Fault Geometry and Lagged GNSS in a Bounded ETAS-Neural Model for Seven-Day M5+ Occurrence in Marmara

A reproducible retrospective evaluation with sparse target events

Alderbeam, Inc.

Company research report | Version 1.3 | 12 October 2026

Company research report; not independently peer reviewed. All empirical findings concern the archived experiment specified below; no new model fitting was performed to prepare this report.

Abstract

Background. Earthquake forecasting models can use abundant small events to estimate changing seismicity rates, yet evidence that additional neural, fault-geometric or geodetic features improve moderate-earthquake occurrence probabilities remains difficult to establish. **Objective.** This study evaluates whether mapped-fault proximity and lagged Global Navigation Satellite System (GNSS) histories improve seven-day source-native M5+ occurrence forecasts in the Marmara region beyond a fitted spatial epidemic-type aftershock sequence (ETAS) reference and an existing ETAS-neural hybrid. **Methods.** Three small extensions were fitted: fault attention, GNSS temporal convolution and their combination. The frozen neural reference contains 625 parameters; the extensions add 217, 80 and 297 parameters, respectively. Training uses 31,168 scored M2+ observations before 2014, following a one-year conditioning period. Evaluation comprises 252 nonoverlapping weekly windows from 2014 to October 2018, containing five native M5+ events in three positive windows. Authentic inputs include 55 selected GEM/EMME fault traces and 17,555 daily GNSS records from ISTA and TUBI. Forecast probabilities integrate clustered future seismicity using a first-target-event survival simulation. **Results.** The combined extension improves mean weekly log score by 0.000383 and Brier loss by 0.00000556 relative to the frozen hybrid. Its descriptive 53-week moving-block interval for log-score gain is −0.000217 to 0.000990. All three extensions reduce mean log score in positive windows. The combined model improves small-event held-out log likelihood by 25.46 natural-log units, but retains 17 elevated daily issue dates without a subsequent M5+ event under an inherited 5% diagnostic threshold. **Interpretation.** The results support a reproducible implementation and a small exploratory average-score gain, predominantly during quiet periods. They do not establish improved M5+ event anticipation, reliable calibration, geographical generalization or prospective skill. Revised catalog, geometry and GNSS vintages prevent an as-issued historical interpretation. The work contributes an auditable, constrained multimodal experiment and explicit negative evidence rather than a demonstrated operational forecasting advance.

Keywords: earthquake forecasting; Marmara; ETAS; neural point process; GNSS; fault geometry; proper scoring rules; catalog uncertainty; retrospective evaluation; reproducibility.

Contents

1. Introduction . . . 3

2. Related work and the scope of novelty . . . 4

3. Estimand, hypothesis and study design . . . 5

4. Data and provenance . . . 7

5. Statistical reference and inherited neural component . . . 9

6. Fault and GNSS extensions . . . 11

7. Fitting and numerical implementation . . . 14

8. Seven-day probability calculation . . . 16

9. Evaluation methods . . . 17

10. Results . . . 18

11. Discussion . . . 25

12. Limitations and threats to validity . . . 26

13. A falsifiable next-stage validation protocol . . . 28

14. Research evidence and public implementation . . . 29

15. Conclusions . . . 30

Data and author statements . . . 30

Appendix A. Implementation specification . . . 31

Appendix B. Data sources and attribution . . . 32

Appendix C. Target-event identity and temporal interpretation . . . 32

Appendix D. Verification and interpretation . . . 33

References . . . 33

1. Introduction

1.1 Scientific problem

A regional earthquake forecast is a probability statement about a defined future event set. Its usefulness depends on whether probabilities discriminate changing conditions, remain calibrated and improve on credible alternatives under information constraints that could have been met at issuance. A visually plausible model, a technically successful deployment and a neural network trained on many earthquakes answer different questions. None by itself demonstrates that a model anticipates moderate earthquakes better than a carefully fitted statistical reference.

The epidemic-type aftershock sequence model represents earthquake clustering through background occurrence and triggering by previous events [1]. Its structure offers an interpretable reference for evaluating additional information. A neural extension can be useful if it captures reproducible departures from that reference. However, a model may improve an overall score by lowering probabilities during many uneventful periods while providing no improvement before target events. This distinction is especially consequential when a study contains thousands of small earthquakes but only a few moderate targets.

The present work investigates a narrow question: does a bounded correction informed by mapped faults and measured, lagged GNSS histories improve seven-day native M5+ occurrence probabilities for a specified Marmara box? The target is deliberately distinguished from earthquake timing prediction, damage forecasting, early warning after rupture onset and fault-specific M7+ recurrence. The experiment does not estimate any of those quantities.

1.2 Motivation for a constrained hybrid

Fault geometry could inform where recent seismicity occurs relative to mapped structures. GNSS histories could describe recent surface displacement patterns that are not represented directly in an earthquake catalog. These are plausible covariates, but neither is an established short-term precursor in this experiment. A mapped trace is an uncertain geometric representation; a GNSS coordinate increment combines tectonic motion, environmental loading, reference-frame effects, instrumentation and processing errors. Inferring impending rupture from either requires more evidence than predictive association.

A constrained hybrid offers a testable intermediate design. The statistical reference remains fixed, and small neural branches modify its intensity within explicit bounds. This limits the extension's capacity, preserves a recoverable reference and makes ablation comparisons possible. The bounds do not confer physical validity. In the fitted Marmara reference, the branching parameter is near its imposed upper limit, so the permitted upward correction is much smaller than the permitted downward correction. This asymmetry is part of the experiment, not a detail that can be omitted when interpreting a negative result.

1.3 Contributions and scope

The study makes four concrete contributions. First, it connects immutable catalog observations, fault geometry, GNSS coordinate products and model outputs through documented data preparation. Second, it implements small fault and GNSS residual branches with explicit latency and missing-data rules. Third, it evaluates moderate-event probabilities through simulated clustered histories rather than substituting a weekly Poisson model. Fourth, it presents every primary comparison, including negative event-window results, false-elevation diagnostics and limitations of retrospective data availability.

The scientific contribution is an implementation and evaluation of a specific hypothesis. Scientific priority for the architecture has not been established. The historical period had already been inspected during earlier project work, and the extension protocol was recorded after those outcomes were known. Therefore, the evaluation is exploratory even though extension fitting and feature normalization use only observations dated before the chronological cutoff.

2. Related work and the scope of novelty

2.1 Statistical and neural earthquake forecasts

ETAS provides a family of models rather than one universal off-the-shelf benchmark [1]. Spatial kernels, background-rate estimation, magnitude productivity, catalog thresholds and simulation methods materially affect comparisons. A neural model should therefore be compared with a disclosed, regionally fitted reference. It should also be evaluated on the probability target of interest rather than only on training loss or classification accuracy.

Neural earthquake forecasting already includes recurrent point-process formulations such as RECAST [11]. The current EarthquakeNPP version reports comparisons involving five neural point-process models and finds that none outperform its ETAS reference under its benchmark [12]. That result does not prove that all neural approaches fail. It does establish the need for strong references and carefully designed evaluations. The distinction between temporal likelihood, spatial likelihood and target-event probability remains essential: improvement in one does not automatically imply improvement in the others.

2.2 Closest multimodal precedents

Published work predates the present project in each broad architectural direction. González Fuentes et al. combine ETAS-related intensity and crustal-velocity information in a multi-column ConvLSTM for Chilean seismicity [7]. Zhang et al. combine ETAS and ConvLSTM forecasting for sequences following three Southern California mainshocks [8]. Tong et al. investigate GNSS-based earthquake alerting with a neural information-transformation method [9]. Kaya et al. use fault-graph representations with magnitude and occurrence-time models in the northern Aegean [10]. Their targets, data and validation designs differ from this study; their reported scores are not interchangeable with the results here.

A particularly close regional comparison is the 2026 EarthArXiv preprint by Alhan and Khabat, which describes a Marmara ETAS/ML evaluation with fault-related features and GNSS extensions [14]. Its authors report negative or non-robust incremental results. This is a preprint, and its numerical findings were not independently reproduced for the present manuscript. Its existence nevertheless limits any claim that adding geodetic or fault information to a Marmara statistical/ML forecast is new in itself.

Table 1. Selected prior work and the relationship to the present experiment. This is a targeted primary-source comparison, not an exhaustive systematic review.

Work	Relevant precedent	Distinction from this experiment
González Fuentes et al., 2022 [7]	ETAS-related intensity, crustal velocity and neural sequence modeling	Chile; next-day seismicity; multi-column ConvLSTM
Zhang et al., 2024 [8]	ETAS and deep-learning hybrid	Three Southern California sequences; ConvLSTM design and different target
Tong et al., 2023 [9]	GNSS histories and neural earthquake alerting	Different transformation method; does not establish this residual architecture
Kaya et al., 2026 [10]	Mapped-fault graph representations	Northern Aegean; magnitude and occurrence-time objectives
Alhan and Khabat, 2026 [14]	Marmara ETAS/ML, fault covariates and GNSS extensions	Preprint; predominantly 30-day targets; results not reproduced here
Present study	Bounded residual branches, lagged GNSS, fault proximity and clustered seven-day simulation	Exact implementation is documented; priority and reliable added skill remain unestablished

2.3 Defensible novelty statement

The appropriate statement is that DepremİstanbuI implements a Marmara-specific ETAS-neural model combining fault-proximity attention and lagged GNSS temporal features for seven-day source-native M5+ occurrence. Established statistical, recurrent, graph-attention and convolutional ideas underlie the implementation [1, 16, 17, 18]. The present contribution lies in their specified integration, source reconciliation and transparent evaluation. No exhaustive literature search, patent search or independent priority review has established that the exact combination is a scientific first.

The targeted review covered primary publisher records, author institutions and preprint records available by 12 October 2026. Full text was inspected when accessible; some comparisons relied on official abstracts or record descriptions. Absence of an exact match in that search cannot establish originality. Architectural difference and forecasting skill are separate empirical claims.

3. Estimand, hypothesis and study design

3.1 Region, horizon and magnitude definition

The target region is the closed longitude-latitude box 26-31° E and 39.5-41.7° N. It is called “Marmara” for this experiment, although a rectangular domain includes neighboring onshore and offshore structures and is not a unique tectonic segmentation. The forecast horizon is seven days. For an issue time T, the binary outcome is one if at least one catalog event with selected native magnitude at least 5 occurs within the region during the half-open interval [T, T + 7 days).

$$Y_T = \mathbf{1}\{N_{M_x \geq 5}(R, [T, T+7)) \geq 1\} \qquad (1)$$

The estimand is the conditional probability of that outcome given the model's permitted history and fixed parameters. “Native M5+” refers to the selected source magnitude variable Mx used

by the code. It is not a posterior statement that physical moment magnitude Mw exceeds 5.0. Reporting that distinction prevents apparent catalog agreement from being mistaken for resolved magnitude uncertainty.

3.2 Falsifiable hypothesis

The operational hypothesis is that at least one added modality improves both mean weekly logarithmic score and Brier loss relative to both the retained spatial ETAS and the frozen neural hybrid, without deterioration in the supporting M2+ marked likelihood or a numerical/causality failure. The three extensions are evaluated separately. The protocol's point criteria identify exploratory progress; they do not define scientific reliability or authorize a claim of independent confirmation.

A stronger hypothesis would require positive expected score gain on unseen, historically available or genuinely prospective data, stable calibration and replication across independent regions. That stronger hypothesis is not established by this study. This separation of development, testing and communication is consistent with current evaluation guidance [13]. The distinction matters because a small positive average difference in an exposed historical period can satisfy the recorded point criteria while remaining compatible with chance, selection effects or a narrow quiet-period adjustment.

3.3 Chronology and prior exposure

The model uses observations from 1 January 1990 onward. The first 365 days condition the temporal history and are excluded from the scored fitting objective. The chronological fitting cutoff is 1 January 2014. Evaluation ends at 1 November 2018, exclusively. The scored training period contains 8,401 days and 31,168 M2+ events; the complete pre-cutoff history contains 31,733 M2+ events. The evaluation period contains 1,765 days and 5,134 M2+ events.

There are 252 complete nonoverlapping seven-day windows, beginning on 1 January 2014 and ending with the window issued on 24 October 2018. There are also 1,759 complete daily issue dates for overlapping seven-day diagnostics. Five native M5+ events occupy three positive weekly windows. The remaining 249 weekly windows have no target event. The final coverage day that does not complete another weekly block is retained in the underlying daily and likelihood analyses where applicable.

The extension protocol was recorded on 12 October 2026 at 13:06:34 UTC before fitting the extension branches. Historical test outcomes were already exposed through earlier research. Consequently, “chronologically later evaluation” is accurate, while “untouched test set,” “prospective validation” and “preregistered independent confirmation” are not. The study does not retroactively convert a known historical sample into independent evidence.

3.4 Model comparisons

Six models are retained: a stationary pre-cutoff native-M5 rate, the regional spatial ETAS reference, the frozen ETAS-GRU/grid hybrid, and three extensions adding faults, GNSS or both. The ETAS fit and base neural weights remain fixed during extension training. The frozen base was previously fitted jointly with Marmara and California observations. No additional California training or evaluation is introduced in this experiment. California therefore cannot be presented here as an independent external validation region. The inherited California data retain SCEDC and SCSN attribution [20, 21].

4. Data and provenance

4.1 Catalog reconciliation and event identity

The earthquake observations derive from the public supplementary release associated with TURHEC [2]. The acquired uncompressed source table contains 377,429 records. The study retains a derived observation table with source lines, ISC event identifiers, original fields, selected native magnitudes and available uncertainty descriptors. Unique ISC identifiers are asserted after filtering and sorting; multiple reports are not naively concatenated as independent earthquakes.

The inherited reconciliation work addressed gaps in a previously acquired AFAD history, including 1993, 1996 and 1997. The public TURHEC source contains 3,242 records in those years within the earlier repair analysis. Filling missing years improves provenance and exposure accounting, but does not establish complete detection. The present model consumes the reconciled public observation product rather than a restricted catalog extension.

Magnitude harmonization remains incomplete in a scientific sense. The observation records preserve reported magnitude summaries, the published equivalent magnitude and absent measurement-error fields. Between-report scatter is not relabeled as a measurement standard deviation. A reported or converted magnitude is not promoted into a calibrated latent-Mw posterior. The model nevertheless uses a deterministic selected Mx for thresholding and productivity. This choice makes the implemented target reproducible while leaving uncertainty near thresholds unresolved.

TURHEC's publication reports a catalog-level completeness estimate on its homogenized magnitude scale [2]. That estimate cannot establish local, time-varying completeness at native M2 in this box. The present M2 threshold is an experimental input choice. Changes in network capability, reporting, magnitude practice and short-term post-event detection can affect both learned background rates and apparent triggering. A complete observation model would need explicit detection probabilities and uncertainty propagation; those are not implemented here.

4.2 Fault geometry

Fault inputs are taken from the GEM Global Active Faults 2019.0 release associated with the database described by Styron and Pagani [3]. The selected representation uses the EMME catalog alone, avoiding the automatic stacking of overlapping EMME and SHARE traces. All represented mechanisms are retained within the study box plus a 30-km buffer. The final graph contains 55 traces and 846 short line elements used for distance calculations.

Trace proximity is a geometric feature, not a rupture assignment. The graph does not establish that nearby faults rupture together, transmit stress with the same sign, share a seismogenic depth or have comparable slip rates. No segment-specific slip rate is invented. The graph includes no historical assignment of the 1766, 1894 or 1912 earthquakes, and none is inferred from the neural attention weights.

The source is a 2019 geometry product used to analyze 2014-2018 outcomes. This is a retrospective geometry sensitivity study. It is not evidence that the same digital representation was available at each historical issue time. Derived geometry and the map carry GEM/EMME attribution and conservative share-alike treatment where applicable; published licensing statements differ, as explained in Appendix B.

4.3 GNSS coordinate products

The GNSS source is the Nevada Geodetic Laboratory's documented IGS20 daily tenv3 coordinate product [4, 5, 6]. The two stations are ISTA at approximately 41.10445° N, 29.01934° E and TUBI at 40.78673° N, 29.45067° E. The acquired files contain 9,140 and 8,415 daily rows, respectively, totaling 17,555. The records extend beyond the historical evaluation period; only appropriately lagged epochs are used in each constructed feature.

These are processed coordinate solutions, not raw carrier-phase observations. Their coordinate representation, reference-frame documentation, formal component uncertainties, correlations and station metadata are retained. Products from different analysis centers are not mixed into a single undocumented time series.

Daily covariance matrices are reconstructed from supplied uncertainties and correlations and checked for admissibility. The pipeline rejects nonfinite values, nonpositive formal standard deviations, invalid correlations, indefinite covariance, horizontal formal uncertainties above 50 mm, duplicate epochs, date/MJD inconsistencies and coordinate-meridian changes. Accepted daily covariance does not provide the missing cross-day and cross-station covariance. The analysis therefore cannot claim that the neural output integrates a complete geodetic error model.

Instrument-change metadata and daily antenna-height changes constrain usable windows. Trend fitting starts after the latest relevant step, and increments crossing a step are masked. Earthquake-related offsets are not invented or removed through a hand-crafted list inferred from future outcomes. This policy reduces one source of analyst discretion, but unresolved processing changes and environmental signals remain possible explanatory alternatives.

4.4 Source versions and access

The catalog uses the public TURHEC supplementary release dated 28 May 2021. Fault geometry uses GEM Global Active Faults 2019.0. The ISTA and TUBI IGS20 coordinate series were accessed on 12 October 2026. Appendix B identifies these sources and their roles. Their publication and access dates do not establish availability at each historical forecast time.

The public TURHEC source is associated with CC BY 4.0. GNSS use is accompanied by provider and station attribution. This report presents derived figures and results; it does not redistribute the original station coordinate series or alter third-party data rights.

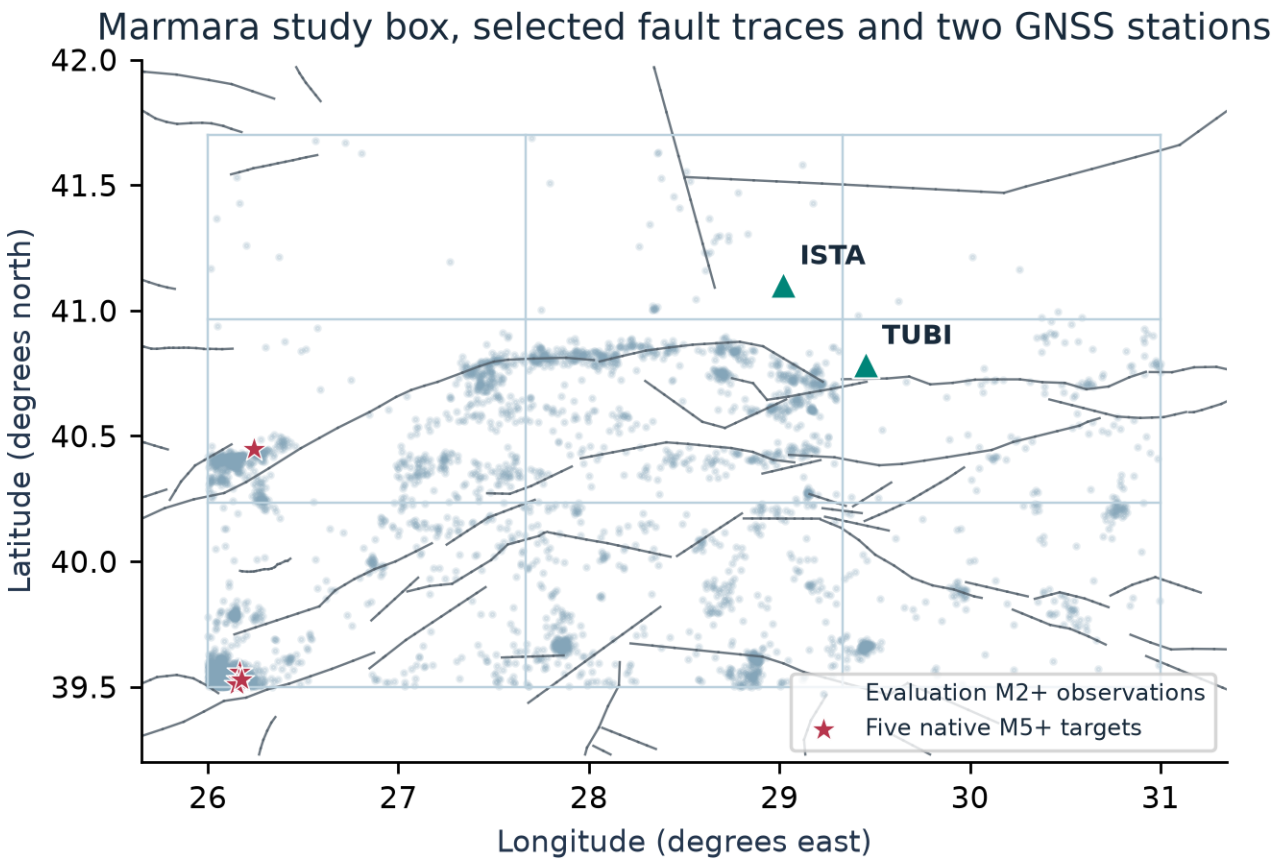

Figure 1. Study region and input geometry. The blue points show the 5,134 evaluation M2+ observations; red stars mark the five source-native M5+ targets. Gray lines are selected GEM/EMME traces, with the 3 × 3 forecast grid overlaid. Green triangles identify ISTA and TUBI. The panel is a coordinate plot without a coastline basemap; line proximity is not evidence of rupture connectivity. Geometry derived from GEM/EMME [3], with attribution and share-alike treatment described in Appendix B.

5. Statistical reference and inherited neural component

5.1 Spatial ETAS intensity

Let c index one of nine rectangular spatial cells and let the history contain events i with time ti, location si and native magnitude-bin mark ki. The retained reference is a cell-integrated marked ETAS intensity. Its background is the product of a regional rate μ and normalized cell weights bc. Triggering combines a normalized mark productivity, a spatial allocation kernel and an Omori-type temporal density.

$$\lambda_c^E(t) = \mu b_c + n \sum_{t_i < t} w_{k_i} K_c(s_i; \sigma) g(t - t_i; c_0, p) \qquad (2)$$

$$g(u; c_0, p) = \frac{p-1}{c_0}\left(1 + \frac{u}{c_0}\right)^{-p}, \quad u > 0 \qquad (3)$$

The notation c0 denotes the temporal offset and is distinct from the cell index. Kc is the integral of an isotropic Gaussian over a rectangular cell, normalized across the nine cells. The region is therefore closed for this spatial allocation: mass is not lost outside the domain, and unobserved outside parents are not explicitly represented. This boundary convention may distort triggering near the edges and should be revisited in a future scientific reference.

Coordinates use a fixed local approximation, x = 111.195 cos(40.6°) times longitude and y = 111.195 times latitude, in kilometers. Cells are approximately 140.7 by 81.5 km. The coarse grid is sufficient to define the implemented regional target, but it is not a detailed fault-source mesh. Spatially precise trace features are pooled into much larger forecast cells, a mismatch that constrains physical interpretation.

The numerical temporal state uses a 161-term exponential quadrature on rates spanning 10^{-12} to 10^{5} per day. This supports efficient state updates and integrated intensities. The approximation is inherited unchanged for all extensions; it is not a new analytic law for earthquake triggering.

5.2 Magnitude marks and productivity

Native magnitudes are grouped into [2,3), [3,4), [4,5) and [5,∞). The empirical mark law uses counts from the full pre-cutoff conditioning/training history with a half-count added to each bin. Let Nk be the bin count and N their total.

$$q_k = \frac{N_k + 0.5}{N + 2}, \qquad w_k = \frac{\exp[a\min(\bar{m}_k - 2, 3.5)]}{\sum_j q_j \exp[a\min(\bar{m}_j - 2, 3.5)]} \qquad (4)$$

This normalization makes average productivity equal to one under the fitted mark law. The model does not estimate a time-varying large-event mark probability. Conditional on the modeled total intensity, a fixed q5 determines the target-event rate. Thus, small-event history changes M5+ occurrence primarily through intensity and triggering rather than through a learned physical mechanism for changing the magnitude tail.

Table 2. Fitted native mark distribution. Counts include the one-year conditioning history. Values are rounded for presentation.

Native magnitude bin	Count	Mean magnitude	Smoothed probability
[2,3)	27,641	2.60222	0.871009926
[3,4)	3,930	3.18102	0.123853789
[4,5)	143	4.27273	0.004521821
[5,∞)	19	5.38947	0.000614464

5.3 Fitted parameters and baseline limitations

The retained spatial ETAS fit uses pre-2014 Marmara observations and a background-weight regularization equivalent to adding 0.5 times the sum of log background weights to the objective. The recorded optimization includes two starting points. Convergence diagnostics support numerical completion within the chosen model, but the selected branching parameter reaches its imposed upper bound of 0.95. Optimizer convergence at a boundary is not evidence that the physical branching ratio is known or that the reference is globally optimal among plausible ETAS formulations.

Table 3. Retained Marmara ETAS parameters.

Parameter	Value	Interpretation
μ	0.8910045 per day	Regional M2+ background rate
n	0.9500000	Mean productivity parameter; at imposed cap
a	0.9384792	Native mark productivity slope
c0	0.0222422 days	Temporal offset
p	1.1266944	Temporal decay exponent
σ	8.2875294 km	Gaussian spatial scale
Background weights	Nine normalized values	Jointly fitted cell allocation; Appendix A

The baseline is fitted, but “properly calibrated” requires empirical evidence beyond optimization. This study reports its probability scores and calibration summaries; it does not establish exact calibration of the reference. Alternative background models, spatial kernels, boundary treatment, detection models and mark laws remain scientifically relevant competitors. Added neural skill relative to this particular ETAS implementation is not equivalent to superiority over the best achievable ETAS model.

5.4 Frozen GRU and grid correction

The inherited neural component contains 625 trainable parameters, frozen during this study. It encodes the last 32 events with an eight-unit gated recurrent unit (GRU) [16]. Each event token contains four magnitude-bin indicators, normalized horizontal coordinates, a transformed age and a validity channel. Seven cell features describe recent counts over 1, 7 and 30 days relative to annual activity, annual rate, current ETAS intensity relative to background and normalized cell coordinates.

A shared grid module combines local cell features, neighboring-cell summaries and the recurrent context. An eight-unit hidden representation produces one raw correction per cell. This is a deliberately small model. Its earlier joint fitting used Marmara and California data; freezing it preserves a numerical control but does not erase prior model selection or make the current historical evaluation independent.

The neural correction modifies total M2+ intensity. It does not output an arbitrary binary M5 classifier score. Seven-day target probabilities are subsequently derived from the resulting marked, clustered process. This separation makes the intensity model and the probability calculation individually inspectable.

6. Fault and GNSS extensions

6.1 Fault representation and event-relative activity

Each selected fault trace has nine features. Five encode fixed geometry: normalized length-weighted centroid coordinates, transformed length and two axial orientation components, cos(2θ) and sin(2θ). The axial form avoids treating a line's reversed digitization direction as a different orientation. Four additional channels summarize recent activity by native magnitude bin using the last 32 events.

For a recent event, activity decays exponentially with age on a seven-day scale and with point-to-trace distance on a 20-km scale. Contributions are assigned to the nearest selected trace. Exact point-to-segment distances are used on the subdivided geometry. These choices are deterministic modeling assumptions, not estimated tectonic interaction lengths.

$$A_{fk}(T) = \sum_{i \in H_{32}(T)} \mathbf{1}\{k_i = k, f_i = f\} \exp[-(T - t_i)/7 - d(s_i, f)/20] \qquad (5)$$

The neural input uses log(1 + Afk). A trace's neighbors are its three nearest eligible traces within 20 km, together with itself. Proximity uses a symmetrized endpoint-to-trace distance rule. This graph is local and computationally small; it is not an independently reviewed rupture-connectivity model. A different trace segmentation or distance rule can change the graph and therefore the covariates, even if the underlying geological interpretation is similar.

An eight-unit linear embedding followed by tanh represents each trace. Four-dimensional query and key projections produce masked, scaled dot-product attention; an eight-dimensional value projection forms messages. A residual sum of local representation and neighbor message passes through tanh and a scalar output layer. This use of attention follows established neural design principles [17], with a project-specific geometry graph and feature set.

$$\alpha_{fj} = \frac{\exp(q_f^{\mathsf{T}} k_j/2)}{\sum_{r \in \mathcal{N}(f)} \exp(q_f^{\mathsf{T}} k_r/2)}, \quad j \in \mathcal{N}(f) \qquad (6)$$

Trace outputs are projected to the nine forecast cells with normalized weights proportional to exp(−distance/20 km). A total of 217 parameters is added. Learned attention weights are not interpreted as stress-transfer coefficients, segment probabilities or causal explanations for an earthquake. The branch tests whether this representation improves forecasting, not whether it reconstructs fault mechanics.

6.2 GNSS feature construction

For an issue time T, the GNSS feature end is U = T − L, where the primary lag L is 21 days. A 35-day lag is evaluated as a fixed-weight sensitivity. The lag is an operational assumption motivated by final-product latency; the exact publication time of each historical coordinate solution is unavailable. Lagging epoch dates cannot undo later reprocessing, so the feature set remains retrospective.

For each station and horizontal component, an ordinary least-squares trend is fitted using valid observations from the later of U − 365 days or the most recent instrument step, up to U − 63 days exclusively. At least 100 observations are required. Eight weekly feature windows then characterize horizontal increments after this earlier trend period. Each increment compares medians in adjacent seven-day windows, requiring at least five observations in each window. An increment crossing an instrument step is invalid.

The implementation subtracts the fitted trend from each observation before calculating weekly medians. This ordering is necessary because the observation dates can be irregular: the median of residuals generally differs from the raw median minus a trend evaluated at an arbitrary representative date. A test caught the earlier incorrect ordering, and the feature arrays were rebuilt before the reported probability evaluation. Retaining that correction in the research record is part of the reproducibility evidence.

Each station-week contains six numerical features plus a validity mask: detrended east and north median increments in millimeters; log-transformed median formal east and north

uncertainties; median east-north correlation; and coverage measured by the smaller of the two weekly observation counts divided by seven. The numerical channels are standardized using valid pre-2014 windows only and clipped to [−10,10]. Invalid entries are zeroed after applying the mask.

This feature design preserves uncertainty indicators but does not statistically integrate over the underlying measurement distribution. Formal uncertainties and correlations serve as covariates. The transformation does not recover a complete time-correlated displacement posterior, infer fault slip, or estimate a strain tensor. With two stations and absent network covariance, such stronger geodetic claims would be unsupported.

6.3 GNSS temporal encoder and missingness

The GNSS tensor has two stations, eight weekly positions and seven channels per station. The six numerical channels pass through a shared one-dimensional convolution with four output channels and a kernel width of three. Two positions of left-only padding preserve causal temporal direction. A tanh nonlinearity is followed by a masked average across valid weeks and a bias-free scalar output. A station with no valid weeks contributes exactly zero.

Station outputs are pooled to cells using fixed normalized weights proportional to exp(−station-to-cell distance/100 km). This is a feature-interpolation convention. It is not a physical Green's function, a slip inversion or a covariance-weighted strain solution. The GNSS branch adds 80 parameters. The joint fault-plus-GNSS extension adds 297 parameters, bringing the combined neural count to 922 when the frozen 625-parameter base is included.

The mask provides a deterministic fallback. When the GNSS-only model receives no valid GNSS input, it reproduces the frozen hybrid to numerical precision. In the combined model, removal of GNSS leaves the fitted fault branch active. Because the branches were fitted jointly, the remaining combined-model fault weights need not equal the separately fitted fault-only model.

6.4 Bounded intensity correction

Let z0 be the raw frozen neural output and zF and zG the added branch outputs. The model adds these before the bounded transformation. The total raw output is z = z0 + zF + zG. Setting a branch output to zero yields the corresponding nested reference at initialization.

$$v_c = \tanh(z_c), \qquad L_+ = \log[\min(2, 0.99/n)] \qquad (7)$$

For nonnegative v, the log multiplier is vL+. For negative v, it is v log 2. The final cell intensity is the reference intensity multiplied by the exponential of this log multiplier.

$$\lambda_c^H(t) = \exp[\ell_c(t)]\lambda_c^E(t) \qquad (8)$$

At n = 0.95, the limiting multiplier range is 0.5 to approximately 1.042105. The upper value was chosen to keep the reference branching scale times the maximum multiplier below 0.99. This numerical constraint limits potential explosive behavior within the implemented model, but it is not a full physical stability proof or calibration guarantee. The asymmetry also makes downward adjustments substantially easier than upward adjustments.

Consequently, a small average gain dominated by lower quiet-period probabilities is compatible with the design's capacity. Conversely, failure to achieve a large pre-event increase cannot refute all possible uses of fault or GNSS data. It tests the specified low-capacity, bounded extension around this fitted baseline. Future comparisons should report whether results

depend on the boundary fit and whether alternative defensible constraints change the conclusion.

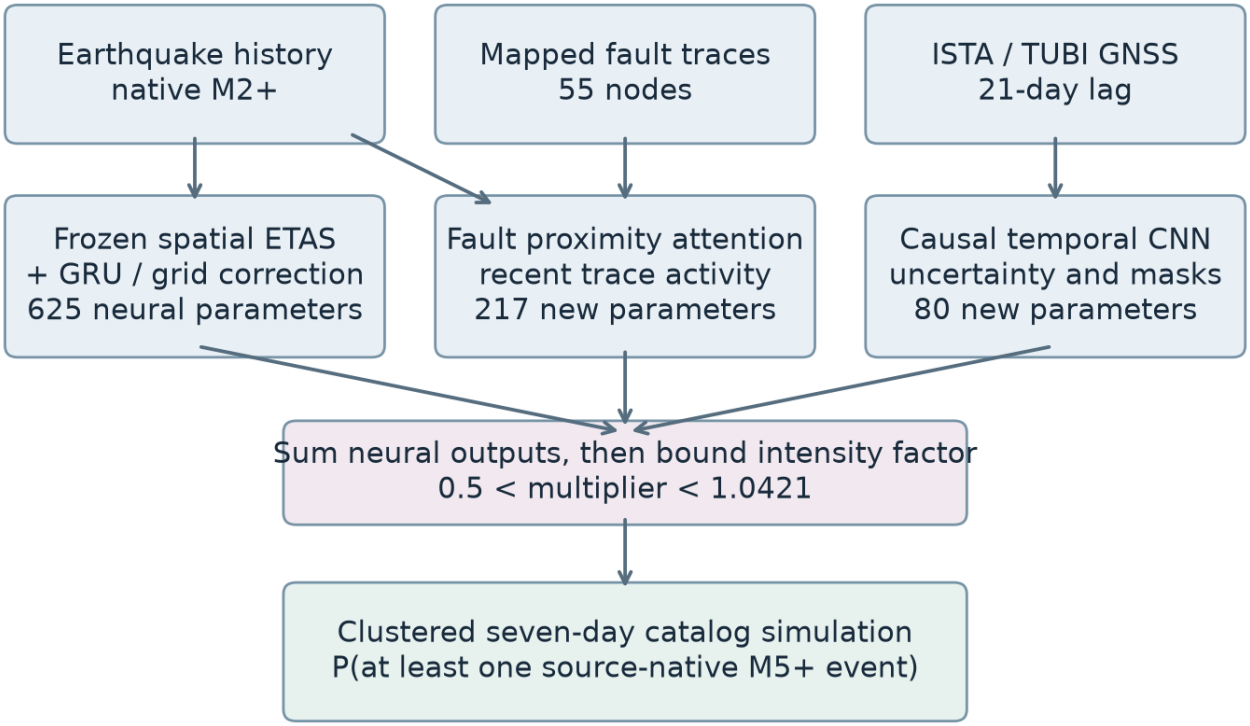

Figure 2. Implemented architecture. The earthquake-history reference and its neural weights remain frozen. Fault and GNSS branches add raw corrections before a shared bounded transformation. GNSS is fixed at the issue-time information cutoff throughout a forecast, while simulated earthquakes update seismicity-dependent features. The diagram represents the actual small model, not a larger Transformer or a physical geodetic inversion.

7. Fitting and numerical implementation

7.1 Likelihood objective

The new branches are fitted on the 8,401 scored pre-2014 days using the realized earthquake history. Let Adc be the reference integrated intensity in day d and cell c, Ndc the corresponding event count and dc the model's daily log multiplier. Relative to the fixed ETAS likelihood, terms independent of the extension cancel. The minimized objective is a normalized negative log-likelihood difference plus a quadratic penalty on the total log multiplier.

$$\mathcal{J}(\theta) = \frac{\sum_{d,c}[A_{dc}(\exp \ell_{dc} - 1) - N_{dc}\ell_{dc}]}{\sum_{d,c} N_{dc}} + 0.01\,\mathrm{mean}_{d,c}\,(\ell_{dc}^2) \qquad (9)$$

The penalty applies to the total correction relative to ETAS, not solely to the added branch relative to the frozen hybrid. Therefore, a new branch can partly cancel the inherited neural correction if this improves the penalized objective. This distinction is relevant to interpreting both nesting and ablation: zero-initialized outputs reproduce the reference at the start, but the fitted model is not constrained to preserve each component's previous contribution.

The mark law is fixed, so its common terms cancel in the extension likelihood difference. Fitting uses all scored native M2+ events, not only the rare M5+ targets. That increases the amount of training information for the intensity model. It also creates an objective mismatch: an improvement in small-event marked likelihood need not improve the binary probability of an M5+ event within a week.

7.2 Optimization and retained fits

All three extensions use full-batch AdamW [18], learning rate 0.01, weight decay 0.01, gradient clipping at norm 10 and 200 epochs. Seeds 11, 29 and 47 are fitted for each variant, and every fit is retained. Selection uses the lowest penalized pre-cutoff training loss. No extension seed is selected by the later weekly M5 scores. All three selected fits use seed 47 and epoch 200.

The selected penalized objectives are approximately -0.0169464 for fault, -0.0171608 for GNSS and -0.0182482 for the combined branch. These losses are not directly comparable to mean M5 log scores; they optimize a different quantity on a different sample. The best loss reaching the final allowed epoch means that the run-length limit was active. It does not establish full convergence or justify silently extending only the best-performing variant after examining evaluation scores.

No isotonic regression, temperature scaling or other post-hoc M5 calibrator is fitted in this extension experiment. With only three positive evaluation windows, selecting and evaluating a calibrator on the same period would be particularly misleading. Model capacity, feature scales and lags are disclosed so that future evaluation can test frozen choices rather than reuse these outcomes for further selection.

7.3 Software verification and limits

The archived research evidence records 15 focused scientific tests, reconstruction of all nine fitted training records and close agreement between the independent NumPy prediction path and the training framework. The maximum raw implementation discrepancy reported for the selected combined log factors is below 2×10^{-7}. Future-array poisoning tests support numerical causality at fractional issue times, and selected simulated forecast paths were replayed within 10^{-12}.

For this manuscript, an additional independent analysis of the archived simulation results recomputed all 7,560 forecast records across 15 forecast series, including 15,162 stored simulation batches when refinement-stage originals are counted. Stored batch means and standard errors matched their recomputation exactly in the checked floating-point representation. All six models' weekly scores and the descriptive block-resampling intervals were independently reproduced within 10^{-14}. These checks establish consistency between saved samples, outcomes and reported arithmetic; they do not demonstrate scientific reliability or independence of the historical experiment.

The numerical audit did not retrain the model or regenerate the complete simulation experiment. Computational checks establish specified numerical properties; they cannot repair sparse target support, catalog bias or lack of a prospective test.

8. Seven-day probability calculation

8.1 Why a direct Poisson conversion is insufficient

For a stationary Poisson model with fixed rate r, the probability of at least one event in seven days is 1 − exp(−7r). The stationary comparison uses that construction with the pre-cutoff native-M5 rate. In a self-exciting process, future small events change later intensity. A conversion based only on a current rate or on an expected event count generally does not equal the probability of zero events under clustering.

The hybrid therefore simulates future seismicity and updates the earthquake-dependent features. This preserves the feedback through which small simulated events affect subsequent triggering and neural corrections. The resulting probability is conditional on the fitted parameters, the fixed mark law and the chosen simulation conventions. It is not a posterior probability marginalized over all plausible parameter, catalog and geodetic uncertainties.

8.2 First-target-event survival estimator

The simulator uses a first-M5 survival construction. It generates histories containing only non-target marks at their corresponding intensity and accumulates the target hazard along each such history. If Jb is the integrated all-mark intensity along path b and q5 is the fixed target-mark probability, the path contributes one minus exp(−q5Jb). Averaging these contributions estimates the probability that a first target event occurs during the horizon.

$$\widehat{P}_T = \frac{1}{B}\sum_{b=1}^{B}\{1-\exp[-q_5 J_b(T,T+7)]\} \qquad (10)$$

The expectation over non-target paths is essential. Non-target earthquakes can trigger further events, so Jb varies across paths and is not replaced by its mean before exponentiation. Once a first M5+ event occurs, later target-triggered events are irrelevant to the binary occurrence question; the survival formulation exploits that fact. It does not estimate the full distribution of the number of M5+ events by itself.

The formal sample standard error is the sample standard deviation of the path contributions divided by the square root of the number of paths. Two independent batches are combined, and their recorded errors are propagated for the mean. This quantifies numerical simulation uncertainty conditional on the model. It excludes fitted-parameter uncertainty, catalog uncertainty, missing geodetic covariance, alternative graph definitions and model misspecification.

8.3 Forecast-time information rules

The earthquake history used to initialize ETAS contains events strictly before issuance. Neural earthquake features update daily during simulation, using simulated histories after issuance rather than observed future events. GNSS features are fixed at issuance for the entire seven-day horizon. This avoids feeding future realized GNSS observations into a forecast whose earthquake history is being simulated.

For a fractional-day issue, the first partial day's neural earthquake features are taken from the start of that UTC day, while the ETAS state conditions on the earthquake history through the exact issue time. This conservative within-day neural lag is part of the event-probe implementation. It should not be hidden by describing the entire model as continuously updated at every earthquake.

Within a simulated cell, locations are sampled uniformly under the implemented spatial convention. The Gaussian allocation governs which cell receives a triggered event, not a finely resolved continuous fault-distance posterior within the cell. Catalog-based testing tools such as pyCSEP support evaluation of simulated forecast distributions [19], but no complete independent CSEP evaluation or certification was performed in this experiment. A path guard rejects excessive event counts rather than silently truncating a simulated cascade. Such rejection is a numerical failure condition, not an acceptable way to suppress extreme model behavior.

8.4 Simulation sizes and reproducibility

Each weekly forecast uses two independent batches of 256 paths, for 512 paths in total. Exact seven-days-before-event probes use two batches of 1,024. Daily diagnostic forecasts initially use two batches of 128; forecasts within the larger of 0.005 probability units or four standard errors of the 5% threshold are refined using two batches of 1,024. The refinement criterion is disclosed and retained with its initial results.

Seeds are derived deterministically from region, issue day and batch index. The archived records preserve compressed path contributions, seed identifiers, probabilities and standard errors. This permits arithmetic verification without rerunning every path and supports selective exact replay. Numerical reproducibility should be distinguished from observational reproducibility: identical computational procedures and inputs reproduce a result, while independent data are needed to test whether the result generalizes.

9. Evaluation methods

9.1 Proper probability scores

The primary scores are the mean Bernoulli logarithmic score and mean Brier loss over the 252 complete weekly windows. Higher logarithmic score is better; lower Brier loss is better. Probabilities are clipped to $[10^{-12},\ 1-10^{-12}]$ for numerical scoring. The two scores weight probability errors differently, which helps reveal a result that depends on one scoring convention [15].

$$S_{\log}(p,y) = y\log p + (1-y)\log(1-p), \qquad S_B(p,y) = (p-y)^2 \qquad (11)$$

All comparisons pair models on exactly the same issue windows and catalog outcomes. Log-score gain is the candidate minus the reference; Brier improvement is the reference loss minus the candidate loss. Scores use natural logarithms. No classification accuracy percentage is treated as a reliability measure. With 249 quiet windows, an always-negative classifier would appear accurate while assigning no useful probability to the target events.

9.2 Dependence and descriptive intervals

The study uses circular moving-block resampling with blocks of 53 weekly positions, 5,000 replicates and a fixed random seed. Each replicate draws enough random block starts to cover the 252-window sequence, wraps around its end and truncates the concatenated series to the original length. The 2.5th and 97.5th percentiles of the paired mean score differences are reported as descriptive intervals.

These are not five independent calendar-year observations or a proof of 95% coverage. The sample includes only three positive weekly windows and a strongly clustered February 2017 sequence. Block choice, circular wrapping, historical model selection and the short evaluation

span constrain interpretation. The intervals provide a sensitivity description of the saved score series, not a reliable basis for declaring independent statistical significance.

9.3 Event, sequence and daily diagnostics

Each observed native M5+ event is paired with a reconstructed probability at its exact origin time minus seven days. These probes answer the diagnostic question about model behavior one week before each event. They are not forecasts actually archived in 2014 or 2017. Moreover, the event falls at the right endpoint of the corresponding half-open interval; the probe is a lead-time diagnostic and is not inserted into the primary weekly score as a positive outcome.

Sequence status is described by whether a native M5+ event occurred in the preceding 30 days. This rule separates two events without such a predecessor from three later events in the February 2017 sequence. It is a catalog-based label, not proven parentage or a seismological mainshock classification. Status is computed both at issue and at event time because the future sequence may not have begun at issuance.

A daily false-elevation audit uses the inherited 5% threshold. It counts issue dates above the threshold that do or do not have a target event during the next seven days. This threshold is a diagnostic inherited from earlier work, not an optimized or recommended public alert cutoff. Overlapping seven-day windows share outcomes; their counts are neither independent alarms nor independent successes.

9.4 Calibration, likelihood and fixed-weight sensitivities

Calibration is described by probability-bin support, average forecast probability and observed positive-window frequency. No claim of calibrated reliability is based on a bin containing one positive observation. The supporting marked M2+ log likelihood measures how well the model describes the much larger small-event sample. Its relationship to M5 skill is examined explicitly rather than assumed.

Sensitivity analyses remove GNSS, increase its lag to 35 days or permute fault assignment at inference. Parameters are held fixed; the sensitivities are not used to select a replacement model after examining the evaluation results. Removing a modality or permuting geography can produce an input distribution unlike training. These tests probe dependence on the encoded inputs, but they do not isolate a causal tectonic mechanism or constitute a valid alternative geological model.

10. Results

10.1 Overall weekly scores

Table 4 gives the primary probability scores. The frozen hybrid improves average log score over the retained ETAS reference in this sample. The fault-only extension slightly worsens both primary scores relative to the frozen hybrid. GNSS-only and combined extensions improve both average scores, with the combined model having the best mean logarithmic score and GNSS-only having the smallest Brier loss. The ranking therefore depends on which proper score is emphasized.

Table 4. Seven-day native-M5 probability scores on 252 nonoverlapping windows. Higher log score and lower Brier loss are better. Probabilities are percentages; score changes are relative to the frozen hybrid.

Model	Mean log score	Brier loss	Mean P (%)	Log gain	Brier improvement
Stationary	-0.065091338	0.011777494	1.5707	-0.009874181	-0.000702568
Spatial ETAS	-0.056722306	0.011142442	1.4534	-0.001505148	-0.000067516
Frozen hybrid	-0.055217158	0.011074926	1.2995	0.000000000	0.000000000
+ Fault	-0.055264743	0.011076955	1.2975	-0.000047585	-0.000002029
+ GNSS	-0.055061999	0.011068521	1.2706	0.000155159	0.000006405
+ Fault + GNSS	-0.054834015	0.011069370	1.2611	0.000383143	0.000005556

The combined model's mean log-score gain over the frozen hybrid is 0.000383143 per week. Its Brier improvement is 0.000005556. Relative to ETAS, the corresponding values are 0.001888291 and 0.000073072. These are point estimates on the exposed historical sample. They do not establish that the added modalities would improve future forecasting.

The mean predicted weekly probability falls from 1.2995% for the frozen hybrid to 1.2611% for the combined model, while the observed positive-window frequency is 1.1905%. Agreement of these two averages does not imply conditional calibration. A model can match an average frequency while assigning probabilities poorly across individual weeks or earthquake sequences.

10.2 Score support and uncertainty

The descriptive block interval for combined-model gain spans −0.000216568 to 0.000989610. The fault and GNSS intervals also include zero. The mean Monte Carlo standard error of an individual weekly probability is approximately 0.000093 for the combined model, equivalent to 0.0093 percentage points. That is a numerical error summary, not a confidence interval for scientific reliability.

Table 5. Log-score gain by outcome and descriptive resampling interval. All gains are relative to the frozen hybrid. Stratum means use three positive and 249 quiet weeks.

Extension	Positive-week gain	Quiet-week gain	Descriptive interval
+ Fault	-0.005562325	0.000018857	[-0.000162818, 0.000065010]
+ GNSS	-0.011512872	0.000295738	[-0.000164081, 0.000450345]
+ Fault + GNSS	-0.000254975	0.000390831	[-0.000216568, 0.000989610]

All extensions have negative mean log-score gain in positive weeks. The combined model's small negative positive-week contribution is outweighed by improvement in quiet weeks. Summing the paired weekly differences, a post-hoc manuscript decomposition gives +0.097317 log units from quiet windows and −0.000765 from positive windows, for +0.096552 overall. This descriptive identity is calculated from the same saved outcomes; it is not an additional independent experiment.

For GNSS-only, the positive-week loss is larger, despite a slightly better overall Brier loss than the combined model. The fault-only extension has a small quiet-week gain and a larger

positive-week loss, giving a negative overall result. These patterns do not support a claim that the extensions successfully anticipate the target events. They show how a model can improve an average score through modest adjustments during the dominant quiet state.

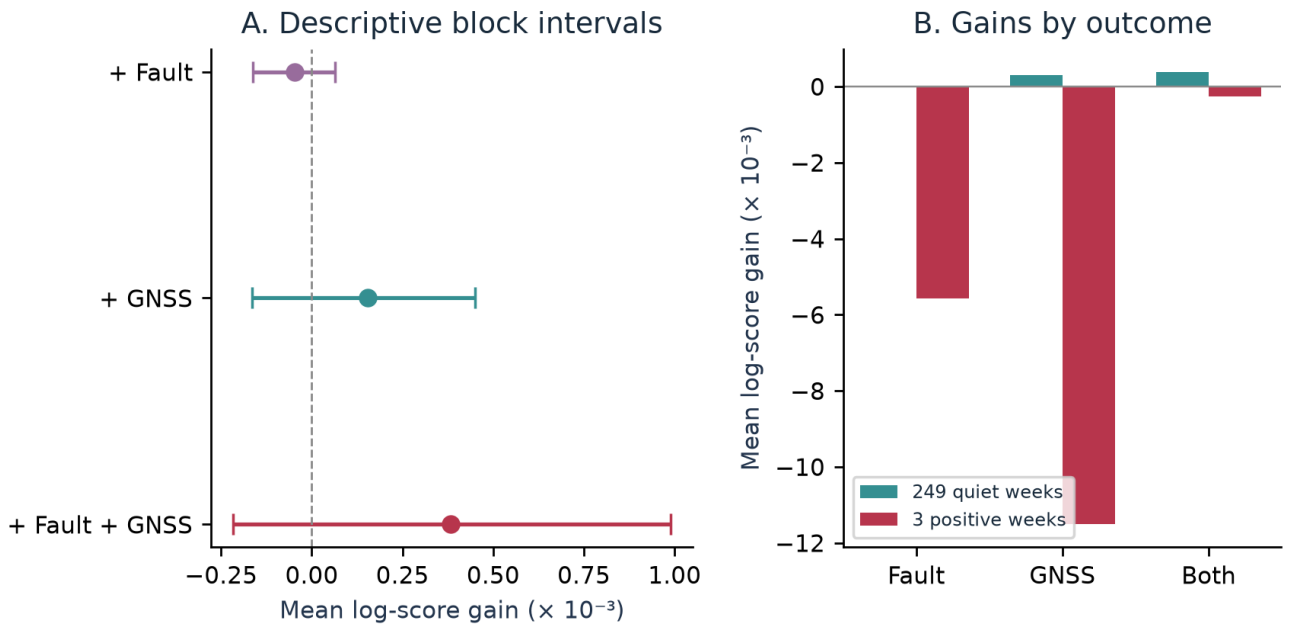

Figure 3. Incremental log-score evidence. Panel A shows paired mean gains with descriptive circular 53-week moving-block intervals. All three intervals include zero. Panel B separates positive and quiet weeks; the extensions' mean contributions in positive weeks are negative. Different support sizes are explicitly labeled, and the bars should not be interpreted as independent population effects.

10.3 Variation through time

The time series in Figure 4 shows a common increase during the February 2017 sequence. This is consistent with a clustering model responding to existing seismicity. It does not imply that the onset of the sequence was predicted. Differences between the frozen and combined hybrids are generally small relative to the shared temporal pattern.

Annual paired gains are heterogeneous. The combined model improves in 2014, 2015, 2017 and the evaluated portion of 2018, but worsens in 2016. The 2017 gain is small despite the concentration of target events in that year. A favorable aggregate result is therefore not a uniform improvement across years, and selecting a subset of favorable years would misrepresent the evaluation.

Table 6. Combined-model annual log-score gain over the frozen hybrid. The final year ends in October. Years with no positive weeks cannot independently establish target-event calibration.

Year	Weekly windows	Positive windows	Mean log-score gain
2014	53	1	0.001034898
2015	52	0	0.000390111
2016	52	0	-0.000436205
2017	52	2	0.000021080
2018	43	0	0.001000070

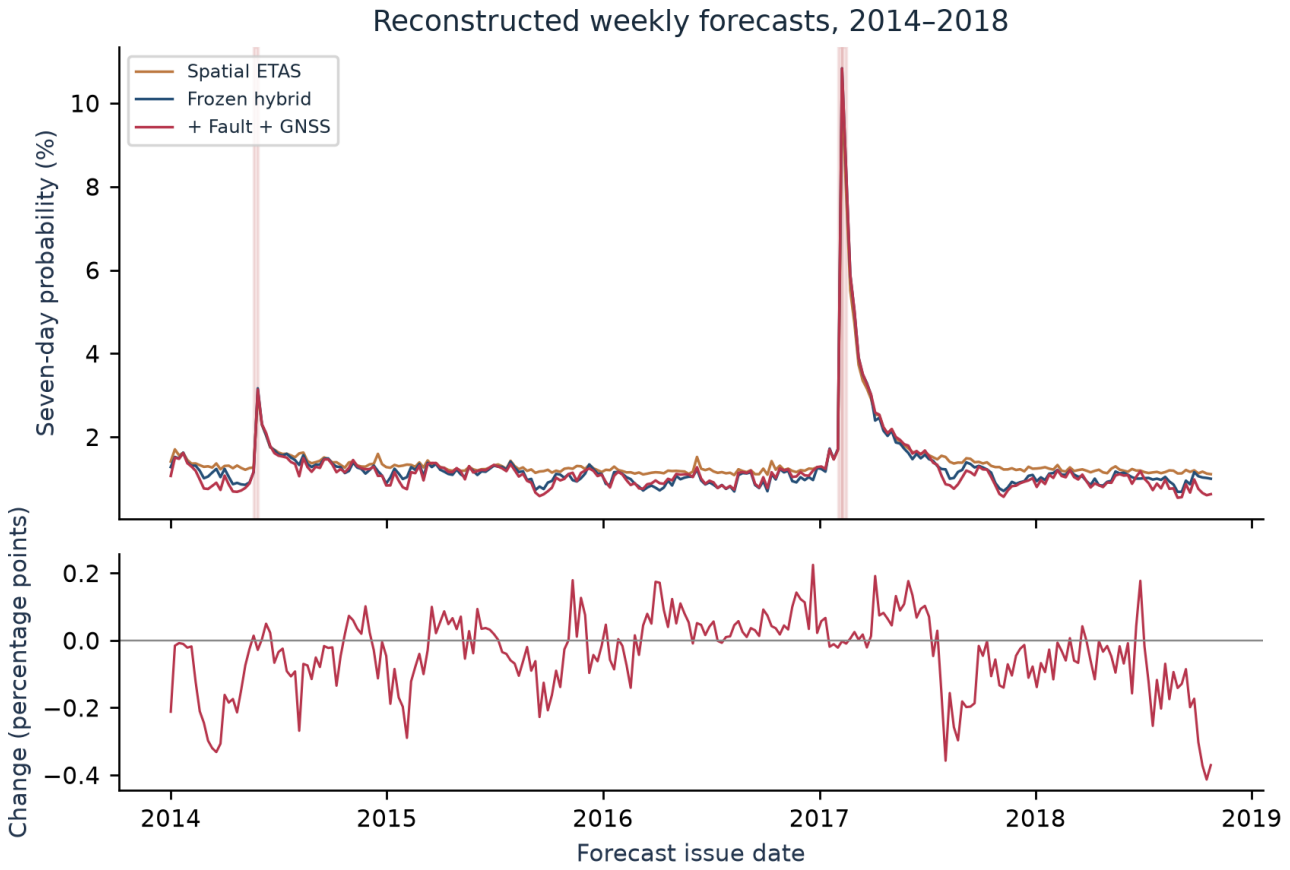

Figure 4. Reconstructed seven-day probabilities. Shading identifies the three positive weekly windows. The lower panel shows the combined-model change relative to the frozen hybrid in percentage points. The largest shared increase occurs within an existing sequence; the plotted series was reconstructed retrospectively and was not issued in these historical years.

10.4 Event-by-event seven-day lead-time audit

Table 7 lists the probability reconstructed exactly seven days before each observed target event. All five issue times precede any native M5+ event in their respective preceding 30-day histories. At event time, the later three February 2017 events have a recent native M5+ predecessor. This difference demonstrates why sequence labels computed after an event should not be used as if they were known at forecast issuance.

Table 7. Reconstructed probability (%) exactly seven days before each native M5+ event. Stationary probability is 1.5707% for every row. “Recent M5” describes event-time history, not issue-time history. ISC identifiers are given in Appendix C.

Event UTC	Mx	ETAS	Frozen	Fault	GNSS	Both	Recent M5
2014-05-24 09:31	5.3	1.255	1.100	1.109	1.181	1.132	No
2017-02-06 03:51	5.3	1.808	1.844	1.831	1.845	1.832	No
2017-02-06 10:58	5.2	1.843	1.872	1.867	1.865	1.862	Yes
2017-02-07 02:24	5.3	1.727	1.758	1.746	1.745	1.716	Yes
2017-02-12 13:48	5.3	1.702	1.757	1.746	1.755	1.772	Yes

The combined model is slightly above the frozen hybrid before the May 2014 event and the final listed February 2017 event, and below it before the other three. None of these combined-model lead-time probes reaches the inherited 5% diagnostic threshold. Their values cluster around 1.1-1.9%, which should be read as conditional model estimates rather than successful warnings.

The May 2014 target provides a concrete example of magnitude-type sensitivity. Its retained source record has a reported ML median of 4.9 and a reported Mw value of 5.3; the selected Mx is 5.3. The event therefore crosses a nominal threshold of 5 depending on which magnitude scale is chosen. The model reads the publication's Mx and Mx_type fields directly rather than constructing a new agency hierarchy. A physical-Mw target would require a separately audited measurement model. In addition, absence of a preceding in-box native-M5 event does not rule out triggering by larger events outside the box.

10.5 Elevated probability without a target event

Across 1,759 daily issue dates, the retained ETAS model exceeds 5% on 19 dates: six have at least one target in the following seven days and 13 do not. The frozen hybrid and all three extensions exceed the threshold on 23 dates: six have a target and 17 do not. The extensions do not improve this diagnostic count relative to the frozen hybrid.

Table 8. Daily 5% threshold diagnostic. There are 20 daily windows with a target outcome, many sharing the same earthquakes. Counts are issue dates, not independent alerts.

Model	Elevated dates	With target	Without target	Episodes
Spatial ETAS	19	6	13	1
Frozen hybrid	23	6	17	2
Stationary	0	0	0	0
+ Fault	23	6	17	2
+ GNSS	23	6	17	2
+ Fault + GNSS	23	6	17	2

The hybrid models produce two elevated episodes, compared with one for ETAS. Seven hybrid dates fall within the recorded threshold-ambiguity margin, and eight do for ETAS. The margin is a deliberately broad refinement/diagnostic rule, so it must not be interpreted as a formal posterior probability that the threshold decision is wrong. Rounding percentages for display can introduce further apparent changes near a threshold.

The same earthquake can satisfy several overlapping future windows. Consequently, the six dates with a target cannot be reported as six independently predicted earthquakes, and 17 non-target dates cannot automatically be read as 17 distinct public alarm episodes. A future operational analysis would need a separately specified alert policy, episode definition and cost model. No such policy is validated here.

10.6 Calibration support

The combined model places 96 windows in the 0.5-1% bin, with no positive outcomes; 145 in the 1-2.5% bin, with two positives; seven in the 2.5-5% bin, with none; three in the 5-10% bin, with none; and one in the 10-25% bin, with one positive. The single high-probability positive window produces an observed bin frequency of 100%, an illustration of sparse support rather than a reliable estimate of calibration.

Table 9. Combined-model descriptive calibration. Probability-bin limits and means are percentages. The observed frequency is positives divided by windows within each bin.

Probability bin (%)	Windows	Positive	Mean P (%)	Observed (%)
0.5-1.0	96	0	0.828	0.000
1.0-2.5	145	2	1.284	1.379
2.5-5.0	7	0	3.139	0.000
5.0-10.0	3	0	6.418	0.000
10.0-25.0	1	1	10.832	100.000

The primary dataset cannot establish stable reliability curves across this probability range. Standard binomial confidence intervals would additionally assume a form of independence that the clustered sample does not supply. The manuscript therefore reports counts and frequencies directly, avoiding an attractive but unsupported calibration-certification graphic. Figure 5 makes the sample sizes visible beside the points.

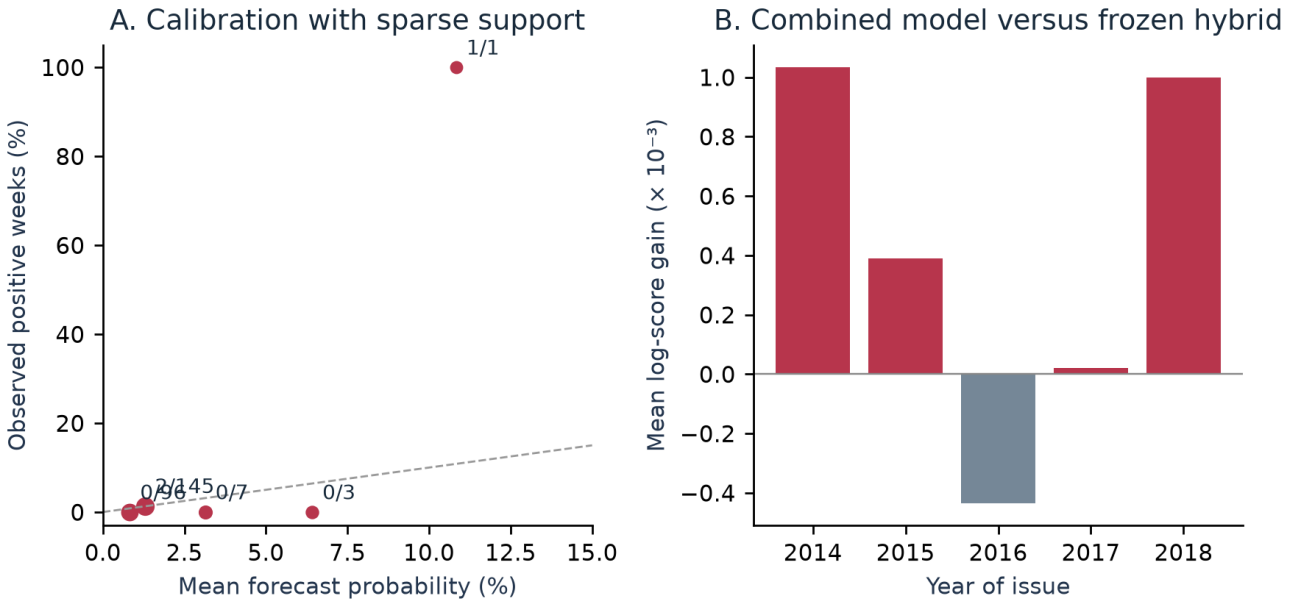

Figure 5. Sparse calibration and annual variation. Panel A labels each combined-model calibration point with positive windows/total windows; the 1/1 point is deliberately shown at 100% observed frequency. Panel B reports annual mean log-score differences from the frozen hybrid. These descriptive panels expose limited support and temporal heterogeneity rather than establish reliable calibration.

10.7 Supporting M2+ likelihood

The fault, GNSS and combined models improve held-out marked M2+ log likelihood over the frozen hybrid by 11.531, 8.120 and 25.463 natural-log units, respectively. Dividing by 5,134 evaluation events gives approximately 0.002246, 0.001582 and 0.004960 per event. The combined model improves by 132.834 log units relative to the retained ETAS reference.

These supporting results concern the modeled occurrence and spatial allocation of abundant smaller events under a fixed mark law. They do not turn the five M5+ events into a large independent sample. The combined model also worsens M2+ likelihood in 2016, matching the broader evidence of temporal heterogeneity. Small-event likelihood and moderate-event probability are related through the model but remain distinct evaluation targets.

10.8 Fixed-weight modality sensitivities

Removing GNSS from the GNSS-only model returns the frozen reference within numerical precision. Removing it from the combined model changes the mean weekly gain to approximately −0.0001878. Permuting the combined model's fault assignment changes the gain to approximately −0.0000157. These outcomes show that the fitted joint prediction depends on its input configuration; they do not prove that the geographic or displacement relationship is causal.

Table 10. Fixed-weight weekly sensitivity results. Gains are relative to the frozen hybrid. These diagnostics are not refitted candidates and were not used to replace the primary 21-day model.

Fixed-weight condition	Mean log score	Brier loss	Log gain
Combined; GNSS removed	-0.055404933	0.011078245	-0.000187776
Combined; faults permuted	-0.055232823	0.011071082	-0.000015665
Combined; 35-day GNSS lag	-0.054956126	0.011071173	0.000261032
Fault; assignment permuted	-0.055252461	0.011074962	-0.000035303
GNSS; GNSS removed	-0.055217158	0.011074926	-0.000000000
GNSS; 35-day lag	-0.054774632	0.011067559	0.000442526

Increasing GNSS lag to 35 days retains a positive combined-model gain of approximately 0.0002610. GNSS-only performs better under this longer lag in the exposed sample, with a gain of approximately 0.0004425. Selecting that result after seeing the evaluation would add another layer of model selection. The study preserves it as a sensitivity and retains the original primary comparison.

11. Discussion

11.1 What the experiment establishes

The experiment establishes that authentic fault and GNSS products can be connected to a small, bounded ETAS-neural extension and evaluated reproducibly with a fixed regional catalog. The preserved outputs support a slight average historical score improvement for the combined model. They also support an unfavorable conclusion about the more practically appealing interpretation: the extensions do not improve average log score in the weeks containing target events, and they do not reduce the daily false-elevation count relative to the frozen hybrid.

This is a useful scientific result because it narrows the claim. The added measurements are real; the code performs an actual fit; the probability calculation respects simulated clustering; and the negative diagnostics are retained. The resulting evidence is still too weak to establish reliable M5+ forecasting. Data authenticity and software correctness are prerequisites for an empirical claim, not substitutes for enough independent target events.

11.2 Why a small average gain can coexist with worse event windows

Proper scores reward calibrated probabilities across all outcomes. In a rare-event problem, lowering probabilities slightly across many quiet windows can outweigh a small deterioration in a few positive windows. That arithmetic is not a defect in proper scoring. It is a reason to report the distribution and source of the gain rather than describing the best average score as “better earthquake prediction” without qualification.

The combined-model decomposition is particularly informative: its quiet-window total gain exceeds its overall gain because the positive-window contribution is negative. The result could reflect a useful correction to the background rate, an adaptation to catalog observation changes, an indirect seasonal or processing signal in GNSS, or chance in a small exposed sample. The present experiment does not identify which explanation is responsible.

The asymmetric multiplier bounds add another interpretive constraint. The model can substantially lower intensity but can raise it by only about 4.2%. It is therefore better equipped to dampen a reference than to express a large new precursor-driven increase. This is not a reason to discard the negative result. It defines the model class for which the negative result is relevant and prevents overgeneralizing it to all geodetic or neural hypotheses.

11.3 Association, mechanism and identifiability

Fault proximity may correlate with catalog location density, network coverage and the fitted background rate. GNSS increments may correlate with season, station availability, common-mode noise and processing transitions. A neural branch can exploit these patterns without representing fault loading. The attention and convolution weights are predictive parameters, not mechanically identified quantities.

No elastic dislocation model maps slip on a reviewed fault surface to station displacement in this experiment. No joint inversion separates tectonic displacement from loading, reference-frame error and instrument motion. No full network covariance is available. The model therefore cannot identify a fault-coupling or slip posterior from the two station histories. Calling its GNSS features “strain” or its fault weights “stress” would exceed the information in the inputs and the structure of the model.

A physically richer future analysis would need fault geometry with depth and kinematics, station reference-frame and epoch consistency, a justified forward model, spatially and

temporally correlated errors, and identifiable parameter combinations. Even such an inversion would estimate deformation or coupling under assumptions; its connection to seven-day M5 occurrence would still require a separate forecast test. The present architecture is statistically constrained and geoscience-informed, but it is not a validated physical rupture simulator.

11.4 Comparison with the literature

The broad concept of combining statistical earthquake models with neural and geodetic information has direct prior examples [7, 8, 9, 10]. The present implementation differs in its small residual branches, native-M5 binary target, fixed GNSS latency and first-event simulation. These details make the work specific and reproducible. They do not independently establish novelty at the level of a scientific principle.

The results are compatible with the broader lesson that neural flexibility alone does not guarantee improvement over ETAS [12]. They also align qualitatively with the caution raised by a close Marmara preprint [14], although datasets, targets and methods differ and its results were not reproduced here. A formal comparative benchmark would need common input versions, target definitions, chronological splits and scoring code. Cross-paper score magnitudes cannot substitute for that experiment.

11.5 Why “reliability in percent” is not a supported summary

A forecast probability is not the percentage reliability of its model. Reliability in probabilistic forecasting concerns whether outcomes occur at the frequencies assigned to comparable forecasts. Discrimination concerns ranking conditions; sharpness concerns concentration away from climatology; decision value depends on actions and consequences. A single accuracy percentage can obscure all of these distinctions.

In this study, only three weekly windows are positive. Reporting the proportion of correctly labeled quiet weeks would mostly measure class imbalance. Reporting a percentage of events “predicted” would depend on an arbitrary threshold and on whether overlapping issue dates are counted repeatedly. Reporting 1.26% as reliability would confuse the mean predicted event probability with validation. The defensible report consists of proper scores, event support, calibration counts, dependence-aware uncertainty limitations and a clearly identified data vintage.

12. Limitations and threats to validity

12.1 Observation completeness and magnitude uncertainty

The model conditions on recorded M2+ events, but the catalog does not establish uniform detection at that threshold. Long-term network changes and short-term incompleteness can alter apparent seismicity in ways that a flexible model may absorb. The fitted mark distribution also assumes stable conditional magnitude frequencies. If reporting or magnitude selection changes over time, a constant q5 can be miscalibrated even when the total-rate model is useful.

Magnitude and location uncertainties are preserved in the observation archive but are not fully marginalized during forecasting. A deterministic boundary and threshold can move uncertain events into or out of the target. This is particularly important near M5 and the regional box edge. A future measurement model should distinguish reported magnitude scatter, agency calibration, missingness and latent physical magnitude. It should avoid treating a published interval or inter-agency range as an independent uniform posterior.

12.2 Revised data and historical availability

The earthquake source incorporates later revisions, the fault product postdates much of the test period and the acquired GNSS solutions were downloaded in 2026. The code can prevent the direct use of future epochs while still using products whose processing depends on later information. These are different forms of information control. Numerical causality tests address the first; they cannot establish the second without archived release timestamps and historical product versions.

The 21-day GNSS lag is not a verified publication-time history. Station-step metadata itself may be revised. A retrospective analysis using current best products can answer a useful scientific question about association under reconstructed histories, but it should not be labeled an as-issued replay. An archive-faithful operational backtest would need the actual contemporaneously available catalog and geodetic releases or a carefully validated release-delay reconstruction.

12.3 Limited independent targets and repeated development

Five native M5+ events provide little information about moderate-event calibration, and four occur in one short sequence. The nominal count of 252 weekly windows overstates the number of independent target situations. Historical outcomes were already examined during earlier project iterations. Although seed selection uses training loss, the broader choice to build these extensions is informed by prior research on the same system and sample.

Three variants, multiple years, two scores and several sensitivities create many opportunities to highlight a favorable result. The paper reports all of them, but disclosure alone does not remove selection effects. No multiplicity-adjusted confirmatory test or independent replication is claimed. The reported block intervals cannot account fully for the complete development history.

12.4 Reference quality and model misspecification

The branching parameter is at its cap, spatial cells are coarse, triggering mass is normalized into a closed box, and the mark law is fixed. The small number of optimization starts does not establish global optimality across model formulations. The reference is a disclosed fitted comparator, not a certified optimal ETAS implementation. Before a strong claim of neural improvement, scientific review should assess whether a more appropriate ETAS reference absorbs the apparent gain.

The neural penalty, clipping, finite history length, graph construction and fixed spatial pooling are additional assumptions. Missing GNSS windows may correlate with time and observation conditions. Uniform simulated locations within large cells can influence fault-distance features during forecast propagation. These choices define the experiment and should be subjected to future sensitivity tests on independent data, rather than tuned repeatedly against this historical sample.

12.5 Incomplete uncertainty propagation

Monte Carlo error bars quantify simulation integration only. They do not include uncertainty in ETAS fitting, neural weights, mark probabilities, fault geometry, event identity, magnitude selection, station coordinate processing or missing observations. The descriptive resampling interval captures variability in the realized paired score sequence under one block scheme. Neither is a comprehensive predictive uncertainty interval.

The experiment does not fit a full posterior over coupled seismic and geodetic states. A percentage displayed to several decimal places would therefore convey numerical precision without equivalent scientific precision. The manuscript uses extra digits only where needed to reconcile scores and demonstrate small differences; it does not treat those digits as a measure of forecast trustworthiness.

12.6 Generalization and independent review

The extension experiment is Marmara-only. California's role in the inherited base is training history, not a new external test. No independently selected strike-slip region, catalog release or target period validates the added modalities. The archived work also does not provide completed independent seismological review or prospective target outcomes under a frozen protocol. These missing forms of evidence are central limitations, not administrative details.

13. A falsifiable next-stage validation protocol

13.1 Freeze the estimand before adding capacity

The next study should first fix the region, magnitude definition, catalog provider and observation-time rules. A physical Mw ≥ 5 target requires a documented measurement model or adjudication process; a source-native threshold requires stable provider semantics and explicit reconciliation with revisions. The choice should precede evaluation and remain visible in every output.

The present model and a reviewed regional ETAS reference should then be frozen, including source versions, feature definitions, lags, simulation budgets and handling of unavailable data. Adding a larger Transformer, more graph layers or more stations should be treated as a separate candidate. Increasing architectural complexity before securing independent evaluation would increase model-selection freedom faster than it increases evidence.

13.2 Independent regional replication

A comparison region should be selected through documented tectonic and observational criteria rather than because preliminary scores look favorable. Relevant criteria include strike-slip character, catalog duration, known observation changes, moderate-event support, station coverage and access to historical data versions. Similar tectonics do not guarantee interchangeable detection thresholds or magnitude scales.

The experimental unit should include complete earthquake sequences when defining training and evaluation exposure. Random event splitting is inappropriate because a later aftershock can enter fitting while its precursor or mainshock enters evaluation. Regional transfer should be tested with normalization and calibration fitted only in permitted development data. A leave-region-out design can assess generalization if all tuning decisions remain outside the final region.

13.3 Prospective shadow forecasts

A genuinely prospective study requires the forecast artifact to exist before its outcome. Each issuance should record the probability, region, horizon, magnitude target, model version, permitted input versions, source arrival times and any missing-data status. An independently verifiable archive should bind the forecast to its issue time. A timestamped source record alone is not a timestamped forecast.

The service should continue issuing on a predefined schedule through quiet periods, sequences and data outages. Skipped or withheld forecasts should remain part of the exposure ledger. Outcomes should be assigned with a predetermined delay and revision policy. A reviewer should be able to reproduce which observations were available, what probability was issued and how the final score was calculated.

This manuscript proposes that protocol; it does not claim that it has been completed. Source records and website availability do not by themselves establish prospective scientific operation. The preserved research records remain marked as unissued and unaccepted, with zero independently accepted scientific forecasts in the reported experiment.

13.4 Decision criteria and stopping

A confirmatory comparison should specify its primary proper score, paired reference, secondary calibration criteria, target exposure and treatment of repeated evaluation in advance. It should preserve sequence dependence in uncertainty calculations and account for repeated looks if results are monitored over time. The number of independent moderate-event situations needed cannot be inferred reliably from five clustered events alone.

Design-stage simulation can explore power under explicit plausible alternatives, but the simulation's assumptions should be reported rather than promoted into evidence for the model. A scientifically credible promotion criterion would require sustained proper-score improvement, no material calibration deterioration, documented observation quality and independent review. A single favorable year, one high-probability sequence or a selected threshold should not determine the conclusion.

13.5 Measurement-model priorities

Before a new geodetic inversion is attempted, the project should acquire documented cross-station and temporal covariance or a scientifically defensible error model, consistent reference frames, epoch metadata and correction histories. Additional stations should improve identifiable deformation modes rather than merely add correlated inputs. Offshore travel times require sound-speed, clock, instrument-motion and geometry corrections before being interpreted as displacement; raw ranging values are not fault slip rates.

For the earthquake catalog, the next priority is an observation model that separates occurrence from detection and magnitude reporting. A hierarchical formulation could represent a latent event and its agency-specific observations, with missingness and time-varying detection explicitly modeled. This is a proposed direction, not an implemented result. It should be judged by whether it changes calibrated out-of-sample target probabilities, not by whether it produces a more elaborate latent representation.

14. Research evidence and public implementation

A public implementation of the combined model became available on 12 October 2026. Website availability is separate from the research experiment: it does not alter the historical sample, confer independent scientific approval or convert reconstructed forecasts into prospective evidence.

The public implementation consumes an AFAD-derived live history, whereas the reported historical experiment uses reconciled TURHEC observations. This provider transfer can change magnitude semantics, detection and revision behavior. Local cross-language agreement

demonstrates that the implementation can reproduce specified computations; it does not validate transfer of scientific calibration from one observation process to another.

The paper consequently does not present a current public probability as an empirical result or an imminent-warning claim. Research acceptance remains distinct from public display. Any public description of the model should preserve the target definition, experimental status, data limitations and absence of proven M5+ reliability. No M7 probability is inferred from this model, and no long-term historical estimate is relabeled as a rolling contemporary probability.

15. Conclusions

A bounded fault/GNSS extension to a frozen ETAS-neural model was fitted and evaluated on a reconciled Marmara catalog with native M2+ inputs and a seven-day native M5+ occurrence target. Authentic fault geometry and GNSS coordinate products were acquired, transformed with explicit masks and lags, and connected to a reproducible simulation and scoring pipeline. The combined extension adds 297 parameters and achieves a small positive mean retrospective log-score gain over both disclosed references.

That aggregate improvement does not establish improved anticipation of the target earthquakes. All extensions worsen mean log score in the three positive weekly windows, the combined-model block interval includes zero, and its 17 daily false-elevation dates equal the frozen hybrid's count. Five target events, clustered into three positive weeks, cannot support a stable claim of probability calibration or a meaningful single reliability percentage.

The defensible contribution is an auditable experiment with explicit constraints, a distinct implementation of established ideas and retained negative evidence. Independent catalog-aware replication and prospective shadow evaluation are required before any claim of reliable incremental M5+ forecasting. Fault-specific M7+ regional 1-, 5-, 10- and 30-year occurrence probabilities remain **not estimable from the verified inputs and model in this study**.

Data and author statements

Data availability. Primary catalog, fault and GNSS sources are identified in the references and Appendix B. This report contains derived figures and forecast-score tables. Upstream data rights remain with their respective providers. Historical source-availability times, full geodetic temporal/network covariance and a calibrated physical-Mw posterior are unavailable.

Reproducibility. The report describes the model, data preparation, simulation settings and evaluation methods. The full research implementation and underlying analysis records are not included in this public report; unrestricted external replication has not been established.

Organizational author. Alderbeam, Inc. is the organizational author of this company research report. The cited data providers and researchers are not represented as authors or endorsers of the report.

Company context. This report is issued by Alderbeam, Inc. about its DepremİstanbuI research project. It presents the company's own model evaluation. Publication as a company report does not constitute independent endorsement of the model or its forecasts.

Report status. This is a company research report. Independent seismological review and confirmatory prospective validation have not been completed. The proposed future protocol in Section 13 is a proposal, not a report of work already performed.

Appendix A. Implementation specification

A.1 Spatial allocation and fixed parameters

The nine normalized background weights in cell order are 0.0767165784, 0.1724237984, 0.2199900096, 0.1255935219, 0.2406003159, 0.0964602862, 0.0060593340, 0.0477619824 and 0.0143941732. The simulation mark law is the empirical four-bin distribution in Table 2. It should not be replaced by an assumed continuous Gutenberg-Richter law when reproducing the reported probabilities.

The frozen base has 625 neural parameters. The fault branch contains an affine 9-to-8 embedding, bias-free query and key 8-to-4 projections, a bias-free 8-to-8 value projection and an affine 8-to-1 output, totaling 217. The GNSS branch contains a 6-to-4, width-3 convolution with bias and a bias-free 4-to-1 output, totaling 80. Fixed graph masks, pooling matrices and normalizers are buffers rather than fitted parameters in these counts.

A.2 GNSS normalization and step rules

The six training means are approximately −0.434518, 0.259457, 0.590759, 0.696335, −0.046617 and 0.987182. The corresponding scales are 3.110116, 2.051366, 0.055478, 0.063231, 0.038724 and 0.049124. They apply in order to east increment, north increment, transformed east uncertainty, transformed north uncertainty, east-north correlation and coverage. All six are fitted using valid pre-cutoff windows; the validity mask is not standardized.

Instrument-step days are expressed relative to 1 January 1990. ISTA uses days 3,632 and 9,625; TUBI uses 3,293, 4,538, 4,615, 4,620, 9,529, 9,542 and 9,656. These recorded days and daily antenna-height changes restrict feature construction. The actual source metadata, rather than rounded calendar dates printed in a narrative, governs a reproducible run.

A.3 Optimization records

Table A1. Best penalized training objective for each seed. These are training losses, not target-event scores. All selected seed-47 models are at epoch 200.

Variant	Seed 11	Seed 29	Seed 47, selected
Fault	−0.0167762	−0.0167688	−0.0169464
GNSS	−0.0167999	−0.0167902	−0.0171608
Combined	−0.0180170	−0.0179953	−0.0182482

A.4 Exposure and simulation ledger

The evaluation ledger contains nine primary forecast series: weekly, daily and event probes for each of three extensions. Six additional weekly series contain fixed-weight sensitivities. Across all 15 forecast series there are 7,560 forecast rows. The manuscript audit includes initial low-draw batches retained before daily refinement, giving 15,162 batch summaries. The batch count is not a count of independent earthquake outcomes.

The archived exact path replay uses five selected combined-model weekly issues: 1 January 2014, 21 May 2014, 1 February 2017, 8 February 2017 and 24 October 2018. These checks cover first and last issues, positive windows and the peak-probability region. The manuscript

arithmetic audit independently recomputes saved samples but does not claim to be a new complete path replay or refit.

Appendix B. Data sources and attribution

Table B1. Principal data sources. Publication versions and access dates identify the inputs used in this retrospective study.

Input	Product or publication version	Source date	Reference
TURHEC catalog	Public supplementary table	28 May 2021	[2]
GEM / EMME geometry	Global Active Faults 2019.0	2019 release	[3]
ISTA coordinates	NGL IGS20 daily tenv3	Accessed 12 October 2026	[4, 5, 6]
TUBI coordinates	NGL IGS20 daily tenv3	Accessed 12 October 2026	[4, 5, 6]

TURHEC is attributed to Tan and its public supplementary release [2]. Fault geometry is attributed to the GEM Global Active Faults Database and the underlying EMME contributions [3]. The map and graph are derived products; conservative CC BY-SA 4.0 treatment is retained for the geometry-derived material because the published licensing statements differ. This treatment does not relicense unrelated portions of the report or the underlying third-party data.

GNSS processing is attributed to the Nevada Geodetic Laboratory and relevant station providers [4, 5, 6]. The original ISTA and TUBI coordinate series remain with their providers and are not redistributed in this report. Source identification does not imply unrestricted redistribution rights.

Appendix C. Target-event identity and temporal interpretation

Table C1. Native-M5 target observations. “Prior at issue” means a native M5+ in the 30 days before the exact seven-day lead-time probe; “prior at event” uses the event's own preceding 30 days.

ISC event ID	Origin UTC	Native Mx	Prior at issue	Prior at event
604653419	2014-05-24T09:31:18.08	5.3	No	No
610246604	2017-02-06T03:51:40.04	5.3	No	No
610197961	2017-02-06T10:58:02.24	5.2	No	Yes
610198715	2017-02-07T02:24:04.19	5.3	No	Yes
610208516	2017-02-12T13:48:15.64	5.3	No	Yes

The first listed event and the first February 2017 event lack a preceding native-M5 event under this rule. The remaining three have a predecessor by event time, but none has one at the seven-day-earlier issue. These labels do not imply that the first two are causally isolated from all smaller-event activity, nor that the later three are proven aftershocks of a particular rupture. They support a transparent sequence-aware audit without inventing causal event assignments.

The exact-lead-time probes should not be pooled into the weekly scoring sample. They are selected using observed event times, and the target event is on the excluded right boundary of the half-open window. A different event-centered scoring convention could be defined, but it would require its own prospective specification and treatment of selection. The primary score remains the complete nonoverlapping weekly exposure.

Appendix D. Verification and interpretation

The report presents five figures, probability and score tables, the experimental protocol and source references. Verification reconciled catalog identities and outcomes, recomputed simulation-sample summaries and checked the reported scores. These computational checks support numerical consistency; they do not constitute independent scientific validation.

The evidence hierarchy is deliberately explicit. Source documentation establishes data origin and processing context. Numerical tests establish specified computational properties. Arithmetic recomputation establishes agreement between saved forecasts and reported scores. Historical evaluation establishes behavior on the exposed reconstructed sample. Independent regional tests and prospective evaluation would provide stronger evidence of generalization and operational probability quality.

No new model training or selection was performed to prepare this report. The preparation work consisted of literature verification, source inspection, separate arithmetic checks, descriptive decomposition, figure generation and writing. The analysis review was completed on 12 October 2026. Scientific reliability has not been established, and confirmatory prospective validation has not been performed.

References

[1] Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association, 83(401), 9-27. Source

[2] Tan, O. (2021). A homogeneous earthquake catalogue for Turkey. Natural Hazards and Earth System Sciences, 21, 2059-2073. Public supplement and catalog documentation. Source

[3] Styron, R., and Pagani, M. (2020). The GEM Global Active Faults Database. Earthquake Spectra, 36(1S), 160-180. Source

[4] Blewitt, G., Hammond, W. C., and Kreemer, C. (2018). Harnessing the GPS data explosion for interdisciplinary science. Eos, 99. Source

[5] Blewitt, G. (2024). An improved equation of latitude and a global system of graticule distance coordinates. Journal of Geodesy, 98. Source

[6] Nevada Geodetic Laboratory (accessed 12 October 2026). IGS20 time-series processing documentation and tenv3 format specification. Documentation for the daily ISTA and TUBI coordinate products. Source

[7] González Fuentes, A., Nicolis, O., Peralta, B., and Chiodi, M. (2022). Spatio-temporal seismicity prediction in Chile using a multi-column ConvLSTM. IEEE Access, 10, 107402-107415. Official institutional abstract inspected; publisher full-text retrieval unavailable in this review. Source

[8] Zhang, H., Ke, S., Liu, W., and Zhang, Y. (2024). A combining earthquake forecasting model between deep learning and epidemic-type aftershock sequence (ETAS) model. Geophysical Journal International, 239(3), 1545-1556. Source

[9] Tong, Y., Hong, R., Zhang, Z., Aihara, K., Chen, P., Liu, R., and Chen, L. (2023). Earthquake alerting based on spatial geodetic data by spatiotemporal information transformation learning. Proceedings of the National Academy of Sciences, 120(37), e2302275120. Source

[10] Kaya, K., Sahin, M., Mangir, F., Beydemir, A. B., Yaltirak, C., and Oguducu, S. G. (2026). Deep Quake Dynamics: A multimodal fault-aware approach to earthquake magnitude and occurrence time forecasting. Geoscience Data Journal, 13(2), e70066. First published 16 March 2026. Source

[11] Dascher-Cousineau, K., Shchur, O., Brodsky, E. E., and Günnemann, S. (2023). Using deep learning for flexible and scalable earthquake forecasting. Geophysical Research Letters. Source

[12] Stockman, S., Lawson, D., and Werner, M. (2026). EarthquakeNPP: A benchmark for earthquake forecasting with neural point processes. arXiv:2410.08226v3, revised 10 March 2026; record reports acceptance to Transactions on Machine Learning Research. Source

[13] Mizrahi, L., et al. (2024). Developing, testing, and communicating earthquake forecasts: Current practices and future directions. Reviews of Geophysics, 62(3). Source

[14] Alhan, B. K., and Khabat, K. (2026). Machine learning versus ETAS for earthquake forecasting in the Sea of Marmara: A leakage-audited negative result and a closed-form scoring artifact. EarthArXiv preprint, version 4, updated 24 July 2026. Non-peer-reviewed; numerical claims not reproduced here. Source

[15] Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378. Source

[16] Cho, K., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP 2014; arXiv:1406.1078v3. Source

[17] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. ICLR 2018; arXiv:1710.10903v3. Source

[18] Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. ICLR 2019; arXiv:1711.05101v3. Source

[19] pyCSEP contributors (accessed 12 October 2026). Evaluations. pyCSEP v0.7.0 documentation. Gridded and catalog-based forecast evaluation methods. Source

[20] Southern California Earthquake Data Center (2013). Southern California Earthquake Data Center. California Institute of Technology. Data-center attribution for inherited California training inputs. Source

[21] California Institute of Technology and United States Geological Survey Pasadena (1926). Southern California Seismic Network. International Federation of Digital Seismograph Networks. Network attribution for inherited California inputs. Source