station–location–time intervals, phase capabilities, and input conditions associated with each output. These records make it possible to distinguish a lack of corroboration from a lack of eligible external processing.

The layers can be combined according to the processing stage under investigation. End-to-end evaluation starts from continuous waveforms and produces associated and located earthquakes. Modular evaluation can instead supply manual arrivals, individual automatic-pick streams, or multi-output phase candidates directly to an associator and locator. This allows downstream methods to be tested with different phase inputs without rerunning waveform-based phase picking. Waveforms, instrument responses, geometry, and retained catalogue magnitudes additionally support magnitude-estimation experiments. Section 7.2 describes these reuse configurations.

3.2 Coverage-aware evaluation specification

Evaluation is specified at both the phase-arrival and earthquake-event levels. Users declare the continuous evaluation domain, waveform selection, preprocessing, and input requirements before matching workflow outputs to references. Waveform availability and the intervals actually processed by each workflow are documented separately. The coverage hierarchy (C0–C3) distinguishes point coverage from progressively stricter window, component, and processing-readiness requirements (Sect. 6.2). Reference eligibility is determined from the declared input requirements independently of whether an automatic pick or event is reported.

At the earthquake level, end-to-end evaluation requires an observable reference-event population defined independently of the assessed workflow. An event may, for example, qualify through a declared minimum number of stations with suitable waveform coverage, a minimum number of coverage-qualified P and S arrivals, or both. The workflow processes the declared continuous input and produces associated and located events, which are matched one-to-one to the qualifying reference events using declared origin-time and epicentral-distance tolerances. Earthquake recall is the fraction of qualifying reference events recovered by the complete workflow. This measure captures the combined outcome of phase detection, association, and location within the specified observation domain.

At the phase level, recall measures the recovery of coverage-qualified canonical manual arrivals using deterministic one-to-one matching by station, location, normalized phase, and arrival time. Phase recall and earthquake recall describe different aspects of workflow performance: recovery of individual arrivals and recovery of their parent events. Reporting both connects

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.


component-level behaviour to the resulting earthquake catalogue. Both measures refer to their declared reference populations, whose completeness is limited by the available annotations.

Two complementary phase-pick precision proxies use the reference and multi-output layers. Precision proxy 1 measures catalogue agreement within merged manual-reference windows intersected with the assessed workflow's processing domain. Precision proxy 2 measures the fraction of eligible assessed picks that meet a declared level of corroboration from eligible external outputs. External eligibility is established from processing records, rather than from the presence of reported picks. When the assessed output contributed to the candidate catalogue, its records are removed and affected candidates are reconstructed before corroboration is calculated.

These proxies provide complementary evidence about phase-pick precision, with precision denoting positive predictive value. Catalogue-unmatched picks may include unlabelled arrivals, while cross-output agreement may reflect dependencies among models or workflows. Similarly, catalogue-unmatched workflow events remain identifiable outputs without being automatically assigned false-event status. Continuous-output reporting therefore accompanies the reference-based metrics, including total pick and event counts and, preferably, picks per processed station-hour. Results for the two monitoring periods are reported separately to retain their contrasting seismicity and monitoring conditions.

Table 1 summarizes the specification. Section 7.1 gives the phase-level calculations and reporting rules, and Sect. 7.2 describes end-to-end and modular earthquake-workflow evaluation.

3.3 Relationship to existing seismic data resources

SeismicX-Cont complements resources that support different stages of earthquake-data analysis. Event- and phase-window datasets provide curated examples for model development and component-level evaluation. CEED supplies event-based waveform and annotation resources, archive-scale automatic-pick products provide extensive phase-output records, and SeisBench standardizes access to datasets and models. Table E1 summarizes these complementary reuse roles.

SeismicX-Cont connects these roles through fixed continuous observations linked to waveform coverage, reference arrivals and events, instrument metadata, and multi-output phase records. This alignment supports a progression from phase-level testing to evaluation of earthquake recovery by complete monitoring workflows. It also allows association and location methods to be evaluated separately with documented phase inputs. The contribution is the reproducible organization of these connected evaluation domains, while the underlying observations and source annotations remain attributable to their originating networks, data centres, and data products.

4 Data-product generation

Generation links continuous waveforms, coverage, source annotations, instrument responses, and automatic phase outputs for phase-level and end-to-end earthquake evaluation.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 1. Evaluation specification supported by SeismicX-Cont for continuous earthquake-monitoring workflows and their individual processing stages.

Evaluation element	SeismicX-Cont specification
Evaluation target	Fixed daily continuous waveforms support evaluation from phase detection and picking through association, event location, and catalogue construction, retaining native gaps, sampling rates, channel families, and source provenance.
Input eligibility	Waveform input is checked against declared channel, window, component, sampling-rate, response, and data-quality requirements. Reference eligibility is established independently of reported picks or events.
Processing domain	The declared evaluation domain and the station–location–time intervals actually processed by each workflow are documented, together with phase capability and valid-input requirements.
Reference information	Canonical manual P/S references retain source memberships and define the primary phase-reference set after coverage qualification. Source event identifiers, origin times, and hypocentres support event matching. Operational automatic P references form a provenance-distinct expanded set.
Phase recall	Reported picks are matched one-to-one to eligible references by station, location, normalized phase, and time. The reference denominator, processing domain, matching tolerance, and recovered-arrival count are reported.
Earthquake recall	An observable reference-event denominator is defined using declared station-coverage and/or reference-arrival requirements. Associated and located workflow events are matched one-to-one by origin time and epicentral distance. The qualifying-event count, recovered-event count, and resulting recall are reported.
Multi-output evidence	Automatic phase records retain their output identities and member-level information within a consolidated candidate catalogue. The output registry and processing-eligibility records identify the external outputs eligible to corroborate each assessed pick.
Precision proxy 1	Catalogue-conditioned reference-window match fraction within the intersection of merged manual-reference windows and the workflow's processing domain. Reports include window size, matching tolerance, effective processed-window duration, and matched and catalogue-unmatched pick counts.
Precision proxy 2	Fraction of assessed picks meeting a declared eligibility-normalized external-support threshold, among picks with sufficient eligible external outputs. The assessed output is excluded and affected candidates are reconstructed before matching. Eligibility criteria, thresholds, matching tolerance, and denominator counts are reported.
Modular evaluation	Manual arrivals, individual automatic-pick streams, or multi-output phase candidates can serve as association and location inputs. Reports identify the input source, phase-selection rules, association settings, velocity model, location method, and matching criteria.
Continuous-output reporting	Total pick and event counts, matched and catalogue-unmatched event counts, and unrecovered qualifying reference events are reported over the declared domain. Picks per processed station-hour are recommended. Results for the two periods are reported separately.
Leakage and calibration	Training-data exposure, checkpoint selection, threshold tuning, and calibration information are reported where documented by the source artifacts; otherwise, the corresponding fields are explicitly marked as not documented.

4.1 Waveform assembly

MiniSEED records from SCEDC and NCEDC were converted to daily HDF5 files using ObsPy (Beyreuther et al., 2010) and HDF5 (The HDF Group, 2000–2026). Each trace segment is stored as a one-dimensional dataset preserving native sam-

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



ples, NSLC identifiers, timing, sampling rate, sample count, and data type. No response removal, filtering, normalization, gap interpolation, or resampling is applied. Gaps and inter-event records remain explicit.

Station groups retain available coordinates, elevation, and validity periods; annotations retain event and source–receiver geometry. Responses are stored separately in JSON and matched by NSLC and epoch. At read time, the PyTorch loader can remove the native response through ObsPy or simulate a target response. Appendix A documents the versioned schema; optional additions preserve existing fields, while breaking changes require a new schema version.

Coordinates retain source-reported geographic latitude and longitude in decimal degrees, with station elevations in metres. Datum identifiers are not uniformly retained, and no datum transformation is applied. SCEDC documents event depths relative to the WGS84 reference ellipsoid, whereas NCEDC specifies the geoid for the NCSS catalogue (Southern California Earthquake Data Center, 2015; Northern California Earthquake Data Center, 2015). CEED depth values and signs, including negative depths, are preserved. Event depths remain in kilometres, with preferred-origin `depth_m` obtained by multiplication by 1000. Response-JSON `depth_m` instead denotes sensor depth relative to the local ground surface. Time strings are in UTC; numerical Unix epoch fields express seconds since 1 January 1970 at 00:00:00 UTC.

4.2 Annotation assembly

CEED event and phase records (Zhu et al., 2025), integrating NCEDC and SCEDC catalogues, were converted to JSON and aligned with the independently assembled waveforms. Events retain origin time, hypocentre, magnitude and type, source, and station metadata; arrivals retain phase, time, label status, distance, azimuth, and auxiliary metadata. Annotation keys use network–station–location (NSL), formatted as `network.station.location`; waveform queries use complete NSLC identifiers.

Events are selected by preferred origin time within the half-open intervals in Sect. 2.2. All associated P/S arrivals are retained without further time filtering; every arrival falls within its corresponding interval in this snapshot. The converted snapshot is version-locked as `data/label/annotations_for_continuous_hdf5.json`, with its SHA-256 digest in `manifest_sha256.txt`. Event identifiers remain unique, and source records are grouped by NSL without additional arrival deduplication or conflict reconciliation.

The `reference_arrivals` table preserves all 290 329 source readings. A separate `evaluation_references` layer collapses exact repeated station–location–normalized-phase–time readings into 274 696 canonical references, including repeats shared by multiple events. The `evaluation_reference_members` table retains all source arrival and event memberships. References with at least one manual member enter the manual evaluation set; all others retain operational-automatic provenance. Canonicalization preserves source events and their links to coverage-qualified arrivals.

The source inventory contains 160 905 P and 129 424 S readings, comprising 239 423 manual and 50 906 operational automatic readings. Among canonical references, 223 820 have manual members and 50 891 have automatic members; 15 have both and are classified as manual. All operational automatic labels are P arrivals. Event magnitudes range from -0.49 to 7.1, with a median of 1.79, and depths range from -1.52 to 38.3 km.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Coverage-qualified canonical manual references define primary phase recall; operational automatic P references support a separately reported expanded setting. Event memberships support observable-event selection for earthquake recall. Source arrivals remain available for association and location tests when waveforms are missing. Exploratory consensus candidates remain separate from the reference sets.

4.3 Multi-output phase-candidate layer

The multi-output extension consolidates 124 403 181 picks from nine non-standardized workflows into 49 689 373 station–location–phase candidate groups. P/Pg/Pn and S/Sg/Sn map to broad P-like and S-like classes; members retain their original phase, output identity, relative time, and probability. Single-output groups retain uncorroborated picks, while multi-output groups provide evidence for precision proxy 2 (Table 5).

Construction is deterministic. Records are partitioned by exact station–location and normalized phase, then sorted by arrival time and output identifier. The earliest unused record anchors a non-overlapping group containing subsequent records strictly less than 1.5 s after the anchor. The next unused record starts another group. Representative times average the per-output mean arrival times, giving each output equal weight. Duplicate picks remain individual members, but each output contributes only one support vote. Probabilities retain their documented quantization and are not averaged across outputs. Channel family is not a grouping key.

Support counts distinct output streams; unique-family support is recorded separately. The `data/label/output_eligibility.sqlite` database contains 169 573 processed station–location intervals with timing, P/S capability, channel family, components, source file, and derivation method. Eligibility follows processed waveform samples independently of reported picks. The `data/label/output_registry.tsv` registry records family/checkpoint labels, interval and station counts, union station-hours, and pick counts. Unavailable training-domain, threshold, and evaluation-overlap metadata are marked *not documented*.

Member records and processed intervals support exclusion of the assessed output and reconstruction of candidates. These records also support corroboration normalized by eligibility. Individual streams and multi-output candidates also provide association and location inputs. Each output is mapped to a serialized artifact under `pickers/`, documented in `pickers/README.md`. The PNSN implementation and models are independently archived (Cai and Yu, 2025); its workflow is described by Cai et al. (2025). Appendix B summarizes the registry.

4.4 Coverage index and validation products

Every HDF5 segment is indexed in SQLite by file, internal path, NSLC, timing, sampling rate, sample count, data type, and station coordinates. The indexed `start_epoch` and `end_epoch` give the first and final sample times, denoted t_0 and t_1. The half-open duration interval is $[t_0, t_1 + \Delta t)$, where Δt is the sample interval. Overlapping or sample-adjacent segments are merged only within the same NSLC. Default queries use HH, BH, EH, and HN and exclude state-of-health channels. Coverage is calculated independently of picker outputs.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



An explicitly specified annotation location must match exactly. Otherwise, candidate locations and families are ranked by observed component count at the arrival time, then by family priority HH, BH, EH, HN, and finally by location code. The selected stream remains fixed for C1–C3, without substitution if it fails a higher coverage level; the resulting denominators describe this fixed-priority selection. Components are never combined across locations or families. Channel suffixes E/1, N/2, and Z/3 map to H1, H2, and Z, respectively; non-standard horizontal orientations are not rotated.

Point availability uses the nearest sample, assigning half-sample ties to the later sample and requiring floating-point values to be finite. Response epochs are left-closed and right-open; response-qualified configurations require exactly one matching epoch per selected component at the arrival time. Validation scripts export point, window, component, and processing-readiness fields for phase-recall and precision-proxy calculations.

Joining coverage fields to event memberships allows users to count distinct qualifying stations and canonical P/S references and construct observable-event subsets under declared requirements. Phase and event exchange formats connect picking, association, and location. The event-comparison utility reports matched, catalogue-unmatched, and unrecovered events using one-to-one origin-time and epicentral-distance matching, with depth differences reported separately. Applied to an observability-qualified reference set, the recovery fraction provides earthquake recall. Sections 7.1 and 7.2 give the evaluation and reporting rules.

5 Data-product contents

5.1 Waveform inventory

The waveform inventory comprises 14 daily HDF5 files named `continuous_waveform_usa_YYYYMMDD.h5`, totalling approximately 387 GiB compressed and 159 916 indexed segments. Across all retained channels, these records represent 158 090 component-days. Component-days are calculated by summing indexed segment durations and dividing by 86 400 s; reported totals are rounded. The HH, BH, EH, and HN families contain 54 400 segments, 52 875 component-days, and 21 034 station-day-family entries.

HN contributes 24 481 segments from 768 network–station keys, HH 12 703 from 372, BH 11 718 from 346, and EH 5498 from 424. Station counts overlap across families. Arrays retain `int32` or `float32` values and native sampling rates, most commonly 40, 100, and 200 Hz. Figure 2 summarizes daily availability.

5.2 Annotation inventory

The annotations contain 7519 events, 172 550 event–station associations, and 180 344 source-lineage records representing 290 329 P/S source readings. Labels without waveform coverage remain available. Source readings, canonical evaluation references, and exploratory consensus candidates are distinguished, with manual and operational automatic provenance retained (Sect. 4.2). Event memberships link arrival-level coverage to reference-event selection for earthquake-recall evaluation.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

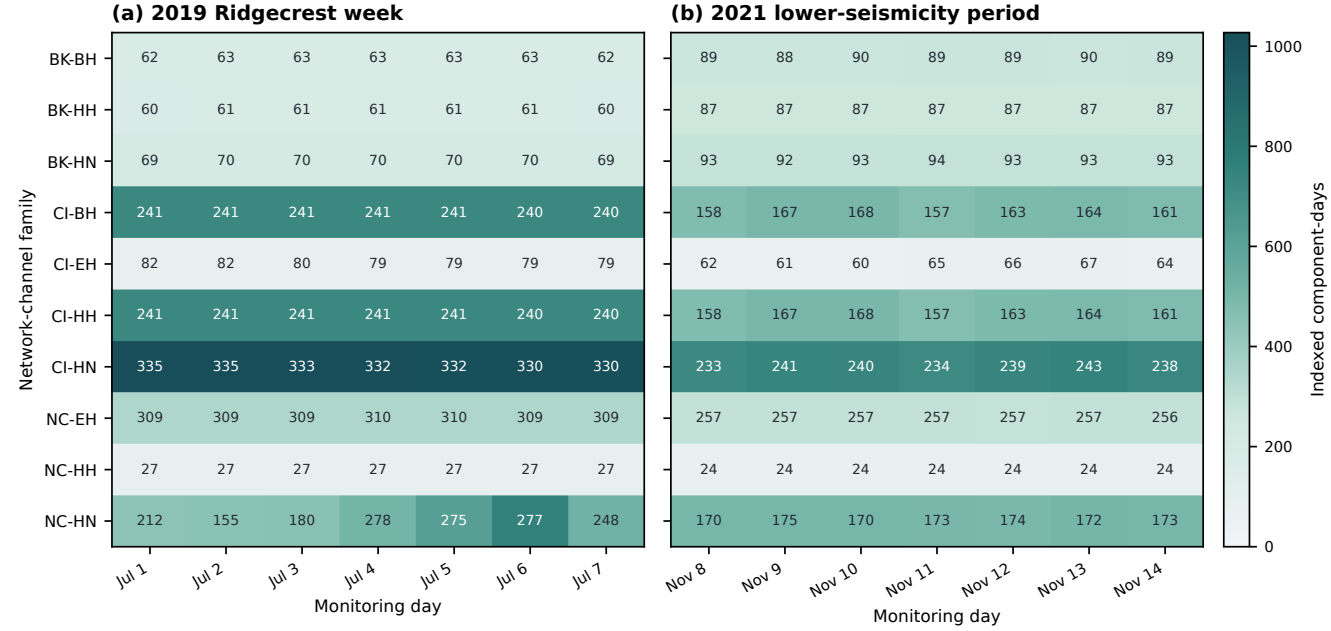

Figure 2. Daily waveform inventory for the HH, BH, EH, and HN channel families during the two monitoring periods, derived from the released SQLite index. Cell colour represents indexed component-days; numbers indicate unique network–station keys. Rows show network–channel-family combinations observed on at least one of the 14 d; blank cells indicate no indexed data. Workflow eligibility additionally depends on exact NSLC intervals, finite-sample availability, and declared input requirements (Sect. 6.2).

5.3 Monitoring-regime characteristics

Figure 3 summarizes source-reading geometry, magnitude–distance sampling, reference density, event timing, and temporal crowding. Arrival statistics use C0-qualified source readings from HH/BH/EH/HN. Operational automatic readings account for 39 943 of 129 288 covered P readings in 2019 (30.9 %) and 2707 of 3835 in 2021 (70.6 %); all S readings are manual. Period and phase comparisons therefore include differences in label provenance. Inter-event statistics use positive intervals between consecutive catalogue origin times.

The point-covered inventory contains 235 481 readings in 2019 (129 288 P; 106 193 S) and 5081 in 2021 (3835 P; 1246 S). Median P/S distances are 43.6/44.1 km in 2019 and 40.6/30.5 km in 2021; the corresponding 90th percentiles are 102.4/99.8 and 106.9/71.4 km. For source magnitudes below 2, 90th-percentile distances are 88.6 km in 2019 and 76.4 km in 2021; for magnitudes of at least 3, they are 135.3 and 127.9 km. Source magnitude types are retained.

Across all catalogued events, the covered-reading count has median, 90th-percentile, and maximum values of 26, 61, and 240 in 2019, compared with 15, 38.5, and 84 in 2021. Median inter-event intervals are 28.2 and 1206.6 s, respectively, with

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 2. Principal released assets and their roles in SeismicX-Cont. Licensing and repository identifiers are given in the Code and data availability section.

Asset group	Status	Principal content and function
Waveform data	Core	Fourteen daily `data/hdf5/continuous_waveform_usa_*.h5` files provide continuous-waveform input for phase picking and end-to-end earthquake-detection workflows (Yu et al., 2026a).
Coverage and reference tables	Core	`data/index/waveform_index.sqlite` indexes exact NSLC intervals. `data/label/reference_arrivals.sqlite` stores source readings, canonical references, event memberships, and point-availability and workflow-configured coverage fields.
Annotations and instrument metadata	Core	`data/label/annotations_for_continuous_hdf5.json` and `data/label/stations.csv` retain CEED-derived events, P/S readings, provenance, geometry, and station metadata. `data/response/instrument_responses.json` provides channel-epoch responses for NSLC-and-time matching.
Validation and access software	Core	`manifest_sha256.txt`, validation summaries, conversion and access scripts, utilities, and `notebooks/seismicx_cont_quickstart.ipynb` support integrity checks, indexed access, phase and event comparison, and reproduction of reported summaries.
Multi-output products	Optional; full release only	The phase-candidate catalogue, output-eligibility database, output registry, phase-matching records, and consensus candidates support precision-proxy calculations and association/location tests. Paths and sizes are documented in the adjacent inventory. These products remain separate from the reference annotations.
Two-hour mini dataset	Optional lightweight release	Two one-hour waveform subsets with matching annotations, index, responses, scripts, and one example automatic-pick stream support tutorials and software checks. Precomputed multi-output and consensus-candidate products are not included (Yu et al., 2026b).

76.1 and 6.0 % of intervals shorter than 60 s. The 2019 10th-percentile interval is 7.1 s. These distributions distinguish the dense association load in 2019 from the longer inter-event monitoring intervals in 2021.

Within each station–location and phase class, 24.8, 34.0, and 56.4 % of covered 2019 source readings have another reading within ± 10, ± 15, and ± 30 s, respectively; the corresponding 2021 fractions are 12.7, 13.2, and 13.2 %. Together with event spacing, these source-level statistics characterize the contrasting monitoring loads available for evaluating phase detection, association, and earthquake recovery.

5.4 File organization

Table 2 summarizes the core assets, optional full-release products, and mini dataset. The core checksum manifest and adjacent multi-output inventory define the versioned release boundary, documenting file identities and associated metadata. Unlisted files are outside the release. Schemas, interfaces, and script inventories are documented in the appendices and repository.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

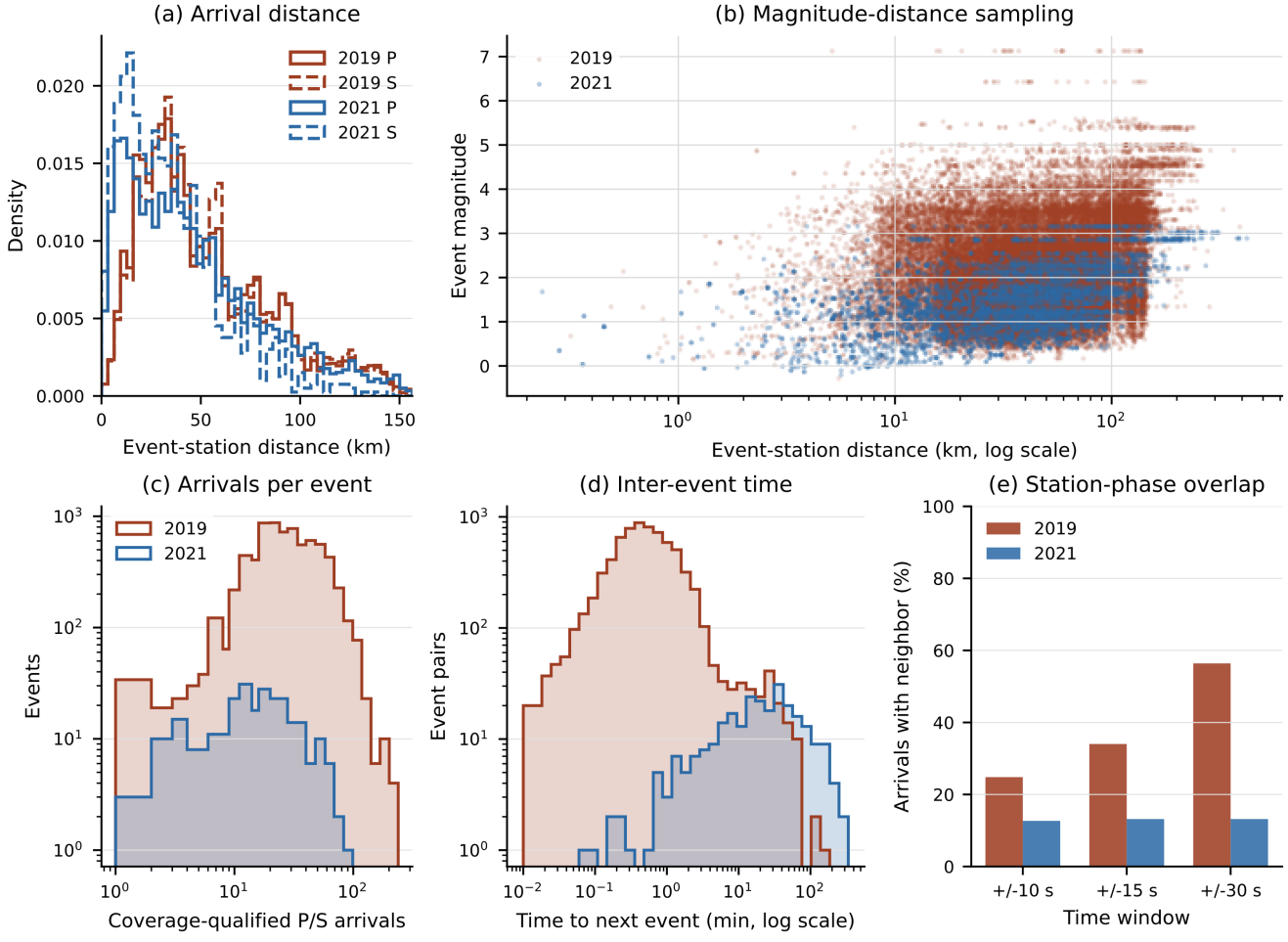

Figure 3. Characteristics of the two monitoring regimes. Panels (a–c, e) use C0-qualified source readings, including manual and operational automatic P readings; panel (d) uses catalogue origin times. (a) Event–station distance distributions for P-like and S-like readings, displayed to the 99.2nd percentile of the pooled covered-distance distribution. (b) Source magnitude versus distance, with a fixed-seed random sample of 60 000 readings from 2019 to limit raster density. (c) Covered P/S readings per catalogued event; zero-count events are omitted only from the logarithmic display. (d) Positive intervals between consecutive catalogue origin times. (e) Fraction of readings with another same-NSL, same-phase-class reading within each symmetric time window. Source magnitude types and label provenance are retained, so the panels describe the released cases without controlling for these differences.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

6 Data quality, uncertainty, and validation

6.1 Inventory validation

The inventory audit regenerates reported counts from the annotation JSON and SQLite index (Table 4). Exact NSLC intervals support gap and overlap checks without assuming uninterrupted station-day coverage. Links to sampled HDF5 datasets are checked separately in Sect. 6.4.

6.2 Waveform coverage validation

Coverage queries combine exact NSLC intervals with valid HDF5 samples under the selection rules in Sect. 4.4. The annotation location – is unspecified and resolves to at most one released location; explicit locations require exact matches. Default queries use HH/BH/EH/HN and exclude state-of-health channels.

Coverage follows four levels. C0 requires a valid nearest sample at the arrival time in the selected location–family stream. C1–C3 use the declared half-open window $[t-b, t+a)$. Gap fractions are calculated separately for each observed component; where multiple channels represent one component, the smallest-gap channel is selected, with channel code breaking ties. C1 requires at least one component to meet the allowed gap fraction; `window_gap_fraction` stores the component minimum. C2 requires all declared components—Z, H1, and H2 in the example—to meet that threshold; `required_component_gap_fraction` stores their maximum. Replicated vertical traces do not constitute three observed components. C3 additionally applies the minimum sampling rate, any declared quality criteria, and, when requested, exactly one matching NSLC response epoch per selected component at the arrival time. C0 depends on the selected inventory; C1–C3 are regenerated for each input configuration.

The released example uses a ± 30 s window, zero allowed gap fraction, observed Z/H1/H2 components, a minimum rate of 40 Hz, and one matching response epoch per component. Of 290 329 source readings, 240 562 satisfy C0 and 49 767 lack point coverage. C1 retains 240 449 readings, and C2/C3 retain 238 594: 113 fail window completeness and a further 1855 fail component requirements. All C2 components meet the rate and response requirements, so C3 removes no further readings. Table 3 gives the breakdown.

For a declared evaluation domain and coverage configuration, phase recall is

$$\mathrm{Recall}_{\mathrm{cov}} = \frac{\mathrm{TP}_{\mathrm{cov}}}{\mathrm{TP}_{\mathrm{cov}} + \mathrm{FN}_{\mathrm{cov}}}, \tag{1}$$

where $\mathrm{TP}_{\mathrm{cov}}$ and $\mathrm{FN}_{\mathrm{cov}}$ count matched and unmatched qualifying canonical references under deterministic one-to-one matching within a declared tolerance.

The C0 source inventory comprises 133 123 P and 107 439 S readings. Collapsing 15 470 repeated C0 manual readings yields 182 442 canonical manual references: 81 751 P and 98 599 S in 2019, and 987 P and 1105 S in 2021. These form the recommended primary C0 reference set. The provenance-mixed expanded set contains 225 068 unique C0 references and is

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Table 3. Source-reading coverage by period, phase, and provenance. Counts retain repeated source readings. C0 requires an exact HH/BH/EH/HN NSLC interval and a valid sample at the arrival time; C1–C3 use the example configuration described in the text.

Period	Phase and provenance	Total	C0	C0 (%)	C1	C2	C3
2019 Ridgecrest	Manual P	108 023	89 345	82.7	89 313	89 249	89 249
2019 Ridgecrest	Operational automatic P	46 201	39 943	86.5	39 915	38 879	38 879
2019 Ridgecrest	Manual S	127 247	106 193	83.5	106 140	106 064	106 064
2021 lower-seismicity	Manual P	1976	1128	57.1	1128	1068	1068
2021 lower-seismicity	Operational automatic P	4705	2707	57.5	2707	2150	2150
2021 lower-seismicity	Manual S	2177	1246	57.2	1246	1184	1184
Full release	All P/S source readings	290 329	240 562	82.9	240 449	238 594	238 594

All operational automatic labels are P arrivals. These source-reading counts are distinct from canonical phase-pick denominators, which count each exact station–location–phase–time reference once while retaining source memberships.

reported separately. Under the stricter example configuration, canonical manual counts are 182 357 at C1 and 182 105 at C2/C3. Any further restriction by workflow processing domain must also be reported.

For earthquake recall, coverage fields are linked to source event memberships to construct an observable-event denominator under declared station and phase requirements (Sect. 7.2).

6.3 Annotation and response validation

Annotation checks cover identifiers, nested counts, phase names, catalogue sources, and provenance; daily counts agree with Fig. 1. All 7519 events contain the required origin, hypocentral, magnitude, type, and source fields, although two magnitude-type values retain blank upstream codes. Coordinates and elevation are present for all 172 550 event–station associations, and phase, time, status, station identity, distance, and azimuth for all 290 329 readings. Uniform arrival-time, hypocentral, depth, and magnitude uncertainties are unavailable and are not inferred.

The response JSON contains 17 979 epochs covering all 14 723 target channels. Each required component for the 238 594 C2-qualified source readings matches exactly one epoch under the example configuration. Other configurations require renewed NSLC-and-time matching, with zero or multiple matches recorded explicitly.

6.4 Waveform and index integrity

All 159 916 index rows pass required-field, sampling-rate, end-time, duplicate-key, and exact-NSLC overlap checks. The index contains no exact-NSLC overlaps and retains 4429 gaps within the selected periods.

The HDF5 audit uses a deterministic sample stratified by period, network, and HH/BH/EH/HN family. Metadata agree with the index for all 200 sampled datasets. Among 1 611 000 values from 600 sampled 30 s windows, 0.82 % are non-finite, 0.65 % are zero, and three windows are constant; no sampled integers equal their data-type extrema. These sample diagnostics do not establish archive-wide waveform-quality rates. Table 4 distinguishes full-index checks from sampled waveform checks.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 4. Summary of validation results and their scope.

Validation target	File-derived result
Waveform and index inventory	14 daily HDF5 files, 159 916 segments, and 979 network–station keys. HH/BH/EH/HN contribute 54 400 segments and 52 875 component-days.
Annotation consistency	7519 events, 172 550 event–station associations, and 290 329 P/S source readings, comprising 239 423 manual and 50 906 operational automatic readings.
Canonical evaluation references	The 290 329 source readings map to 274 696 unique station–location–phase–time references. The primary C0 manual set contains 182 442 references, each linked to its source arrivals and events.
Metadata completeness	Required fields are present for all events, event–station associations, and source readings. Two magnitude-type values retain blank upstream codes; uniform source uncertainties are unavailable.
C0 point coverage	240 562 source readings have qualifying NSLC coverage and a valid arrival-time sample. Of these, 238 745 have at least three observed components, 15 have two, and 1802 have one.
Configured C1–C3 example	240 449 C1, 238 594 C2, and 238 594 C3 source readings; successive reductions after C0 are 113, 1855, and 0.
Waveform and index integrity	All 159 916 index rows pass required-field, sampling, end-time, duplicate-key, and overlap checks. The selected periods contain 4429 exact-NSLC gaps and no exact-NSLC overlaps.
Stratified waveform sample	Metadata agree for 200 sampled HDF5 datasets. Among 1 611 000 values from 600 sampled 30 s windows, 0.82 % are non-finite, 0.65 % are zero, and three windows are constant. Results describe the sample only.
Instrument responses	17 979 epochs cover all 14 723 target channels. Each required component for the 238 594 C2-qualified source readings has exactly one matching epoch under the example configuration.
Software interoperability	The mini dataset exercises HDF5 access, SQLite queries, JSON/JSONL exchange, and single-output smoke tests. Full-release multi-output products additionally exercise deterministic one-to-one phase matching.

6.5 Phase-pick precision proxies and multi-output interoperability validation

Validation examines catalogue agreement within processed reference windows and corroboration by eligible external outputs after leave-one-output reconstruction (Sect. 7.1). Precision denotes positive predictive value. Candidate-level analyses establish the observed support–agreement relationship; workflow-level analyses demonstrate the two proxies within their declared domains.

Phase-matching records under `data/validation/multimodel_phase_matching/` are generated from newline-delimited JSON (JSONL) outputs and independently calculated C0 flags. Deterministic one-to-one matching uses station–location and normalized phase, with P/S results and reference-centred recall reported separately from pick-centred catalogue agreement. For precision proxy 1, matching and pick counting are restricted to merged manual-reference windows intersected with the workflow processing domain. Candidate JSONL and SQLite products retain member times, original phases, outputs, probabilities, and construction provenance; utilities also accept compatible user-generated outputs.

Table 5 independently matches each support stratum to canonical C0 manual references within 0.5 s. Candidate match fractions increase from support $= 1$ to ≥ 2 and ≥ 3 in every period–phase group. Reference recovery at support ≥ 2 is 82.7–

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 5. Agreement between multi-output candidates and canonical C0 manual references. Each support stratum is matched independently by exact station–location and normalized phase within 0.5 s; the ≥ 2 and ≥ 3 strata are nested. Candidate match uses candidates as its denominator; reference recovery uses all canonical C0 manual references in the corresponding period–phase group.

Period	Phase	Support	Candidates	Candidate match n (%)	Reference recovery (%)	Median $\lvert\Delta t\rvert$ (s)
2019	P	$= 1$	13 040 519	3294 (0.0253)	4.0	0.0687
2019	P	≥ 2	3 587 802	69 615 (1.940)	85.2	0.0377
2019	P	≥ 3	1 781 088	67 244 (3.775)	82.3	0.0373
2019	S	$= 1$	7 817 926	5200 (0.0665)	5.3	0.0791
2019	S	≥ 2	5 014 983	81 572 (1.627)	82.7	0.0839
2019	S	≥ 3	2 816 294	77 625 (2.756)	78.7	0.0839
2021	P	$= 1$	9 039 351	3 (< 0.001)	0.3	0.0587
2021	P	≥ 2	2 108 343	959 (0.0455)	97.2	0.0362
2021	P	≥ 3	916 423	958 (0.1045)	97.1	0.0362
2021	S	$= 1$	5 740 890	5 (< 0.001)	0.5	0.2997
2021	S	≥ 2	3 339 557	1046 (0.0313)	94.7	0.0813
2021	S	≥ 3	1 716 295	1044 (0.0608)	94.5	0.0813

Median $\lvert\Delta t\rvert$ uses matched candidates only. Two single-output P candidates, 15.2 and 15.245 s outside the selected-period boundaries, are retained for provenance but excluded from this analysis.

85.2 % in 2019 and 94.7–97.2 % in 2021; at ≥ 3, it is 78.7–82.3 % and 94.5–97.1 %. These all-output candidate comparisons provide empirical support for corroboration; workflow-level assessment additionally excludes the assessed output.

The relationship persists after eligibility normalization (Table B2). Candidates with at least six eligible outputs are grouped by $q = k/m$, where k counts supporting outputs and m counts eligible outputs at the candidate time. All in-period candidates have $7 \leq m \leq 9$; the two out-of-period candidates have $m = 0$. At $q \geq 0.5$, reference recovery is 74.5–79.3 % in 2019 and 94.3–96.8 % in 2021, compared with 0.3–5.3 % at $q < 0.25$.

Workflow-level leave-one-output validation covers all nine streams with $q_0 = 0.5$, $m_0 = 6$, and $\delta = 0.5$ s (Appendix Table B3). The assessed output is excluded before candidate reconstruction, and eligibility is evaluated at each assessed-pick time. For `seismicxm.01`, precision proxy 2 ranges from 8.26 to 19.36 % across period–phase groups, exceeding mean null fractions of 0.18–0.38 % from twenty station-day circular time-shift replicates (Appendix Table B3).

For the same output, precision proxy 1 uses $\epsilon = 0.5$ s and window half-widths $T = 1$, 5, 10, and 30 s (Appendix Table B4). Values decrease from 37.1–40.9 % at $T = 1$ s to 27.3–38.8 % at $T = 30$ s as the windows include more continuous-output context.

The nine runs have non-standardized preprocessing, thresholds, training domains, channel selections, inference settings, checkpoints, and operating points, with available metadata recorded in the output registry. Their results demonstrate catalogue agreement and eligible cross-output reproducibility within specified domains. Catalogue incompleteness and dependence among outputs limit interpretation as full-stream positive predictive value; catalogue-unmatched picks remain unclassified as physical false positives.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



7 Recommended use

SeismicX-Cont supports end-to-end earthquake detection and modular evaluation of phase picking, association, location, and magnitude estimation. Evaluations use a declared continuous station–time domain and report the two monitoring periods separately.

7.1 Coverage-aware phase-picking evaluation

Phase-picking evaluation combines coverage-aware recall against canonical manual references with two pick-centred precision proxies. The declared input configuration determines reference eligibility throughout the evaluation. Expanded reference settings may include operational automatic arrivals, with their provenance reported explicitly.

7.1.1 Precision proxy 1: catalogue-conditioned agreement

Let $\mathcal{R}_{s,\phi}^{(L)}$ contain canonical manual reference times satisfying coverage configuration L, where s is an exact station–location key and ϕ a normalized phase class. Let $\mathcal{E}_{s,\phi}^{(w)}$ denote intervals in which workflow w processed valid input and could report that phase. References within this domain define the window union

$$\mathcal{W}_{s,\phi}^{(w,L)}(T) = \bigcup_{\substack{r\in\mathcal{R}_{s,\phi}^{(L)} \\ t_r\in\mathcal{E}_{s,\phi}^{(w)}}} [t_r - T, t_r + T]. \tag{2}$$

Overlapping windows are merged within each station–location and phase stream to avoid double-counting durations or picks. Intersecting the union with the processing domain gives

$$\Omega_{\phi}^{(w,L)}(T) = \bigcup_{s} \left[\{s\} \times \left(\mathcal{W}_{s,\phi}^{(w,L)}(T) \cap \mathcal{E}_{s,\phi}^{(w)}\right)\right]. \tag{3}$$

Picks in this domain are matched deterministically one-to-one to eligible manual references by station–location and phase within $0 < \epsilon \leq T$. Precision proxy 1 is

$$\mathrm{CatMatch}_{\mathrm{win},\phi}^{(w,L)}(T,\epsilon) = \frac{N_{\mathrm{matched},\phi}\left(\Omega_{\phi}^{(w,L)}(T),\epsilon\right)}{N_{\mathrm{reported},\phi}\left(\Omega_{\phi}^{(w,L)}(T)\right)}. \tag{4}$$

The numerator counts matched picks; the denominator includes all reported picks in the effective domain. Unmatched picks retain catalogue-unmatched status, and an empty denominator gives an undefined proxy. Increasing T includes more continuous context and potentially unlabelled signals or unrelated triggers. Reports therefore specify L, T, ϵ, initial and workflow-eligible reference counts, raw and effective window durations, and matched and unmatched pick counts.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

7.1.2 Precision proxy 2: eligible cross-output corroboration

For workflow w, let $\mathcal{P}^{(w)}$ contain its picks in the declared domain and $\mathcal{A}_p^{(-w)}$ the external outputs eligible at assessed pick p. Eligibility requires valid processed input at the same station–location and time, compatible phase capability, and the documented component requirements. The eligible-output count is

$$m_p = \left|\mathcal{A}_p^{(-w)}\right|. \tag{5}$$

Eligibility is queried at the assessed-pick time independently of reported support. Outputs without qualifying processing coverage are excluded.

If the assessed output contributed to the candidate catalogue, remove its members and reconstruct affected groups and representative times using the original clustering rule. Subtracting one support vote alone is insufficient because removal can split groups or shift their times. No exclusion is needed when the assessed output did not contribute.

Assessed picks and external candidates are matched deterministically one-to-one by station–location and phase within δ. Let $\mathcal{S}_p^{(-w)}(\delta) \subseteq \mathcal{A}_p^{(-w)}$ contain eligible outputs contributing to the matched candidate. Then

$$k_p(\delta) = \left|\mathcal{S}_p^{(-w)}(\delta)\right|. \tag{6}$$

Unmatched picks, including duplicates left unmatched by one-to-one assignment, have $k_p = 0$. For $m_p > 0$, the eligibility-normalized support score is

$$q_p(\delta) = \frac{k_p(\delta)}{m_p}. \tag{7}$$

For a support threshold $0 < q_0 \leq 1$ and minimum eligible-output count $m_0 \geq 1$, precision proxy 2 is

$$\mathrm{Corrob}_{q_0,m_0}^{(w)}(\delta) = \frac{\left|\left\{p \in \mathcal{P}^{(w)} : m_p \geq m_0,\ q_p(\delta) \geq q_0\right\}\right|}{\left|\left\{p \in \mathcal{P}^{(w)} : m_p \geq m_0\right\}\right|}. \tag{8}$$

The denominator counts picks with sufficient eligible external outputs; the numerator counts those meeting the support threshold. An empty denominator gives an undefined proxy. Thus, q_p describes individual-pick support, while the aggregate fraction describes workflow-level corroboration.

Candidate span is strictly less than 1.5 s. Matching tolerance δ is separate and should not exceed 1.5 s without justification. The stored field `n`, or $\mathrm{Support}_{\mathrm{obs}}$, counts contributing outputs; calculating m_p additionally requires exact station–location and time joins to `data/label/output_eligibility.sqlite`, with phase, component, and valid-input requirements applied. The relationship between corroboration and full-stream positive predictive value depends on shared errors among outputs.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

7.1.3 Minimum reporting requirements

Report the release revision, waveform days, networks and NSLC selection, coverage configuration, reference provenance, processing domain, preprocessing, model/checkpoint, operating point, matching tolerance, recall, and total pick volume. Picks per processed station-hour support comparison across processing domains. Training-data exposure and tuning information should be reported where documented, with missing information identified explicitly.

For proxy 1, report the reference-window quantities specified above. For proxy 2, include the candidate version, assessed-output identifier, exclusion and reconstruction status, δ, eligibility rule, q_0, m_0, eligible and corroborated counts, and resulting fraction. Report both periods separately.

7.2 End-to-end and modular earthquake-workflow reuse

End-to-end evaluation applies the complete workflow to the declared continuous waveforms. JSONL phase picks can be passed to REAL, GaMMA, PyOcto, or another associator and a user-selected locator. Observable reference events are defined independently of the evaluated workflow using minimum requirements for stations with qualifying waveforms, coverage-qualified P/S references, or both. Associated and located events are matched one-to-one to these references using declared origin-time and epicentral-distance tolerances. Earthquake recall is the fraction of observable reference events recovered. Reports include association parameters, velocity model, locator, minimum phase requirements, and matching criteria.

The comparison utility reports `catalogue_matched_events`, `catalogue_unmatched_events`, and `unrecovered_catalogue_events`. Its `catalogue_recovery_fraction` uses reference events as the denominator and gives earthquake recall when the supplied reference set satisfies the declared observability criteria. The separate `catalogue_match_fraction` uses reported workflow events as the denominator. Catalogue-unmatched events retain unresolved physical status, so earthquake recall remains the primary event-level measure.

Modular tests can supply manual P/S annotations directly to association and location, including readings without released waveforms. Individual automatic-pick streams or multi-output candidates provide alternative inputs without rerunning picking. Report the input source, provenance, phase selection, and probability or support thresholds. Mini-dataset examples exercise these exchange interfaces.

For magnitude studies, retrieve waveform intervals and match channel-epoch responses for removal at read time. Report waveform selection, units, response processing, components, distance range, and retained or reconciled catalogue magnitude types.

7.3 FAIR implementation

Versioned full- and mini-release DOIs and machine-readable citations support findability. Independently downloadable layers and the full-release ModelScope mirror support accessibility without author approval or personalized credentials. Documented HDF5, SQLite, comma-separated values (CSV), JSON, and JSONL schemas support interoperability; source lineage, validity masks, response epochs, validation software, and layer-specific licenses support reuse. Indexed and modular access allows

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.


annotation-only and partial-data studies without downloading the complete waveform archive. Repository identifiers and terms are consolidated in the Code and data availability statement.

8 Limitations of the current release

The release covers two 7 d intervals in California and the Nevada border region. The Ridgecrest and lower-seismicity cases differ in season, network and channel coverage, noise, and annotation composition. They support repeated operational evaluation but do not constitute a balanced training sample or support multiyear seismicity statistics. Transfer to other tectonic and monitoring settings requires separate assessment.

Reference annotations inherit source-catalogue incompleteness and uncertainty. Manual and operational automatic readings retain distinct provenance, with automatic labels restricted to P. Earthquake recall is conditional on the declared observable reference-event set. Catalogue-unmatched picks may represent real signals, duplicates, or noise, while unmatched events lack verified physical status. Precision proxy 1 depends on catalogue completeness, window half-width T, and processing domain; proxy 2 depends on output diversity and eligibility. Its candidate span of < 1.5 s is a grouping rule, separate from matching tolerance δ. The station-day time-shift null tests temporal coincidence for one example output but does not resolve dependencies arising from shared training, architectures, preprocessing, or responses to the same signals.

Gaps, partial component coverage, native rates, sensor types, and non-finite samples remain explicit. C0 establishes point availability; C1–C3 apply workflow-specific window, component, and processing requirements. Readings without waveforms remain available for annotation-based tests. Users should document filling, component replication, resampling, rotation, and filtering. Index checks cover all rows, whereas waveform-value checks use a stratified sample.

Amplitude-sensitive reuse requires exact NSLC-and-epoch response matching, documented units, and explicit handling of missing or ambiguous responses. Native amplitudes and source magnitude types are retained. Coordinate-datum metadata are not uniform, and source depth-reference conventions require attention in location and depth comparisons. Indexed, streamed, or staged access supports use of the 387 GiB archive; the mini dataset provides a lightweight entry point for tutorials and smoke tests.

9 Conclusions

SeismicX-Cont provides a versioned benchmark for evaluating earthquake-detection workflows on continuous observations, extending tests beyond event-centred excerpts. It links 14 d of waveforms from 979 network–station keys and 159 916 indexed segments with CEED-derived annotations for 7519 events and 290 329 P/S source readings, channel-epoch responses, and multi-output phase records. Exact coverage checks identify 240 562 point-covered source readings. The primary phase-recall reference set contains 182 442 canonical C0-qualified manual arrivals, with stricter denominators available for declared input requirements.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



The linked observations, coverage, and provenance support coverage-aware phase recall and two complementary precision proxies: catalogue-conditioned agreement and eligible cross-output corroboration. Validation spans four reference-window sizes, nine leave-one-output assessments, and a 20-replicate station-day time-shift null. Manual-reference agreement increases with candidate support, and corroboration for the null-tested output exceeds the time-shift baseline. Member records, the output registry, and processing-eligibility intervals support reproduction of these calculations.

The Ridgecrest and lower-seismicity weeks provide contrasting monitoring loads for repeated end-to-end and modular evaluation. Shared data and exchange interfaces support phase picking, association, location, earthquake recall over declared observable-event populations, and magnitude studies. SeismicX-Cont connects these tasks through a common continuous-observation framework while complementing the SCEDC, NCEDC, CEED, and contributing-network products and preserving their source attribution.

10 Code and data availability

Full release and optional products. The full SeismicX-Cont release is available at https://huggingface.co/datasets/cangyeone/SeismicX-Cont. Its DOI is https://doi.org/10.57967/hf/10283 (Revision a9d5f3d; Yu et al., 2026a). A mirror is available at https://www.modelscope.cn/datasets/cangyeone/SeismicX-Cont. The release contains 14 daily HDF5 waveform files (387 GiB compressed), the SQLite coverage index, annotation and response metadata, station and reference-arrival tables, validation scripts, and the quick-start notebook. The reference-arrival database preserves all source readings and event memberships and includes the canonical evaluation-reference layer used for recall. Optional full-release products include the multi-output phase-candidate catalogue, consensus candidates, output-eligibility and registry files, manual-reference agreement summaries, the four-window precision-proxy sensitivity results, the nine leave-one-output results, the station-day time-shift null, and their regeneration utilities. The adjacent metadata files and `manifest_sha256.txt` define the versioned file boundary and support complete or partial download verification. The repositories are accessible without author approval or personalized credentials.

Mini dataset. A two-hour mini dataset is available from Hugging Face, DOI https://doi.org/10.57967/hf/9664 (Revision 883e092; Yu et al., 2026b). It contains one hour from each monitoring period, together with matching annotations, coverage and response metadata, rebuild scripts, and an example stream of automatic picks. It supports inspection, tutorials, and software tests but does not include the full-release multi-output, consensus, eligibility, or agreement products.

Code and documentation. SeismicX-Cont access and validation code is versioned with the full release under `scripts/`, `utils/`, and `essd_scripts/`, with the quick-start notebook under `notebooks/`. It is released under CC BY 4.0 and cited through the full-release DOI; no separate code DOI is assigned. Serialized picker artifacts are distributed under `pickers/`; the PNSN implementation and related phase-detection models are additionally archived as software at https://doi.org/10.5281/zenodo.17510174 (Cai and Yu, 2025). The documentation describes the schemas, access procedures, preprocessing, and transformations. Event and phase metadata derive from CEED (Zhu et al., 2025). Waveform processing used ObsPy (Beyreuther et al., 2010), HDF5 (The HDF Group, 2000–2026), and NumPy (Harris et al., 2020). The code regen-

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



erates inventory, coverage and quality audits, core figures, coverage-aware phase recall, and both precision proxies, including candidate-level agreement and workflow-level leave-one-output validation.

Licensing and attribution. In both releases the waveform sample arrays follow the applicable source-provider terms for non-commercial scientific research. For arrays derived from NCEDC records, the release applies NCEDC-compatible CC BY-NC 4.0 conditions. Arrays derived from SCEDC records remain subject to SCEDC data-use and citation requirements. Except for those waveform arrays, all assets in both releases are licensed under CC BY 4.0, including metadata, schemas, coverage indexes, event and phase annotations, response metadata, manifests, validation summaries, automatic-pick and consensus-candidate products, figures, documentation, code, and notebooks. A repository-level license label does not override the applicable source-provider terms for waveform arrays. Source lineage and citation requirements remain applicable to source-attributed fields and derived records. Users should cite the SeismicX-Cont release used, the networks and archives listed in Sect. 2.1, and CEED when using its event or phase annotations. The full-release source metadata and mini-release source metadata give the network and archive citations and the reuse terms for each layer. Each record identifies the data revision to which it applies. SeismicX-Cont is a processed continuous-monitoring representation rather than an official catalogue, an authoritative waveform archive, or a replacement for the source products.

Appendix A: HDF5, response, and index reference

Waveforms are stored using the following HDF5 path pattern:

```
/<year_id>/<day_id>/stations/<net.sta.loc>/
    waveform/<channel>/<segment_N>
```

Each `segment_N` is a one-dimensional native-sample array with lossless gzip level 4 compression and the shuffle filter. Missing segments represent gaps; floating-point non-finite samples require validity checks. The SQLite index maps each segment's path, NSLC, timing, rate, count, and data type. ISO-8601 year/day identifiers also permit direct browsing. Response JSON records channel epochs, provenance, sensor and logger descriptions, sensitivity, stages, and identifiers for matching at read time. Table A1 lists the required stable fields; minor schema releases may add optional attributes.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table A1. Primary HDF5 attributes, response-JSON fields, and SQLite index fields used for audit, waveform retrieval, and response-aware preprocessing.

Object	Field or attribute	Description
Station group	`network`	Network code
Station group	`schema_version`	Product schema version where present
Station group	`station`	Station code
Station group	`location`	Location code
Station group	`latitude`	Geographic latitude (decimal degrees)
Station group	`longitude`	Geographic longitude (decimal degrees)
Station group	`elevation`	Station elevation (m); source vertical reference retained
Segment dataset	`start_epoch`	Segment start time (Unix epoch seconds)
Segment dataset	`end_epoch`	Final sample time (Unix epoch seconds)
Segment dataset	`sampling_rate`	Native sampling rate
Segment dataset	`npts`	Number of samples
Segment dataset	`dtype`	Stored waveform data type
Segment dataset	`source_file`	Source MiniSEED path
Response JSON	`response_id`	Stable response epoch identifier
Response JSON	`epoch_start`, `epoch_end`	Response validity interval
Response JSON	`instrument_sensitivity`	Overall sensitivity and units
Response JSON	`stages`	Serialized response stages for ObsPy use
SQLite index	`dataset_path`	Internal HDF5 path
SQLite index	`h5_file`	Daily HDF5 file path
SQLite index	`network`, `station`, `location`, `channel`	Network–station–location–channel identity used for waveform queries
SQLite index	`start_epoch`, `end_epoch`	First and final sample times (Unix epoch seconds) used for station-time coverage checks
SQLite index	`sampling_rate`, `npts`, `dtype`	Native sampling and storage information used for audit and dataloader setup
SQLite index	`latitude`, `longitude`	Geographic coordinates (decimal degrees) for subsetting and audit

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Appendix B: Multi-output registry and eligibility

Table B1 lists the nine output streams. The tab-separated values (TSV) registry adds available training-domain, overlap, and threshold metadata; missing fields are marked *not documented*. Each stream maps to a full-release picker artifact, with PNSN and related models independently archived (Cai and Yu, 2025). Run-specific reporting follows Sect. 7.1.

Table B1. Registry of the nine outputs contributing to the multi-output phase-candidate layer.

ID	Output	Family	Checkpoint	Intervals	Locations	Station-hours	Picks
0	lppnm	LPPNM	not documented	21 200	994	245 641.1	4 628 109
1	phasenet	PhaseNet	not documented	20 884	993	238 245.2	41 503 939
2	pnsn_v1	PNSN	v1	21 200	994	243 937.4	6 783 670
3	pnsn_v3_5120	PNSN	v3_5120	21 200	994	242 308.6	17 933 533
4	pnsn.v3.diff	PNSN	v3.diff	21 200	994	242 794.0	17 805 916
5	seismicxm.01	SeismicXM	01	21 200	994	241 677.0	17 309 291
6	seist	SeisT	not documented	21 200	994	241 588.8	13 027 713
7	skynet	SkyNet	not documented	21 200	994	248 646.1	4 302 639
8	ustc_sc	USTC-SC	not documented	289	111	2664.0	1 108 371

Family and checkpoint labels are limited to information supported by the output identifiers. Intervals and locations are processed-sample intervals and unique exact station–location keys recorded in `output_eligibility.sqlite`; station-hours are unions of overlapping channel-family intervals. Exact station–location counts can exceed the 979 network–station keys because one station may contain multiple location codes. All outputs contain P and S records and use subsets of the BH, EH, HH, and HN channel families.

Table B2. Eligibility-normalized multi-output agreement with canonical C0-qualified manual references.

Period	Phase	Support score $q = k/m$	Candidates	Candidate match n (%)	Reference recovery (%)
2019	P	$q < 0.25$	13 154 155	3294 (0.0250)	4.0
2019	P	$0.25 \le q < 0.5$	2 407 524	4853 (0.2016)	5.9
2019	P	$q \ge 0.5$	1 066 642	64 825 (6.077)	79.3
2019	S	$q < 0.25$	7 883 046	5200 (0.0660)	5.3
2019	S	$0.25 \le q < 0.5$	3 277 115	8310 (0.2536)	8.4
2019	S	$q \ge 0.5$	1 672 748	73 415 (4.389)	74.5
2021	P	$q < 0.25$	9 039 351	3 (< 0.001)	0.3
2021	P	$0.25 \le q < 0.5$	1 655 323	5 (< 0.001)	0.5
2021	P	$q \ge 0.5$	453 020	955 (0.2108)	96.8
2021	S	$q < 0.25$	5 740 890	5 (< 0.001)	0.5
2021	S	$0.25 \le q < 0.5$	2 429 737	9 (< 0.001)	0.8
2021	S	$q \ge 0.5$	909 820	1042 (0.1145)	94.3

Results are regenerated from the candidate JSONL, output-eligibility database, and canonical evaluation-reference layer and are restricted to candidates with at least six eligible outputs ($m \ge 6$). Here, k is the number of distinct supporting outputs, m is the number of eligible outputs, and $q = k/m$ is the candidate-level support score. Each stratum is matched independently to canonical C0-qualified manual references within 0.5 s. Candidate match uses the candidates in each row as its denominator; reference recovery uses all canonical C0-qualified manual references in the corresponding period–phase group.

Higher manual-reference agreement at $q \ge 0.5$ across both periods and phase classes supports the candidate-level basis of precision proxy 2. At workflow level, m_p is queried at the assessed-pick time and reconstructed candidates are assigned one-to-one within δ, ensuring that each candidate corroborates at most one assessed pick.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table B3. Workflow-level leave-one-output precision proxy 2 for the nine released output streams. Output rows report corroborated/eligible picks and the resulting percentage for $q_0 = 0.5$, $m_0 = 6$, $\delta = 0.5$ s, and a strict external-candidate span below 1.5 s. The indented row reports the station-day time-shift null for `seismicxm.01` as the mean and 5th–95th percentile range across 20 deterministic replicates.

Output	2019 P	2019 S	2021 P	2021 S
lppnm	511 780/1 999 472 (25.60 %)	511 552/1 372 044 (37.28 %)	144 811/830 784 (17.43 %)	162 741/425 809 (38.22 %)
phasenet	447 305/21 907 706 (2.04 %)	709 458/5 285 505 (13.42 %)	103 240/12 257 931 (0.84 %)	253 123/2 052 793 (12.33 %)
pnsn_v1	615 214/2 309 194 (26.64 %)	814 285/2 392 757 (34.03 %)	184 775/973 492 (18.98 %)	343 611/1 108 224 (31.01 %)
pnsn_v3_5120	664 838/4 835 339 (13.75 %)	969 148/6 139 590 (15.79 %)	214 473/2 961 345 (7.24 %)	438 666/3 997 252 (10.97 %)
pnsn.v3.diff	638 934/4 348 147 (14.69 %)	921 692/6 417 019 (14.36 %)	206 682/2 675 876 (7.72 %)	429 436/4 364 874 (9.84 %)
seismicxm.01	592 736/3 061 896 (19.36 %)	927 538/7 771 634 (11.93 %)	172 970/1 434 090 (12.06 %)	416 384/5 041 671 (8.26 %)
time-shift null	0.378 % [0.374, 0.382] %	0.320 % [0.316, 0.324] %	0.181 % [0.177, 0.189] %	0.202 % [0.198, 0.204] %
seist	523 430/3 800 775 (13.77 %)	777 906/4 318 114 (18.01 %)	160 334/2 169 179 (7.39 %)	358 494/2 739 640 (13.09 %)
skynet	235 344/1 206 027 (19.51 %)	320 831/1 686 721 (19.02 %)	38 949/511 520 (7.61 %)	105 147/898 371 (11.70 %)
ustc_sc	3968/441 341 (0.90 %)	10 304/667 030 (1.54 %)	–	–

The assessed output is removed before external candidates are reconstructed. Eligibility is queried at each assessed-pick time, and assessed picks and reconstructed candidates are assigned one-to-one. Each null replicate preserves the assessed output's station-day pick count and circularly shifts its times by a deterministic offset between 60 and 600 s before repeating the eligibility and one-to-one corroboration calculation. Differences among output rows reflect non-standardized example outputs and are not interpreted as a workflow ranking. Complete integer results are distributed in `data/validation/leave_one_output_precision_proxy.tsv`.

Table B4. Reference-window sensitivity of precision proxy 1 for the `seismicxm.01` example output. Entries are catalogue-conditioned proxy values in percent for $\epsilon = 0.5$ s.

Period	Phase	$T = 1$	$T = 5$	$T = 10$	$T = 30$
2019	P	37.3	35.3	34.5	28.6
2019	S	37.1	34.9	33.9	27.3
2021	P	40.9	40.0	39.8	38.8
2021	S	39.0	37.9	37.5	35.8

The canonical C0 manual counts are 81 751, 98 599, 987, and 1105 for 2019 P, 2019 S, 2021 P, and 2021 S, respectively; the assessed workflow is eligible for 81 749, 98 597, 987, and 1103 of these references. Complete matched/reported counts, effective window durations, and window counts are distributed in `data/validation/reference_window_precision_proxy_sensitivity.tsv`.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Appendix C: Reproducibility software

The software covers conversion, indexed access, responses, C0–C3 and canonical-reference generation, quality audits, figure regeneration, and both precision proxies. Reproduction and interactive entry points are:

```
essd_scripts/reproduce_manuscript_outputs.sh
notebooks/seismicx_cont_quickstart.ipynb
```

The repository README supplies the script inventory and commands. Core objects remain readable through standard HDF5, SQLite, and JSON libraries.

Appendix D: Access and exchange interfaces

Access examples cover station–time queries, masked loading, and response-aware processing. Pick JSONL records contain station, phase, UTC time, probability, and waveform source. Utilities map P/Pg/Pn and S/Sg/Sn to broad classes and implement one-to-one matching, coverage-aware recall, and the two precision proxies. Repository documentation specifies fields, commands, coverage/eligibility rules, and required T, ϵ, q_0, and m_0 declarations.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Appendix E: Complementary seismic data resources

Table E1. Primary released objects, temporal organization, availability relations, and complementary roles of selected seismic data resources. The descriptions refer to defining features of the released objects; individual resources may provide additional metadata or interfaces.

Resource	Primary released object	Temporal organization	Waveform-availability relation	Complementary role
STEAD, INSTANCE, DiTing, LEN-DB, and TXED (Mousavi et al., 2019; Michelini et al., 2021; Zhao et al., 2023; Magrini et al., 2020; Chen et al., 2024)	Curated event, phase, and noise excerpts	Event- or phase-centred windows	Input availability is defined within each extracted window rather than over a fixed continuous station-time domain	Model training and component-level evaluation
PNW-ML (Ni et al., 2023)	Curated event and exotic-event waveform records spanning 21 years	Event-centred records spanning 21 years	Waveform availability follows the released regional record structure	Regional model development and long-duration event-data access
Global automatic-pick product (Ni et al., 2025)	Automatic P/S picks derived from continuous archives	Archive-scale continuous processing	Processing coverage and pick provenance are represented; waveforms are not the primary released object	Archive-scale pick access and seismicity analysis
CEED (Zhu et al., 2025)	California event catalogue and event-based waveform products	Event-centred	Waveform records and annotations are organized around catalogued events	Event-based reuse and source lineage for SeismicX-Cont annotations
SeisBench (Woollam et al., 2022)	Dataset and model access framework	Dataset-dependent	Availability and metadata relations are inherited from each integrated dataset	Standardized data access, model access, and component-level evaluation
SeismicX-Cont	Daily HDF5 waveforms linked to an index, annotations, and responses	Two fixed seven-day continuous periods with native gaps retained	Exact NSLC segment intervals are linked to reference arrivals and channel-epoch responses	Coverage-aware continuous-stream evaluation under contrasting monitoring loads

Author contributions. Z.Y. conceived the study, designed the dataset structure, implemented the waveform indexing, ESSD validation and reporting scripts, reporting utilities, and picker runner, and wrote the manuscript. Y.C. performed waveform retrieval and HDF5 conversion. X.L. contributed to annotation conversion and quality control. Y.A. contributed to data validation and quality control. F.M. contributed to

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

network coverage analysis and data validation. T.D. contributed to HDF5 format design and the machine-learning data loader. T.L. contributed to reference catalogue preparation and cross-validation. All authors reviewed and approved the final manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the Northern California Earthquake Data Center (NCEDC, Northern California Earthquake Data Center 2014), the Southern California Earthquake Data Center (SCEDC, Southern California Earthquake Data Center 2013), and the operators of the CI, BK, and NC networks for providing and maintaining the source waveform archives used to construct SeismicX-Cont. We also thank NCEDC and SCEDC staff for clarifying the data-use terms applicable to the processed and structured derivative release. During manuscript preparation, the authors used OpenAI Codex for assistive English-language editing, consistency checks, and figure-format refinement. The tool was not used to generate scientific data, results, or interpretations. All analyses, citations, figures, and final manuscript text were verified and approved by the authors.

Financial support. This research has been supported by the Deep Earth Probe and Mineral Resources Exploration–National Science and Technology Major Project (grant no. 2024ZD1000300).

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



References

BDSN: Berkeley Digital Seismic Network, UC Berkeley Seismological Laboratory [data set], https://doi.org/10.7932/BDSN, 2014.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J.: ObsPy: A Python Toolbox for Seismology, Seismological Research Letters, 81, 530–533, https://doi.org/10.1785/gssrl.81.3.530, 2010.

Cai, Y. and Yu, Z.: cangyeone/pnsn: PnSn models and other phase detection models, Zenodo [software], https://doi.org/10.5281/zenodo.17510174, 2025.

Cai, Y., Yu, Z., Liu, X., An, Y., Zhang, Y., Li, L., Zhang, Y., and Ma, X.: A Deep Learning Framework for Pg/Sg/Pn/Sn Phase Picking and Its Nationwide Implementation in Chinese Mainland, Journal of Geophysical Research: Machine Learning and Computation, 2, e2025JH000 944, https://doi.org/10.1029/2025JH000944, 2025.

California Institute of Technology and United States Geological Survey Pasadena: Southern California Seismic Network, International Federation of Digital Seismograph Networks [seismic network], https://doi.org/10.7914/SN/CI, 1926.

Chen, Y., Savvaidis, A., Saad, O. M., Dino Huang, G.-C., Siervo, D., O'Sullivan, V., McCabe, C., Uku, B., Fleck, P., Burke, G., Alvarez, N. L., Domino, J., and Grigoratos, I.: TXED: The Texas Earthquake Dataset for AI, Seismological Research Letters, 95, 2013–2022, https://doi.org/10.1785/0220230327, 2024.

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 Río, J. F., Wiebe, M., Peterson, P., Gérard-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.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P.: Machine Learning in Seismology: Turning Data into Insights, Seismological Research Letters, 90, 3–14, https://doi.org/10.1785/0220180259, 2019.

Magrini, F., Jozinović, D., Cammarano, F., Michelini, A., and Boschi, L.: Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale, Artificial Intelligence in Geosciences, 1, 1–10, https://doi.org/10.1016/j.aiig.2020.04.001, 2020.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V.: INSTANCE – The Italian seismic dataset for machine learning, Earth System Science Data, 13, 5509–5544, https://doi.org/10.5194/essd-13-5509-2021, 2021.

Mousavi, S. M. and Beroza, G. C.: Deep-learning seismology, Science, 377, eabm4470, https://doi.org/10.1126/science.abm4470, 2022.

Mousavi, S. M. and Beroza, G. C.: Machine Learning in Earthquake Seismology, Annual Review of Earth and Planetary Sciences, 51, 105–129, https://doi.org/10.1146/annurev-earth-071822-100323, 2023.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C.: STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, IEEE Access, 7, 179 464–179 476, https://doi.org/10.1109/ACCESS.2019.2947848, 2019.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C.: Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking, Nature Communications, 11, 3952, https://doi.org/10.1038/s41467-020-17591-w, 2020.

Münchmeyer, J.: PyOcto: A high-throughput seismic phase associator, Seismica, 3, https://doi.org/10.26443/seismica.v3i1.1130, 2024.

Ni, Y., Hutko, A., Skene, F., Denolle, M., Malone, S., Bodin, P., Hartog, R., and Wright, A.: Curated Pacific Northwest AI-ready Seismic Dataset, Seismica, 2, https://doi.org/10.26443/seismica.v2i1.368, 2023.

Ni, Y., Denolle, M. A., Thomas, A. M., Hamilton, A., Münchmeyer, J., Wang, Y., Bachelot, L., Trabant, C., and Mencin, D.: A Global-scale Database of Seismic Phases from Cloud-based Picking at Petabyte Scale, Seismica, 4, https://doi.org/10.26443/seismica.v4i2.1738, 2025.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Northern California Earthquake Data Center: Northern California Earthquake Data Center (NCEDC), UC Berkeley Seismological Laboratory [data set], https://doi.org/10.7932/NCEDC, 2014.

Northern California Earthquake Data Center: The NCSS now reports earthquake depth relative to the geoid (sea level), https://www.ncedc.org/blog/2015/10/07/the-ncss-now-reports-earthquake.html, 7 October 2015; last access: 13 September 2026, 2015.

Perol, T., Gharbi, M., and Denolle, M.: Convolutional neural network for earthquake detection and location, Science Advances, 4, e1700 578, https://doi.org/10.1126/sciadv.1700578, 2018.

Ross, Z. E., Meier, M.-A., Hauksson, E., and Heaton, T. H.: Generalized Seismic Phase Detection with Deep Learning, Bulletin of the Seismological Society of America, 108, 2894–2901, https://doi.org/10.1785/0120180080, 2018.

Ross, Z. E., Idini, B., Jia, Z., Stephenson, O. L., Zhong, M., Wang, X., Zhan, Z., Simons, M., Fielding, E. J., Yun, S.-H., Hauksson, E., Moore, A. W., Liu, Z., and Jung, J.: Hierarchical Interlocked Orthogonal Faulting in the 2019 Ridgecrest Earthquake Sequence, Science, 366, 346–351, https://doi.org/10.1126/science.aaz0109, 2019.

Southern California Earthquake Data Center: Southern California Earthquake Data Center (SCEDC), Caltech [data set], https://doi.org/10.7909/C3WD3xH1, 2013.

Southern California Earthquake Data Center: News and Updates: 2015 earthquake depth datum announcement, https://scedc.caltech.edu/news/index.html, announcement dated 22 September 2015; last access: 13 September 2026, 2015.

The HDF Group: Hierarchical Data Format, version 5, The HDF Group [software], https://www.hdfgroup.org/solutions/hdf5/, 2000–2026.

USGS Menlo Park: USGS Northern California Seismic Network, International Federation of Digital Seismograph Networks [seismic network], https://doi.org/10.7914/SN/NC, 1966.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., Michelini, A., Saul, J., and Soto, H.: SeisBench—A Toolbox for Machine Learning in Seismology, Seismological Research Letters, 93, 1695–1709, https://doi.org/10.1785/0220210324, 2022.

Yu, Z., Liu, X., and Cai, Y.: SeismicX-Cont (Revision a9d5f3d), Hugging Face [data set], https://doi.org/10.57967/hf/10283, 2026a.

Yu, Z., Liu, X., and Cai, Y.: SeismicX-Cont-mini (Revision 883e092), Hugging Face [data set], https://doi.org/10.57967/hf/9664, 2026b.

Yue, H., Sun, J., Wang, M., Shen, Z., Li, M., Xue, L., Lu, W., Zhou, Y., Ren, C., and Lay, T.: The 2019 Ridgecrest, California Earthquake Sequence: Evolution of Seismic and Aseismic Slip on an Orthogonal Fault System, Earth and Planetary Science Letters, 570, 117 066, https://doi.org/10.1016/j.epsl.2021.117066, 2021.

Zhang, M., Ellsworth, W. L., and Beroza, G. C.: Rapid Earthquake Association and Location, Seismological Research Letters, 90, 2276–2284, https://doi.org/10.1785/0220190052, 2019.

Zhao, M., Xiao, Z., Chen, S., and Fang, L.: DiTing: A Large-Scale Chinese Seismic Benchmark Dataset for Artificial Intelligence in Seismology, Earthquake Science, 36, 84–94, https://doi.org/10.1016/j.eqs.2022.01.022, 2023.

https://doi.org/10.5194/essd-2026-753
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Zhu, W. and Beroza, G. C.: PhaseNet: a deep-neural-network-based seismic arrival-time picking method, Geophysical Journal International, 216, 261–273, https://doi.org/10.1093/gji/ggy423, 2019.

Zhu, W., McBrearty, I. W., Mousavi, S. M., Ellsworth, W. L., and Beroza, G. C.: Earthquake phase association using a Bayesian Gaussian mixture model, Journal of Geophysical Research: Solid Earth, 127, e2021JB023 249, https://doi.org/10.1029/2021JB023249, 2022.

Zhu, W., Hou, A. B., Yang, R., Datta, A., Mousavi, S. M., Ellsworth, W. L., and Beroza, G. C.: QuakeFlow: a scalable machine-learning-based earthquake monitoring workflow with cloud computing, Geophysical Journal International, 232, 684–693, https://doi.org/10.1093/gji/ggac355, 2023.

Zhu, W., Wang, H., Rong, B., Yu, E., Zuzlewski, S., Tepp, G., Taira, T., Marty, J., Husker, A., and Allen, R. M.: California Earthquake Dataset for Machine Learning and Cloud Computing, arXiv [preprint], https://doi.org/10.48550/arXiv.2502.11500, 2025.