CIRRELT

CIRRELT-2026-14

# Integrated Service Network Design and Pricing of Regional Air Mobility with Passenger Travel Mode Choice

**Zhongyi Jin**
**Okan Arslan**
**Kam K.H. Ng**
**Chenliang Zhang**
**Lingxiao Wu**

**June 2026**

**Bureau de Montréal**
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
**Tél** : 1-514-343-7575
**Télécopie** : 1-514-343-7121

**Bureau de Québec**
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
**Tél** : 1-418-656-2073
**Télécopie** : 1-418-656-2624

# Integrated Service Network Design and Pricing of Regional Air Mobility with Passenger Travel Mode Choice

**Zhongyi Jin[1], Okan Arslan[2-3], Kam K.H. Ng[1*], Chenliang Zhang[1], Lingxiao Wu[1]**

[1] Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR, China.

[2] Department of Decision Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, H3T 2A7, QC, Canada.

[3] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT).

**Abstract.** We consider a Regional Air Mobility (RAM) Service Network Design and Pricing (RAM-SNDP) problem, in which an operator offering RAM services determines the service network, aircraft fleet size, and service pricing. While these decisions aim to attract travelers, the incorporation of the new RAM service may inadvertently cannibalize demand for the operator's existing expressway taxi services, or travelers may switch to competing modes such as trains or private cars. To solve the RAM-SNDP problem, we propose a bilevel programming model. At the upper level, the operator makes strategic decisions for two types of RAM services, namely regular air travel ride-sharing and priority dedicated service, by maximizing revenue subject to a budget constraint. At the lower level, passengers choose the travel alternative that minimizes their generalized cost among all available options. We reformulate the problem into an equivalent single-level mixed-integer linear program. To solve it efficiently, we derive exact best response sets for passengers, introduce variable reduction techniques, and develop valid inequalities. We test our solution approach on instances of different sizes and build a case study using real-world data from New York taxi operations. The results show that incorporating passenger mode choice yields significant benefits. Finally, we conduct a sensitivity analysis to examine how key model parameters affect the outcomes.

**Keywords**: regional air mobility (RAM), bilevel programming, service network design and pricing, travel mode choice.

**Acknowledgements.** This work was supported by grants from the Research Grants Council, the Hong Kong Government (Grant No. PolyU15223425, F-PolyU501/25).



* Corresponding author: kam.kh.ng@polyu.edu.hk

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2026

## 1. Introduction

Regional Air Mobility (RAM) is emerging as a transformative paradigm for intercity and metropolitan transportation (Justin et al. 2022). This new mode of travel is largely enabled by advancements in electric Vertical Takeoff and Landing (eVTOL) aircraft, which offer the potential for fast, frequent, and scalable air services over short- to medium-haul distances, such as connections between districts within major metropolitan regions or between neighboring cities (Sun et al. 2018, Straubinger et al. 2021, Wang et al. 2022, Jin et al. 2024). As a result, RAM has rapidly gained growing commercial interest. For instance, Uber has announced plans to operate all-electric air taxis directly within its app, with the first commercial launch expected in Dubai in 2026 (Uber Newsroom 2026). Recently, Joby Aviation, in partnership with the Port Authority and Blade Air Mobility, successfully completed the first point-to-point eVTOL demonstration flights in New York City, connecting JFK Airport to Manhattan heliports in under ten minutes (Joby 2026).

Introducing this new mode of transportation requires substantial capital investment. High upfront costs are incurred for building vertiports, deploying eVTOL fleets, and developing new aerial routes. Specifically, vertiports play a critical role in the RAM system, serving as the infrastructure for eVTOL takeoffs and landings, passenger boarding and alighting, as well as eVTOL maintenance and charging (Jin et al. 2024, Li et al. 2025). At the same time, aerial routes need to be carefully designed to ensure safety and efficiency (Weng et al. 2026). This infrastructure is essential for an efficient RAM system.

For a multimodal operator such as Uber, launching such a service can have complex implications for its existing expressway taxi offerings, especially when facing competition from opt-out alternatives. The addition of RAM may either cannibalize the operator's traditional business or create a win-win situation, depending on how passenger travel behavior adjusts to the new travel mode. Specifically, an attractive RAM design and pricing strategies can draw passengers away from the opt-out option, generating additional revenue. Conversely, some passengers who would have used the expressway taxi may instead switch to the opt-out alternative, leading to revenue loss. Additionally, real-world platforms such as Uber (UberX vs. Uber Black) and Didi (Express vs. Premier) already demonstrate the viability of tiered service offerings. Applying this concept to RAM, we consider both regular and priority service levels. These two types differ in operational modes: priority service offers dedicated eVTOL aircraft without ride-sharing to ensure a premium experience, whereas regular service operates on a ride-sharing basis. In this paper, we study how a multimodal operator designs an efficient RAM service system to maximize revenue under a budget constraint while responding to passengers who minimize their generalized cost when choosing among travel modes.

We contribute to the current research on RAM service system design from multiple aspects:

- We formulate the RAM Service Network Design and Pricing (RAM-SNDP) problem from the perspective of a multimodal operator.
- We model the RAM-SNDP problem as a bilevel program, where the leader represents the operator's integrated decision-making across vertiport location, aerial route opening, eVTOL fleet deployment, and pricing for both priority and regular RAM services to maximize total revenue. The follower problem captures passengers' mode choices, each minimizing their generalized cost independently.
- Methodologically, we employ duality theory to transform the bilevel formulation into an equivalent mixed-integer linear programming model. We also derive explicit expressions for passengers' best responses across different mode choice alternatives and develop solution acceleration techniques, including mode-specific variable reduction techniques and several families of valid inequalities.
- We conduct extensive experiments to evaluate the efficiency of our solution procedure. Using a real-world dataset based on New York taxi operations, we investigate the benefits of the proposed bilevel program and examine how system revenue and passenger choice are influenced by key factors such as expressway price, inconvenience cost, and passenger valuation of time.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature. The problem is formulated as a bilevel program in Section 3, and the solution methodology is developed in Section 4. An extensive numerical study is presented in Section 5, followed by concluding remarks and future research directions in Section 6.

## 2. Literature Review

As defined by the Federal Aviation Administration, the RAM constitutes a pivotal component of Advanced Air Mobility, providing passenger and cargo transportation services within metropolitan and regional networks. Transitioning from conceptual frameworks to a functional and efficient RAM system entails system design and operational planning. This section reviews the literature relevant to the integrated design and pricing of RAM services. Section 2.1 examines the vertiport location and RAM service network design problem, and Section 2.2 summarizes models of passenger mode choice in the field of transportation. Finally, Section 2.3 reviews the integrated location and pricing problem.

### 2.1. Vertiport Location Problem and RAM Service Network Design

The vertiport serves as the essential ground infrastructure for takeoff, landing, and passenger transfer in the RAM system. Consequently, the Vertiport Location Problem (VLP) has emerged as a critical and active research stream within RAM service network design. Existing research can be

broadly categorized into two complementary approaches. One strand of literature focuses on data-driven location analysis using clustering or p-median algorithms (Lim and Hwang 2019, Rajendran and Zack 2019, Willey and Salmon. 2021). Another stream employs mathematical optimization techniques to solve the VLP. For instance, Chen et al. (2022) proposed a vertiport selection model that explicitly incorporates airspace and geographical restrictions by excluding prohibited grid cells from candidate hub locations, and developed a grid-based variable neighborhood search heuristic to solve the resulting optimization problem.

The vertiport location is not merely a point on a map but a pivotal decision that constrains or enables subsequent network and operational planning (Jin et al. 2024). This characteristic shares key similarities with the interdependencies observed in the integrated planning of electric vehicle infrastructure, logistics hub networks, and bike-sharing station systems (Çelebi et al. 2018, Baloch and Gzara 2020, Chen and Liu 2023). In these domains, facility location cannot be decoupled from decisions regarding fleet deployment, pricing, and service operations. This fact has motivated a growing body of research aimed at integrating strategic facility location decisions with tactical and operational aspects of RAM service network design. For example, Wang et al. (2022) optimizes the locations and capacities of vertiports in urban air mobility systems, explicitly capturing the interdependencies between strategic vertiport deployment and tactical operations under demand uncertainty. Similarly, Jin et al. (2024) proposed an integrated model that concurrently addresses vertiport scale, eVTOL fleet sizing, and vehicle repositioning operations. The model developed in this paper is situated within this integrated paradigm. It addresses a problem closely related to, yet extending beyond, the classic VLP by comprehensively considering the strategic placement of vertiports, the establishment of the air network, the deployment of the eVTOL fleet, service pricing, and passenger mode choice.

### 2.2. Passenger Travel Mode Choice Modeling

Incorporating passenger travel choice is essential for developing transportation infrastructure models that are both scientifically sound and policy-relevant (Zhang et al. 2004). A well-established stream of research utilizes discrete choice models to simulate traveler decision-making within multimodal networks characterized by multiple competing attributes (Park et al. 2025, Qi et al. 2026).

Building upon this behavioral foundation, a significant body of work has embedded passenger mode choice models within optimization frameworks for transportation system planning. This integrated approach is exemplified in domains such as airline scheduling, public transit planning, and railway system design, where demand-responsive models have been employed to improve service attractiveness (Wei et al. 2020, 2022, Hartleb et al. 2023). The coupled framework ensures that demand responds endogenously to supply-side planning variables. However, in the emerging field of

RAM transportation, such integrated studies remain relatively scarce. One of the related literature in this area is by Rath and Chow (2022), who formulated the vertiport location problem as a single-allocation p-hub median model incorporating exogenous passenger travel mode choice behavior. This research gap highlights a critical opportunity to advance toward a more holistic planning framework. In response, this paper develops an integrated optimization model that synthesizes the interdependent dimensions of RAM service network design under endogenous passenger choice.

While random-utility discrete choice models are widely used in transportation mode-choice analysis (Ye et al. 2017, Zhao and Feng 2025), we do not adopt an unconstrained probabilistic demand model in this paper. The reason is that RAM services are capacity-limited, so standard choice probabilities reflect passengers' willingness to use a service but do not by themselves guarantee feasible seat allocations when eVTOL capacity is scarce. For the prescriptive design-and-pricing problem studied here, we therefore represent passenger response through a generalized-cost-based lower-level problem with explicit availability and capacity constraints. This formulation preserves the central behavioral trade-offs among fare, travel time, and inconvenience, while remaining compatible with an exact bilevel reformulation and tractable MILP solution procedures. We thus view the deterministic generalized-cost model as a first-order benchmark for strategic RAM planning in an emerging market with limited revealed-preference data. Extending the framework to stochastic or random-utility choice models with richer scarcity effects is an important direction for future research.

### 2.3. Integrated Network Design and Pricing Problem

In tactical level planning, service network design and pricing have traditionally been addressed independently. In contrast, service pricing research has increasingly adopted bilevel programming as a suitable framework for modeling hierarchical decision-making structures. Bilevel programs mathematically formalize the Stackelberg game concept, wherein a subset of variables is constrained to be an optimal solution of a second optimization problem, which itself is parameterized by the remaining variables from the upper level (Bracken and McGill 1973, Arslan et al. 2018).

In the context of transportation networks, early applications of bilevel pricing models can be traced to Labbé et al. (1998). Their work considers a setting in which a highway authority, at the upper level, sets tolls on a subset of links to maximize revenue, while travelers, at the lower level, choose minimal-cost paths in response to the imposed tolls and other associated costs. Building upon this research direction, numerous studies in transportation and logistics systems operations planning have subsequently integrated service network design and pricing into combined optimization frameworks. For example, Brotcorne et al. (2008) formulated a joint service design and pricing problem in the airline or telecommunication industries as a mixed-integer bilevel program and developed a novel Lagrangian relaxation algorithm capable of solving large-scale instances. Similarly, in

the field of intermodal freight transportation, research has addressed the intertwined questions of service network design and pricing, incorporating considerations of service quality from the perspective of a transport operator, and adopted bilevel programming approaches for such problems. In the context of logistics network design, Li et al. (2023) proposed a bilevel model that captures the strategic interaction between a biofuel producer and farmers in a dynamic decision-making setting.

In contrast to existing studies on integrated network design and pricing, our proposed framework provides a comprehensive model for designing a market-responsive RAM network. It endogenizes both pricing and passenger mode choice as variables, allowing their mutual influence to shape a dynamic market equilibrium, as pricing decisions directly influence passengers' travel choices in the transportation system. Overall, our work captures passenger travel choice within a bilevel RAM network design and pricing model in this endogenous manner.

## 3. Problem Statement and Mathematical Model

This section describes the sequence of decisions in Section 3.1 and introduces the RAM-SNDP problem in Section 3.2. Then, a bilevel program for the RAM-SNDP problem is presented in Section 3.3, where the multimodal service operator acts as the leader and passengers independently make decisions as followers.

### 3.1. Sequence of Decisions

Launching large-scale RAM services requires three large upfront investments: (1) vertiport construction, (2) eVTOL fleet deployment, and (3) air route opening. In practice, the operator first raises capital, which represents a fixed initial investment budget. It then uses this budget to cover the three upfront costs, which are largely sunk costs independent of the passenger volume captured. Investment occurs upfront, returns accrue over time, and break-even is typically reached only after several years. Hence, we focus on a revenue maximization problem under an investment budget, particularly during the market entry phase of the RAM service.

We now introduce the sequence of decisions in this Stackelberg game. The multimodal service operator acts as the leader, and passengers are the followers. At the upper level, the operator makes strategic decisions to maximize total revenue subject to an investment budget. These decisions include vertiport location selection, aerial route activation, and eVTOL fleet deployment. The operator also posts route prices and controls the number of available seats on each open aerial route, subject to fleet capacity. Although the operator does not directly assign passengers to seats, it can effectively determine which passengers ultimately use the RAM service through its pricing and seat allocation decisions. By adjusting the posted prices for regular and priority services, the operator influences the generalized cost faced by each passenger, thereby shaping their mode choices.

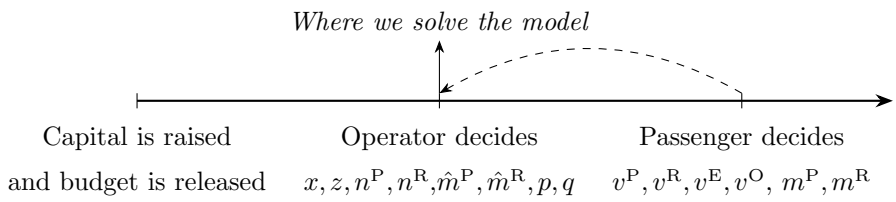



**Figure 1 Sequence of decisions**

Once the infrastructure is in place and prices are posted, passengers arrive and make their mode choice according to a deterministic generalized cost minimization principle at the lower level. This lower-level representation is intended as a tractable behavioral model for a capacity-constrained RAM system. In particular, passenger choices are restricted not only by generalized costs but also by the service availability induced by the operator's network, pricing, and seat-reservation decisions. At equilibrium, the seats endogenously reserved by the leader are precisely occupied by those passengers for whom the RAM service is the most attractive option. The operator commits to costly upfront infrastructure decisions, anticipating that passengers will respond optimally to the resulting service landscape. The sequence of events, along with the variables that will be introduced in the next section, is shown in Figure 1. We formally capture this leader–follower interaction using a bilevel framework, which we present in the following subsection.

### 3.2. Problem Setting

This paper investigates the RAM-SNDP problem faced by an innovative multimodal service operator that intends to offer integrated mobility services. We model the problem as a single-leader, multiple-follower Stackelberg game. The multimodal service operator, as the leader, first makes integrated decisions at both strategic and tactical levels. Strategic-level decisions include the design of the RAM network, including vertiport location selection, aerial route activation, and the deployment of eVTOL aircraft to operational routes. At the tactical level, the operator determines the pricing for both regular and priority RAM services. Passengers, as followers, subsequently make their travel mode choices in response to these service and pricing decisions. Consequently, the proposed framework holistically co-optimizes long-term infrastructure investment and pricing mechanisms to deliver an efficient and market-responsive RAM service network. Regarding the decision horizon, this paper focuses on long-term infrastructure planning at the strategic level; at the tactical/operational level, we model pricing, passenger mode choice, and capacity utilization over a representative operating period, referred to as a time block, such as a morning or evening rush-hour period demand window.

First, we introduce the notation for the leader's (operator's) problem. Let $\mathcal{V}$ denote the set of candidate vertiport locations. For each candidate location $k \in \mathcal{V}$, the operator decides whether to build

a vertiport. This decision is represented by a binary variable $x_k$. Constructing a vertiport incurs an infrastructure-related cost $f_k$, which includes both construction and maintenance expenses. Each operational vertiport has a capacity of $g_k$, representing the maximum number of eVTOLs that can take off and land within a unit time block. Furthermore, for each vertiport pair $(k, h)$, the operator determines whether to establish an aerial route, indicated by the binary variable $z_{kh}$, with an associated fixed cost $c_{kh}$. It should be noted that the aerial routes are directed and $k \neq h$. Each aerial route $(k, h)$ is constrained by low-altitude airspace management capacity $Q_{kh}$ within a unit time block. Furthermore, the operator determines eVTOL fleet deployment through non-negative integer variables $n^{\mathrm{P}}_{kh}$ and $n^{\mathrm{R}}_{kh}$, representing the number of eVTOL aircraft deployed on aerial route $(k, h)$ for priority and regular RAM services, respectively. The two service types are differentiated by their operational characteristics: priority service provides dedicated eVTOL aircraft without ride-sharing, offering a premium experience; regular service operates on a ride-sharing basis, with each eVTOL equipped with $b$ seats to accommodate multiple passengers simultaneously. Each deployed eVTOL incurs a deployment cost, denoted by $d^{\mathrm{P}}$ for priority service and $d^{\mathrm{R}}$ for regular service. Note that the deployment costs also include the operating costs of eVTOLs. The combined costs of vertiport construction, aerial route operation, and eVTOL fleet deployment must remain within the total available budget $B$. The operator also sets pricing for both service types through continuous variables $p_{kh}$ and $q_{kh}$, representing the prices for regular and priority RAM service on aerial route $(k, h)$. These prices are constrained by minimum values $\underline{p}_{kh}$ and $\underline{q}_{kh}$, which realistically represent the price floors dictated by market conditions and constraints. Finally, the binary variables $\hat{m}^{\mathrm{R}}_{khs}$ and $\hat{m}^{\mathrm{P}}_{khs}$ represent the operator's supply-side decision to reserve capacity for passenger $s \in \mathcal{S}$ to regular or priority RAM service on aerial route $(k, h)$, thereby linking passenger routing with service operations.

The follower's problem involves a set of passengers $\mathcal{S}$, where each passenger $s \in \mathcal{S}$ has a specific travel demand characterized by an origin node $\mathscr{O}(s)$ and a destination node $\mathscr{D}(s)$. Each passenger can choose among four travel modes: the operator's two RAM options (priority and regular service), the operator's expressway service, and opt-out options such as self-driving or train services. We summarize the available travel modes and provide detailed explanations in Table 1.

Each passenger $s$ has an individual value of time (VOT), denoted by $w_s$, representing the time cost. For the priority and regular RAM services, passengers take a multimodal trip: they first travel from their origin to a departure vertiport, then take an eVTOL flight to an arrival vertiport, and finally proceed to their destination. We introduce a route-specific travel time $l_{khs}$ for passenger $s$ when using the vertiport pair $(k, h)$. To capture the generalized cost associated with this trip, we introduce a fixed inconvenience cost $l^{\mathrm{P}}_s$ for priority service and $l^{\mathrm{R}}_s$ for regular service for each passenger $s$. For the expressway service, we consider a fixed inconvenience cost, denoted by $l^{\mathrm{E}}_s$.

**Table 1 Summary of available travel modes**

| Travel mode | Description |
|---|---|
| Expressway (E) | Door-to-door expressway taxi: the operator provides door-to-door service with reliable travel time, offering a convenient alternative for passengers who value simplicity and direct routing. |
| Regular RAM (R) | Ride-sharing air taxi: passengers first travel from their origin to the departure vertiport, then board an eVTOL aircraft. Passengers are responsible for both ground access legs. Upon landing at the arrival vertiport, passengers continue their journey to the final destination. This service follows a ride-sharing operating model, where multiple passengers are accommodated on the same eVTOL (with a capacity of $b$ seats per aircraft). |
| Priority RAM (P) | Dedicated air taxi: the overall process is the same as the regular RAM service. Passengers access vertiports and enjoy exclusive use of an eVTOL aircraft without sharing with other travelers. This premium service offers greater privacy and comfort at a higher fare. |
| Opt-out (O) | Alternative modes: passengers choose other transportation options not operated by the multimodal provider, such as trains, private cars, or other transportation modes. |

The expressway service has a fare of $r_s$ and a travel time of $t_s$. Additionally, the disutility cost of opt-out options is captured by $u_s^{\mathrm{O}}$ for passenger $s$.

Based on the aforementioned parameters, each passenger $s$ makes a travel mode choice according to their individual generalized cost. To formulate this travel choice mathematically, we define a set of binary decision variables. Specifically, let $v_s^{\mathrm{R}}$, $v_s^{\mathrm{P}}$, $v_s^{\mathrm{E}}$, and $v_s^{\mathrm{O}}$ be binary variables that equal one if and only if passenger $s$ chooses the regular RAM service, priority RAM service, expressway service, or an opt-out option, respectively. Furthermore, for passengers who select either RAM service, the routing decision is incorporated in our optimization framework. This routing decision is captured by the binary variables $m_{khs}^{\mathrm{R}}$ and $m_{khs}^{\mathrm{P}}$, which indicate whether passenger $s$ travels on the aerial route from vertiport $k$ to $h$ using the regular or priority RAM service, respectively. We also summarize each component of the generalized cost for available travel modes in Table 2. The notations utilized in this paper are summarized in Table A.1.

In this paper, the generalized-cost formulation serves as a reduced-form representation of passenger behavior that is especially suitable for the present bilevel design problem, because it captures the key supply-side levers affected by the operator (price, route availability, and seat capacity) without introducing the additional nonlinearities of stochastic demand specifications.

### 3.3. Bilevel Program for the RAM-SNDP problem

The bilevel program for the RAM-SNDP problem is formulated as follows:

$$\text{(BP)} \quad \text{maximize} \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} p_{kh} m_{khs}^{\mathrm{R}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} q_{kh} m_{khs}^{\mathrm{P}} + \sum_{s\in\mathcal{S}} r_s v_s^{\mathrm{E}} \tag{1a}$$

**Table 2 Summary of generalized cost components of travel modes**

| Travel mode | Monetary fare | VOT | Inconvenience cost | Generalized cost |
|---|---|---|---|---|
| Expressway (E) | $r_s$ | $w_s t_s$ | $l_s^{\mathrm{E}}$ | $r_s + w_s t_s + l_s^{\mathrm{E}}$ (for passenger $s$) |
| Regular RAM (R) | $p_{kh}$ | $w_s l_{khs}$ | $l_s^{\mathrm{R}}$ | $p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}}$ (for passengers $s$ who chooses route $(k,h)$) |
| Priority RAM (P) | $q_{kh}$ | $w_s l_{khs}$ | $l_s^{\mathrm{P}}$ | $q_{kh} + w_s l_{khs} + l_s^{\mathrm{P}}$ (for passenger $s$ who choose route $(k,h)$) |
| Opt-out (O) | – | – | – | $u_s^{\mathrm{O}}$ (for passenger $s$) |

$$\text{subject to} \quad \sum_{k\in\mathcal{V}} f_k x_k + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} d^{\mathrm{R}} n_{kh}^{\mathrm{R}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} d^{\mathrm{P}} n_{kh}^{\mathrm{P}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} c_{kh} z_{kh} \le B, \tag{1b}$$

$$z_{kh} \le x_k, \ \ k,h \in \mathcal{V}, \tag{1c}$$

$$z_{kh} \le x_h, \ \ k,h \in \mathcal{V}, \tag{1d}$$

$$n_{kh}^{\mathrm{R}} + n_{kh}^{\mathrm{P}} \le Q_{kh} z_{kh}, \ \ k,h \in \mathcal{V}, \tag{1e}$$

$$\sum_{h\in\mathcal{V}} n_{hk}^{\mathrm{P}} + \sum_{h\in\mathcal{V}} n_{hk}^{\mathrm{R}} + \sum_{h\in\mathcal{V}} n_{kh}^{\mathrm{P}} + \sum_{h\in\mathcal{V}} n_{kh}^{\mathrm{R}} \le g_k x_k, \ \ k \in \mathcal{V}, \tag{1f}$$

$$\sum_{s\in\mathcal{S}} \hat{m}_{khs}^{\mathrm{R}} \le b n_{kh}^{\mathrm{R}}, \ \ k,h \in \mathcal{V}, \tag{1g}$$

$$\sum_{s\in\mathcal{S}} \hat{m}_{khs}^{\mathrm{P}} \le n_{kh}^{\mathrm{P}}, \ \ k,h \in \mathcal{V}, \tag{1h}$$

$$p_{kh} \ge \underline{p}_{kh} z_{kh}, \ \ k,h \in \mathcal{V}, \tag{1i}$$

$$q_{kh} \ge \underline{q}_{kh} z_{kh}, \ \ k,h \in \mathcal{V}, \tag{1j}$$

$$x_k, z_{kh} \in \{0,1\}, \ \ k,h \in \mathcal{V}, \tag{1k}$$

$$\hat{m}_{khs}^{\mathrm{R}}, \hat{m}_{khs}^{\mathrm{P}} \in \{0,1\}, \ \ k,h \in \mathcal{V}, s \in \mathcal{S}, \tag{1l}$$

$$n_{kh}^{\mathrm{R}}, n_{kh}^{\mathrm{P}} \in \mathbb{Z}^+, \ \ k,h \in \mathcal{V}, \tag{1m}$$

$$p_{kh}, q_{kh} \ge 0, \ \ k,h \in \mathcal{V}, \tag{1n}$$

where $v_s^{\mathrm{E}}$, $m_{khs}^{\mathrm{P}}$ and $m_{khs}^{\mathrm{R}}$ are optimal solutions of the follower model $s$, for all $s \in \mathcal{S}$,

$$\begin{aligned}\text{minimize} \quad & \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} p_{kh} m_{khs}^{\mathrm{R}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} q_{kh} m_{khs}^{\mathrm{P}} + r_s v_s^E \\ & + (l_s^{\mathrm{E}} + w_s t_s) v_s^{\mathrm{E}} + l_s^{\mathrm{P}} v_s^{\mathrm{P}} + l_s^{\mathrm{R}} v_s^{\mathrm{R}} + u_s^{\mathrm{O}} v_s^{\mathrm{O}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} w_s l_{khs} (m_{khs}^{\mathrm{R}} + m_{khs}^{\mathrm{P}})\end{aligned} \tag{2a}$$

$$\text{subject to} \quad v_s^{\mathrm{E}} + v_s^{\mathrm{O}} + v_s^{\mathrm{R}} + v_s^{\mathrm{P}} = 1, \tag{2b}$$

$$\sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} m_{khs}^{\mathrm{R}} = v_s^{\mathrm{R}}, \tag{2c}$$

$$\sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} m^{\mathrm{P}}_{khs} = v^{\mathrm{P}}_{s}, \tag{2d}$$

$$m^{\mathrm{R}}_{khs} \le \hat{m}^{\mathrm{R}}_{khs}, \;\; k,h\in\mathcal{V}, \tag{2e}$$

$$m^{\mathrm{P}}_{khs} \le \hat{m}^{\mathrm{P}}_{khs}, \;\; k,h\in\mathcal{V}, \tag{2f}$$

$$v^{\mathrm{R}}_{s}, v^{\mathrm{P}}_{s}, v^{\mathrm{O}}_{s}, v^{\mathrm{E}}_{s}, m^{\mathrm{R}}_{khs}, m^{\mathrm{P}}_{khs} \in \{0,1\}, \;\; k,h\in\mathcal{V}. \tag{2g}$$

The objective function of the leader's problem (1a) is to maximize the revenue. Constraints (1b) impose a budget limit on vertiport construction, eVTOL deployment, and aerial route opening. Constraints (1c) and (1d) ensure that an aerial route can only be operational if both of its endpoint vertiports are open. Constraints (1e) enforce that the number of eVTOLs deployed on each route does not exceed the low-altitude airspace management capacity. Constraints (1f) limit the number of eVTOLs taking off and landing at each vertiport to its service capacity. Constraints (1g) and (1h) guarantee that the seat supply for both regular and priority RAM services cannot exceed the total available seat capacity. Constraints (1i) and (1j) define the minimum prices of regular and priority RAM service. Constraints (1k)–(1n) define the variable domains of the leader's problem.

The objective function (2a) of the follower's problem is to minimize the generalized cost of each passenger. Constraints (2b) ensure that each passenger selects only one travel mode. For a passenger choosing the regular or priority RAM service, Constraints (2c) and (2d) require that their trip must be completed using an aerial route. Furthermore, Constraints (2e) and Constraint (2f) stipulate that a passenger can be reserved capacity on an aerial route only if sufficient supply is available on it. Finally, Constraints (2g) define the variable domains of the follower's problem.

Model (2) can be equivalently reformulated as a linear program, as described in the following proposition.

**Proposition 1** *The constraint matrix of model (2) is totally unimodular. Therefore, model (2) with relaxed integrality requirements always has an integral optimal solution.*

*Proof.* See Appendix B. □

## 4. Solution Procedure

To solve the bilevel program introduced in Section 3, we first reformulate it into an equivalent single-level program using the strong duality theorem in Section 4.1. We then derive the exact best-response sets of passengers in Section 4.2. To improve computational efficiency, we tighten the Big-M coefficients in our models in Section 4.3, and propose several variable reduction techniques in Section 4.4 as well as valid inequalities in Section 4.5.

### 4.1. Reformulation as a Single-Level Program

Based on Proposition 1, the integrality constraints on the variables $v_s^{\mathrm{E}}$, $v_s^{\mathrm{O}}$, $v_s^{\mathrm{P}}$, $v_s^{\mathrm{R}}$, $m_{khs}^{\mathrm{P}}$, and $m_{khs}^{\mathrm{R}}$ can be relaxed. That is, these variables are allowed to be continuous and satisfy $\{v_s^{\mathrm{E}}, v_s^{\mathrm{O}}, v_s^{\mathrm{P}}, v_s^{\mathrm{R}}, m_{khs}^{\mathrm{P}}, m_{khs}^{\mathrm{R}} \geq 0\}$. The upper bounds $\{v_s^{\mathrm{E}}, v_s^{\mathrm{O}}, v_s^{\mathrm{P}}, v_s^{\mathrm{R}}, m_{khs}^{\mathrm{P}}, m_{khs}^{\mathrm{R}} \leq 1\}$ can be omitted because they are implied by Constraints (2b)–(2f). Because model (2) is bounded and feasible, the strong duality can be leveraged to represent the followers' optimal decisions. Specifically, let $\alpha_s$, $\beta_s$, $\gamma_s$, $\theta_{khs}$, and $\kappa_{khs}$ be dual variables associated with Constraints (2b)–(2f), respectively. Then the dual of each model $s \in \mathcal{S}$ is

$$\text{maximize } \alpha_s - \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} \hat{m}_{khs}^{\mathrm{R}}\theta_{khs} - \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} \hat{m}_{khs}^{\mathrm{P}}\kappa_{khs} \tag{3a}$$

$$\text{subject to } \alpha_s \leq l_s^{\mathrm{E}} + r_s + w_s t_s, \tag{3b}$$

$$\alpha_s \leq u_s^{\mathrm{O}}, \tag{3c}$$

$$\alpha_s - \gamma_s \leq l_s^{\mathrm{P}}, \tag{3d}$$

$$\alpha_s - \beta_s \leq l_s^{\mathrm{R}}, \tag{3e}$$

$$\beta_s - \theta_{khs} \leq p_{kh} + w_s l_{khs}, \ \ k,h \in \mathcal{V}, \tag{3f}$$

$$\gamma_s - \kappa_{khs} \leq q_{kh} + w_s l_{khs}, \ \ k,h \in \mathcal{V}, \tag{3g}$$

$$\theta_{khs}, \kappa_{khs} \geq 0, \ \ k,h \in \mathcal{V}. \tag{3h}$$

We then reformulate the bilevel program into an equivalent single-level program as follows:

$$\text{maximize } \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} p_{kh} m_{khs}^{\mathrm{R}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} q_{kh} m_{khs}^{\mathrm{P}} + \sum_{s\in\mathcal{S}} r_s v_s^{\mathrm{E}} \tag{4a}$$

$$\text{subject to } (1b)-(1n), (2b)-(2f), (3b)-(3h),$$

$$\alpha_s - \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} \hat{m}_{khs}^{\mathrm{R}}\theta_{khs} - \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} \hat{m}_{khs}^{\mathrm{P}}\kappa_{khs} \geq \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} (p_{kh} + w_s l_{khs}) m_{khs}^{\mathrm{R}}$$
$$+ \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} (q_{kh} + w_s l_{khs}) m_{khs}^{\mathrm{P}} + (l_s^{\mathrm{E}} + r_s + w_s t_s) v_s^{\mathrm{E}} + u_s^{\mathrm{O}} v_s^{\mathrm{O}} + l_s^{\mathrm{P}} v_s^{\mathrm{P}} + l_s^{\mathrm{R}} v_s^{\mathrm{R}}, \tag{4b}$$

$$v_s^{\mathrm{R}}, v_s^{\mathrm{P}}, v_s^{\mathrm{O}}, v_s^{\mathrm{E}}, m_{khs}^{\mathrm{R}}, m_{khs}^{\mathrm{P}} \geq 0, \ \ k,h \in \mathcal{V}, \tag{4c}$$

where Constraints (1b)–(1n) are constraints of the operator model, Constraints (2b)–(2f) ensure the primal feasibility and (3b)–(3h) guarantee the dual feasibility of the passenger model, Constraints (4b) ensure the optimality of the passenger model, and Constraints (4c) are variable domains.

Because of the bilinear terms $p_{kh} m_{khs}^{\mathrm{R}}$, $q_{kh} m_{khs}^{\mathrm{P}}$, $\hat{m}_{khs}^{\mathrm{R}}\theta_{khs}$ and $\hat{m}_{khs}^{\mathrm{P}}\kappa_{khs}$, the above bilevel model is a mixed-integer nonlinear program. Therefore, we define auxiliary variables $\rho_{khs} := p_{kh} m_{khs}^{\mathrm{R}}$,

$\xi_{khs} \coloneqq q_{kh} m^{\mathrm{P}}_{khs}$, $\sigma_{khs} \coloneqq \hat{m}^{\mathrm{R}}_{khs} \theta_{khs}$, $\tau_{khs} \coloneqq \hat{m}^{\mathrm{P}}_{khs} \kappa_{khs}$ and add the following linear constraints to the bilevel model, for all $k, h \in \mathcal{V}$ and $s \in \mathcal{S}$:

$$0 \le \rho_{khs} \le p_{kh}, \tag{5a}$$

$$0 \le \xi_{khs} \le q_{kh}, \tag{5b}$$

$$p_{kh} - M_1^{kh}(1 - m^{\mathrm{R}}_{khs}) \le \rho_{khs} \le M_1^{kh} m^{\mathrm{R}}_{khs}, \tag{5c}$$

$$q_{kh} - M_2^{kh}(1 - m^{\mathrm{P}}_{khs}) \le \xi_{khs} \le M_2^{kh} m^{\mathrm{P}}_{khs}. \tag{5d}$$

$$0 \le \sigma_{khs} \le \theta_{khs}, \tag{5e}$$

$$0 \le \tau_{khs} \le \kappa_{khs}, \tag{5f}$$

$$\theta_{khs} - M_3^{s}(1 - \hat{m}^{\mathrm{R}}_{khs}) \le \sigma_{khs} \le M_3^{s} \hat{m}^{\mathrm{R}}_{khs}, \tag{5g}$$

$$\kappa_{kh} - M_4^{s}(1 - \hat{m}^{\mathrm{P}}_{khs}) \le \tau_{khs} \le M_4^{s} \hat{m}^{\mathrm{P}}_{khs}, \tag{5h}$$

where $M_1^{kh}$, $M_2^{kh}$, $M_3^{s}$, and $M_4^{s}$ are sufficiently large constants.

The single-level program is then reformulated as follows:

$$\text{(LSP1)} \quad \text{maximize} \quad \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} \rho_{khs} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} \xi_{khs} + \sum_{s\in\mathcal{S}} r_s v^{\mathrm{E}}_s \tag{6a}$$

$$\text{subject to} \quad (1\text{b}) - (1\text{n}), (2\text{b}) - (2\text{f}), (3\text{b}) - (3\text{h}), (5\text{a}) - (5\text{d}), (5\text{e}) - (5\text{h}), (4\text{c})$$

$$\alpha_s - \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} \sigma_{khs} - \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} \tau_{khs} \ge \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} (\rho_{khs} + w_s l_{khs} m^{\mathrm{R}}_{khs} + \xi_{khs} + w_s l_{khs} m^{\mathrm{P}}_{khs})$$
$$+ (l^{\mathrm{E}}_s + r_s + w_s t_s) v^{\mathrm{E}}_s + u^{\mathrm{O}}_s v^{\mathrm{O}}_s + l^{\mathrm{P}}_s v^{\mathrm{P}}_s + l^{\mathrm{R}}_s v^{\mathrm{R}}_s. \tag{6b}$$

The program LSP1 is an equivalent mixed-integer linear program (MILP) reformulation of the bilevel program BP.

### 4.2. Passenger Best Response Set

We now derive a linear programming representation of the optimal solution set for each passenger, commonly referred to as the best response set, in order to enhance the computational efficiency of solving the bilevel program. Since each passenger faces a finite set of feasible choices, it is possible to explicitly enumerate the objective values corresponding to all possible decisions, thereby verifying whether a given response is indeed optimal. Before we proceed, we first show that customer decisions match the operator's seat reservation decisions. Observe that, because offering inventory that is not selected provides no benefit and consumes capacity, we can restrict attention to solutions where all offered inventory is used. The following proposition formalizes that, without loss of generality, the optimal solutions never admit unused capacity.

**Proposition 2** *For the bilevel program, there exists an optimal solution in which* $\hat{m}^R_{khs} = m^R_{khs}$ *and* $\hat{m}^P_{khs} = m^P_{khs}$ *for all* $k, h \in \mathcal{V}$ *and* $s \in \mathcal{S}$*. Hence, imposing* $\hat{m}^R_{khs} = m^R_{khs}$ *and* $\hat{m}^P_{khs} = m^P_{khs}$ *is without loss of optimality.*

*Proof.* See Appendix C. □

We next present the explicit expressions for the best response of passengers.

**Proposition 3** *Considering passenger* $s \in \mathcal{S}$ *and fixed values of* $p_{kh}$ *and* $q_{kh}$*, a solution* $(v^E_s, v^O_s, v^P_s, v^R_s, m^R_{khs}, m^P_{khs})$ *is optimal for model (2) if and only if it satisfies Constraints (2b)–(2f) and there exists* $\Psi_s \geq 0$ *such that*

$$\Psi_s \geq \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(p_{kh} + w_s l_{khs})m^R_{khs} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(q_{kh} + w_s l_{khs})m^P_{khs} + (r_s + l^E_s + w_s t_s)v^E_s + u^O_s v^O_s + l^P_s v^P_s + l^R_s v^R_s, \tag{7a}$$

$$\Psi_s \leq l^E_s + r_s + w_s t_s, \tag{7b}$$

$$\Psi_s \leq u^O_s, \tag{7c}$$

$$\Psi_s \leq p_{kh} + w_s l_{khs} + l^R_s + M^s_5(1 - m^R_{khs}), \quad k, h \in \mathcal{V}, \tag{7d}$$

$$\Psi_s \leq q_{kh} + w_s l_{khs} + l^P_s + M^s_6(1 - m^P_{khs}), \quad k, h \in \mathcal{V}, \tag{7e}$$

*where* $M^s_5$ *and* $M^s_6$ *are sufficiently large numbers.*

*Proof.* See Appendix D. □

Based on Proposition 3, we can reformulate the bilevel program into an equivalent single-level program.

$$\text{maximize} \quad \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} p_{kh} m^{\mathrm{R}}_{khs} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} q_{kh} m^{\mathrm{P}}_{khs} + \sum_{s\in\mathcal{S}} r_s v^{\mathrm{E}}_s \tag{8a}$$

$$\text{subject to} \quad (1b)-(1f), (1i)-(1j), (1k), (1m)-(1n), (2b)-(2d), (2g), (7a)-(7e),$$

$$\sum_{s\in\mathcal{S}} m^{\mathrm{R}}_{khs} \leq b n^{\mathrm{R}}_{kh}, \quad k, h \in \mathcal{V}, \tag{8b}$$

$$\sum_{s\in\mathcal{S}} m^{\mathrm{P}}_{khs} \leq n^{\mathrm{P}}_{kh}, \quad k, h \in \mathcal{V}, \tag{8c}$$

$$\Psi_s \in \mathbb{R}, \quad s \in \mathcal{S}. \tag{8d}$$

We then linearize the bilinear terms to derive a linear single-level program by letting $\rho_{khs} = p_{kh} m^{\mathrm{R}}_{khs}$ and $\xi_{khs} = q_{kh} m^{\mathrm{P}}_{khs}$ and by adding the linear constraints (5a)–(5d) into the model:

$$(\texttt{LSP2}) \quad \text{maximize} \quad \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} \rho_{khs} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}} \xi_{khs} + \sum_{s\in\mathcal{S}} r_s v^{\mathrm{E}}_s \tag{9a}$$

$$\text{subject to} \quad (1b)-(1f), (1i)-(1k), (1m)-(1n), (2b)-(2d), (2g), (7b)-(7e), (5a)-(5d), (8b)-(8d),$$

$$\Psi_s \geq \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(\rho_{khs} + w_s l_{khs} m^{\mathrm{R}}_{khs}) + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(\xi_{khs} + w_s l_{khs} m^{\mathrm{P}}_{khs}) + (r_s + l^{\mathrm{E}}_s + w_s t_s)v^{\mathrm{E}}_s + u^{\mathrm{O}}_s v^{\mathrm{O}}_s + l^{\mathrm{P}}_s v^{\mathrm{P}}_s + l^{\mathrm{R}}_s v^{\mathrm{R}}_s. \tag{9b}$$

**Remark 1** *The proposed* `LSP2` *can be readily extended to a stochastic setting. By introducing a finite set of scenarios* $\Omega$ *representing demand realizations, the deterministic model can be extended as a stochastic program. This extension naturally captures the linkage between strategic decisions (e.g., infrastructure investment) and operational decisions (e.g., passenger mode choice) that adapt to realized demand. We present a stochastic formulation in the Appendix H. However, since the stochastic extension does not alter the essential nature of the problem, we focus on the deterministic case (i.e., a single scenario) and explore the bilevel structure of the problem in this paper for clarity.*

### 4.3. Big-M Tightening

To enhance computational performance, we tighten the six sets of Big-M coefficients in our formulation. Loosely chosen large values create a weak linear programming relaxation, which degrades the efficiency of the solution process. Three propositions are in order.

**Proposition 4** *The* $M_1^{kh}$ *and* $M_2^{kh}$ *in Constraints* (5c) *and* (5d) *can be selected, for* $k, v \in \mathcal{V}$*, as*

$$M_1^{kh} = \max_{s\in\mathcal{S}} \left\{ \min \left\{ u_s^{\mathrm{O}} - w_s l_{khs} - l_s^{\mathrm{R}}, r_s + l_s^{\mathrm{E}} + w_s t_s - w_s l_{khs} - l_s^{\mathrm{R}} \right\}, 0 \right\}, \tag{10a}$$

$$M_2^{kh} = \max_{s\in\mathcal{S}} \left\{ \min \left\{ u_s^{\mathrm{O}} - w_s l_{khs} - l_s^{\mathrm{P}}, r_s + l_s^{\mathrm{E}} + w_s t_s - w_s l_{khs} - l_s^{\mathrm{P}} \right\}, 0 \right\}. \tag{10b}$$

*Proof.* See Appendix E. □

**Proposition 5** *The* $M_3$ *and* $M_4$ *in Constraints* (5g) *and* (5h) *can be selected, for* $s \in \mathcal{S}$*, as*

$$M_3^s = M_4^s = u_s^{\mathrm{O}}. \tag{11}$$

*Proof.* See Appendix F. □

**Proposition 6** *The* $M_5$ *and* $M_6$ *in Constraints* (7d) *and* (7e) *can be selected, for* $s \in \mathcal{S}$*, as*

$$M_5^s = M_6^s = \min \left\{ l_s^{\mathrm{E}} + r_s + w_s t_s, u_s^{\mathrm{O}} \right\}. \tag{12}$$

*Proof.* See Appendix G. □

### 4.4. Variable Reduction

In this subsection, we propose several variable reduction techniques to enhance the computational efficiency of models `LSP1` and `LSP2`. The variable reduction techniques eliminate solutions that are operationally feasible but suboptimal from the passengers' perspective.

**Proposition 7** *A passenger* $s \in \mathcal{S}$ *will not choose the expressway service if the generalized cost of expressway service, comprising the fare* $r_s$*, the travel time cost* $w_s t_s$*, and the inconvenience cost* $l_s^{\mathrm{E}}$*, exceeds the utility* $u_s^{\mathrm{O}}$ *of the opt-out option. That is, for each passenger* $s$*, if*

$$r_s + l_s^{\mathrm{E}} + w_s t_s \geq u_s^{\mathrm{O}}, \tag{13}$$

*then* $v_s^{\mathrm{E}} = 0$*, and the variable* $v_s^{\mathrm{E}}$ *can be removed from the model.*

*Proof.* The proof of the Proposition 7 is straightforward. □

**Proposition 8** *A passenger $s \in \mathcal{S}$ will not choose regular RAM service if the minimum generalized cost of regular RAM travel, which includes the minimum price across all the aerial routes $\min_{k\in\mathcal{V},h\in\mathcal{V}} \underline{p}_{kh}$, inconvenience cost $l_s^{\mathrm{R}}$ and the minimum time cost $w_s \min_{k,h} l_{khs}$, exceeds the disutility cost of the opt-out option $u_s^{\mathrm{O}}$ or the generalized cost of expressway service, comprising the fare $r_s$, the travel time cost $w_s t_s$, and the inconvenience cost $l_s^{\mathrm{E}}$. That is, for each passenger $s$, if*

$$\min_{k\in\mathcal{V},h\in\mathcal{V}} \underline{p}_{kh} + l_s^{\mathrm{R}} + w_s \min_{k\in\mathcal{V},h\in\mathcal{V}} l_{khs} \geq \min\left\{u_s^{\mathrm{O}}, r_s + l_s^{\mathrm{E}} + w_s t_s\right\}, \tag{14}$$

*then $v_s^{\mathrm{R}} = 0$, and the variable $v_s^{\mathrm{R}}$ can be removed from the model; Similarly, for each passenger $s$, if*

$$\min_{k\in\mathcal{V},h\in\mathcal{V}} \underline{q}_{kh} + l_s^{\mathrm{P}} + w_s \min_{k\in\mathcal{V},h\in\mathcal{V}} l_{khs} \geq \min\left\{u_s^{\mathrm{O}}, r_s + l_s^{\mathrm{E}} + w_s t_s\right\}, \tag{15}$$

*then $v_s^{\mathrm{P}} = 0$, and the variable $v_s^{\mathrm{P}}$ can be removed from the model.*

*Proof.* The proof of the Proposition 8 is straightforward. □

**Proposition 9** *A passenger $s \in \mathcal{S}$ choosing regular RAM service will not select an aerial route from vertiport $k$ to vertiport $h$ if it is dominated by other alternatives. This occurs its generalized cost, including the route-specific minimum fare $\underline{p}_{kh}$, the inconvenience cost $l_s^{\mathrm{R}}$, and the time cost $w_s l_{khs}$) exceeds the disutility $u_s^{\mathrm{O}}$ of the opt-out option or exceeds the generalized cost of the expressway service. That is, for each aerial route $(k,h)$ and each passenger $s$, if*

$$\underline{p}_{kh} + l_s^{\mathrm{R}} + w_s l_{khs} \geq \min\left\{u_s^{\mathrm{O}}, r_s + l_s^{\mathrm{E}} + w_s t_s\right\}, \tag{16}$$

*then $m_{khs}^{\mathrm{R}} = 0$, and the variable $m_{khs}^{\mathrm{R}}$ can be removed from the model. Similarly, for each aerial route $(k,h)$ and each passenger $s$, if*

$$\underline{q}_{kh} + l_s^{\mathrm{P}} + w_s l_{khs} \geq \min\left\{u_s^{\mathrm{O}}, r_s + l_s^{\mathrm{E}} + w_s t_s\right\}, \tag{17}$$

*then $m_{khs}^{\mathrm{P}} = 0$, and the variable $m_{khs}^{\mathrm{P}}$ can be removed from the model.*

*Proof.* The proof of the Proposition 9 is straightforward. □

### 4.5. Valid Inequalities

Models `LSP1` and `LSP2` can be strengthened by two families of valid inequalities. The first family of valid inequalities is derived by coupling passenger routing decisions with vertiport location choices.

Specifically, if a node is not selected as a vertiport, then passengers cannot be scheduled to travel via that node. This logical relationship is formalized as follows:

$$m^{\mathrm{R}}_{khs} \le x_k, \;\; k,h \in \mathcal{V}, s \in \mathcal{S}, \tag{18a}$$

$$m^{\mathrm{R}}_{khs} \le x_h, \;\; k,h \in \mathcal{V}, s \in \mathcal{S}, \tag{18b}$$

$$m^{\mathrm{P}}_{khs} \le x_k, \;\; k,h \in \mathcal{V}, s \in \mathcal{S}, \tag{18c}$$

$$m^{\mathrm{P}}_{khs} \le x_h, \;\; k,h \in \mathcal{V}, s \in \mathcal{S}. \tag{18d}$$

In the second family of valid inequalities, we apply a bound tightening to strengthen the formulation. Specifically, for variables $\rho_{khs}$ and $\xi_{khs}$, we derive and impose a valid bound and obtain the following valid inequalities.

$$\rho_{khs} \ge \underline{p}_{kh} m^{\mathrm{R}}_{khs}, \;\; k,h \in \mathcal{V}, s \in \mathcal{S}, \tag{19a}$$

$$\xi_{khs} \ge \underline{q}_{kh} m^{\mathrm{P}}_{khs}. \;\; k,h \in \mathcal{V}, s \in \mathcal{S}. \tag{19b}$$

## 5. Numerical Study

In this section, we first assess the efficiency of the solution procedures in Section 5.1 and then conduct a case study inspired by real-world data in Section 5.2. The solutions were implemented in Python with GUROBI 13.0.1, and experiments were performed on a computer with a Windows operating system featuring an AMD Ryzen 9 9950X3D 16-Core Processor and 96 GB of RAM.

### 5.1. Computational Efficiency Evaluation

We now examine the computational performance. The experimental parameters and data settings are first introduced, followed by the computational efficiency analysis.

#### 5.1.1. Parameter and Data Setting

To evaluate the performance of the proposed solution methodology, we conduct numerical experiments based on the high-traffic Hong Kong–Shenzhen corridor in China. We consider a transportation network consisting of 129 population centers represented by nodes, distributed across Hong Kong and Shenzhen. Among these nodes, 13 population centers are candidate vertiport locations (seven in Hong Kong and six in Shenzhen). We consider intercity round-trip demand between Hong Kong and Shenzhen. Specifically, each passenger is associated with a unique origin–destination (OD) pair generated randomly from the 129 population centers, with the origin located in one city and the destination in the other. We test the model with passenger sets of varying sizes: $|\mathcal{S}| \in \{50, 500, 1000, 2000, 6000, 10000, 20000, 50000\}$. For each passenger scale, we generate five instances, resulting in a total of 40 instances for the experimental study. We introduce the parameter settings as follows:

- The total investment budget is US$40,000,000. The unit opening cost of a vertiport is set to $1,000,000 (Delaware Aeronautics 2024). The operational cost of an aerial route is $5000. The unit deployment cost for the four-seat eVTOL aircraft utilized for regular and priority service is respectively set to $338,000 and $371,800 per aircraft (Richard 2024).
- The expressway price $r_s$ is set to $0.8 per kilometer multiplied by the ground distance between the origin and destination for the passenger $s$. The ground distance is assumed to be 1.2 times the Euclidean distance between two points (Bekli et al. 2021). The minimum prices for the regular and priority RAM services are set to 0.8 and 0.7 times their respective maximum prices.
- Regarding passengers' VOT, we classify travelers into two categories. For personal travel, VOT is randomly generated from $31.0 to $46.5 per hour; for business travel, VOT is randomly generated from $50.6 to $75.8 per hour (Federal Aviation Administration 2016).
- The travel time of the RAM service consists of four components: first-mile and last-mile travel time, air leg time, and transfer time. The air leg time is calculated by dividing the Euclidean distance by the cruising speed of the eVTOLs, which is set to 200 km/h (André and Hajek 2019). The first-mile and last-mile travel times are determined by dividing the ground distance from the passenger's origin to the departure vertiport and from the arrival vertiport to the destination, respectively, by the ground vehicle speed of 80 km/h. The transfer time at vertiports is set to 10 minutes. Similarly, the travel time of the expressway service $t_s$ is calculated by dividing the origin–destination ground distance of passenger $s$ by the ground vehicle speed, also set at 80 km/h.
- The inconvenience costs for expressway and RAM services are randomly generated within the range of $10 to $30, with the condition that $l_s^{\mathrm{E}} > l_s^{\mathrm{R}} > l_s^{\mathrm{P}}$. The disutility cost of the opt-out option $u_s^{\mathrm{O}}$ is generated from $50 to $70 for personal travel passengers and from $70 to $100 for business travel passengers.
- The capacity-related parameters are set as follows. For vertiport operational capacity, the total capacity per vertiport is set to 60 (Zahn et al. 2023). Meanwhile, the airspace management capacity is set to 20 eVTOL aircraft (Jin et al. 2024).

#### 5.1.2. Effectiveness and Efficiency of Solution Procedures

We now compare the computational efficiency of the proposed models `LSP1` and `LSP2` as well as the enhancement methodologies (variable reduction techniques and valid inequalities). Table 3 reports the performance of `LSP1` and `LSP2`. All experiments were conducted with a time limit of 7200 seconds. The reported metrics include the objective value (Obj), the time when the last incumbent solution was found (Inc.Time), the LP relaxation gap (LP-R), the CPU time (Time), and the optimality gap (Gap). Specifically, LP-R is defined as $\frac{|Z^*_{LP} - Z^*_{LB}|}{Z^*_{LB}} \times 100$, and Gap is defined as $\frac{|Z^*_{UB} - Z^*_{LB}|}{Z^*_{LB}} \times 100$, where $Z^*_{LP}$, $Z^*_{UB}$, and $Z^*_{LB}$ denote the optimal objective value of the LP relaxation, the upper bound, and the lower bound, respectively.

The results in Table 3 show that the LSP2 yields moderate improvements over LSP1. For small instances with up to 500 passengers, both models solve to optimality within a short time. However, as instance size increases, the computational advantages of LSP2 become more pronounced. Across all instances, the average solution time decreases from 5579.0 seconds for LSP1 to 5329.7 seconds for LSP2, while the average optimality gap decreases from 12.90% to 12.68%.

Table 4 reports the computational performance after applying the enhancement techniques to LSP2. The incorporation of variable reduction techniques in LSP2+VR yields substantial performance gains. The average solution time decreases from 5329.7 seconds for LSP2 to 4837.6 seconds for LSP2+VR, while the average optimality gap drops from 12.68% to 0.25%. More importantly, LSP2+VR demonstrates the capability to solve large-scale instances with up to 50,000 passengers, a task that LSP2 cannot accomplish. Further improvements are achieved by incorporating valid inequalities into the formulation (LSP2+VR+VI). Across all instances, LSP2+VR+VI consistently outperforms the other variants. The average CPU time and the average Inc.Time are further reduced, and the LP-R decreases to 8.71%, representing a substantial tightening of the formulation compared to the 16.93% observed for LSP2+VR. Notably, LSP2+VR+VI also yields the highest average objective function value. The decrease in LP-R for larger instances should be interpreted cautiously because LP-R is a relative measure. As $|S|$ grows while the budget and capacity parameters remain fixed, the integer objective increases with market size, whereas the LP's fractional advantage is still limited by the same investment and capacity constraints. Thus, the relative gap may decrease even if the formulation is not inherently tighter. In addition, larger passenger pools create a demand-aggregation effect: attractive routes and service patterns have enough similar passengers to fill integer aircraft and seat capacities efficiently, so the integer solution more closely approximates the LP relaxation. Overall, our enhancement methodologies can boost performance in terms of both efficiency and effectiveness.

**Table 3** **Performance of model `LSP1` and `LSP2`**

| ID | $\|\mathcal{S}\|$ | `LSP1` | | | | | `LSP2` | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Inc.Time(s) | Obj | LP-R(%) | Time(s) | Gap(%) | Inc.Time(s) | Obj | LP-R(%) | Time(s) | Gap(%) |
| 1 | 50 | 0.6 | 975.3 | 0.57 | 0.6 | 0.00 | 0.6 | 975.3 | 0.57 | 0.6 | 0.00 |
| 2 | 500 | 995.2 | 3701.1 | 26.01 | 1645.6 | 0.00 | 944.2 | 3701.1 | 25.97 | 1019.3 | 0.00 |
| 3 | 1,000 | 3548.8 | 5005.0 | 26.69 | 5364.9 | 0.13 | 1435.4 | 5005.0 | 26.73 | 3747.7 | 0.03 |
| 4 | 2,000 | 3728.2 | 6855.3 | 27.33 | 7200.0 | 0.23 | 4210.6 | 6856.3 | 27.32 | 7200.0 | 0.09 |
| 5 | 6,000 | 6227.4 | 13519.0 | 23.70 | 7200.0 | 0.70 | 4476.0 | 13524.7 | 23.62 | 7200.0 | 0.20 |
| 6 | 10,000 | 5618.4 | 18786.9 | 19.47 | 7200.0 | 1.52 | 5270.4 | 18829.8 | 19.19 | 7200.0 | 0.44 |
| 7 | 20,000 | 5344.6 | 29426.8 | 9.31 | 7200.0 | 0.61 | 5790.0 | 29408.8 | 9.38 | 7200.0 | 0.68 |
| 8 | 50,000 | – | – | – | 7200.0 | 100.00 | – | – | – | 7200.0 | 100.00 |
| **Avg** | | – | – | – | 5579.0 | 12.90 | – | – | – | 5329.7 | 12.68 |

**Table 4** **Performance of enhancement methodologies**

| ID | $\|\mathcal{S}\|$ | `LSP2+VR` | | | | | `LSP2+VR+VI` | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Inc.Time(s) | Obj | LP-R(%) | Time(s) | Gap(%) | Inc.Time(s) | Obj | LP-R(%) | Time(s) | Gap(%) |
| 1 | 50 | 0.2 | 975.3 | 0.57 | 0.24 | 0.00 | 0.4 | 975.3 | 0.57 | 0.4 | 0.00 |
| 2 | 500 | 416.8 | 3701.1 | 25.88 | 560.0 | 0.00 | 406.8 | 3701.1 | 12.64 | 453.8 | 0.00 |
| 3 | 1,000 | 1587.2 | 5005.0 | 26.51 | 3333.0 | 0.03 | 2099.8 | 5005.0 | 11.83 | 2923.4 | 0.01 |
| 4 | 2,000 | 2943.6 | 6855.7 | 26.98 | 6007.4 | 0.10 | 4522.0 | 6856.2 | 13.43 | 6032.5 | 0.07 |
| 5 | 6,000 | 5297.8 | 13526.0 | 23.56 | 7200.0 | 0.17 | 5176.2 | 13527.9 | 14.01 | 7200.0 | 0.15 |
| 6 | 10,000 | 5277.6 | 18841.4 | 19.13 | 7200.0 | 0.36 | 4745.4 | 18838.3 | 9.71 | 7200.0 | 0.40 |
| 7 | 20,000 | 5190.0 | 29412.1 | 9.37 | 7200.0 | 0.62 | 3951.8 | 29422.8 | 5.56 | 7200.0 | 0.61 |
| 8 | 50,000 | 5399.8 | 57145.5 | 3.42 | 7200.0 | 0.70 | 5184.6 | 57147.0 | 1.91 | 7200.0 | 0.76 |
| **Avg** | | 3264.1 | 16932.8 | 16.93 | 4837.6 | 0.25 | 3260.9 | 16934.2 | 8.71 | 4776.3 | 0.25 |

## 5.2. New York Case Study

### 5.2.1. Dataset

In this section, we conduct a case study inspired by the New York taxi trip dataset to demonstrate the utility of the proposed model. The dataset is from the NYC Taxi & Limousine Commission (2019). OD pairs are extracted for the morning rush hours (6:00 AM to 10:00 AM) from the New York taxi dataset. Following Wang et al. (2022), we remove OD pairs with a distance of less than 18 kilometers, as RAM is unlikely to be competitive over such short distances. We also exclude OD pairs that appear fewer than ten times in the dataset. Unless otherwise stated, all other parameters remain consistent with those defined in the above setting.

### 5.2.2. Value of incorporating endogenous passenger mode choice

First, we aim to validate the value of incorporating endogenous passenger travel mode choice. This analysis constructs ten New York taxi trip datasets, covering the period from July to November 2025, with data aggregated at half-month intervals. Specifically, we first compute the objective value of the proposed bilevel program. We then evaluate the objective value of the high-point relaxation (HPR) of the bilevel program, which is obtained by removing the objective function of the customers while keeping their constraints. This represents centralized planning. In other words, the HPR represents exogenous demand data instead of endogenous demand as in the bilevel model. This section helps us quantify the benefits of the bilevel modeling approach. Once we obtain the optimal solution of the HPR, we fix this solution $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{n}^{\mathrm{P}}, \boldsymbol{n}^{\mathrm{R}})$, and re-optimize within the bilevel model to compute the corresponding objective value. We refer to this final evaluation as an *ex post* analysis. We report the objective values obtained from the bilevel program, HPR, and the *ex post* analysis in Table 5.

**Table 5 Value of incorporating endogenous passenger mode choice**

| Dataset | $\lvert\mathcal{S}\rvert$ | Bilevel program ($) | HPR ($) | *Ex post* analysis ($) | Benefit (%) |
|---|---|---|---|---|---|
| July 1–15 | 2,636 | 24,557.4 | 57,168.6 | 20,348.5 | 17.1 |
| July 16–31 | 2,973 | 27,000.6 | 60,958.3 | 22,852.6 | 15.4 |
| August 1–15 | 2,742 | 25,085.0 | 57,310.3 | 20,985.6 | 16.3 |
| August 16–31 | 3,239 | 29,495.3 | 65,393.9 | 25,179.5 | 14.6 |
| September 1–15 | 3,154 | 28,351.5 | 63,516.5 | 24,577.1 | 13.3 |
| September 16–30 | 3,255 | 29,597.1 | 64,447.4 | 25,014.2 | 15.5 |
| October 1–15 | 3,214 | 29,185.3 | 65,670.6 | 24,579.7 | 15.8 |
| October 16–31 | 3,567 | 32,170.6 | 68,762.5 | 28,261.5 | 12.2 |
| November 1–15 | 3,171 | 29,184.0 | 62,947.0 | 25,132.7 | 13.9 |
| November 16–30 | 2,844 | 25,236.7 | 59,405.4 | 21,121.4 | 16.3 |
| Average | 3,080 | 27,986.4 | 62,558.1 | 23,805.3 | 15.0 |

The HPR yields the highest objective value among all three reported values, as it represents an optimistic, yet unrealistic, scenario in which passenger mode choices are fully controlled by the operator. In this idealized setting, the operator can freely assign passengers to the most revenue-generating services without considering individual preferences, thereby achieving an upper bound on potential revenues. However, this outcome is not attainable in practice, since passengers are self-interested and will choose the alternative that minimizes their own generalized cost. The HPR thus provides a useful benchmark but serves as an "unconstrained best case" that overlooks the fundamental tension between operator objectives and passenger behavior.

Additionally, it can be observed that the objective value derived from the *ex post* analysis is, on average, 15% lower than that obtained from the bilevel model. This gap highlights the benefit of our bilevel program. Although the HPR solution appears favorable to the operator, it does not account for the fact that passengers will not comply with assignments that are not in their best interest. Ignoring passenger travel mode choice in decision-making leads to substantial profit losses in the *ex post* analysis. By embedding this choice mechanism directly into the planning stage, the bilevel model produces outcomes that are both more robust and more reflective of real-world market dynamics.

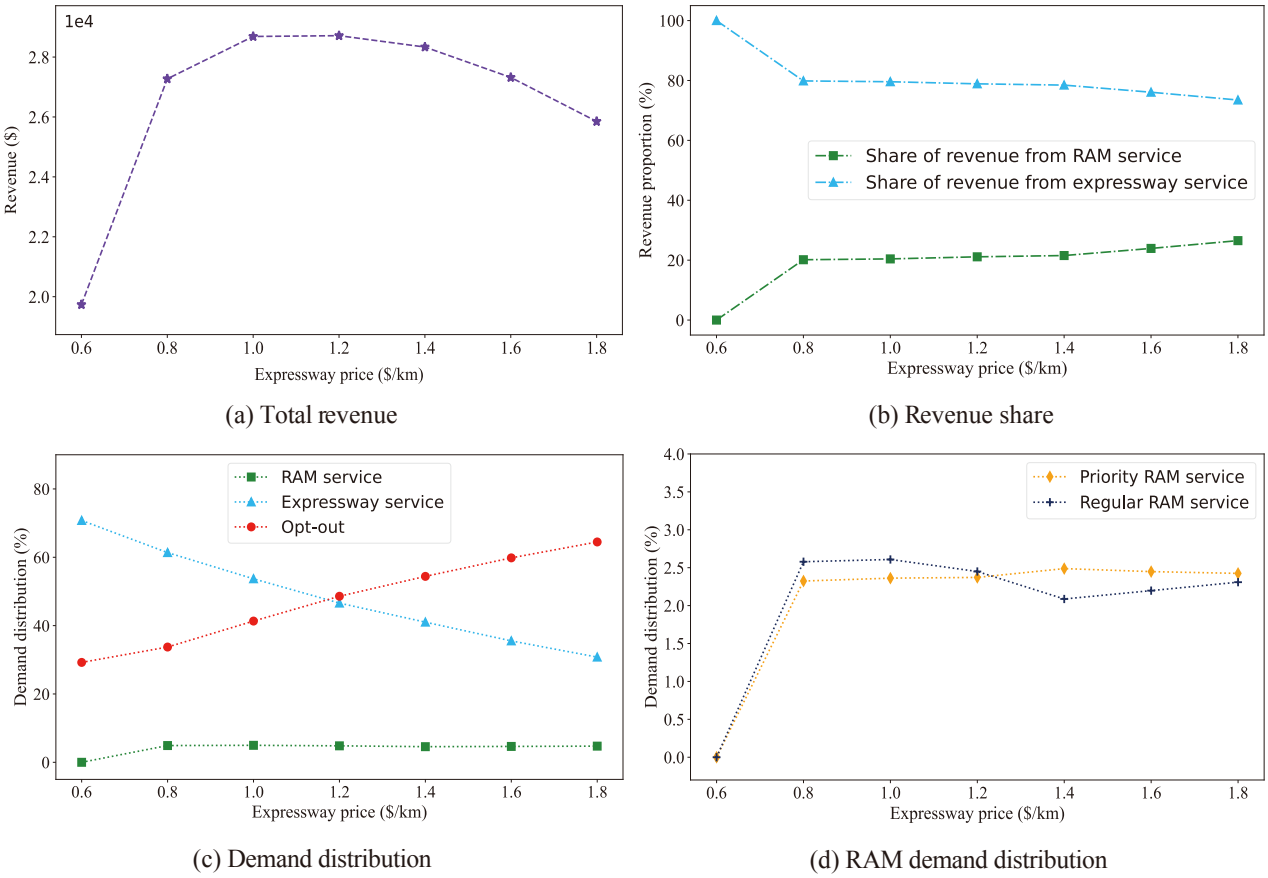
#### 5.2.3. Sensitivity analysis

We conduct sensitivity analysis by solving model `LSP2` with enhancement techniques within a two-hour time limit. The analysis utilized data extracted from the first two weeks of each month from July to November 2025 from the New York Taxi Dataset (NYC Taxi & Limousine Commission 2019).

**(1) The impact of the expressway service price**

The price of expressway taxi services is characterized by volatility. Therefore, we investigate the impact of the price by increasing the unit price per kilometer from \$0.6 to \$1.8 and analyze the resulting changes in revenue and passenger mode choice, which is illustrated in Figure 2.

Increasing the expressway fare raises the generalized cost of taking taxis, prompting some travelers to reconsider their modal choices. For a subset of passengers, the cost gap between expressway service and RAM narrows at relatively low price levels, making RAM comparatively more attractive. This substitution effect is reflected in the initial rise in RAM demand shown in Figure 2(c). Simultaneously, the sharp rise in total revenue depicted in Figure 2(a) indicates that the operator successfully captures these higher-margin RAM trips, forming the left side of the inverted U-shaped curve. However, this positive trend does not persist indefinitely. As the expressway price continues to rise, the generalized cost of expressway taxis becomes extremely high. Many price-sensitive travelers, rather than upgrading to RAM, turn to opt-out modes. This phenomenon is evident in




(a) Total revenue

(b) Revenue share

(c) Demand distribution

(d) RAM demand distribution

**Figure 2 The impact of the expressway service price.**

the plateauing of RAM demand in Figure 2(c) and the eventual decline in total revenue in Figure 2(a). The revenue share analysis in Figure 2(b) further confirms that while RAM can absorb some of the displaced demand, it cannot fully compensate for the loss of expressway taxi users. Most price-sensitive passengers perceive RAM as too expensive and ultimately exit the operator. At this point, the total revenue curve transitions to the downward slope of the inverted U-shaped curve, with revenue decreasing as expressway service price continues to rise.

Figure 2(d) illustrates the demand trends for the two RAM services as expressway service prices vary. As the expressway taxi price increases, the demand for both regular and priority services first rises sharply and then stabilizes. This initial increase is primarily driven by passengers switching from the expressway to regular RAM, as indicated by the steeper rise of regular demand at low price levels. This pattern reflects a substitution effect where cost-sensitive travelers first seek the more affordable regular RAM option. As the expressway price continues to rise beyond the crossover point, priority RAM demand stabilizes and eventually exceeds regular demand. The relative insensitivity of priority demand to further price increases suggests that priority passengers are

less price-sensitive. They value the higher comfort and time savings of priority service, and once the expressway becomes sufficiently expensive, their mode choice remains largely unchanged.

**(2) The impact of the budget**

We next examine the impact of the investment budget. We increase the budget and explore its effect on total revenue and investment scale, measured by the number of vertiports and active aerial routes (denoted as #VP and #AE, respectively). As shown in Table 6, at lower budget levels, the infrastructure capacity is limited, as reflected by the relatively small number of vertiports and aerial routes. Consequently, the RAM service can only accommodate a small fraction of total demand, and the majority of passengers continue to rely on expressway taxis or opt out of the operator altogether. This is evident in Figure 3(c), where the share of RAM demand remains low, and the opt-out proportion is relatively high. Additionally, from Table 6, the apparent imbalance between the million-dollar investment budget and the reported revenue reflects a difference in time scale. The budget is an upfront capital investment, whereas the revenue is measured for a single operating period; repeated operations over multiple periods and many days can generate cumulative revenue sufficient to recover the investment over time

**Table 6 The impact of budget on infrastructure scale**

| Budget (million USD) | 0 | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|---|
| Revenue (USD) | 22,537.8 | 24,083.3 | 25,247.1 | 26,295.0 | 27,272.6 | 28,155.4 | 29,017.8 |
| #VP | 0 | 3.6 | 5.4 | 5.8 | 6.4 | 7.6 | 8.0 |
| #AR | 0 | 5.8 | 11.8 | 14.2 | 17.0 | 25.0 | 30.0 |

As the investment budget increases, capacity expansion leads to a gradual increase in both RAM revenue share and RAM demand, as shown in Figure 3(a) and (c). Notably, the additional RAM ridership primarily comes from passengers who would have otherwise opted out of traveling altogether, as can be observed in Figure 3(c). Meanwhile, expressway demand remains relatively stable.

In Figures 3(b) and (d), as the investment budget increases, both the priority RAM revenue share and its demand distribution exceed those of the regular RAM service. This shift occurs because the increased budget allows for the deployment of more VTOL aircraft dedicated to RAM services. When the budget is constrained, operators tend to favor the regular RAM service, which accommodates ride-sharing and utilizes vehicles more efficiently, over operating individual vehicles for the priority service. As capacity expands, the operator can increasingly accommodate the more flexible and higher-margin priority service.

**(3) The impact of the vertiport capacity**

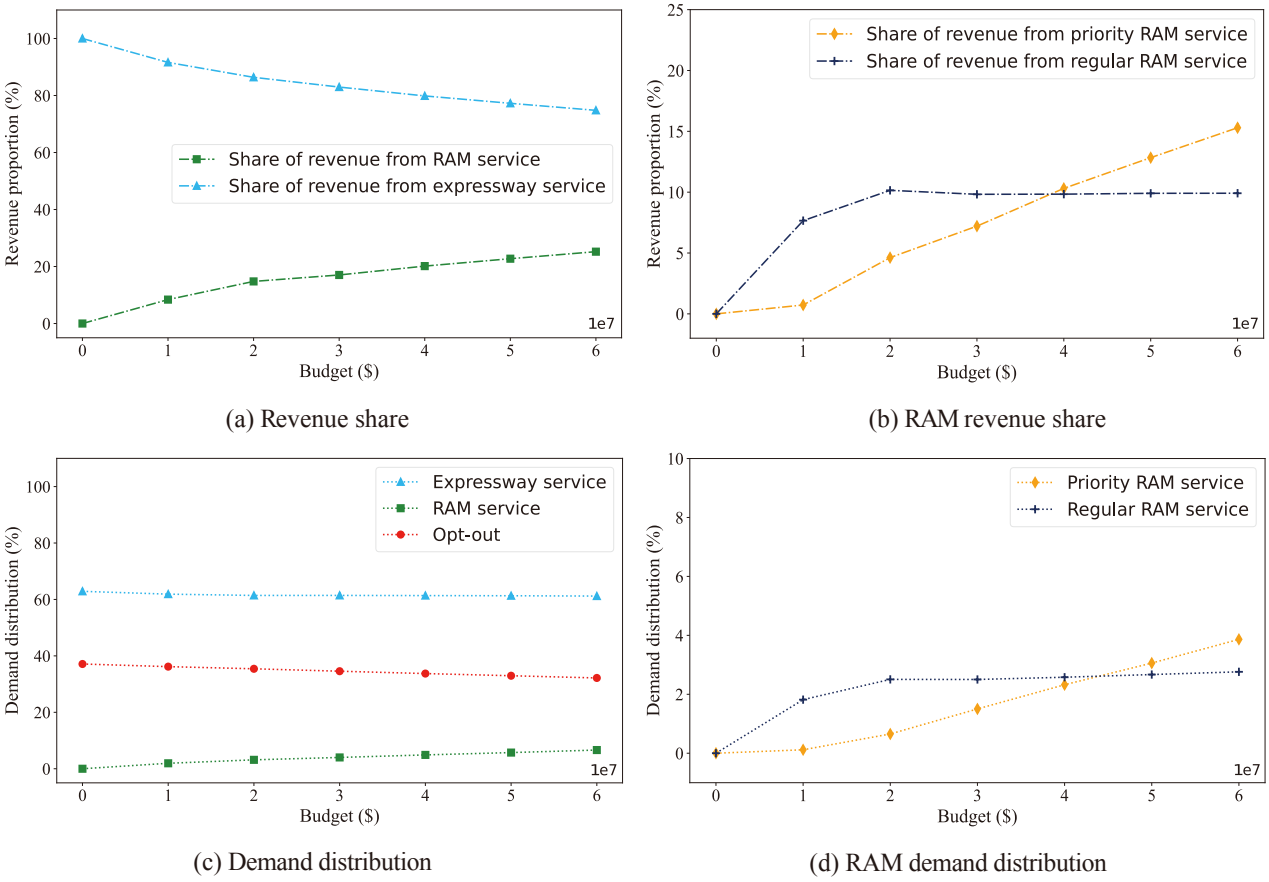



**Figure 3 The impact of the budget.**

We next analyze the impact of vertiport capacity, which can be constrained by factors such as weather conditions that affect the throughput capacity of vertiports. We vary the vertiport capacity from 10 to 60. Figure 4(a) shows that as vertiport capacity increases, system revenue rises modestly. This is because capacity expansion primarily releases potential demand that was previously suppressed by capacity constraints, leading to a slight increase in the total number of passengers served by RAM. Figure 4(b) illustrates that as vertiport capacity increases, the number of vertiports constructed and the number of opened aerial routes both decrease. This is because higher individual vertiport capacity reduces the number of facilities required to accommodate the same level of demand. Notably, it can be observed that revenue share and demand distribution do not change significantly from Figures 4(c) and (d). This indicates that the overall demand distribution is relatively insensitive to such variations.

**(4) The impact of the passenger composition**

In this section, we analyze the impact of passenger composition. Passengers are categorized into business travelers and leisure travelers, who have distinct VOT. Therefore, we increase the percentage of leisure travelers from 20% to 80% and analyze the resulting effects. As shown in

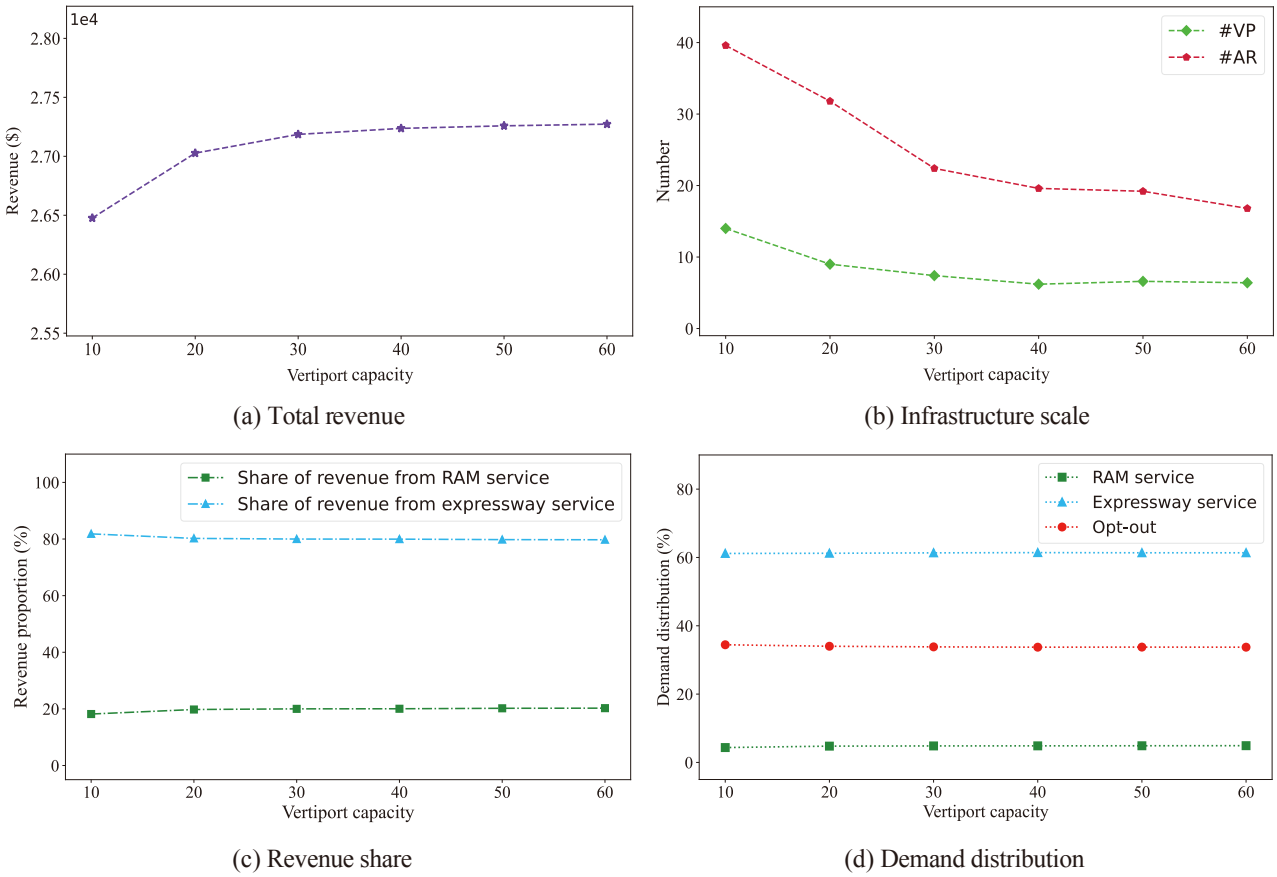
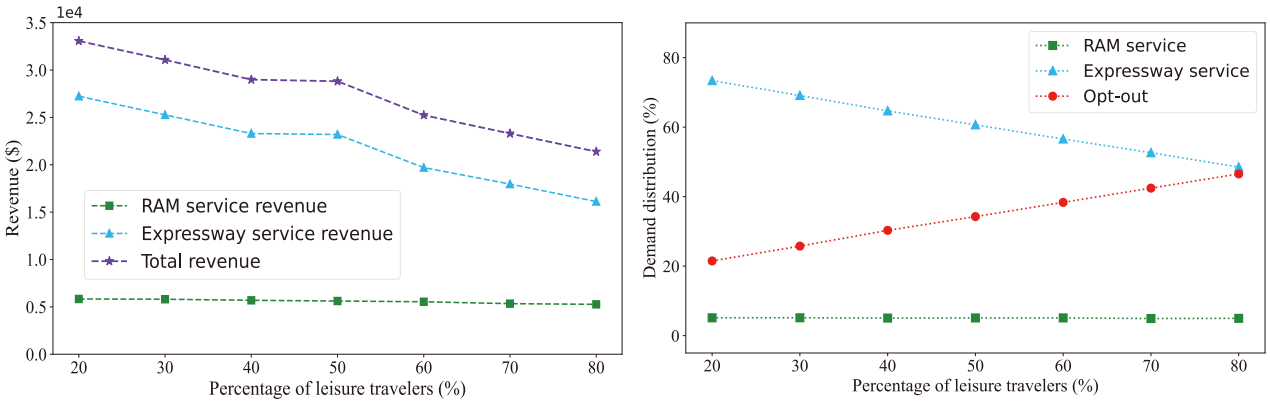



**Figure 4 The impact of the vertiport capacity.**

Figure 5(a), as the proportion of leisure travelers increases, the total revenue of the operator decreases significantly, the revenue of the expressway service also declines substantially, while the revenue of RAM services shows a slight decrease. The reason for this phenomenon can be observed in Figure 5(b), where the proportion of passengers opting out increases significantly as the share of leisure travelers rises.

At a deeper level, this phenomenon arises because different groups of passengers have different VOT, which in turn leads to distinct behavioral patterns. Business travelers have a higher VOT, and their travel decisions prioritize timeliness; they are more inclined to choose expressway and RAM services first. In contrast, leisure travelers are less time-sensitive and unwilling to pay high prices for RAM or expressway taxi services, so they tend to choose the opt-out option. This shift in passenger composition directly leads to the decline in overall operator revenue. It is worth noting that the decline in RAM service revenue in Figure 5(a) is significantly smaller than that of expressway revenue, indicating that the RAM passenger group exhibits relatively high stability. This is because passengers who choose RAM services inherently have specific service preferences, which are difficult to fully substitute with other transportation modes. Therefore, even as a large




(a) Revenue (b) Demand distribution

**Figure 5 The impact of the passenger composition.**

number of passengers opt out, the core passenger base of RAM services remains relatively intact. Overall, to maintain the profitability of the RAM system, the development site should have a certain proportion of business travelers, rather than being located in leisure-oriented areas.

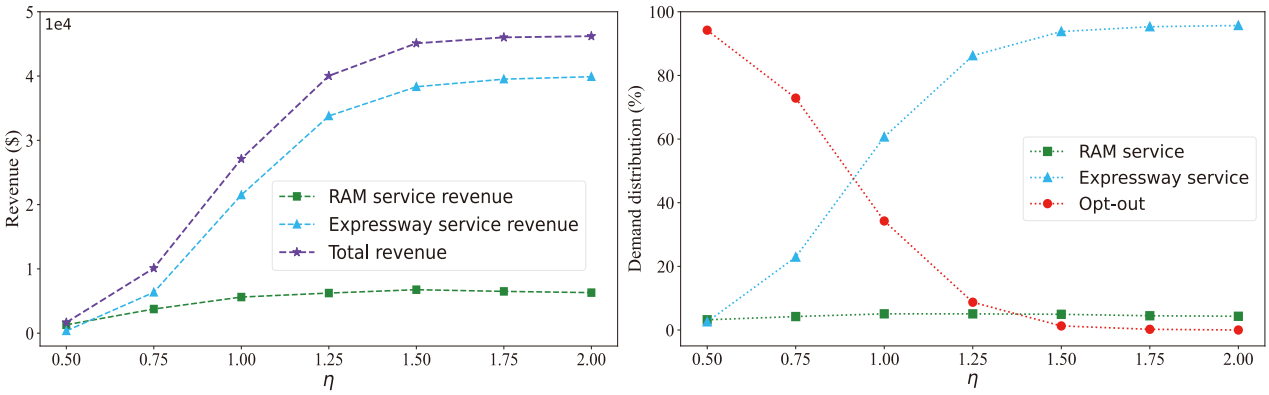
**(5) The impact of the opt-out disutility cost**

Next, we analyze the impact of the opt-out disutility cost by scaling it with a coefficient $\eta$. Specifically, the disutility cost of the opt-out option ranges from $\$50\eta$ to $\$70\eta$ for leisure travelers and from $\$70\eta$ to $\$100\eta$ for business travelers. We vary $\eta$ from 0.5 to 2.0 and examine the resulting effects.

As shown in Figure 6(a), as the opt-out disutility cost increases, both total operator revenue and expressway service revenue increase significantly, while RAM service revenue also exhibits a certain degree of growth. When the opt-out disutility cost increases, the attractiveness of external alternatives diminishes, and passengers become more inclined to remain within the operator and choose either RAM services or expressway services, which can be found in Figure 6(b). It is worth noting that the impact of the opt-out disutility cost varies across different service types. The revenue increase for expressway services is the most pronounced, indicating a higher sensitivity to this cost. In contrast, RAM services, with their differentiated service positioning and relatively stable passenger base, experience a more moderate revenue increase, though they still benefit from the overall retention of passengers.

**(6) The impact of the transfer time**

RAM is a multimodal transportation system. Consequently, transfer time plays a critical role in service performance. Therefore, we conduct a sensitivity analysis by varying the transfer time from 0 to 10 minutes to evaluate its impact. From Figure 7(a), we can find that the total revenue and the revenue from RAM services both gradually decline.



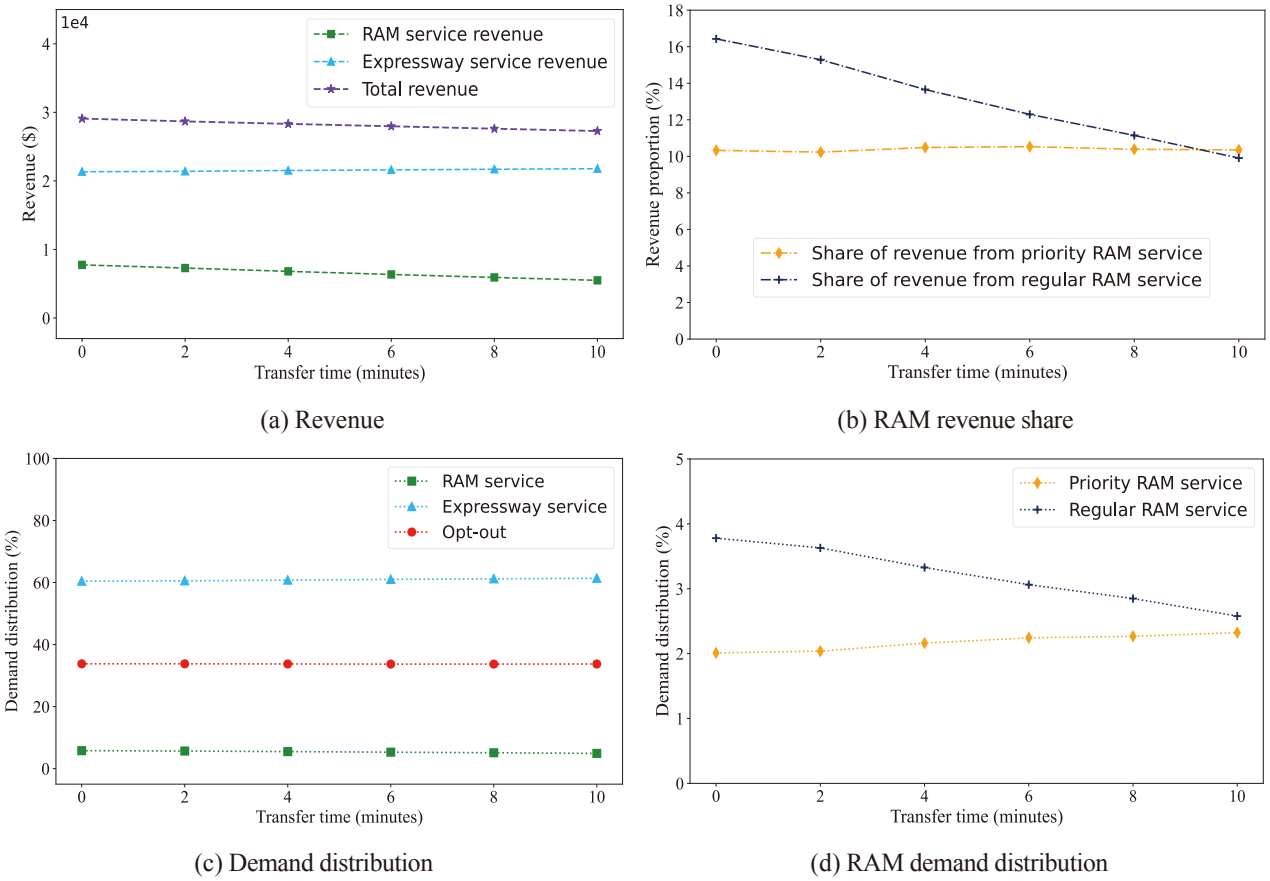

(a) Revenue (b) Demand distribution

**Figure 6 The impact of the opt-out disutility cost.**

Figure 7(b) reveals that the number of priority service users remains stable and largely unaffected by changes in transfer time, while the number of regular users declines sharply. This divergence can be attributed to differences in passenger preferences and tolerance for transfer time. Priority users are more tolerant of longer transfer times because they continue to value the time savings and convenience benefits that RAM provides. In contrast, regular users, being more cost-sensitive, are less willing to accept extended transfer times. However, a small portion of regular RAM users switch to the priority RAM service, as the remaining time savings of RAM become more valuable to them relative to the increased transfer time. Meanwhile, other regular users find RAM no longer affordable or attractive due to the longer transfer time and thus shift to the alternatives. Reducing transfer time could therefore enhance RAM competitiveness by expanding its potential customer base and strengthening its appeal to time-sensitive travelers, which in turn supports the long-term sustainability of RAM operations.

**(7) The impact of the average ground traffic speed**

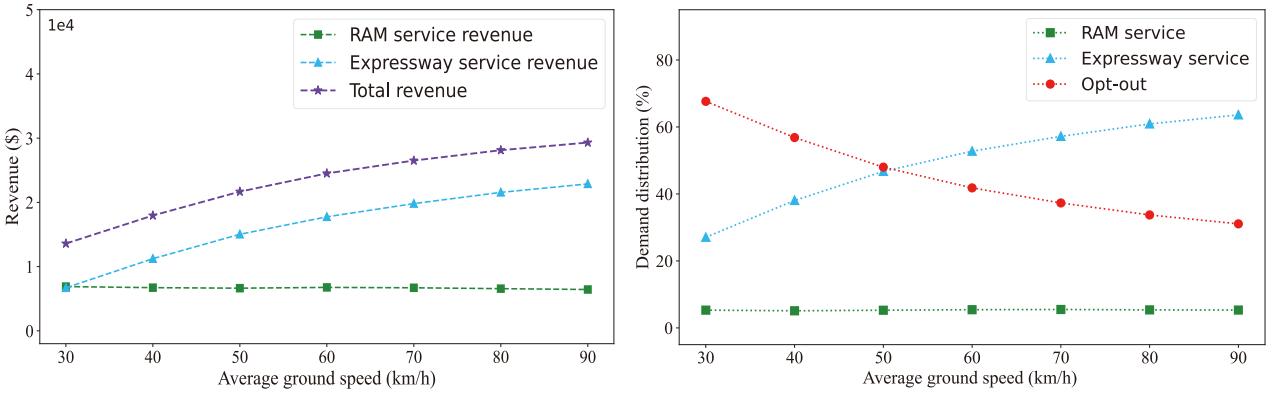
We now discuss the impact of average ground traffic speed. We set the transfer time of the RAM service to five minutes. The ground speed is increased from 30 km/h to 90 km/h in increments of 10 km/h, and the results are shown in Figure 8. As shown in Figure 8(a), when the average ground traffic speed increases, both total operator revenue and revenue from the expressway service increase significantly, while revenue from RAM services decreases slightly. This pattern can be explained by Figure 8(b): as ground traffic speed increases, more passengers choose the expressway service, fewer passengers opt out, and the number of RAM passengers remains relatively stable. From a managerial perspective, the RAM service should not be positioned as a direct competitor to expressway taxi in regions with high average traffic speeds. Instead, its value proposition is most compelling in congested urban environments where ground speeds are low, and the time savings offered by RAM are substantial.




(a) Revenue (b) RAM revenue share

(c) Demand distribution (d) RAM demand distribution

**Figure 7** **The impact of the transfer time.**

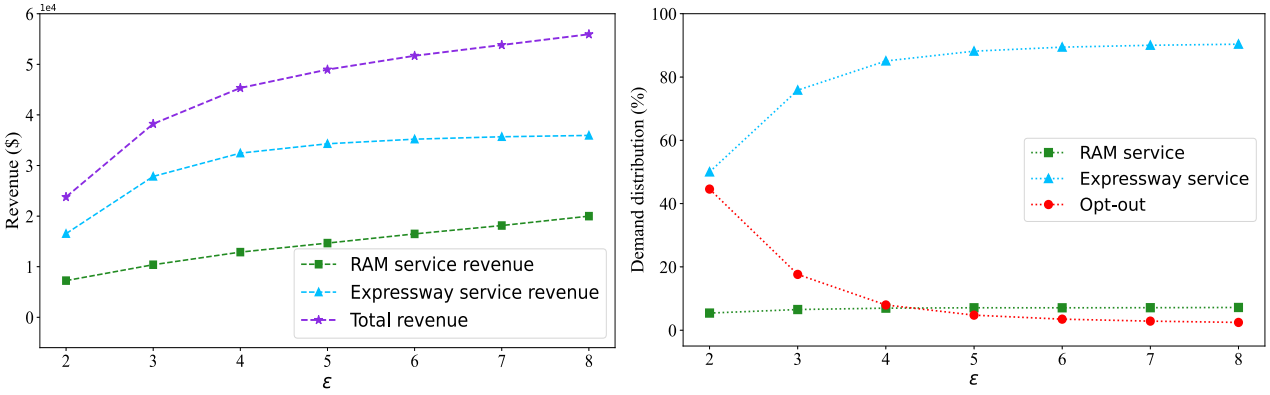



(a) Revenue (b) Demand distribution

**Figure 8** **The impact of the average ground speed.**

### (8) The impact of the VOT

Finally, we discuss the impact of passenger VOT. We scale the VOT by a coefficient $\varepsilon$ based on the settings in Section 5.1.1. Specifically, the VOT of the opt-out option ranges from \$31.0$\varepsilon$ to




(a) Revenue (b) Demand distribution

**Figure 9 The impact of the VOT.**

$46.5\varepsilon$ for leisure travelers and from $50.6\varepsilon$ to $75.8\varepsilon$ for business travelers, where $\varepsilon$ varies from 2 to 8. Since the opt-out utility $u_s^{\mathrm{O}}$ already incorporates travel time costs, scaling passenger VOT requires correspondingly scaling $u_s^{\mathrm{O}}$ to maintain internal consistency. Accordingly, in this sensitivity analysis, we update $u_s^{\mathrm{O}}$ to $[50\varepsilon \times a,\ 70\varepsilon \times a]$ for personal travel passengers and to $[70\varepsilon \times b,\ 100\varepsilon \times b]$ for business travelers, where we assume that $a \sim \mathrm{Uniform}(0.5, 0.6)$ and $b \sim \mathrm{Uniform}(0.6, 0.7)$.

As shown in Figure 9(a), as the VOT increases, total operator revenue as well as revenues from both expressway and RAM services increase significantly. This phenomenon can be explained by Figure 9(b): with a higher VOT, the number of passengers choosing the expressway service increases substantially, the number choosing RAM services increases slightly, while the number opting out decreases dramatically, eventually falling below the number choosing RAM services. From a managerial perspective, several insights emerge. First, targeting regions or time windows with a higher concentration of time-sensitive passengers (e.g., business districts during peak hours) can substantially improve the operator's revenue performance. When entering markets with a high proportion of leisure travelers, the revenue potential may be limited. Second, the RAM system should be positioned as a premium service for high-VOT passengers, as these travelers are more likely to value its time-saving benefits and less likely to opt out.

## 6. Conclusion

This paper proposes a bilevel optimization framework for the Regional Air Mobility Service Network Design and Pricing (RAM-SNDP) problem faced by a multimodal operator entering the RAM market, considering endogenous passenger travel mode choice. The proposed model captures the strategic interaction between the operator and passengers over a sequential decision timeline: the operator first commits to costly upfront investments (vertiport location, aerial route activation,

and fleet deployment) and sets service prices, while passengers then choose among regular RAM, priority RAM, expressway taxi, and opt-out alternatives based on their generalized travel cost.

From a methodological perspective, we reformulate the bilevel program into an equivalent single-level program using duality theory. To enhance computational tractability, we derive the best response sets for passengers and develop several acceleration techniques, including variable reduction strategies and valid inequalities that strengthen the formulation. Computational experiments on instances of varying scale demonstrated the effectiveness of these enhancement techniques in improving solution efficiency. Beyond validating the proposed framework, the numerical experiments based on the New York taxi trip dataset provide several managerial insights for RAM deployment. First, accounting for endogenous passenger travel mode choice is critical, as ignoring endogenous mode choice leads to a suboptimality loss of roughly 15% in realized revenue. Second, the sensitivity analysis reveals a substitution relationship between expressway services and RAM: moderate increases in expressway taxi prices can shift some time-sensitive travelers to RAM, while excessive price increases may instead reduce overall operator demand. Third, RAM demand remains relatively stable across different passenger composition sets and vertiport capacities, indicating that it mainly serves a niche market of travelers with high values of time and strong preferences for travel reliability. Finally, development sites should include a certain proportion of high-value-of-time business travelers, rather than being located in leisure-oriented areas.

Several directions for future research merit attention. First, the current model adopts a deterministic generalized-cost representation of passenger choice. This provides a tractable benchmark for a capacity-constrained RAM design problem, but it abstracts from unobserved taste heterogeneity and richer stochastic substitution patterns. Extending the lower level to stochastic or random-utility choice formulations, potentially with scarcity-related service penalties, would provide a richer behavioral foundation. Second, the model considers a static planning horizon, but RAM operations involve dynamic demand patterns and vehicle repositioning decisions over time. Developing a time-expanded formulation that captures these temporal dynamics would enhance operational realism. Finally, competition among multiple multimodal operators is an important consideration for mature markets, suggesting the need for equilibrium models that capture strategic interactions among multiple service operators.

## References

Arslan, O., Jabali, O., & Laporte, G. (2018). Exact solution of the evasive flow capturing problem. *Operations Research*, 66(6), 1625-1640.

André, N., & Hajek, M. (2019). Robust environmental life cycle assessment of electric VTOL concepts for urban air mobility. *AIAA Aviation 2019 Forum* (p. 3473).

Bekli, S., Boyacı, B., & Zografos, K. G. (2021). Enhancing the performance of one-way electric carsharing systems through the optimum deployment of fast chargers. *Transportation Research Part B: Methodological*, 152, 118-139.

Bracken, J., & McGill, J. T. (1973). Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1), 37-44.

Brunelli, M., Ditta, C. C., & Postorino, M. N. (2023). New infrastructures for urban air mobility systems: A systematic review on vertiport location and capacity. *Journal of Air Transport Management*, 112, 102460.

Brotcorne, L., Labbé, M., Marcotte, P., & Savard, G. (2008). Joint design and pricing on a network. *Operations Research*, 56(5), 1104-1115.

Baloch, G., & Gzara, F. (2020). Strategic network design for parcel delivery with drones under competition. *Transportation Science*, 54(1), 204-228.

Chen, L., Wandelt, S., Dai, W., & Sun, X. (2022). Scalable vertiport hub location selection for air taxi operations in a metropolitan region. *INFORMS Journal on Computing*, 34(2), 834-856.

Chen, Y., & Liu, Y. (2023). Integrated optimization of planning and operations for shared autonomous electric vehicle systems. *Transportation Science*, 57(1), 106-134.

Çelebi, D., Yörüsün, A., & Işık, H. (2018). Bicycle sharing system design with capacity allocations. *Transportation Research Part B: Methodological*, 114, 86-98.

Delaware Aeronautics (2024). Advanced Air Mobility (AAM) Future Outlook for Delaware. Accessed May 23, 2026, https://deldot.gov/Programs/aeronautics/pdfs/sys_plan/Advanced%20Air%20Mobility%20(AAM)%20Future%20Outlook%20Study.pdf.

Federal Aviation Administration (2016). Treatment of time. Accessed April 8, 2026, https://www.faa.gov/sites/faa.gov/files/regulations_policies/policy_guidance/benefit_cost/econ-value-section-1-tx-time.pdf.

Hartleb, J., Schmidt, M., Huisman, D., & Friedrich, M. (2023). Modeling and solving line planning with mode choice. *Transportation Science*, 57(2), 336-350.

Jin, Z., Ng, K. K., Zhang, C., Wu, L., & Li, A. (2024) Integrated optimisation of strategic planning and service operations for urban air mobility systems. *Transportation Research Part A: Policy and Practice*, 183, 104059.

Joby Aviation (2026). Joby brings electric air taxis to New York City in week-long flight campaign. Accessed May 14, 2026, https://www.jobyaviation.com/news/joby-brings-electric-air-taxis-to-new-york-city-in-week-long-flight-campaign.

Justin, C. Y., Payan, A. P., & Mavris, D. N. (2022). Integrated fleet assignment and scheduling for environmentally friendly electrified regional air mobility. *Transportation Research Part C: Emerging Technologies*, 138, 103567.

Li, S., Zhang, T., Xiao, Y., & Li, D. (2025). On-demand ridesharing based on dynamic scheduling in urban air mobility. *Transportation Research Part C: Emerging Technologies*, 175, 105111.

Li, Q., Üster, H., & Zhang, Z. H. (2023). A bilevel model for robust network design and biomass pricing under farmers' risk attitudes and supply uncertainty. *Transportation Science*, 57(5), 1296-1320.

Lim, E., & Hwang, H. (2019). The selection of vertiport location for on-demand mobility and its application to Seoul metro area. *International Journal of Aeronautical and Space Sciences*, 20(1), 260-272.

Labbé, M., Marcotte, P., & Savard, G. (1998). A bilevel model of taxation and its application to optimal highway pricing. *Management Science*, 44(12-part-1), 1608-1622.

NYC Taxi & Limousine Commission. (2019). TLC trip record data. Accessed April 8, 2026, https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Park, S., Kim, H., Kang, J., Lee, J., Kim, Y., Joh, C. H., & Kim, J. (2025). Identifying potential market of advanced ride-hailing services with taxi-to-taxi transfer using a latent class choice model. *Transportation Research Part A: Policy and Practice*, 198, 104543.

Qi, Q., Rasouli, S., & Feng, T. (2026). Modeling the effects of additional crowd-sourced delivery service on transportation mode choice in mobility-as-a-service. *Travel Behaviour and Society*, 42, 101159.

Rajendran, S., & Zack, J. (2019). Insights on strategic air taxi network infrastructure locations using an iterative constrained clustering approach. *Transportation Research Part E: Logistics and Transportation Review*, 128, 470-505.

Rath, S., & Chow, J. Y. (2022). Air taxi skyport location problem with single-allocation choice-constrained elastic demand for airport access. *Journal of Air Transport Management*, 105, 102294.

Richard, P. (2024). EHang Prices its EH216-S eVTOL at RMB2.4 million in China. Accessed April 8, 2026, https://flightplan.forecastinternational.com/2024/02/05/ehang-prices-its-eh216-s-evtol-at-rmb2-4-million-in-china/.

Straubinger, A., Verhoef, E. T., & De Groot, H. L. (2021). Will urban air mobility fly? The efficiency and distributional impacts of UAM in different urban spatial structures. *Transportation Research Part C: Emerging Technologies*, 127, 103124.

Sun, X., Wandelt, S., & Stumpf, E. (2018). Competitiveness of on-demand air taxis regarding door-to-door travel time: A race through Europe. *Transportation Research Part E: Logistics and Transportation Review*, 119, 1-18.

Uber Newsroom (2026). Get ready for takeoff with Uber and Joby. Accessed April 8, 2026, https://www.uber.com/us/en/newsroom/uber-air/.

Weng, C., Chen, C., Tan, J., Pan, T., & Zhong, R. (2026). Real-time traffic simulation and management for large-scale urban air mobility: Integrating route guidance and collision avoidance. *Transportation Research Part C: Emerging Technologies*, 183, 105477.

Wang, K., Alexandre, J., & Vikrant V. (2022). Vertiport planning for urban aerial mobility: An adaptive discretization approach. *Manufacturing & Service Operations Management*, 24(6), 3215-3235.

Willey, L. C., & Salmon, J. L. (2021). A method for urban air mobility network design using hub location and subgraph isomorphism. *Transportation Research Part C: Emerging Technologies*, 125, 102997.

Wolsey, L. A., & Nemhauser, G. L. (1999). Integer and combinatorial optimization. New York: *John Wiley & Sons*.

Wei, K., Vaze, V., & Jacquillat, A. (2020). Airline timetable development and fleet assignment incorporating passenger choice. *Transportation Science*, 54(1), 139-163.

Wei, K., Vaze, V., & Jacquillat, A. (2022). Transit planning optimization under ride-hailing competition and traffic congestion. *Transportation Science*, 56(3), 725-749.

Ye, X., Garikapati, V. M., You, D., & Pendyala, R. M. (2017). A practical method to test the validity of the standard Gumbel distribution in logit-based multinomial choice models of travel behavior. *Transportation Research Part B: Methodological*, 106, 173-192.

Zhang, J., Timmermans, H., Borgers, A., & Wang, D. (2004). Modeling traveler choice behavior using the concepts of relative utility and relative interest. *Transportation Research Part B: Methodological*, 38(3), 215-234.

Zahn, D., Naru, R., Guion, A., & Eggum, S. (2023). UAM instrument flight procedure design and evaluation in the joby flight simulator. Accessed April 8, 2026, https://ntrs.nasa.gov/citations/20230003478.

Zhao, Y., & Feng, T. (2025). Commuter choice of UAM-friendly neighborhoods. *Transportation Research Part A: Policy and Practice*, 192, 104338.

## Appendix A: Notations

**Table A.1 Notations**

| Notation | Definition |
|---|---|
| **Sets** | |
| $\mathcal{S}$ | Set of passengers. |
| $\mathcal{V}$ | Set of candidate vertiports. |
| **Parameters** | |
| $f_k$ | Infrastructure-related cost for constructing and maintaining vertiport $k$. |
| $c_{kh}$ | Unit opening cost of aerial route from vertiport $k$ to vertiport $h$. |
| $l_s^{\mathrm{E}}$ | Inconvenience cost when passenger $s$ travels via expressway. |
| $l_s^{\mathrm{P}}$ | Inconvenience cost when passenger $s$ travels via priority RAM service. |
| $l_s^{\mathrm{R}}$ | Inconvenience cost when passenger $s$ travels via regular RAM service. |
| $u_s^{\mathrm{O}}$ | Disutility cost when passenger $s$ chooses other travel modes. |
| $l_{khs}$ | Travel time when passenger chooses RAM service via vertiport $k$ and vertiport $h$ for passenger $s$. |
| $t_s$ | Travel time when passenger $s$ chooses expressway service. |
| $d^{\mathrm{R}}$ | Unit deployment cost of the eVTOL utilised to provide regular RAM service. |
| $d^{\mathrm{P}}$ | Unit deployment cost of the eVTOL utilised to provide priority RAM service. |
| $r_s$ | Price of expressway travel for passenger $s$. |
| $b$ | Number of seats of the eVTOL. |
| $B$ | Economic budget of investments. |
| $g_k$ | Service capacity of vertiport $k$. |
| $Q_{kh}$ | Low-altitude airspace management capacity of aerial route $(k,h)$. |
| $w_s$ | VOT of passenger $s$. |
| $\underline{p}_{kh}$ | Minimum price of regular RAM service between vertiports $(k,h)$. |
| $\underline{q}_{kh}$ | Minimum price of priority RAM service between vertiports $(k,h)$. |
| **Variables** | |
| $x_k$ | 1, if vertiport $k$ is open; 0, otherwise. |
| $z_{kh}$ | 1, if aerial route between vertiport $k$ and vertiport $h$ is operational; 0, otherwise. |
| $n_{kh}^{\mathrm{P}}$ | Number of eVTOLs deployed to aerial route between vertiport $k$ and vertiport $h$ for priority RAM service. |
| $n_{kh}^{\mathrm{R}}$ | Number of eVTOLs deployed to aerial route between vertiport $k$ and vertiport $h$ for RAM regular RAM service. |
| $p_{kh}$ | The price of regular RAM service from vertiport $k$ to vertiport $h$. |
| $q_{kh}$ | The price of priority RAM service from vertiport $k$ to vertiport $h$. |
| $\hat{m}_{khs}^{\mathrm{P}}$ | 1, if passenger $s$ is provided priority RAM service from vertiport $k$ to vertiport $h$. |
| $\hat{m}_{khs}^{\mathrm{R}}$ | 1, if passenger $s$ is provided regular RAM service from vertiport $k$ to vertiport $h$. |
| $v_s^{\mathrm{E}}$ | 1, if passenger $s$ chooses to travel via the expressway; 0, otherwise. |
| $v_s^{\mathrm{P}}$ | 1, if passenger $s$ chooses the priority RAM service; 0, otherwise. |
| $v_s^{\mathrm{R}}$ | 1, if passenger $s$ chooses the regular RAM service; 0, otherwise. |
| $v_s^{\mathrm{O}}$ | 1, if passenger $s$ chooses opt-out modes (such as self-driving, train, etc.); 0, otherwise. |
| $m_{khs}^{\mathrm{R}}$ | 1, if passenger $s$ travels via aerial route $(k,h)$ when opting for the regular RAM service; 0, otherwise. |
| $m_{khs}^{\mathrm{P}}$ | 1, if passenger $s$ travels via aerial route $(k,h)$ when opting for the priority RAM service; 0, otherwise. |

## Appendix B: Proof of Proposition 1

The coefficient matrix $\mathbb{A}$ of Constraints (2b)–(2f) can be represented as

$$\mathbb{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)} & \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)} \\ 0 & 0 & 0 & -1 & \mathbf{1}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)} & \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)} \\ 0 & 0 & -1 & 0 & \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)} & \mathbf{1}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)} \\ \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{I}_{|\mathcal{V}|(|\mathcal{V}|-1)\times\mathcal{V}|(|\mathcal{V}|-1)} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times\mathcal{V}|(|\mathcal{V}|-1)} \\ \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times 1} & \mathbf{0}_{|\mathcal{V}|(|\mathcal{V}|-1)\times\mathcal{V}|(|\mathcal{V}|-1)} & \mathbf{I}_{|\mathcal{V}|(|\mathcal{V}|-1)\times\mathcal{V}|(|\mathcal{V}|-1)} \end{bmatrix}$$

where $\mathbf{0}_{m\times n}$ and $\mathbf{1}_{m\times n}$ denote the zero matrix and the matrix of all ones, respectively, and $\mathbf{I}_{n\times n}$ denotes the identity matrix.

We use the Ghouila-Houri characterization of TU matrices (see, e.g., Wolsey and Nemhauser. (1999), Theorem 2.7). A matrix $\mathbb{A}$ is totally unimodular if and only if for every subset $R$ of its rows, there exists a partition of $R$ into two subsets $R_1$ and $R_2$ such that for every column $\ell$ of $\mathbb{A}$, $\sum_{r\in R_1}\mathbb{A}_{r\ell} - \sum_{r\in R_2}\mathbb{A}_{r\ell} \in \{-1,0,1\}$. Let $R$ be an arbitrary non-empty subset of the five rows of $\mathbb{A}$. Define $R_1 = R\cap\{1,2,3\}$ and $R_2 = R\cap\{4,5\}$. For each column type:

- Columns of $v_s^{\mathrm{E}}, v_s^{\mathrm{O}}, v_s^{\mathrm{R}}, v_s^{\mathrm{P}}$: Only rows 1–3 have non-zero entries. The sum over $R_1$ is 0 or $\pm1$; $R_2$ contributes 0. Thus, the differences are all in $\{-1,0,1\}$.
- Columns of $m_{khs}^{\mathrm{R}}$: Row 2 has coefficient 1, row 4 has coefficient 1. Hence $\sum_{R_1}=1$ iff $2\in R$, and $\sum_{R_2}=1$ iff $4\in R$. Their differences are all in $\{-1,0,1\}$.
- Columns of $m_{khs}^{\mathrm{P}}$: Same argument gives difference $\in\{-1,0,1\}$.

With this partition, we observe that for each column of $\mathbb{A}$, the difference between the sum of the elements in $R_1$ and $R_2$ sets is $-1$, 0, or 1. Therefore, the coefficient matrix $\mathbb{A}$ is TU. □

The proof of Proposition 1 is completed.

## Appendix C: Proof of Proposition 2

For simplicity, we let $\hat{\boldsymbol{m}} = (\hat{\boldsymbol{m}}^{\mathrm{P}}, \hat{\boldsymbol{m}}^{\mathrm{R}})$, $\boldsymbol{m} = (\boldsymbol{m}^{\mathrm{P}}, \boldsymbol{m}^{\mathrm{R}})$ and $\boldsymbol{v} = (\boldsymbol{v}^{\mathrm{P}}, \boldsymbol{v}^{\mathrm{R}}, \boldsymbol{v}^{\mathrm{E}}, \boldsymbol{v}^{\mathrm{O}})$. Consider an optimal solution of bilevel program $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{n}^{\mathrm{P}}, \boldsymbol{n}^{\mathrm{R}}, \boldsymbol{p}, \boldsymbol{q}, \hat{\boldsymbol{m}}; \boldsymbol{m}, \boldsymbol{v})$ with $\hat{\boldsymbol{m}} \neq \boldsymbol{m}$, i.e., there exists some $k\in\mathcal{V}, h\in\mathcal{V}, s\in\mathcal{S}$ such that $\hat{m}_{khs}^{\mathrm{R}} > m_{khs}^{\mathrm{R}}$ or $\hat{m}_{khs}^{\mathrm{P}} > m_{khs}^{\mathrm{P}}$. Note that $\hat{\boldsymbol{m}}$ appears only in the follower availability constraints $m_{khs}^{\mathrm{R}} \le \hat{m}_{khs}^{\mathrm{R}}$ and $m_{khs}^{\mathrm{P}} \le \hat{m}_{khs}^{\mathrm{P}}$ (for all $k\in\mathcal{V}, h\in\mathcal{V}, s\in\mathcal{S}$) and in the leader capacity constraints $\sum_{s\in\mathcal{S}}\hat{m}_{khs}^{\mathrm{R}} \le b\, n_{kh}^{\mathrm{R}}$ and $\sum_{s\in\mathcal{S}}\hat{m}_{khs}^{\mathrm{P}} \le n_{kh}^{\mathrm{P}}$ (for all $k\in\mathcal{V}, h\in\mathcal{V}$), and $\hat{\boldsymbol{m}}$ does not appear in the leader's objective.

Tighten the leader inventory by setting $\hat{m}_{khs}^{\mathrm{R}} \leftarrow m_{khs}^{\mathrm{R}}$ and $\hat{m}_{khs}^{\mathrm{P}} \leftarrow m_{khs}^{\mathrm{P}}$ for each $k\in\mathcal{V}, h\in\mathcal{V}, s\in\mathcal{S}$, keeping all other leader variables and the follower decision $(\boldsymbol{m}, \boldsymbol{v})$ unchanged. Leader feasibility is preserved because each capacity left-hand side $\sum_{s\in\mathcal{S}}\hat{m}_{khs}^{\mathrm{R}}$ and $\sum_{s\in\mathcal{S}}\hat{m}_{khs}^{\mathrm{P}}$ weakly decreases, while no other leader constraint depends on $\hat{m}$. The follower decision $(\boldsymbol{m}, \boldsymbol{v})$ remains feasible since the availability constraints become $\boldsymbol{m} \le \hat{\boldsymbol{m}}$ at equality; moreover, since the follower minimizes generalized cost and its feasible set can only shrink when $\hat{\boldsymbol{m}}$ is tightened, $(\boldsymbol{m}, \boldsymbol{v})$ remains optimal. Finally, the leader's objective value is unchanged because it depends on realized choices $\boldsymbol{m}$ (and prices/design variables), not on $\hat{\boldsymbol{m}}$. Hence, we obtain another optimal solution with $\hat{\boldsymbol{m}} = \boldsymbol{m}$, as desired.

The proof of Proposition 2 is completed.

## Appendix D: Proof of Proposition 3

We let $\Psi_s^*$ be the optimal objective value of model (2). The optimal solution set is as

$$\Big\{(v_s^E, v_s^O, v_s^P, v_s^R, m_{khs}^R, m_{khs}^P) \;\Big|\; (2b)-(2f),\ \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(p_{kh}+w_s l_{khs})m_{khs}^{\mathrm{R}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(q_{kh}+w_s l_{khs})m_{khs}^{\mathrm{P}} + (r_s + l_s^{\mathrm{E}} + w_s t_s)v_s^{\mathrm{E}} + u_s^{\mathrm{O}} v_s^{\mathrm{O}} + l_s^{\mathrm{P}} v_s^{\mathrm{P}} + l_s^{\mathrm{R}} v_s^{\mathrm{R}} \le \Psi_s^*\Big\}.$$

We start with the case $s\in\mathcal{S}$ and investigate the fact that there are four feasible solutions for model (2). We let $\hat{\boldsymbol{m}}_s^{\mathrm{R}} = \{\hat{m}_{khs}^{\mathrm{R}} \mid \forall k\in\mathcal{V}, \forall h\in\mathcal{V}\}$ and $\hat{\boldsymbol{m}}_s^{\mathrm{R}} = \{\hat{m}_{khs}^{\mathrm{R}} \mid \forall k\in\mathcal{V}, \forall h\in$

$\mathcal{V}\}$. When passenger $s$ chooses expressway service (i.e., $v_s^E = 1$), we have the feasible solutions as $\chi_1 := \left\{v_s^{\mathrm{E}} = 1, v_s^{\mathrm{O}} = 0, v_s^{\mathrm{P}} = 0, v_s^{\mathrm{R}} = 0, \boldsymbol{m}_s^{\mathrm{P}} = \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)}, \boldsymbol{m}_s^{\mathrm{R}} = \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)}\right\}$. Similarly, if passenger $s$ chooses opt-out service (i.e., $v_s^{\mathrm{O}} = 1$), we have the feasible solutions as $\chi_2 := \left\{v_s^{\mathrm{E}} = 0, v_s^{\mathrm{O}} = 1, v_s^{\mathrm{P}} = 0, v_s^{\mathrm{R}} = 0, \boldsymbol{m}_s^{\mathrm{P}} = \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)}, \boldsymbol{m}_s^{\mathrm{R}} = \mathbf{0}_{1\times|\mathcal{V}|(|\mathcal{V}|-1)}\right\}$.

Additionally, if passenger $s$ selects the priority RAM service (i.e., $v_s^{\mathrm{P}} = 1$), then exactly one vertiport pair $(k,h) \in \mathcal{V}\times\mathcal{V}$ satisfies $m_{khs}^{\mathrm{P}} = 1$, while all other elements in the association vector $\boldsymbol{m}_s^{\mathrm{P}}$ are zero. Consequently, the set of feasible solutions is characterized by $\chi_3(s,k,h) := \left\{v_s^{\mathrm{E}} = 0,\ v_s^{\mathrm{O}} = 0,\ v_s^{\mathrm{P}} = 1,\ v_s^{\mathrm{R}} = 0,\ \boldsymbol{m}_s^{\mathrm{P}} = \boldsymbol{e}_{(k,h)},\ \boldsymbol{m}_s^{\mathrm{R}} = \mathbf{0}\right\}$, where $\boldsymbol{e}_{(k,h)}$ denotes the unit row vector of length $|\mathcal{V}|(|\mathcal{V}|-1)$ with its only nonzero element (equal to 1) at the position corresponding to the specific pair $(k,h)$.

Similarly, if passenger $s$ chooses the regular RAM service (i.e., $v_s^{\mathrm{R}} = 1$), the feasible solution set is given by $\chi_4(s,k,h) := \left\{v_s^{\mathrm{E}} = 0,\ v_s^{\mathrm{O}} = 0,\ v_s^{\mathrm{P}} = 0,\ v_s^{\mathrm{R}} = 1,\ \boldsymbol{m}_s^{\mathrm{P}} = \mathbf{0},\ \boldsymbol{m}_s^{\mathrm{R}} = \boldsymbol{e}_{(k,h)}\right\}$.

Therefore, by substituting the explicit values from $\chi_1$, $\chi_2$, $\chi_3$, and $\chi_4$ into the original objective function, the subproblem $\Psi_s^*$ reduces to selecting the minimum cost among these four mutually exclusive cases:

$$\Psi_s^* = \min\Big\{r_s + l_s^{\mathrm{E}} + w_s t_s,\ u_s^{\mathrm{O}},\ \min_{\substack{k,h\in\mathcal{V}\\ k\neq h}}\left(p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}}\right),\ \min_{\substack{k,h\in\mathcal{V}\\ k\neq h}}\left(q_{kh} + w_s l_{khs} + l_s^{\mathrm{P}}\right)\Big\}.$$

To reformulate this explicit minimum operation into a linear programming form, we apply the Big-M linearization technique. This introduces auxiliary constants $M_5$ and $M_6$ that are sufficiently large. The equivalent linear formulation is:

$$\Psi_s^* = \min\Big\{r_s + l_s^{\mathrm{E}} + w_s t_s,\ u_s^{\mathrm{O}},\ p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}} + M_5(1 - m_{khs}^{\mathrm{R}}),\ q_{kh} + w_s l_{khs} + l_s^{\mathrm{P}} + M_6(1 - m_{khs}^{\mathrm{P}})\Big\}.$$

where the terms involving $M_5$ and $M_6$ act as switches: $p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}}$ (or $q_{kh} + w_s l_{khs} + l_s^{\mathrm{P}}$) is active only when its associated binary variable $m_{khs}^{\mathrm{R}}$ (or $m_{khs}^{\mathrm{P}}$) equals 1; otherwise, the large penalty $M_5$ (or $M_6$) dominates, ensuring that particular choice is not selected in the minimization. The proof of Proposition 3 is completed.

## Appendix E: Proof of Proposition 4

The big-M coefficients $M_1^{kh}$ and $M_2^{kh}$ denote the maximum prices for the regular and priority RAM services for air route $(k,h)$, respectively. If the price of a service exceeds a certain threshold, no passenger will select it. Our goal is to identify these threshold values to establish tight bounds for $M_1$ and $M_2$.

We begin by analyzing $M_1^{kh}$. For a passenger $s$, the regular RAM service will not be chosen if its price is too high compared to alternative options. Specifically:

- Compared to the opt-out travel mode, passenger $s$ will not choose regular RAM if the price exceeds $u_s^{\mathrm{O}} - w_s l_{khs} - l_s^{\mathrm{R}}$.
- Compared to the expressway travel mode, passenger $s$ will not choose regular RAM if the price exceeds $r_s + l_s^{\mathrm{E}} + w_s t_s - w_s l_{khs} - l_s^{\mathrm{R}}$.

Furthermore, since prices must be non-negative, the effective price threshold for passenger $s$ is the minimum of the two upper bounds derived from the alternatives, truncated at zero. Thus, the maximum acceptable price for passenger $s$ for the regular service from vertiport $k$ to $h$ is $P_s =$

$\min\left\{\max\left\{u_s^{\mathrm{O}} - w_s l_{khs} - l_s^{\mathrm{R}},\ 0\right\},\ \max\left\{r_s + l_s^{\mathrm{E}} + w_s t_s - w_s l_{khs} - l_s^{\mathrm{R}},\ 0\right\}\right\}$. Consequently, if the price is set to the maximum value of $P_s$ across all passengers, i.e., $M_1^{kh} = \max_{s\in\mathcal{S}} P_s$, then no passenger will choose the regular RAM service.

The proof for bounding the value of $M_2^{kh}$ follows a similar line of reasoning. This completes the proof of Proposition 4.

## Appendix F: Proof of Proposition 5

In the dual of the follower's problem (3), the zero vector $(\alpha_s, \beta_s, \boldsymbol{\gamma}_s, \boldsymbol{\theta}_s, \boldsymbol{\kappa}_s) = \mathbf{0}$ constitutes a feasible solution, as every constraint is satisfied when all dual variables are set to zero. The objective value equals 0 under this solution. Considering problem (3) is a maximization problem, its optimal value is at least the value of any feasible solution. Hence, we have $\alpha_s - \sum_{k\in\mathcal{V},h\in\mathcal{V}}\left(\hat{m}_{khs}^{\mathrm{R}}\theta_{khs} + \hat{m}_{khs}^{\mathrm{P}}\kappa_{khs}\right) \geq 0$. From the Constraints (3c), we have $\sum_{k\in\mathcal{V},h\in\mathcal{V}}\left(\hat{m}_{khs}^{\mathrm{R}}\theta_{khs} + \hat{m}_{khs}^{\mathrm{P}}\kappa_{khs}\right) \leq \alpha_s \leq u_s^{\mathrm{O}}$.

Due to the non-negative property, we have $\hat{m}_{khs}^{\mathrm{R}}\theta_{khs} + \hat{m}_{khs}^{\mathrm{P}}\kappa_{khs} \leq \sum_{k'\in\mathcal{V},h'\in\mathcal{V}}\left(\hat{m}_{k'h's}^{\mathrm{R}}\theta_{k'h's} + \hat{m}_{k'h's}^{\mathrm{P}}\kappa_{k'h's}\right)$ for any fixed $(k,h)$. Therefore, we can obtain $\hat{m}_{khs}^{\mathrm{R}}\theta_{khs} + \hat{m}_{khs}^{\mathrm{P}}\kappa_{khs} \leq u_s^{\mathrm{O}}$. Considering $\hat{m}_{khs}^{\mathrm{R}}, \hat{m}_{khs}^{\mathrm{P}} \in \{0,1\}$, we have $\theta_{khs} \leq u_s^{\mathrm{O}}$ and $\kappa_{khs} \leq u_s^{\mathrm{O}}$ for each $\forall k,h \in \mathcal{V}$ and $\forall s \in \mathcal{S}$. In the case when both $\hat{m}_{khs}^{\mathrm{R}}$ and $\hat{m}_{khs}^{\mathrm{P}}$ equal one, the $\theta$ and $\kappa$ are unbounded, and there exist multiple optimal solutions. However, as we consider an optimistic bilevel model, keeping only one of these multiple optimal solutions is sufficient, and the specific $M_1^{kh}$ and $M_2^{kh}$ values we consider here ensure the existence of at least one such optimal solution. This completes the proof of Proposition 5.

## Appendix G: Proof of Proposition 6

From Constraints (7b) and (7c), we have $\Psi_s \leq l_s^{\mathrm{E}} + r_s + w_s t_s$ and $\Psi_s \leq u_s^{\mathrm{O}}$. Therefore, it follows that $\Psi_s \leq p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}} + l_s^{\mathrm{E}} + r_s + w_s t_s$ and $\Psi_s \leq p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}} + u_s^{\mathrm{O}}$. Now we consider Constraint (7d). When $m_{khs}^{\mathrm{R}} = 0$, the constraint becomes $\Psi_s \leq p_{kh} + w_s l_{khs} + l_s^{\mathrm{R}} + M_5^s$. If we set $M_5^s = \min\left\{l_s^{\mathrm{E}} + r_s + w_s t_s,\ u_s^{\mathrm{O}}\right\}$, then the term $M_5$ ensures that the constraint does not become overly restrictive when the binary variable is inactive. A similar reasoning applies to Constraint (7e).

Hence, this completes the proof of Proposition 6.

## Appendix H: Stochastic program

This stochastic program is based on model `LSP2`. We let $\Omega$ denote the scenario set and $\omega \in \Omega$. Then, $S_\omega$ represents the passenger set under scenario $\omega$. Let $\pi_\omega$ denote the occurrence probability of scenario $\omega$. Therefore, we have the following stochastic program.

$$\text{maximize} \sum_{\omega\in\Omega}\pi_w\left\{\sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}_\omega}\rho_{khs} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}\sum_{s\in\mathcal{S}_\omega}\xi_{khs} + \sum_{s\in\mathcal{S}_\omega} r_s v_s^{\mathrm{E}}\right\} \tag{20a}$$

$$\text{subject to} \sum_{k\in\mathcal{V}} f_k x_k + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} d^{\mathrm{R}} n_{kh}^{\mathrm{R}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} d^{\mathrm{P}} n_{kh}^{\mathrm{P}} + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} c_{kh} z_{kh} \leq B, \tag{20b}$$

$$z_{kh} \leq x_k,\ \ k,h \in \mathcal{V}, \tag{20c}$$

$$z_{kh} \leq x_h,\ \ k,h \in \mathcal{V}, \tag{20d}$$

$$n_{kh}^{\mathrm{R}} + n_{kh}^{\mathrm{P}} \leq Q_{kh} z_{kh},\ \ k,h \in \mathcal{V}, \tag{20e}$$

$$\sum_{h\in\mathcal{V}} n^{\mathrm{P}}_{hk} + \sum_{h\in\mathcal{V}} n^{\mathrm{R}}_{hk} + \sum_{h\in\mathcal{V}} n^{\mathrm{P}}_{kh} + \sum_{h\in\mathcal{V}} n^{\mathrm{R}}_{kh} \le g_k x_k,\ \ k\in\mathcal{V}, \tag{20f}$$

$$\sum_{s\in\mathcal{S}_\omega} m^{\mathrm{R}}_{khs} \le b n^{\mathrm{R}}_{kh},\ \ k,h\in\mathcal{V},\omega\in\Omega, \tag{20g}$$

$$\sum_{s\in\mathcal{S}_\omega} m^{\mathrm{P}}_{khs} \le n^{\mathrm{P}}_{kh},\ \ k,h\in\mathcal{V},,\omega\in\Omega, \tag{20h}$$

$$x_k, z_{kh}\in\{0,1\},\ \ k,h\in\mathcal{V}, \tag{20i}$$

$$n^{\mathrm{R}}_{kh}, n^{\mathrm{P}}_{kh}\in\mathbb{Z}^+,\ \ k,h\in\mathcal{V}, \tag{20j}$$

$$p_{kh}, q_{kh}\ge 0,\ \ k,h\in\mathcal{V}, \tag{20k}$$

$$v^{\mathrm{E}}_s + v^{\mathrm{O}}_s + v^{\mathrm{P}}_s + v^{\mathrm{R}}_s = 1, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20l}$$

$$\sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} m^{\mathrm{R}}_{khs} = v^{\mathrm{R}}_s, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20m}$$

$$\sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}} m^{\mathrm{P}}_{khs} = v^{\mathrm{P}}_s, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20n}$$

$$v^{\mathrm{R}}_s, v^{\mathrm{P}}_s, v^{\mathrm{O}}_s, v^{\mathrm{E}}_s, m^{\mathrm{R}}_{khs}, m^{\mathrm{P}}_{khs}\in\{0,1\},\ \ k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20o}$$

$$\Psi_s \le l^E_s + r_s + w_s t_s, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20p}$$

$$\Psi_s \le u^O_s, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20q}$$

$$\Psi_s \le p_{kh} + w_s l_{khs} + l^R_s + M^s_5(1-m^R_{khs}),\quad k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20r}$$

$$\Psi_s \le q_{kh} + w_s l_{khs} + l^P_s + M^s_6(1-m^P_{khs}),\quad k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20s}$$

$$\Psi_s \ge \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(\rho_{khs} + w_s l_{khs} m^{\mathrm{R}}_{khs}) + \sum_{k\in\mathcal{V}}\sum_{h\in\mathcal{V}}(\xi_{khs} + w_s l_{khs} m^{\mathrm{P}}_{khs}) + (r_s + l^{\mathrm{E}}_s + w_s t_s)v^E_s + u^{\mathrm{O}}_s v^{\mathrm{O}}_s + l^{\mathrm{P}}_s v^{\mathrm{P}}_s + l^{\mathrm{R}}_s v^{\mathrm{R}}_s,$$
$$s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20t}$$

$$\Psi_s\in\mathbb{R},\quad s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20u}$$

$$0\le\rho_{khs}\le p_{kh},\ \ k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20v}$$

$$0\le\xi_{khs}\le q_{kh},\ \ k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20w}$$

$$p_{kh} - M^{kh}_1(1-m^{\mathrm{R}}_{khs}) \le \rho_{khs} \le M^{kh}_1 m^{\mathrm{R}}_{khs},\ \ k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega, \tag{20x}$$

$$q_{kh} - M^{kh}_2(1-m^{\mathrm{P}}_{khs}) \le \xi_{khs} \le M^{kh}_2 m^{\mathrm{P}}_{khs},\ \ k,h\in\mathcal{V}, s\in\mathcal{S}_\omega,\omega\in\Omega. \tag{20y}$$