American Journal of Traffic and Transportation Engineering
2026, Vol. 11, No. 1, pp. 1–14
https://doi.org/10.11648/j.ajtte.20261101.11

Review Article

Performance Approximation for Highway Emergency Rescue Systems

Song Jianhui[1], Shu Xiaofeng[2], Zhong Jie[1], Jiang Lanqing[3, *], Zhang Wenchao[4], Shi Guoqing[4]

[1]Jiangsu Sutong Bridge Co., Ltd, Soochow, China
[2]Transportation Engineering Construction Bureau, Soochow, China
[3]Jiangsu Baisheng Information Technology Co., Ltd, Soochow, China
[4]China Communications Tunnel Engineering Bureau Co., Ltd, Soochow, China

Abstract

Ensuring prompt responses to spatially uncertain traffic accidents is critical for highway emergency rescue systems, aiming to address emergencies and to prevent network-wide traffic disruptions. This paper develops an approximation model to characterize priority-based rescue systems for RV fleets. By representing each RV's state as either idle or busy, we formulate an approximation model in which the dispatch probability of an RV is dynamically influenced by the utilization of higher-priority RVs and its own availability. The proportionality constant governing this relationship is derived analytically from an M/M/N queuing model, incorporating a random server selection mechanism without replacement until the first available RV is identified. The system of equations is simplified into a tractable set of N linear simultaneous equations, enabling efficient computation of RV workloads. We further propose a solution algorithm to estimate critical performance metrics, including individual RV workloads, system response times, and cross-dispatch frequencies. A practical case study based on the G15 Highway (Sutong Bridge section) in China illustrates the model's applicability in quantifying performance measures such as individual RV response times, cross-regional dispatch ratios, and workload distribution under varying vehicle allocations and priority levels. The proposed model provides a practical and computationally efficient tool to support RV location planning, response section partitioning, and overall performance evaluation in highway emergency services, offering valuable insights for optimizing rescue efficiency and resource allocation.

Keywords

Highway Emergency, RV Deployment, RV Dispatch Priority, Rescue Work Intensity

1. Introduction

With the ongoing expansion of the worldwide Highway network, the number of traffic accidents and related casualties has been steadily increasing [1]. Against this backdrop, establishing an efficient Highway Rescue System (HRS) has become a crucial measure for reducing accident-related fatalities and injuries [2]. The efficiency of the HRS largely de-

*Corresponding author: 1223586577@qq.com (Jiang Lanqing), 2099230995@qq.com (Jiang Lanqing)

Received: 11 November 2025; **Accepted:** 9 December 2025; **Published:** 19 January 2026

pends on the response speed of Rescue Vehicles (RVs) to traffic incidents. Essentially, the HRS is a demand-responsive system that addresses traffic accidents, which are randomly distributed in space, by deploying RVs [3, 4]. Therefore, deriving its operational performance requires careful analysis of the spatially stochastic accidents and the dispatch and operation characteristics of RVs. Accurately evaluating the performance of the HRS is of significant practical importance for enhancing the overall emergency response capabilities of society.

The development of methodologies for highway emergency response systems has evolved through several phases, marked by successive refinements to address computational and practical constraints. Larson introduced the Hypercube Queuing Model (HQM), which characterizes systems such as emergency services via continuous-time Markov chains and provides performance metrics like workload and service rates for fixed server configurations [5, 6]. However, despite its descriptive power, the HQM is limited to analysis rather than optimization [7]. Early assumptions—such as service time being independent of call and server locations—simplified modeling but led to exponential state growth as server numbers increased. Subsequent efforts focused on overcoming these limitations: Larson proposed heuristics for larger systems, later adapted by Jarvisto incorporate server- and customer-dependent service times [8, 9]. Further extensions included multi-state models, partial state representations, and station-level approximations for systems with collocated RVs [10-13].

Subsequent studies on emergency systems address three main themes: strategic RV location planning, operational dispatch and routing, and hybrid approaches. The first two themes address long-term planning and daily operations, respectively. Studies on strategic RV location focus on optimal placement under limits like budgets [14, 15]. Some papers improve classic models to include costs for stations, RVs, and relocation. They show that spending on both new stations and new RVs works better than only buying more RVs [16-18]. Other work adds goals like fairness or handling travel time uncertainty [19, 20].

Studies on dispatch and routing develop better real-time assignment rules. Some research creates methods for pandemics, using two-step models to reduce infection risk while maintaining service [21]. Other work provides flexible assignment policies for systems with multiple RVs and shared duties [1, 3, 22-26]. Recent extensions to dispatch modeling include cooperative service scenarios where multiple RVs respond simultaneously to incidents [31], approximate algorithms for large-scale systems with variable service rates [4], and dedicated server policies for high-priority calls with efficient computational methods [4, 27, 28].

Hybrid approaches explicitly model spatial interactions between RVs, formulating integrated optimization models to simultaneously determine server locations, districting plans, and multiple-dispatch policies [12, 29, 30]. Overall, these studies commonly employ the HQM-based models as an underlying analytical tool, which is then embedded within metaheuristic frameworks such as genetic algorithms to solve the resulting complex optimization problems.

A major limitation of these models is their computational complexity. Given n servers each with m states, the system has m^n unique states (e.g., 10 servers with 2 states each yield 2^{10}=1024 states). Consequently, the solution space grows exponentially with m and n. For large m and n values, the computation time becomes prohibitively long, and individual state probabilities become infinitesimally small. As a result, solving models for larger-scales emergency systems is often computationally infeasible on personal computers [8, 30]. This intractability arises because traditional HQM-based approaches require enumerating all possible system states.

In this paper, we develop an approximate variant of the hypercube queueing model for HMS that simultaneously addresses three key limitations in the literature: (i) the incorporation of spatial demand characteristics, including server availability constraints; (ii) a reduction in model complexity to a system of only N linear equations—a significant simplification from the 2^N equations required by the exact formulation—coupled with an efficient heuristic solution algorithm; and (iii) the flexibility to accommodate pre-defined dispatching preferences. The proposed model serves as a practical tool for evaluating critical performance measures of a given system configuration.

This paper is structured as follows. Section 1 presents the approximation of the hypercube model. Section 2 presents the solution approach for the proposed approximation model. Section 3 presents the application of the model on a highway. Section 4 concludes this paper with its contributions and suggestions for future research.

2. Model Formulation

2.1. Problem Description and Assumptions

To elucidate the emergency rescue mechanism in a highway, we analyze the bidirectional section in Figure 1 as an illustrative example. The segment is divided into three emergency response districts, with upstream and downstream lanes separated by a median barrier. Strategically placed openable barrier gaps enable cross-district rescue deployment. RVs 1, 2, and 3 are stationed at emergency posts in districts 1, 2, and 3, respectively. Upon receiving a distress call from district 1, the dispatch center assigns RV 1 based on proximity. After completing the rescue, the unit returns to its post. For emergencies in district 2, RV 2 serves as the primary responder. However, if RV 2 is engaged—leaving the district temporarily uncovered—the system may reallocate RV 1 for cross-district intervention. When multiple incidents occur simultaneously across districts, temporary resource shortages may leave some districts without available units. In such cases, idle RVs from adjacent districts must provide cross-district support. This protocol demonstrates that both standby positioning and the quantity of RVs critically influence performance metrics, particularly average emergency response time.

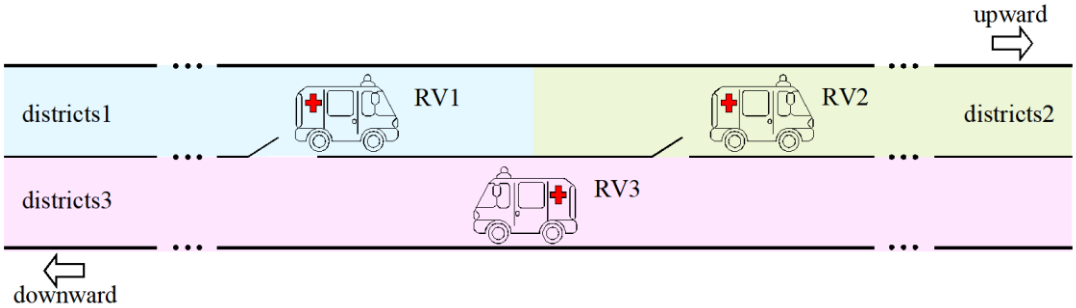


Figure 1. *Disposition of Districts and RVs for the Illustrative Example.*

Based on the aforementioned rescue process, the modeling assumptions are as follows:

A1. Assume that the highway section is divided into K districts (emergency districts), and a total of N RVs are deployed throughout the section to handle rescues.

A2. When an emergency call is received, the emergency station dispatches the nearest available RV to the location of the call on a first-come-first-served (FCFS) basis.

A3. It is assumed that the RVs are homogeneous, and their rescue duration follows a negative exponential distribution with a mean of $1/\mu$, independent of the location and time of the incident.

A4. In district *j*, the emergency demand follows a Poisson distribution with an average of $\lambda f_j \equiv \lambda_j$, where λ is the emergency demand occurrence rate, f_j is the proportion of district j's emergency demand to the total emergency demand occurrence rate, and λ_j is the average rate of service calls generated from district *j*.

The symbols used in the following text are explained in Table 1.

Table 1. *Notation Table.*

N	Number of RVs.
K	Number of emergency districts.
μ^{-1}	Average rescue time for emergency calls.
f_j	Proportion of district j's emergency demand to the total emergency demand occurrence rate ($\sum_j f_j = 1$).
λ	Average rate of emergency demand occurrence.
λ_j	Average rate of emergency demand generated from district j, $\lambda f_j \equiv \lambda_j$
τ_{lj}	Average travel time from district l to district j.
ρ	Service intensity of the queuing system, $\rho=\lambda/(N\mu)$
r	Service intensity of the queuing system (for M/M/N/∞ system, r = $\rho=\lambda/(N\mu)$).
B_i	Event: The i-th selected RV is busy, i = 1, 2,..., N.
F_i	Event: The i-th selected RV is idle, i = 1, 2,..., N.
S_k	State where exactly k RVs are busy, k = 1, 2,..., N.
P_k	Probability that k RVs are busy, k = 1, 2,..., N.
$Q(N,\rho,i)$	Multiplication factor used to calculate the probability that the (i+1)-th selected RV is the first available one, given N RVs and service intensity, i = 1, 2,..., N-1.
w_i	Event: RV i is in a busy state.
$\bar{w}_i$	Event: RV i is in an idle state.
ρ_i	Work intensity of RV i.
η_i	Total rate at which RV i is assigned to calls (per unit time).
$\bar{\eta}_i$	Effective rate at which RV i is assigned to calls when idle.
λ_D	Intensity of rescue tasks awaiting assignment to a RV when all RVs are busy, $\lambda_D = \lambda P_N / N$.

2.2. Approximation Approach

After an accident occurs, RVs are dispatched on a first-come-first-rescue basis. Based on assumptions A2-A4, the system of emergency calls and N RVs can be modeled as an M/M/N/∞ queuing system. And $P\{B_1B_2\cdots B_iF_{i+1}\}$ denotes the probability that the i-th selected RV is idle while the preceding i-1 selected RVs are all busy. Consequently, according to the law of total probability for conditional probability, it follows that

$$P\{B_1 B_2\cdots B_i F_{i+1}\} = \sum_{k=0}^{N} P\{B_1 B_2\cdots B_i F_{i+1}|S_k\} P_k \quad (1)$$

The state probability of P_k is

$$P_N = N^N \rho^N P_0 / N!(1\text{-}\rho) \quad (2)$$

The probability that all N RVs are either fully busy or completely idle is

$$P_N = N^N \rho^N P_0 / N!(1\text{-}\rho) \quad (3)$$

$$P_0 = \left[\sum_{i=0}^{N-1} N^i \rho^i / i! + N^N \rho^N \big/ N!(1\text{-}\rho)\right]^{-1} \quad (4)$$

In Equation (1), $\sum_{k=0}^{N} P\{B_1 B_2\cdots B_i F_{i+1}|S_k\}$ represents the probability that the i-th selected RV is idle while the preceding i-1 selected RVs are all busy, given that k RVs are in a busy state. According to the multiplication principle of conditional probability, this can be expressed as

$$\begin{aligned} P\{B_1B_2\cdots B_iF_{i+1} \mid S_k\} &= P\{F_{i+1} \mid B_1B_2\cdots B_iS_k\} \\ &\cdot P\{B_i \mid B_1B_2\cdots B_{i-1}S_k\}\cdots P\{B_1 \mid S_k\} \end{aligned} \quad (5)$$

Given that k RVs are busy, the probability that the first selected RV is busy is $P\{B_1|S_k\}$. Clearly, $P\{B_1|S_k\} = k/N$, By analogy, it follows that:

$$P\{B_i|B_1 B_2\cdots B_{i-1}S_k\} = [k-(i-1)]/[N-(i-1)], \quad (6)$$

Similarly,

$$P\{F_{i+1}|B_1 B_2\cdots B_iS_k\} = (N-k)/(N-i) \quad (7)$$

Substituting equations (2)-(7) into equation (1), we obtain

$$P\{B_1B_2\cdots B_iF_{i+1}\} = \sum_{k=i}^{N-1} \frac{k}{N}\cdot\frac{k\text{-}1}{N\text{-}1}\cdots\frac{k\text{-}(i\text{-}1)}{N\text{-}(i\text{-}1)}\cdot\frac{N\text{-}k}{N\text{-}i}P_k,\quad i = 0,1,\cdots,N\text{-}1 \quad (8)$$

Substituting equations (2) and (3) into equation (8), we have

$$\begin{aligned} P\{B_1B_2\cdots B_iF_{i+1}\} &= \sum_{k=i}^{N-1} \frac{k}{N}\cdot\frac{k\text{-}1}{N\text{-}1}\cdots\frac{k\text{-}(i\text{-}1)}{N\text{-}(i\text{-}1)}\cdot\frac{N\text{-}k}{N\text{-}i}(N^k\rho^k/k!)P_0 \\ &= \sum_{k=i}^{N-1} \frac{(N-i-1)!(N-k)}{(k-i)!}\cdot\frac{N^k}{N!}\cdot\rho^{k-i}\cdot\rho^i(1-\rho)[P_0/(1-\rho)] \end{aligned} \quad (9)$$

Then, rewrite equation (9) as

$$P\{B_1B_2\cdots B_iF_{i+1}\} = Q(N,\rho,i)\rho^i(1\text{-}\rho) \quad (10)$$

where

$$Q(N,\rho,i) = \frac{\sum_{k=j}^{N-1}\{(N-i-1)!(N-k)(k-i)!\}(N^k/N!)\rho^{k-i}}{(1-\rho)[\sum_{i=0}^{N-1}(N^i/i!)\rho^i] + N^N\rho^N/N!} \quad (11)$$

Employing the multiplication rule of probability, Equation (5) facilitates the computation of $P\{B_1B_2\cdots B_iF_{i+1}\}$ through the introduction of an isolated term denoted as $Q(N,\rho,i)$. This term functions as a multiplicative factor in the process of determining the probability that the (i+1)-th selected RV is the first available one. It serves to quantify the extent to which the probabilities ρ or 1-ρ deviate from the true conditional probabilities—overestimating or underestimating the actual likelihood of a server being busy or free—under the condition that both N and ρ are known. Thus, the conditional probability $P\{B_1B_2\cdots B_iF_{i+1}\}$ can be expressed as

$$P\{B_1B_2\cdots B_iF_{i+1}\} = P\{F_{i+1} \mid B_1B_2\cdots B_i\}P\{B_i \mid B_1B_2\cdots B_{i-1}\}\cdots P\{B_1\} \tag{12}$$

$$Q(N,\rho,i) = [P\{F_{i+1} \mid B_1B_2\cdots B_i\}/(1\text{-}\rho)]\cdot[P\{B_i \mid B_1B_2\cdots B_{i\text{-}1}\}/\rho]\cdots[P\{B_1\}/\rho] \tag{13}$$

Using the multiplication factor $Q(N,\rho,i)$, the working intensity of a RV can be calculated based on Formula (8). This intensity represents the proportion of time the RV spends in rescue operations—including travel to the incident location and on-site rescue time—relative to the total available time. This metric facilitates the evaluation of various performance indicators of the rescue system.

First, the working intensity of a RV is determined by analyzing the transition between its busy and idle states. For RV *i*, it can only be in one of two states: busy or idle. Let $p\{w_i\} = \rho_i$ denote the probability of the busy state and $p\{\bar{w}_i\} = 1-\rho_i$ the probability of the idle state. The demand rate at which RV i transitions from idle to busy due to an assigned rescue task is denoted as η_i. However, rescue tasks can only be assigned during the RV's idle time. Therefore, the effective rate is expressed as

$$\bar{\eta}_i = \eta_i(1-\rho_i)$$

Based on the state transition diagram of the RVs in Figure 2, the state balance equations are derived: $P\{w_i\}\cdot\mu = P\{\bar{w}_i\}\cdot\eta_i = (1-P\{w_i\})\cdot\eta_i$, After reorganization, it can be written as

$$\rho_i = \eta_i/(\mu+\eta_i) \tag{14}$$

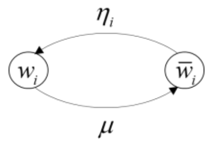


Figure 2. *Busy-Idle State Transition Diagram for a RV.*

Next, we examine the working intensity of RVs assigned to different districts under varying levels of system busyness. The working intensity of a RV depends on two factors: (1) the priority of the rescue location (determined by the proximity of the RV to the incident site, with closer RVs assigned higher priority) and (2) the busy states of other RVs. Specifically, the working intensity of a given RV is the sum of the intensities corresponding to all priority levels and the intensity when all RVs are busy. Based on the above analysis, we derive the following:

$$\bar{\eta}_i = \sum_{j\in G_i^1}\lambda_j P\{\bar{w}_i\} + \sum_{j\in G_i^2}\lambda_j P\{w_{n_{j1}}\bar{w}_i\} + \cdots + \sum_{j\in G_i^N}\lambda_j P\{w_{n_{j1}}w_{n_{j2}}\cdots w_{n_{j(N\text{-}1)}}\bar{w}_i\} + \lambda_D \tag{15}$$

For each emergency response district, the primary unit dispatched (i.e., the first-priority unit) is the RV within the same district. The secondary unit (i.e., the second-priority unit) is the RV from the nearest adjacent district. Subsequent priorities (third, fourth, etc.) follow the same logic. It is important to note that priority assignment is not strictly based on proximity; it can also be determined by administrative district or other practical considerations.

Based on Equation (8), Equation (15) can be rewritten as

$$\bar{\eta}_i = \sum_{j\in G_i^1}\lambda_j(1-\rho_i) + \sum_{j\in G_i^2}\lambda_j Q(N,\rho,1)\rho_{n_{j1}}(1-\rho_i) + \cdots + \sum_{j\in G_i^N}\lambda_j Q(N,\rho,N-1)\rho_{n_{j1}}\rho_{n_{j2}}\cdots\rho_{n_{j(N\text{-}1)}}(1-\rho_i) + \lambda_D \tag{16}$$

Substituting Equations (13), (14) into Equation (16), the following can be derived

$$\rho_i = 1-\mu/[\mu + \sum_{j\in G_i^1}\lambda_j + \sum_{j\in G_i^2}\lambda_j Q(N,\rho,1)\rho_{n_{j1}} + \cdots + \sum_{j\in G_i^N}\lambda_j Q(N,\rho,N\text{-}1)\rho_{n_{j1}}\rho_{n_{j2}}\cdots\rho_{n_{j(N\text{-}1)}} + \lambda_D/(1-\rho_i)] \tag{17}$$

Equation (16) represents the working intensity of an emergency response district, reflecting the saturation level of its rescue operations. This equation can be simplified into a system of linear equations with N unknowns, which is computationally straightforward. By reducing the original system of 2^N unknowns to a system of N unknowns, the simplified model significantly decreases computational complexity.

3. Solution Algorithm

3.1. Working Intensity Solution Algorithm

Equation (17) represents a system of N simultaneous linear equations that can be solved iteratively to determine the workload intensity for each RV. Letting m denote the iteration count, we obtain the following iterative equations.

$$1-\rho_i(m+1)=\mu/[\mu+\sum_{j\in G_i^1}\lambda_j+\sum_{j\in G_i^2}\lambda_j Q(N,\rho,1)\rho_{n_{j1}}(m)+\cdots \\ +\sum_{j\in G_i^N}\lambda_j Q(N,\rho,N\text{-}1)\rho_{n_{j1}}(m)\rho_{n_{j2}}(m)\cdots\rho_{n_{j(N\text{-}1)}}(m)+\lambda_D/(1-\rho_i(m+1))] \tag{18}$$

Here, the parameters λ_j, μ, and λ_D are all known. The value of $Q(N,\rho,i)$ can be determined using Equation (18). The term $\rho_i(m)$, which represents the result after m iterations, is also available. Thus, Equation (18) forms a system of N linear equations with N unknowns. The iterative procedure is described in detail below.

Step 1: Input data: number of RVs N, number of emergency districts K, average demand intensity for accident rescue λ, average service level of RVs μ, distance matrix D from emergency points to districts, and the rescue demand proportion matrix f_j for each district;

Step 2: Calculate necessary values: service intensity of the rescue ρ, average service intensity r= ρ, probabilities of all RVs being idle or busy P_0 and P_N, and construct the matrix multiplication factor matrix Q with assigned values;

Step 3: Determine the priority order based on the distance from emergency points to districts and derive the priority matrix G.

Step 4: Initialize the iteration for the service intensity (ρ_i) of RV i, setting m=0, $\hat{\rho}_i(0)=r, i=1,2,...,N$;

Step 5: Iterate over $\hat{\rho}_i(m)$, where $\hat{\rho}_i(m)$ represents the m-th iteration of ρ_i, $m \leftarrow m+1$, Calculate $\hat{\rho}_i(m)$ according to Equation (16), substituting ρ_j on the right-hand side with $\hat{\rho}_j(m-1)$;

Step 6: Normalize the obtained $\hat{\rho}_i(m)$ such that $N^{-1}\sum_{i=1}^{i=N}\hat{\rho}_i(m)=r$, set $\Gamma\equiv[N^{-1}\sum_{i=1}^{i=N}\hat{\rho}_i(m)/r]^{-1}$, $\hat{\rho}_i(m)\leftarrow\Gamma\hat{\rho}_i(m)$;

Step 7: Use $|\hat{\rho}_i(m)-\hat{\rho}_i(m-1)|\le\varepsilon$ to determine convergence. If convergence is achieved, extract the final iteration result of ρ_i; otherwise, return to Step 5.

The equations that the algorithm solves are a system of N linear equations with N unknowns. Under the condition of fixed rescue district boundaries, if the number of deployed RVs N increases, both the time complexity and space complexity of the algorithm exhibit order $O(N)$ growth. Therefore, the approximate model offers considerable computational advantages for emergency response systems equipped with a large number of RVs.

3.2. Performance Evaluation

In practice, emergency response to highway accidents must ensure that RVs arrive at the incident location within a specified time threshold. Additionally, the working intensity of each RV and the response time for accidents across different emergency districts should be as balanced as possible. Significant disparities in RV working intensity or response times among districts indicate an allocation of rescue resources, necessitating adjustments to the spatial distribution plan of RVs. To address this, this paper proposes several performance metrics for evaluating RV efficiency.

Emergency rescue operations are critical to preserving human life, necessitating a performance evaluation framework that accounts for both efficiency and equity. In terms of efficiency, shorter response times correspond to higher system performance, while an elevated rate of cross-district dispatches indicates that more RVs are required to travel longer distances, thereby reducing operational efficiency. With regard to equity, significant workload imbalance among RVs—where some remain idle for extended periods while others are overburdened—may result in unmet service demands. Similarly, substantial variation in response times across districts implies inconsistent service delivery, with prolonged delays in some areas adversely affecting survival rates. Furthermore, pronounced disparities in cross-district dispatch ratios suggest uneven resource allocation: some districts benefit from prompt local response, whereas others depend heavily on RVs dispatched from farther away, increasing response times. Thus, a comprehensive evaluation of emergency rescue performance requires quantifying key metrics such as response time, workload distribution, and cross-district dispatch proportion. The computational methods for these performance indicators are provided in Table 2.

Table 2. *Performance Metrics and Formulas.*

Performance metrics	Meaning of performance metrics	Formula
Response time	The duration from the emergency station receiving a rescue call to the RV arriving at the incident location.	$\overline{T}=\sum_{i=1}^{N}\sum_{j=1}^{J}\rho_{ij}^{[1]}t_{ij}+P_N\overline{T}_Q$

Performance metrics	Meaning of performance metrics	Formula
Proportion of cross-district emergency dispatches	Ratio of cross-district dispatch probability of RV *i* to total dispatch probability	$Fl_i = \sum_{j \neq h} \rho_{ij} \Big/ \sum_{j=1}^{J} \rho_{ij}$
Proportion of rescue from external district	Ratio of cross-district dispatch probability to total rescue probability of the district *j*	$FO_j = \sum_{i \notin j} \rho_{ij} \Big/ \sum_{i=1}^{N} \rho_{ij}$
Total cross-district rescue	Total cross-district dispatch probability of RV *i* (from district *h*) to district *j*.	$FT = \sum_{i=1}^{N} \sum_{j \neq h} \rho_{ij}$
District response time	Mean response time to district *j*	$\overline{T}_j = \sum_{i=1}^{N} \rho_{ij}^{[1]} t_{ij} \Big/ \sum_{i=1}^{N} \rho_{ij}^{[1]} \cdot (1 - P_N) + \sum_{l=1}^{J} f_l \tau_{lj} P_N$
RV response time	Response time of RV *i*	$\overline{Tu}_i = \left(\sum_{j=1}^{J} \rho_{ij}^{[1]} t_{ij} + \overline{T}_Q P_N / N \right) \Big/ \left(\sum_{j=1}^{J} \rho_{ij}^{[1]} + P_N / N \right)$

It is noteworthy that the performance indicators are interdependent. A higher rate of cross-district dispatches indicates that most emergency calls require longer travel distances, which increases travel time, extends response times, and raises workload—ultimately leading to longer waiting times. Therefore, reducing the cross-dispatch rate will also lower workload, response times, and waiting times. Disparities in these indicators offer valuable insights for system optimization, such as improving RV deployment and redefining district boundaries. This involves determining three key variables: the number of RVs, their locations, and the partitioning of districts. In practice, however, increasing the number of RVs is often infeasible due to a shortage of emergency professionals. Station location planning is also complex, requiring the consideration of multiple practical factors. Similarly, redistricting must account for administrative boundaries and geographical constraints such as rivers and lakes. Purely theoretical models that ignore these factors often yield inapplicable results. Based on our analysis and the performance metrics derived from the model, districts with significantly higher workloads or cross-district dispatch rates likely have either insufficient RVs or excessively large territories. By considering these interdependencies alongside administrative and geographical constraints, managers can adjust RV allocations and district boundaries. The model can then reassess the revised plan's performance to evaluate whether system efficiency and equity have improved.

4. Case Study

4.1. The Data and Study Site

District	1	2	3	4
1	0.1875	0.25	—	—
2	0.25	0.1875	—	—
3	—	—	0.1875	0.25
4	—	—	0.25	0.1875

Figure 3. *Travel Times (hours) Within and Between Rescue District.*

The Sutong Bridge of G15 Shenhai highway's section spans 25 km, with the bridge itself covering 8.2 km and handling a peak daily traffic volume of 136,000 RVs. This section is a high-frequency accident district in China. According to traffic police statistics, during holidays such as National Day and Spring Festival, the average accident rates are λ = 1.02 incidents per hour in the Jiangsu-to-Shanghai direction and λ = 1.143 incidents per hour in the Shanghai-to-Jiangsu direction. The average duration for RV dispatch and on-site operations is 30 minutes. Currently, four emergency response points are established along this section. As illustrated in Figure 3, this study proposes four candidate locations (a, b,..., k) for emergency points, along with the zoning of rescue districts and the average travel times within and across these districts. Two priority levels are defined: a single priority level, where RVs are restricted to their own districts, and a dual priority level, where RVs can respond to incidents in adjacent districts.

4.2. Analysis of Rescue Performance

This section employs the performance metrics calculation method detailed in Section 2.1 to evaluate the rescue system along the G15 Shenhai Highway Sutong Bridge section. Figures 4 and 5 display the average working intensity of rescue vehicles (RVs) across different districts under single-priority and dual-priority levels, respectively, with varying RV allocation quantities. A comparison of the two figures reveals that, overall, the average working intensity under the dual-priority scheme is higher than under the single-priority scheme across districts. In both scenarios, District 2 shows the highest average working intensity. Additionally, when only one RV is assigned per district, working intensity is highly uneven among districts. As the number of RVs allocated increases, the intensity becomes progressively more balanced. Notably, as shown in Figures 4 and 5, the curves for 4 and 5 RVs nearly overlap, indicating diminishing marginal returns from adding additional RVs, which in turn promotes a more balanced distribution of working intensity. Moreover, the average working intensity of RVs in Districts 1 and 2 is consistently higher than that in Districts 3 and 4. This suggests that greater RV deployment is warranted in the Jiangsu-to-Shanghai direction compared to the opposite direction.

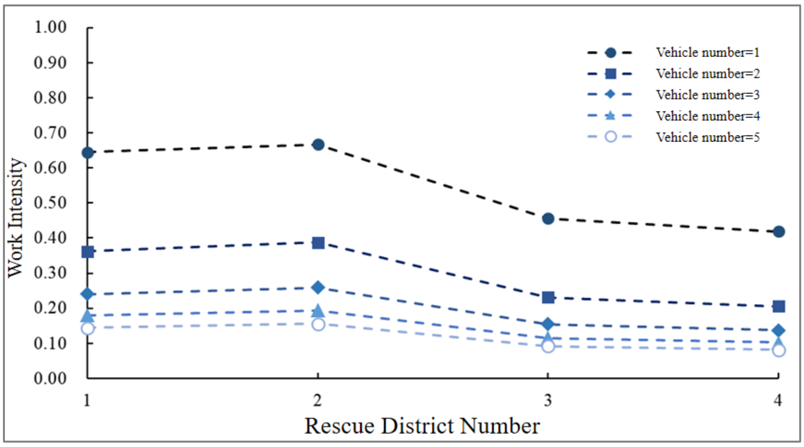


Figure 4. *Statistical Chart of Working Intensity for In-District RVs.*

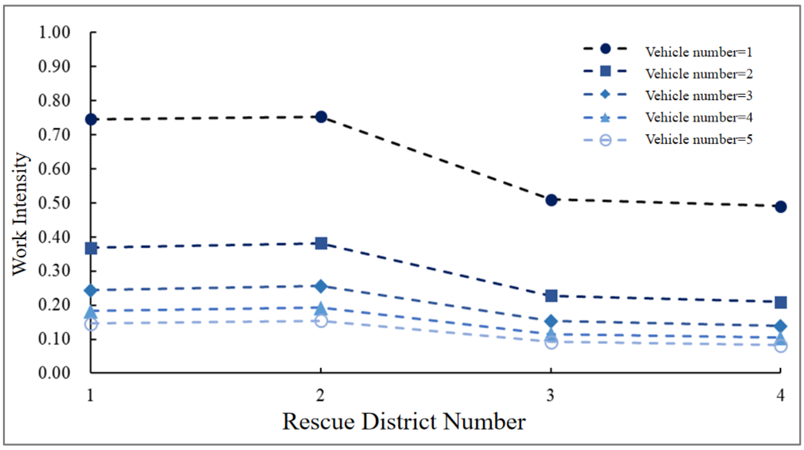


Figure 5. Statistical Chart of Working Intensity for Cross-District RVs.

Table 3. Cross-District Rescue Ratio by RV Allocation.

Number of RVs Allocated per District	Jiangsu to Shanghai	Shanghai to Jiangsu
1	0.429	0.334
2	0.259	0.170
3	0.191	0.123
4	0.153	0.097
5	0.128	0.079

Due to the high demand for accident rescue services on highways, emergency districts often face situations where no RVs are available for dispatch, necessitating the assignment of RVs from other districts. As shown in Table 3, as the number of allocated RVs in each district increases, the overall cross-district rescue ratio decreases, significantly reducing the need to dispatch RVs from other districts for accidents occurring within a given district.

Figure 6 illustrates the cross-district rescue ratios for each district under single- and dual-priority levels. As the priority level increases, the cross-district rescue ratio also rises. When only one RV is assigned per district, the average ratio is generally high across all districts except District 3, with District 1 recording the highest ratio. This indicates a significant likelihood that RVs from District 1 are dispatched to District 2. As the number of RVs per district increases to two, the average cross-district rescue ratio declines, suggesting a more sufficient allocation of vehicles within each district. If a cross-district rescue strategy is implemented for the G15 Shenhai Highway Sutong Bridge section, the cross-district rescue ratio is expected to range between 10% and 45%. Therefore, increasing the number of RVs would effectively reduce the frequency of cross-district dispatches.

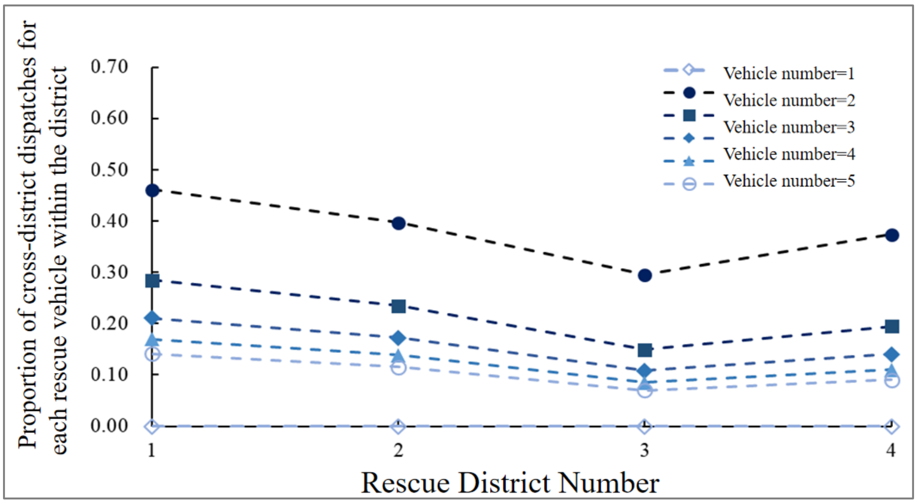


Figure 6. Cross-District Rescue Ratios for RVs.

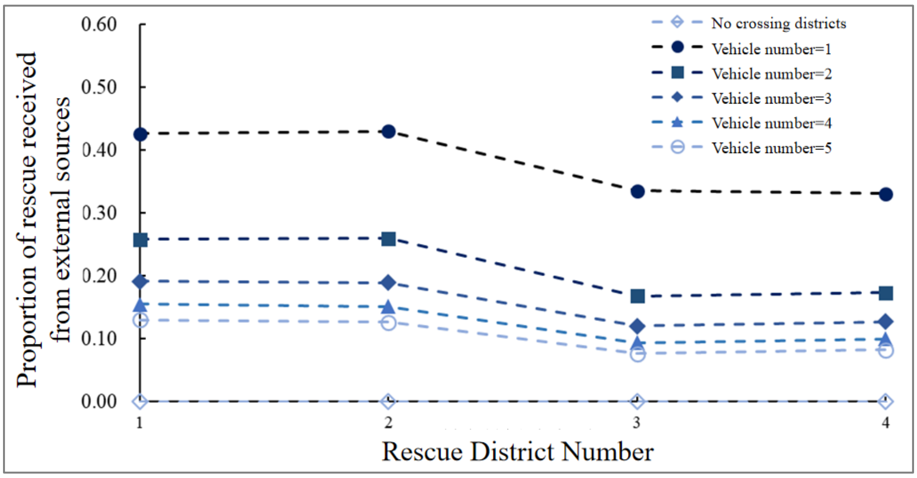


Figure 7*. External Rescue Proportions for Emergency Districts.*

The demand for accident rescue services in an emergency district is not only met by its own RVs but also relies on cross-district assistance from RVs outside the district. As shown in Figure 7, when only one RV is allocated per district, Districts 1 and 2 exhibit higher cross-district rescue ratios, which increase with the priority level. At a priority level of 2, District 2 receives relatively less cross-district assistance, while, as indicated in Figure 6, District 4's RVs are frequently dispatched to other districts. Combining these two metrics reveals that District 2 has an insufficient allocation of RVs, whereas District 4 has a surplus. Given a fixed total number of RVs, one RV could be reallocated from District 4 to District 2 to address this imbalance. Furthermore, as the number of allocated RVs per district increases, the proportion of external rescue assistance gradually decreases. Specifically, when two RVs are allocated per district, the external rescue ratio reaches a relatively low level.

Response time is a critical performance metric for highway emergency rescue systems. As indicated in Figure 8, rescue response time increases with the number of priority levels. This suggests that cross-district sharing of RVs elevates the cross-district rescue ratio, leading to longer rescue distances and, consequently, prolonged response times for highway accident rescues. Moreover, when only one RV is allocated per district, the variation in response times across different priority levels is pronounced. However, this variation decreases as the number of RVs allocated increases. In this case study, the average response time of the rescue system for the G15 Sutong Bridge section is approximately 10 minutes, which falls within an acceptable range.

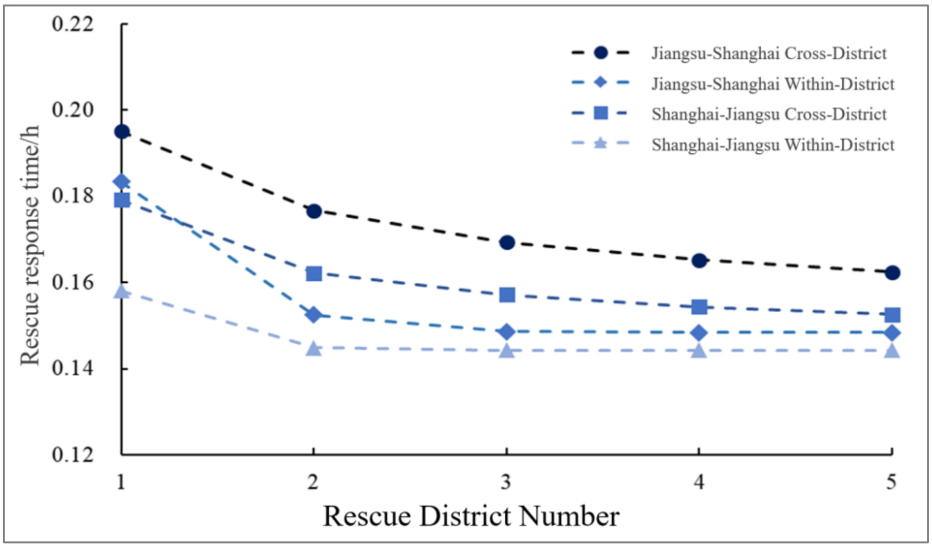


Figure 8*. Rescue Response Times.*

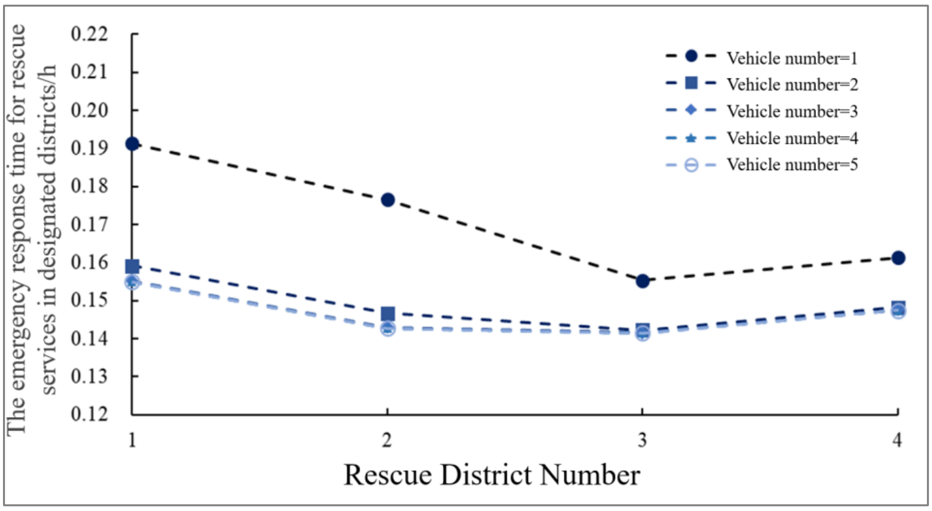


Figure 9. Response Times for Emergency Districts Under Priority Level 1.

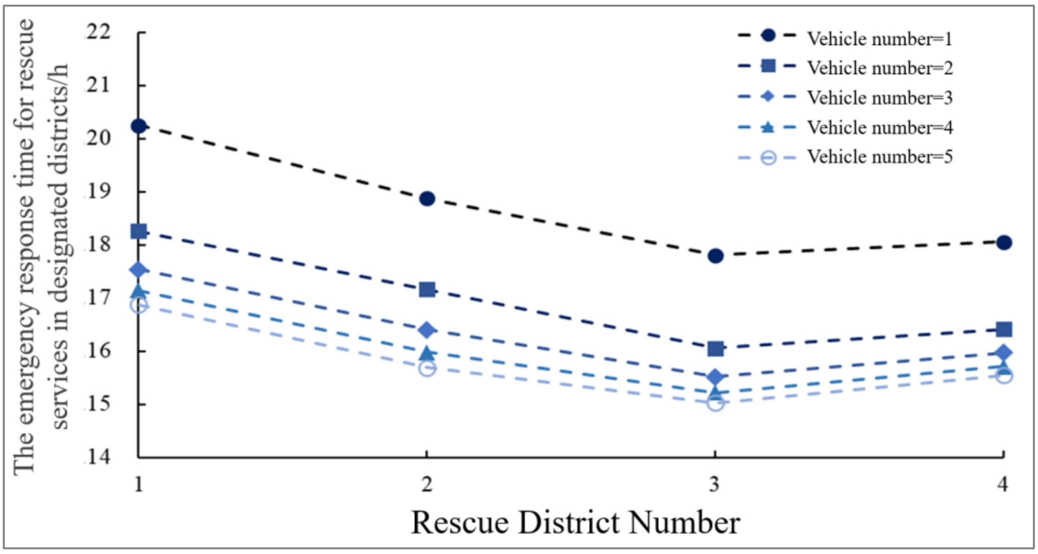


Figure 10. Response Times for Emergency Districts Under Priority Level 2.

Figures 9 and 10 present the response times for emergency districts under two priority levels. It can be observed that as the number of priority levels increases, the response time for rescue operations also lengthens, as RVs are shared across more districts. When 2 to 5 RVs are allocated per district, the response times across districts become nearly equal and show a noticeable reduction, indicating that accidents in each district can be addressed promptly. Furthermore, allocating 2 RVs per district achieves a relatively efficient rescue state.

Figures 11 and 12 illustrate the average rescue response times for RVs within each district under two priority levels. It is evident that as the number of priority levels increases, RVs are required to serve more districts, leading to longer average response times. Overall, the RVs in Districts 1 and 4 exhibit the longest average response times, primarily due to their frequent cross-district assistance to Districts 2 and 3, resulting in a higher total number of rescue operations. When 2 to 5 RVs are allocated per district, the average response times within each district become nearly equal, indicating that accidents are promptly addressed. Moreover, allocating 2 RVs per district achieves an effective rescue state.

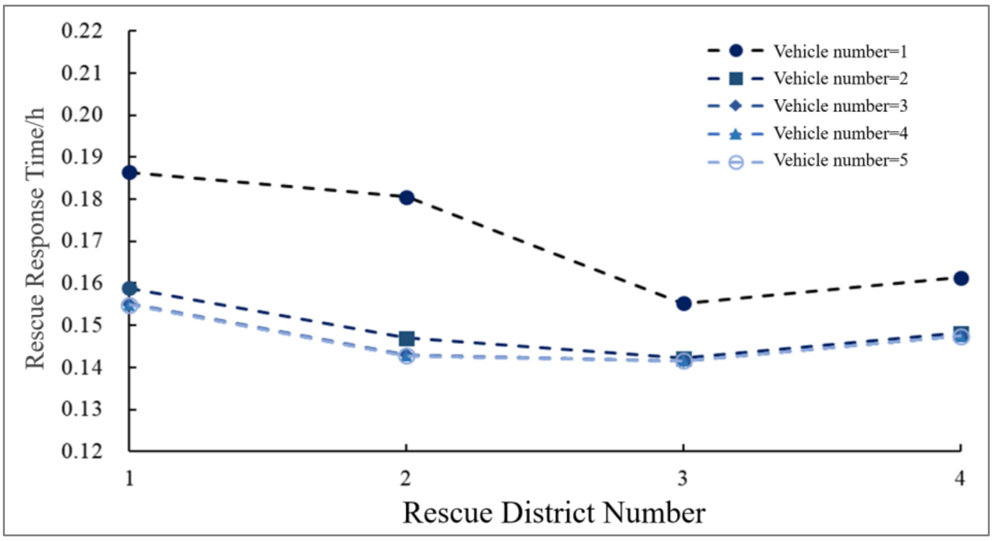


Figure 11. Rescue Response Times for RVs Under Priority Level 1.

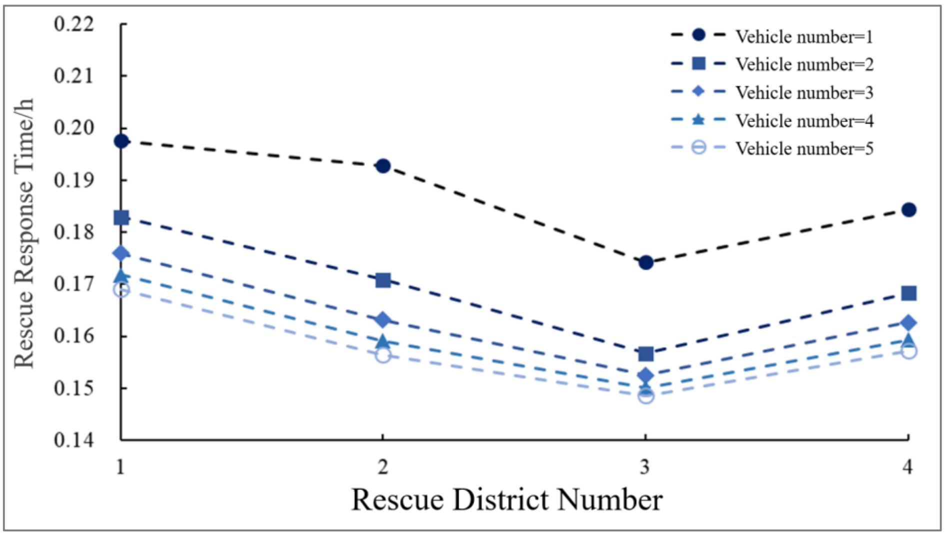
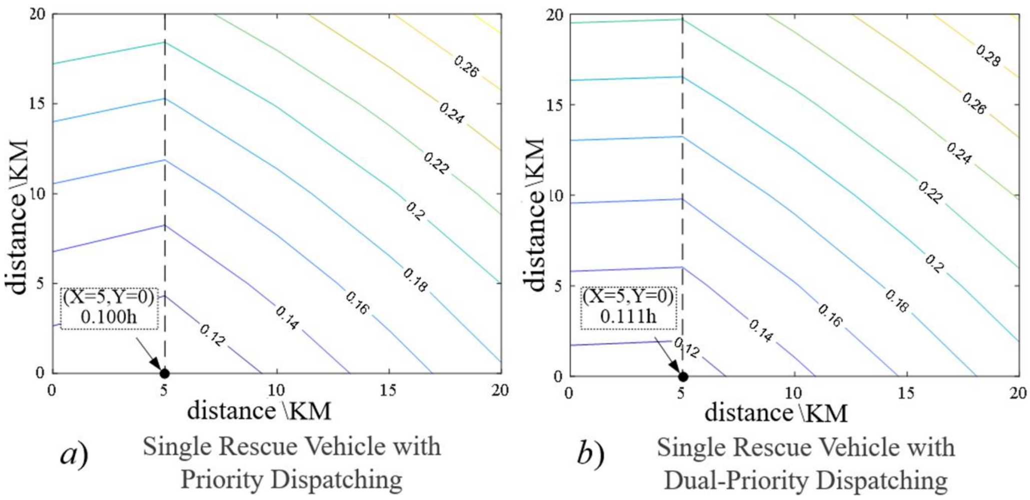


Figure 12. Rescue Response Times for RVs Under Priority Level 2.



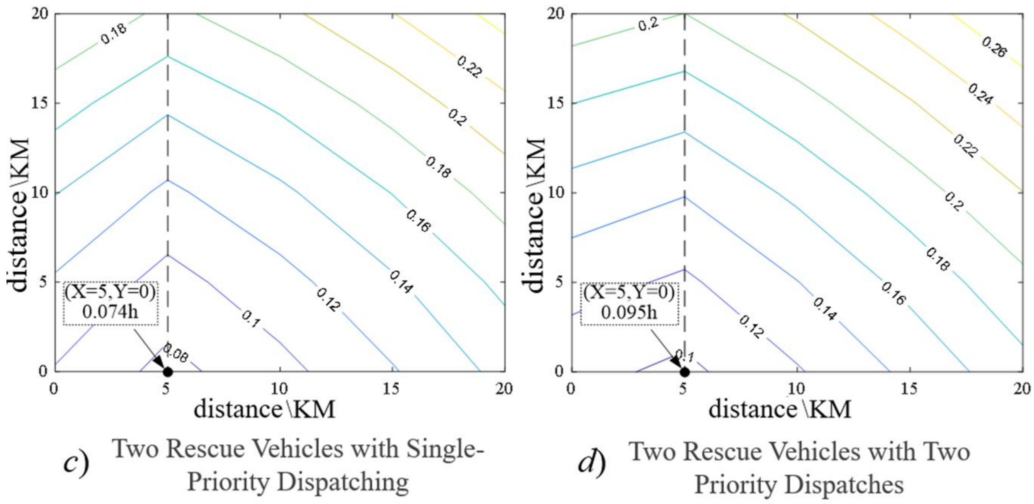


Figure 13. *Rescue Response Times for the Jiangsu-to-Shanghai Direction.*

Figure 13 presents contour maps of response times for Districts 1 and 2 in the Jiangsu-to-Shanghai direction, considering different numbers of RVs and priority levels. Based on the four subplots in Figure 13, the optimal locations for rescue performance are identified as points d and g (see Figure 3). Comparing the four subplots in Figure 13, it is evident that increasing the number of RVs reduces the overall response times by 26.0% and 14.4%, respectively. This demonstrates that under the optimal location scheme, augmenting the number of RVs effectively decreases response times, thereby providing robust support for highway emergency rescue operations.

5. Conclusion

This study models the rescue dispatch process for highway emergency vehicles by establishing a state-space representation and probabilistic characterization of rescue fleets. Based on stochastic birth–death process theory, a balance equation for rescue states is formulated, and an approximate model with a solution algorithm for rescue vehicle (RV) working intensity is developed. Furthermore, calculation methods are proposed for key performance metrics, including the cross-district rescue ratio, the proportion of external rescue assistance received by each district, and rescue response times. Using the G15 Sutong Bridge section as a case study, the emergency response performance of this highway segment is evaluated. With the dual objectives of balancing RV workload and minimizing response times, optimization schemes for RV allocation and district partitioning are proposed without increasing the total number of RVs. The findings of this study can be generalized to optimize RV deployment on various highway networks, with the aim of shortening response times and balancing operational intensity, demonstrating broad practical applicability.

However, this study does not account for the impact of post-accident traffic congestion, which may extend RV travel times. Future research could further investigate this factor to enhance the model's realism and applicability.

Abbreviations

RVs	Rescue Vehicles
HRS	Highway Rescue System
HQM	Hypercube Queuing Model

Conflicts of Interest

The authors declare that there is no conflicts of interest.

References

[1] Iannoni, A. P., & Morabito, R. (2007). A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. Transp. Res. Part E Logist. Transp. Rev., 43(6), 755–771.

[2] Nakata, H. (2012, January 6). Domestic robots failed to ride to rescue after No. 1 plant blew. The Japan Times. Retrieved November 18, 2018, from https://www.japantimes.co.jp/news/2012/01/06/national/domestic-robots-failed-to-ride-to-rescue-after-no-1-plant-blew

[3] Andradóttir, S., Ayhan, H., & Down, D. G. (2003). Dynamic server allocation for queueing networks with flexible servers. Oper. Res., 51(6), 952–968.

[4] Beojone, C. V., Máximo de Souza, R., & Iannoni, A. P. (2021). An efficient exact hypercube model with fully dedicated servers. Transp. Sci., 55(1), 222–237.

[5] Larson, R. C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. Comput. Oper. Res., 1(1), 67–95.

[6] Larson, R. C., & Odoni, A. R. (1981). Urban operations research. Prentice-Hall.

[7] Galvão, R. D., & Morabito, R. (2008). Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. Int. Trans. Oper. Res., 15, 525–549.

[8] Larson, R. C. (1975). Approximating the performance of urban emergency service systems. Oper. Res., 23(5), 845–868.

[9] Jarvis, J. P. (1985). Approximating the equilibrium behavior of multi-server loss systems. Manage. Sci., 31(2), 235–239.

[10] Larson, R. C., & McKnew, M. A. (1982). Police patrol-initiated activities within a system queueing model. Manage. Sci., 28(7), 759–774.

[11] Atkinson, J. B., Kovalenko, I. N., Kuznetsov, N. Y., & Mikhalevich, K. V. (2008). A hypercube queueing loss model with customer-dependent service rates. Eur. J. Oper. Res., 191(1), 223–239.

[12] Iannoni, A. P., Morabito, R., & Saydam, C. (2009). An optimization approach for ambulance location and the districting of the response segments on highways. Eur. J. Oper. Res., 195(2), 528–542.

[13] Budge, S., Ingolfsson, A., & Erkut, E. (2009). Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. Oper. Res., 57(1), 251–255.

[14] Morabito, R., Chiyoshi, F., & Galvão, R. D. (2008). Non-homogeneous servers in emergency medical systems: Practical applications using the hypercube queueing model. Socioecon. Plann. Sci., 42(4), 255–270.

[15] Iannoni, A. P., Morabito, R., & Saydam, C. (2011). Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. Socioecon. Plann. Sci., 45(3), 105–117.

[16] Takeda, R. A., Widmer, J. A., & Morabito, R. (2007). Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. Comput. Oper. Res., 34(3), 727–741.

[17] Khoshgehbari, F., & Al-e, S. M. J. M. (2023). Ambulance location routing problem considering all sources of uncertainty: Progressive estimating algorithm. Comput. Oper. Res., 160, 106400.

[18] Frichi, Y., Aboueljinane, L., & Jawab, F. (2025). Ambulance location and relocation under budget constraints: Investigating coverage-maximization models and ambulance sharing to improve emergency medical services performance. Health Care Manage. Sci., 1–24.

[19] Grot, M., Nagel, L., Becker, T., Fiebrandt, P. M., & Werners, B. (2022). Fairness or efficiency-managing this conflict in emergency medical services location planning. Comput. Ind. Eng., 173, 108664.

[20] Gago-Carro, I., Aldasoro, U., Lee, D. J., & Merino, M. (2025). Hierarchical compromise optimization of ambulance locations under stochastic travel times. Comput. Oper. Res., 107208.

[21] Rautenstrauss, M., Martin, L., & Minner, S. (2023). Ambulance dispatching during a pandemic: Tradeoffs of categorizing patients and allocating ambulances. Eur. J. Oper. Res., 304(1), 239–254.

[22] Mandelbaum, A., & Stolyar, A. L. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. Oper. Res., 52(6), 836–855.

[23] Tsai, Y. C., & Argon, N. T. (2008). Dynamic server assignment policies for assembly-type queues with flexible servers. Nav. Res. Logist., 55(3), 234–251.

[24] Iannoni, A. P., Chiyoshi, F., & Morabito, R. (2015). A spatially distributed queuing model considering dispatching policies with server reservation. Transp. Res. Part E Logist. Transp. Rev., 75, 49–66.

[25] Wang, J., Liu, H., An, S., & Cui, N. (2016). A new partial coverage locating model for cooperative fire services. Inf. Sci., 373, 527–538.

[26] Rodrigues, L. F., Morabito, R., Chiyoshi, F. Y., Iannoni, A. P., & Saydam, C. (2017). Towards hypercube queuing models for dispatch policies with priority in queue and partial backup. Comput. Oper. Res., 84, 92–105.

[27] Liu, H., Yin, H., Zhou, Y., & Li, M. (2021). Cooperative hypercube queuing model for emergency service systems. J. Adv. Transp., 2021(1), 6653573.

[28] Gu, Y., Jiang, L., Liu, H., Zhang, X., Wei, S., & Qi, Q. (2025). Advancing road traffic safety through stochastic modeling: An innovative assessment of emergency road rescue systems with the mn hypercube queuing model. J. Adv. Transp., 2025(1), 8145358.

[29] Geroliminis, N., Karlaftis, M. G., & Skabardonis, A. (2009). A spatial queuing model for the emergency vehicle districting and location problem. Transp. Res. Part B Methodol., 43(7), 798–811.

[30] Geroliminis, N., Kepaptsoglou, K., & Karlaftis, M. G. (2011). A hybrid hypercube–genetic algorithm approach for deploying many emergency response mobile units in an urban network. Eur. J. Oper. Res., 210(2), 287–300.

[31] Zhen, L., Wu, J., Chen, F., & Wang, S. (2024). Traffic emergency vehicle deployment and dispatch under uncertainty. Transportation Research Part E: Logistics and Transportation Review, 183, 103449.