A Parsimonious Monte Carlo Model for Verifying Ambulance System Dynamics and Time-Dependent Blocking from Dispatch Records

Max Pernklau
University of Hagen
max.pernklau@fernuni-hagen.de

Sabine Folz-Weinstein
University of Hagen
sabine.folz-weinstein@fernuni-hagen.de

Christian Beecks
University of Hagen
christian.beecks@fernuni-hagen.de

ABSTRACT

Emergency medical service (EMS) systems operate under finite vehicle availability and stochastic, time-varying demand. When all ambulances are occupied, additional calls experience blocking. Dispatch records typically document served missions but might not capture unserved incidents.
We present a parsimonious Monte Carlo (MC) simulation to estimate time-dependent blocking risk from dispatch data alone. The model fits two components: A censoring-corrected incident rate and a time-dependent duration distribution. Duration parameters correlate strongly with the lagged incident rate, which we exploit by collapsing per-bin duration fits into a compact linear model. Together with vehicle duty schedules estimated from the same source, these form a complete simulation.
Unlike piecewise Erlang B formulas, it captures transient effects such as missions spanning shift and hour boundaries. Validated against one year of data from a mid-sized German city, the simulation reproduces the observed distribution of concurrent active missions – a quantity not used directly in calibration.

Keywords

emergency medical services, Monte Carlo simulation, queueing theory, resource planning

INTRODUCTION

Emergency medical services (EMS) operate under a fundamental constraint: The supply of ambulances is finite and governed by staffing schedules, while emergency demand is stochastic and time-varying. When all available vehicles are simultaneously committed to active missions, incoming emergencies experience *blocking*: No vehicle can be dispatched immediately. Consequences range from delayed response to improvised alternatives such as dispatching a fire engine or requesting mutual aid from neighboring jurisdictions. Quantifying how often and when blocking occurs is a prerequisite for informed capacity planning.

In some EMS systems, including the municipal ambulance service studied here, dispatch software records the missions that *are* served, but does not generate a record when all resources are exhausted and a call cannot be assigned to a vehicle. The fraction of time spent at full capacity can be extracted directly from the dispatch data, and rudimentary counting methods can aggregate blocked cases over a year. However, these observations alone do not reveal *when* individual incidents are missed, *when* during the day blocking risk concentrates, or *how* a change in fleet size or shift design would alter the risk profile. Analytical tools such as the Erlang B formula can answer these questions in principle, but rely on stationarity assumptions whose validity is difficult to establish for a specific system. Therefore, our research question is: Given estimated incident rates and durations from dispatch records,

does a simple Monte Carlo simulation reproduce the observed system dynamics – and if so, what time-dependent blocking rate does it imply?

In this paper, we introduce a parsimonious Monte Carlo (MC) simulation framework for estimating time-dependent blocking probability using only routinely available dispatch data. This minimal-input design is deliberate: Some mid-sized EMS services maintain only these minimal records, with no data that would support richer models. The framework relies on three statistical inputs extracted from a single data source (the dispatch log): a time-varying incident rate, a time-varying distribution of mission durations, and the duty schedule of each vehicle. It imposes no spatial structure and does not incorporate dispatch priority rules. This architecture keeps the model transparent and easily implementable while remaining empirically grounded in standard dispatch records such as those available from the German *Funkmeldesystem*. The aim is not to replicate operational detail, but to capture aggregate system dynamics (in particular, the frequency and timing of blocked missions) with the smallest set of structural assumptions. The model produces three outputs: The expected number of missed missions over time (the primary quantity of interest, not directly observable in dispatch records), the distribution of concurrent active missions by time of day (the validation target), and scenario comparisons under modified fleet configurations.

Our contribution is twofold. First, we show that a simple MC simulation, calibrated from dispatch records alone, can reproduce the empirically observed distribution of concurrent active missions – a quantity not used directly in calibration – and thereby produce time-resolved blocking estimates. The censoring correction we apply to recover the true incident rate during preprocessing is itself based on the Erlang loss principle; the novelty lies not in the correction but in the ability to *verify* whether the resulting model faithfully reproduces the observed system dynamics. Second, we propose an empirically motivated parameterization of the mission duration model via the incident rate. Per-bin duration parameters correlate strongly with the lagged incident rate (Pearson r between 0.6 and 0.9), an empirical regularity consistent with known mechanisms such as traffic congestion and hospital crowding during peak demand. This reduces the duration model from 48 parameters to nine while preserving the time-of-day structure.

The model is validated against one year of operational data from a mid-sized city in western Germany (250,000 inhabitants), operating a fleet of 9 ambulances (*Rettungswagen*, RTW). By fitting smooth parametric distributions to aggregate data rather than relying on individual event records, the model tolerates the noisy timestamps, missing status codes, and minor data entry errors typical of operational dispatch logs. Beyond blocking estimation, the simulation serves as a lightweight tool for verifying analytical results from EMS planning reports and for rapid scenario exploration.

RELATED WORK

We review the most relevant literature, organized by its relationship to the problem addressed here. For a comprehensive survey of EMS planning models, including Erlang-type formulas and their extensions, see Ingolfsson (2013).

Analytical and Simulation-Based Approaches

The most direct antecedent to our work is Rastpour et al. (2020), who model 'red alert' durations (periods when all ambulances in a system are busy) using Erlang loss models with state-dependent service rates. Their research question is closest to ours, but differs in two respects: They use analytical queueing models rather than simulation, and their dispatch systems explicitly track when all ambulances are busy, so alert onset and duration are directly observable. In our setting, the *Funkmeldesystem* records only served missions, and blocking events are invisible in the data. Restrepo et al. (2009) apply the Erlang loss formula to static ambulance deployments across bases, using blocking probability as an objective for fleet allocation. Whitt and Zhao (2017) develop approximations for blocking in loss models with non-Poisson time-varying arrivals, but target large-scale systems and assume structural properties (e.g., Gaussian limits) that need not hold for small municipal fleets such as ours.

Simulation-based approaches avoid these distributional assumptions at the cost of requiring explicit specification of system dynamics. Yang et al. (2019) combine discrete-event simulation with Gaussian mixture model clustering of spatial demand for ambulance allocation optimization. These models target spatial dispatch decisions and response time optimization; our model addresses a narrower problem: Estimating the *frequency* of blocking events with deliberately minimal inputs. When the available data consists only of dispatch timestamps, a simpler model that is calibrated from what *is* recorded fills a practical gap.

Time-of-Day Variation in Service Times

A central element of our model is the observation that mission durations vary systematically with time of day and that this variation can be predicted from the incident rate. Vandeventer et al. (2011) show that hospital turnaround times, which are a major component of total mission duration, are significantly associated with time of day, with emergency department crowding as a likely contributing factor. Other plausible mechanisms include daytime traffic congestion extending drive times and shifts in the severity mix between day and night. Indeed, Ingolfsson (2013) identifies load-dependent service times as an open research question; Rastpour et al. (2020) quantify the effect empirically, finding that service times increase measurably with the number of busy ambulances.

Most EMS simulation models either ignore this time dependence or handle it by fitting independent distributions per time period, multiplying the number of parameters. Our approach differs: We model per-bin duration parameters as linear functions of the *lagged incident rate*, yielding considerably fewer parameters than independent per-bin fits. To the best of our knowledge, using the incident rate itself as a covariate for service time distribution parameters has not been exploited in prior EMS modeling work.

Data-Driven EMS Planning in German Systems

The German EMS context presents specific features relevant to this work. The *Funkmeldesystem* – standardized radio status codes transmitted by vehicle crews – provides a structured record of mission phases, but as a vehicle-based system, it cannot track demand that goes unserved. Capacity planning in German EMS has traditionally relied on Poisson-based Erlang B models (Lindemann 2021), which estimate the required number of vehicles from the incident rate and a target recurrence interval for blocking events.

Degel et al. (2015) address a related problem from the Ruhr area of Germany: They derive time-dependent 'empirically required coverage' from observed distributions of concurrent active missions and embed this in an integer programming model for ambulance location. Their use of empirical concurrent mission data is conceptually close to our validation approach, but they use it as an *input* to optimization, not as a *validation target* for a generative model. They also assume constant service times, while we model them explicitly. Where Degel et al. ask 'given the observed load pattern, where should vehicles be stationed?', we ask 'given estimated incident rates and durations, does a simple simulation reproduce the observed load pattern – and if so, what blocking rate does it imply?'

In summary, analytical approaches such as Erlang B models assume piecewise stationarity and cannot capture transient effects from missions spanning hour or shift boundaries. Existing simulation-based approaches model spatial and operational detail but require many inputs, such as road networks, dispatch policies, and demand distributions by zone. To the best of our knowledge, no prior work has attempted to reconstruct time-dependent blocking from dispatch data alone using a model of the simplicity proposed here.

DATA

Source and Collection

We extracted data from the computer-aided dispatch system of a city of 250,000 inhabitants in western Germany, covering the entire year 2022 with 25,931 ambulance missions recorded.

Each record combines automatic timestamps from the dispatch center with semi-automated status reports from vehicle crews. The *Funkmeldesystem*, a dispatch and communications system common in Germany, defines standardized status codes that operators transmit at key points during a mission. For this paper, the time ambulance crews are scrambled (alert time) and the return to availability in the field (Status 1) or at the station (Status 2) are most relevant.

Preprocessing and Filtering

We restrict the analysis to ambulance (*Rettungswagen*, RTW) missions operated by city-owned vehicles. We exclude weekends and holidays, which exhibit a different demand regime; they are deferred to the full paper. After filtering, the dataset contains 18,978 missions across 260 weekdays.

Service Time

The simulation requires a measure of how long each mission occupies a vehicle. We define this *service time* as the interval during which the vehicle is unavailable to other calls; specifically, from the alert timestamp to the earliest of Status 1 (available in the field) or Status 2 (available at the station). The return trip to the station is not part of the mission; Status 1 marks the vehicle as available immediately after handover, while Status 2 is used when the vehicle reports availability only upon return. This quantity encompasses drive time, on-scene treatment, transport, and patient handover. When Status 1 is missing (12.4% of records), we impute it using the median offset between Status 2 and Status 1 observed in complete records.

Duty Schedule

The dispatch data does not contain explicit vehicle shift schedules. Rather than relying on nominal schedules – which may not reflect actual operational behavior, particularly regarding the latest time a vehicle still accepts missions before the shift ends – we estimate each vehicle's duty window directly from the data. We take the 2nd percentile of first-mission times and the 98th percentile of last-mission times across all observed days, rounded to the closest half-hour. This procedure yields an estimated duty schedule for 9 vehicles, of which 7 operate around the clock. We assume that these schedules are fixed across the entire observation period.

METHOD

Our model treats the EMS system as a time-dependent loss queue; in Kendall's notation, this corresponds to an $M(t)/G(t)/c(t)/0$ queue: Arrivals (incidents) $M(t)$ follow an inhomogeneous Poisson process with a time-of-day-dependent rate $\lambda(t)$. Service times (mission durations) $G(t)$ follow a time-dependent mixture distribution $D(t)$. The number of servers (vehicles) $c(t)$ depends on the time of day and the vehicles' shifts V. There is no queue: When all vehicles are occupied, arriving emergencies are *blocked*, i.e. lost from the system without being served or waiting. In plain terms: Incidents arrive randomly at a time-varying rate, each occupying a vehicle for a random duration; the number of available vehicles changes with shift schedules, and any incident that finds all vehicles busy is lost.

Analytical solutions exist only for simplified versions of this system under restrictive assumptions (Waldmann and Helm 2016). We therefore use MC simulation, which can handle the full time-dependent system directly.

The model requires three fitted components: the incident rate function $\lambda(t)$, the time-dependent mission duration distribution $D(t)$, and the vehicle duty schedule V. We describe each in turn.

Incident Rate

The rate at which emergencies occur varies strongly with time of day. In this section, we estimate the incident rate as a two-component wrapped generalized normal mixture, fitted to censoring-corrected hourly counts.

Censoring Correction

When considering the mean number of starting missions $\lambda_{\text{obs}}(t)$ as a proxy for the true incident rate $\lambda_{\text{true}}(t)$, a complication arises: During periods when all vehicles are occupied, *emergencies* still occur but do not generate *mission* records. The observed mission start rate therefore *underestimates* the true incident rate when the system is under heavy load.

To correct for non-recorded incidents, we compute the empirical fraction of time at which all on-duty vehicles are simultaneously occupied. At minute resolution, we observe the number of concurrently active missions and group these observations into 24 hourly bins (pooling across all observed days). Within each bin we compute

$$f_{\text{full}}(t) = \frac{\text{minutes at full capacity in bin } t}{\text{total minutes in bin } t},$$

i.e., the fraction of time the system was fully occupied. The corrected incident rate estimate is then

$$\lambda_{\text{true}}(t) \approx \frac{\lambda_{\text{obs}}(t)}{1 - f_{\text{full}}(t)} := \lambda(t)$$

This first-order correction assumes that non-recorded, i.e. blocked, emergencies arrive at roughly the same rate as served ones during the blocking interval. In our case, the correction is negligible during low-demand hours (where $f_{\text{full}}(t) \approx 0$) and most pronounced during the day.

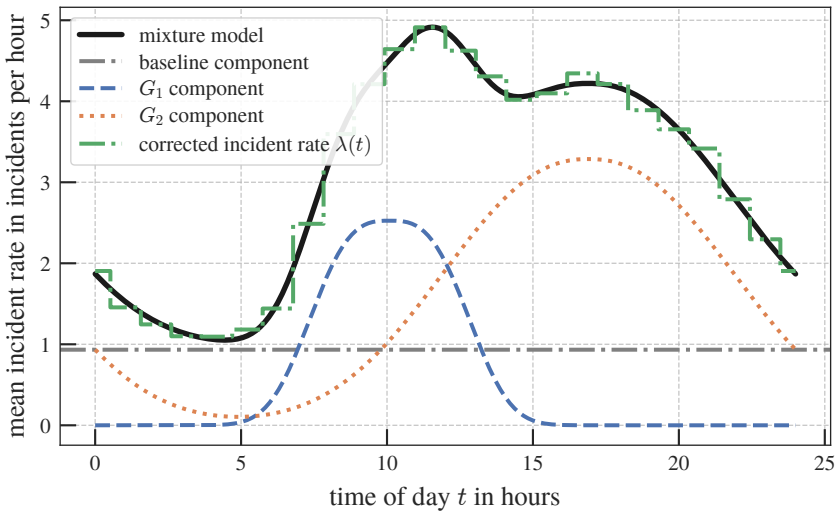

Figure 1. Fitted incident rate model: Two wrapped generalized normal components (dashed, dotted) and a baseline component (dash-dotted). The resulting mixture distribution (solid) is compared against the corrected hourly incident rate (dash-dotted step function). The model clearly identifies the morning and afternoon peaks that are typical for incident data.

Parametric Model

We model the corrected incident rate $\lambda(t)$ as a sum of two wrapped generalized normal (WGN) components plus a constant baseline:

$$\lambda(t) \approx b + \sum_{i=1}^{2} w_i \, f_{\mathrm{WGN}}(t;\, \mu_i, \alpha_i, \beta_i),$$

where b is the baseline rate capturing time-independent demand, and the two WGN components, weighted by w_i, capture the time-dependent demand. The wrapped generalized normal distribution is defined as

$$f_{\mathrm{WGN}}(t;\, \mu, \alpha, \beta) = \sum_{k=-\infty}^{\infty} \frac{\beta}{2\,\alpha\,\Gamma(1/\beta)} \exp\left(-\left|\frac{t-\mu-kT}{\alpha}\right|^{\beta}\right).$$

This distribution extends the wrapped normal distribution with a shape parameter β that controls tail behavior ($\beta = 2$ recovers the wrapped normal; smaller values produce heavier tails). We choose a period $T = 24\,\mathrm{h}$. The wrapping enforces strict periodicity $\lambda(t) = \lambda(t + 24\,\mathrm{h})$, ensuring a smooth transition across midnight.

The two WGN components yield nine free parameters $(b, \mu_1, \alpha_1, \beta_1, w_1, \mu_2, \alpha_2, \beta_2, w_2)$ estimated by nonlinear least-squares regression against the corrected incident rate $\lambda(t)$ in one-hour bins. The binning suppresses the noise present at finer temporal resolutions. The best fit is selected from multiple random initializations; the resulting decomposition is shown in Figure 1.

Mission Duration

The time an ambulance spends on a single emergency mission, from dispatch to patient handover, varies considerably with the time of day. We model mission duration as a time-dependent gamma-exponential mixture $D(t)$ whose parameters are linked to the incident rate $\lambda(t)$, reducing a potentially high-dimensional model to nine free parameters. Since we fit the *aggregate* distribution of durations rather than modeling operational details, the approach captures the overall shape (how long vehicles are bound, and with what variability) without requiring detailed data on mission phases.

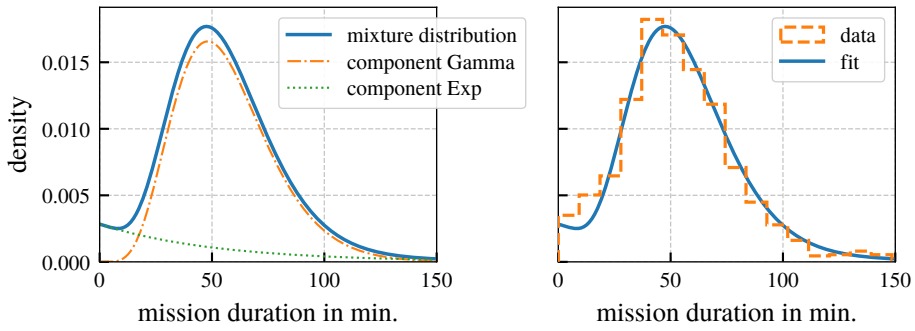

Figure 2. Gamma-exponential mixture fitted to observed mission durations. We only show the fit for one of the 12 bins, by way of example, choosing the (22:00–24:00)-bin that exhibits the worst fit (highest EMD). *Left:* The mixture components (dotted, dash-dotted) combine into the mixture distribution (solid). *Right:* The same mixture distribution (solid), compared to the empirical duration histogram (dashed steps). In this example, the fit underestimates the frequency of short missions slightly.

		$\lvert\mathrm{corr}(\lambda(t-\Delta t),\,\cdot)\rvert$			
Dataset	Δ in hrs.	μ_t	α_t	β_t	w_t
Ours	2.0	1.0	0.6	0.6	0.8
Manhattan	1.6	0.9	0.7	0.8	0.1

Table 1. Pearson correlation between fitted, per-bin duration parameters and the lagged incident rate.

Per-Bin Mixture Fitting

As a preliminary step, we divide the day into $B = 12$ equal bins and fit a separate mixture distribution to the durations of missions starting in each bin. The mixture model $D(t)$, expressed as a probability density function $f(d \mid t)$ that captures how likely a given duration d is to be observed at time bin t, has two components:

$$f(d \mid t) = (1 - w_t)\,\Gamma(d;\,\mu_t, \alpha_t) + w_t\,\mathrm{Exp}(d;\,\beta_t),$$

where μ_t and α_t are the mean and shape parameters of the gamma component, β_t is the rate parameter of the exponential component, and w_t is the mixture weight. The gamma component captures the bulk of typical missions, while the exponential component absorbs unusually short missions such as false alarms and early cancellations. Each per-bin fit has four free parameters $\Theta(t) = \{\mu_t, \alpha_t, \beta_t, w_t\}$, yielding $4 \times 12 = 48$ parameters in total – far too many for a parsimonious model. An independent per-bin model would, by construction, fit the training data better than a constrained model, but it offers no explanatory power beyond interpolation: With 48 free parameters, overfitting is a concern, and generalization to modified demand scenarios is not justified.

Linking Parameters to the Incident Rate

Thus, rather than using B independent mixture models, we exploit an empirical regularity: All per-bin parameters Θ correlate strongly with the time-shifted incident rate λ (Pearson r between 0.6 and 0.9; the probability of observing $|r| \geq 0.6$ by chance alone is at most $p = 0.04$). This correlation is consistent with known mechanisms linking call volume to service time: Emergency department crowding increases hospital turnaround times (Vandeventer et al. 2011), daytime traffic congestion extends drive times, and the mission severity mix shifts between day and night. What appears to be new in our approach is not the observation of time-of-day variation itself, but the use of the incident rate as a direct predictor of the duration distribution parameters.

A full evaluation of this mechanism is beyond the scope of this work-in-progress paper; however, to motivate that it is not merely a statistical coincidence, we compare the per-bin correlations from our study city against an EMS dataset covering Manhattan (City of New York 2026), which differs substantially in fleet size and operational protocols. Table 1 shows that the correlations are of comparable magnitude for at least three parameters.

We model each of the mixture parameters $\Theta_i(t)$ as a linear function of the incident rate evaluated at a learned time lag Δ:

$$\Theta_i(t) = m_i \cdot \lambda(t - \Delta) + c_i,$$

where m_i and c_i are slope and intercept for parameter i. The time lag $\Delta = 2.0\,\text{h}$ is chosen once to maximize the Pearson correlation between the lagged incident rate and the per-bin gamma mean μ_t, as this is the dominant parameter governing overall mission duration. The fit is shown in Figure 2 for one exemplary bin. This linearization reduces the model from 48 per-bin parameters to nine free parameters (four slopes, four intercepts, and the shared time lag Δ), while enabling smooth interpolation between bins and extrapolation under modified demand scenarios. The tradeoff is that nonlinear and bin-specific deviations are lost, but the smaller parameter count yields better interpretability and less overfitting. Reduced overfitting is particularly important for cities with smaller EMS systems and for year-over-year comparisons.

To draw a mission duration for an incident at time t, we evaluate $\lambda(t - \Delta)$ from the fitted incident rate model, compute the four mixture parameters from their linear fits, construct the gamma-exponential mixture, and draw from it.

MC Simulation

Given the fitted incident rate $\lambda(t)$ and the time-dependent mission duration distribution $D(t)$ derived in the previous sections, together with the estimated duty schedule V, we construct an MC simulation of EMS vehicle operations. The simulation proceeds in discrete time steps $\Delta = 1$ min, generates incidents from the inhomogeneous Poisson process with rate $\lambda(t)$, assigns each a duration drawn from $D(t)$, and dispatches available vehicles. When no vehicle is available, the incident is recorded as blocked. Algorithm 1 summarizes the procedure.

Algorithm 1: MC simulation of EMS vehicles

```
Input:
  λ(t)      – arrival rate function for emergencies at time t
  D(t)      – time-dependent probability distribution of mission durations
  V         – set of vehicles with a known duty schedule
  Δt        – simulation time step
  T_end     – simulation end time

Output:
  missions  – list of records (vehicle, t_start, t_end),
              where vehicle = null if no vehicle was available

Procedure Simulate():
  t ← 0
  missions ← empty list

  while t < T_end:
    n ← draw from Poisson(λ(t)) # number of new incidents in [t, t+Δt)

    for each new incident j = 1, …, n:
      d ← draw from D(t) # mission duration conditioned on arrival time
      v ← find a vehicle in V that is both on duty and not on a mission at time t

      if v ≠ null:
        mark v as on mission during [t, t + d]

      append (vehicle: v, t_start: t, t_end: t + d) to missions

    t ← t + Δt
```

The simulation runs over multiple consecutive days, with a burn-in period of 30 days to allow the system state to reach a realistic distribution before collecting statistics.

A vehicle that begins a mission during its duty period completes that mission even if the shift nominally ends before the vehicle returns. This behavior is consistent with our duty schedule estimation, which derives shift boundaries from observed mission times rather than nominal schedules.

This design is deliberately simple; there is no spatial structure, no dispatch priority logic, and no interaction between concurrent missions. This allows us to test a focused question: Is the statistical structure captured by three fitted inputs, together with the simple system dynamics defined through Algorithm 1, sufficient to reproduce the aggregate behavior of the real system?

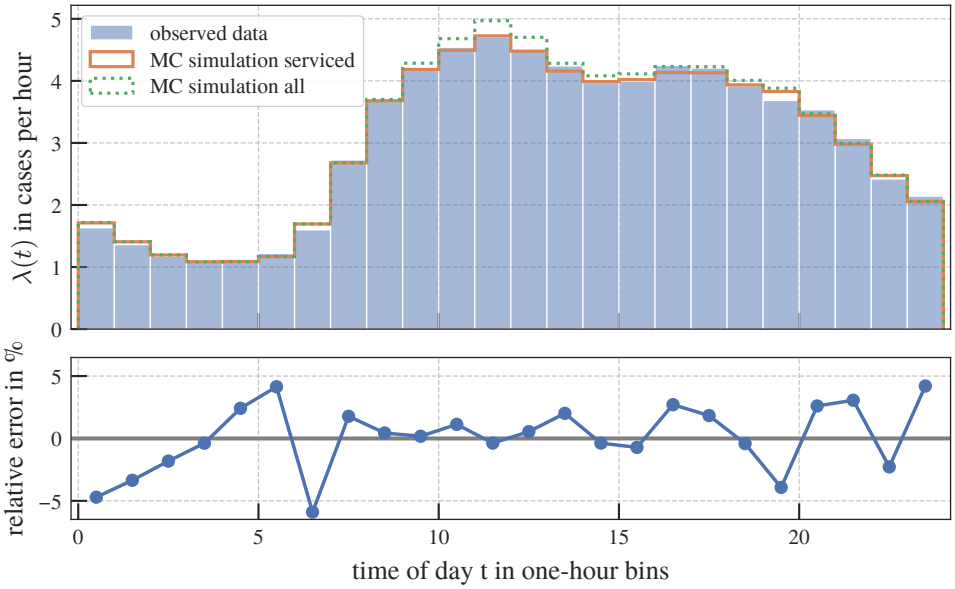

Figure 3. ***Top:*** **Mean hourly mission start rate from observed data (shaded bars) and simulation (solid line), together with the estimated true incident rate used internally by the simulation (dotted line).** ***Bottom:*** **Relative error between observed and simulated start rates.**

EVALUATION

Overall, the simulation reproduces the observed system behavior well: Our model with 18 fitted distributional parameters and no spatial structure recovers the time-of-day distribution of concurrent active missions to within a mean absolute error of 1.0%pt.

We fit $\lambda(t)$ and $D(t)$ on the entire dataset of 260 weekdays and generate 3900 days of synthetic data, which takes 0.7 seconds on a consumer laptop. Together with the duty schedules V, these 18 fitted distributional parameters are the only inputs to the simulation; the duty schedule is estimated separately from the data and contributes additional per-vehicle parameters.

Below, we first validate the simulation against observed data, then present the blocking probability estimates and a counterfactual scenario.

Incident Rate

As a first consistency check, we compare the mean hourly mission start rate between observed data and the simulation's served missions (Figure 3). Since the simulation generates incidents at the corrected rate $\lambda(t)$, but only records those that find an available vehicle, the simulated start rate should recover the observed censored rate if the model is consistent. The observed agreement is close, with a mean relative absolute error of 2.1%; the largest discrepancies occur during low-demand hours where absolute errors are small.

This check is primarily driven by the incident rate model; the duration distribution has only a minor effect as long as mean mission times are modeled correctly. Increasing the number of components in $\lambda(t)$ by one would reduce the mean relative absolute error to 1.5 %.

Concurrent Active Missions

The primary validation target is the distribution of concurrent active missions $\hat{p}(n,t)$: The probability that exactly n vehicles are simultaneously engaged at time-of-day t. This quantity emerges from the dynamic interaction of all three model inputs and was not used directly in calibrating any of them. The only indirect link is through $f_{\text{full}}(t)$, which adjusts the magnitude of $\lambda(t)$ but does not constrain the distributional shape across occupancy levels; $\hat{p}(n,t)$ therefore serves as a largely out-of-sample test.

We compute $\hat{p}(n,t)$ at hourly resolution for both data and simulation, comparing them as a family of curves per occupancy level n (Figure 4). For the empirical data, 95 % confidence intervals are computed from day-to-day

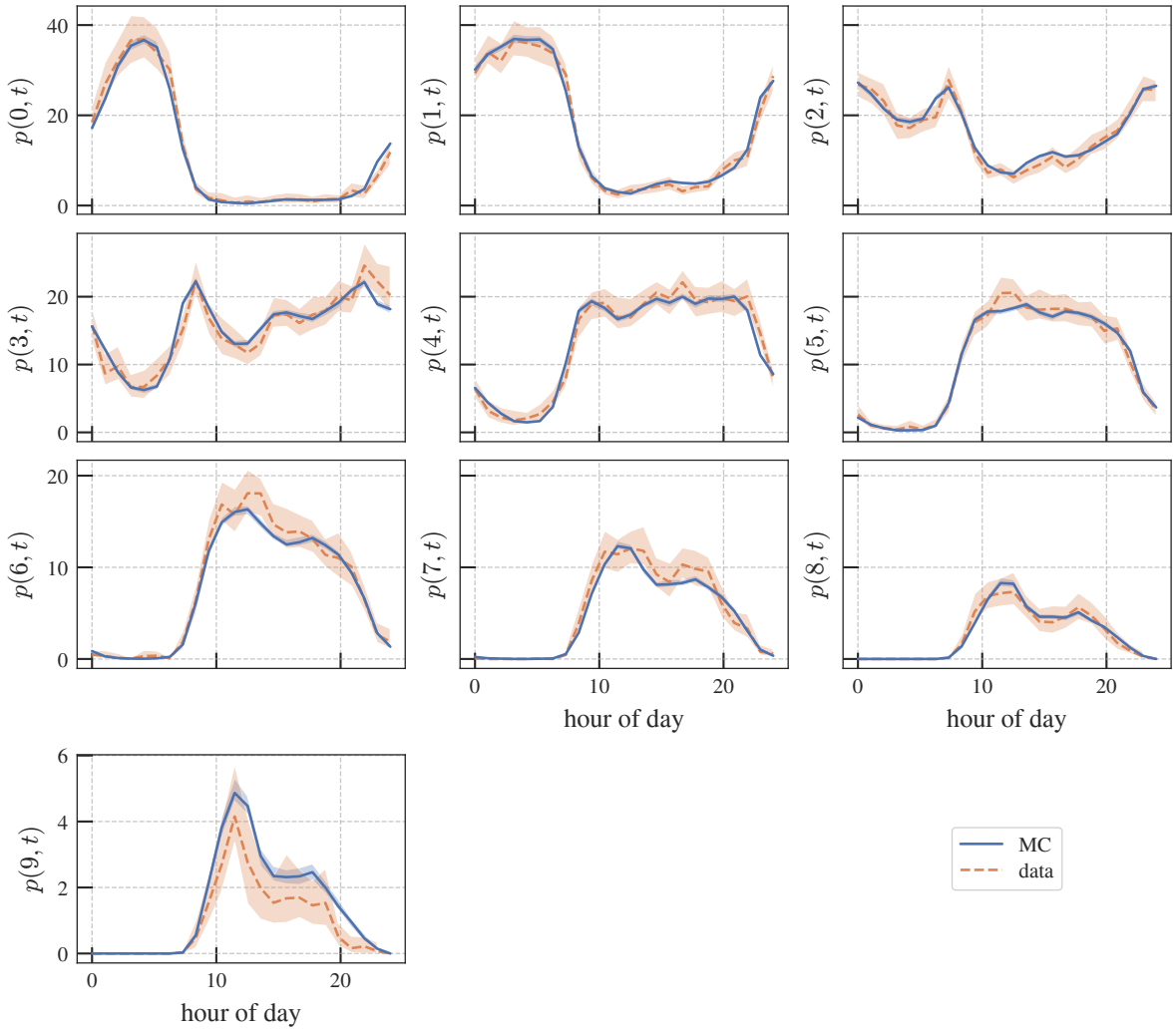

Figure 4. Probability $\hat{p}(n,t)$ of exactly $n = 0,\ldots,9$ concurrent active missions by time of day, in percent. Dashed lines: empirical estimate from observed data (shaded: 95 % CI). Solid lines: simulation. Note that the Y-axis scales are shared within a row, with the bottom row's values being much smaller.

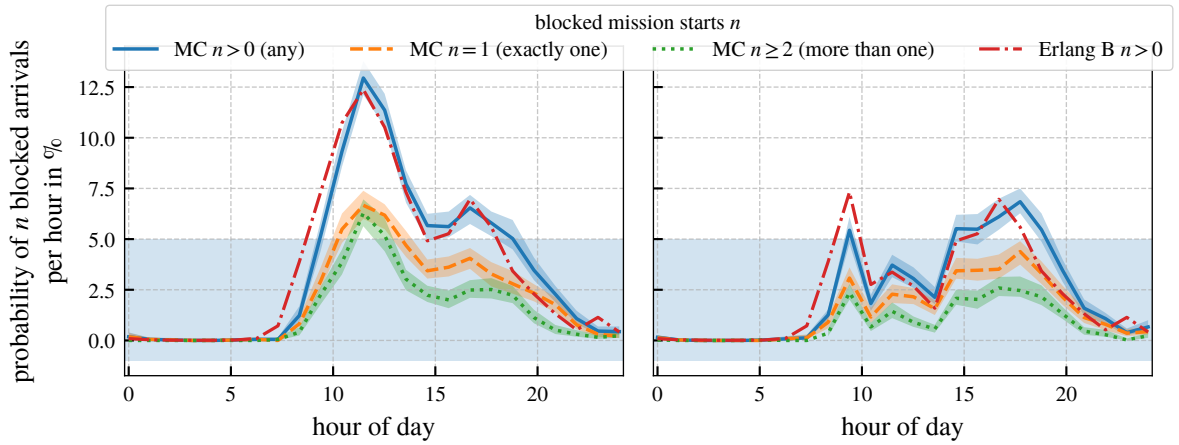

Figure 5. Estimated per-hour probability that at least one blocked (solid), exactly one (dashed), and two or more (dotted) arriving incidents are blocked, using the MC simulation, with 95 % CIs. An estimate using Erlang B (dash-dotted) is provided for comparison. *Left:* Baseline scenario with the current fleet configuration. *Right:* Counterfactual scenario with two additional ambulances deployed between 10:00 and 14:00. The noon blocking peak is substantially reduced, bringing the hourly risk close to 5 %.

variance via bootstrap resampling; for the simulation, enough MC samples are generated that simulation uncertainty is negligible. About 77% of simulated values fall within the empirical 95 % confidence intervals, and the mean absolute error across all (n, t) pairs is 1.0%pt. For a parsimonious model, this level of agreement is encouraging, even in light of the statistically significant differences between the model and the data. The model captures the rise and fall of occupancy levels throughout the day, including the afternoon peak where five or more vehicles are frequently active simultaneously. The largest discrepancies occur at $n = 9$ (full occupancy), where the simulation slightly but systematically overestimates the probability of all vehicles being engaged simultaneously.

Blocking Probability and Scenario Analysis

The simulation estimates 1.5 blocked missions per weekday on average, out of roughly 72.7 total observed daily missions. Each blocked case represents a delayed or degraded emergency response. The left panel of Figure 5 shows the hourly blocking risk, decomposed by the number of simultaneously blocked cases. The probability of at least one blocked mission in a given hour peaks during midday, just after the period of highest concurrent vehicle engagement in Figure 4. Overnight, the blocking risk is negligible. About half of blocking events are isolated (dashed line); simultaneous blocks (dotted line) are slightly less common, suggesting the system operates near but not far beyond its capacity limit.

To illustrate the model's utility for scenario exploration, we simulate two additional ambulances with shifts from 10:00 to 14:00 (Figure 5, right panel). Recall that these times define the earliest and latest mission starts; a mission dispatched at 13:59 is still served to completion. This intervention brings the hourly blocking risk close to 5 % during the new shift, reducing blocked cases to 0.9 per day. Cases after 15:00 are virtually unaffected. Quantifying this marginal benefit of specific fleet changes on time-resolved blocking risk is the primary operational application of the model and is not achievable from dispatch records alone. Note that the model lacks spatial structure, and thus cannot capture secondary effects such as reduced drive times from stationing additional vehicles at specific locations.

Comparison with Piecewise Erlang B

To place the simulation in context, we compare against the Erlang B formula applied as a pointwise stationary approximation (PSA) to hourly bins, which is a standard analytical tool in German EMS planning. PSA treats each hour as an independent $M/M/c/0$ system in steady state.

Total blocked cases over a full day agree (1.4 cases/day), but hour by hour, PSA overestimates blocking risk on the rising flank of demand and underestimates on the falling flank (Figure 5, dash-dotted). This is consistent with the transient lag it ignores by construction. Operationally, shifts dimensioned from PSA therefore start and end too early. The same limitation shows up in $\hat{p}(n, t)$: Only 53 % of PSA predictions fall within the empirical 95 % CI (not shown in Figure 4 to avoid clutter), versus 77% for the MC model.

DISCUSSION

The model's primary practical value is as a lightweight tool for verifying and extending analytical EMS capacity planning results. As shown in the Erlang B comparison, the simulation captures transient effects that piecewise stationary approximations miss by construction.

Fitting parametric distributions to aggregate data (binned counts, pooled durations, shift percentiles) rather than individual records makes the model tolerant to noisy, low-resolution timestamps and missing status codes often found in operational dispatch logs. The duration–rate linkage that reduces 48 to nine parameters is not only a parameter-saving device: It encodes a plausible mechanism, since emergency department crowding, traffic congestion, and severity mix jointly drive both call volume and service time. Notably, the duration model D contains no explicit time-of-day dependency; time enters only indirectly through the arrival rate, i.e. $D(\lambda(t))$ rather than $D(t, \lambda(t))$. That this formulation suffices to reproduce the observed load pattern (Figure 4) is consistent with demand-driven service times. Establishing causality requires further investigation. This structural simplicity also makes the model a candidate for transfer to other cities, although our correlation study on the Manhattan dataset (Table 1) hints that the linear coefficients require re-fitting from local data. The w_t parameter does not transfer at all, probably due to operational differences.

Limitations

While the goal of this work-in-progress paper is to establish that the duration–incident-rate correlation can serve as the basis for a parsimonious simulation, several limitations should be noted:

First, the model treats the city as a single service zone with a shared vehicle pool. In reality, vehicles are stationed at 6 locations, and dispatch decisions account for geographic proximity. This means our model cannot capture local blocking (all nearby vehicles occupied while others are free), nor can the scenario analysis reflect reduced drive times from improved geographic distribution. For a compact urban area, this simplification may be acceptable; for larger, more dispersed service regions, spatial structure must be incorporated.

Second, the incident rate model assumes that emergency demand follows the same pattern every weekday. Seasonal variation, day-of-week effects beyond the weekday–weekend distinction, and special events are not captured. In particular, we exclude weekends entirely and treat all weekdays as interchangeable, which introduces artifacts around Friday and Monday midnight where the real system transitions to and from a different demand regime.

Third, the model cannot be validated against the quantity it is designed to estimate. This is inherent to the problem setting: The absence of recorded blocking events is precisely the gap that motivates the simulation. Our validation is therefore indirect: We show that the model reproduces the observed *load* distribution and argue that a model which correctly predicts load should also produce reasonable estimates of *overflow* and system dynamics. This argument is strongest when blocking is rare and weakest when the system is chronically overloaded.

Fourth, and most fundamentally, the model has been calibrated on a single dataset from one mid-sized German city. The cross-city correlation check on the Manhattan data (Table 1) is encouraging but addresses only one model component in isolation; it does not establish that the full simulation pipeline (censoring correction, rate fit, duration linkage, finite-pool dynamics) transfers end-to-end. Whether the model structure holds for systems with substantially different fleet sizes, demand volumes, geography, or operational protocols is an open question requiring multi-site validation, which we defer to the full paper.

CONCLUSION AND FUTURE WORK

We have presented a parsimonious MC simulation framework for modeling urban ambulance fleet dynamics from routine dispatch records. The model combines a time-varying Poisson incident process, a gamma-exponential mixture duration model linked to the incident rate, and estimated duty schedules into a simulation that reproduces the observed distribution of concurrent active missions. A key empirical finding is that mission duration parameters correlate strongly with the lagged incident rate, which we exploit to reduce the duration model from 48 per-bin parameters to nine while preserving the time-of-day structure. Initial validation against one year of operational data from a mid-sized German city shows good agreement between observed and simulated behavior.

This work is in progress, and several extensions are planned. Most importantly, we plan to validate the full simulation pipeline on additional EMS systems to assess whether the model structure generalize beyond the single system studied here. The cross-city correlation evidence we presented here is encouraging but insufficient: A complete transfer study requires running the simulation end-to-end on independent data and comparing blocking estimates. Secondly, we intend to incorporate spatial structure by modeling station catchment areas and distance-dependent

dispatch, allowing estimation of *local* blocking probabilities, and to extend the incident rate model to weekend, seasonal, and holiday effects. Third, we aim to calibrate across multiple years to assess temporal stability and demand trends, and to conduct systematic scenario analyses covering a range of fleet sizes and shift configurations. Fourth, we plan to obtain aggregate blocking counts from the cooperating institution as independent validation, and to add uncertainty propagation through bootstrap resampling to quantify confidence intervals on blocking estimates. Finally, we plan an ablation of the duration model against two simpler variants – a constant mean service time, and a single gamma-exponential mixture pooled across bins – to quantify the value added by the rate-correlated parameterization.

Acknowledgements

We thank our cooperating EMS agency, which wishes to stay anonymous, for providing the dispatch data used in this study.

The authors used Claude 4.6 under close human supervision during drafting and revision to restructure prose, refine phrasing and style, verify consistency, and to extensively refactor the analysis source code. All scientific content, including modeling decisions, idea generation, and analysis, was produced solely by the authors.

REFERENCES

City of New York (2026). *EMS Incident Dispatch Data*. http://catalog.data.gov/dataset/ems-incident-dispatch-data.

Degel, D., Wiesche, L., Rachuba, S., and Werners, B. (2015). Time-Dependent Ambulance Allocation Considering Data-Driven Empirically Required Coverage. *Health Care Management Science* 18.4, 444–458. https://doi.org/10.1007/s10729-014-9271-5.

Ingolfsson, A. (2013). EMS Planning and Management. *Operations Research and Health Care Policy*. Ed. by G. S. Zaric. New York, NY: Springer, 105–128. https://doi.org/10.1007/978-1-4614-6507-2_6.

Lindemann, T. (2021). *Feuerwehrbedarfsplanung*. 1st ed. Stuttgart: W. Kohlhammer GmbH. https://doi.org/10.17433/978-3-17-035396-1.

Rastpour, A., Ingolfsson, A., and Kolfal, B. (2020). Modeling Yellow and Red Alert Durations for Ambulance Systems. *Production and Operations Management* 29.8, 1972–1991. https://doi.org/10.1111/poms.13190.

Restrepo, M., Henderson, S. G., and Topaloglu, H. (2009). Erlang Loss Models for the Static Deployment of Ambulances. *Health Care Management Science* 12.1, 67–79. https://doi.org/10.1007/s10729-008-9077-4.

Vandeventer, S., Studnek, J. R., Garrett, J. S., Ward, S. R., Staley, K., and Blackwell, T. (2011). The Association Between Ambulance Hospital Turnaround Times and Patient Acuity, Destination Hospital, and Time of Day. *Prehospital Emergency Care* 15.3, 366–370. https://doi.org/10.3109/10903127.2011.561412.

Waldmann, K.-H. and Helm, W. E. (2016). *Simulation stochastischer Systeme*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-49758-6.

Whitt, W. and Zhao, J. (2017). Many-Server Loss Models with Non-Poisson Time-Varying Arrivals. *Naval Research Logistics (NRL)* 64.3, 177–202. https://doi.org/10.1002/nav.21741.

Yang, W., Su, Q., Huang, S. H., Wang, Q., Zhu, Y., and Zhou, M. (2019). Simulation Modeling and Optimization for Ambulance Allocation Considering Spatiotemporal Stochastic Demand. *Journal of Management Science and Engineering* 4.4, 252–265. https://doi.org/10.1016/j.jmse.2020.01.004.