BENCHMARKING CONFLICT–EVENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conflict-event forecasting is a high-stakes task with direct implications for humanitarian response, early warning, and policy planning. Despite the growing availability of fine-grained conflict-event data, the task remains underexplored in the machine learning community. We introduce CONFLICTCAST, the first machine learning benchmark for conflict event forecasting. Given an observed history of geo-referenced conflict-events, models predict an ensamble of future event trajectories, with their times, locations, and types; sampled trajectories can then be evaluated directly or aggregated to the space–time grids used by existing operational forecasting systems. CONFLICTCAST builds on four conflict-event datasets and includes neural spatio-temporal point processes, conflict-tailored neural models, tree-ensemble binned regressors, and simple baselines. We evaluate models under different conditioning regimes and a variety of metrics. Across datasets, we identify where current architectures succeed and fail, highlight directions for future work, and make standardized evaluation results publicly available through our website.

1 INTRODUCTION

[Code repository, Reproduction Paths, Artifacts]

[Website & Leaderboard]

Armed conflict remains a consequential feature of the international landscape. Although war-related deaths have declined substantially since the world wars (Herre et al., 2024), organized violence has not been eliminated, and recent years have witnessed a renewed increase in conflict activity worldwide. By 2016, more countries were experiencing violent conflict than at almost any point in nearly 30 years (United Nations, 2020). In 2024, the Heidelberg Institute for International Conflict Research recorded 27 wars and 23 limited wars (HIIK, 2025). By mid-2025, the United Nations estimated that armed conflict and organised violence had displaced more than 100 million people worldwide (UNHCR, 2025).

Humanitarian actors, peace operations, and crisis-response planners seek to reduce the harms of armed conflict by monitoring local conditions and allocating resources preemptively. In practice, this often translates into making forecasts of how violence in a conflict zone is likely to evolve over the coming days and weeks.

In recent years, the growth of fine-grained conflict-event datasets has made this forecasting problem increasingly suitable for machine-learning methods (Hegre et al., 2019; Vesco et al., 2022). These datasets represent conflict as a continuous stream of discrete violent incidents, each defined by time, location, and incident type. Beyond this shared structure, however, they also differ substantially in coverage, inclusion criteria and data collection. For example, the Armed Conflict Location and Event Data Project (ACLED) codes reports from media, non-governmental organisations, and local partners into structured records of political violence and protest (Raleigh et al., 2010). The UCDP Georeferenced Event Dataset (UCDP-GED) records incidents of organised violence from newswire and related sources (Sundberg & Melander, 2013). United Nations peacekeeping missions collect first-hand reports through the SAGE system in conflict zones where UN peacekeepers are deployed.

Despite progress on both data and modeling fronts, conflict-event forecasting still lacks a unified benchmark across datasets and evaluation protocols. This lack stands in contrast to other societally important forecasting domains, such as weather prediction (Rasp et al., 2020) and climate modelling (Watson-Parris et al., 2022), where benchmark suites have helped standardise evaluation, compare modelling approaches, and foster interdisciplinary research across communities. In these fields,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shared benchmarks have been a primary driver of methodological progress, enabling cumulative gains that would be difficult to achieve under fragmented and individual efforts. Conflict-event forecasting has yet to benefit from such common foundation.

In this paper, we introduce CONFLICTCAST, a benchmark for conflict-event forecasting. We formulate the task in its most general form: predicting future conflict-event trajectories. Given a sequence of observed conflict events up to a forecast origin, a model generates an *ensemble* of trajectories over a fixed forecast horizon. Each trajectory consists of a sequence of future events, with each event characterized by a time, location, and type, and represents one possible continuation of the observed historical sequence.

Existing evaluations of in conflict research (e.g., Hegre et al., 2019) typically score aggregated counts at a fixed horizon, which obscures event-level structure and ignores how forecasts are revised as new data arrive. We therefore pair CONFLICTCAST with an evaluation framework built around three simulation regimes that differ in how much of the realised event history is revealed during the forecast window: after every event (*fully-conditioned*), at regular intervals, e.g., such as in ACLED's weekly releases (*periodic-update*), or not at all (*no-update*). Within each regime, we evaluate a range of metrics including distributional accuracy, operational calibration, and standard grid-count metrics.

Practically, we implement CONFLICTCAST through an end-to-end open-source pipeline (see Fig. 1) designed to standardize data-modelling procedures, facilitate covariate integration, and enable transparent, reproducible evaluation. More broadly, beyond machine learning researchers, we envision this infrastructure as supporting practitioners—such as data scientists at the United Nations—by providing a public reference point for current modelling approaches and their performance on these datasets. It can also inform evidence-based decisions about the collection, use, and regulation of such data by quantifying the extent to which conflict events can currently be predicted via machine learning.[1]

We initialize CONFLICTCAST with four datasets spanning two diverse conflict settings and the principal reporting mechanisms used in quantitative conflict research: primary information collection, media-based collection, and mixed-source collection. Our website and leaderboard provide standardized comparisons and document how to contribute evaluations on additional regions and conflict-event datasets, making CONFLICTCAST an extensible community resource.

Our experiments reveal three main patterns. First, conflict-tailored Hawkes models perform best overall, with their clearest advantage in spatial prediction. Second, neural and self-exciting point processes excel at next-event prediction, but their advantage diminishes as forecasts are rolled forward with less frequent conditioning, where simpler count-based and persistence-style models become more competitive. Third, across model families, simulated trajectories remain imperfectly calibrated on operational properties such as event volume, burstiness, spatial extent, and hotspot structure. Together, these results suggest that models with explicit spatio-temporal dynamics can capture short-horizon conflict evolution well, but translating this strength into reliable, calibrated longer-horizon forecasts remains an open challenge.

2 TASK AND EVALUATION

The task is to forecast future event streams: given the conflict events observed up to a forecast origin, predict the events that will happen in a fixed forecast window. Each event has a time, a location, and a type, and a trajectory is an ordered sequence of events. CONFLICTCAST implements this task in two stages:

Simulation. `stpp-sim run <args>` takes a trained model, a forecast window, an updating regime, and a requested number of Monte Carlo trajectories. It returns an ensemble of simulated trajectories generated under that regime.

Evaluation. `stpp-eval score <args>` compares the simulated ensemble with the realised held-out trajectory using the metric families described below.

[1] For safety reasons, we deliberately restrict our experiments and any future internal or external addition to historical datasets, avoiding the deployment or evaluation of forecasts for ongoing conflicts, where predictions could influence real-world decisions or inform malicious actors.

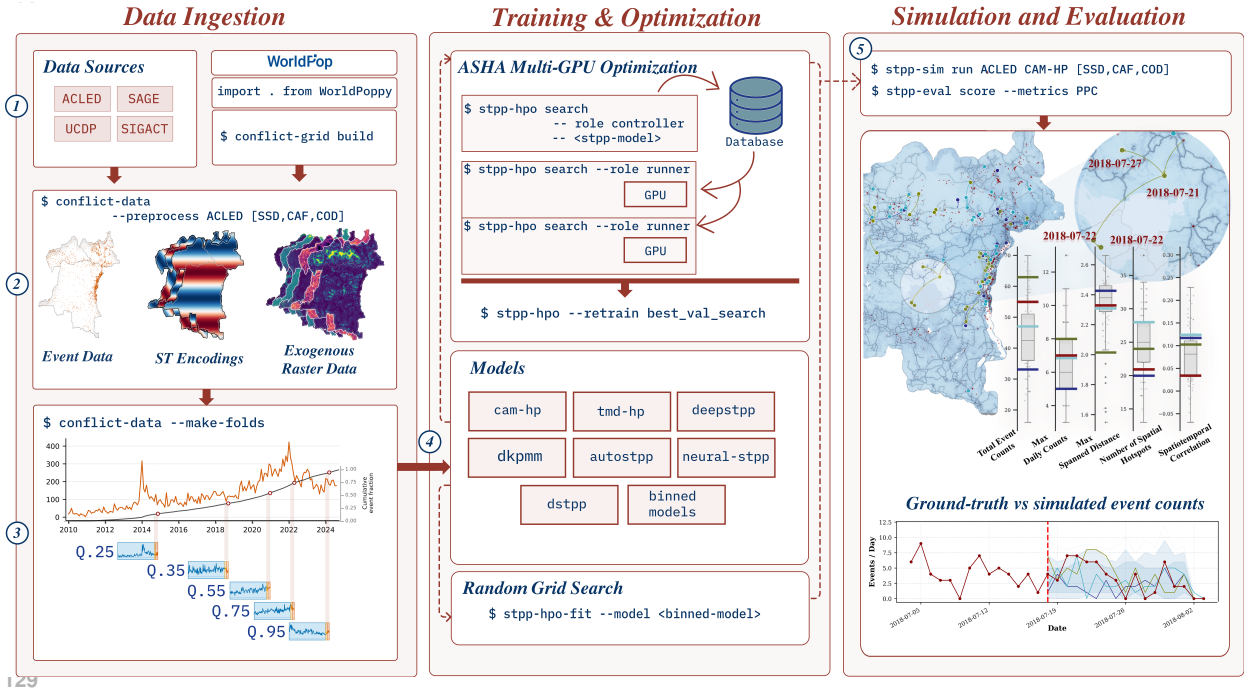


Figure 1: **End-to-end CONFLICTCAST pipeline.** The figure shows the full CONFLICTCAST pipeline, instantiated here for ACLED over South Sudan, the Central African Republic, and the Democratic Republic of the Congo (SSD, CAF, COD), our Central Africa conflict regime. Raw event records are converted into the benchmark event-stream format **(1)**, enriched with spatio-temporal encodings and raster covariates **(2)**, and split into temporal folds by placing forecast origins at empirical event-time quantiles **(3)**. Models are tuned through a shared optimization interface, using ASHA-based multi-GPU search for neural point processes and grid search for binned models **(4)**. Trained forecasters are then evaluated by simulating future trajectories under the benchmark updating regimes **(5)**. Panel **(5)** shows three CAM-HP sample trajectories (, ,) and the realised trajectory (), evaluated through posterior predictive checks and binned event-count comparisons.

This section describes metrics by keeping mathematical notation to a minimum. §2.1 defines the three updating regimes used during simulation. §2.2 then describes the metric families used to score the resulting trajectory ensembles. Full metric definitions are given in §G.

2.1 UPDATING REGIMES

```
stpp-sim run --regimes ["FC", "PU", "NU"]
```

The updating regime specifies how much realised test history is made available while generating the future, with each regime corresponding to a different operational setting. Our primary forecast horizon is $W_0 = 6$ weeks, motivated by the short-term horizons used in operational conflict early-warning systems (Rød et al., 2024); for completeness, we also evaluate 12- and 24-week horizons (§E.1). We use three regimes:

- **Fully conditioned (FC).** The simulator is reconditioned on the realised history after every realised event. This corresponds to next-event prediction across the test window, and reflects settings with near-real-time access to conflict reports.
- **Periodic update (PU).** The simulator is reconditioned on the realised history only at fixed update intervals; within each interval, trajectories evolve autoregressively from the history available at its start. This reflects settings in which newly collected events are released in batches. We set the interval to 7 days, matching the weekly publication cycle of ACLED.
- **No update (NU).** The simulator is conditioned only on the history available at the start of W_0 and then generates the full test window from simulated events alone. This reflects longer-horizon planning, where decisions must be made using only the information available at the forecast origin.

2.2 METRICS

```
stpp-eval score --metrics ["energy", "ppc", "grid", "NLL"]
```

Energy score with Wasserstein distance. The energy score (Gneiting & Raftery, 2007) is part of the standard toolkit for probabilistic forecast evaluation (Gneiting & Katzfuss, 2014)[2] It evaluates whether the simulated ensemble reproduces the temporal and spatial *shape* of the realised trajectory. We compute three such scores: a temporal score using one-dimensional Wasserstein-1 distance between event-time distributions (Villani, 2009); a spatial score using sliced Wasserstein-1 distance between spatial event distributions (Bonneel et al., 2015); and a joint spatio-temporal score using sliced Wasserstein-1 distance over standardised time–space coordinates. We report the former in days and metres and the latter as skill relative to climatology, $1 - \mathrm{ES}_{\mathrm{model}}/\mathrm{ES}_{\mathrm{climatology}}$. Full details are provided in Appendix L.1.

Posterior predictive checks (PPC). We use posterior predictive checks (Gelman et al., 1995) to ask whether interpretable properties of the realised trajectory are plausible under the simulated trajectories. A *check* takes one statistic of interest, such as total event count, and compares its value on the realised trajectory with the empirical distribution of values across the simulated trajectories. This yields an empirical PPC p-value for each statistic: values near 0 or 1 indicate that the realised statistic lies in the tails of the simulated distribution, while values near 0.5 indicate that it lies near the centre.

We use five checks, each targeting a distinct quantity of interest: total event count (*volume*), maximum daily count (*burstiness*), maximum pairwise distance between events (*spatial extent*), number of connected components when events within 50 km are linked (*spatial hotspots*), and Pearson correlation between pairwise spatial and temporal distances (*space–time coupling*).

We summarise calibration with two *pass rates*: `PPC-in-90`, the share of checks with $p \in [0.05, 0.95]$, and the stricter `PPC-in-40`, the share with $p \in [0.30, 0.70]$. For a calibrated simulator, the expected value of these pass rates are 90% and 40%.

Grid-aggregated count metrics and point-process likelihood. We also evaluate forecasts after aggregation to the format used by existing conflict-forecasting systems: event counts on spatial grids and fixed time bins. Our primary count metric is the continuous ranked probability score (CRPS) of the simulated counts in each cell and week, which also ranks the probabilistic forecasts of the VIEWS prediction challenge (Hegre et al., 2025); we compute it on 25 km and 50 km cells with 7-day bins under the PU and NU rollouts (§G.3), and report $\mathrm{MAE}^{+}_{\mathrm{Real}}$ and $\mathrm{MAE}^{+}_{\mathrm{Pred}}$ as secondary errors. We omit grid-count metrics under the fully conditioned (FC) regime, where the simulator is reconditioned after each observed event and the simulated event count therefore equals the realised one by construction.

To connect with the point-process literature, we also report next-event point-process negative log-likelihood. This evaluates next-event predictive density under the realised history and is included because most neural spatio-temporal point processes are trained by maximum-likelihood objectives.

3 DATA SOURCES AND DATA PROCESSING

3.1 DATA SOURCES AND CONFLICT REGIMES

The benchmark covers three public conflict-event datasets and event data collected by United Nations (UN) peacekeeping missions. These sources span the main reporting mechanisms used in quantitative conflict research, including primary field reporting, media-based collection, and mixed-source pipelines. Geographically, we focus on two different conflict regions with distinct dynamics.

Central Africa (diffusive civil-war regime). We first focus on the interconnected conflict system of the Democratic Republic of the Congo (DRC), the Central African Republic (CAR), and South Sudan (SSD). Many conflict actors in this region operate transnationally, including the Lord's Resistance Army and various regional militias. We choose this region because it hosts UN peacekeeping missions in which conflict early warning is operationalized (i.e., see SAGE below). For this region, we evaluate three distinct event datasets:

[2]The energy score is a standard generalisation of the continuous ranked probability score (CRPS) to multivariate or structured outcomes (Scheuerer & Hamill, 2015).

- **ACLED (2010–2024) (Raleigh et al., 2010):** Sourced from media reports and information transmitted by local partner organisations.
- **UCDP-GED (2010–2024) (Sundberg & Melander, 2013):** Sourced exclusively from global newswire reports with strict inclusion criteria (requiring at least one fatality per event). This dataset is sparser than ACLED but more consistent (Eck, 2012).
- **SAGE (2020–2024):** Event data compiled by UN peacekeeping missions in the CAR (MINUSCA), DRC (MONUSCO), and South Sudan (UNMISS), reflecting on-the-ground conflict reporting by each mission.

Iraq (high-intensity military-reporting regime) (Shaver & Bollfrass, 2023). As a counterpoint to the Central Africa setting, we use U.S. military **SIGACT** field reports from the Iraq War Logs (2004–2009), which were made public in 2010. These records provide dense incident-level observations of violence during the Iraq War and expose models to high-frequency burstiness and tactical micro-dynamics that are often smoothed out in civilian or news-based reporting. This setting therefore tests whether forecasting models can handle a very different observation process and temporal scale from those found in sparser event datasets.

3.2 Geospatial Raster-Data

`conflict-grid build --card path-to-card`

Conflict dynamics are strongly associated with exogenous socioeconomic and geographic factors, including population density, terrain ruggedness, accessibility, and human activity (Buhaug & Rød, 2006; Raleigh & Hegre, 2009; Tollefsen et al., 2012). Existing forecasting systems such as ViEWS incorporate such covariates at the resolution of their output grid, typically country-month or grid-cell-month units at approximately 55×55 km (Hegre et al., 2019). In contrast, ConflictCast supports covariate construction at user-specified spatial resolutions. Given a region of interest, the pipeline constructs a spatial grid and retrieves raster covariates on demand through `WorldPoppy`, a separate Python package that we developed to manage raster access independently from the benchmark code. This design makes the covariate interface easier to maintain and extend, since new raster sources can be added by updating `WorldPoppy` without changing the core benchmark pipeline.[3]

For the Central African datasets (ACLED, UCDP and SAGE), we build a 1km-cell grid with $D = 17$ covariate channels [4]. Four static channels describe terrain and infrastructure: slope, elevation, distance to roads, and distance to waterways. Nine yearly channels describe socioeconomic context at three spatial aggregation scales: population, nightlights, and built surface area. Four additional dynamic channels describe proximity to land-use features: parks, urban areas, cultivated land, and woody cover. For dynamic covariates, we use a year-minus-one alignment rule: events in year Y are matched to raster values from year $Y - 1$, preventing lookahead bias. SIGACT-Iraq uses a reduced set of the two terrain channels, slope and elevation, because all other rasters postdate its 2004–2009 events: thirteen channels come from products that start in 2015 and the road and waterway distances from a 2016 snapshot, so using them would give the models information from after the events they forecast. Further details on the raster catalogue and grid construction are given in §I.

3.3 Data Processing and Training Folds

`conflict-data preprocess --card path-to-dataset-card`

This step proceeds in two stages. First, a source-specific `adapter` module processes each dataset into a canonical event schema; adding a new dataset to ConflictCast therefore amounts to writing a new adapter for that source. Second, raster covariates from the pre-built master grid are sampled at event locations, and Fourier and calendar-based spatio-temporal encodings are appended to each event. Source-specific ingest details are given in §J.

`conflict-data make-folds --fold-spec path-to-fold-spec`

This step constructs the temporal folds used for model selection and evaluation. We use activity-based stratification, placing forecast origins at quantiles of event time rather than at fixed calendar dates.

[3] Planned extensions include temperature and precipitation rasters.

[4] `Worldpoppy` currently supports 54 covariates, yet they have different coverage in space and time;

This choice reflects the uneven temporal structure of conflict-event streams, where long periods of low activity can be followed by short periods of intense escalation. By stratifying folds by observed activity, CONFLICTCAST evaluates models at successive stages of each conflict trajectory while ensuring that test windows contain enough events to support meaningful evaluation. We benchmark our models on an up-to-five-fold specification, which places forecast origins at the 15th, 35th, 55th, 75th, and 95th percentiles of event time (see Fig. 1, 3). Each fold contains a 728-day training window, a 28-day validation window, and a 42-day held-out test window. For each fold, feature transformations are fit on the training window only and then applied to validation, test, and simulation inputs. This prevents lookahead bias and ensures that all models receive covariates produced by the same fold-specific preprocessing pipeline. Feature transform details can be found in Tab. 15 and the fold-spec card schema are in §K.

4 MODELS

CONFLICTCAST implements two broad model families: neural spatio-temporal point processes (NSTPPs) and binned count models. NSTPPs (Chen et al., 2021; Zhou et al., 2022; Yuan et al., 2023; Zhou & Yu, 2023; Zhang et al., 2023) model the event-stream directly: each model defines a conditional distribution over the time, location, and type of the next event given the observed history. Binned count models first aggregate the continuous event stream into fixed spatial and temporal cells and then predict event counts for each cell.

4.1 NSTPPS

We port five recent neural spatio-temporal point process models that represent complementary modelling strategies in the literature: **Neural STPP** (Chen et al., 2021), which uses Neural ODE and continuous-time normalising-flow parameterisations for history-dependent temporal and spatial distributions; **DSTPP** (Yuan et al., 2023), which models the spatio-temporal joint next-event distribution with a denoising diffusion process; **DeepSTPP** (Zhou et al., 2022), which represents the space-time intensity as a kernel mixture governed by a latent stochastic process; **AutoSTPP** (Zhou & Yu, 2023), which uses an automatic-integration architecture with a ProdNet-based integral network to make likelihood computation tractable; and **DKMPP** (Zhang et al., 2023), which extends deep mixture point processes with deep kernels and multimodal covariates, trained through integration-free score matching.

Each NSTPP requires a model-specific simulator to support trajectory-level evaluation. This is a non-trivial requirement: most existing NSTPPs are evaluated primarily through likelihood-based or next-event prediction metrics rather than simulated rollout quality (Mukherjee et al., 2025), and their public implementations typically do not provide trajectory simulators. We implement simulators for all five ported NSTPPs. DeepSTPP, DSTPP, and DKMPP expose their intensities or next-event samplers directly; for AutoSTPP and Neural STPP, whose public implementations ship no trajectory sampler, we implement exact thinning samplers at the wrapper level (§F.4). Neural STPP is simulated from checkpoints retrained with physical time units (days; six hours on SIGACT), because the as-released time normalisation caps its intensity at about 0.07 events per day and makes free-running rollouts degenerate.

4.2 PROPOSED CONFLICT MODELS

To broaden the class of neural models supported by the benchmark, we adapt ideas from the temporal and spatio-temporal PP literature into two further architectures that we add to the point process class.

TMD-HP Model. TMD-HP (Transformer Mixture Density Hawkes Process) is a transformer-based spatio-temporal point-process baseline designed to test whether covariate-aware neural sequence models can improve conflict-event forecasting. It extends TransFeat-TPP (Meng et al., 2024) from temporal to marked spatio-temporal events and uses a per-mark spatial mixture head with learned hotspot anchors to model recurring conflict locations. Details are provided in App. F.3.

CAM-HP. CAM-HP (Covariate Adaptive Mixture Hawkes Process) is a covariate-aware multivariate Hawkes baseline that models conflict diffusion through mixtures of spatio-temporal triggering

kernels. Excitation strengths are modulated by event-level raster covariates, allowing the model to adapt triggering patterns to local geographic and socioeconomic context. We include it to test whether explicit Hawkes-like multi-scale diffusion remains a useful inductive bias for conflict-event forecasting (Kibris, 2021; Schutte & Weidmann, 2011). Full specifications are provided in §F.

4.3 Binned models

Tree ensemble regressors from operational forecasting pipelines. Tree ensemble models are a central reference class in operational conflict forecasting systems (Hegre et al., 2019; Vestby et al., 2022). We include two such baselines. **XGBoost** (Chen & Guestrin, 2016) and **Hurdle-GBM** (Vestby et al., 2022). Hurdle-GBM separates the prediction problem into two stages: a binary classifier first predicts whether a cell contains any events, and a count regressor then predicts the event count conditional on cell occupancy. Both stages are implemented with XGBoost.

Simulating from binned models To simulate event trajectories from such models, we proceed as follows. For each cell–bin, the predicted count is interpreted as the rate of a homogeneous Poisson process over that bin. We sample an event count from this process and then draw that many event locations and times uniformly at random from the cell's spatial extent and the bin's time interval, respectively. Pooling the sampled events across all cells and bins and sorting them by time yields a simulated trajectory. This procedure is justified by regarding binned models as a particular subclass of point-process models, namely those whose intensity function is piecewise constant over the spatio-temporal domain. A formal definition of the intensity function is given in §2, and full details of binned models are provided in §F.4.

4.4 Other baselines.

For reference, we also implement a **multivariate Hawkes process** (Hawkes & Oakes, 1974) and three simple binned baselines: **Persistence**, which predicts each cell's next-bin count as its most recent bin count; **Climatology**, which uses the long-run training-window mean per cell; and **Moving average**, which uses a rolling mean over a configurable temporal window.

4.5 Model packages `stpp-*`

The `stpp-*` packages provide the infrastructure for benchmarking NSTPPs independently of the conflict-event domain. New models can be added by implementing the standard `train_fn` interface described in §M, which connects the model to our hyperparameter-optimization and evaluation pipeline.

For models with available public code, we keep the upstream implementation under an `_upstream/` directory and expose it through a thin ConflictCast wrapper. For models without suitable public code, we provide our own implementation. Implementation details for each model are reported in Tab. 12.

5 Hyperparameter optimization

For each $(\text{dataset}, \text{fold}, \text{model})$ cell, we run an independent hyperparameter search and select the configuration with the lowest validation loss. We then retrain the selected configuration on the combined training and validation windows before evaluating it once on the held-out test window. For neural models, the factorial structure of the benchmark—datasets, folds, and model families—makes exhaustive search impractical. We therefore use Asynchronous Successive Halving (**ASHA**) (Li et al., 2018), implemented with a multi-GPU runner that dispatches trials in parallel across the available devices. ASHA allocates small initial budgets to many configurations and progressively prunes poorly performing trials at predefined rungs, allowing us to evaluate several hundred candidate configurations within a practical wall-clock budget. For binned baselines, we use grid search over their respective hyperparameter spaces. Search budgets and rung schedules are held fixed across most neural models for direct comparability; Neural STPP uses a smaller schedule because of its higher per-trial cost (§N).

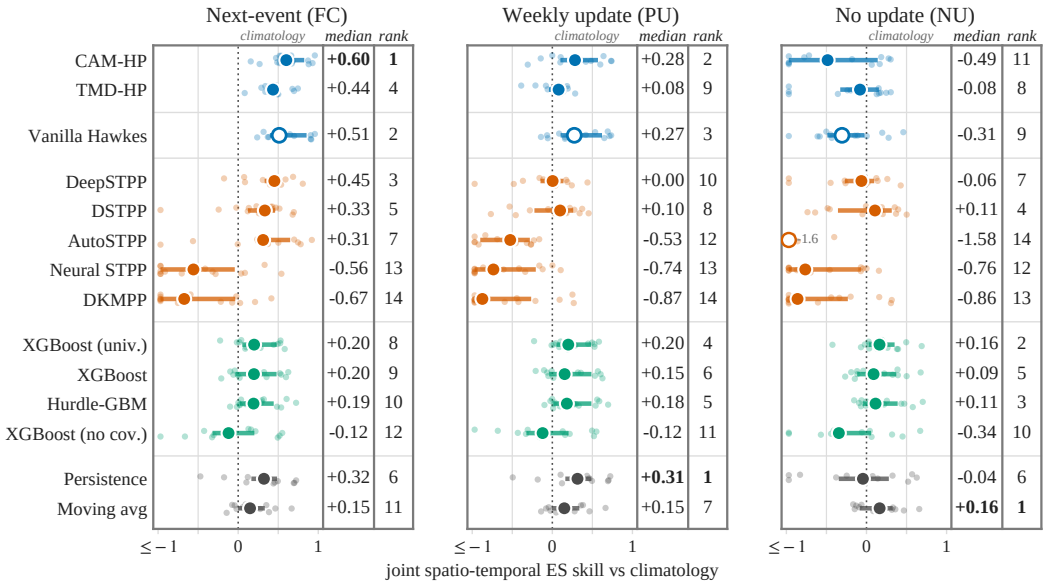


Figure 2: **Joint spatio-temporal trajectory fidelity across rollout regimes.** Skill of the joint energy score against climatology (0 = climatology, 1 = perfect; §G.2) under the fully conditioned next-event regime (FC), the 7-day periodic-update regime (PU) and the 6-week no-update regime (NU). Rows are grouped by model family and colour marks the family: ● conflict-tailored neural point processes (CAM-HP, TMD-HP), ○ the vanilla Hawkes process, ● off-the-shelf NSTPPs, ● binned regressors, ● naive baselines. Each light dot is one dataset–fold cell, the bar spans the interquartile range, the marker is the median, and the median and its rank within the panel are printed at the right edge (bold for the best model in the panel). Hollow markers at the axis floor denote medians below -1.

6 Results

Conflict-tailored Hawkes models rank highest overall, with their clearest advantage in spatial prediction. Tab. 1 ranks the 13 models within each dataset–fold combination for every metric. We summarize performance by the median rank across these cells and average the five metric families with equal weight. CAM-HP ranks first overall (3.13), followed by vanilla Hawkes (3.83), TMD-HP (4.89), XGBoost (5.53), and Hurdle-GBM (5.69); Neural STPP and DKMPP rank last. The advantage of Hawkes-based models is clearest in the spatial component: CAM-HP holds median ES-Space ranks of 1.5–3.0 under all three regimes (Tab. 5), and CAM-HP or vanilla Hawkes achieves the best spatial NLL term on all four datasets (Tab. 3). This suggests that explicitly modeling spatial diffusion provides a useful inductive bias for conflict-event forecasting. Our kernel ablations support this interpretation: for CAM-HP, power-law temporal decay and a multi-scale mixture beat exponential and single-scale kernels in every ACLED fold (Tab. 8); for TMD-HP, the heavier-tailed Laplace spatial kernel beats the Gaussian one in four of five folds, although not decisively (Tab. 9).

Neural and self-exciting point processes excel at next-event prediction, but their advantage fades as conditioning becomes stale. The FC panel of Fig. 2 shows that point-process models are strongest when forecasting the next event from the full observed history. This is consistent with their test likelihoods: AutoSTPP has the best median likelihood rank (1.0), followed by CAM-HP (2.0), TMD-HP (3.0), and DeepSTPP (4.0; Tab. 1). The picture changes as forecasts are rolled forward with less frequent conditioning. Importantly, without updates, the self-exciting models fall below climatology (CAM-HP -0.49, vanilla Hawkes -0.31, TMD-HP -0.08), whereas XGBoost, its univariate variant, Hurdle-GBM (0.09–0.16), and the moving average (0.16) remain above it. Thus, the models that are strongest at next-event prediction are not necessarily the most robust under long open-loop rollouts.

No model is calibrated on the posterior predictive checks. A calibrated simulator would place the realised value of each summary statistic inside the central 90% of its simulated distribution in

Table 1: **Median rank per metric and overall ranking** (1 = best of 13, median over the 14 dataset–fold cells; best entry per column in bold).

	Likelihood	ES-time			ES-space			Counts		Calibration		Overall	
Model	NLL	FC	PU	NU	FC	PU	NU	PU	NU	PU	NU	Score	Rank
● CAM-HP	2.0	2.5	3.5	10.5	**2.0**	**1.5**	3.0	**1.0**	4.5	**3.0**	**3.5**	3.13	1
○ Vanilla Hawkes	5.0	2.5	**2.5**	5.5	**2.0**	3.5	4.0	2.0	5.5	4.0	**3.5**	3.83	2
● TMD-HP	3.0	**1.0**	6.5	11.5	4.0	3.5	3.0	7.5	7.0	5.2	**3.5**	4.89	3
● XGBoost	5.0	10.0	6.5	6.5	3.5	3.5	**2.0**	4.0	**3.0**	8.2	8.8	5.53	4
● Hurdle-GBM	6.0	10.0	7.0	5.0	4.5	4.0	3.5	4.0	4.0	7.0	7.2	5.69	5
● DeepSTPP	4.0	6.0	7.0	6.0	8.5	9.5	10.0	6.0	7.5	5.2	6.5	6.46	6
● AutoSTPP	**1.0**	5.0	7.0	7.5	9.5	12.0	12.0	8.5	10.0	11.2	9.0	7.61	7
● Moving avg	11.0	11.0	8.5	8.5	8.0	8.0	8.0	5.5	5.0	5.0	6.5	7.87	8
● DSTPP	9.0	6.0	6.0	**4.5**	9.0	7.5	8.0	10.0	8.5	8.0	7.5	7.93	9
● Climatology	8.5	11.0	9.5	7.0	10.5	9.0	9.5	6.0	6.0	7.8	7.8	8.22	10
● Persistence	13.0	10.5	7.5	5.0	7.0	7.0	7.0	13.0	11.5	6.8	7.5	9.41	11
● Neural STPP	11.0	6.0	10.0	7.0	12.5	12.0	11.0	12.0	11.5	9.8	10.0	10.42	12
● DKMPP	9.5	11.5	12.0	10.5	10.5	10.0	9.0	11.0	11.0	11.0	11.0	10.53	13

NLL, ES-time, ES-space: energy scores (§G.2). Counts: count-CRPS on 25 km × 7-day cells (§G.3); FC is omitted for counts and calibration because it fixes the simulated count to the observed one. Calibration: share of the five checks passed at `PPC-in-90` (§G.4). Score: mean over the five families of the family's average median rank. AutoSTPP's NLL covers 12 cells. ● conflict-tailored NPPs, ○ vanilla Hawkes, ● off-the-shelf NSTPPs, ● binned regressors, ● naive baselines.

90% of the checks (`PPC-in-90`, §G.4). Pooled over the five checks and all cells, the best model reaches 70% with weekly updates and 66% without them (CAM-HP in both), and the median model 54% and 51%; 44 of 48 model–dataset combinations fall short of the target with weekly updates and 47 of 52 without (Fig. 8). Under the stricter `PPC-in-40` band the best pass rate is 23% with weekly updates and 29% without, against a target of 40%.

Further results. For the remaining metrics, we organize extended results by metric (§D: likelihood in §D.1, energy scores in §D.2, grid counts in §D.3, posterior predictive checks in §D.4) and provide a small set of takeaways intended to guide practitioners in their interpretation and use. The full results are also available on our website in the public leaderboard.

7 Conclusion and Outlook

ConflictCast offers a foundation for machine-learning research on conflict-event forecasting. Across four datasets and three updating regimes, we identify two central failure modes of current event-stream models when applied to the conflict-setting: next-event gains can degrade under longer-horizon rollouts, and simulated trajectories are often miscalibrated over operational summary statistics. Hence, we highlight two directions for future work.

Long-horizon event-stream learning. Neural point processes are commonly trained and evaluated through next-event prediction (Shchur et al., 2021; Mukherjee et al., 2025). This objective is useful for measuring local predictive fit, but it is not sufficient for operational forecasting. In operational settings, models are often rolled forward for several days or weeks with limited reconditioning on the realised event history. Small errors in event time, location, or type can then accumulate through the simulated history and distort the resulting forecast trajectory. A central open problem is therefore to train point-process models whose predictive quality is stable under autoregressive multi-week simulation, not only under fully conditioned next-event evaluation.

Objectives aligned with operational summaries. Likelihood-based training should be complemented by objectives that directly reward calibrated uncertainty over quantities practitioners use, including event volume, burstiness, spatial extent, hotspot structure, and grid-cell counts. By exposing these gaps in a shared benchmark and pipeline, ConflictCast provides a testbed for models and evaluation protocols better aligned with decision support under uncertainty.

AI USE STATEMENT

We used generative AI tools (LLM-based coding assistants) to help write, refactor, and document the benchmark code, to orchestrate and monitor experiments, and to draft analysis scripts and figures; we also used them to edit and polish text. All AI-generated code was reviewed and tested by the authors, all numerical claims were checked against the underlying experiment artefacts, and the research questions, model choices, and interpretations are our own. We did not use generative AI tools to generate research ideas or to write the paper's content without author review. We take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

REFERENCES

Eduardo Albrecht. Predictive technologies in conflict prevention: Practical and policy considerations for the multilateral system. Technical report, United Nations University Centre for Policy Research, 2024. URL https://unu.edu/publication/predictive-technologies-conflict-prevention-practical-and-policy-considerations.

André Victor Ribeiro Amaral, Jonatan A González, and Paula Moraga. Spatio-temporal modeling of infectious diseases by integrating compartment and point process models. *Stochastic Environmental Research and Risk Assessment*, 37(4):1519–1533, 2023.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Patrick T. Brandt, Vito D'Orazio, Latifur Khan, Yi-Fan Li, Javier Osorio, and Marcus Sianan. Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions*, 48(4):800–822, 2022. doi: 10.1080/03050629.2022.2036987.

Halvard Buhaug and Jan Ketil Rød. Local determinants of african civil wars, 1970–2001. *Political Geography*, 25(3):315–335, 2006.

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. In *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2011.04583. URL https://openreview.net/forum?id=XQQA6-So14.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 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*, pp. 1555–1564, 2016.

Kristine Eck. In data we trust? a comparison of ucdp ged and acled conflict events datasets. *Cooperation and Conflict*, 47(1):124–141, 2012.

Felix Ettensperger. Forecasting conflict using a diverse machine-learning ensemble: Ensemble averaging with multiple tree-based algorithms and variance promoting data configurations. *International Interactions*, 48(4):555–578, 2022.

Quansheng Ge, Mengmeng Hao, Fangyu Ding, Dong Jiang, Jürgen Scheffran, David Helman, and Tobias Ide. Modelling armed conflict risk under climate change with machine learning and time-series data. *Nature Communications*, 13:2839, 2022. doi: 10.1038/s41467-022-30356-x.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974. doi: 10.2307/3212693.

Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, et al. Views: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174, 2019. doi: 10.1177/0022343319823860.

Håvard Hegre, Paola Vesco, Michael Colaresi, Jonas Vestby, Alexa Timlick, et al. The 2023/24 VIEWS prediction challenge: Predicting the number of fatalities in armed conflict, with uncertainty. *Journal of Peace Research*, 62(6), 2025. doi: 10.1177/00223433241300862.

Bastian Herre, Lucas Rodés-Guirao, and Max Roser. War and peace. *Our World in Data*, 2024. https://ourworldindata.org/war-and-peace.

HIIK. Conflict Barometer 2024. Technical report, Heidelberg Institute for International Conflict Research, 2025. URL https://hiik.de/wp-content/uploads/2026/03/coba-2024.pdf.

Mikyoung Jun and Scott Cook. Flexible multivariate spatiotemporal hawkes process models of terrorism. *The Annals of Applied Statistics*, 18(2):1378–1403, 2024.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Arzu Kibris. The geo-temporal evolution of violence in civil conflicts: A micro analysis of conflict diffusion on a new event data set. *Journal of Peace Research*, 58(5):885–899, 2021.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. Massively parallel hyperparameter tuning. *CoRR*, abs/1810.05934, 2018. URL http://arxiv.org/abs/1810.05934.

Luca Macis, Marco Tagliapietra, Elena Siletti, and Paola Pisano. Predicting fatalities with pre-trained temporal transformers: A time series regression approach. Contribution to the VIEWS Prediction Challenge 2023/2024, University of Turin, 2024. URL https://viewsforecasting.org/wp-content/uploads/Unito_VIEWSPredictionChallenge2023.pdf.

Iris Malone. Recurrent neural networks for conflict forecasting. *International Interactions*, 48(4): 614–632, 2022.

Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

Zizhuo Meng, Boyu Li, Xuhui Fan, Zhidong Li, Yang Wang, Fang Chen, and Feng Zhou. Transfeat-tpp: An interpretable deep covariate temporal point processes. *arXiv preprint arXiv:2407.16161*, 2024.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the american statistical association*, 106(493):100–108, 2011.

Sumantrak Mukherjee, Mouad Elhamdi, George Mohler, David A Selby, Yao Xie, Sebastian Vollmer, and Gerrit Grossmann. Neural spatiotemporal point processes: Trends and challenges. *arXiv preprint arXiv:2502.09341*, 2025.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Yosihiko Ogata. Seismicity analysis through point-process modeling: A review. *Pure and applied geophysics*, 155(2):471–507, 1999.

Maya Okawa, Tomoharu Iwata, Takeshi Kurashima, Yusuke Tanaka, Hiroyuki Toda, and Naonori Ueda. Deep mixture point processes: Spatio-temporal event prediction with rich contextual information. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 373–383, 2019.

Maya Okawa, Tomoharu Iwata, Yusuke Tanaka, Takeshi Kurashima, Hiroyuki Toda, and Hisashi Kashima. Context-aware spatio-temporal event prediction via convolutional hawkes processes. *Machine Learning*, 111(8):2929–2950, 2022.

Aristeidis Panos. Decomposable transformer point processes. *Advances in Neural Information Processing Systems*, 37:88932–88955, 2024.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Benjamin J. Radford. High resolution conflict forecasting with spatial convolutions and long short-term memory. *International Interactions*, 48(4):739–758, 2022. doi: 10.1080/03050629.2022.2031182.

Clionadh Raleigh and Håvard Hegre. Population size, concentration, and civil war. a geographically disaggregated analysis. *Political geography*, 28(4):224–238, 2009.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660, 2010. doi: 10.1177/0022343310378914.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), 2020. doi: 10.1029/2020MS002203.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6), 2024. doi: 10.1029/2023MS004019.

Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

Espen Geelmuyden Rød, Tim Gåsste, and Håvard Hegre. A review and comparison of conflict early warning systems. *International Journal of Forecasting*, 40(1):96–112, 2024.

Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

Sebastian Schutte and Nils B Weidmann. Diffusion patterns of violence in civil wars. *Political Geography*, 30(3):143–152, 2011.

Andrew Shaver and Alexander K. Bollfrass. Disorganized political violence: A demonstration case of temperature and insurgency. *International Organization*, 77(2):440–474, 2023. doi: 10.1017/S0020818323000024.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020. arXiv:1909.12127.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pp. 4585–4593. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/623. URL https://doi.org/10.24963/ijcai.2021/623. Survey Track.

Ralph Sundberg and Erik Melander. Introducing the ucdp georeferenced event dataset. *Journal of peace research*, 50(4):523–532, 2013.

Andreas Forø Tollefsen, Håvard Strand, and Halvard Buhaug. Prio-grid: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374, 2012.

UNHCR. Mid-Year Trends 2025. Technical report, United Nations High Commissioner for Refugees, 2025. URL https://www.unhcr.org/media/mid-year-trends-2025.

United Nations. A new era of conflict and violence. UN75 Issue Brief, 2020. URL https://www.un.org/sites/un2.un.org/files/2020/01/un75_conflict_violence.pdf.

Gilian van den Hengel and Philip Hans Franses. Forecasting social conflicts in africa using an epidemic type aftershock sequence model. *Forecasting*, 2(3):284–308, 2020.

Paola Vesco, Håvard Hegre, Michael Colaresi, Remco Bastiaan Jansen, Adeline Lo, Gregor Reisch, and Nils B. Weidmann. United they stand: Findings from an escalation prediction competition. *International Interactions*, 48(4):860–896, 2022. doi: 10.1080/03050629.2022.2029856. URL https://doi.org/10.1080/03050629.2022.2029856.

Jonas Vestby, Jürgen Brandsch, Vilde Bergstad Larsen, Peder Landsverk, and Andreas Forø Tollefsen. Predicting (de-)escalation of sub-national violence using gradient boosting: Does it work? *International Interactions*, 48(4):841–859, 2022. doi: 10.1080/03050629.2022.2021198. URL https://doi.org/10.1080/03050629.2022.2021198.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Julian Walterskirchen, Sonja Häffner, Christian Oswald, and Marco Nicola Binetti. Taking time seriously: Predicting conflict fatalities using temporal fusion transformers, 2024.

Michael D Ward, Brian D Greenhill, and Kristin M Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375, 2010. ISSN 00223433. URL http://www.jstor.org/stable/20752194.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10): e2021MS002954, 2022.

Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3173–3184, 2023.

Yixuan Zhang, Quyu Kong, and Feng Zhou. Integration-free training for spatio-temporal multimodal covariate deep kernel point processes. *Advances in Neural Information Processing Systems*, 36: 25031–25049, 2023.

Zihao Zhou and Rose Yu. Automatic integration for spatiotemporal neural point processes. *Advances in Neural Information Processing Systems*, 36:50237–50253, 2023.

Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. Neural point process for learning spatiotemporal event dynamics. In Roya Firoozi, Negar Mehr, Esen Yel, Rika Antonova, Jeannette Bohg, Mac Schwager, and Mykel Kochenderfer (eds.), *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 777–789. PMLR, 23–24 Jun 2022. URL https://proceedings.mlr.press/v168/zhou22a.html.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pp. 11692–11702. PMLR, 2020.

Appendix Contents

A Ethics statement **15**

B Reproducibility statement **15**

Limitations **15**

C Related work **16**

D Extended results **18**

D.1 Per-event log-likelihood 18

D.2 Energy score 20

D.3 Grid-aggregated counts 23

D.4 Posterior predictive checks 27

E Ablations and sensitivity studies **31**

E.1 Multi-horizon evaluation 31

E.2 Reporting delay under periodic updates 32

E.3 Ablation on covariate effect 32

E.4 Ablations of conflict-model design choices 32

E.5 Resolution of the binned baselines 33

E.6 Sensitivity analysis on PPC pass rates 34

F Model specifications **34**

F.1 Notation 35

F.2 CAM-HP: Covariate-Adaptive Mixture Hawkes Process 36

F.3 TMD-HP: Transformer Mixture-Density Hawkes Process 36

F.4 Binned models 37

G Metric specifications **38**

G.1 Per-event log-likelihood 38

G.2 Energy score 38

G.3 Grid-aggregated counts 39

G.4 Posterior predictive checks 40

H Extended data statistics **40**

I Covariate catalog and year alignment **41**

I.1 Building the master grid 41

I.2 Catalog of channels 42

I.3 Available raster products 42

I.4 Year alignment 42

J Ingest preprocessing and Fourier encodings **42**

J.1 Source-specific ingest rules 42

J.2 Fourier-and-calendar encodings 42

K Feature transformations and fold construction **43**

K.1 Per-fold feature transformations 43

K.2 Fold-spec card schema 44

L Datasheet for CONFLICTCAST **44**

M Repository architecture and reproducibility pipeline **45**

N Training protocol and search spaces **46**

O Compute statement **46**

A ETHICS STATEMENT

This research proposes forecasting models and evaluation techniques for conflict-event data, contributing to the scientific understanding of violent conflict (Ward et al., 2010) and the development of humanitarian early-warning systems (Hegre et al., 2019). While ML-based systems aim to enhance civilian protection and the safety of humanitarian personnel, their deployment carries significant ethical and operational risks. These include the potential for dual-use in state repression, automation bias in high-pressure environments, and the risk of systematic "blind spots" arising from lower data coverage in remote regions (Albrecht, 2024).

We characterise conflict-event forecasting at this stage as foundational research. Current early-warning systems are not yet sufficiently reliable to serve as a prominent basis for humanitarian resource allocation. While the field matures, humanitarian organisations will find more immediate utility of machine learning in other areas, such as supply-chain optimisation and population estimation. We nevertheless recommend that humanitarian actors remain engaged with the forecasting literature and that ML models of conflict events be developed through a close, trans-disciplinary dialogue. This collaboration should focus on: (1) the continued re-evaluation of potential risks; (2) the development of model-evaluation schemes that align with the operational needs of decision makers; and (3) the design of human-centric analysis workflows that can accommodate the probabilistic nature of ML models.

B REPRODUCIBILITY STATEMENT

The benchmark, training pipeline, raster-fetching layer, and all per-fold evaluation artefacts are released at an anonymised URL for review and will be deposited at a permanent archive on acceptance. Manifests are Pydantic-validated; every artefact carries SHA256 provenance. The full pipeline from raw data to benchmark numbers is executable as the seven CLI commands documented in §M and artifacts are accessible through the repository. Configuration files for every model, dataset, and grid are version-controlled alongside the code, and the per-fold leakage-prevention invariants are formalised in §N.

LIMITATIONS

We explicitly surface the following limitations of CONFLICTCAST.

(i) *Observation processes.* Conflict-event datasets do not provide a complete record of underlying violence: they reflect different reporting mechanisms, inclusion criteria, geographic coverage, and delays. Because CONFLICTCAST intentionally preserves these source-specific characteristics, some differences in model performance may reflect the observation process as well as the underlying conflict dynamics. Including multiple datasets over the same region partly exposes this variation, but does not eliminate it.

(ii) *Geographic scope.* The current benchmark focuses on two conflict settings where we were able to work closely with practitioners and domain experts familiar with the regions, helping us define coherent spatial domains and assemble comparable data sources across reporting mechanisms. This necessarily limits the geographic scope of the present evaluation, and model performance may differ in other conflict settings. At the same time, CONFLICTCAST is designed as an extensible framework: researchers studying particular conflicts can contribute new regions and datasets directly, or request their inclusion through the contact form on the benchmark website (we will put it up with de-anonymization), while retaining the same training, simulation, and evaluation protocols for comparable evaluation across settings.

(iii) *Exogenous covariate coverage.* The current benchmark does not yet include several dynamic covariates that may help explain conflict dynamics, including temperature, precipitation, food prices, displacement flows, and major political or international events. In particular, multimodal signals such as news streams or diplomatic-event data could support future research on event-stream models that combine structured event histories with external textual and geopolitical context. We are extending the raster interface to include weather covariates and plan to support additional exogenous event sources in future benchmark versions.

(iv) *Operational abstraction.* Our three updating regimes provide controlled abstractions of different forecasting settings, but real operational pipelines can involve irregular reporting schedules, delayed observations, retrospective corrections, and missing data. To assess this issue, we provide a reporting-delay ablation in the appendix that considers one such departure from the idealised setting within the data collection protocol of the ACLED dataset.

(v) *Historical evaluation.* For safety and reproducibility, CONFLICTCAST evaluates models only on historical conflicts and does not deploy forecasts for ongoing conflicts. As a result, the benchmark measures forecasting performance under retrospectively reconstructed information sets rather than effectiveness in live decision-support systems.

C RELATED WORK

Benchmark precedents. Our methodological template comes from scientific-ML benchmarks that established multi-horizon probabilistic evaluation in adjacent domains: medium-range weather forecasting (Rasp et al., 2020; 2024), climate-impact projection (Watson-Parris et al., 2022), probabilistic weather-and-climate prediction (Price et al., 2025), sub-seasonal-to-seasonal climate forecasting (Nathaniel et al., 2024), and general AI forecasting (Karger et al., 2025). From these we adopt the conventions of multi-horizon evaluation, ensemble-based scoring of predictive distributions, and per-cell aggregated operational metrics, and we instantiate this template for armed-conflict events.

Computational conflict research. Geo-referenced conflict-event datasets such as ACLED, UCDP, and SIGACT (Raleigh et al., 2010; Sundberg & Melander, 2013; Shaver & Bollfrass, 2023) record incidents with continuous time, location, and semantic attributes, making them a natural fit for spatiotemporal point processes. Nevertheless, most statistical forecasting work discretises space and time into bins and predicts binned counts or rates (Hegre et al., 2019; Radford, 2022; Brandt et al., 2022; Ge et al., 2022; Vestby et al., 2022; Malone, 2022). In this regime, tree ensembles such as XGBoost and random forests are often the strongest baselines (Vestby et al., 2022; Ettensperger, 2022), although recent work from forecasting competitions (Ettensperger, 2022) also explores deep sequence models, including RNNs and Transformers (Radford, 2022; Brandt et al., 2022; Macis et al., 2024; Walterskirchen et al., 2024). Following these findings, CONFLICTCAST adopts a multivariate XGBoost baseline with exogenous covariates for binned prediction; a binned-Transformer variant (Macis et al., 2024) was implemented to provide a neural-binned counterpart but failed to learn under fine-grained spatiotemporal sparsity, and is therefore excluded from the main roster (Tab. 12). More broadly, binning induces a resolution–sparsity trade-off: coarse bins obscure short-term triggering

and diffusion, while fine bins are extremely sparse, limiting operationally relevant early-warning signal. Related work from the statistics community models political violence with multivariate spatiotemporal Hawkes processes (Jun & Cook, 2024), but likelihood-based inference there typically relies on expensive grid approximations that do not scale to our large-area simulation setting. Conflict data remain relatively understudied in neural point processes, with few exceptions such as Okawa et al. (2022).

Spatiotemporal point processes and self-excitation. Spatiotemporal point processes (STPPs) model events occurring in continuous time and space via a conditional intensity. A major classical family is self-exciting STPPs, which extend Hawkes processes to capture triggering and diffusion in both time and space. These models underpin applications in seismology (Ogata, 1999), epidemiology (Amaral et al., 2023), and crime (Mohler et al., 2011), and offer an interpretable decomposition into background rates and history-dependent excitation terms. Comprehensive statistical overviews of self-exciting STPPs and their applications are given by Reinhart (2018).

Neural temporal point processes. Neural point processes (NPPs) increase modelling flexibility by parameterising the history dependence of intensities (or equivalent conditional inter-arrival distributions) with deep networks. Early works embed event histories with recurrent architectures, e.g. RMTPP (Du et al., 2016), while the Neural Hawkes Process (Mei & Eisner, 2017) introduces continuous-time recurrent dynamics to capture complex excitation and inhibition patterns. Attention-based models such as the Transformer Hawkes Process (Zuo et al., 2020) improve long-range dependency modelling and allow for parallel training, but incur quadratic computational complexity during simulation, since the intensity function must attend to the full history at every step of the thinning algorithm. Orthogonally, intensity-free approaches avoid explicit intensity integration by modelling the conditional inter-event time distribution directly (Shchur et al., 2020; Meng et al., 2024; Panos, 2024). This latter class enables direct sampling from the predicted distributions, bypassing the rejection sampling inherent to thinning algorithms; given our simulation-based evaluation framework, we adopt this intensity-free design for the temporal component of our Transformer-based baseline, TMD-HP.

Neural spatiotemporal point processes. Extending NPPs to continuous space adds representation and tractability challenges: likelihoods generally require multi-dimensional (space–time) integration, and simulation must repeatedly evaluate the intensity. A highly expressive approach combines continuous-time neural dynamics with normalising flows, as in Neural STPPs (Chen et al., 2021), which learn flexible conditional distributions over times and locations. A complementary line retains kernel structure for tractability: Deep Mixture Point Processes (Okawa et al., 2019) model the intensity as a mixture of kernels with weights produced by deep networks from rich context, enabling analytic integration for likelihood-based training. Recent work extends Okawa et al. (2019) by increasing flexibility via deep kernels and likelihood-free / integration-free estimators, e.g. DKMPP (Zhang et al., 2023), which we employ as the inhomogeneous neural baseline. Other designs pursue nonparametric intensities with latent dynamics while retaining closed-form integration, e.g. DeepSTPP (Zhou et al., 2022), which models the spatial component using both the locations of past events in the history and a set of representative points in space learned by the model; we adopt the same design choice in TMD-HP, as it facilitates learning persistent spatial hotspots (e.g. cities). Methods such as Spatio-Temporal Diffusion Point Processes (Yuan et al., 2023) argue that factorising time and space can be limiting and instead learn joint spatiotemporal distributions via diffusion-based generative modelling. CAM-HP follows the kernel-structured direction: we use neural networks to adapt event-level excitation parameters (including branching ratios) while preserving analytic compensators.

Exogenous context and covariate-aware STPPs. Conflict-event risk depends strongly on geographic and socio-environmental context. Recent work incorporates rich exogenous signals into spatiotemporal Hawkes-style models, including unstructured inputs such as maps and satellite imagery (Okawa et al., 2019; 2022). In both cases these covariates do not modulate excitation at the level of individual past events; instead, they enter either through an inhomogeneous Poisson background rate or as a multiplicative modulation of the excitation term. Related temporal-only work on deep covariate-TPPs aims to model event–covariate interactions with improved interpretability, e.g. Transformer-based covariate TPPs (Meng et al., 2024), from which TMD-HP takes inspiration in

its covariate-conditioning mechanism. Many covariate-aware STPP models still face a scalability trade-off between spatial resolution and tractable likelihood or simulation; in contrast, our approach conditions excitation parameters on high-resolution raster covariates at event locations, avoiding grid-based integration while retaining a generative point-process interpretation.

D EXTENDED RESULTS

This appendix reports per-fold extended results for each (model, dataset) cell, organised by metric: per-event log-likelihood (§D.1), energy score on the temporal and spatial marginals (§D.2), MAE on grid-aggregated counts (§D.3), and posterior predictive checks (§D.4). Simulation-based metrics are reported per dataset and per updating regime (FC, PU, NU). DKMPP is a non-self-exciting inhomogeneous Poisson process, so FC and PU collapse to NU and its FC and PU cells repeat the NU value;

D.1 PER-EVENT LOG-LIKELIHOOD

Per-event negative log-likelihood — main takeaways

1. Classical spatio-temporal point processes dominate the spatial term: CAM-HP or Hawkes achieves the best spatial NLL on all four datasets.
2. AutoSTPP is strongest temporally on three of four datasets: best temporal NLL on ACLED, SAGE, and UCDP; TMD-HP is best on SIGACT.
3. Overall Neural Point Processes win on NLL: AutoSTPP wins on ACLED and UCDP, CAM-HP on SAGE, and DeepSTPP on SIGACT.

All binned rows use 25 km cells with 7-day bins, the resolution that is also simulated; §E.5 reports the coarser alternative.

Table 2: Per-event NLL (total), best variant per model

lower is better

Model	ACLED	SAGE	UCDP	SIGACT
Conflict NPPs				
CAM-HP (power law)	2.27±0.45	**1.46**±0.45	2.35±0.28	−3.39±0.50
TMD-HP	2.66±1.23	2.65±0.57	2.83±0.52	−3.19±0.67
NSTPP community				
DKMPP (50 km)	5.73±0.97	3.75±0.36	22.85±24.70	0.33±0.62
DeepSTPP	3.03±0.77	2.93±0.61	3.38±0.42	**−3.52**±0.52
DSTPP	4.68±1.51	4.87±0.35	4.81±1.79	−1.28±0.49
AutoSTPP	**1.18**±1.98	2.37±1.08	**1.12**±1.40	−2.68±5.53
Neural STPP	10.47±0.91	9.29±0.37	9.91±1.22	4.50±0.20
Hawkes baselines				
Hawkes-Exp	3.07±0.46	2.18±0.56	3.73±0.90	−2.59±0.48
Hawkes-PL	2.98±0.34	2.18±0.55	3.65±1.21	−2.62±0.51
Binned				
Persistence (25 km, 7 d)	18.04±1.43	16.26±2.45	21.23±1.28	−1.10±1.15
Climatology (25 km, 7 d)	5.49±1.72	3.79±0.80	5.59±1.45	−2.40±0.47
Mov. average (25 km, 7 d)	7.52±1.88	5.59±1.43	7.62±0.24	−2.50±0.56
XGBoost (25 km, 7 d)	3.37±0.71	1.83±0.50	3.16±0.14	−2.74±0.50
Hurdle GBM (25 km, 7 d)	3.66±0.82	1.82±0.51	3.47±0.03	−2.74±0.50

Mean ± SD across rolling-quantile folds; units $100\,\mathrm{km}^2 \cdot \mathrm{day}$; lower is better; best per column bold and shaded. In Tab. 3, grey mark cells sit at $\log K$ (no mark head) and a dash means not applicable.

Table 3: Per-event NLL decomposition (T temporal, S spatial-conditional, M mark)

(a) ACLED and SAGE

lower is better

	ACLED			SAGE		
Model	T	S	M	T	S	M
Conflict NPPs						
CAM-HP	−1.96±0.75	3.03±0.38	**1.21**±0.10	−1.46±0.12	**2.66**±0.32	**0.26**±0.02
TMD-HP	−3.15±0.75	4.43±0.49	1.38±0.18	−1.44±0.20	3.21±0.36	0.87±0.02
NSTPP community						
DKMPP	−0.53±0.49	4.86±0.42	1.41±0.17	−1.18±0.15	4.59±0.19	0.34±0.04
DeepSTPP	−2.54±0.99	3.99±0.50	1.61±0.00	−1.46±0.15	3.30±0.48	1.10±0.00
DSTPP	−2.31±1.00	5.38±0.52	1.61±0.00	−1.43±0.24	5.20±0.19	1.10±0.00
AutoSTPP	**−3.77**±1.96	3.35±0.60	1.61±0.00	**−1.57**±0.11	2.84±0.97	1.10±0.00
Neural STPP	3.27±0.37	5.59±0.59	1.61±0.00	2.82±0.11	5.37±0.40	1.10±0.00
Hawkes baselines						
Hawkes-Exp	−0.76±0.48	**2.45**±0.37	1.38±0.11	−1.43±0.15	3.18±0.37	0.44±0.07
Hawkes-PL	−0.80±0.51	**2.45**±0.40	1.34±0.12	−1.43±0.15	3.19±0.37	0.42±0.04
Binned						
Persistence	−0.75±0.49	17.18±1.29	1.61±0.00	−1.43±0.16	16.59±2.29	1.10±0.00
Climatology	−0.76±0.49	4.64±1.30	1.61±0.00	−1.43±0.15	4.13±0.65	1.10±0.00
Mov. average	−0.75±0.49	6.66±1.42	1.61±0.00	−1.44±0.16	5.93±1.26	1.10±0.00
XGBoost	−0.76±0.50	2.92±0.25	**1.21**±0.18	−1.44±0.16	3.01±0.35	**0.26**±0.03
Hurdle GBM	−0.68±0.59	2.94±0.29	1.41±0.42	−1.44±0.16	3.00±0.33	**0.26**±0.05

(b) UCDP and SIGACT (SIGACT has one event type, $K=1$, hence no mark term)

lower is better

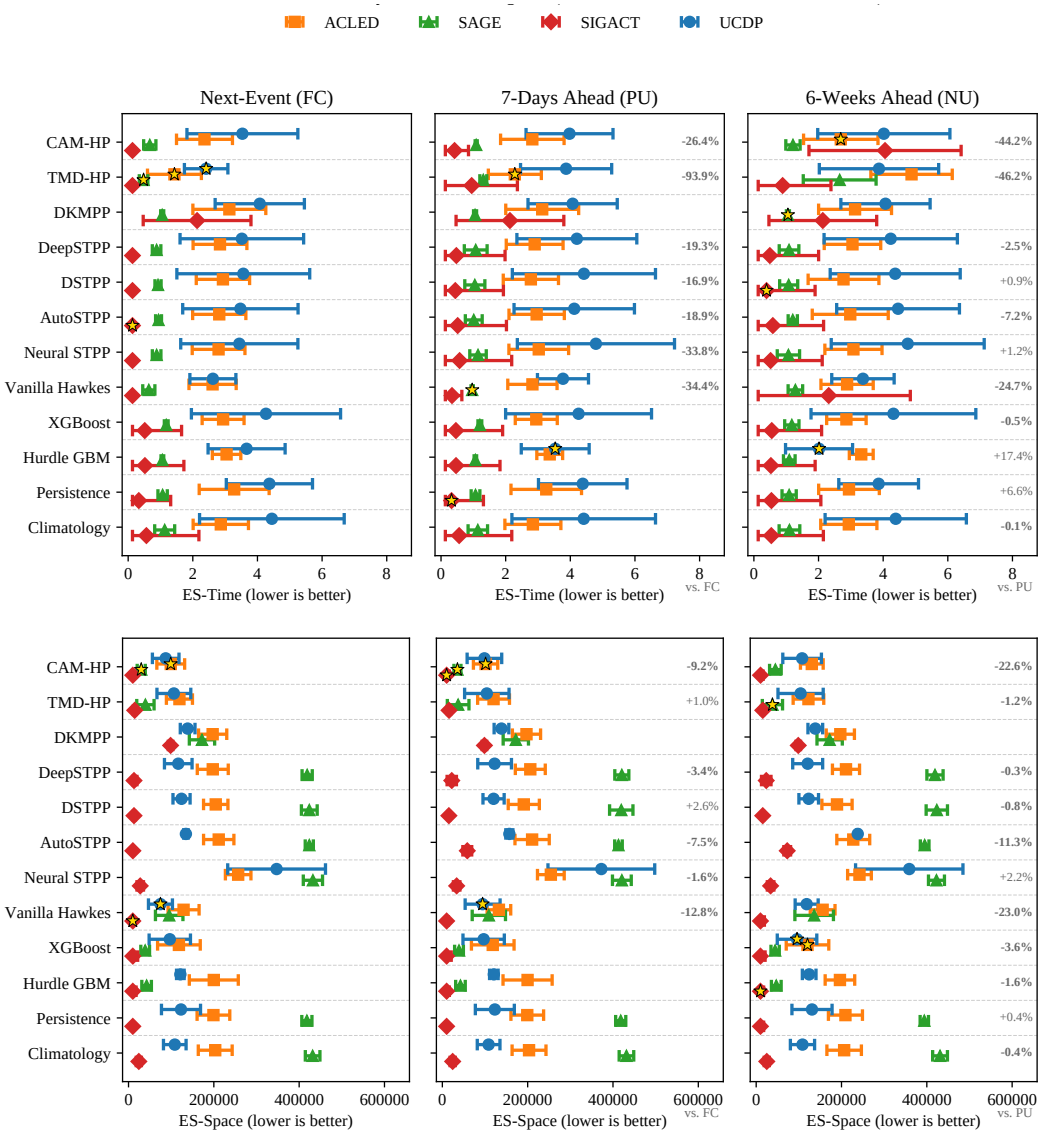
	UCDP			SIGACT		
Model	T	S	M	T	S	M
Conflict NPPs						
CAM-HP	−0.28±0.32	**2.05**±0.02	0.59±0.07	−3.30±0.40	**−0.09**±0.33	—
TMD-HP	−1.28±0.37	3.31±0.03	0.80±0.12	**−3.79**±0.58	0.60±0.31	—
NSTPP community						
DKMPP	17.51±23.79	4.88±1.57	**0.47**±0.66	−3.05±0.62	3.38±0.02	—
DeepSTPP	−0.36±0.02	3.01±0.08	1.10±0.00	−3.44±0.45	−0.08±0.27	—
DSTPP	−0.28±1.14	3.99±0.66	1.10±0.00	−3.26±0.45	1.98±0.11	—
AutoSTPP	**−2.87**±1.68	2.90±0.28	1.10±0.00	−2.69±5.42	0.01±0.51	—
Neural STPP	4.25±0.10	4.57±1.12	1.10±0.00	2.68±0.00	1.82±0.19	—
Hawkes baselines						
Hawkes-Exp	0.35±0.11	2.54±0.84	0.83±0.05	−3.26±0.39	0.67±0.24	—
Hawkes-PL	0.39±0.16	2.41±1.11	0.84±0.05	−3.28±0.41	0.66±0.25	—
Binned						
Persistence	0.43±0.04	19.70±1.31	1.10±0.00	−3.33±0.49	2.23±0.67	—
Climatology	0.39±0.13	4.11±1.32	1.10±0.00	−3.21±0.45	0.81±0.08	—
Mov. average	0.38±0.10	6.14±0.14	1.10±0.00	−3.31±0.49	0.81±0.18	—
XGBoost	0.38±0.11	2.12±0.01	0.66±0.02	−3.33±0.49	0.59±0.20	—
Hurdle GBM	0.39±0.10	2.18±0.17	0.90±0.04	−3.32±0.49	0.58±0.21	—

D.2 ENERGY SCORE

Energy score — main takeaways

1. Under teacher forcing (FC) the conflict models lead on both marginals: TMD-HP has the best ES-Time on ACLED, SAGE, and UCDP, and CAM-HP the best ES-Space on ACLED and SAGE, with vanilla Hawkes best on SIGACT.
2. Neural models without tailored spatial components—such as the spatial diffusion kernels and hotspot points used in CAM-HP and TMD-HP—perform significantly worse than all other models on ES-Space.

Figure 3: Energy score per model, dataset, and regime: ES-Time (top, days) and ES-Space (bottom, metres)



Markers: median over folds per dataset; whiskers: ±1 SD across folds; gold stars: best model per dataset and regime; numbers in the PU and NU panels: percentage change relative to the previous regime, averaged across datasets.

Table 4: Temporal energy score, ES-Time (days) under the FC, PU, and NU rollout regimes

(a) ACLED and SAGE

Model	ACLED			SAGE		
	FC	PU	NU	FC	PU	NU
Conflict NPPs						
CAM-HP	2.36±0.87	2.82±0.98	**2.69**±1.15	0.67±0.19	1.09±0.03	1.21±0.22
TMD-HP	**1.43**±0.83	**2.28**±0.82	4.88±1.26	**0.48**±0.13	1.31±0.11	2.65±1.13
NSTPP community						
DKMPP	3.13±1.13	3.13±1.13	3.13±1.13	1.05±0.05	1.05±0.05	**1.05**±0.05
DeepSTPP	2.84±0.83	2.90±0.88	3.05±0.87	0.88±0.13	1.07±0.35	1.09±0.30
DSTPP	2.93±0.83	2.78±0.85	2.77±1.10	0.92±0.10	1.05±0.31	1.08±0.28
AutoSTPP	2.82±0.82	2.96±0.86	2.98±1.18	0.94±0.09	1.01±0.26	1.20±0.14
Neural STPP	2.80±0.82	3.02±0.92	3.08±0.88	0.88±0.13	1.14±0.25	1.07±0.35
Hawkes baselines						
Vanilla Hawkes	2.61±0.73	2.83±0.76	2.88±0.81	0.64±0.19	**0.96**±0.04	1.28±0.22
Binned						
Persistence	3.28±1.08	3.26±1.09	2.94±0.94	1.07±0.14	1.07±0.13	1.09±0.22
Climatology	2.87±0.86	2.84±0.87	2.94±0.87	1.13±0.31	1.14±0.30	1.10±0.32
Mov. average	3.11±0.88	3.15±0.89	3.03±0.92	1.11±0.28	1.10±0.27	1.09±0.33
XGBoost	2.94±0.65	2.95±0.65	2.86±0.61	1.18±0.06	1.20±0.06	1.17±0.22
Hurdle GBM	3.04±0.44	3.36±0.40	3.32±0.37	1.06±0.04	1.06±0.04	1.10±0.18

lower is better

(b) UCDP and SIGACT

Model	UCDP			SIGACT		
	FC	PU	NU	FC	PU	NU
Conflict NPPs						
CAM-HP	3.53±1.72	3.98±1.34	4.02±2.04	0.01±0.01	0.42±0.43	4.06±2.35
TMD-HP	**2.41**±0.67	3.87±1.41	3.87±1.85	0.02±0.01	0.95±1.41	0.89±1.50
NSTPP community						
DKMPP	4.07±1.38	4.07±1.38	4.07±1.38	2.13±1.67	2.13±1.67	2.13±1.67
DeepSTPP	3.52±1.91	4.20±1.85	4.23±2.06	0.01±0.00	0.47±1.51	0.49±1.51
DSTPP	3.56±2.06	4.42±2.21	4.37±2.01	0.01±0.01	0.44±1.49	**0.39**±1.50
AutoSTPP	3.47±1.78	4.12±1.86	4.46±1.90	**0.01**±0.00	0.52±1.51	0.59±1.57
Neural STPP	3.44±1.81	4.79±2.43	4.76±2.37	0.01±0.00	0.57±1.61	0.52±1.60
Hawkes baselines						
Vanilla Hawkes	2.62±0.71	3.77±0.79	3.37±0.97	0.01±0.01	0.34±0.30	2.32±2.52
Binned						
Persistence	4.37±1.33	4.38±1.37	3.86±1.23	0.33±0.99	**0.33**±0.99	0.54±1.53
Climatology	4.45±2.24	4.41±2.22	4.39±2.19	0.56±1.62	0.56±1.63	0.54±1.61
Mov. average	4.58±2.34	4.57±2.32	4.55±2.26	0.56±1.73	0.55±1.73	0.53±1.74
XGBoost	4.26±2.31	4.26±2.26	4.32±2.55	0.51±1.14	0.46±1.45	0.55±1.54
Hurdle GBM	3.66±1.19	**3.53**±1.05	**2.02**±1.04	0.52±1.21	0.46±1.36	0.53±1.37

lower is better

Median ± SD across rolling-quantile folds (ACLED 5, SAGE 3, UCDP 2, SIGACT 4); lower is better; best per column bold and shaded. DKMPP is history-free (inhomogeneous Poisson), so its FC and PU cells reuse the NU value and are greyed. AutoSTPP and Neural STPP use the wrapper-level samplers of §F.4; Neural STPP rows use the physical-time-unit checkpoints.

Table 5: Spatial energy score, ES-Space (km) under the FC, PU, and NU rollout regimes

(a) ACLED and SAGE

Model	ACLED			SAGE		
	FC	PU	NU	FC	PU	NU
Conflict NPPs						
CAM-HP	**99.0**±32.6	**101.3**±28.5	130.4±26.6	**29.8**±9.4	**35.4**±8.2	45.0±13.6
TMD-HP	119.3±30.5	120.0±37.3	122.5±35.9	39.6±20.3	37.0±25.4	**38.1**±23.7
NSTPP community						
DKMPP	197.2±33.1	197.2±33.1	197.2±33.1	172.2±29.7	172.2±29.7	172.2±29.7
DeepSTPP	197.5±36.3	206.2±35.1	210.5±31.8	418.8±11.9	420.7±16.4	419.2±19.0
DSTPP	204.5±28.6	190.9±36.5	189.4±35.4	423.8±18.5	419.4±27.4	423.4±25.0
AutoSTPP	211.5±35.5	210.4±40.2	227.6±38.6	423.7±9.5	412.9±8.9	394.9±10.1
Neural STPP	257.0±29.9	254.3±31.3	242.1±27.8	432.2±22.6	420.8±22.0	422.6±18.7
Hawkes baselines						
Vanilla Hawkes	129.1±36.4	132.6±27.7	155.8±29.0	95.3±32.2	109.0±39.0	135.9±45.2
Binned						
Persistence	199.1±38.3	199.1±38.3	209.2±39.9	418.1±11.9	418.1±11.9	394.1±9.8
Climatology	203.3±39.7	203.2±39.7	206.2±40.4	431.6±17.1	431.6±17.1	431.0±17.3
Mov. average	206.1±35.7	206.2±35.8	205.8±35.7	424.0±18.2	424.0±18.2	424.2±15.4
XGBoost	118.2±50.0	118.1±50.0	**120.3**±49.9	39.0±10.2	39.0±10.3	44.4±9.7
Hurdle GBM	199.9±57.4	199.8±57.6	196.3±34.6	42.4±11.2	42.4±11.2	47.1±11.4

lower is better

(b) UCDP and SIGACT

Model	UCDP			SIGACT		
	FC	PU	NU	FC	PU	NU
Conflict NPPs						
CAM-HP	87.2±31.2	98.6±40.5	107.7±45.3	3.0±0.6	**3.9**±6.1	10.2±6.0
TMD-HP	106.3±39.7	104.3±52.1	104.1±53.0	14.5±3.4	15.7±3.7	15.5±3.4
NSTPP community						
DKMPP	138.6±17.2	138.6±17.2	138.6±17.2	98.6±5.3	98.6±5.3	98.6±5.3
DeepSTPP	116.6±32.4	122.5±39.5	120.7±35.1	12.9±6.9	22.1±10.9	23.6±11.1
DSTPP	124.3±20.1	120.1±24.8	123.1±23.0	12.8±3.4	15.3±2.4	15.4±2.7
AutoSTPP	134.5±5.2	157.1±8.3	238.1±2.3	10.2±1.6	58.5±10.4	72.8±7.9
Neural STPP	347.1±114.6	372.7±125.1	358.8±125.9	27.5±7.7	33.5±7.6	33.8±7.7
Hawkes baselines						
Vanilla Hawkes	**74.9**±28.1	**94.4**±40.9	118.3±27.0	**2.8**±0.6	4.7±4.2	9.1±9.2
Binned						
Persistence	123.0±45.8	123.0±45.7	130.8±47.0	4.5±1.5	4.5±1.5	7.5±7.5
Climatology	108.4±26.5	108.5±26.6	108.5±28.4	24.2±6.1	24.2±6.1	24.8±6.0
Mov. average	117.8±35.5	117.8±35.5	120.2±30.8	14.6±7.0	14.6±6.9	15.1±7.8
XGBoost	96.8±48.4	96.8±48.4	**95.8**±46.3	6.1±11.4	6.3±10.8	9.1±10.8
Hurdle GBM	121.5±8.9	120.7±10.0	124.2±16.0	5.3±7.7	6.1±6.4	**7.2**±7.4

lower is better

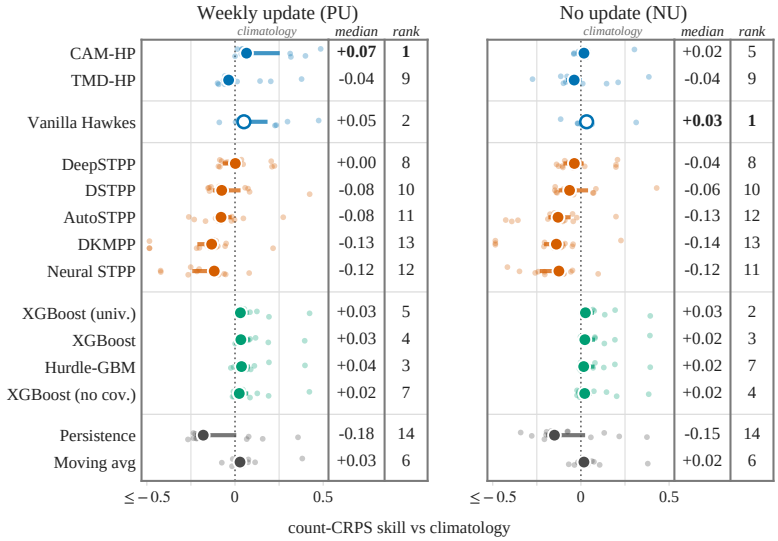
Median ± SD across rolling-quantile folds (ACLED 5, SAGE 3, UCDP 2, SIGACT 4); lower is better; best per column bold and shaded. DKMPP is history-free (inhomogeneous Poisson), so its FC and PU cells reuse the NU value and are greyed. AutoSTPP and Neural STPP use the wrapper-level samplers of §F.4; Neural STPP rows use the physical-time-unit checkpoints.

D.3 GRID-AGGREGATED COUNTS

Grid-aggregated counts — main takeaways

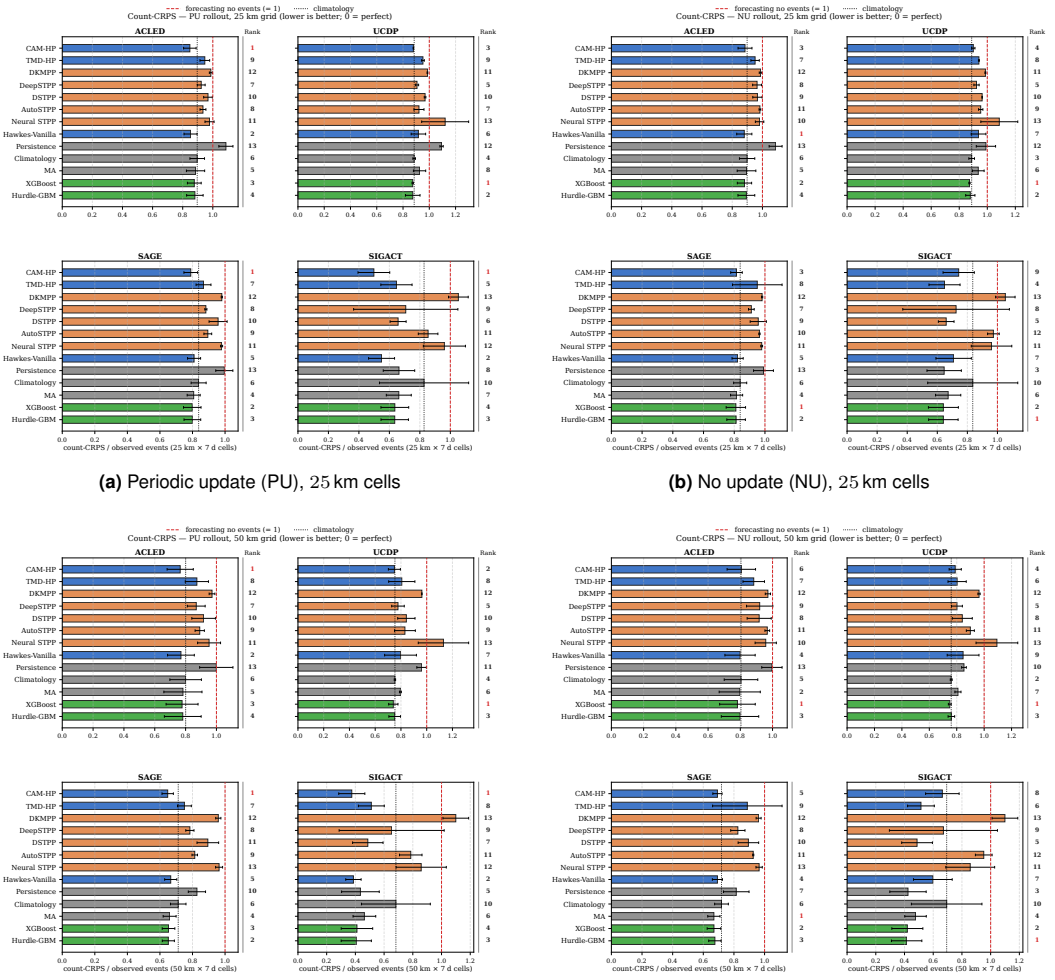
1. With weekly updates CAM-HP forecasts grid counts best: lowest count-CRPS on ACLED, SAGE, and SIGACT and third on UCDP.(Fig. 5a).
2. Without updates the lead passes to the binned models: XGBoost ranks first on SAGE and UCDP and second on ACLED and SIGACT, while CAM-HP drops to third or lower (ninth on SIGACT, where Hurdle-GBM leads); the 50 km cells tell the same story (Figs. 5b and 5d).
3. The MAE$^+$ grids (Fig. 6) are kept for continuity with prior work and interpretability. (§G.3).

Figure 4: Count-CRPS skill against climatology per model and rollout regime



Skill is $1 - \mathrm{CRPS}_{\mathrm{model}}/\mathrm{CRPS}_{\mathrm{climatology}}$ per dataset–fold cell on 25 km × 7-day cells (§G.3); 0 is climatology and higher is better. Layout, rows and colours as in Fig. 2: each light dot is one cell, the bar spans the interquartile range, the marker is the median, and the median and its rank within the panel are printed at the right (bold for the best model). FC is omitted because it fixes the simulated count to the observed one; hollow markers at the axis floor denote medians below -0.5.

Figure 5: Count-CRPS on grid-aggregated counts per rollout regime, at 25 km (top) and 50 km (bottom) cells



(a) Periodic update (PU), 25 km cells

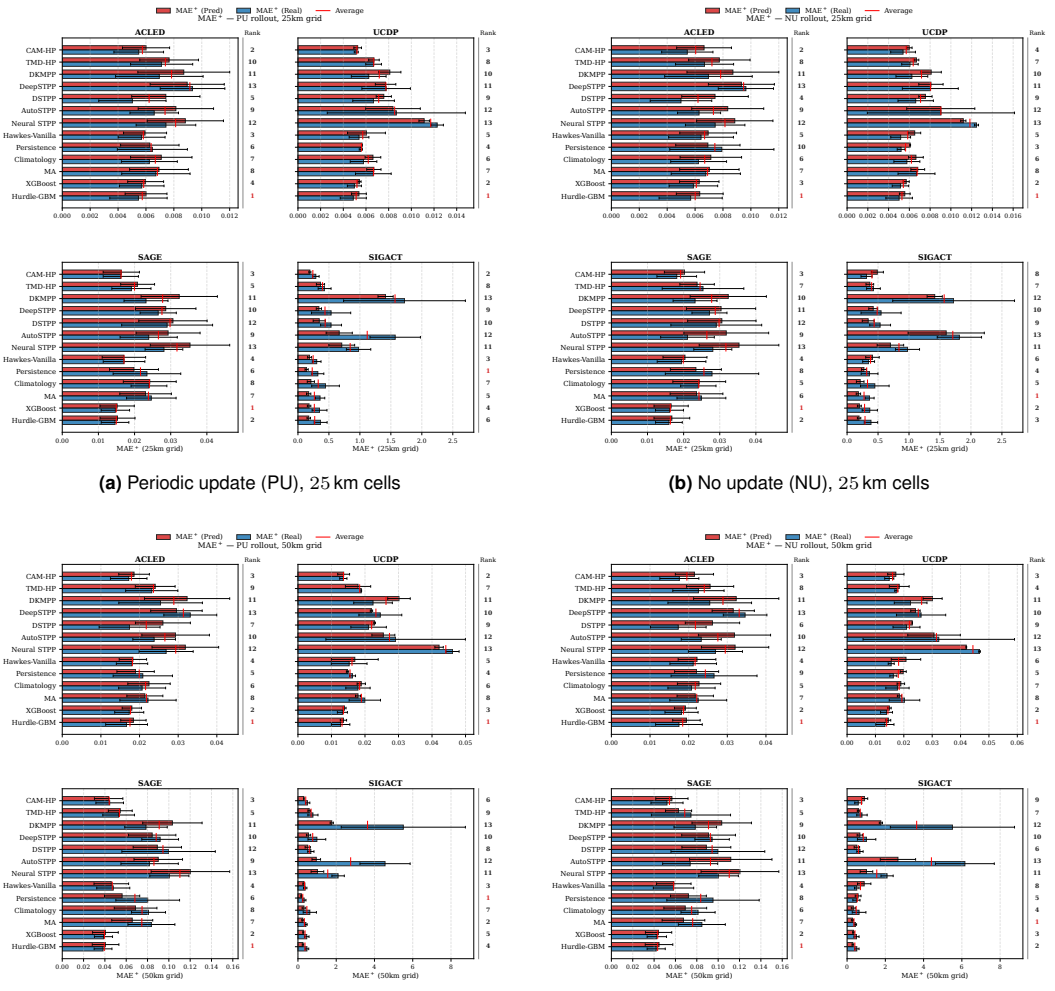
(b) No update (NU), 25 km cells

(c) Periodic update (PU), 50 km cells

(d) No update (NU), 50 km cells

Each bar is the count-CRPS (§G.3) summed over all cell–weeks of the fixed study-region grid and divided by the number of observed events, mean ± SD over the dataset's folds: 0 is a perfect forecast and 1 (red dashed line) is the score of forecasting no events at all; the dotted line marks climatology. Lower is better; the right-most column ranks models within the dataset. Levels depend on the cell size, so compare models within a resolution. The fully conditioned (FC) regime is omitted: FC draws one event per observed event, so simulated counts equal observed counts by construction. DKMPP is history-free (inhomogeneous Poisson), so its PU and NU values coincide.

Figure 6: MAE^{+} on grid-aggregated counts per rollout regime and cell size



(a) Periodic update (PU), 25 km cells

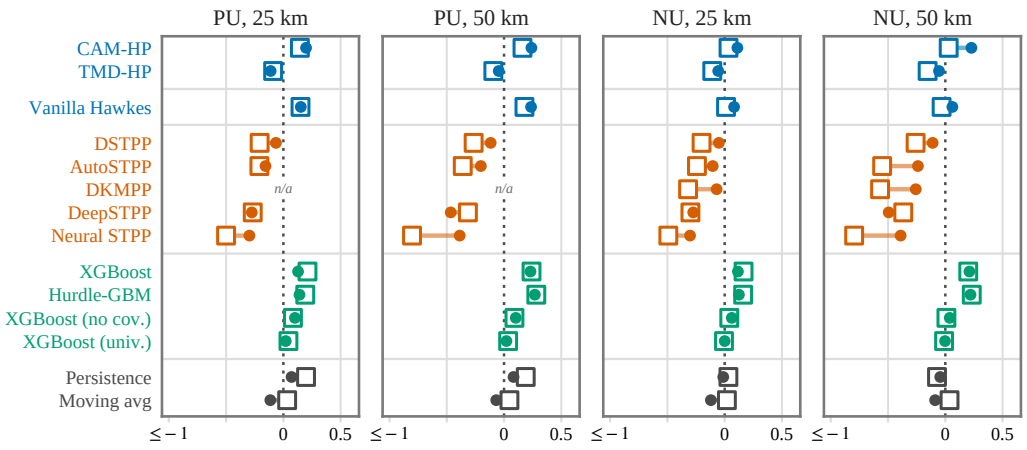
(b) No update (NU), 25 km cells

(c) Periodic update (PU), 50 km cells

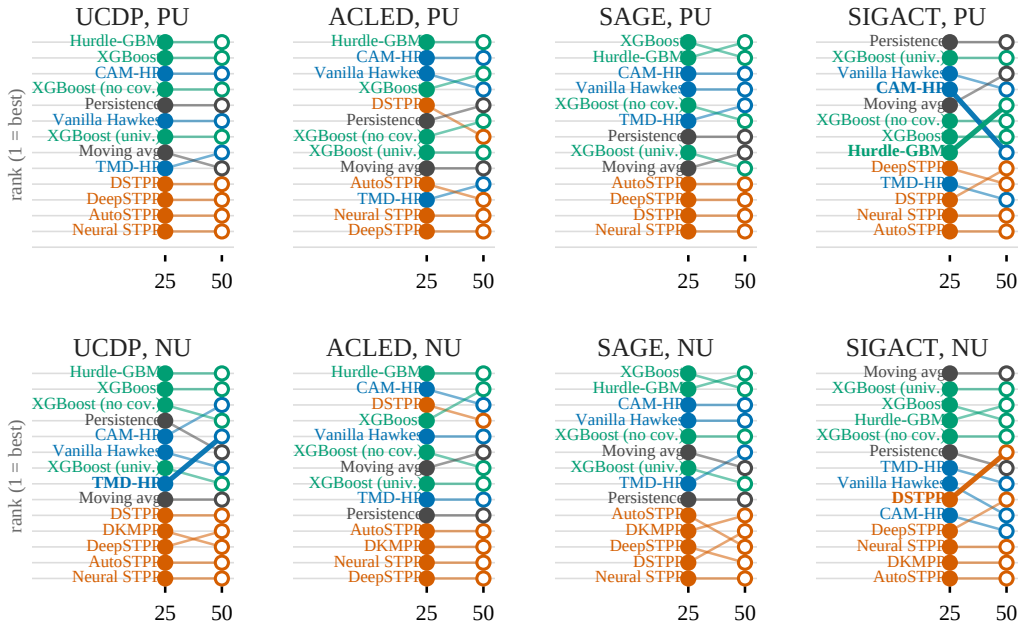
(d) No update (NU), 50 km cells

Each panel ranks models within its dataset (right-most column); model order follows the family ordering used throughout. Lower is better. DKMPP is history-free, so its PU and NU values coincide; FC is omitted because history anchoring keeps simulated totals close to observed counts. MAE^{+} levels are not comparable across resolutions (coarsening to 50 km raises the level about $3\times$ for every model, §E.6); orderings agree across resolutions with a median Spearman correlation of 0.94, except SIGACT $\mathrm{MAE}^{+}_{\mathrm{Real}}$, where CAM-HP is first at 25 km and sixth at 50 km.

Figure 7: Grid-aggregated counts: skill against climatology and resolution stability



(a) Median skill per model on $\mathrm{MAE}^{+}_{\mathrm{Real}}$ (●) and $\mathrm{MAE}^{+}_{\mathrm{Pred}}$ (□); a square left of its dot marks an under-active simulator, right of it an over-active one.



(b) Rank on the mean of the two skills per dataset and regime (left column: 25 km cells, right column: 50 km cells); models that move three or more places are named in bold.

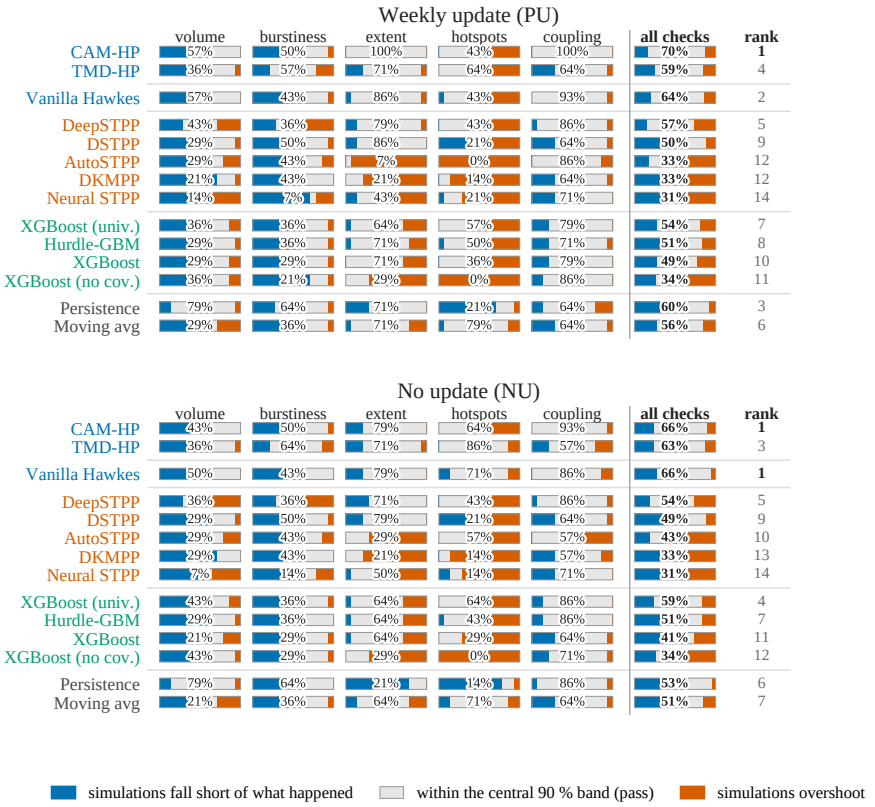
Skill is $1 - \mathrm{MAE}^{+}_{\mathrm{model}} / \mathrm{MAE}^{+}_{\mathrm{climatology}}$ per dataset–fold cell, which removes the dataset scale and the resolution level of Fig. 6. $\mathrm{MAE}^{+}_{\mathrm{Real}}$ normalises by the observed count and forgives too few events, $\mathrm{MAE}^{+}_{\mathrm{Pred}}$ by the predicted count and forgives too many, so the gap between the two skills tracks the count bias (Spearman $+0.57$ at 25 km, $+0.47$ at 50 km with the log ratio of simulated to observed events). Colours: ● conflict-tailored NPPs and vanilla Hawkes, ● off-the-shelf NSTPPs, ● binned regressors, ● naive baselines; DKMPP is history-free and has no PU entry.

D.4 Posterior predictive checks

Posterior predictive checks — main takeaways

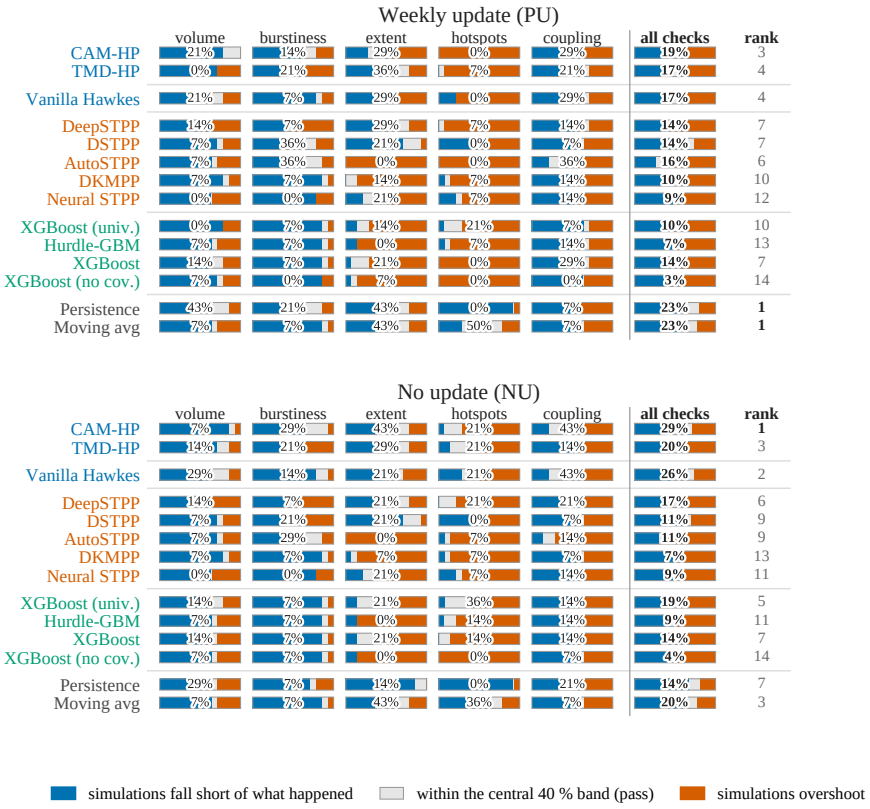
1. No simulator is calibrated over operational statistics: the best pooled `PPC-in-90` pass rate is 70% (CAM-HP, PU) against a 90% target, and under the stricter `PPC-in-40` band the best model reaches 29% (CAM-HP, NU) against 40%, with the tree ensembles at 7–19% (Figs. 8 and 9).
2. violetThe conflict-tailored models and vanilla Hawkes are the best calibrated: CAM-HP and vanilla Hawkes have the highest pooled `PPC-in-90` pass shares in both regimes (70% and 64% with weekly updates, 66% each without), and TMD-HP follows at 63% without updates and 59% with them, just behind persistence (60%; Figs. 8 and 10);
3. The failure modes are shared across families: nearly every model under-generates event volume and daily bursts (blue ratio in the volume and burstiness column in Fig. 8)

Figure 8: Where the simulators miss, per check (`PPC-in-90` band)



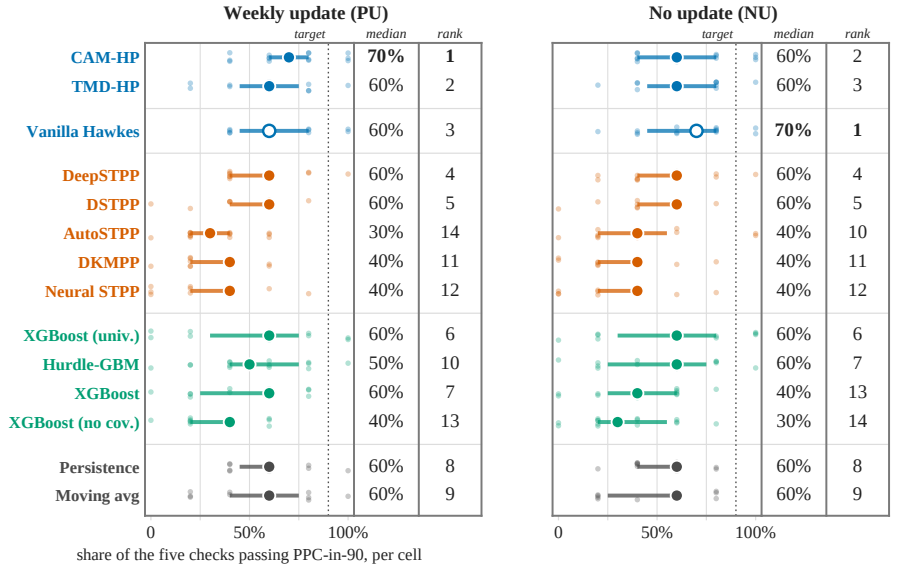
For each model and check, the bar shows how the 14 dataset–fold cells split: blue, the realised statistic lies above the central 90% of the simulated distribution (simulations fall short of what happened); grey, inside it (pass); orange, below it (simulations overshoot). The printed number is the share of cells passing, the last bar pools the five checks, and the rank orders models by that pooled share within the regime (1 = best; ties share a rank). Rows are grouped by family and ordered by PU pass rate within family; labels coloured by family (● conflict-tailored NPPs and vanilla Hawkes, ● off-the-shelf NSTPPs, ● binned regressors, ● naive baselines). Volume = event count, burstiness = maximum daily count, extent = maximum pairwise distance, hotspots = connected components within 50 km, coupling = space–time correlation; definitions in §G.4.

Figure 9: Where the simulators miss, per check (stricter `PPC-in-40` band)



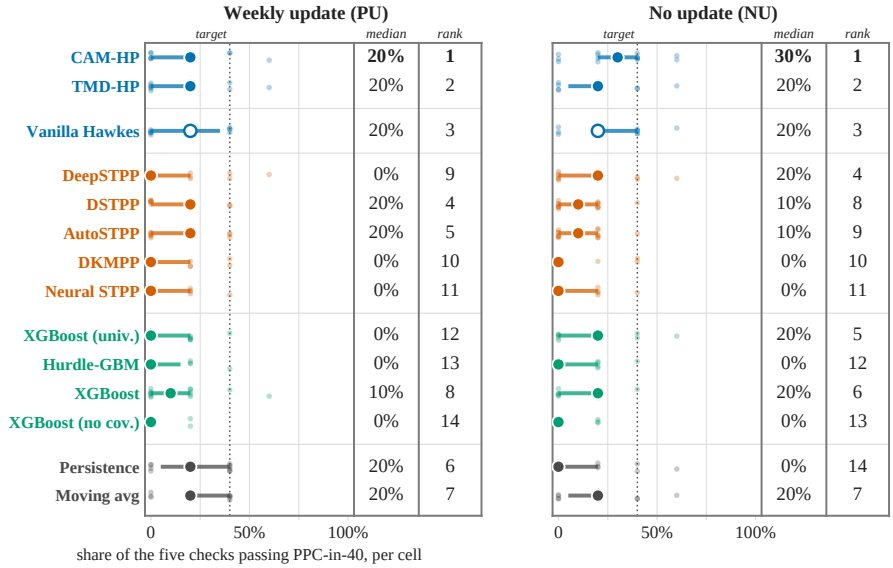
As Fig. 8 with the central 40% band: grey now means the realised statistic lies inside the central 40% of the simulated distribution, whose calibrated pass share is 40%.

Figure 10: Per-cell `PPC-in-90` pass rate by model and regime



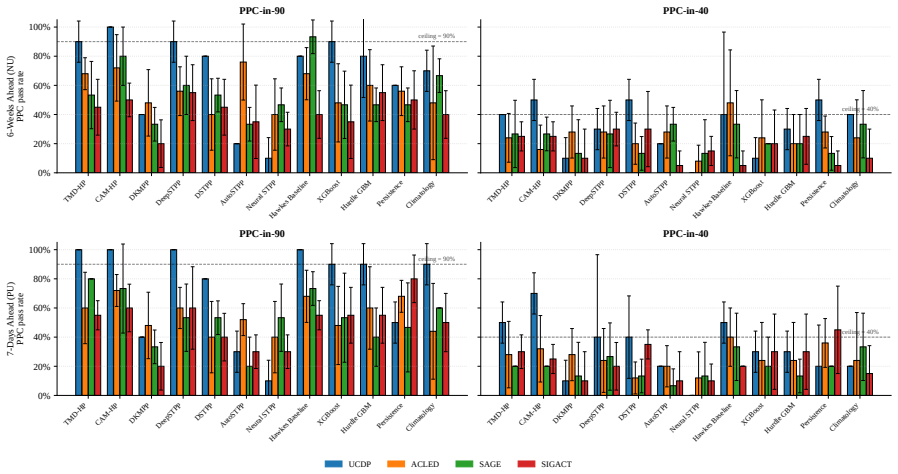
Share of the five checks passing in each dataset–fold cell (light dots), interquartile range (bar) and median (marker), with the median and its rank printed at the right; the dotted line is the 90% calibration target. Families and colours as in Fig. 8. With five checks per cell the rate is quantised to fifths, which is why several medians tie.

Figure 11: Per-cell `PPC-in-40` pass rate by model and regime



Same layout as Fig. 10 for the stricter band (realised statistic inside the central 40% of the simulated distribution); the dotted line is the 40% calibration target.

Figure 12: PPC pass rates by model, regime, and dataset



Within each panel, every model is split into four bars (one per dataset) with whiskers at ± 1 SD across folds. Left column: `PPC-in-90`, share of checks whose realised statistic lies inside the simulator's central 90% interval (the two-sided $\alpha = 0.05$ PPC test); right column: `PPC-in-40`, the stricter central 40% interval. Dashed lines mark the calibration ceilings (90% and 40%) of a perfectly calibrated model.

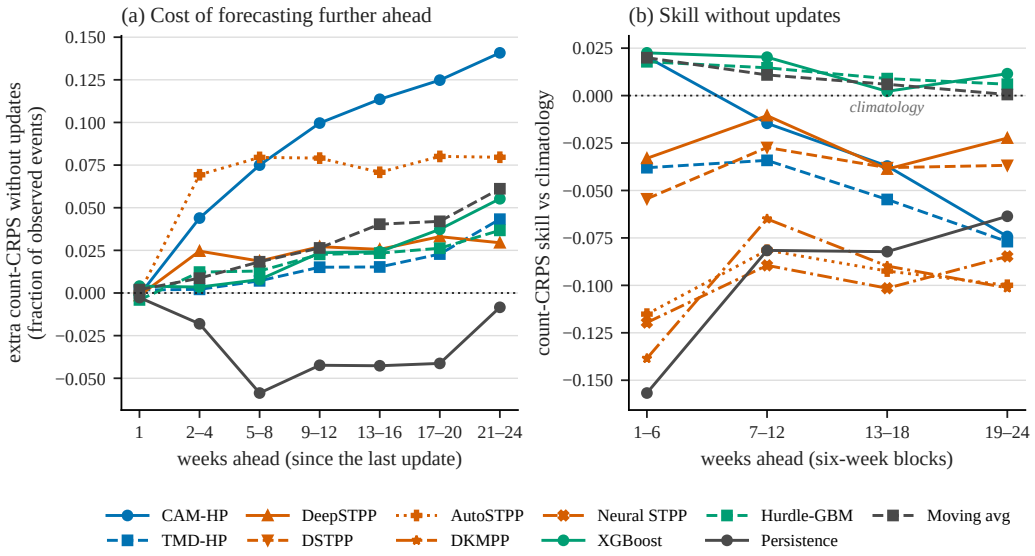


Figure 13: **Count-CRPS over 24-week rollouts.** Scores use 25 km × 7-day cells and are aggregated by the median over the cells that extend to 24 weeks. (a) Cost of forecasting further ahead, measured as count-CRPS without updates minus count-CRPS with weekly updates, normalized by the number of observed events and grouped by weeks ahead. DKMPP, DSTPP, and Neural STPP remain within ± 0.002 and are omitted. (b) Count-CRPS skill of no-update forecasts relative to climatology, aggregated in six-week blocks; 0 denotes equal performance to climatology.

E Ablations and sensitivity studies

E.1 Multi-horizon evaluation

E.1.1 How does forecast quality change with the forecast horizon?

We study how forecast quality changes with the forecast horizon, and in particular the cost of rolling a model forward for multiple weeks without incorporating new observations. While the main results use a 42-day (six-week) forecast window, here we extend each fold's test window to 12 and 24 weeks and re-simulate every published checkpoint under FC, PU, and NU using the paper's original seeds and 100 trajectories

We score each week of the rollout separately using count-CRPS on 25 km cells (§G.3), against the events observed in that week. PU always forecasts one week ahead before updating, whereas NU forecasts progressively farther from its last observation. Thus, in rollout week k, the difference NU−PU measures the cost of forecasting k weeks ahead rather than repeatedly forecasting one week ahead. As a sanity check, PU and NU issue the same forecast in week 1, since the first PU update occurs only at the end of that week; their scores differ by a median of only 0.5% (at most 3.5%), consistent with simulation noise.

E.1.2 Results.

The cost of forecasting further ahead grows for models that rely on recent history. The gap between no-update and weekly-update forecasts generally widens with the number of weeks forecast ahead (Fig. 13a). For CAM-HP, the excess count-CRPS grows from 0.04 of the observed event count in weeks 2–4 to 0.14 in weeks 21–24. AutoSTPP deteriorates earlier, reaching a gap of 0.07 within four weeks and remaining near 0.08, while XGBoost, Hurdle-GBM, and TMD-HP degrade more gradually, reaching 0.06, 0.04, and 0.04, respectively, by weeks 21–24. By contrast, DKMPP, DSTPP, and Neural STPP show essentially no gap between NU and PU, indicating that weekly updates provide little benefit to their forecasts in this experiment; for Persistence, the gap is negative.

Without updates, CAM-HP loses its advantage over climatology after about six weeks. CAM-HP's no-update skill relative to climatology falls from 0.02 in weeks 1–6 to -0.01, -0.04, and -0.07 over the next three six-week blocks (Fig. 13b). XGBoost, Hurdle-GBM, and the moving average, in contrast, remain near climatology throughout, with skill between 0.00 and 0.02. The deterioration is strongest on SIGACT: CAM-HP's no-update count-CRPS rises from 0.51 of the observed count in week 1 to 0.75 by week 4 and 0.98 by week 24, whereas its weekly-updated forecast remains between 0.39 and 0.51.

E.2 Reporting delay under periodic updates

E.2.1 How robust are the gains from periodic updating to realistic reporting delays?

Our PU regime assumes that all events up to each weekly update boundary are immediately available to the forecaster. In practice, conflict-event data can arrive with a short delay: ACLED releases new events in weekly batches, and updates may lag the end of the reporting period by one or two days. We therefore repeat the PU simulations of CAM-HP and TMD-HP on ACLED and UCDP folds with a two-day reporting lag. At each weekly boundary, the simulator observes only events that occurred more than two days earlier and rolls forward through the unobserved gap, marginalizing over the missing recent history rather than treating it as event-free.

E.2.2 Results

A two-day reporting lag has a small effect on forecast accuracy. For CAM-HP, the median count-CRPS and count MAE increase by 1.5% and 2.0%, while TMD-HP changes by -0.2% and $+0.5\%$, and the energy-score changes are mixed (Tab. 6). Only the 2.0% increase in the count error of CAM-HP is significant: it occurs in all 7 folds (sign test $p = 0.016$), whereas no other change in Tab. 6 is consistent enough across the seven folds to reach $p < 0.05$.

Table 6: **Effect of a two-day reporting lag on weekly-update forecasts** (released weights). Median relative change over the 7 folds (ACLED 5, UCDP 2); in parentheses, the number of folds in which the lagged forecast is worse. * sign test $p < 0.05$.

	Count-CRPS	Count MAE (observed cells)	ES-time	ES-space
CAM-HP	$+1.5\%$ (6/7)	$+2.0\%$ (7/7)*	$+4.9\%$ (4/7)	-2.2% (1/7)
TMD-HP	-0.2% (2/7)	$+0.5\%$ (6/7)	-3.7% (3/7)	0.0% (3/7)

E.3 Ablation on covariate effect

We study covariates by comparing the released CAM-HP variants with and without covariates, each searched separately, on per-event test log-likelihood (Tab. 7). Covariates improve the likelihood on SAGE ($+0.08$ nats per event, 3 of 3 folds) and UCDP ($+0.14$, 2 of 2 folds) and leave it essentially unchanged on ACLED (-0.01, 2 of 5 folds).

Table 7: **Covariate ablation for CAM-HP.** Per-event test log-likelihood (nats per event, higher is better) of the released CAM-HP variants with and without covariates, each searched separately; mean $\pm$ SD over folds. The covariate gain is positive when the covariates help; the last column counts the folds in which they did.

Dataset	With covariates (paper)	Without covariates	Covariate gain	Folds helped
ACLED	-2.27 ± 0.45	-2.26 ± 0.36	-0.01	2 of 5
SAGE	-1.46 ± 0.45	-1.54 ± 0.49	$+0.08$	3 of 3
UCDP	-2.35 ± 0.28	-2.50 ± 0.24	$+0.14$	2 of 2

E.4 Ablations of conflict-model design choices

Several of our conflict-specific design choices encode assumptions about how violence propagates across space and time. In particular, prior work has modeled conflict diffusion using ETAS-style and

power-law spatial triggering kernels (van den Hengel & Franses, 2020). We therefore ablate two such choices: CAM-HP's multi-scale triggering kernel and TMD-HP's heavier-tailed spatial kernel.

E.4.1 Does CAM-HP benefit from multi-scale triggering?

We compare a single triggering kernel with CAM-HP's multi-scale mixture, under both exponential and power-law temporal decay. On ACLED, the mixture improves test NLL by 1.05 nats per event with exponential decay and 1.46 with power-law decay, with gains in every fold (Tab. 8). Power-law decay also outperforms exponential decay, and the full-spectrum mixture is selected in 9 of the 10 mixture searches. Thus, CAM-HP's performance depends strongly on representing conflict triggering at multiple spatial and temporal scales.

E.4.2 Does a heavier-tailed spatial kernel help TMD-HP?

Motivated by prior work allowing long-range spatial triggering in conflict events, we replace TMD-HP's Gaussian spatial mixture with a heavier-tailed Laplace mixture while keeping the search space and evaluation fixed. On ACLED, the Laplace variant improves spatial log-likelihood in 4 of 5 folds, for an event-weighted gain of 0.16 nats per event and 0.22 nats on total log-likelihood (Tab. 9). The corresponding 95% bootstrap intervals include zero, however, so the evidence favors Laplace directionally but does not distinguish the two kernels decisively.

Table 8: CAM-HP kernel ablation (ACLED, 5 folds): test NLL, nats per event

(a) Mean over folds

Temporal decay	Single	Mixture	Mixture gain
Exponential	4.69 ± 1.54	3.64 ± 1.07	$+1.05$ (5/5)
Power-law (paper)	4.62 ± 1.41	$\mathbf{3.16 \pm 1.18}$	$+1.46$ (5/5)
Power-law gain	$+0.06$ (2/5)	$+0.48$ (5/5)	

(b) Per fold

	Single kernel		Multi-scale mixture		Mixture gain
Fold	exponential	power-law	exponential	power-law	(power-law)
0	7.41	7.12	5.54	**5.25**	$+1.87$
1	4.18	4.22	3.43	**2.90**	$+1.31$
2	3.71	3.75	3.08	**2.57**	$+1.18$
3	4.33	4.18	3.19	**2.53**	$+1.65$
4	3.82	3.86	2.96	**2.56**	$+1.30$
Mean	4.69	4.62	3.64	3.16	$+1.46$

Lower is better; a gain is the NLL reduction from the paper's choice, with the number of folds (of 5) in which it helped. Levels predate the evaluation-history fix of §E.6; the paired differences are unaffected.

Table 9: TMD-HP spatial-kernel ablation (ACLED, 5 folds): per-event test log-likelihood, nats per event

		Spatial term			Total		
Fold	Events	Laplace	Gaussian	Laplace gain	Laplace	Gaussian	Laplace gain
0	167	$\mathbf{-5.27}$	-5.70	$+0.43$	$\mathbf{-4.61}$	-5.00	$+0.39$
1	132	$\mathbf{-4.81}$	-4.88	$+0.07$	$\mathbf{-3.68}$	-3.82	$+0.13$
2	326	$\mathbf{-4.13}$	-4.49	$+0.36$	$\mathbf{-1.89}$	-2.61	$+0.72$
3	411	$\mathbf{-4.20}$	-4.26	$+0.06$	$\mathbf{-1.45}$	-1.54	$+0.09$
4	321	-4.32	$\mathbf{-4.31}$	-0.01	-2.48	$\mathbf{-2.32}$	-0.16
Event-weighted mean	1357	$\mathbf{-4.40}$	-4.57	$+0.16$	$\mathbf{-2.41}$	-2.63	$+0.22$
95 % interval of the gain				$[-0.04, +0.38]$			$[-0.10, +0.58]$
Laplace better in				4 of 5 folds			4 of 5 folds

Higher is better; identical search space, seeds and evaluation; better arm in bold, Laplace gain = Laplace − Gaussian. Each arm selects its own winner, so only the spatial term isolates the kernel family; the interval is a fold-clustered bootstrap over the 1357 paired test events.

E.5 Resolution of the binned baselines

The finer grid is better for both tree models on every dataset. The binned baselines are fitted at 25 km cells with 7-day bins throughout; refitting XGBoost and Hurdle-GBM at 50 km cells with 14-day bins, with their own search over the same space, raises the per-event NLL by 0.3–0.6 nats on every dataset (Tab. 10), so the finer resolution is used for the likelihood tables and for the simulations.

Table 11: **Sensitivity of the PPC pass rates to the number of simulated paths** (2080 checks). Every check is recomputed from B of the 100 stored paths; per-check quantities are exact (hypergeometric), pooled shares are medians $\pm$ the 95 % half-width over 2000 joint resamples. †Lattice effect: at $B = 40$ and 80 a p-value falls exactly on the in-90 boundary and counts as a pass; $B = 20$ and 60 are convention-free.

		Verdict flips (%)		PPC-in-90 pass share (%)		PPC-in-40 pass share (%)	
B	median $\lvert p_B - p_{100}\rvert$	in-90	in-40	PU	NU	PU	NU
20	0.040	11	14	52.3 ± 1.7	51.3 ± 1.6	26.5 ± 1.9	25.7 ± 1.8
40	0.020	8	8	$59.8 \pm 1.3^{\dagger}$	$58.7 \pm 1.3^{\dagger}$	25.0 ± 1.5	23.8 ± 1.5
60	0.013	5	5	55.1 ± 1.1	53.4 ± 1.1	24.7 ± 1.2	23.2 ± 1.1
80	0.010	4	3	$58.2 \pm 0.8^{\dagger}$	$56.5 \pm 1.0^{\dagger}$	24.5 ± 1.0	23.0 ± 0.9
100 (reported)	–	–	–	56.1	53.6	24.5	22.6

Table 10: Resolution ablation of the tree-based binned baselines: per-event NLL (total)

lower is better

Model	Resolution	ACLED	SAGE	UCDP	SIGACT
XGBoost	25 km, 7 d	**3.37**±0.71	**1.83**±0.50	**3.16**±0.14	**−2.74**±0.50
	50 km, 14 d	4.00±0.75	2.25±0.47	3.51±0.12	−2.11±0.52
Hurdle GBM	25 km, 7 d	**3.66**±0.82	**1.82**±0.51	**3.47**±0.03	**−2.74**±0.50
	50 km, 14 d	4.25±0.83	2.26±0.46	3.80±0.17	−2.12±0.51

Mean $\pm$ SD over folds of the per-event NLL (nats per event); each resolution has its own search over the same space; the better resolution per dataset is in bold. The paper uses the 25 km, 7-day fits throughout.

E.6 Sensitivity analysis on PPC pass rates

The PPC pass rates do not depend on the number of simulated paths. Recomputing every check of the 15-model roster from subsamples of B of the 100 stored paths (2080 checks; exact hypergeometric p-values, 2000 joint resamples per cell for the pooled shares), the median absolute change of a p-value is 0.040, 0.020, 0.013 and 0.010 for $B = 20, 40, 60, 80$, the share of checks that flip their PPC-in-90 verdict is 11, 8, 5 and 4 %, and the pooled PPC-in-90 pass share stays within four points of the reported 56 % (PU) and 54 % (NU) at the convention-free sizes $B = 20$ and 60 (Tab. 11); the PPC-in-40 shares move by at most two points. PPC-in-40 is sensitive to the closed-interval convention because 48 of 2080 checks fall exactly on its boundary; we report the closed interval.

F Model specifications

This appendix collects the notation used throughout (§F.1), the full specifications of the two proposed conflict models, CAM-HP (§F.2) and TMD-HP (§F.3), and the baselines (§F.4): implementation notes for the ported neural STPPs and the specifications of the binned models. Tab. 12 lists the model roster.

Table 12: Model taxonomy. **Impl. (Model / Sampler)** indicates what we contributed: *ours* = implemented from scratch; *ported* = upstream source kept byte-clean in an `_upstream/` directory with a thin wrapper for benchmark integration; *ext.* = upstream model extended (e.g. multivariate marks and a new sampler). Neural STPP and AutoSTPP ship no sampler upstream; ours draw exact thinning samples from the byte-clean upstream intensities.

Family	Model	Reference	Impl.	
			Model	Sampler
Proposed Conflict Models	CAM-HP	this work	ours	ours
	TMD-HP	this work	ours	ours
NSTPP baselines	Neural STPP	Chen et al. (2021)	ported	ours
	DSTPP	Yuan et al. (2023)	ported	ported
	DeepSTPP	Zhou et al. (2022)	ported	ours
	AutoSTPP	Zhou & Yu (2023)	ported	ours
	DKMPP	Zhang et al. (2023)	ext.	ours
Binned baselines	Hurdle-GBM	Vestby et al. (2022)	ours	ours
	XGBoost	Chen & Guestrin (2016)	ours	ours
	Persistence	-	ours	ours
	Climatology	-	ours	ours
	Moving average	-	ours	ours

F.1 Notation

Geo-reference databases. A geo-referenced conflict-event database is a finite sequence $\boldsymbol{\mathcal{E}}_N = \{\boldsymbol{e}_n\}_{n=1}^N$ of **events** $\boldsymbol{e}_n = (\boldsymbol{s}_n, t_n, m_n)$ collected over a spatial region $R \subset \mathbb{R}^2$ and a time interval $[0, W]$, where $\boldsymbol{s}_n \in R$ is a location, $t_n \in (0, W]$ a time, and $m_n \in [K] \stackrel{\text{def}}{=} \{1, \dots, K\}$ a mark (event type).

Point processes. $\boldsymbol{\mathcal{E}}_N$ is a realisation of a marked spatiotemporal point process Φ on $R \times [0, W] \times [K]$, fully characterised by its **conditional intensity** $\lambda_k(\boldsymbol{s}, t \mid \boldsymbol{\mathcal{H}}_t)$, defined so that $\lambda_k(\boldsymbol{s}, t \mid \boldsymbol{\mathcal{H}}_t)\, \mathrm{d}\boldsymbol{s}\, \mathrm{d}t$ is the expected number of mark-k events in the infinitesimal volume $\mathrm{d}\boldsymbol{s} \times \mathrm{d}t$ given the strict history $\boldsymbol{\mathcal{H}}_t \stackrel{\text{def}}{=} \{\boldsymbol{e}_n \in \boldsymbol{\mathcal{E}}_N : t_n < t\}$. We formulate conflict-event forecasting as the problem of estimating the conditional law of the point process restricted to a future forecast window. Formally, given a forecast origin τ and horizon Δ, and given the restricted process $\Phi_{\tau,\Delta} \stackrel{\text{def}}{=} \Phi \cap (R \times (\tau, \tau + \Delta] \times [K])$, we define the quantity our forecasting target

$$\mathcal{L}(\Phi_{\tau,\Delta} \mid \boldsymbol{\mathcal{H}}_\tau)\,,$$

$\mathcal{L}$ jointly determines the expected number of future events in any subset of the spatio-temporal domain, their times, locations, and marks, and all other quantities practitioners may care about such hotspot maps, displacement risk by district etc.

Next-event density. The conditional intensity induces a density over the next event. Given the current history, the density assigns probability to the next event occurring at a particular time, location, and mark. For a candidate next event $(\boldsymbol{s}, t, k)$,

$$p(\boldsymbol{s}, t, k \mid \boldsymbol{\mathcal{H}}_t) = \lambda_k(\boldsymbol{s}, t \mid \boldsymbol{\mathcal{H}}_t) \exp\left(- \int_t^t \sum_{j=1}^K \int_R \lambda_j(\boldsymbol{s}', u \mid \boldsymbol{\mathcal{H}}_t)\, \mathrm{d}\boldsymbol{s}'\, \mathrm{d}u \right).$$

Binned models as point process models. Let $\mathcal{G} = \{R_c\}_{c=1}^C$ be a partition of the spatial region $\mathbb{R}^2$ into C cells, and let $\mathcal{T} = \{(u_{b-1}, u_b]\}_{b=1}^B$ be a partition of the time horizon into intervals of length Δt. We define

$$\lambda_k(\boldsymbol{s}, t \mid \boldsymbol{\mathcal{H}}_{t^-}) = \sum_{c=1}^C \sum_{b=1}^B \frac{f^k\big(\boldsymbol{\mathcal{H}}_{b-1}^{\text{bin}}, \mathbf{z}_{c,b}\big)}{|R_c|\, \Delta t}\, \mathbf{1}\{\boldsymbol{s} \in R_c, t \in (u_{b-1}, u_b]\}, \tag{1}$$

where $f^k(\boldsymbol{\mathcal{H}}^{\text{bin}}_{b-1}, \mathbf{z}_{c,b})$ is the predicted count of mark-k events in $(R_c, (u_{b-1}, u_b])$ given binned history $\boldsymbol{\mathcal{H}}^{\text{bin}}_{b-1}$ (e.g., lagged counts) and bin-level covariates $\mathbf{z}_{c,b}$, and f^k is a nonnegative count estimator mapping binned history and bin-level covariates to an expected event count in $R_c \times (u_{b-1}, u_b]$.

Neural spatio-temporal point processes. Neural spatio-temporal point processes usually follow one of two strategies. Some models parameterise the conditional intensity $\lambda_k(\boldsymbol{s}, t \mid \boldsymbol{\mathcal{H}}_t)$ directly with a neural network. The next-event density is then obtained from the intensity through the survival formula above, which requires integrating the total rate over time and space. Other models parameterise the next-event density $p(\boldsymbol{s}, t, k \mid \boldsymbol{\mathcal{H}}_t)$ directly, avoiding an explicit intensity integral but giving up the intensity-based representation used by classical point-process models.

F.2 CAM-HP: Covariate-Adaptive Mixture Hawkes Process

Here we use the notation introduced in §F.1. For each event j in the history, we construct a feature vector $\boldsymbol{x}_j$ that combines spatiotemporal encodings, mark encodings, and the seventeen geographic context features. We map $\boldsymbol{x}_j$ to a latent state $\boldsymbol{h}_j \in \mathbb{R}^{D_h}$ using a shared encoder MLP. The conditional intensity for mark k is an additive mixture of L separable spatiotemporal components:

$$\lambda_k(\boldsymbol{s}, t \mid \boldsymbol{\mathcal{H}}_{t^-}) = \mu_k + \sum_{j:\, t_j < t} \sum_{\ell=1}^{L} \phi^t_\ell\big(\Delta t_j;\, \eta^j_{k,\ell}, \boldsymbol{\theta}^{t,j}_{k,\ell}\big)\, \phi^s_\ell\big(\Delta \boldsymbol{s}_j;\, \boldsymbol{\theta}^{s,j}_{k,\ell}\big), \tag{2}$$

where $\Delta t_j = t - t_j$, $\Delta \boldsymbol{s}_j = \boldsymbol{s} - \boldsymbol{s}_j$, and the per-event parameters $\eta^j_{k,\ell}, \boldsymbol{\theta}^{t,j}_{k,\ell}, \boldsymbol{\theta}^{s,j}_{k,\ell}$ are obtained by projecting $\boldsymbol{h}_j$ through linear layers with $\mathrm{softplus}$ link functions enforcing positivity.

Hierarchical branching allocation. Subcriticality is enforced by construction: the per-event branching ratio is decomposed into a sigmoid-bounded total budget allocated across marks and scales,

$$\eta_{k,\ell}(\boldsymbol{h}_j) = \eta_{\text{tot}}(\boldsymbol{h}_j) \cdot p_k(\boldsymbol{h}_j) \cdot q_{\ell|k}(\boldsymbol{h}_j), \tag{3}$$

with $\eta_{\text{tot}}(\boldsymbol{h}_j) = \sigma(\boldsymbol{w}_{\text{tot}}^\top \boldsymbol{h}_j + b_{\text{tot}}) \cdot \kappa$ for $\kappa \in (0, 1)$, so that $\sum_{k,\ell} \eta_{k,\ell}(\boldsymbol{h}_j) < \kappa < 1$ for every event j, guaranteeing stability of long-horizon simulations. The implementation uses $\kappa = 0.95$.

Kernels and multi-scale modelling. We primarily use a shifted power-law temporal kernel,

$$\phi^t_\ell(\Delta t) = \alpha_{k,\ell} (\Delta t + c_{k,\ell})^{-p_{k,\ell}}, \tag{4}$$

with $p_{k,\ell} > 1$ enforced via $1 + \mathrm{softplus}(\cdot)$ so that the branching-ratio integral is finite, and $c_{k,\ell}, p_{k,\ell}$ obtained from $\boldsymbol{h}_j$ via affine projections. We anchor each component ℓ at fixed temporal and spatial scales T_ℓ, S_ℓ (e.g. $T_\ell \in \{1, 7, 30\}$ days; $S_\ell \in \{0.2, 1, 2\}$ 100 km) so that mixture components represent distinct spatiotemporal diffusion regimes; the set of anchors and the number L of components are selected via hyperparameter optimisation. We also support exponential temporal kernels $\phi^t_\ell(\Delta t) = \alpha_{k,\ell} \exp(-\beta_{k,\ell} \Delta t)$ for the temporal component. For the spatial component, we use an isotropic Gaussian spatial kernels $\phi^s_\ell(\Delta \boldsymbol{s}) = \mathcal{N}(\Delta \boldsymbol{s}; 0, \sigma^2_{k,\ell} \mathbf{I})$ as alternatives.

F.3 TMD-HP: Transformer Mixture-Density Hawkes Process

TMD-HP is an autoregressive intensity-free model that adapts the covariate-aware temporal Transformer of TransFeat-TPP (Meng et al., 2024) to the spatiotemporal marked-event setting by adding a per-mark spatial mixture-density head with anchored components.

Continuous spatiotemporal representations. For each observed event $(\boldsymbol{s}_n, t_n, m_n)$ we construct an event token

$$\begin{aligned} \boldsymbol{u}_n = \mathrm{LN}\Big(& \boldsymbol{\xi}_t(t_n) + \boldsymbol{\xi}_x(x_n) + \boldsymbol{\xi}_y(y_n) + \boldsymbol{\xi}_{\Delta t}(\Delta t_n) + \boldsymbol{\xi}_{\Delta x}(\Delta x_n) + \boldsymbol{\xi}_{\Delta y}(\Delta y_n) \\ & + \mathbf{e}_{m_n} + \mathbf{W}_Z\, \mathbf{Z}(\boldsymbol{s}_n, t_n) \Big), \end{aligned} \tag{5}$$

where $\boldsymbol{\xi}_\bullet$ are sinusoidal encodings, $\mathbf{e}_{m_n}$ is a learned mark embedding, and $\mathbf{Z}(\boldsymbol{s}, t)$ is the seventeen-channel covariate vector. A causal Transformer encoder produces contextual representations $\boldsymbol{h}_n$ that parameterise the next-event distribution.

Mixture-density heads. Three heads decode $\boldsymbol{h}_i$ into the next-event distribution: a temporal head, a spatial head, and a mark head.

1. Temporal head (log-normal mixture). For each event type k the inter-arrival time Δt is modelled as a mixture of M_t log-normal components: $f_t^k(\Delta t \mid \boldsymbol{\mathcal{H}}_i) = \sum_{m=1}^{M_t} \pi_{k,m}^{(t)}(\boldsymbol{h}_i) \cdot \mathcal{LN}(\Delta t;\, \mu_{k,m}^{(t)}(\boldsymbol{h}_i), \sigma_{k,m}^{(t)}(\boldsymbol{h}_i))$.

2. Spatial head (anchored Laplace mixture). Conditioned on the next-event type k, we model the next location $\boldsymbol{s}_{i+1} \in \mathbb{R}^2$ as a mixture of M_s bivariate Laplace components, each anchored at a learned spatial reference point. One component is anchored at the most recent event location $\boldsymbol{s}_i$ to capture short-range continuity; the remaining $M_s - 1$ are anchored at fixed K-means centroids of the type-k training events to capture longer-range hotspots. Each component predicts a residual shift relative to its anchor:

$$f_s^k(\boldsymbol{s} \mid \boldsymbol{\mathcal{H}}_i) = \sum_{m=1}^{M_s} \pi_{k,m}^{(s)}(\boldsymbol{h}_i)\, \mathrm{Laplace}\Big(\boldsymbol{s};\, \boldsymbol{\mu}_{k,m}^{(a)} + \Delta\boldsymbol{\mu}_{k,m}(\boldsymbol{h}_i),\, \boldsymbol{b}_{k,m}(\boldsymbol{h}_i)\Big), \tag{6}$$

where $\boldsymbol{\mu}_{k,1}^{(a)} = \boldsymbol{s}_i$ (recency anchor), $\{\boldsymbol{\mu}_{k,m}^{(a)}\}_{m=2}^{M_s}$ are K-means centroids of type-k training events frozen during training, and $\boldsymbol{b}_{k,m} > 0$ are diagonal Laplace scale parameters. We use the Laplace family rather than Gaussian because the resulting mixture has heavier tails, improving robustness to the spatially diffuse hotspots characteristic of conflict data.

3. Mark head (residual frequency bias). The event type m is modelled via a residual classifier that corrects a smoothed empirical frequency baseline,

$$\log P(m = k \mid \boldsymbol{\mathcal{H}}_i) \propto \mathbf{w}_k^\top \boldsymbol{h}_i + \log\left(\sum_{j \le i} \mathbb{I}(m_j = k) + \epsilon\right), \tag{7}$$

allowing the Transformer to focus on learning deviations from historical base rates under severe class imbalance.

F.4 Binned models

Five binned models are reported in the main roster (Tab. 12): Hurdle-GBM, single-stage XGBoost, Persistence, Climatology, and Moving average. All operate on spatiotemporal count tensors at a fixed resolution of 25×25 km cells with 7-day bins; the coarser 50×50 km, 14-day alternative is an ablation (§E.5).

Hurdle-GBM. For each (cell, time-bin, event-type) triple, the per-cell rate prediction is decomposed into a presence component and a conditional-count component:

$$\hat{\lambda}_{c,k} = p_{c,k} \cdot \mu_{c,k}, \qquad p_{c,k} = \mathrm{XGB}_{\text{cls}}(\mathbf{x}_c), \quad \mu_{c,k} = \mathrm{XGB}_{\text{reg}}(\mathbf{x}_c \mid p_{c,k} > 0), \tag{8}$$

where $\mathbf{x}_c$ concatenates the seventeen raster covariates at cell c with lag counts over a configurable history horizon. Stage 1 (`XGBClassifier`, `objective="binary:logistic"`) is trained on the indicator $\mathbb{I}[N_{c,k} > 0]$ with full sampling; Stage 2 (`XGBRegressor`, `objective="count:poisson"`) is trained on the non-zero subset $\{N_{c,k} : N_{c,k} > 0\}$ only. The two-stage decomposition follows Vestby et al. (2022) faithfully. Simulation draws independent $\mathrm{Poisson}(\hat{\lambda}_{c,k})$ counts per cell.

XGBoost. A single-stage Poisson regressor: $\hat{\lambda}_{c,k} = \mathrm{XGB}_{\text{reg}}(\mathbf{x}_c)$ with `objective="count:poisson"` fitted on the full count distribution (zeros included). Input features are identical to Hurdle-GBM: seventeen raster covariates concatenated with lag counts over a configurable history horizon.

Persistence. $\hat{\lambda}_{c,k} = N_{c,k-1}$. The next-bin count of each cell equals the count from the same cell in the most recent bin. No parameters; deterministic given the training data.

Climatology. $\hat{\lambda}_{c,k} = \bar{N}_c$, where $\bar{N}_c$ is the mean count of cell c over the training window. No temporal dynamics; purely the long-run intensity field. A model that fails to beat Climatology has not learned anything beyond the per-cell base rate.

Moving average. $\hat{\lambda}_{c,k} = \frac{1}{w}\sum_{i=1}^{w} N_{c,k-i}$ over a configurable window w (default $w = 4$ for 7-day bins, i.e. a 28-day moving average). Bridges Persistence ($w = 1$) and Climatology ($w = \infty$), capturing slow drift of the per-cell intensity without explicit covariates.

G METRIC SPECIFICATIONS

This appendix gives the full definitions of the metrics summarised in §2, in the order of §D: the per-event point-process log-likelihood and its decomposition (§G.1), the energy score on trajectory shape (§G.2), the grid-aggregated count metrics (§G.3), and the posterior predictive checks (§G.4).

G.1 PER-EVENT LOG-LIKELIHOOD

Using the notation introduced in §F.1, the per-event NLL is computed as $\bar{\ell} = N^{-1}\sum_{i=1}^{N}(\ell_i^{(t)} + \ell_i^{(s)} + \ell_i^{(m)})$, where $\ell_i^{(t)} = \log p_t(t_i \mid \boldsymbol{\mathcal{H}}_{t_i^-})$, $\ell_i^{(s)} = \log p_s(\boldsymbol{s}_i \mid t_i, \boldsymbol{\mathcal{H}}_{t_i^-})$, and $\ell_i^{(m)} = \log p_m(m_i \mid t_i, \boldsymbol{s}_i, \boldsymbol{\mathcal{H}}_{t_i^-})$ are the temporal, spatial-conditional, and mark-conditional log-densities at event i. Per-event means are reported in shared benchmark units of $(100\,\mathrm{km}^2, \mathrm{day})$: a model parameterised in $(\mathrm{m}^2, \mathrm{s})$ converts via $\ell^{(t)}_{\text{benchmark}} = \ell^{(t)}_{\text{model}} + \log(86\,400/t_{\text{scale}})$ and $\ell^{(s)}_{\text{benchmark}} = \ell^{(s)}_{\text{model}} + 2\log(10^5/s_{\text{scale}})$. The factor of two on the spatial offset reflects the two-dimensional area conversion.

G.2 ENERGY SCORE

We compute energy scores separately for the temporal and spatial marginals. For each event type k, let Φ_k be the realised sequence projected onto the marginal being evaluated. Thus, Φ_k is the sequence of realised event times for the temporal score, and the sequence of realised event locations for the spatial score. Let $\widetilde{\Phi}_k^{(r)}$ denote the corresponding projected sequence from simulated trajectory $r \in \{1, \dots, R\}$.

For any finite non-empty sequence $\Phi = \{\phi_i\}_{i=1}^{n}$, let

$$\mu_\Phi = \frac{1}{n}\sum_{i=1}^{n} \delta_{\phi_i}$$

be its empirical distribution. The per-type energy score is

$$\widehat{S}_k = \frac{1}{R}\sum_{r=1}^{R} W\left(\mu_{\widetilde{\Phi}_k^{(r)}}, \mu_{\Phi_k}\right) - \frac{1}{2R(R-1)}\sum_{r \neq r'} W\left(\mu_{\widetilde{\Phi}_k^{(r)}}, \mu_{\widetilde{\Phi}_k^{(r')}}\right). \tag{9}$$

Here, W is the distance used for the marginal: exact one-dimensional Wasserstein-1 for time, reported in days, and sliced Wasserstein-1 for space, reported in metres.

The first term measures how close the simulated trajectories are to the realised trajectory. The second term compares simulated trajectories to one another and penalises ensembles that collapse to nearly identical samples. Lower scores are better.

We report one score per marginal by averaging over event types with realised-count weights. If n_k is the number of realised events of type k, then

$$\widehat{S} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j}\widehat{S}_k.$$

Event types with no realised events in the forecast window receive zero weight.

Computational details. The fidelity term is computed exactly by comparing the realised trajectory with each of the R simulated trajectories. For the diversity term, computing all pairwise distances between simulated trajectories can be costly. We therefore estimate the ordered-pair average with a Monte Carlo subsample of 500 ordered pairs drawn uniformly without replacement from the $R(R-1)$ available pairs. This gives an unbiased estimate of the diversity term while keeping computation tractable at $R = 100$. If fewer than 500 ordered pairs are available, we use all pairs.

Joint spatio-temporal variant. The two marginal scores cannot tell whether events are placed at the right locations at the right times. Fig. 2 therefore reports a joint variant in which each event is mapped to the unit cube,

$$u = \left(\tfrac{t-t_0}{t_1-t_0},\ \tfrac{x-x_{\min}}{x_{\max}-x_{\min}},\ \tfrac{y-y_{\min}}{y_{\max}-y_{\min}} \right),$$

with $[t_0, t_1]$ the forecast window and $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$ the fixed study region of the dataset, and W in Eq. (9) is the sliced Wasserstein-1 distance between the empirical distributions of these triples, averaged over $P = 100$ fixed unit directions on the sphere (a Fibonacci lattice). All event types are pooled, so the joint score carries no type weighting; the same $R = 100$ trajectories and 500 sampled pairs are used. Because the sliced Wasserstein-1 distance is a metric of negative type, the resulting energy score remains a proper scoring rule. We report it as a skill against climatology, $1 - \widehat{S}_{\text{model}}/\widehat{S}_{\text{climatology}}$, computed per cell and regime, so that 0 is the climatology forecast and 1 a perfect one. Projecting onto the time axis alone recovers the temporal score up to the window rescaling, and onto the two spatial axes the spatial score up to the region rescaling.

G.3 Grid-aggregated counts

We evaluate aggregate event counts on grids of square spatial cells. Let $\mathcal{G} = \{R_c\}_{c=1}^{C}$ be the spatial grid, and let $\mathcal{T} = \{I_b\}_{b=1}^{B}$ be a partition of the forecast window into time bins. For each spatial cell R_c and time bin I_b, let $N_{c,b}$ denote the realised number of events in that cell–bin, and let $\widetilde{N}_{c,b}^{(r)}$ denote the corresponding count in the r-th of the R simulated trajectories.

Count-CRPS (primary grid metric). The grid covers the fixed study region of the dataset with 25 km cells (50 km as a robustness check) and 7-day bins from the start of the forecast window; simulated events outside the region are collected in one additional cell per bin, so that they count as errors. In every cell–bin the R trajectories form an ensemble forecast of the count, which we score with the continuous ranked probability score in its unbiased ensemble form (Gneiting & Raftery, 2007),

$$\mathrm{CRPS}_{c,b} = \frac{1}{R}\sum_{r=1}^{R}\left|\widetilde{N}_{c,b}^{(r)} - N_{c,b}\right| - \frac{1}{2R(R-1)}\sum_{r\neq s}\left|\widetilde{N}_{c,b}^{(r)} - \widetilde{N}_{c,b}^{(s)}\right|, \tag{10}$$

and report the total relative to the number of observed events, $\sum_{c,b}\mathrm{CRPS}_{c,b}/\sum_{c,b}N_{c,b}$. The first term is the mean absolute count error of a single trajectory; the second credits the ensemble for its own spread, and for a deterministic forecast the score reduces to the absolute error. Summed over cell–bins, it is the energy score of Eq. (9) applied to the vector of cell–bin counts with the L_1 distance, so it is a proper scoring rule; unlike the trajectory energy scores, which compare normalised empirical distributions of event times and locations, it penalises missing and surplus events. A forecast of no events scores exactly 1, because the CRPS of a point mass at zero equals the observed count, and a perfect forecast scores 0.

Relative count errors (secondary). The MAE^{+} pair below is computed with a single time bin ($B = 1$, the whole forecast window) on a 25 km or 50 km grid spanning the bounding box of the observed events of the window, separately for each event type and averaged over types; simulated events outside that box are not counted. A smoothing constant of 1 in the denominator handles cells with zero events. We use them for being interpretable, in support to the count-CRPS.

- $\mathrm{MAE}^{+}_{\mathrm{Real}}$. This score normalises the error by the realised count, so it is dominated by simulated events in cells where nothing was observed, whereas a missed event costs at most one per cell:

$$\mathrm{MAE}^{+}_{\mathrm{Real}} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{CB}\sum_{c=1}^{C}\sum_{b=1}^{B}\frac{|N_{c,b} - \widetilde{N}_{c,b}^{(r)}|}{N_{c,b}+1}. \tag{11}$$

- $\mathrm{MAE}^{+}_{\mathrm{Pred}}$. This score normalises the error by the simulated count, so it is dominated by observed events in cells where the trajectory placed none, whereas a surplus event costs at most one per cell:

$$\mathrm{MAE}^{+}_{\mathrm{Pred}} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{CB}\sum_{c=1}^{C}\sum_{b=1}^{B}\frac{|N_{c,b} - \widetilde{N}_{c,b}^{(r)}|}{\widetilde{N}_{c,b}^{(r)}+1}. \tag{12}$$

Table 13: The four datasets as used in the benchmark: curator, period and event count after preprocessing, folds, event types (K), access and what the benchmark redistributes. SAGE is closed access and included for comparison only.

Dataset	Curator	Period used	Events	Folds	Event types (K)	Access; redistributed
ACLED	Armed Conflict Location & Event Data project	2010-01 to 2024-08	23 158	5	Battles; explosions / remote violence; violence against civilians; protests; riots (5)	registration, ACLED EULA; trained weights only
UCDP-GED	Uppsala Conflict Data Program	2010-01 to 2024-12	4 906	2	State-based; non-state; one-sided violence (3)	open, CC BY 4.0; folds and weights
SAGE	UN peacekeeping missions MINUSCA, MONUSCO, UNMISS	2020-01 to 2024-09	19 785	3	Battles; violent unrest; other violence (3)	closed, per-user research agreement; nothing derived
SIGACT-Iraq	US military reports (Iraq War Logs deposit)	2004-01 to 2009-01	102 168	4	Enemy-action reports, one class (1)	open, CC0; folds and weights

G.4 Posterior predictive checks

To assess further validity of our models, we perform posterior predictive checks (PPCs). We compare the realised value of specific test statistics $T(\Phi)$ against the distribution of those statistics generated by the model's simulated trajectories $\{T(\widetilde{\Phi}^{(r)})\}_{r=1}^{R}$. We use five interpretable statistics that capture complementary aspects of conflict dynamics: volume, burstiness, spatial extent, spatial clustering, and spatiotemporal dependence.

1. **Total event count (N).** Measures the global volume of conflict over a forecast window: $T_{\text{count}}(\Phi) = |\Phi(W_0)|$.
2. **Maximum daily intensity ($N_{\text{max-day}}$).** Measures temporal burstiness and a model's ability to generate extreme peaks in activity: $T_{\text{max-day}}(\Phi) = \max_{d \in W_0} \text{Count}(\Phi, d)$.
3. **Max-spanned distance (D_{max}).** Measures the geographic extent of the conflict: $T_{\text{max-dist}}(\Phi) = \max_{e_i, e_j \in \Phi} \|\boldsymbol{s}(e_i) - \boldsymbol{s}(e_j)\|_2$.
4. **Spatial connected components (N_{cc}).** Measures spatial clustering versus scattering. We construct a geometric graph in which an edge exists between events at distance ≤ 50 km (default theatre-diversity threshold). The statistic $T_{\text{cc}}(\Phi)$ is the number of connected components.
5. **Spatiotemporal correlation coefficient (ρ_{st}).** The Pearson correlation between Δ-space and Δ-time over event pairs, capturing whether the model reproduces the observed coupling between space and time dynamics.

For each statistic we report the empirical p-value of the realised value under the simulated distribution; values near 0 or 1 flag specific failure modes (collapsed volume, missed bursts, geographic over- or under-extension, scattering vs. contiguity, decoupled space–time). The empirical PPC results—signed miscalibration map, per-cell pass rates, and pass-rate barchart—are reported in §D alongside the other simulation-based extended results.

H Extended data statistics

Tab. 13 summarises the four datasets as the benchmark uses them: ACLED, UCDP-GED and SAGE cover South Sudan, the Central African Republic and DR Congo, SIGACT-Iraq covers Iraq. Each processed dataset records the SHA-256 of its raw input and its preprocessing date (snapshot 2026-04); per-source porting notes are on the website's data-porting pages. Marks are integer-encoded $0, \ldots, K-1$ in the order listed.

Tab. 14 reports per-fold event counts for each dataset, broken out by the train, validation, and test windows of the rolling-quantile fold spec (728/28/42 days; §3.3). The number of folds per dataset is set adaptively by the rolling-quantile placement: ACLED admits five folds (the configured maximum),

Table 14: Per-fold event counts per dataset for the rolling-quantile fold spec used throughout the benchmark (728-day train, 28-day validation, 42-day test). Counts are events that survive precision filtering and projection-clip (§3.3).

Dataset	Fold	Train	Val	Test
ACLED	0	2,312	84	167
ACLED	1	2,740	125	132
ACLED	2	3,954	222	326
ACLED	3	6,029	376	411
ACLED	4	5,333	167	321
SAGE	0	8,776	345	418
SAGE	1	7,549	293	480
SAGE	2	7,724	418	572
SIGACT-Iraq	0	32,122	1,150	2,093
SIGACT-Iraq	1	37,082	3,144	4,485
SIGACT-Iraq	2	48,184	3,130	5,250
SIGACT-Iraq	3	61,096	914	2,090
UCDP-GED	0	891	47	75
UCDP-GED	1	1,368	76	85

SIGACT-Iraq four, SAGE three, and UCDP-GED two. Fewer-than-configured folds are an expected outcome of quantile placement on sparse data, not a truncation artefact.

I COVARIATE CATALOG AND YEAR ALIGNMENT

I.1 BUILDING THE MASTER GRID

The master grid is the substrate every covariate-aware model shares. It is built once per region, projection, resolution and catalog by `conflict-grid build --card <grid card>` from a grid card that lists the ISO-3 countries of the region (their ADM0 union, buffered by 20 km, is the region of interest), the target projected CRS (UTM 35N for Central Africa), the cell size (1 km) and the catalog identifier. The grid identifier is a content hash of these fields, so rebuilding from the same card is idempotent and any catalog edit that changes a channel yields a new grid. The production cards ship with the repository; the Build the grid page of the website walks through the command and its outputs.

I.2 CATALOG OF CHANNELS

Table 15: Covariate catalog used in CONFLICTCAST. Eleven features are instantiated on the user-specified spatial grid, yielding seventeen channels per year: four static features, three dynamic features emitted with $\{10\,\text{km}, 50\,\text{km}\}$ Gaussian aggregates ($3 \times 3 = 9$ channels), and four dynamic signed-distance-field features without aggregates. The *per-fold transform* column records the post-projection numeric transform fitted at fold-construction time on the training split only and applied to validation and test splits at inference time.

Feature	Source	Native res.	Year range	Per-fold transform
Static (terrain + infrastructure)				
slope	SRTM-derived	∼100 m	–	log1p+standardize
elevation	SRTM-derived	∼100 m	–	standardize
dist_road	worldpoppy (g1)	∼100 m	–	log1p+standardize
dist_water	worldpoppy (g1)	∼100 m	–	log1p+standardize
Dynamic with {10, 50} km Gaussian aggregates				
population	WorldPop	100 m	2000–2020	quantile-normal
nightlights	VIIRS-DNB (filtered)	∼500 m	2015–2023	quantile-normal
built	ESA built-surface	∼1 km	2015–2030	quantile-normal
Dynamic, signed-distance fields (no aggregates)				
dist_park	IUCN Cat-I	∼1 km	2015–2022	signed-log1p+standardize
dist_urban	ESA built-surface	∼1 km	2015–2030	signed-log1p+standardize
dist_cult	ESA WorldCover (cult.)	∼1 km	2015–2022	signed-log1p+standardize
dist_woody	ESA WorldCover (woody)	∼1 km	2015–2022	signed-log1p+standardize

I.3 AVAILABLE RASTER PRODUCTS

The eleven products in Tab. 15 are a curated subset of the 55 raster products (administrative areas, population counts, covariates and grid-cell surface areas) that the bundled `worldpoppy` client fetches on demand; the full manifest ships with the client and is listed on the Raster features page. Adding a product to the benchmark is a single catalog entry.

I.4 YEAR ALIGNMENT

At feature-sampling time we follow the rule: each event in year Y is enriched with the year-$(Y-1)$ raster of every dynamic feature, and if year $Y-1$ lies outside the feature's coverage window the event is sampled from the nearest available year. The $Y-1$ shift avoids look-ahead leakage. SIGACT-Iraq uses only the two terrain channels, which predate its events (§3.2).

J INGEST PREPROCESSING AND FOURIER ENCODINGS

This appendix lists the source-specific ingest preprocessing applied at stage (1) of §3.3 and the deterministic Fourier-and-calendar encodings appended at stage (3).

J.1 SOURCE-SPECIFIC INGEST RULES

Each supported source has its own ingest module under `conflict_data.adapters`. Adding a new dataset adds one row to Tab. 16; everything downstream of the ingest is dataset-agnostic.

The temporal jitter applies only to sources coded at daily resolution (ACLED, UCDP-GED, SAGE), where the events of one day share the timestamp 00:00:00; a uniform shift into the first hour breaks the ties without crossing day boundaries, and SIGACT-Iraq, which carries true sub-daily stamps, is left untouched. The spatial jitter applies only to sources that geocode to administrative centroids (ACLED, UCDP-GED) and would otherwise create spurious clusters there; SIGACT-Iraq and SAGE carry sub-kilometre precision. Both jitters are seeded from the global seed.

J.2 FOURIER-AND-CALENDAR ENCODINGS

Tab. 17 lists every column appended at stage (3).

Table 16: Source-specific ingest preprocessing applied at stage (1) of §3.3. Rules are deterministic given the global seed (default 42).

Source	Raw format	Time jitter	Space jitter	Filter
ACLED	CSV (;)	U[0, 3600] s	± 50 m	time and location precision $= 1$
UCDP-GED	CSV (,)	U[0, 3600] s	± 50 m	date and location precision $= 1$; known violence type
SIGACT-Iraq	tabular log	none	none	none
SAGE	encrypted `.crypt`	U[0, 3600] s	none	categories Battles, Violent unrest, Other violence

Table 17: Spatio-temporal encodings appended at stage (3) of §3.3. Default values shown; the dataset card may override every knob via `TemporalConfig` and `SpatialConfig`.

Channel group	**Source**	**Cols**	**Default**
Temporal (`t_*`, 15 columns total)			
day-of-week (sin, cos)	event time	2	on
month-of-year (sin, cos)	event time	2	on
week-of-month (one-hot)	event time	5	on
annual Fourier ($k = 1, 2, 3$)	event time	6	on; $K = 3$ harmonics
hour-of-day (sin, cos)	event time	2	off; *on for SIGACT-Iraq*
linear trend	event time	1	off (requires `start_date`, `end_date`)
Spatial (`s_*`, 50 columns total)			
z-score (x, y) within ROI	projected coords	2	on
2-D Fourier on unit box	projected coords	48	on; $K_{\max} = 4$, mixed terms enabled
raw (x, y)	projected coords	2	off

The 2-D Fourier basis comprises 8 pure-x components ($k_x \in \{1, 2, 3, 4\}$, sin and cos), 8 pure-y components ($k_y \in \{1, 2, 3, 4\}$, sin and cos), and 32 mixed components ($(k_x, k_y) \in \{1, \dots, 4\}^2$, sin and cos), totalling 48 channels. Coordinates are normalised to the unit box of the dataset's projected ROI before evaluation, so the same basis transfers across datasets without retuning $K_{\max}$.

K Feature transformations and fold construction

K.1 Per-fold feature transformations

Every `g_*` (raw geospatial) channel in the catalog (Tab. 15) carries a *per-fold transform* declaration. The transform is a sklearn pipeline fit on the training events of each fold; the same fitted state is then applied (transform-only) to the fold's validation events, test events, and the per-fold simulation grid. The four transform families used in production are:

`standardize` Subtract the train-set mean, divide by the train-set standard deviation. Used for near-Gaussian channels (`srtm_elevation_g1`).

`log1p_standardize` Apply $\log(1 + x)$, then standardize. Used for non-negative right-skewed channels (slope, distance to roads / waterways).

`signed_log1p_standardize` Apply $\mathrm{sign}(x) \cdot \log(1 + |x|)$, then standardize. Used for signed-distance fields where negative values indicate "inside the feature" (`dist_park`, `dist_urban`, `dist_cult`, `dist_woody`). Monotone, sign-preserving, and cannot extrapolate outside $\pm \log(1 + |\max|)$.

`quantile_normal` Map the train-set marginal to a standard normal via the empirical quantile transform. Used for zero-heavy channels (population, nightlights, built surface) where the raw marginal is far from Gaussian.

Per-channel assignments are listed in the *per-fold transform* column of Tab. 15. The fitted state is persisted with the fold, and the per-fold simulation grid stores the same transformed channels on every cell of the master grid, so simulators see covariates with the distribution seen at training time.

K.2 FOLD-SPEC CARD SCHEMA

The fold-spec card (`quantroll_728_28_42`) declares the window lengths (728, 28 and 42 days), the five forecast origins at the 0.15, 0.35, 0.55, 0.75 and 0.95 quantiles of event time, and minimum-event guards of 50 training and 40 test events; the guards are what reduce the fold count on sparser sources (UCDP-GED 2, SAGE 3, SIGACT-Iraq 4, ACLED 5). The Folds page documents the card schema.

L DATASHEET FOR CONFLICTCAST

We follow the structure of Gebru et al. (2021), organised into seven sections.

Motivation. CONFLICTCAST standardises conflict-event forecasting, a task that operational deployments run under ad-hoc protocols and on which binned regression, classical Hawkes and neural point-process models had not been compared under one evaluation. It targets researchers and practitioners who need cross-family rankings under several operational regimes and a reproducible path from raw event records to fold-level artefacts.

Composition. Four event datasets (ACLED, UCDP-GED, SAGE, SIGACT-Iraq; Tab. 13) covering 2004–2024 in two regions (Central Africa: DR Congo, the Central African Republic and South Sudan; Iraq). Each event carries a time, a projected location, a mark (event type) and a country code; the Central African datasets carry seventeen covariate channels per event (four static terrain and infrastructure rasters and thirteen dynamic socio-economic rasters aligned to the previous year), SIGACT-Iraq the two terrain channels that predate its events (§3.2). Each dataset is cut into up to five rolling-quantile folds with 728-day training, 28-day validation and 42-day test windows (Tab. 14).

Collection mechanism. Each dataset is collected by its curator under a documented mechanism: ACLED by media monitoring with local-partner verification, UCDP-GED by newswire coding with a one-fatality threshold, SAGE by UN peacekeeping operational reporting, SIGACT by US military field reports declassified in the Iraq War Logs (2004–2008). The mechanisms induce distinct, known reporting biases, which the benchmark exposes rather than reconciles (§3).

Preprocessing. `conflict-data preprocess` applies the source-specific precision filters and jitters of Tab. 16, deduplicates, projects to the regional metric CRS and clips to the boundary union; covariates are sampled at the event location under the year-minus-one rule (§I.4).

Uses. Intended: benchmarking probabilistic event forecasters, comparing continuous-time and binned model families, and ablating components of a forecasting pipeline. Not intended: substituting for on-the-ground analysis, targeting decisions of any kind, horizons beyond the 42-day window for which no calibration is provided, or single-event predictions that ignore the predictive-distribution framing.

Distribution. Released at an anonymised URL for review and deposited at a permanent archive on acceptance. Derived per-fold artefacts are redistributed where the source licence allows: UCDP-GED and SIGACT-Iraq folds ship, ACLED folds do not (trained weights do), and nothing derived from SAGE ships beyond the pipeline scripts and fold manifests (Tab. 13). Per-package licences are in each `NOTICE` file.

Maintenance. The repository accepts datasets, models and metric families through pull requests against the layered architecture of §M; only the application layer is domain-specific, so a new event-stream domain re-implements that layer alone. The deposited artefacts are maintained for the duration of the proceedings policy, and external results enter the leaderboard through the submission protocol of §M.

Layer 1: Infrastructure (domain-agnostic) `stpp-core` (types, manifests, registries) · `stpp-grid` (raster grids) · `stpp-hpo` (ASHA) · `stpp-hpo-fit` (grid/random search) · `stpp-eval` (metrics + simulation dispatch)
Layer 2: Application (conflict-specific) `conflict-grid` (covariate rasters via the anonymised raster-fetching layer of §3.3, plus VIIRS, SRTM, OSM) · `conflict-data` (source adapters, ETL, fold generation)
Layer 3: Models (one package per model) CAM-HP · TMD-HP · DKMPP · DeepSTPP · DSTPP · AutoSTPP · Neural STPP · Hurdle-GBM (binned)

Table 18: Three-layer repository architecture. Infrastructure defines contracts (`TrainResult`, `SimFn`); the application layer implements the `FeatureCatalog` protocol for conflict rasters; model packages plug in via auto-import registration.

M REPOSITORY ARCHITECTURE AND REPRODUCIBILITY PIPELINE

CONFLICTCAST is a modular multi-package Python repository in three layers (Tab. 18): a new domain replaces only the application layer, a new model implements two functions, a new metric extends only the evaluation package.

Infrastructure layer. Five domain-agnostic packages define the core abstractions. `stpp-core` provides Pydantic-validated manifest schemas (grid, dataset, fold), the `TrainResult` contract, and auto-import registries for HPO and simulation. `stpp-grid` implements generic raster grid construction and covariate fetching via a `FeatureCatalog` protocol. `stpp-hpo` orchestrates ASHA hyperparameter search with a controller/runner architecture backed by SQLite state, plus a `retrain` step that trains the best configuration on train+validation data. `stpp-hpo-fit` provides grid and random search for fast-fitting models (the binned baseline). `stpp-eval` implements the evaluation suite (NLL, PPC, energy score, grid metrics) and dispatches simulation to model-specific samplers via the `SimFn` protocol.

Application layer. Two packages contain all conflict-domain-specific logic. `conflict-grid` implements the `FeatureCatalog` protocol for conflict-relevant rasters and provides a CLI for building covariate grids. `conflict-data` implements source adapters for the four event datasets, applies standardised preprocessing, and generates temporal folds with train-only transformer fitting. Replacing these two packages—and writing a new `FeatureCatalog`—adapts the entire benchmark to a different domain (e.g., seismology, epidemiology) without modifying any infrastructure or model code.

Adding a new model. A model package must implement two functions and register them at import time:

1. A `train_fn(hparams, budget, start_epoch, ...)` → `TrainResult` that trains for `budget` epochs and returns validation loss plus model state. The ASHA controller calls this function at each rung boundary; checkpoint-resume across rungs is handled via `start_epoch` and a checkpoint path.
2. A `sim_fn(checkpoint, config, fold, modes, n_paths, ...)` → `dict[str, DataFrame]` that loads a trained checkpoint and generates simulated trajectories under each requested mode (FC, PU, NU) in the canonical `(path_id, time, x, y, mark)` schema.

Both functions are discovered via an auto-import registry: importing `<model>.hpo` or `<model>.sim` triggers registration. No changes to infrastructure packages are required.

Reproducibility and submissions. Every artefact carries a versioned manifest that records its inputs, software versions and creation time, and the full pipeline is seven commands:

`conflict-grid build` → `conflict-data preprocess` → `conflict-data make-folds` → `stpp-hpo search` + `finalize` + `retrain` → `stpp-sim run` → `stpp-eval score`.

The website offers three reproduction tiers with their cost: regenerating every table and figure from the released per-fold scores (minutes on a CPU), re-running simulation and scoring from the released

checkpoints and processed folds (about 100 GPU-hours for the full campaign, minutes for one cell), and the full pipeline from raw data (§O); see Reproduce our results. External models join the leaderboard through the same protocol (frozen folds, benchmark units, declared regimes, no tuning on the test window) by submitting the per-fold likelihood and simulation artefacts described on the submission page; a model without a sampler is listed as likelihood-only.

N Training protocol and search spaces

Per-fold data flow. A fold is three non-overlapping windows of 728, 28 and 42 days (training, validation, test; §3). The data processor is fit on the training window only (temporal anchor, coordinate centring, unit scales of one day and 100 km, category map, region bounding box) and applied unchanged to the other two windows.

Search, retrain, evaluate. For each model and fold, ASHA (Li et al., 2018) draws configurations from the model's grid (Tab. 19), trains each on the training window and records the validation loss per epoch; the rung schedule is $[5, 50, 150]$ epochs at reduction factor $\eta = 10$ ($[5, 15, 30]$ and $\eta = 3$ for Neural STPP), promotion compares the running minimum of the validation loss, and the winner is the trial with the lowest validation loss at any rung. Its configuration is retrained from scratch on the training and validation windows for $b^\star + 1$ epochs, $b^\star$ being the winner's best validation epoch, without snapshot selection; the final state is evaluated once on the test window, for the likelihood decomposition and for the simulations. The test window is never read during search or retraining. The tree models are fitted by random search over 100 configurations per fold and resolution profile, selected on the validation likelihood of a teacher-forced rollout; the naive baselines have no tuned parameter except the moving-average window. The configuration files ship with the repository; the Train pages describe the controller.

O Compute statement

Hardware. All neural models were trained on a shared academic cluster on single-GPU partitions, NVIDIA RTX 4090 (24 GB) for the bulk of the campaign and RTX Pro 6000 (96 GB) for the largest runs (DKMPP at 25 km, AutoSTPP and Neural STPP on SIGACT-Iraq); each ASHA sweep ran one controller and three runner processes on one GPU (AutoSTPP: three GPUs per search job). The binned family, evaluation aggregation and scoring ran on CPU nodes.

Measured cost. **??** gives the GPU-hours of the main campaign from the cluster accounting of every job of the 209 roster cells: 2 077 GPU-hours in total, of which 1 440 in jobs that completed; hyper-parameter search accounts for 1 448 and simulation for 574. AutoSTPP alone consumed 627 GPU-hours, the two conflict-tailored models 373 together. The ablations of §E add 582 GPU-hours for the forecast-horizon study (AutoSTPP 371, Neural STPP 117) and 254 for the leak-free SIGACT re-run.

Expected cost to reproduce. The three reproduction tiers of §M cost minutes on a CPU (tables and figures from the released scores), about 100 GPU-hours (simulation and scoring from the released checkpoints) and about 1 400 GPU-hours for one clean pass of the full pipeline on equivalent hardware; a first sanity check on UCDP-GED with its two folds needs of the order of 50 GPU-hours.

Table 19: Hyper-parameter search space of every tuned model. Learning rates are log-spaced between the listed end points; "grid" is the number of distinct configurations and "trials" the number evaluated per fold (drawn with replacement, master seed 42). ASHA rungs are $[5, 50, 150]$ epochs at $\eta = 10$ for every model except Neural STPP ($[5, 15, 30]$, $\eta = 3$). Layer widths are written $n \times w$ (n layers of width w).

Model	Tuned hyper-parameters (values)	Grid	Trials
CAM-HP	learning rate $\{10^{-4}, \dots, 10^{-2}\}$ (5); weight decay $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$; batch $\{16, 32, 64, 128\}$; encoder {linear, MLP}; kernel mixture: nine scenarios of 1–7 components with temporal scales in $\{1, 3, 7, 14, 30\}$ days and spatial scales in $\{0.2, 1, 2\} \times 100$ km	1 440	80
Vanilla Hawkes	as CAM-HP without the kernel mixture (single kernel)	160	80
TMD-HP	learning rate (5); weight decay $\{0, 10^{-4}, 10^{-3}\}$; batch $\{16, 32, 64\}$; model width $\{32, 64, 128\}$; heads $\{2, 4\}$; layers $\{1, 2, 3\}$; dropout $\{0, 0.1, 0.2\}$; spatial mixtures $\{3, 5, 8\}$; temporal mixtures $\{1, 3, 5\}$	21 870	300
DKMPP (50 km)	learning rate $\{10^{-4}, \dots, 10^{-2.5}\}$ (4); weight decay (4); batch $\{16, 32, 64\}$; latent dimension $\{32, 64, 128\}$; kernel MLP $\{2\times64, 2\times128, 3\times64\}$; base-kernel scale $\{5, 10, 20, 50, 100\}$; mixture MLP $\{2\times64, 2\times128, 3\times64\}$; positional basis $k_{\max} \in \{0, 4, 8\}$	19 440	300
DeepSTPP	learning rate (5); weight decay $\{0, 10^{-4}, 10^{-3}\}$; batch $\{16, 32, 64\}$; embedding $\{32, 64, 128\}$; hidden $\{32, 64, 128\}$; inducing points $\{10, 20, 40\}$; KL weight $\{0.1, 0.5, 1\}$; KL schedule {fixed, linear warm-up} with warm-up $\{5, 10, 20\}$ epochs; dropout $\{0, 0.1, 0.2\}$	65 610	300
DSTPP	learning rate (5); weight decay (3); batch (3); model width $\{32, 64, 128\}$; heads $\{2, 4\}$; layers $\{1, 2, 3\}$; dropout (3); RNN width $\{128, 256\}$	4 860	300
AutoSTPP	learning rate $\{5\times10^{-4}, 10^{-3}, 4\times10^{-3}\}$; weight decay $\{0, 10^{-4}\}$; batch $\{64, 128\}$; ProdNets $\{2, 4, 10\}$; width $\{128, 256\}$; layers $\{2, 3\}$; tanh; history truncated to 20 events	144	300
Neural STPP	learning rate $\{10^{-4}, 5\times10^{-4}, 10^{-3}\}$; weight decay $\{0, 10^{-4}\}$; CNF hidden layers $\{2\times64, 3\times64, 2\times128\}$; TPP hidden layers $\{(8, 20), (16, 32)\}$	36	20
XGBoost, Hurdle-GBM	lags $\{3, 7, 14\}$; depth $\{3, 4, 6, 8\}$; learning rate log-uniform $[0.01, 0.2]$; trees $\{500, 1000, 1500, 2000\}$; minimum child weight uniform $[0, 5]$ (Hurdle-GBM: drawn per stage)	random	100
Moving average	window $\{4, 8, 12, 24\}$ bins	4	4