MagST-PP: Magnitude-Conditioned Spatio-Temporal Point Processes for Earthquake Forecasting

Hao Lin
Mira Costa High School
Manhattan Beach, California, USA
hl_092@usc.edu

Yan Liu
University of Southern California
Los Angeles, California, USA
yanliu.cs@usc.edu

Abstract

Probabilistic earthquake forecasting is commonly framed as a spatio-temporal point-process problem, yet hazard-relevant evaluation centers on rare moderate-to-large events rather than the small events that dominate catalog likelihoods—a mismatch we call the likelihood trap. Classical epidemic-type aftershock sequence (ETAS) models capture space-time clustering via a self-exciting mechanism but treat magnitudes through a separable Gutenberg–Richter component, which does not directly optimize short-window, location-specific exceedance queries. Many neural temporal point processes likewise either ignore continuous space or reduce locations to categorical marks. We introduce MagST-PP (Magnitude-Conditioned Spatio-Temporal Point Process), a continuous neural framework with a 1-D Gaussian mixture decoder for next-event inter-arrival time and a 2-D Gaussian mixture decoder for next-event location, both conditioned on recent seismic history. A tail-weighted magnitude head and an autoregressive Monte Carlo sampler convert these event-level densities into fixed-window exceedance scores such as "at least one $M > 4.5$ event within 7 days and 20 km." On an observed California–Nevada catalog, MagST-PP improves held-out spatial log-likelihood over ETAS and a neural spatio-temporal point-process (NSTPP) baseline while reducing training time by approximately 36× and inference time by approximately 75×. For the primary 7-day / 20-km / $M > 4.5$ exceedance query under a chronological 1980–2008 / 2008–2016 / 2016–2024 split, MagST-PP achieves AUC-ROC 0.79, compared with 0.503 for a random baseline and 0.73 for a coordinate-only background forecast. Ablations show that flexible event-count context windows improve temporal likelihood, while structured cluster and mainshock labels outperform naive fault-distance regularization. Space-time histories alone also contain recoverable information about large-event occurrence.

CCS Concepts

• **Information systems → Data mining**; • **Computing methodologies → Neural networks**; • **Mathematics of computing → Stochastic processes**.

Keywords

spatio-temporal point processes, earthquake forecasting, Gaussian mixture models, magnitude-conditioned forecasting, irregular time series

MiLeTS@KDD '26, Jeju, Republic of Korea
2026.

ACM Reference Format:
Hao Lin and Yan Liu. 2026. MagST-PP: Magnitude-Conditioned Spatio-Temporal Point Processes for Earthquake Forecasting. In *MiLeTS@KDD '26, August 10, 2026, Jeju, Republic of Korea.* ACM, New York, NY, USA, 7 pages.

1 Introduction

Earthquake catalogs are irregular, nonstationary event streams localized in continuous time and space. A scientifically useful model is probabilistic: it estimates where, when, and at what magnitude future seismicity may occur, without making deterministic claims that exceed available evidence.

Statistical seismology has long used point-process models for this purpose. The ETAS (epidemic-type aftershock sequence) family is especially influential: it represents earthquakes as a self-exciting process in which each event raises the local rate of later events, reproducing the clustered space-time patterns seen in real catalogs [10, 11]. Magnitudes are typically handled separately through Gutenberg–Richter (GR) statistics [5]. We do not dispute the validity of this long-run frequency–magnitude relationship; our target is narrower—whether recent, local space-time activity contains useful information for short-window exceedance screening under a fixed magnitude threshold.

This separation creates what we call a likelihood trap. Event-level log-likelihood is dominated by the most frequently observed part of a catalog, while hazard-oriented screening asks a different empirical question: for a specified area, horizon, and magnitude threshold, does recent seismic activity change the risk ranking relative to a static background? We therefore keep likelihood evaluation for point-process fit, but also evaluate short-window exceedance queries that better match the way practical screening questions are posed. Earthquake forecasting is also a natural MiLeTS problem: the data are irregularly sampled, spatially localized, and strongly nonstationary because local triggering patterns evolve over time.

Recent neural temporal point processes have improved flexibility by learning history-dependent event distributions [4, 8, 12, 14]. Earthquake-specific systems such as RECAST [3] demonstrate that deep point-process methods can scale to seismicity modeling. However, many neural formulations focus on time-only forecasting, discretize space into categorical marks, or do not directly answer practical area-time-magnitude queries. In earthquake forecasting this is costly: relative distances and local clustering are central to the process.

Table 1: High-level positioning of related model families. MagST-PP keeps space continuous and evaluates magnitude-conditioned area-time queries; entries summarize the formulation used in this paper rather than all possible extensions.

Model	Cont. space	Hist.-cond. M	Area-time Q
ETAS	✓	–	partial
AMDN-style TPP	–	optional	–
RECAST	partial	–	partial
MAGNET	queried	✓	partial
MagST-PP	✓	✓	✓

We present MagST-PP, a continuous neural spatio-temporal point process for earthquake forecasting. The model adapts the mixture-density idea behind AMDN-HAGE [13] from social-media event modeling to seismicity, replacing discrete spatial marks with a continuous 2-D spatial density. It first learns a probabilistic next-event model for time and location, then uses a magnitude-aware head and autoregressive sampling to answer fixed-window exceedance queries (Figure 1). **Our contributions are threefold:**

- We formulate earthquake forecasting for irregular, nonstationary event streams as continuous neural spatio-temporal point-process learning with separate Gaussian-mixture decoders for next-event time and 2-D location.
- We connect event-density learning to magnitude-aware area-time queries through a magnitude head and an autoregressive Monte Carlo sampler.
- We provide a broad retrospective evaluation covering held-out likelihood, computational efficiency, context-window ablations, domain-supervision comparisons, magnitude-tail reweighting, space-time history diagnostics, fixed-window AUC-ROC, radius/horizon sensitivity, and qualitative forecast maps.

2 Related Work

Statistical earthquake point processes. ETAS models represent seismicity as a self-exciting point process closely related to Hawkes processes [7]. ETAS remains valuable because its parameters encode interpretable triggering behavior, and extensions support residual analysis, declustering, and seismic productivity studies [16, 17]. For the present task, the issue is not that GR statistics are invalid, but that the separable ETAS formulation does not train directly for short-window, location-specific exceedance queries.

Neural temporal and spatio-temporal point processes. Neural point processes learn conditional event distributions using recurrent, attention, flow, ODE, or Transformer-style encoders [4, 8, 9, 12, 14]. Continuous spatio-temporal extensions model events jointly in time and space [2], though scaling to large catalogs remains challenging. RECAST [3] demonstrated flexible, scalable seismicity modeling. Importantly, recent benchmarks show that generic neural point-process implementations do not automatically outperform ETAS on spatio-temporal likelihood [15]; we therefore do not claim universal superiority and focus on continuous spatial density and magnitude-aware short-horizon queries.

Magnitude-aware earthquake modeling. Recent work has begun to test whether earthquake catalogs encode magnitude-relevant information beyond a global frequency–magnitude baseline. MAGNET [1] forecasts the magnitude distribution of a future event conditioned on queried space-time coordinates. MagST-PP is complementary: it first models the joint distribution of future event times and locations, then uses a magnitude-conditioned exceedance module and autoregressive sampling for area-time-window forecasts where future coordinates are not known in advance. Table 1 summarizes the relationships.

3 Problem Definition

Let an earthquake catalog be the ordered sequence $\mathcal{S} = \{e_i\}_{i=1}^N$, where $e_i = (t_i, \mathbf{s}_i, m_i)$ contains origin time t_i, two-dimensional projected location $\mathbf{s}_i = (x_i, y_i) \in \mathbb{R}^2$, and magnitude m_i. Given the causal history before event i,

$$\mathcal{H}_i = \{(t_j, \mathbf{s}_j, m_j) : j < i\}, \tag{1}$$

we model the conditional next-event distribution as

$$p(t_i, \mathbf{s}_i \mid \mathcal{H}_i) = p(\Delta t_i \mid \mathcal{H}_i)\, p(\mathbf{s}_i \mid \mathcal{H}_i, \Delta t_i), \tag{2}$$

where $\Delta t_i = t_i - t_{i-1} > 0$ is the inter-arrival time.

The practical forecasting query is not just the next event. For query time t, location $\mathbf{q}$, horizon ΔT, radius Δr, and magnitude threshold M_0, define

$$\begin{aligned} Y(t, \mathbf{q}) = 1\{\exists e : t_e \in [t, t + \Delta T],\ \|\mathbf{s}_e - \mathbf{q}\| \le \Delta r, \\ m_e > M_0\}. \end{aligned} \tag{3}$$

The main binary task is to estimate $P(Y = 1 \mid \mathcal{H}_t)$. Our primary setting is $\Delta T = 7$ days, $\Delta r = 20$ km, and $M_0 = 4.5$.

Terminology. Magnitude-conditioned means the exceedance query is parameterized by M_0 and sampled event trajectories are scored by a magnitude module; it does not assert a deterministic relationship between past seismicity and future magnitudes. In this implementation, the magnitude module is a tail-weighted point-estimate scorer rather than a full conditional magnitude distribution, so the fixed-window output is evaluated as a retrospective ranking score rather than an operationally calibrated probability.

4 Method

4.1 History Encoder

For each target event or forecast query, MagST-PP encodes a causal context window—no future events are ever visible. A flexible event-count window selects a fixed number of the most recent events rather than a fixed time gap, preventing underrepresentation during seismically quiet periods (validated in Section 5.4).

Input features are normalized inter-event time and projected spatial coordinates. Magnitudes and potency features may be appended for ablation studies, but our experiments show that the space-time pattern alone encodes useful magnitude-related structure (Section 5.7). Let $\mathbf{h}_i = f_\theta(\mathcal{H}_i)$ be the history embedding, produced by a causal LSTM over the irregular, chronologically ordered event sequence. Replacing the LSTM with attention or state-space encoders is a natural extension, but not evaluated here.

4.2 Temporal and Spatial Mixture Decoders

Temporal decoder. A K_t-component 1-D Gaussian mixture over normalized inter-arrival time:

$$p_\theta(\Delta t_i \mid \mathcal{H}_i) = \sum_{k=1}^{K_t} \pi_{ik}^{(t)}\, \mathcal{N}\Big(\Delta t_i; \mu_{ik}^{(t)}, (\sigma_{ik}^{(t)})^2\Big). \tag{4}$$

Because inter-arrival times are positive, autoregressive simulation uses rejection resampling from the temporal GMM until $\Delta t > 0$. The corresponding positive-truncated density is $p_\theta^+(\Delta t \mid \mathcal{H}_i) = p_\theta(\Delta t \mid \mathcal{H}_i)/Z_i^+$, where Z_i^+ is the GMM probability mass on the positive half-line. The reported temporal LL values are implementation-level log scores on observed positive increments under the same convention across all variants, rather than strict normalized likelihoods over $\mathbb{R}_+$.

Spatial decoder. A K_s-component 2-D Gaussian mixture over projected coordinates:

$$p_\theta(\mathbf{s}_i \mid \mathcal{H}_i, \Delta t_i) = \sum_{k=1}^{K_s} \pi_{ik}^{(s)}\, \mathcal{N}_2\Big(\mathbf{s}_i; \boldsymbol{\mu}_{ik}^{(s)}, \boldsymbol{\Sigma}_{ik}^{(s)}\Big). \tag{5}$$

Locations remain continuous: relative distances and clustering structure are preserved, unlike mark-based formulations that bin locations into grid cells or category IDs. The training objective is

$$\mathcal{L}_{\text{st}} = -\sum_i \left[\log p_\theta(\Delta t_i \mid \mathcal{H}_i) + \log p_\theta(\mathbf{s}_i \mid \mathcal{H}_i, \Delta t_i)\right]. \tag{6}$$

The number of mixture components K_t and K_s is treated as a validation-selected hyperparameter.

4.3 Magnitude-Aware Forecasting

Magnitude regression with tail reweighting. Because small events dominate earthquake catalogs, a standard MAE loss underweights the high-magnitude tail. The current magnitude head predicts a point estimate $\hat{m}_i = g_\phi(\mathbf{h}_i, \Delta t_i, \mathbf{s}_i)$ and is trained with

$$\mathcal{L}_m = \sum_i w(m_i)\, \ell(\hat{m}_i, m_i), \tag{7}$$

where $w(m) = 1$ (normal), $w(m) = m$ (linear), or $w(m) = \exp(m)$ (exponential). Exponential reweighting improves MAE on rare high-magnitude subsets at the cost of all-event MAE (Section 5.6).

Fixed-window scoring via Monte Carlo sampling. MagST-PP samples B future event sequences autoregressively and

Algorithm 1 Autoregressive Monte Carlo fixed-window query

Require: history $\mathcal{H}_t$, center $\mathbf{q}$, horizon ΔT, radius Δr, threshold M_0, sample count B
for $b = 1, \ldots, B$ **do**
  $\mathcal{H}_{\text{local}} \leftarrow \mathcal{H}_t$; $\tau \leftarrow 0$; $c_b \leftarrow 0$
  while $\tau < \Delta T$ **do**
    sample $\Delta t \sim p_\theta(\Delta t \mid \mathcal{H}_{\text{local}})$; then $\hat{\mathbf{s}} \sim p_\theta(\mathbf{s} \mid \mathcal{H}_{\text{local}}, \Delta t)$
    $\tau \leftarrow \tau + \Delta t$
    if $\tau \le \Delta T$ **then**
      compute $\hat{m} = g_\phi(f_\theta(\mathcal{H}_{\text{local}}), \Delta t, \hat{\mathbf{s}})$ via the tail-weighted magnitude head
      append $(t + \tau, \hat{\mathbf{s}}, \hat{m})$ to $\mathcal{H}_{\text{local}}$
      if $\|\hat{\mathbf{s}} - \mathbf{q}\| \le \Delta r$ and $\hat{m} > M_0$ **then**
        $c_b \leftarrow 1$
      end if
    end if
  end while
end for

return $\widehat{P}(Y=1) = B^{-1} \sum_{b=1}^{B} c_b$

Table 2: Data and task construction. Splits are chronological when time ranges are given.

Setting	Construction / purpose	Metric
ETAS synthetic	train/val/test 2920/973/973; sanity check	LL
Hauksson	S.CA $M > 2$; 4720/1573/1573; fault-distance aux.	Time LL
S.CA 1932–2024	$M > 2$; 7953/2651/2651; visualization	maps
CA–NV (main)	$M \ge 2.5$; 1980–08/08–16/16–24	LL, MAE, AUC

estimates

$$\widehat{P}(Y=1) = \frac{1}{B} \sum_{b=1}^{B} 1\{\text{sample } b \text{ satisfies the query}\}. \tag{8}$$

This estimator marginalizes over sampled event times and locations, but in the current implementation thresholds a point-estimated magnitude. We therefore use the resulting values primarily for AUC-based retrospective ranking rather than calibrated probability evaluation. The procedure is given in Algorithm 1.

4.4 Domain-Informed Auxiliary Supervision

When domain labels are available, the total training loss is

$$\mathcal{L} = \mathcal{L}_{\text{st}} + \lambda_m \mathcal{L}_m + \lambda_d \mathcal{L}_{\text{domain}}, \tag{9}$$

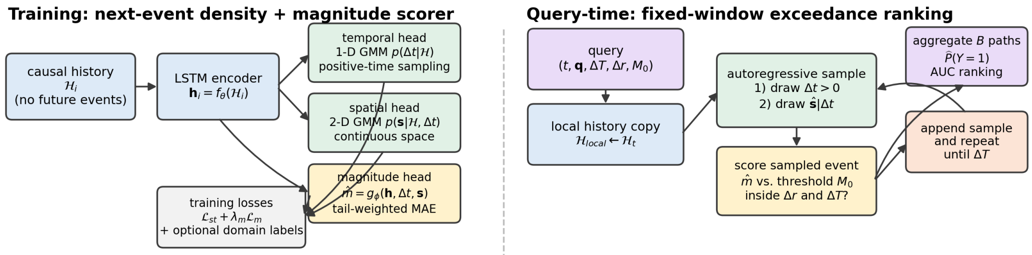

Figure 1: MagST-PP workflow. The diagram separates training-time event-density learning from query-time autoregressive Monte Carlo: a causal encoder feeds temporal and spatial GMM heads, a tail-weighted magnitude scorer evaluates sampled events against M_0, and repeated samples produce a fixed-window exceedance ranking score.

Table 3: ETAS synthetic catalog: test temporal log-likelihood. Higher is better. Both models are compared under the same RECAST-protocol ETAS synthetic benchmark.

Model	Test time LL
RECAST [3]	-4.891
MagST-PP	-4.411

with λ weights tuned on validation data. We test two auxiliary signals: fault-distance supervision (regression for shortest distance to mapped faults) and sequence-structure supervision (cross-entropy for family, mainshock, or cluster labels from seismic declustering catalogs). All auxiliary labels are training-only signals; no future labels are visible during inference or sampling.

5 Experiments

5.1 Data, Tasks, and Evaluation Protocol

We use three observed California–Nevada earthquake datasets: (1) a western U.S. catalog used in seismic-productivity analysis [16]; (2) the Hauksson et al. [6] southern California relocated catalog for fault-distance supervision; and (3) a 1932–2024 southern California catalog for qualitative visualization. Event coordinates are projected into a two-dimensional metric system.

For the main CA–NV evaluation, events are restricted to $M \geq M_c = 2.5$ to reduce sensitivity to network incompleteness. All experiments use chronological splits to prevent leakage across clustered sequences. The fixed-window exceedance task uses 1980–2008 (train), 2008–2016 (validation), and 2016–2024 (test); every prediction uses only history available before the query time. Table 2 summarizes all settings.

Table 4: Conditional log-likelihood on the CA–NV catalog. Higher is better. ETAS parameters are fit by MLE on the combined training/validation fitting sequence; hence separate validation values are omitted.

Model	Val t	Val s	Test t	Test s
ETAS	–	–	0.82	-8.29
NSTPP	16.89	-1.72	17.07	-1.41
MagST-PP	16.52	-0.92	22.09	0.75

Table 5: Computational efficiency on the CA–NV catalog.

Method	Training / epoch	Inference
NSTPP	161.2 s	120.8 s
MagST-PP	4.5 s	1.6 s

5.2 ETAS Synthetic Catalog: Baseline Compatibility

To verify that MagST-PP's point-process backbone is competitive with established neural seismicity models, we evaluate on ETAS synthetic catalogs following the RECAST preprocessing protocol [3]: a train/validation/test split of 2920/973/973 sequences. Table 3 compares temporal log-likelihood with RECAST. MagST-PP achieves -4.411 vs. RECAST's -4.891, confirming that the GMM-based temporal decoder matches or exceeds the Weibull mixture used by RECAST on synthetic data. This result provides a sanity check before evaluating on the more challenging observed catalog.

5.3 Spatio-Temporal Likelihood and Efficiency

Table 4 reports held-out conditional log-likelihoods on the observed CA–NV catalog. ETAS provides an interpretable baseline. In this setup, it has lower held-out spatial likelihood, consistent with the limited ability of its globally fitted

Table 6: Window-ablation temporal score in a separate normalized ablation pipeline. Higher is better. Values are comparable within this table only, not to Table 4.

Model variant	Val	Test
MagST-PP, fixed time-gap window	1.12	1.09
MagST-PP, flexible window	2.04	1.97
+ magnitude input	1.79	2.01
+ potency input	1.95	1.93

Table 7: Domain-informed auxiliary supervision. Higher time LL is better. Values are comparable only within each dataset/setting.

Dataset	Auxiliary signal	Time LL
Hauksson	none	-3.1753
Hauksson	fault distance	-3.1710
Hauksson	$M > 4.5$ fault distance	-3.2307
CA–NV	none	-5.7161
CA–NV	family ID	-5.6768
CA–NV	mainshock label	-5.5198
CA–NV	cluster label	-5.5011
CA–NV	$M > 4.5$ fault distance	-6.0373

parametric spatial kernel to adapt to heterogeneous local clustering structure. NSTPP substantially improves spatial likelihood. MagST-PP further improves both test temporal and test spatial log-likelihood, achieving the strongest spatial result among the three models.

Table 5 reports computational efficiency. MagST-PP is approximately 36× faster per training epoch and approximately 75× faster at inference than NSTPP. This matters because the Monte Carlo exceedance estimator requires many sampled sequences per query. The efficiency gain arises because the temporal and spatial heads are explicit finite Gaussian-mixture densities: density evaluation, mixture-component sampling, and coordinate sampling can be vectorized across queries and Monte Carlo paths, avoiding the numerical integration loops of some neural-ODE STPP formulations [2]. We do not claim operational real-time forecasting; rather, the observed speedups make repeated retrospective Monte Carlo exceedance queries computationally practical.

5.4 Flexible Context Windows and Magnitude Inputs

Table 6 compares context-window variants. The flexible event-count window improves temporal likelihood over the fixed time-gap window on both validation and test. Adding magnitude or potency directly as input features does not reliably improve temporal likelihood—magnitude slightly helps on test but hurts on validation, while potency hurts on test. This motivates a separate magnitude-aware head rather than naive feature appending; the underlying reason is investigated in Section 5.7.

Table 8: Magnitude prediction MAE. Lower is better.

Weighting	All	$M > 4$	$M > 5$	$M > 6$
None	0.32	0.97	2.02	3.05
Linear	0.32	0.99	2.05	3.04
Exponential	0.43	0.73	1.71	2.66

Table 9: Magnitude information latent in space-time histories. AUC-ROC reported.

Setting	1yr/Val	2yr/Test	3yr	4yr
Space-time only (no mag. input)	0.68	0.72	–	–
Direct transfer, train 2006–15	0.72	0.59	0.39	0.27
Rolling forecast, 1996–15	0.62	0.55	0.51	0.50
Rolling + future spatial	0.72	0.68	0.65	0.62

5.5 Domain-Informed Auxiliary Supervision

Table 7 reports auxiliary-supervision results. On the Hauksson catalog, fault-distance regression yields only a marginal improvement. On the CA–NV label catalog, family, mainshock, and cluster supervision all improve time likelihood over the no-auxiliary baseline, with cluster labels performing best. Notably, fault-distance regularization hurts after magnitude filtering ($M > 4.5$), indicating that domain knowledge must be structurally aligned with the prediction target—appending geologic features to the loss without regard for the goal is insufficient.

5.6 Magnitude Regression

Table 8 shows MAE under the three reweighting strategies. Normal and linear losses minimize all-event MAE, as expected. Exponential reweighting increases all-event MAE but substantially reduces MAE on $M > 4.0$, $M > 5.0$, and $M > 6.0$ subsets. This tail-focused ablation supports the need to optimize the magnitude module for thresholded high-magnitude screening rather than only for average error over the many small events.

5.7 Are Magnitudes Encoded in Space-Time Histories?

We test the hypothesis that large-event information is implicitly encoded in space-time activity patterns. Using only time and coordinate sequences (no magnitude features), an LSTM binary classifier predicts whether a sequence contains at least one $M > 4.5$ event. Validation and test AUC-ROC are 0.68 and 0.72 (Table 9). These results support the presence of meaningful large-event information in the space-time pattern and motivate applying magnitude-aware forecasting on top of an LSTM that encodes seismic clustering patterns from time and coordinate histories.

Table 10: Fixed-window $M > 4.5$ exceedance ranking on the 2016–2024 test period ($\Delta T = 7$ days, $\Delta r = 20$ km). Values are AUC-ROC from one chronological split.

Method	AUC-ROC
Random baseline	0.503
Coordinate-only background forecast	0.730
MagST-PP (history-aware)	0.790

Table 11: AUC-ROC sensitivity to query radius Δr and forecast horizon ΔT.

Δr / ΔT	3 d	7 d	14 d
5 km	0.67	0.74	0.73
10 km	0.80	0.82	0.76
20 km	0.83	0.79	0.81

Table 9 also summarizes horizon-dependent diagnostics. A model trained on 2006–2015 degrades under direct transfer to later horizons, consistent with catalog nonstationarity. Rolling autoregressive training stabilizes performance; adding future-window GMM spatial samples to magnitude imputation yields the strongest horizon profile, supporting this routing choice.

5.8 Magnitude-Aware Fixed-Window Forecasting

Table 10 reports the main hazard-oriented ranking result on the held-out 2016–2024 test period: binary prediction of at least one $M > 4.5$ event within 7 days and 20 km of the query location. MagST-PP achieves AUC-ROC 0.79, compared with 0.503 for a random baseline and 0.73 for a coordinate-only background model. This comparison shows improvement over a strong static-location forecast, but it does not establish superiority over ETAS/Hawkes models under the same fixed-window query because a task-matched ETAS Monte Carlo wrapper is not included.

Table 11 reports sensitivity to query radius and forecast horizon. Performance is non-monotone because the task changes with geometry: smaller radii are spatially precise but yield sparser positive labels, while longer horizons include more positives but may dilute short-term history effects. The sweep uses one chronological split with rare positives, so it is a sensitivity diagnostic rather than a significance or calibration analysis.

5.9 Qualitative Retrospective Forecast Maps

Figure 2 provides qualitative spatial diagnostics: a 10-year $M > 6$ probability surface and bin-level information gain relative to a background forecast. The maps suggest that learned probability mass concentrates in historically active seismic zones rather than spreading uniformly, but they are not public-facing risk products and do not support stronger geologic claims on their own.

6 Discussion

Our experiments suggest four conclusions. First, continuous spatial modeling matters: the 2-D GMM decoder improves cluster-aligned spatial likelihood without grid discretization (Table 4). Second, magnitude-aware objectives must address class imbalance: exponential reweighting improves high-magnitude MAE while increasing all-event MAE (Table 8). Third, recent seismic history adds value beyond static location: fixed-window AUC improves from 0.73 to 0.79, and space-time histories alone carry latent magnitude-relevant information (Table 9). Fourth, domain knowledge requires structural alignment: cluster and mainshock labels help, while naive fault-distance regularization provides little benefit or hurts after magnitude filtering (Table 7).

Scope. These findings are retrospective, not operational warnings. Table 4 evaluates likelihood, while Table 10 evaluates fixed-window ranking against random and coordinate-only backgrounds. The present evaluation does not include a task-matched ETAS Monte Carlo baseline, AUC confidence intervals, or Brier/reliability calibration.

7 Conclusion

We presented MagST-PP, a magnitude-conditioned continuous neural spatio-temporal point process for earthquake forecasting. Combining 1-D temporal and 2-D spatial GMM decoders with a tail-aware magnitude head and an autoregressive Monte Carlo sampler, the model improves held-out spatial log-likelihood and achieves AUC-ROC 0.79 on the primary retrospective exceedance-ranking query. Ablations identify the roles of flexible context windows, structured domain labels, and future spatial samples. Overall, earthquake forecasting is a natural MiLeTS problem—irregular, spatial, nonstationary, and imbalanced—where continuous event-history models can complement ETAS-style seismicity modeling.

References

[1] Neri Berman, Oleg Zlydenko, Oren Gilon, Yossi Matias, and Yohai Bar-Sinai. 2026. Earthquake magnitudes depend on seismic history, as revealed by a neural network analysis. *Scientific Reports* (2026). doi:10.1038/s41598-026-48563-7

[2] Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. 2021. Neural Spatio-Temporal Point Processes. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XQQA6-So14

[3] Kelian Dascher-Cousineau, Oleksandr Shchur, Emily E. Brodsky, and Stephan Guennemann. 2023. Using Deep Learning for Flexible and Scalable Earthquake Forecasting. *Geophysical Research Letters* 50, 16 (2023), e2023GL103909. doi:10.1029/2023GL103909

[4] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent Marked Temporal Point Processes: Embedding Event History to Vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 1555–1564. doi:10.1145/2939672.2939875

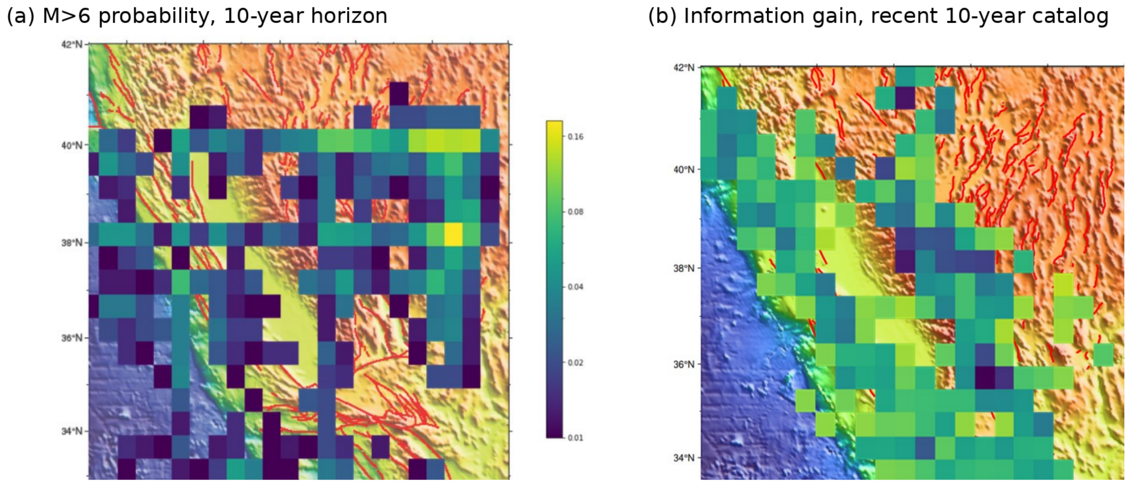

Figure 2: Retrospective spatial forecast diagnostics: (a) $M > 6$ probability over a 10-year horizon and (b) bin-level information gain relative to a background forecast. The panels are qualitative and use separate scales.

[5] Beno Gutenberg and Charles F. Richter. 1944. Frequency of Earthquakes in California. *Bulletin of the Seismological Society of America* 34, 4 (1944), 185–188.

[6] Egill Hauksson, Wenzheng Yang, and Peter M. Shearer. 2012. Waveform Relocated Earthquake Catalog for Southern California (1981 to June 2011). *Bulletin of the Seismological Society of America* 102, 5 (2012), 2239–2244. doi:10.1785/0120120010

[7] Alan G. Hawkes. 1971. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika* 58, 1 (1971), 83–90. doi:10.1093/biomet/58.1.83

[8] Hongyuan Mei and Jason M. Eisner. 2017. The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Red Hook, NY, USA, 6754–6764. https://proceedings.neurips.cc/paper/2017/hash/6463c88460bd63bbe256e495c63aa40b-Abstract.html

[9] Sumantrak Mukherjee, Mouad Elhamdi, George Mohler, David Antony Selby, Yao Xie, Sebastian Josef Vollmer, and Gerrit Grossmann. 2025. Neural Spatiotemporal Point Processes: Trends and Challenges. *Transactions on Machine Learning Research* (2025), 1–27. https://openreview.net/forum?id=N69lSYWkMw

[10] Yosihiko Ogata. 1988. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *J. Amer. Statist. Assoc.* 83, 401 (1988), 9–27. doi:10.1080/01621459.1988.10478560

[11] Yosihiko Ogata. 1998. Space-Time Point-Process Models for Earthquake Occurrences. *Annals of the Institute of Statistical Mathematics* 50, 2 (1998), 379–402. doi:10.1023/A:1003403601725

[12] Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. 2019. Fully Neural Network Based Model for General Temporal Point Processes. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Red Hook, NY, USA, 2120–2129. https://papers.nips.cc/paper/8485-fully-neural-network-based-model-for-general-temporal-point-processes

[13] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying Coordinated Accounts on Social Media Through Hidden Influence and Group Behaviours. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 1441–1451. doi:10.1145/3447548.3467391

[14] Oleksandr Shchur, Ali Caner Tuerkmen, Tim Januschowski, and Stephan Guennemann. 2021. Neural Temporal Point Processes: A Review. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 4585–4593. doi:10.24963/ijcai.2021/623

[15] Samuel Stockman, Daniel J. Lawson, and Maximilian J. Werner. 2026. EarthquakeNPP: A Benchmark for Earthquake Forecasting with Neural Point Processes. *Transactions on Machine Learning Research*, Article 6105 (2026). https://openreview.net/forum?id=dIcNAg6ZuZ

[16] Daniel T. Trugman and Yehuda Ben-Zion. 2023. Coherent Spatial Variations in the Productivity of Earthquake Sequences in California and Nevada. *The Seismic Record* 3, 4 (2023), 322–331. doi:10.1785/0320230039

[17] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. 2002. Stochastic Declustering of Space-Time Earthquake Occurrences. *J. Amer. Statist. Assoc.* 97, 458 (2002), 369–380. doi:10.1198/016214502760046925