Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fifty-year seismicity analysis of the East Anatolian Fault System (1976–2026): leakage-safe probabilistic forecasting for seismic-risk resilience

Mehmet Tevfik AĞDAŞ

mtagdas@munzur.edu.tr

Munzur University

Research Article

Keywords: East Anatolian Fault System, Seismicity, Operational Earthquake Forecasting, Seismic-Risk Resilience, ETAS, Rare-Event Machine Learning, Calibration, Pseudo-Prospective Validation

Posted Date: August 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10650673/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Fifty-year seismicity analysis of the East Anatolian Fault System (1976–2026): leakage-safe probabilistic forecasting for seismic-risk resilience

Mehmet Tevfik AĞDAŞ

Çemişgezek Vocational School, Munzur University, Tunceli, Türkiye

Corresponding author: Mehmet Tevfik AĞDAŞ; e-mail: mtagdas@munzur.edu.tr

Manuscript type: Original Research Article

Abstract

Background: The East Anatolian Fault System (EAFS) has experienced pronounced changes in seismicity rate, catalog sensitivity, and sequence behavior over five decades. We aimed to characterize this 1976–2026 record and test whether historical seismicity supports reproducible short-term fault-trace probability updates relevant to seismic-risk preparedness.

Methods: USGS and AFAD catalogs were merged with source provenance, conservative duplicate matching, pre-test magnitude harmonization, and time-varying completeness control. The dense EAF-corridor catalog contained 73 532 events; 5 026 events met the common Mc=3.4 modeling threshold. These events were associated computationally with 157 source-traceable active-fault geometries, yielding 1 509 869 fault-trace-day observations. The primary outcome was at least one M≥4.0 event associated with a trace during the following seven days. Training ended in 2019, model selection used 2020–2022, 2023 was locked for testing, and 2024–29 July 2026 was retained as a forward holdout. Formal inference used non-overlapping seven-day origins and 28-day moving-block bootstrap confidence intervals.

Results: The validation-selected History+ETAS logistic forecast achieved PR-AUC values of 0.3479 in 2023 and 0.03261 in the forward holdout, compared with 0.0388 and 0.01127 for a Bayesian historical fault-rate reference. Paired PR-AUC gains were 0.314 (95% CI 0.157–0.443) and 0.0201 (0.00888–

0.03844), respectively. However, prespecified incremental tests did not support unique added value from the frozen ETAS parametrization, graph/stress proxies, boosted trees, or deep learning.

Conclusions: A source-auditable 50-year EAFS benchmark indicates that recent seismic history contains reproducible short-term information beyond a static fault-rate reference, including after the exceptional 2023 sequence. More complex model components should not be assigned physical or operational significance until they demonstrate independent prospective skill. The framework supports cautious probabilistic risk communication rather than deterministic earthquake prediction.

Keywords: East Anatolian Fault System; Seismicity; Operational Earthquake Forecasting; Seismic-Risk Resilience; ETAS; Rare-Event Machine Learning; Calibration; Pseudo-Prospective Validation

Background

The East Anatolian Fault System is one of the principal active continental fault systems of Türkiye and has generated destructive earthquakes across markedly different instrumental eras. A longitudinal analysis spanning 1976–2026 therefore offers two complementary opportunities: to characterize how recorded seismicity evolved over half a century and to test whether that historical information can support reproducible short-term probabilistic forecasts. The 24 January 2020 Elazığ earthquake and the February 2023 Kahramanmaraş sequence demonstrate why both perspectives are necessary: the system exhibits strong temporal clustering, multi-fault rupture, changing network sensitivity, and abrupt distribution shifts that can make a model appear successful when it is evaluated only within a single active sequence.

Recent EAFS studies have provided high-resolution descriptions of the 2023 rupture zone, aftershock migration, and small-scale segmentation [1, 2], while modern curated regional data sets improve event-location and waveform quality over recent periods [3]. These resources are essential for process studies, but a different problem remains underdeveloped: constructing a long-horizon, source-auditable benchmark that spans changing catalog regimes and can evaluate forecast generalization before, during, and after major sequences. A 50-year design must therefore separate descriptive long-term seismicity from the denser period used for model features, explicitly control magnitude completeness, and prevent later high-quality observations from retrospectively improving preprocessing applied to earlier forecast experiments.

Operational earthquake forecasting (OEF) does not seek deterministic prediction; it estimates transparent probabilities for specified locations, magnitudes, and time windows [4, 5]. ETAS-type models provide a principled statistical description of clustering [6], whereas machine-learning and deep-learning models can capture nonlinear interactions. The latter flexibility is useful only if it survives future-data isolation. In rare-event settings such as fault-trace-associated M≥4 occurrence, ROC-AUC alone can overstate practical value; precision-recall skill, calibration, proper scoring rules, and uncertainty on paired model differences are required [8]. The CSEP framework similarly emphasizes prospective or delayed pseudo-prospective evaluation, frozen forecasts, and reproducible comparison [9, 10].

Research gap and contribution

The originality of this study is the integration of a 50-year EAFS seismicity analysis with a leakage-safe fault-trace forecasting experiment. Rather than beginning with a preferred algorithm, the workflow first constructs a longitudinal seismic frame with explicit source provenance, duplicate control, magnitude harmonization fitted only on pre-test data, annual completeness analysis, and deterministic fault-trace identities. It then converts the dense 2000–2026 portion into more than 1.5 million daily fault-trace observations and evaluates a prespecified hierarchy of history, ETAS-style, graph, stress-proxy, boosted-tree, and deep-learning models. The design combines a locked 2023 test, an untouched 2024–2026 forward holdout, non-overlapping seven-day inference, moving-block bootstrap, rolling-origin tests, leave-mainshock-sequence-out tests, spatial blocks, fault holdouts, and independent evidence gates for each added modeling component. The resulting contribution is both seismological and methodological: it quantifies half a century of EAFS catalog behavior while testing which components remain predictive after the exceptional 2023 sequence is no longer the only evaluation regime.

Accordingly, the central question is not whether a high-capacity model can reproduce the 2023 aftershock field, but whether information learned from the historical EAFS record retains calibrated probabilistic skill under later, unseen conditions. This framing makes negative results informative: failure of a fixed ETAS parametrization, physics proxy, graph feature family, or deep architecture to add independent skill constrains interpretation rather than being hidden as an unsuccessful experiment.

Study Objectives and Hypotheses

- H1: A calibrated dynamic seismic-history forecast should outperform a static Bayesian fault-rate reference in both the locked 2023 test and the later forward holdout.
- H2: ETAS-style triggering features should add skill beyond lagged seismic history alone.
- H3: fault-graph and static-stress proxy features should add skill beyond history plus ETAS.
- H4: boosted-tree or deep neural models should outperform the calibrated logistic reference if meaningful nonlinear predictive structure exists.
- H5: any auxiliary magnitude or conformal uncertainty claim should be retained only when its own independent quality gate is passed.

Beyond its geophysical relevance, a transparent short-term forecasting benchmark has environmental-management implications for disaster preparedness, infrastructure continuity, land-use planning, and the resilience of communities exposed to cascading earthquake impacts. The study therefore treats forecast probabilities as decision-support information for risk reduction, not as deterministic event predictions.

Methods

Study Area and Temporal Scope

The acquisition box covered 34.0–42.5°N and 34.0–44.5°E, with final event inclusion refined by a 120-km buffer surrounding an EAF-system anchor polyline from the Karlıova/Bingöl region toward Hatay. Crustal modeling was limited to depths ≤50 km. The long-term acquisition period extended from 6 August 1976 through 6 August 2026, whereas the dense catalog began on 1 January 2000. The primary forecast target was binary fault-trace activation: at least one M≥4.0 event assigned to a trace during days t+1 through t+7. Sensitivity targets of M≥3.5 and M≥4.5 and horizons of 1, 3, and 30 days were generated but not used to redefine the primary hypothesis after test inspection.

Earthquake Catalogs and Source Audit

The executable workflow queried USGS and AFAD services in temporal chunks and attempted KOERI acquisition through the available FDSN client. In the executed run, KOERI exposed waveform/station-

related FDSN services but no compatible event-service endpoint, and therefore contributed zero catalog events; the FDSN data-center registry was rechecked on 8 August 2026 [21]. This is reported as a service-availability limitation rather than interpreted as zero seismicity. The potential impact is not uniform across the study: the primary machine-learning interval begins in 2000, AFAD contributed 109,206 raw dense-period records, the forecasting target is M≥4.0, and a common Mc=3.4 threshold is imposed on feature-generating seismicity. These choices reduce—but do not eliminate—the sensitivity of the primary results to missing smaller KOERI events. The omission may still affect near-threshold short-window counts, hypocentral precision, duplicate confirmation, and early-period completeness, so the study does not claim network-complete national catalog coverage.

The source-level catalog audit and availability status are summarized in Table 1.

Table 1. Source-Level Catalog Audit.

Source	Rows	First event	Last event	Status
USGS long (M≥4.0)	2 246	22 Aug 1976	18 Jul 2026	Available
USGS dense (M≥2.0)	2 159	2 Jan 2000	18 Jul 2026	Available
AFAD long (M≥4.0)	2 154	26 Feb 1990	31 Jul 2026	Available
AFAD dense (M≥2.0)	109 206	12 Jan 2000	5 Aug 2026	Available
KOERI long	0	—	—	Event service unavailable in executed client
KOERI dense	0	—	—	Event service unavailable in executed client

This table documents the source-level composition of the earthquake data before cross-catalog duplicate merging. AFAD dominates the dense post-2000 record, whereas USGS provides an independent international source and longer historical coverage for larger events; KOERI contributed no events because the executed route did not expose a compatible event service. This audit makes the provenance imbalance explicit and supports the later decision to control completeness rather than assume that all source periods have equivalent detection sensitivity.

Counts are raw source-level records after initial geographic/value-range quality control, before cross-catalog duplicate merging. The absence of KOERI data is an executed-run limitation, not an imputed zero-seismicity interval.

The temporal variation in source-level recording density is shown in Fig. 1.

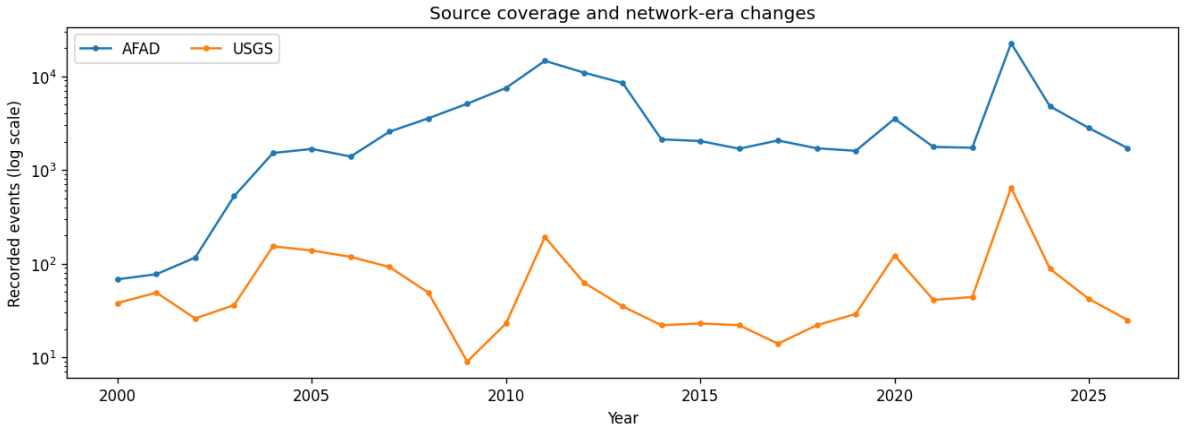


Figure 1. Annual event counts by catalog source. The logarithmic scale shows strong temporal changes in recording density and motivates year-specific completeness estimation rather than stationary catalog sensitivity.

This figure demonstrates that catalog density is strongly nonstationary across the study period, with AFAD contributing substantially more events than USGS during the dense era. This observation is methodologically important because a single fixed low-magnitude threshold would mix periods with different detection capabilities. The year-specific completeness analysis and the conservative common modeling threshold were therefore necessary to make longitudinal seismic-history features more comparable across network eras.

The spatial distribution of the 50-year M≥4 seismicity together with the active-fault geometry is shown in Fig. 2.

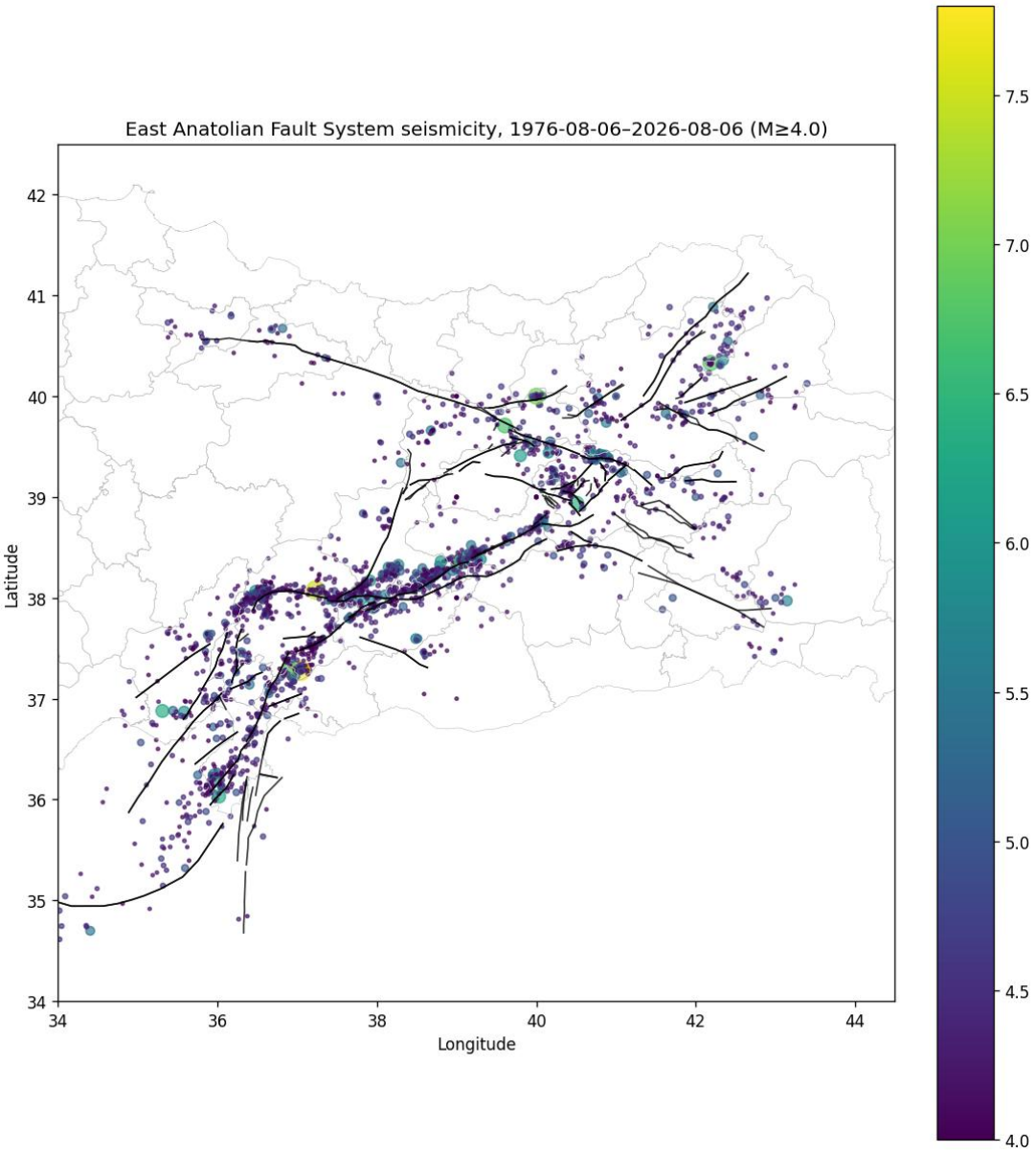


Figure 2. Fifty-year EAFS seismicity and active-fault geometry. M≥4 events from 1976–2026 are shown with active-fault geometry. Named-segment labels were not assigned because the executed GEM subset was source-traceable but unnamed.

This figure provides the spatial context for the full 50-year analysis and shows that M≥4 seismicity is concentrated along and around the East Anatolian fault corridor rather than being uniformly distributed across the acquisition box. The coexistence of dense earthquake clusters and multiple nearby fault traces also explains why the analysis uses computational trace identifiers and an explicit distance-based

association rule. Because the underlying GEM subset did not provide authoritative segment names, the figure supports regional fault-trace forecasting but not named-segment hazard claims.

Quality Control, Duplicate Matching and Event Consolidation

Records with missing time, latitude, longitude, or magnitude were removed, as were coordinates outside the study box and physically implausible magnitude/depth values. Potential cross-catalog duplicates were matched conservatively using a maximum origin-time separation of 45 s, epicentral distance ≤25 km, and magnitude difference ≤0.6. The dense merge yielded 108,393 unique event groups: 106,720 single-source groups and 1,673 groups supported by two sources. The merged long and dense regional catalogs contained 3,028 and 108,393 unique events, respectively; after the EAF corridor filter, 2,039 long-catalog and 73,532 dense-catalog events remained.

For each duplicate group, location/time attributes were selected using a fixed source-priority and reported-uncertainty rule, while the preferred magnitude estimate prioritized reported moment magnitude and then harmonized alternatives. The algorithm retained member IDs and all contributing sources so that merged events remained auditable.

Magnitude Harmonization without Test Leakage

Magnitude conversions were estimated only from matched events dated on or before 31 December 2019. Robust Huber regression was used to map non-Mw types onto a moment-magnitude reference constructed from overlapping reported Mw values. This restriction prevents the 2020–2026 validation/test/holdout periods from informing the magnitude conversion itself. Unsupported magnitude types were retained only as flagged raw fallbacks with enlarged uncertainty and were excluded from strict sensitivity interpretations when appropriate.

The pre-test magnitude-conversion models used for harmonization are summarized in Table 2.

Table 2. Pre-Test Magnitude-Conversion Models.

Source	Input type	Pairs	Intercept	Slope	RMSE	Input range
AFAD	ML	41	−0.132	1.037	0.206	4.1–6.7
USGS	MB	99	1.122	0.732	0.173	3.4–5.1

This table shows that magnitude harmonization was estimated only from pre-test matched events, thereby preventing 2020–2026 observations from informing the conversion equations. The RMSE values indicate

moderate conversion uncertainty, and the relatively small numbers of paired observations—especially for AFAD ML—are the reason these equations are treated as nuisance harmonization models rather than universal scaling laws. Restricting conversions to their observed input ranges further limits extrapolation error near the M≥4 target threshold.

Mapping form: Mw = intercept + slope × Minput. Fits use overlap events available through the training cutoff only.

The two fitted conversion relations are based on 41 AFAD ML–Mw pairs and 99 USGS mb–Mw pairs. They are therefore treated as nuisance harmonization models, not universal magnitude-scaling laws. Their use is restricted to the observed input ranges reported in Table 2; unsupported or out-of-range magnitudes retain explicit uncertainty flags. Because these conversions are fitted only before 2020 and are not selected by test-period performance, uncertainty in the small paired samples can reduce feature precision but cannot introduce future-information leakage.

Active-Fault Geometry and Provenance

Active-fault geometries were obtained from the GEM Global Active Faults Database (GAF-DB), a worldwide, source-attributed compilation of active fault traces [13]. The workflow preserved available catalog identifiers, references, names, kinematic fields, and source-file provenance. Multipart features were exploded into computational traces. Because a catalog feature can contain multiple geometries, each trace received a deterministic identity formed from the catalog identifier and a SHA-1 geometry hash. Exact duplicate geometries were removed before graph construction.

Within the regional download, 411 trace geometries remained after cleaning; 219 exact duplicates were removed and all 411 retained traces had catalog identifiers. The EAF-system corridor selected 157 candidate traces. None carried a usable fault name in the executed GEM subset, although all remained source-traceable through stable identifiers. Consequently, the manuscript deliberately avoids named-segment performance claims. A named-segment interpretation would require an independently curated MTA or equivalent authoritative layer.

Magnitude Completeness and Modeling Threshold

Catalog completeness was estimated annually using maximum-curvature (MAXC) with a +0.2 magnitude correction together with a goodness-of-fit diagnostic. The dense EAF catalog contained 73,532 events, of which 39,082 satisfied the time-varying completeness criterion. To preserve rate comparability across years, the machine-learning catalog used a common conservative completeness threshold of Mc=3.4 from 2000 onward. This produced 5,026 complete events for lagged history and sequence-feature construction. The primary target remained M≥4.0, above the modeling completeness threshold.

The year-specific variation in magnitude of completeness is shown in Fig. 3.

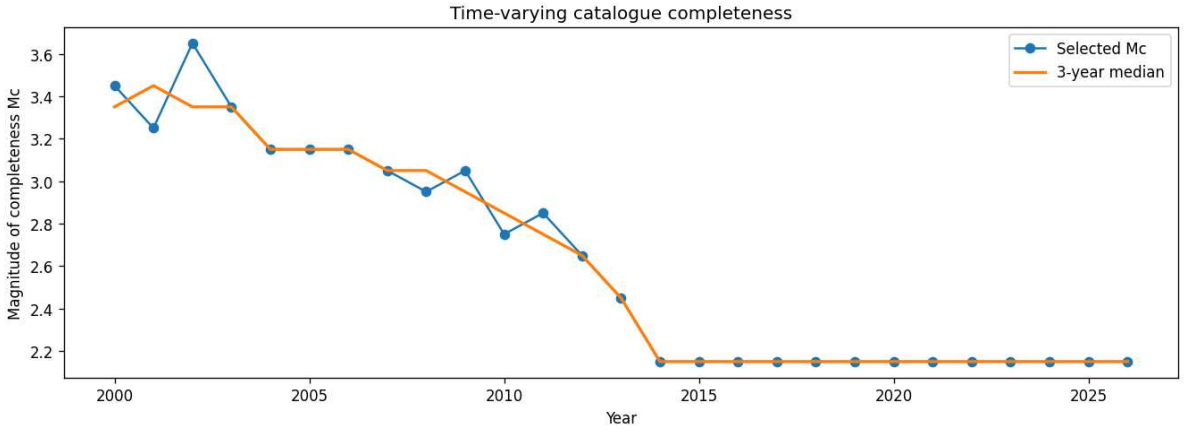


Figure 3. Year-specific magnitude of completeness. Annual Mc estimates show nonstationary detection capability associated with network and catalog changes.

This figure confirms that the estimated magnitude of completeness changes through time, reflecting evolving network coverage and catalog sensitivity. This variation is directly relevant to the forecasting design because short-window event counts can be artificially inflated when later, denser catalog periods are compared with earlier periods. Adopting a common Mc=3.4 threshold sacrifices smaller events but reduces this source of temporal measurement bias in the primary modeling catalog.

The Gutenberg–Richter frequency–magnitude relationship used as a descriptive completeness diagnostic is shown in Fig. 4.

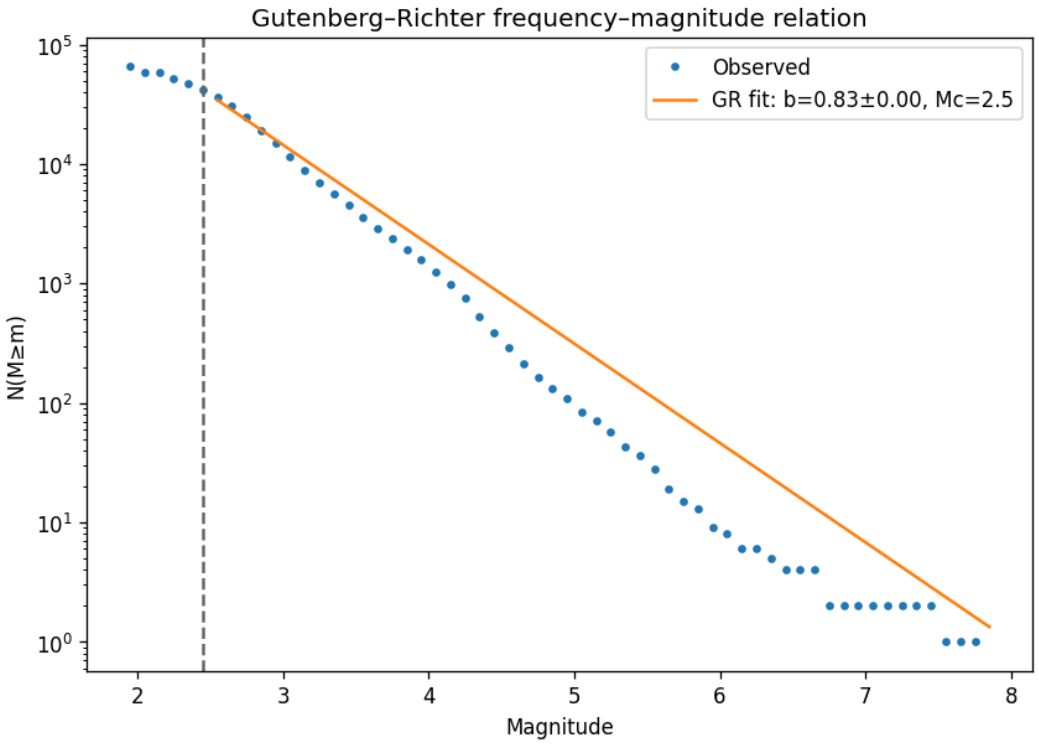


Figure 4. Gutenberg–Richter frequency–magnitude diagnostic. The descriptive fit produced b≈0.83 with a lower descriptive Mc near 2.5; forecasting used the conservative cross-year Mc=3.4 threshold for rate comparability.

This figure summarizes the frequency–magnitude structure of the merged EAFS catalog and shows that the descriptive Gutenberg–Richter relation is broadly compatible with a b-value near 0.83. The lower descriptive completeness level is not used as the forecasting threshold because it is not stable across all years. The distinction between descriptive Mc and the more conservative modeling Mc=3.4 is therefore intentional and protects the forecast experiment from changes in catalog detectability.

The annual Gutenberg–Richter b-value estimates and their temporal variability are shown in Fig. 5.

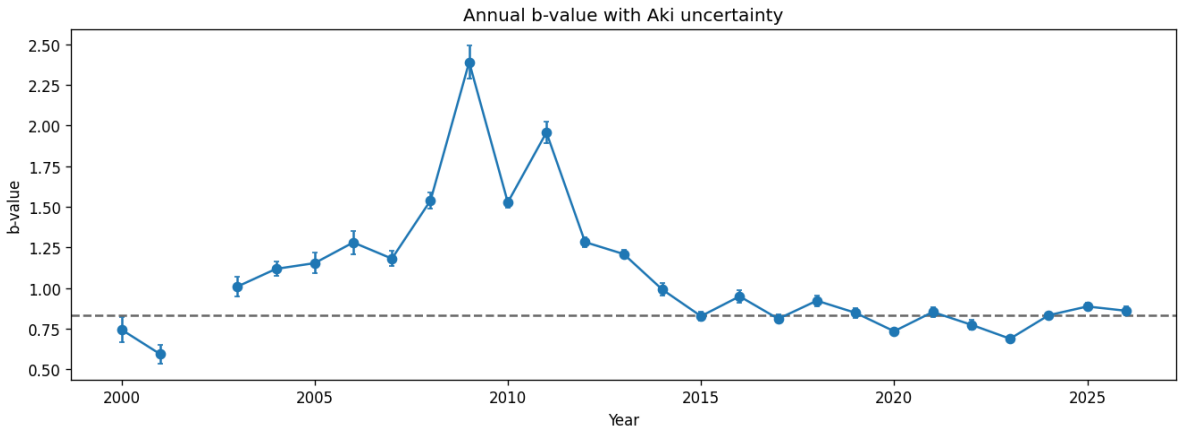


Figure 5. Annual Gutenberg–Richter b-values. Aki uncertainty [15] accompanies annual b-value estimates. Variability is descriptive and was not used as a future-informed tuning signal.

This figure shows meaningful year-to-year variation in estimated b-values, with uncertainty increasing in years containing fewer complete events. The pattern should not be interpreted as a direct deterministic precursor; rather, it documents temporal heterogeneity in the observed seismic regime and catalog composition. Importantly, these annual estimates were used as descriptive diagnostics and were not allowed to tune the future test or holdout models.

Forecast Design and Modeling

Forecast Unit and Leakage-Safe Fault-Trace-Day Construction

Each complete earthquake was projected into a suitable UTM coordinate system and assigned to its nearest EAF candidate trace if the nearest-trace distance was ≤40 km. A daily panel was then constructed for every retained fault trace. For a forecast origin t, historical predictors were computed only from data strictly before t by an explicit one-day shift. The binary target equaled 1 when any M≥4.0 event assigned to the trace occurred during t+1,…,t+7. The final seven days of each series, for which the future target could not be fully observed, were removed.

The resulting pre-enrichment panel contained 1,509,869 rows and 31 fields, with an overall target-event rate of 0.394%. After enriching the panel with ETAS-style and stress-proxy predictors, the complete feature set

contained 24 predictors. The very low event rate makes accuracy inappropriate as a primary metric and motivates precision-recall analysis, proper probability scores, and explicit calibration.

The chronological training, validation, locked-test, and forward-holdout partitions are summarized in Table 3.

Table 3. Chronological data partitions and prevalence.

Partition	Dates	Rows	Positive rate
Training	2000-03-31 to 2019-12-24	1 131 656	0.250%
Validation	2020-01-01 to 2022-12-24	170 973	0.367%
Locked test	2023-01-01 to 2023-12-24	56 206	2.875%
Forward holdout	2024-01-01 to 2026-07-29	147 737	0.577%
Validation, formal non-overlap	2020–2022	24 492 (88 positive)	0.359%
Test, formal non-overlap	2023	8 007 (234 positive)	2.922%
Holdout, formal non-overlap	2024–2026	21 195 (121 positive)	0.571%

This table makes the chronological evaluation design explicit and shows the major prevalence shift between training, validation, the 2023 locked test, and the 2024–2026 forward holdout. The 2023 positive rate is several times higher than the later holdout rate because of the Kahramanmaraş sequence, which explains why absolute PR-AUC values differ strongly across periods. The formal non-overlapping subsets preserve the same chronology while preventing repeated evaluation of overlapping seven-day target windows.

A seven-day purge prevented forecast labels from crossing chronological boundaries. Formal inferential subsets retained only forecast origins spaced by the full seven-day horizon.

Predictor families

The predictor families included in the fault-trace forecasting panel are summarized in Table 4.

Table 4. Predictor families in the fault-trace panel.

Family	Fields	Definition	Unit
History	count_1d/3d/7d/30d/90d	Number of complete events in each trailing window	count
History	maxmw_1d/3d/7d/30d/90d	Largest harmonized Mw in each trailing window	Mw
History	moment_1d/3d/7d/30d/90d	Sum of scalar seismic moment in each trailing window	N m
Fault/graph	fault_length_km	Projected trace length	km
Fault/graph	fault_azimuth_deg	Trace azimuth	degrees
Fault/graph	neighbor_count_1d	Lagged graph-neighbor activity	weighted count
Physics proxy	stress_proxy_sum	Summed $M0/(4\pi r^3)$ proxy from prior events	Pa-like proxy
Physics proxy	signed_transfer_sum	Stress proxy × cos(2Δstrike)	proxy

Family	Fields	Definition	Unit
ETAS-style	etas_parent_prob_sum	Sum of normalized candidate-parent probabilities	0–1 aggregate
ETAS-style	etas_parent_prob_max	Maximum candidate-parent probability	0–1
Calendar	day_of_year_sin/cos	Cyclic seasonality controls	dimensionless

This table groups the 24 predictors according to the information they encode and clarifies that all dynamic variables are lagged relative to the forecast origin. History features capture recent event frequency, maximum magnitude, and accumulated scalar moment; graph, ETAS-style, and stress-proxy terms test whether additional contextual structure improves prediction. This organization is central to the ablation and claim-gate strategy because each feature family can be evaluated against a simpler nested reference rather than credited merely for appearing in a high-capacity model.

All dynamic predictors are lagged. The two stress variables are explicitly treated as proxies and not as Coulomb ΔCFS.

Fault graph

Candidate traces were connected when their minimum geometric separation was $\leq$25 km. Edge weight decayed exponentially with distance, $w=\exp(-d/15\ \text{km})$. The executed graph contained 157 nodes and 876 edges. Neighbor activity was computed from prior-day seismicity only, preventing same-day target leakage.

The graph constructed from the 157 source-traceable EAFS candidate fault traces is shown in Fig. 6.

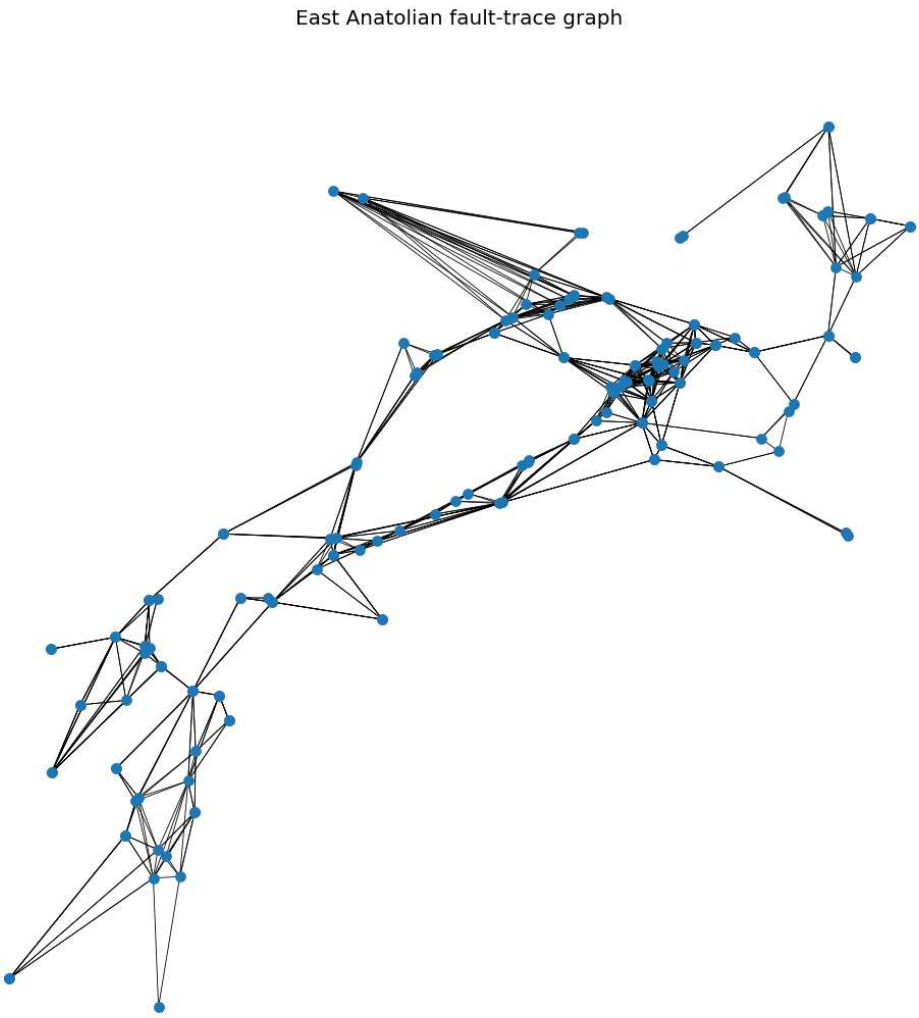


Figure 6. Graph of EAFS candidate fault traces. The 157 nodes are source-traceable candidate traces. Edges connect geometrically proximal traces and support lagged neighbor-activity predictors; the graph does not establish causal stress transfer.

This figure converts the 157 retained fault traces into a proximity graph that allows prior activity on neighboring traces to enter the forecast as a lagged contextual predictor. The graph is intentionally geometric rather than causal: an edge indicates spatial proximity and a decaying weight, not verified mechanical stress transfer. This distinction is important because the subsequent failure of the graph/physics increment gate cannot be interpreted as evidence against physically based interaction models that use finite-fault mechanics.

Omori–Utsu and ETAS-Style Sequence Features

Mainshock candidates were identified from complete seismicity, and sequence windows were defined with a 30-day duration and a magnitude-dependent radius using Wells–Coppersmith [16] strike-slip rupture-length scaling. Aftershock-rate decay [17] was fitted as $n(t)=K/(t+c)^p$ by bounded least squares. The February 2023 Mw7.8 sequence yielded K=2277.24, c=4.836 d, p=0.713 and 9,585 sequence events in the Omori diagnostic; the 2020 Mw6.7 Elazığ sequence yielded K=260.82, c=0.396 d and p=1.107.

A temporal ETAS maximum-likelihood model was also fitted on eligible pre-test sequences. Candidate parent-child links for the complete modeling catalog were scored using a simplified ETAS-style kernel proportional to $K \exp[\alpha(M_{parent}-M_c)]/(\Delta t+c)^p/(r+d)^q$ and normalized against a background term. Parameters were drawn from pre-2020 fitted sequence behavior. Each child retained at most three candidate parents within 30 days and 250 km. The executed workflow generated 13,593 candidate links. These probabilities are predictors, not genealogical truth.

Sequence-level Omori–Utsu fits show substantial heterogeneity: for example, the 2023 Mw7.8 sequence yielded K=2277.24, c=4.8358 d, and p=0.7125, whereas the 2020 Mw6.7 Elazığ sequence yielded K=260.82, c=0.3957 d, and p=1.1071; other sequences were less productive still (Supplementary Table S2). The forecasting features nevertheless use parameters estimated only from pre-2020 behavior so that the 2020–2026 evaluation periods cannot adapt the triggering kernel retrospectively. Consequently, the ETAS comparison tests a fixed, pre-test ETAS-style parametrization. Failure of its incremental gate must not be interpreted as evidence that all sequence-adaptive ETAS formulations lack value; the observed sequence heterogeneity is a plausible reason why a single frozen parametrization can be too rigid.

The fitted Omori–Utsu decay behavior for the identified major earthquake sequences is shown in Fig. 7.

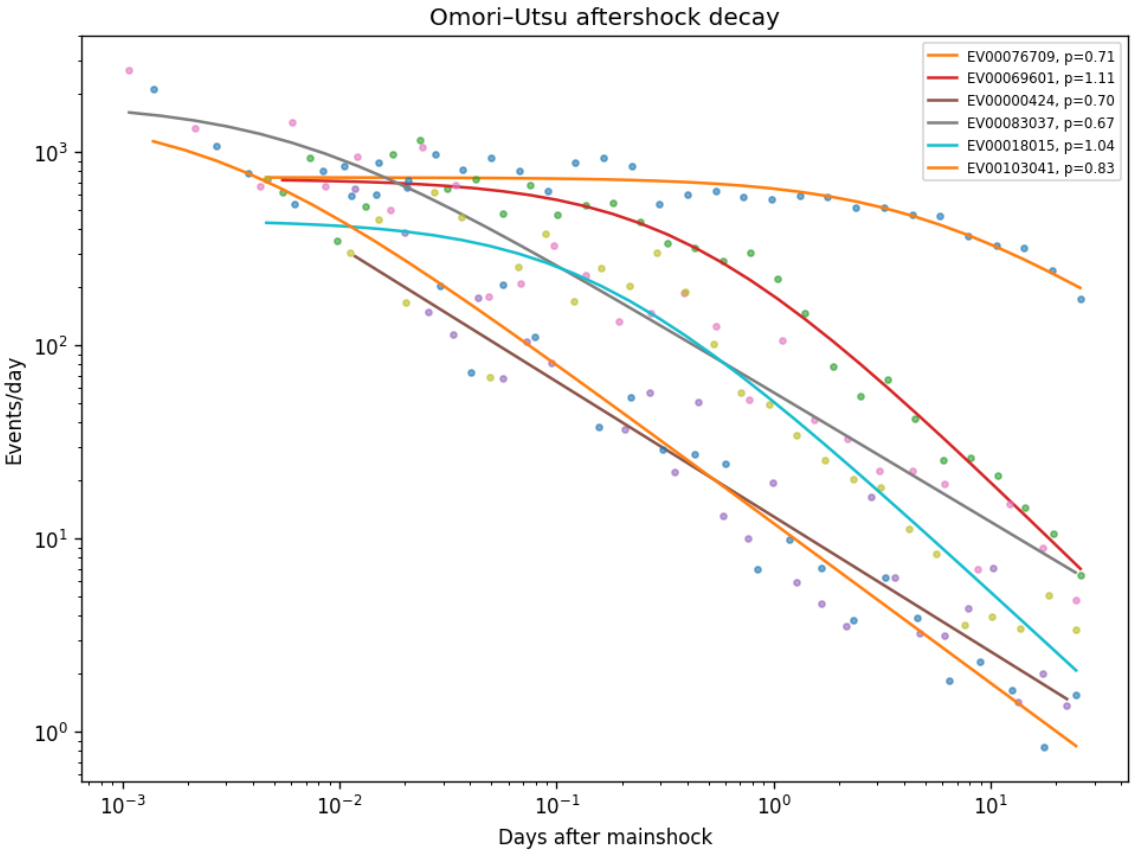


Figure 7. Omori–Utsu fits for major sequences. Variation in p, c, and sequence productivity illustrates heterogeneity among EAFS earthquake sequences.

This figure illustrates the pronounced sequence-to-sequence heterogeneity of aftershock decay in the EAFS. The large differences in productivity, c, and p between the 2020 Elazığ and 2023 Kahramanmaraş sequences show why a single frozen triggering regime may be too restrictive. This heterogeneity provides a plausible explanation for why fixed pre-test ETAS-style predictors did not demonstrate a stable incremental gain over lagged seismic history in both future evaluation periods.

The Båth-law diagnostic across the identified earthquake sequences is shown in Fig. 8.

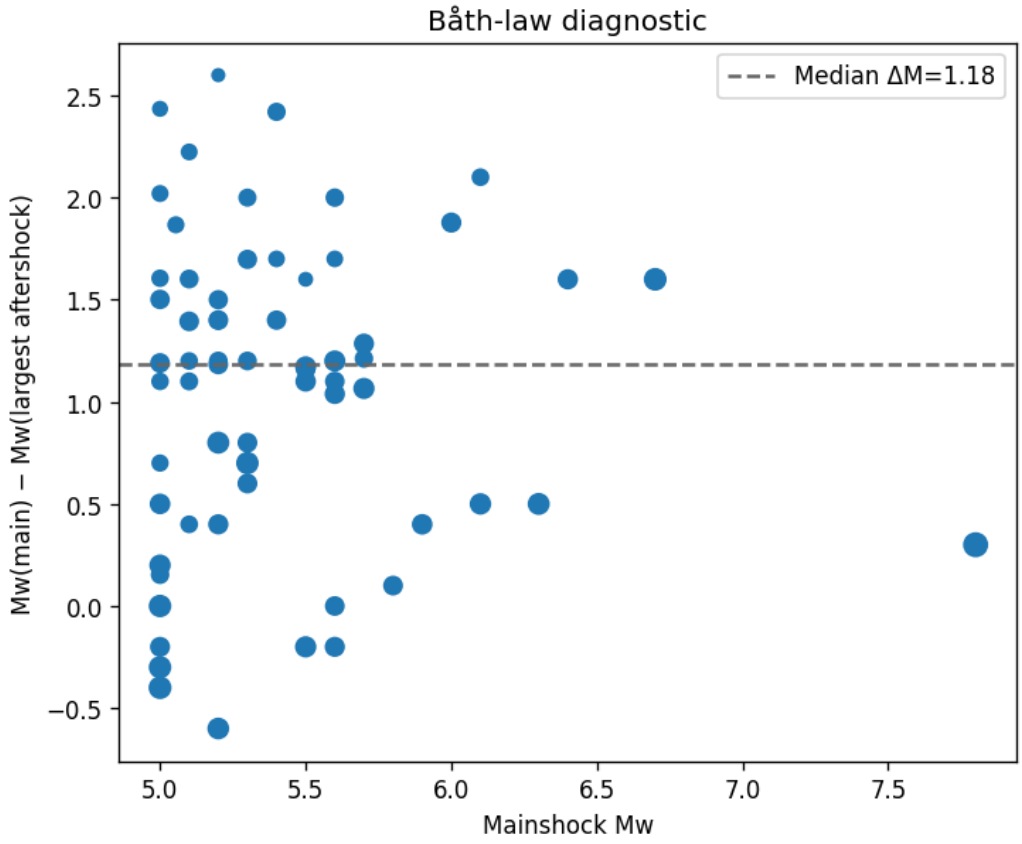


Figure 8. Båth-law diagnostic across identified sequences. The observed median mainshock–largest-aftershock magnitude difference was approximately 1.18.

This figure shows that the median mainshock-to-largest-aftershock magnitude difference is approximately 1.18 across the identified sequences, broadly consistent with the descriptive expectation of Båth-type behavior. The scatter around this central value is substantial, emphasizing that individual EAFS sequences differ in relative aftershock productivity. The diagnostic is therefore used to characterize the catalog and sequence structure rather than to impose a deterministic magnitude relationship on the forecast model.

The temporal evolution of aftershock migration following the 6 February 2023 mainshock is shown in Fig. 9.

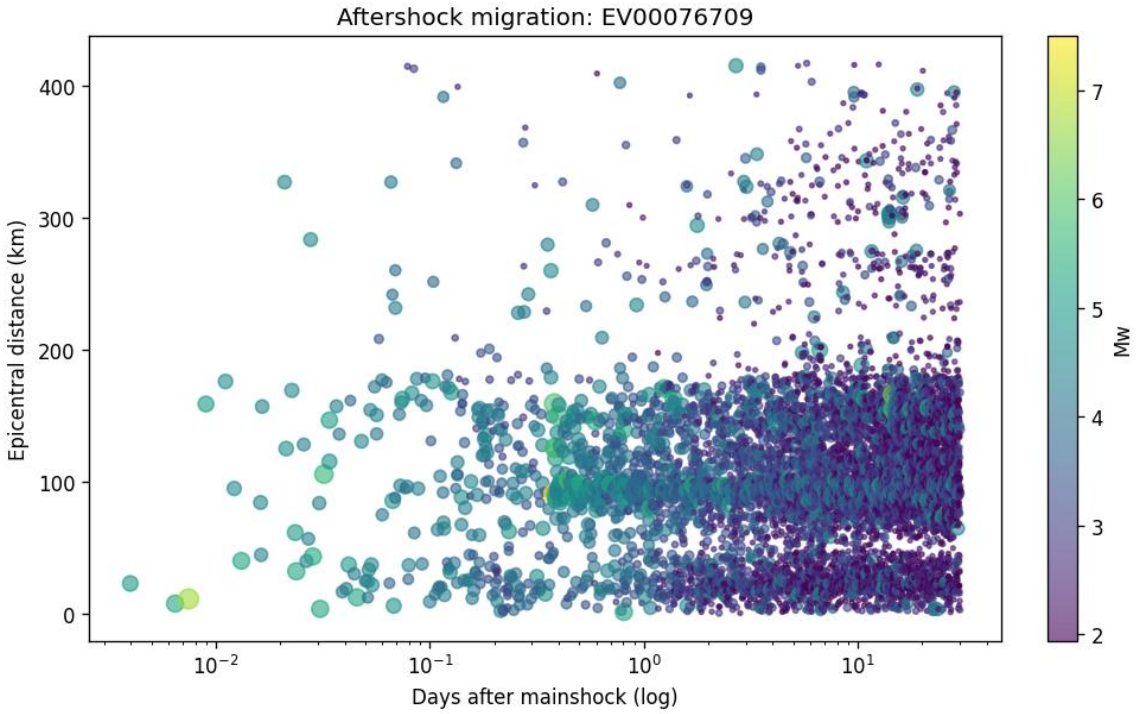


Figure 9. Aftershock migration after the 6 February 2023 mainshock. Event distance from the mainshock is shown through time as descriptive sequence evidence, not as a tuned forecast predictor.

This figure visualizes the spatiotemporal expansion of the 6 February 2023 aftershock sequence relative to the mainshock. The migration pattern underscores the rapid redistribution of seismic activity across a geometrically complex rupture system and helps explain the strong distribution shift observed in the locked 2023 test set. Because the diagnostic is not used as a fitted predictor, it provides independent descriptive context rather than a source of test-period information leakage.

Physics-guided stress proxies

Scalar seismic moment was calculated as M0=10^(1.5Mw+9.1) N m. A distance-decay proxy was defined as $S=M0/(4\pi r^3)$, with r bounded below at 1 km, and a signed orientation proxy used cos(2Δstrike) as a compatibility factor. These quantities provide parsimonious physical inductive bias but omit finite-fault slip distribution, receiver-fault geometry, friction, and elastic Green functions. They are therefore not Coulomb stress changes. The workflow exports Coulomb-ready templates but deliberately prohibits a ΔCFS claim until published finite-slip models and receiver-fault parameters are supplied.

Forecast Models and Calibration Hierarchy

The benchmark hierarchy was deliberately ordered from simple to complex. The static reference was an empirical-Bayes fault-specific historical rate with a global prior strength of 250 observations. Dynamic candidates included History Logistic (lagged history plus calendar terms), History+ETAS Logistic, Full Logistic (all 24 predictors), histogram gradient boosting, and XGBoost. Rare-event fitting retained every positive case and a temporally stratified negative sample capped a priori at 35 negatives per positive; inverse sampling weights restored the original negative mass during model fitting. Logistic candidates used median imputation, standardization, and a fixed C=0.25. This value was prespecified as mild L2 regularization and was applied identically to all logistic candidates rather than tuned on the validation/test periods. Probability calibration was then fitted on the full-prevalence, unsampled 2020–2022 validation set using Platt scaling when sufficient positive events were available. Thus case-control sampling is used for computationally stable estimation, while the reported probability scale is learned from an untouched prevalence distribution. The executed study did not select the 35:1 ratio by holdout performance, and no claim is made that this ratio is uniquely optimal.

A separate CAL-SEQ challenger used a transformer encoder over standardized feature tokens, a fault-trace embedding, and a bounded residual logit added to the Bayesian historical-rate logit. Auxiliary heads modeled conditional maximum magnitude and discrete seven-day event hazard. Event training used focal binary cross-entropy, the magnitude head used positive-event-only Gaussian negative log-likelihood, and the survival head used a right-censored discrete likelihood. CAL-SEQ contained 233,770 trainable parameters. Importantly, it was treated as a challenger; it could not replace the simpler operational model unless it passed the same validation, calibration, locked-test, and forward-holdout gates.

Evaluation Protocol and Model-Selection Discipline

The chronological partitions were fixed before final evaluation: training through 2019, validation in 2020–2022, locked testing in 2023, and forward holdout from 2024 through 6 August 2026 (observable targets ended 29 July 2026). Model selection used validation data only. Because daily seven-day forecasts overlap, formal inference retained origins separated by seven days. Daily-overlapping metrics were preserved only as secondary operational diagnostics.

The validation-selected forecast was defined by a deterministic deployment rule: among models meeting prespecified calibration guardrails (Brier score [18] no more than 1.05 times the prevalence reference and ECE≤0.005), the highest validation PR-AUC was selected. This ranking is not itself treated as evidence that the winning validation model is statistically superior to a near-tied candidate. In particular, the later similarity of History Logistic and History+ETAS is interpreted as model-selection uncertainty rather than retroactive evidence that the validation rule was incorrect. Scientific inference is based on the locked test and forward holdout, not on the magnitude of the validation rank difference. Pairwise uncertainty was estimated by a 28-day moving temporal block bootstrap over forecast-origin dates, preserving serial dependence within approximately four formal seven-day origins. A primary skill claim required positive 95% confidence intervals for ΔPR-AUC and Brier skill relative to the Bayesian historical-rate reference in both the locked test and the forward holdout.

Additional generalization analyses included rolling-origin yearly tests from 2018 through 2025, leave-one-mainshock-sequence-out evaluation, five spatial blocks, and leave-one-fault-out summaries. Claim-specific comparisons separately tested the ETAS increment (History+ETAS vs History), physics/graph increment (Full vs History+ETAS), boosting increment (XGBoost vs History+ETAS), and deep increment (CAL-SEQ vs History+ETAS). H1 was the single confirmatory performance claim and required concordant success in two future periods. H2–H5 were secondary falsification gates: no positive mechanistic or complexity conclusion was permitted unless its own paired comparison passed. Because all secondary gates failed even before multiplicity adjustment, applying a Holm family-wise correction cannot turn any of them into a positive result. The manuscript therefore reports no multiplicity-sensitive positive secondary inference [19].

Statistical Inference, Multiplicity, and Selection Uncertainty

The analysis distinguishes deployment selection from scientific hypothesis testing. Validation PR-AUC determines which eligible model is frozen, whereas evidential claims are evaluated only on future partitions using paired temporal-block resampling. The confirmatory claim is intentionally conjunctive: improvement must

be present in both 2023 and 2024–2026, which is more conservative than declaring success from either period alone. Secondary component tests are interpreted as nested falsification analyses and are not pooled to construct a composite post-hoc success criterion.

Evaluation Metrics

Discrimination was summarized using ROC-AUC and, more importantly for the rare-event setting, precision-recall AUC. Probability quality was evaluated using Brier score [18], logarithmic loss, and expected calibration error (ECE). Proper scores are central to earthquake forecast comparison because they reward calibrated probabilities rather than only thresholded classifications [8]. Decision-curve and Molchan-style diagnostics were retained as secondary visualizations. Conditional magnitude regression was evaluated by MAE, RMSE, and R^2, and conformal set reliability by empirical coverage, class-conditional coverage, and set size.

Use of Generative AI in Manuscript Preparation

OpenAI ChatGPT was used during manuscript preparation for language editing, document-formatting assistance, and code-review support. It was not used to create, alter, or fabricate earthquake observations or fault geometries. All executable analyses were run by the author, and all numerical outputs, references, interpretations, and final manuscript decisions were reviewed and verified by the author, who assumes full responsibility for the work.

Results

Fifty-Year Seismicity and Geographic Distribution

The long-term M≥4 catalog confirms that large-event representation is strongly concentrated in provinces intersecting the EAF system and in the 2023 rupture region. For example, Kahramanmaraş contributed 428 M≥4 events with a maximum Mw of 7.8, Malatya 296 with a maximum Mw of 6.1, Elazığ 137 with a maximum Mw of 6.7, Bingöl 130 with a maximum Mw of 6.4, and Hatay 110 with a maximum Mw of 6.3. These counts are

descriptive and should not be interpreted as normalized hazard because network sensitivity, fault length, and exposure differ among provinces.

Selected province-level descriptive M≥4 seismicity statistics are summarized in Table 5.

Table 5. Province-level M≥4 seismicity statistics.

Province	N	Max Mw	Median depth (km)
Kahramanmaraş	428	7.8	7.58
Malatya	296	6.1	8.33
Elazığ	137	6.7	11.00
Bingöl	130	6.4	10.00
Adıyaman	130	5.6	7.92
Erzurum	129	5.9	10.00
Hatay	110	6.3	10.00
Erzincan	104	7.0	10.00
Adana	97	6.3	10.38
Tunceli	39	6.1	10.50

This table summarizes the descriptive geographic distribution of M≥4 earthquakes and highlights the concentration of large-event activity in provinces intersecting the EAFS and the 2023 rupture region. These counts are not normalized for fault length, exposure, or network sensitivity and therefore should not be interpreted as comparative administrative hazard estimates. Their role is to provide geographic context for the 50-year seismicity record, while the predictive unit remains the source-traceable fault geometry.

Province assignment is descriptive. The forecasting unit is the source-traceable fault geometry, not the administrative province.

Catalog completeness and sequence structure

The catalog-density figure shows a pronounced increase in AFAD reporting after 2000, while the year-specific Mc analysis confirms that a fixed low-magnitude threshold would have mixed incompatible network eras. The conservative common Mc=3.4 therefore trades sample size for longitudinal comparability. Descriptive Gutenberg–Richter [14] behavior was broadly stable enough for a common target design, but annual b-values varied and the 2023 sequence produced a major temporary rate regime shift.

Sixty-five mainshock candidates generated 19,944 sequence-event rows. The 2023 Mw7.8 sequence was by far the most productive and exhibited a relatively low Omori p estimate, whereas the 2020 Mw6.7 Elazığ sequence

had p≈1.11. These differences motivated retaining dynamic history and triggering features while also requiring multi-year evaluation so that the 2023 aftershock cloud could not dominate the entire inference.

Exploratory daily model diagnostics

Before the prespecified non-overlapping protocol was applied, a broad daily-overlapping benchmark compared logistic regression, random forest, histogram gradient boosting, XGBoost, LightGBM, CatBoost, and a historical-rate baseline. The logistic model achieved the strongest forward-holdout PR-AUC (0.0235) in this exploratory set and was used for the ROC, PR, calibration, confusion-matrix, decision-curve, Molchan, and explainability visual diagnostics below. These plots are secondary; the primary inferential results are the non-overlapping, calibrated comparisons in Section 4.3.

The secondary daily-overlapping ROC performance for the locked 2023 period is shown in Fig. 10.

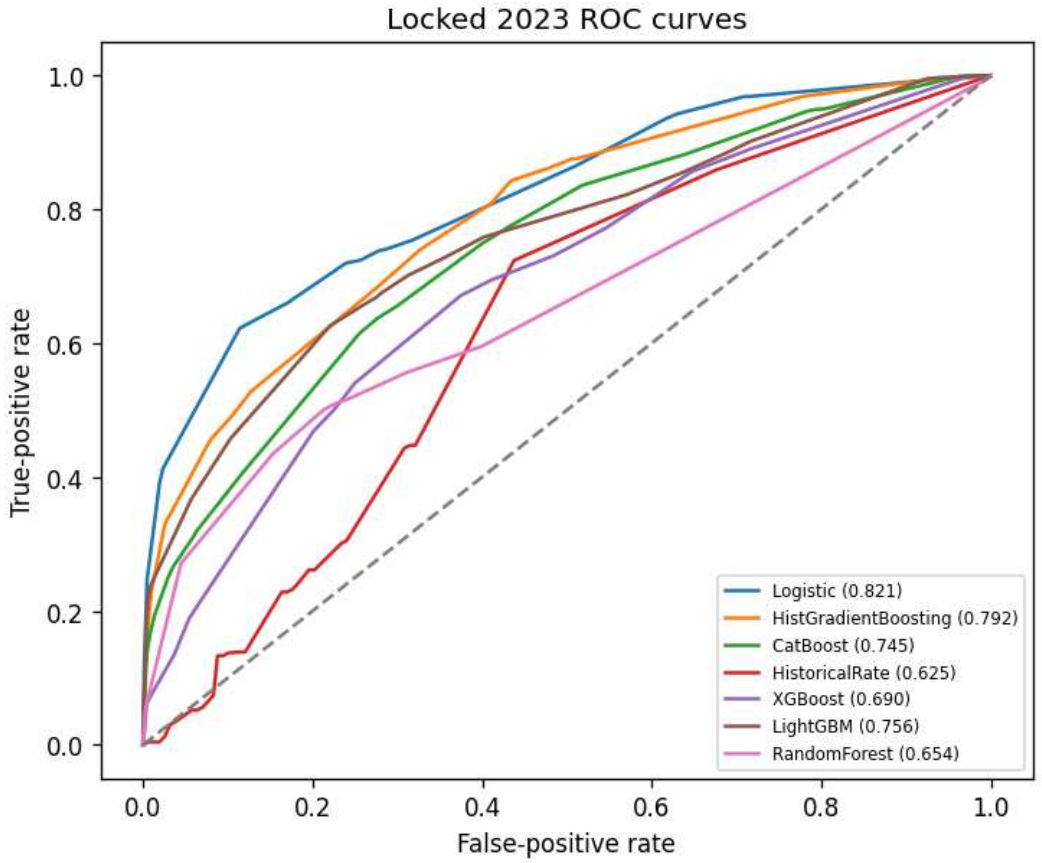


Figure 10. ROC curves for the locked 2023 period. Daily-overlapping discrimination is shown as a secondary diagnostic; formal inference uses non-overlapping seven-day origins.

This figure shows that several models can achieve apparently strong ROC discrimination during the highly active 2023 period, but ROC-AUC alone is not sufficient in this rare-event setting because the negative class dominates. The figure is therefore treated as a secondary diagnostic, while the formal comparison relies more heavily on precision–recall performance, calibration, and proper scoring rules. This prevents global ranking across abundant negatives from being mistaken for operationally meaningful rare-event skill. The corresponding precision–recall performance for the locked 2023 period is shown in Fig. 11.

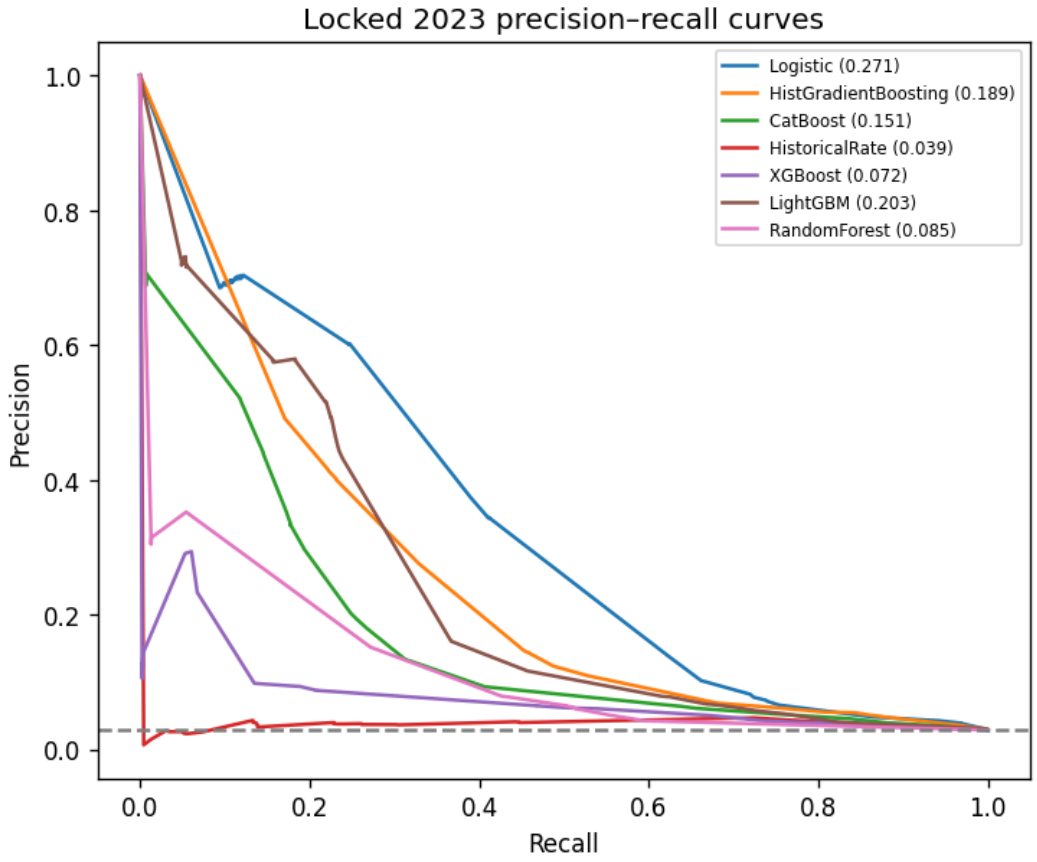


Figure 11. Precision–recall curves for the locked 2023 period. Precision–recall behavior is emphasized because the forecast target is strongly imbalanced.

This figure provides the more informative discrimination view for the imbalanced target by relating precision directly to recall. The separation between the dynamic models and the low-prevalence reference is much clearer in precision–recall space than in raw accuracy terms. The figure also illustrates why PR-

AUC is interpreted only within the same period: the much higher 2023 prevalence mechanically raises achievable precision compared with the later holdout.

The reliability behavior of the exploratory logistic forecast in the locked 2023 period is shown in Fig. 12.

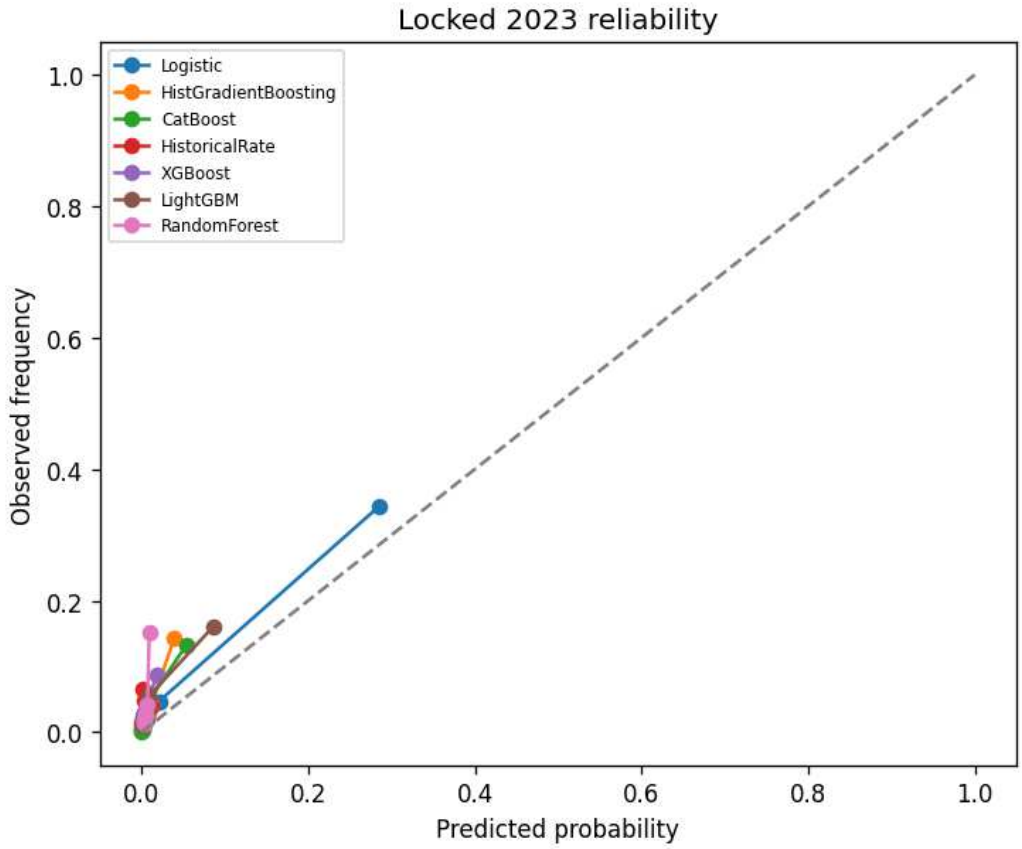


Figure 12. Reliability diagram for the locked 2023 period. Calibration is evaluated separately from ranking performance.

This figure demonstrates that ranking skill and probability calibration are separate properties of the forecast. A model can order higher-risk traces correctly while still over- or under-estimating absolute event probabilities, which is particularly consequential when probabilities are intended for risk communication. This motivates the study's use of Brier score, log loss, and ECE in addition to ROC-AUC and PR-AUC.

Threshold-based classification outcomes for the exploratory logistic model are summarized in Fig. 13.

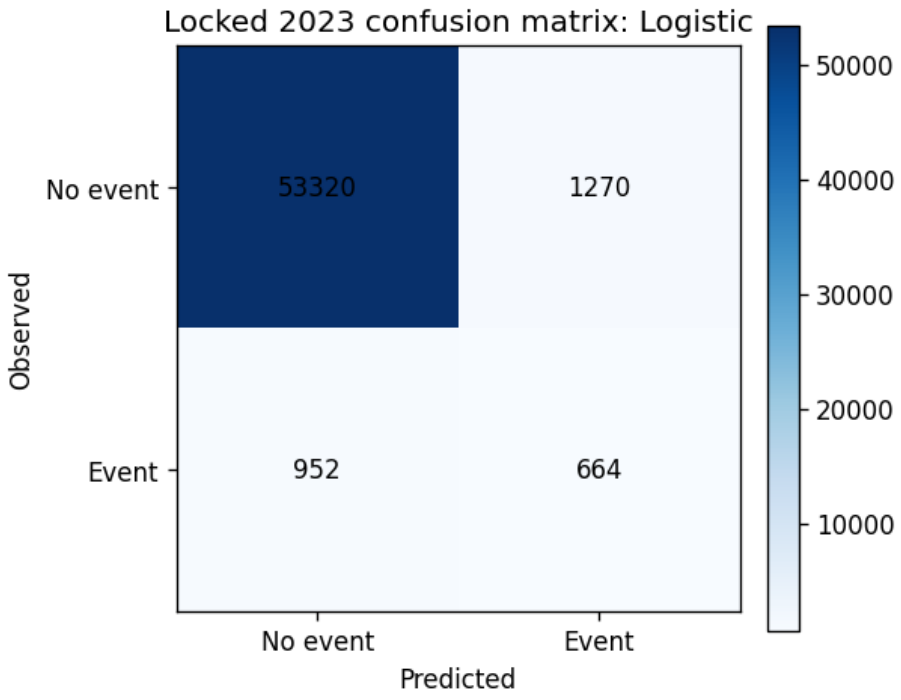


Figure 13. Confusion matrix for the exploratory logistic model. At the plotted threshold there were 53 320 true negatives, 1 270 false positives, 952 false negatives, and 664 true positives. Thresholded counts are diagnostic only.

This figure translates the exploratory logistic forecast into thresholded classifications and reveals the expected trade-off between false alarms and missed events under severe class imbalance. The large number of true negatives is not, by itself, evidence of useful forecasting performance because non-events dominate the data set. For this reason, the confusion matrix is retained only as a diagnostic and the study's principal conclusions are based on threshold-free probabilistic metrics.

The decision-curve analysis of the exploratory logistic forecast is shown in Fig. 14.

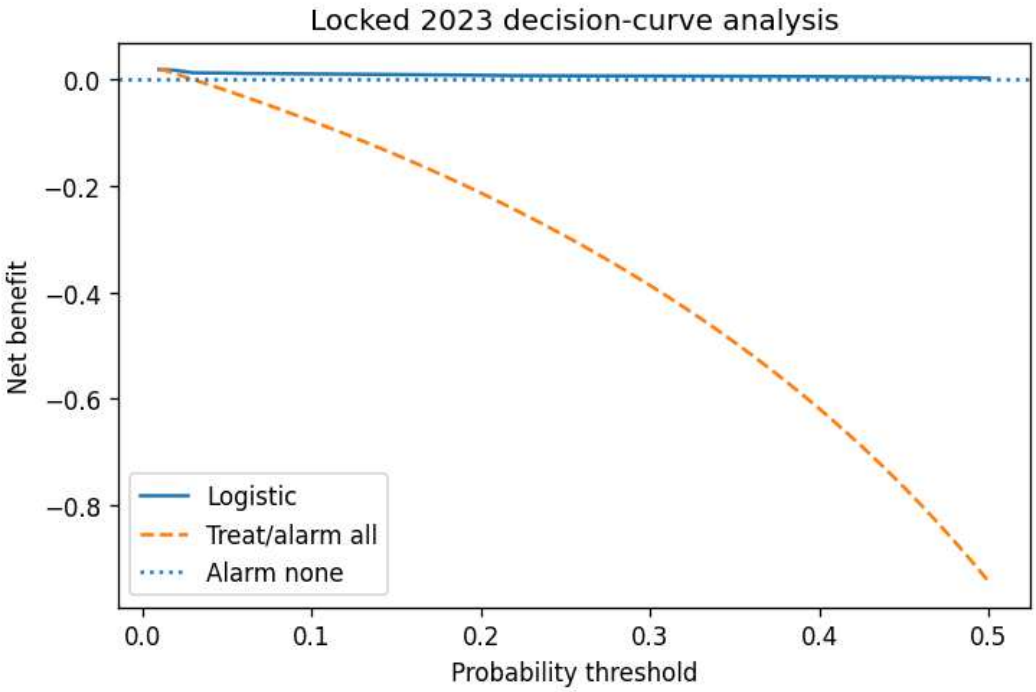


Figure 14. Decision-curve analysis for the exploratory logistic forecast. Net benefit depends on user-specific risk thresholds and was not used for model selection.

This figure evaluates the potential decision value of the exploratory forecast over a range of probability thresholds. Net benefit varies with the decision threshold, showing that a forecast can only be operationally useful when linked to a specific cost–loss or preparedness context. Because no single stakeholder utility function was prespecified, decision-curve results are not used to select the final model and are interpreted as supporting evidence only.

The Molchan-style error trade-off for the exploratory logistic forecast is shown in Fig. 15.

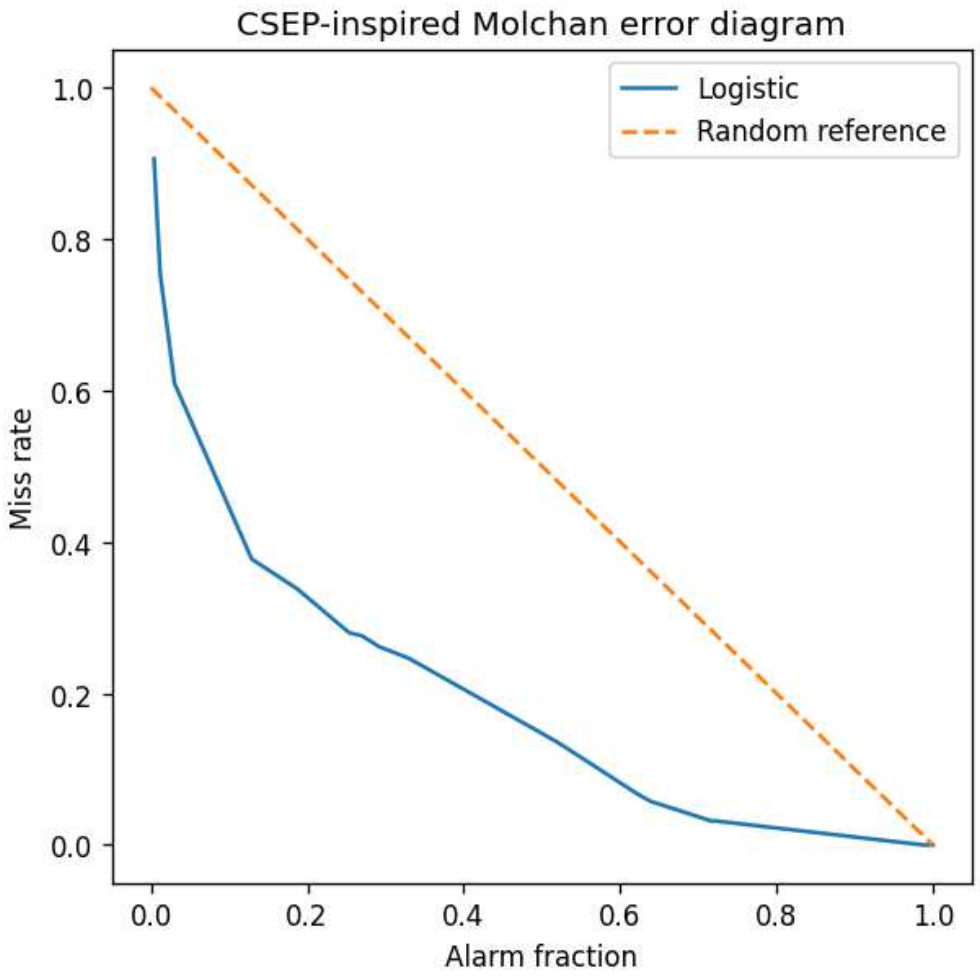


Figure 15. Molchan error diagram for the exploratory logistic model. The diagram summarizes the trade-off between alarm space/time and missed targets relative to a random strategy.

This figure summarizes the balance between the fraction of space–time placed under alarm and the fraction of target events missed. Performance below the random-strategy reference indicates that the model concentrates a portion of future events into a smaller alarm fraction, but this does not establish deterministic predictability. The Molchan diagnostic therefore complements, rather than replaces, calibrated probabilistic evaluation.

Primary Non-Overlapping Forecast Performance

Under formal non-overlapping seven-day evaluation, History+ETAS Logistic was selected exclusively from validation data (validation PR-AUC=0.0405). In the locked 2023 test it achieved ROC-AUC=0.8510, PR-

AUC=0.3479, Brier=0.02408, log loss=0.11921, and ECE=0.01842. The Bayesian historical-rate reference achieved PR-AUC=0.0388 and Brier=0.02906. In the 2024–2026 forward holdout, History+ETAS achieved ROC-AUC=0.7086, PR-AUC=0.03261, Brier=0.005637, log loss=0.03285, and ECE=0.001962; the historical-rate reference yielded PR-AUC=0.01127 and Brier=0.005674.

Absolute PR-AUC was much larger in 2023 because the event prevalence increased to 2.92% during the Kahramanmaraş sequence, compared with 0.571% in the forward holdout. Consequently, cross-period PR-AUC values should not be compared as if prevalence were constant. The key evidence is paired skill against the same-period reference and its block-bootstrap uncertainty.

The primary non-overlapping seven-day forecast performance of all benchmark models is summarized in Table 6.

Table 6. Non-overlapping seven-day forecast performance.

Model	Period	ROC-AUC	PR-AUC	Brier	Log loss	ECE
History+ETAS Logistic	2023 test	0.8510	0.3479	0.02408	0.11921	0.01842
History Logistic	2023 test	0.8190	0.3547	0.02379	0.11889	0.01695
Full Logistic	2023 test	0.8474	0.3353	0.02436	0.11999	0.01880
XGBoost	2023 test	0.8227	0.3088	0.02480	0.11709	0.01884
Bayesian historical rate	2023 test	0.6219	0.0388	0.02906	0.18825	0.02672
History+ETAS Logistic	2024–2026 holdout	0.7086	0.03261	0.005637	0.03285	0.001962
History Logistic	2024–2026 holdout	0.7082	0.03129	0.005629	0.03256	0.001789
Full Logistic	2024–2026 holdout	0.7709	0.03011	0.005645	0.03300	0.002015
XGBoost	2024–2026 holdout	0.7411	0.02338	0.005656	0.03332	0.002274
Bayesian historical rate	2024–2026 holdout	0.7156	0.01127	0.005674	0.03583	0.003208

This table contains the main out-of-sample performance results and shows that both History Logistic and History+ETAS substantially outperform the Bayesian historical-rate reference in PR-AUC during 2023 and the forward holdout. The small differences between the two history-based logistic models are important: History Logistic is slightly stronger in 2023, whereas History+ETAS is marginally stronger in the holdout. This pattern justifies retaining the validation-selected History+ETAS forecast while explicitly rejecting a separate claim that ETAS features add stable independent skill.

The validation-selected forecast was selected on validation only. History Logistic slightly exceeded History+ETAS in 2023 PR-AUC, while History+ETAS slightly exceeded it in the forward holdout; the incremental ETAS claim did not pass paired uncertainty gates.

Paired Moving-Block Skill Against the Historical Reference

The primary forecast-skill claim passed because the validation-selected History+ETAS model outperformed the Bayesian historical-rate reference with confidence intervals entirely above zero for both PR-AUC difference and Brier skill in both formal evaluation periods. In 2023, median ΔPR-AUC was 0.314 (95% CI 0.157–0.443), Brier skill 0.00503 (0.000692–0.01177), and log-loss skill 0.0698 (0.0223–0.1449). In the forward holdout, median ΔPR-AUC was 0.0201 (0.00888–0.03844), Brier skill 3.3×10^{-5} (1.8×10^{-5}–5.0×10^{-5}), and log-loss skill 0.00280 (0.00134–0.00446). The smaller holdout gains emphasize that statistically detectable ranking improvement does not imply a large absolute change in calibrated rare-event probability.

The paired 28-day moving-block bootstrap skill relative to the Bayesian historical-rate reference is summarized in Table 7.

Table 7. Moving-block skill against the historical-rate reference.

Period	Metric	Median	95% CI low	95% CI high
2023	ΔPR-AUC	0.3143	0.1565	0.4434
2023	Brier skill	0.005025	0.000692	0.011768
2023	Log-loss skill	0.06977	0.02231	0.14491
2024–2026	ΔPR-AUC	0.02006	0.008884	0.03844
2024–2026	Brier skill	0.000033	0.000018	0.000050
2024–2026	Log-loss skill	0.002800	0.001339	0.004463

This table provides the strongest inferential support for the primary claim because uncertainty is calculated on paired predictions using moving temporal blocks. The 95% confidence intervals for ΔPR-AUC, Brier skill, and log-loss skill remain above zero in both 2023 and 2024–2026, showing that improvement over the static historical-rate reference persists outside the exceptional 2023 sequence. The much smaller holdout effect sizes also demonstrate that statistical detectability should not be confused with a large absolute change in rare-event probability.

Rolling-Origin and Sequence Generalization

Eight rolling-origin years (2018–2025) were evaluated without tuning on the target year. History Logistic had the highest median PR-AUC (0.0364), followed closely by History+ETAS (0.0345) and XGBoost (0.0340). This ranking reinforces the central complexity-audit result: the dynamic signal is real, but a simpler history model is

at least as competitive as more elaborate formulations across years. Worst-year PR-AUC remained low for every model, indicating substantial temporal heterogeneity.

The 2018–2025 rolling-origin performance across candidate models is summarized in Table 8.

Table 8. Rolling-origin performance, 2018–2025.

Model	Years	Median PR-AUC	IQR PR-AUC	Median Brier	Worst PR-AUC
History Logistic	8	0.03638	0.07013	0.003263	0.008024
History+ETAS Logistic	8	0.03453	0.08018	0.003266	0.007482
XGBoost	8	0.03396	0.04146	0.003433	0.007049
Full Logistic	8	0.02664	0.06710	0.003271	0.004961
HistGradientBoosting	8	0.01159	0.02434	0.003339	0.007220
Bayesian historical rate	8	0.009491	0.008390	0.003254	0.004679

This table evaluates temporal generalization across eight rolling-origin years and shows that the simpler History Logistic model has the highest median PR-AUC, followed closely by History+ETAS and XGBoost. No model performs uniformly well in every year, as reflected by low worst-year PR-AUC values and wide interquartile variation. This result strengthens the paper's parsimony argument: dynamic seismic history is useful, but added model complexity does not provide consistently superior year-to-year performance.

Leave-one-mainshock-sequence-out evaluation further demonstrated heterogeneity. The History+ETAS model strongly outperformed the historical rate in PR-AUC around the 2023 Mw7.8 sequence (0.673 vs 0.147) and the 2020 Mw6.7 sequence (0.525 vs 0.0245), but Brier skill was slightly negative for the 2024 Mw6.0 and June 2020 Mw5.9 sequence windows. Thus, sequence ranking skill was not equivalent to uniformly improved probability calibration.

Selected leave-one-mainshock-sequence-out results are summarized in Table 9.

Table 9. Leave-one-mainshock-sequence-out results.

Mainshock date	Mw	Positives	Model PR	Baseline PR	ΔPR	Brier skill
2023-02-06	7.8	87	0.6728	0.1474	+0.5253	+0.04335
2020-01-24	6.7	4	0.5250	0.0245	+0.5005	+0.00223
2003-05-01	6.4	11	0.1641	0.0175	+0.1466	+0.00010
2023-02-20	6.3	52	0.6001	0.0830	+0.5171	+0.02631
2010-03-08	6.1	5	0.2518	0.0449	+0.2069	+0.00017
2024-10-16	6.0	6	0.0477	0.0162	+0.0315	−0.00002
2020-06-14	5.9	1	0.0833	0.0096	+0.0737	−0.00265

This table shows that forecast behavior varies substantially across individual mainshock sequences. Large positive PR-AUC gains around the 2023 Mw7.8 and 2020 Mw6.7 sequences indicate strong ranking

capability during highly clustered activity, whereas slightly negative Brier skill in some smaller sequences shows that ranking improvement does not guarantee better probability calibration. The sequence-wise analysis therefore supports general signal detection while cautioning against assuming uniform performance across all earthquake regimes.

Spatial and Fault-Level Externalization

Spatial-block analysis produced positive PR-AUC differences over the historical reference in all five blocks, but the magnitude of gain varied markedly (approximately 0.0005–0.346). One block had slightly negative Brier skill (-1.3×10^{-5}), again separating ranking improvement from probability improvement. Leave-one-fault-out summaries were similarly heterogeneous, especially for traces with very few positives. These results support regional forecast skill but argue against presenting a single uniform "fault accuracy."

The five-block spatial externalization results are summarized in Table 10.

Table 10. Five-block spatial externalization.

Block	N	Positives	Model PR	Baseline PR	ΔPR	Brier skill
0	7 480	105	0.2166	0.0140	+0.2026	+0.001477
1	8 976	32	0.0823	0.00357	+0.0788	+0.000014
2	5 049	26	0.0119	0.00515	+0.00671	+0.000008
3	4 488	181	0.3865	0.0403	+0.3462	+0.007946
4	3 366	14	0.00469	0.00416	+0.000535	−0.000013

This table assesses spatial externalization and shows positive PR-AUC differences relative to the historical reference in all five blocks, but with large variation in effect magnitude. The weakest block approaches the baseline and shows slightly negative Brier skill, indicating that some parts of the study region remain difficult to calibrate even when ranking is improved. This spatial heterogeneity is why the manuscript avoids reporting a single universal 'fault accuracy' and instead presents block- and trace-level uncertainty.

Feature Contribution and Explainability

Ablation in the exploratory daily workflow showed that the history-only model had the strongest holdout PR-AUC (0.0323), while adding fault-graph, physics, and ETAS predictors reduced holdout PR-AUC to approximately 0.0234–0.0235 despite improving ROC-AUC. Permutation importance and SHAP diagnostics

likewise assigned the greatest importance to long-window seismic-history features, especially 90-day scalar moment and event counts. This pattern is consistent with the strict claim gates: additional physics/graph and ETAS feature families did not establish independent incremental forecast value.

The permutation-importance ranking of the exploratory logistic predictors is shown in Fig. 16.

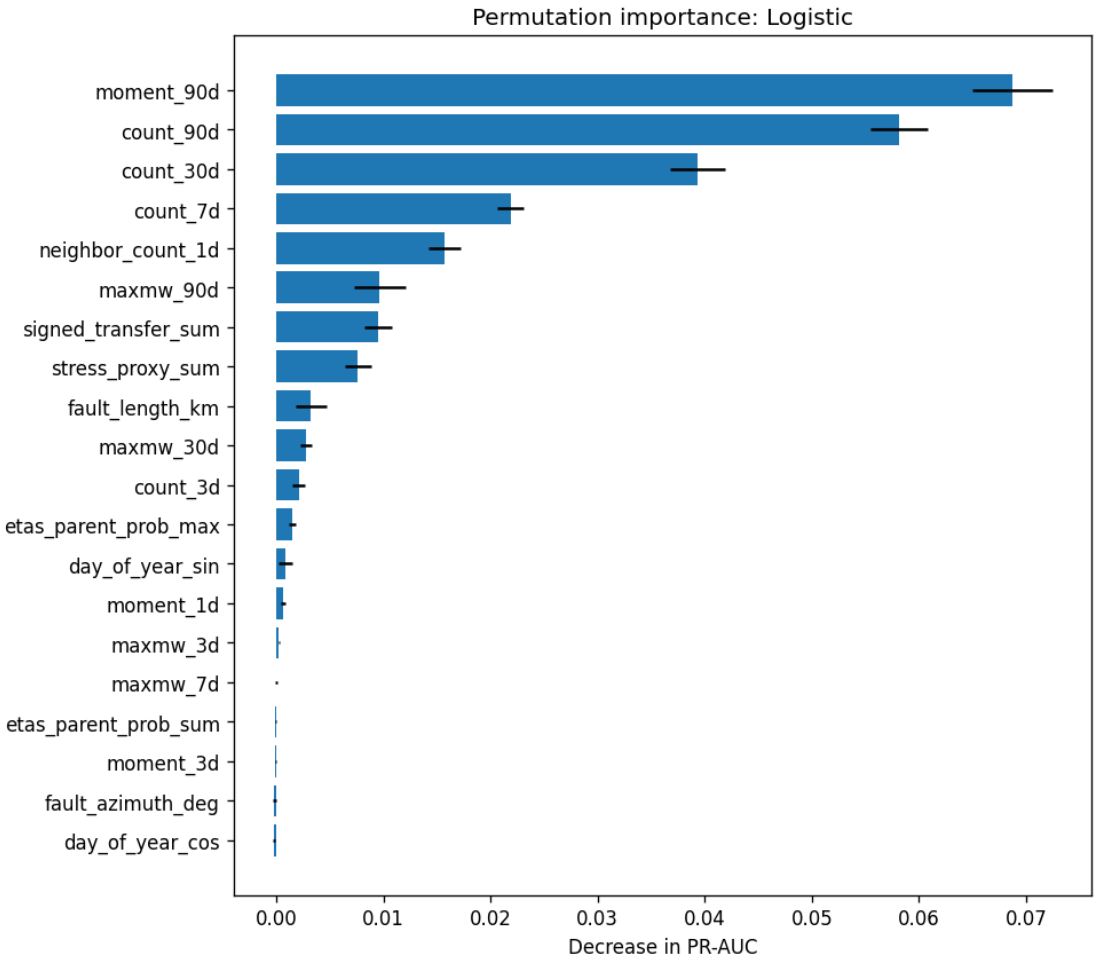


Figure 16. Permutation importance for the exploratory logistic model. Long-window scalar moment and event-count history dominate, with smaller contributions from graph, ETAS, and stress-proxy variables.

This figure shows that long-window seismic-history variables, especially accumulated scalar moment and event counts, dominate permutation importance in the exploratory logistic analysis. The comparatively smaller influence of graph, ETAS, and stress-proxy variables is consistent with the formal incremental-claim tests, in which these additional feature families did not demonstrate reproducible future-period

improvement. Thus, the explainability analysis supports the study's central parsimony result rather than introducing a separate mechanistic claim.

The SHAP-based feature-impact diagnostic is shown in Fig. 17.

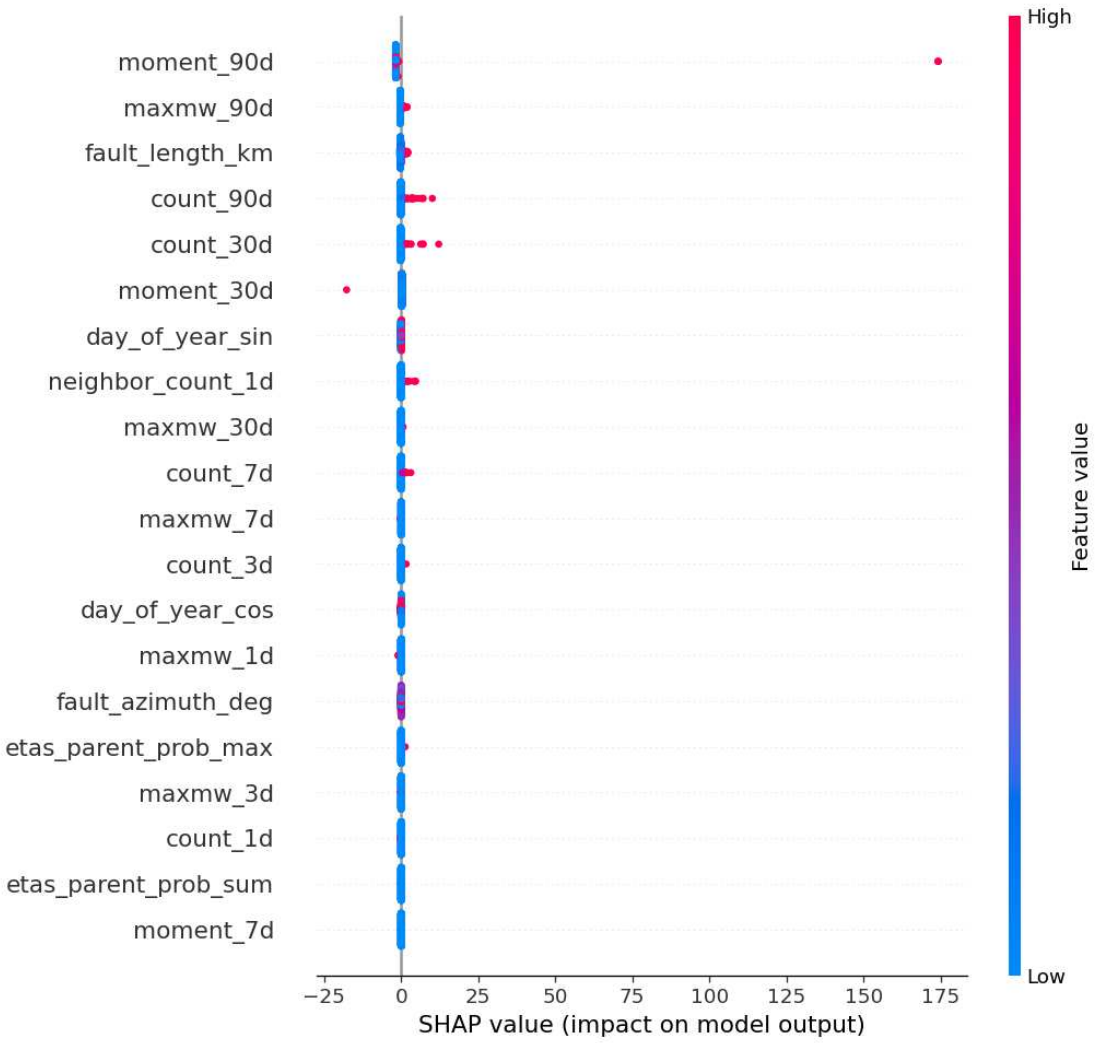


Figure 17. SHAP impact diagnostic. The logged plotting warning occurred after SHAP values were computed and did not alter model parameters, predictions, or numerical SHAP values. The figure is qualitative only; SHAP follows Lundberg and Lee [20].

This figure provides a second, model-specific view of feature influence using SHAP values and again emphasizes long-term seismic-history information. The qualitative agreement between SHAP and permutation importance strengthens confidence that the apparent importance ranking is not an artifact of a single explainability method. Nevertheless, SHAP values describe associations within the fitted model and are not interpreted as causal earthquake-triggering effects.

Complexity Audit and Incremental-Claim Tests

Claim-specific moving-block comparisons were intentionally stricter than the primary comparison against the static historical-rate reference. None of the four incremental hypotheses passed simultaneously in the locked test and forward holdout: ETAS did not establish a statistically reliable increment over History Logistic; Full Logistic did not establish a physics/graph increment over History+ETAS; XGBoost did not establish a boosting increment; and CAL-SEQ did not establish a deep-learning increment. Therefore, the manuscript makes no claim that ETAS, graph/physics features, boosting, or deep learning are superior components, even though the validation-selected operational model contains ETAS predictors.

The prespecified incremental-claim audit and the evidential status of each model component are summarized in Table 11.

Table 11. Prespecified incremental-claim audit.

Claim	Status	Evidence rule
Primary dynamic forecast vs Bayesian historical rate	Supported	History+ETAS Logistic selected on validation; paired ΔPR and Brier skill >0 in test and holdout
ETAS increment over seismic history	Not supported	History+ETAS vs History failed joint paired gates
Physics/graph increment	Not supported	Full Logistic vs History+ETAS failed joint paired gates
Boosting increment	Not supported	XGBoost vs History+ETAS failed joint paired gates
Deep increment	Not supported	CAL-SEQ vs History+ETAS failed joint paired gates
Named-segment interpretation	Not supported	0/157 candidate traces carried usable names in executed GEM subset
Coulomb ΔCFS interpretation	Not supported	Only scalar-moment/distance/orientation proxies were used
External NAF/WAF transfer	Not tested	Candidate geometries were identified but external system runs were optional and not executed

This table is the manuscript's central guard against overinterpretation because each scientific extension must satisfy its own evidence rule. Only the primary dynamic forecast versus the Bayesian historical-rate reference is supported; ETAS, physics/graph, boosting, deep-learning, named-segment, Coulomb, and external-transfer claims are either unsupported or untested. This explicit audit ensures that the conclusions follow the demonstrated evidence rather than the complexity of the modeling pipeline.

CAL-SEQ, Conditional Magnitude, and Conformal Uncertainty

CAL-SEQ early-stopped at epoch 11. After temperature calibration, its non-overlapping PR-AUC was 0.0395 in 2023 and 0.0133 in the holdout, both far below the validation-selected classical model. Its conditional

maximum-magnitude head was not scientifically usable: R^2 was −6.67 in validation, −6.37 in 2023, and −10.09 in the holdout, with MAE values above 1.0 magnitude unit. The magnitude claim was therefore disabled. The negative result is retained rather than omitted because it directly tests whether increased representation capacity improved the stated task.

Mondrian/class-conditional conformal sets targeted 90% coverage but achieved 74.1% in the 2023 test and 86.9% in the forward holdout. Class-1 coverage was high (98.7% and 94.2%), whereas class-0 coverage was 73.3% and 86.9%. The conformal-reliability claim was therefore disabled. These results highlight the effect of distribution shift and very low event prevalence on post-hoc uncertainty sets.

The CAL-SEQ validation learning trajectory and early-stopping behavior are shown in Fig. 18.

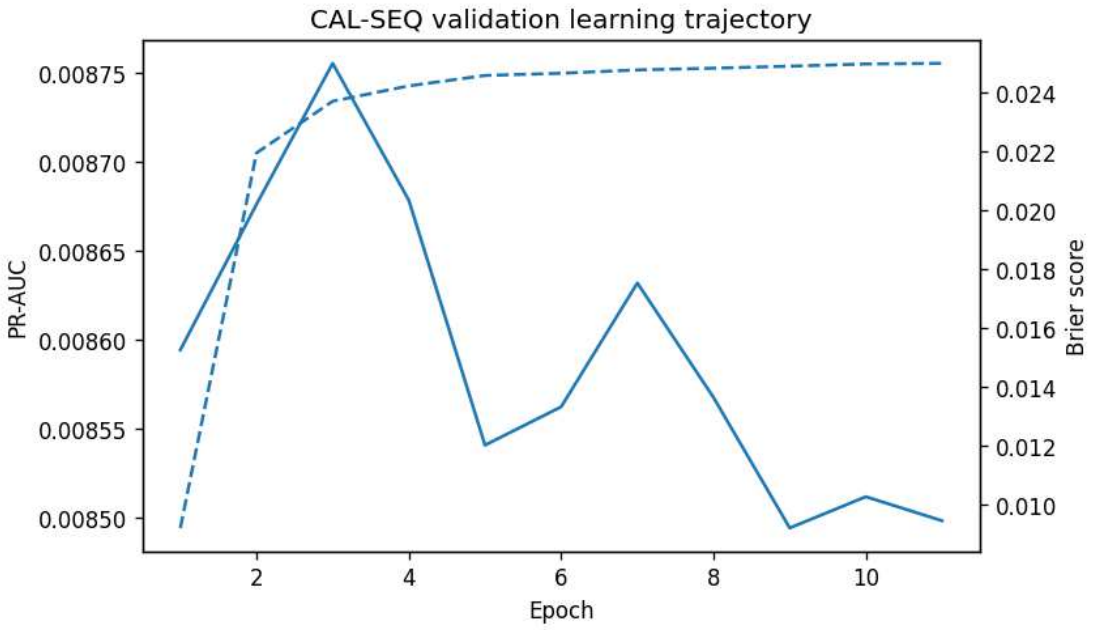


Figure 18. CAL-SEQ validation learning trajectory. Early stopping occurred at epoch 11; additional training complexity did not yield competitive non-overlapping PR-AUC.

This figure shows that CAL-SEQ reached its best validation behavior early and was stopped before unnecessary additional training. Despite the more expressive architecture, its formal non-overlapping test and holdout PR-AUC values remained below those of the simpler logistic reference. The learning trajectory therefore reinforces a central result of the study: additional representation capacity does not warrant adoption unless it produces reproducible out-of-sample skill.

The conditional maximum-magnitude regression diagnostic is shown in Fig. 19.

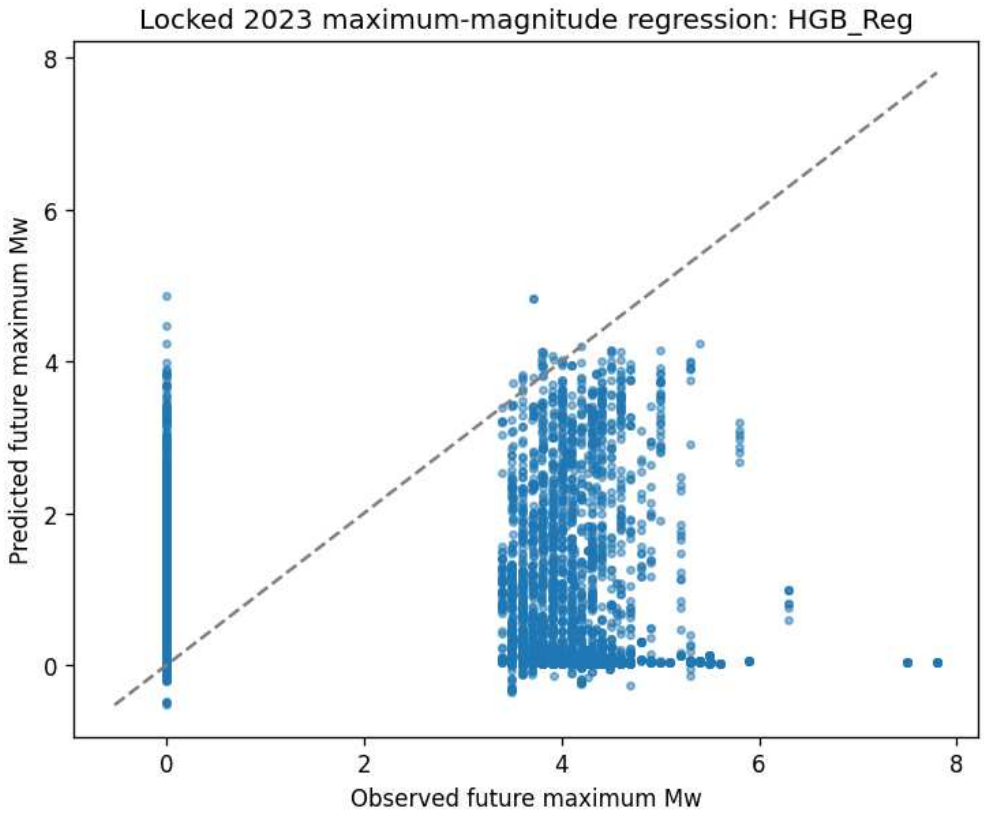


Figure 19. Conditional maximum-magnitude regression diagnostic. The broad scatter and negative conditional R^2 values led to disabling the magnitude-forecast claim.

This figure demonstrates why the auxiliary maximum-magnitude prediction claim was rejected. Predictions show broad dispersion around the observed values and the conditional R^2 values were negative in validation, test, and holdout periods, indicating worse performance than a simple mean-based reference for this task. Retaining this negative diagnostic prevents selective reporting and clearly separates the successful binary probability component from an unsupported magnitude-forecast extension.

Operational Probability Surface and Scenario Interpretation

The executable pipeline exported the validation-selected History+ETAS model as the primary operational artifact and generated a fault-trace probability map for the latest observable holdout origin. A counterfactual Mw6.0 scenario on 29 July 2026 produced top baseline probabilities of approximately 0.7–1.9% across the highest-ranked traces. Because the validation-selected forecast does not use the stress-proxy feature family,

counterfactual model probabilities were unchanged by the proxy perturbation. The scenario is therefore correctly interpreted as a ranking demonstration, not as causal Coulomb transfer or deterministic earthquake prediction.

The seven-day fault-trace probability surface at the latest observable holdout origin is shown in Fig. 20.

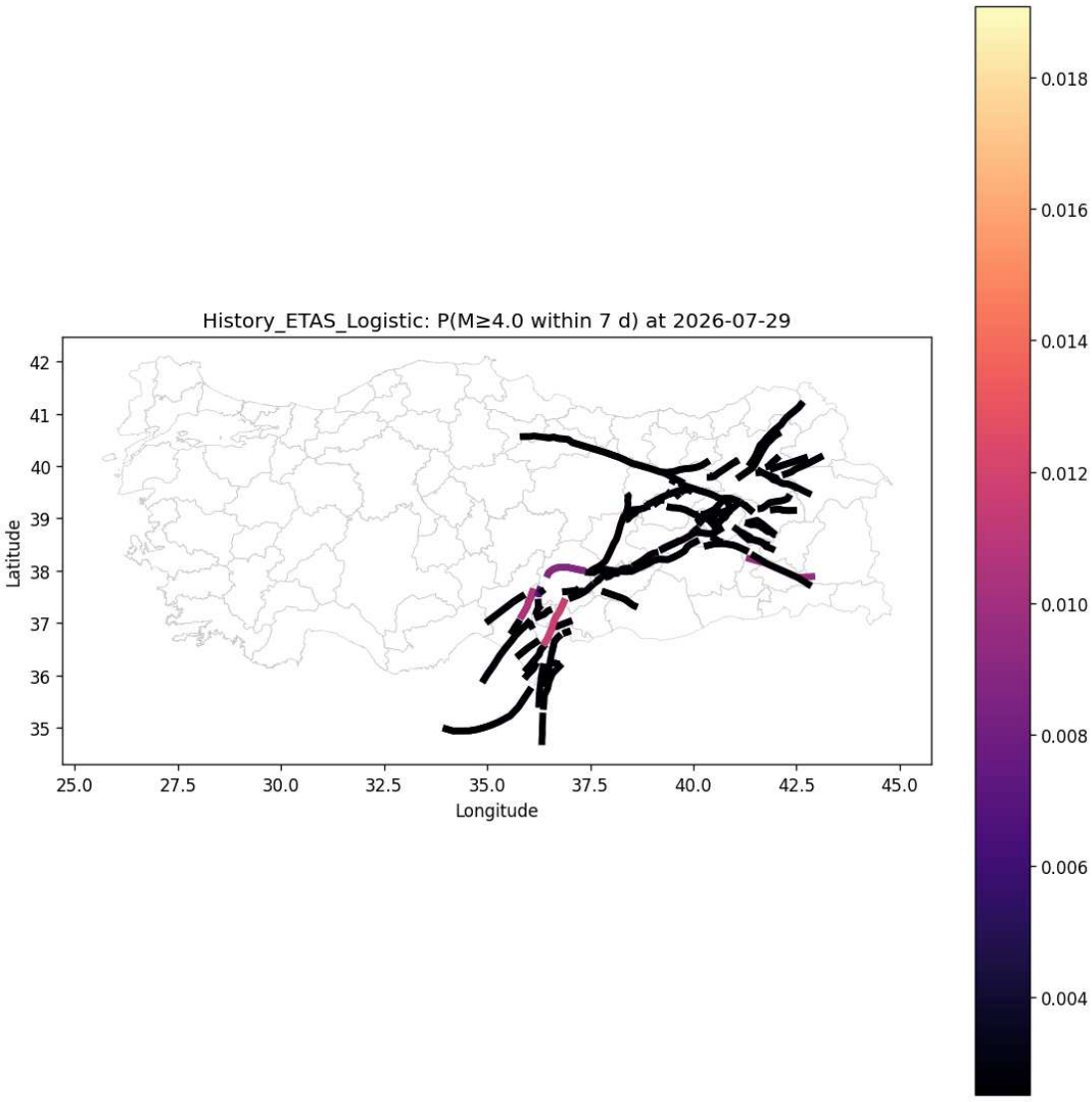


Figure 20. Seven-day fault-trace probability map. Probabilities at the latest observable holdout origin are from the validation-selected History+ETAS logistic forecast. They are not named-segment hazard values or deterministic earthquake predictions.

This figure illustrates how the validation-selected model can be translated into a spatial probability surface for a specific seven-day forecast origin. The highest probabilities remain on the order of only a few percent,

which is appropriate for a rare-event problem and should be communicated as a relative prioritization layer rather than an earthquake alarm. The map is therefore most relevant as a potential input to broader exposure, infrastructure, and emergency-planning workflows, not as a deterministic statement about which fault will rupture next.

Discussion

Interpretation of the Primary Result

Across the 50-year EAFS analysis, the principal predictive result is deliberately narrow: recent seismic history contains reproducible short-term information beyond a static fault-specific occurrence rate. The advantage was visible during the extraordinary 2023 sequence and remained positive in the substantially lower-prevalence 2024–2026 period. This persistence matters more than the absolute 2023 PR-AUC because it shows that the result is not confined to a single aftershock-rich year. Positive moving-block confidence intervals for PR-AUC difference, Brier skill, and logarithmic-score skill further indicate that the result is not an artifact of independently resampling strongly autocorrelated daily forecast origins.

This conclusion is well aligned with modern OEF practice, where the goal is not deterministic prediction but transparent, testable probability updates based on current seismicity [5]. It also complements CSEP's prospective benchmarking philosophy [9, 10]: the appropriate question is whether an issued probability distribution improves on a prespecified reference under future-data isolation.

Why Model Complexity Did Not Add Reproducible Skill

The most important complexity result is the absence of evidence that additional modeling layers earned an independent claim. History Logistic slightly outperformed History+ETAS in 2023 PR-AUC and in the median rolling-origin summary, whereas History+ETAS was marginally better in the forward holdout and had the highest validation PR-AUC. The sequence-level Omori fits simultaneously show pronounced variation in K, c, and p. These observations together support a constrained interpretation: a fixed pre-test ETAS-style parametrization did not provide stable incremental skill beyond lagged history. They do not constitute a general

rejection of adaptive ETAS models, and they suggest that sequence heterogeneity may partly explain the weak incremental result.

Similarly, the Full Logistic model achieved higher holdout ROC-AUC (0.771) than the history models but lower PR-AUC and no demonstrated paired physics/graph gain. In a rare-event application, this distinction matters: improved global ranking across abundant negatives can coexist with poorer prioritization of the scarce positive class. The complex CAL-SEQ model also failed to outperform the classical reference despite an architecture explicitly designed to learn bounded residual corrections to a historical baseline. This finding argues for parsimony until a complex component wins a genuinely delayed prospective experiment.

Relation to EAFS Seismology and Environmental-Risk Resilience

The present data product serves a different purpose from recent high-resolution EAF catalogs. Ding et al. [1] resolved 29,519 events during February 2023 to study rupture-zone structure and early aftershock migration, while Colavitti et al. [3] released 9,442 carefully processed events from 2019 through February 2024 with strong-motion products. Those resources provide substantially higher event-location fidelity in their target windows. In contrast, the present framework sacrifices some local precision to build a longitudinal forecast benchmark spanning network eras, enforce completeness control, and test generalization before and after the 2023 sequence. These approaches are complementary rather than competing.

Palo and Zollo [2] demonstrated that the 2023 rupture contained small-scale segmentation and complex rupture-speed behavior. That evidence reinforces why the present graph and stress proxies should not be overinterpreted. Geometry proximity and $M0/r^3$ terms cannot substitute for finite-fault slip, receiver-fault orientation, or elastic stress-transfer calculations. Future physical extensions should therefore ingest published finite-slip models and authoritative named receiver faults before asking whether ΔCFS adds prospective forecast skill.

For environmental-risk management, the practical implication is not an earthquake alarm but a reproducible probability layer that can be combined with exposure, infrastructure, land-use, and emergency-planning

information. Such integration is relevant to resilient-community planning because the forecast retains uncertainty explicitly and avoids converting model ranking into a deterministic hazard statement.

Methodological Strengths

Several design choices directly address common reasons for inflated machine-learning results in earthquake forecasting. First, magnitude harmonization and ETAS parameter estimation were restricted to pre-test data. Second, dynamic features were shifted so the forecast origin did not contain future target information. Third, split boundaries were purged by the forecast horizon. Fourth, formal origins did not overlap. Fifth, uncertainty used moving temporal blocks rather than independent row bootstrap. Sixth, the extreme 2023 sequence was complemented by a 2024–2026 forward holdout, eight rolling years, major-sequence holdouts, spatial blocks, and fault holdouts. Finally, every substantive interpretation was separated into its own pass/fail gate.

The workflow also retained negative results. The magnitude head, conformal sets, and deep challenger were not removed from the manuscript merely because they failed. This is important for reproducibility and limits the "researcher degrees of freedom" that can otherwise arise when unsuccessful auxiliary tasks are silently omitted.

Limitations

- KOERI events were not retrievable through a compatible event-service endpoint in the executed client. This limitation is most likely to affect smaller-event completeness, location precision, and cross-network confirmation rather than the gross $M \geq 4$ target record, because the predictive interval begins in 2000, AFAD dominates dense-period raw counts, and $Mc = 3.4$ is imposed. Nevertheless, a future sensitivity study using an independently obtained KOERI/ISC event bulletin is warranted.
- The 157 selected GEM traces are source-traceable but unnamed in the executed subset. Results are computational trace-ID forecasts, not authoritative Palu, Pütürge, Erkenek, Pazarcık, or other named-segment forecasts. A versioned MTA or equivalent authoritative named-fault layer is required before named-segment interpretation.
- The static-stress variables are $M0/r^3$ and orientation proxies, not Coulomb ΔCFS. No physical stress-transfer conclusion is drawn from them.

- The study is delayed pseudo-prospective rather than a live prospective deployment. The final operational model should be frozen and evaluated on genuinely future data through a CSEP-compatible workflow.
- The 2023 period is a major distribution shift: target prevalence was approximately five times the forward-holdout prevalence. Cross-period PR-AUC values therefore cannot be compared without same-period references, motivating the paired skill analysis.
- External North Anatolian and West Anatolian candidate fault geometries were identified, but independent frozen-model transfer tests were not executed. Cross-fault-system transfer is not claimed.
- Nearest-trace assignment within 40 km is a computational association rule and not proof that an earthquake ruptured the assigned trace, particularly in geometrically complex multi-fault sequences.
- The common Mc=3.4 improves longitudinal comparability but discards smaller earthquakes that may contain useful short-term information in modern dense-network periods; this is an explicit comparability-versus-resolution trade-off.
- Sequence-specific Omori parameters vary strongly, whereas the ETAS-style forecast features are intentionally frozen from pre-2020 behavior. The negative ETAS increment is therefore specific to this leakage-safe fixed parametrization and should not be generalized to sequence-adaptive ETAS formulations.
- Case-control fitting uses a prespecified 35:1 negative cap with inverse sampling weights and full-prevalence validation calibration. The out-of-sample calibration scores are favorable, but the executed manuscript does not contain a dedicated 20:1/full-data sampling-ratio sensitivity experiment; the ratio should therefore be regarded as a computational design choice rather than a uniquely optimized value.
- The AFAD ML and USGS mb magnitude conversions are supported by 41 and 99 pre-test paired events, respectively. They are restricted to observed input ranges and are not interpreted as universal conversion laws; remaining conversion uncertainty can propagate into near-threshold history features.

Prospective Extensions and External Validation

The next scientific step is not to add another unconstrained neural architecture. A stronger prospective extension should freeze the present benchmark; replace unnamed GEM traces with a curated, versioned named MTA fault layer; assimilate authoritative finite-fault slip products for true Coulomb calculations; execute NAF/WAF

transfer without refitting; and publish rolling forecasts through a timestamped CSEP/floatCSEP-compatible archive. Such a design would allow future work to test whether physical stress transfer or transferable graph representations add incremental skill beyond the strong history baseline rather than merely improving retrospective fit.

Conclusions

This study provides a 50-year seismicity analysis of the East Anatolian Fault System from 1976 to 2026 and translates that record into a reproducible probabilistic forecasting benchmark. The long-term catalog documents changing instrumental coverage and major-event activity across five decades, while the dense 2000–2026 catalog supports year-specific completeness analysis and construction of 1,509,869 daily observations across 157 source-traceable fault geometries. By separating descriptive long-term seismicity from the denser feature-generating period, the framework preserves the historical scope of the EAFS analysis without treating changing network sensitivity as stationary information.

Within that 50-year framework, the strongest predictive finding is that calibrated recent seismic history improves seven-day fault-trace-associated M≥4 probability estimates relative to a static Bayesian fault-rate reference. The validation-selected History+ETAS logistic model showed positive paired skill in both the locked 2023 test and the 2024–2026 forward holdout. However, the evidence did not support attributing that gain specifically to the fixed ETAS-style parametrization, fault-graph/stress proxies, boosted trees, or deep learning. History-only logistic regression remained highly competitive, and the auxiliary CAL-SEQ magnitude and conformal claims failed their prespecified quality gates.

The main scientific contribution is therefore twofold. First, the study establishes a source-auditable half-century EAFS data and evaluation framework that connects catalog completeness, sequence heterogeneity, fault geometry, and strict temporal validation. Second, it demonstrates that model complexity should be accepted only when it adds reproducible future-period skill beyond strong historical baselines. For the present EAFS benchmark, recent seismic history is a defensible operational reference, whereas mechanistic, graph-based, and deep-learning extensions remain hypotheses for frozen prospective testing rather than established improvements.

This interpretation is consistent with the scope stated in the title: a 50-year analysis of the East Anatolian Fault System whose forecasting conclusions are constrained by the evidence actually retained across time.

List of Abbreviations

AFAD: Disaster and Emergency Management Authority of Türkiye; CAL-SEQ: calibrated sequence-learning challenger; CI: confidence interval; CSEP: Collaboratory for the Study of Earthquake Predictability; EAFS: East Anatolian Fault System; ECE: expected calibration error; ETAS: epidemic-type aftershock sequence; GAF-DB: Global Active Faults Database; KOERI: Kandilli Observatory and Earthquake Research Institute; Mc: magnitude of completeness; OEF: operational earthquake forecasting; PR-AUC: area under the precision–recall curve; ROC-AUC: area under the receiver operating characteristic curve; USGS: United States Geological Survey.

Additional files

Additional file 1

Format: DOCX. Title: Supplementary material. Description: source-traceable fault records, sequence-specific Omori–Utsu parameters, scientific-integrity checks, and the KOERI service-audit note.

Additional file 2

Format: ZIP. Title: Reproducibility package. Description: the executable PAPER_FULL notebook and SHA-256 integrity manifest used to reproduce the reported workflow.

Declarations

Ethics Approval and Consent to Participate

Not applicable. The study uses publicly available geophysical event catalogs and fault-geometry data and does not involve human participants, human data, animals, or biological material.

Consent for publication

Not applicable.

Availability of Data and Materials

The source earthquake catalogs are publicly accessible through the USGS Earthquake Catalog web service [11] and the AFAD Earthquake Department data services [12]. Active-fault geometries were obtained from the GEM Global Active Faults Database [13]. The executable notebook and reproducibility manifest are supplied with this submission as Additional file 2. Derived tables and figures reported in the manuscript are reproducible from the cited public sources and the supplied workflow. KOERI event retrieval was unavailable through the executed event-service route and is documented explicitly in Additional file 1.

Competing Interests

The author declares that he has no competing interests.

Funding

No external funding was received for this study.

Authors' contributions

MTA conceived the study, designed the methodology, developed and reviewed the software, curated the data, performed the analyses and validation, interpreted the results, prepared the visualizations, and wrote and revised the manuscript. The author read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Ding H, Zhou Y, Ge Z, Taymaz T, Ghosh A, Xu H, et al. High-resolution seismicity imaging and early aftershock migration of the 2023 Kahramanmaraş (SE Türkiye) Mw7.9 & 7.8 earthquake doublet. Earthquake Sci. 2023;36(6):417–432. doi:10.1016/j.eqs.2023.06.002.
2. Palo M, Zollo A. Small-scale segmented fault rupture along the East Anatolian fault during the 2023 Kahramanmaraş earthquake. Commun Earth Environ. 2024;5:431. doi:10.1038/s43247-024-01597-z.
3. Colavitti L, Bindi D, Tarchini G, Scafidi D, Picozzi M, Spallarossa D. A high-quality data set for seismological studies in the East Anatolian Fault Zone, Türkiye. Earth Syst Sci Data. 2025;17:3089–3108. doi:10.5194/essd-17-3089-2025.
4. Jordan TH, Chen YT, Gasparini P, Madariaga R, Main I, Marzocchi W, et al. Operational earthquake forecasting: state of knowledge and guidelines for utilization. Ann Geophys. 2011;54(4). doi:10.4401/ag-5350.
5. Mizrahi L, Dallo I, van der Elst NJ, Christophersen A, Spassiani I, Werner MJ, et al. Developing, testing, and communicating earthquake forecasts: current practices and future directions. Rev Geophys. 2024;62:e2023RG000823. doi:10.1029/2023RG000823.
6. Ogata Y. Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc. 1988;83(401):9–27.
7. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.
8. Serafini F, Naylor M, Lindgren F, Werner MJ, Main I. Ranking earthquake forecasts using proper scoring rules: binary events in a low probability environment. Geophys J Int. 2022;230(2):1419–1440. doi:10.1093/gji/ggac124.

9. Serafini F, Bayona JA, Silva F, Savran W, Stockman S, Maechling PJ, et al. A benchmark database of ten years of prospective next-day earthquake forecasts in California from the Collaboratory for the Study of Earthquake Predictability. Sci Data. 2025;12:1501. doi:10.1038/s41597-025-05766-3.

10. Iturrieta P, Savran WH, Herrmann M, Bayona JA, Gerstenberger MC, Graham K, et al. floatCSEP: an application to deploy and conduct reproducible prospective earthquake forecasting experiments. J Open Source Softw. 2026;11(118):9408. doi:10.21105/joss.09408.

11. U.S. Geological Survey. Earthquake Catalog FDSN Event Web Service. https://earthquake.usgs.gov/fdsnws/event/1/. Accessed 9 Aug 2026.

12. Disaster and Emergency Management Authority (AFAD). Earthquake Department data and event web services. https://deprem.afad.gov.tr/download. Accessed 9 Aug 2026.

13. Styron R, Pagani M. The GEM Global Active Faults Database. Earthquake Spectra. 2020;36(1 Suppl):160–180. doi:10.1177/8755293020944182.

14. Gutenberg B, Richter CF. Frequency of earthquakes in California. Bull Seismol Soc Am. 1944;34:185–188.

15. Aki K. Maximum likelihood estimate of b in the formula log N=a−bM and its confidence limits. Bull Earthq Res Inst. 1965;43:237–239.

16. Wells DL, Coppersmith KJ. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. Bull Seismol Soc Am. 1994;84(4):974–1002.

17. Utsu T, Ogata Y, Matsu'ura RS. The centenary of the Omori formula for a decay law of aftershock activity. J Phys Earth. 1995;43:1–33.

18. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1–3.

19. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.

20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.

21. International Federation of Digital Seismograph Networks. KOERI Data Centre. FDSN Data Centers Registry. Accessed 8 Aug 2026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx