

Highlights

Integrated Prediction and Distributionally Robust Optimization for Air Traffic Management

Haochen Wu, Xinting Zhu, Lishuai Li, Max Z. Li

- Propose an integrated prediction and distributionally robust optimization framework for Air Traffic Management.
- Show that leveraging upstream predictions significantly reduce system-wide delay costs.
- Demonstrate that distributionally robust optimization models improve performance under distribution shifts relative to upstream predictions.

Integrated Prediction and Distributionally Robust Optimization for Air Traffic Management

Haochen Wu^a, Xinting Zhu^b, Lishuai Li^b, Max Z. Li^a

^a*University of Michigan, Aerospace Engineering, 500 S State St, Ann Arbor, 48109, Michigan, United States of America*

^b*City University of Hong Kong
Kowloon Tong, Kowloon Tong
, 83 Tat Chee Ave, Kowloon Tong, 523808, Hong Kong,*

Abstract

Strategic Traffic Management Initiatives (TMIs) such as Ground Delay Programs (GDPs) play a crucial role in mitigating operational costs associated with air traffic demand-capacity imbalances. However, GDPs can only be planned (e.g., duration, delay assignments) with confidence if the future capacities at constrained resources (i.e., airports) are predictable. In reality, such future capacities are uncertain, and predictive models may provide predictions that are vulnerable to errors and distribution shifts. Motivated by the goal of planning optimal GDPs that are *distributionally robust* against airport capacity prediction errors, we study a fully integrated learning-driven optimization framework. We design a deep learning-based prediction model capable of forecasting arrival and departure capacity distributions across a network of airports. We incorporate the predictions into a distributionally robust formulation of the multi-airport ground holding program (DR-MAGHP). Our results demonstrate that DR-MAGHP can achieve up to a 15.6% improvement over the stochastic programming formulation (SP-MAGHP) under airport capacity distribution shifts. We conclude by outlining future research directions aimed at enhancing both the learning and optimization components of the framework.

Keywords: Air traffic management; Ground Delay Programs (GDPs); Airport capacity prediction; Distributionally robust optimization

1. Introduction

Congestion in air transportation systems results from demand-capacity imbalances, often due to airport capacity reductions. Within the US National Airspace System (NAS), the strategic implementation of Traffic Management Initiatives (TMIs) seeks to reduce the operational costs of such imbalances. A prominent example of TMIs is Ground Delay Programs (GDPs), which aims to delay flights on the ground at the origin airport and mitigate costly airborne delays in response to arrival capacity reductions at the destination airport. The optimal implementation of GDPs is the purview of airport ground holding optimization problems, or GHPs. Airport capacities at future time periods play a significant role in GDP implementations. If such capacities are known, the optimal delay allocation decisions can be found by solving GHPs [1]. However, in practice, it is extremely difficult for traffic management decision-makers to ascertain future airport capacities (i.e., Airport Arrival Rates and Airport Departure Rates) due to myriad uncertainties. Such uncertainties stem from environmental (e.g., convective weather predictions [2]) and operational (e.g., runway availability and traffic volume [3]) sources. Although a large variety of prior work focuses on, e.g., airport runway configuration prediction [4], weather impact predictions [2], and TMI implementation predictions [5], all predictions result in a probability distribution over potential outcomes. If such predictions were to be incorrect (e.g., due to distribution shifts or misspecification), the resultant GDP implementation may be sub-optimal.

1.1. Motivation and research problem

In this work, we are motivated by the dual objective of harnessing advances in machine learning (ML) for airport capacity prediction while maintaining a *cautiously optimistic* stance in prescribing ground delay policies. On the one hand, ML models enable data-driven forecasting, which aligns with many future concepts of operations for air traffic management such as the Federal Aviation Administration’s Information-Centric NAS vision [6] and SESAR’s European ATM Master Plan 2025 Edition [7]. On the other hand, robust decision-making remains essential, as such predictions are inherently imperfect.

ML has seen rapid progress and growing applications in aviation, particularly in airport and airspace operations. By leveraging historical and real-time data, ML models have been used to anticipate and mitigate delays, predict traffic flows, estimate fuel consumption and airport throughput, and optimize flight paths [8, 9, 10], thereby enhancing overall airspace efficiency. Programs such as SESAR in

Europe and NextGen in the United States are actively working to integrate ML into advanced air traffic management (ATM) systems [6, 7].

However, the practical use of ML predictions in downstream optimization can be problematic when decision models are highly sensitive to forecast errors. A central challenge is *distribution shift*, i.e., when the underlying data distribution differs between training and deployment [11]. Such shifts often degrade decision-making performance, as illustrated by autonomous vehicles trained in sunny conditions performing poorly in adverse weather [12], recommendation systems facing seasonal changes in user preferences [13], or fraud and spam detection models being circumvented through adversarial concept drift [14]. In the context of airport capacity prediction, distribution shifts may lead to inaccurate predictions and, in turn, suboptimal or even unsafe ground holding policies.

To address these challenges, GDPs formulations must explicitly account for uncertainty in upstream predictions. Traditional optimization approaches include the deterministic multi-airport ground holding problem (MAGHP), which assumes fixed capacities, and the stochastic MAGHP, which relaxes this assumption by modeling capacity as a random variable. In this work, we advance this line of research by developing a *distributionally robust* formulation, termed the distributionally robust multi-airport ground holding problem (DR-MAGHP). Distributionally robust optimization (DRO) seeks solutions that perform well under the worst-case distribution within a predefined *ambiguity set*, which in our case is constructed based on the predicted distribution of airport capacities generated by the ML model. Section 3 provides a rigorous overview of these concepts.

1.2. Contributions

In this work, our contributions are as follows:

1. We develop a deep learning framework to provide practical probabilistic predictions of departure and arrival capacity at each airport. Based on experiments with the 30 major US core airports, our model demonstrates that the framework is effective in predicting capacity distributions across diverse airport environments and operational characteristics.
2. We develop a distributionally robust multi-airport ground holding problem (DR-MAGHP) that incorporates Wasserstein ambiguity sets constructed from upstream distributional predictions. We derive a tractable reformulation of DR-MAGHP and introduce a scenario reduction technique to mitigate the computational challenges posed by the large-scale scenario trees arising from time-series capacity predictions.

3. We perform a comprehensive sensitivity analysis to evaluate the performance of DR-MAGHP under varying degrees of airport capacity reductions, meant to mimic distributional uncertainty and mispredictions. This analysis highlights how the choice of the *size* of the ambiguity sets influences robustness and cost efficiency, thereby providing insights into the trade-offs involved in robust policy design.

We note that a preliminary version of this work has appeared in the 11th International Conference on Research in Air Transportation[15].

2. Literature review

2.1. *Distributional prediction for airport capacity*

Airport capacity can be defined as the maximum sustainable throughput for arriving (the Airport Arrival Rate, or AAR) and departing (the Airport Departure Rate, or ADR) flights [16]. In contrast to declared capacities obtained from theoretical analyses or statistical approaches [17], real-time capacity is dynamic and challenging to predict in advance. It is influenced by several interconnected operational and environmental factors [18]. As the prediction time horizon increases, forecast uncertainty grows as well, rendering accurate long-term predictions difficult [19].

Traditional prediction models in aviation have predominantly relied on deterministic approaches. A sampling of previous work includes analytical approaches such as the Integrated Airport Capacity Model (IACM) [16], and more recently, data-driven approaches such as the AAR Distribution Prediction Model (ADPM) [20]. However, these deterministic methods often fail to capture the inherent uncertainties in air traffic systems, leading to suboptimal decision-making under uncertain conditions.

The emergence of probabilistic and distributional prediction methods has addressed these limitations by explicitly quantifying prediction uncertainties. A non-exhaustive set of recent approaches utilizes probabilistic methods to account for uncertainties from various sources in aircraft trajectory prediction [21], delay predictions [22], and traffic flow prediction [23, 24], thereby enabling more robust air traffic management decisions. Quantile regression techniques have been developed to capture the full delay distribution rather than simple point estimates, with models demonstrating superior performance compared to statistical baselines by learning various quantiles of departure and arrival delay distributions using features available several days before operations during the pretactical phase [25].

Ensemble methods, particularly Variational LSTM models incorporating Monte Carlo Dropout, achieve median absolute errors in delay predictions of 5.8 minutes per day across multiple airports while providing well-calibrated prediction intervals crucial for risk management applications [26]. Bayesian approaches, including Bayesian structural time series models for air traffic demand forecasting [27], demonstrate efficacy in incorporating external factors and handling non-linear demand patterns. Additionally, probabilistic aircraft trajectory prediction approaches using Bayesian Neural Networks [21, 28] can predict aircraft flight paths while quantifying the uncertainties in those predictions, which is particularly valuable for managing weather-related disruptions. Separately, Monte Carlo simulation techniques [29] enable evaluation of post-resolution conflict probabilities, supporting optimization under uncertainty for air traffic conflict resolution.

Few works have been conducted on predicting real-time airport capacities [30], and flight delay distribution predictions [10]. The field continues to evolve with increasing integration of probabilistic methods, advanced deep learning architectures, and multi-source data fusion approaches.

2.2. *Mathematical Programs in Air Traffic Management*

Mathematical programming, particularly integer programming (IP), has been extensively applied in Air Traffic Management (ATM) to address challenges arising from capacity-constrained airspace and airports. One central problem is the Air Traffic Flow Management Problem (ATFMP), which seeks to optimize aircraft flows and reduce congestion while respecting capacity limits. Odoni [31] was among the first to formalize the Flow Management Problem (FMP), modeling it as a network-based formulation to manage congestion across airports, airways, waypoints, and sectors. Building on this, Bertsimas and Patterson [32] developed an IP model for ATFMP that minimizes total delay costs by optimizing flight arrival and departure times within airspace and airport capacity constraints. Addressing dynamic routing under uncertain weather conditions, Bertsimas and Sim [33] proposed a dynamic multicommodity network flow model, which they solved using a Lagrangian generation algorithm. In a related approach, Sun et al. [34] introduced a cell-transmission model that recasts air traffic flow as a linear time-invariant dynamical system, solved using a relaxed integer program to manage sector-level aircraft counts over time.

A key subclass of ATFMP is the Ground Holding Problem (GHP), which specifically addresses delays caused by limited capacity at departure and arrival airports. Unlike ATFMP, which accounts for system-wide capacity constraints, GHP models focus on airport-level restrictions. The Multi-Airport Ground-Holding

Problem (MAGHP) and its variations—both deterministic and stochastic—have been widely studied [1]. Stochastic formulations incorporate two-stage stochastic programming and chance-constrained programming to model airport capacities as random variables [35, 36]. More recent advances include data-driven control frameworks that minimize delay costs while redistributing delays spatially to improve operational resilience [37]. Additional research introduces fairness- and passenger-oriented metrics to ensure equitable treatment of flights across airlines and time slots [38, 39].

Another critical ATM problem is airport slot allocation, especially at highly congested hubs where demand exceeds available runway capacity. This involves assigning takeoff and landing slots to airlines in a manner that adheres to regulatory and operational constraints [40] while balancing efficiency and fairness. Pellegrini et al. introduced the Simultaneous Slot Allocation Problem (SSAP) and proposed metaheuristic algorithms that outperform traditional ILP methods on large-scale instances by efficiently handling inter-airport coordination [41]. They later extended this work through the SOSTA (Simultaneous Optimisation of the airport Slot Allocation) model, which incorporates aircraft turnaround constraints and conforms to European regulatory standards, demonstrating high computational efficiency (i.e., shorter solution times) on real operational data [42]. From a market design perspective, Ball et al. [43] developed a quantity-contingent auction framework in which slot values vary with allocation volume, incorporating constraints to mitigate the risk of market power abuse. Zografos and Jiang [44] introduced a bi-objective optimization model that jointly optimizes schedule efficiency and fairness, providing trade-off analyses under different regulatory regimes. Addressing computational scalability, Ribeiro et al. [45] proposed a large-scale neighborhood search algorithm that delivers near-optimal slot assignments for busy airports with significant runtime improvements over conventional ILP methods.

2.3. Distributionally Robust Optimization

Distributionally robust optimization (DRO) provides a robust framework for decision-making under uncertainty by optimizing decisions against the worst-case distribution within a specified ambiguity set. Delage and Ye [46] introduced a foundational DRO framework based on moment information, where ambiguity sets are defined using means and covariances, and tractable conic reformulations are derived with statistical guarantees. Esfahani and Kuhn [47] extended this approach by defining ambiguity sets using the Wasserstein distance, offering finite-sample performance guarantees and tractable convex reformulations grounded in

duality. Kim [48] focused on solving two-stage distributionally robust mixed-integer programs under Wasserstein ambiguity using a dual decomposition approach, demonstrating its scalability through discretization and Lagrangian relaxation. Cheramin et al. [49] proposed computationally efficient approximations for DRO models that combine moment-based and Wasserstein-based ambiguity, significantly improving tractability for large-scale problems. Jiang et al. [50] applied Wasserstein DRO to appointment scheduling under service time and no-show uncertainty, using copositive and linear programming reformulations to achieve strong out-of-sample performance. Hanasusanto and Kuhn [51] developed conic programming reformulations for two-stage DRO problems over Wasserstein balls, enabling tractable multistage decision-making. Lastly, Wiesemann, Kuhn, and Sim [52] presented a unifying DRO framework for convex optimization problems, covering a wide range of ambiguity sets and establishing general conditions for tractable reformulations.

3. Methodology

Figure 1 illustrates the proposed learning-driven optimization framework. At each decision epoch t , the capacity prediction model generates a probability distribution, where the support corresponds to the set of possible capacity outcomes $\hat{\xi}_t$ and the associated probabilities are denoted by \hat{p}_t . This predicted distribution then serves as input parameters for the downstream DR-MAGHP, which leverages them to derive robust ground delay policies. In what follows, we first describe the development of the capacity prediction model, including the derivation of actual capacities (Section 3.1.1) and the architecture of the probabilistic prediction model (Section 3.1.2). We then explain how DR-MAGHP incorporates these predictive distributions into the decision-making process, focusing on the construction of Wasserstein ambiguity sets (Section 3.2) and the derivation of a tractable reformulation (Section 3.3).

3.1. Airport capacity distribution prediction

We develop a deep learning model for airport capacity distribution prediction. The model predicts airport arrival or departure capacity distributions across a 12-hour prediction horizon, discretized into 15-minute intervals. The 12-hour horizon is selected based on operational requirements in air traffic management, where medium-term capacity predictions are critical for strategic flow management and ground delay program decisions [53, 54]. Additionally, this horizon aligns with

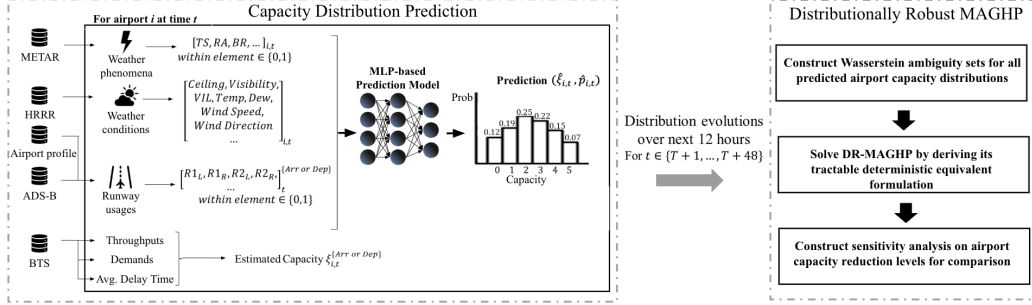


Figure 1: Learning-driven airport capacity distribution prediction and distributionally robust GDP optimization framework.

typical airport coordination timescales and provides sufficient lead time for proactive traffic management while maintaining reasonable prediction accuracy [55]. The 15-minute interval granularity balances computational efficiency with operational relevance, as it corresponds to standard air traffic management coordination intervals [56, 57] and captures meaningful variations in airport capacity without excessive noise from minute-by-minute fluctuations [17, 58].

To account for airport-specific operational characteristics and infrastructure differences [59], we build separate prediction models for each airport. Furthermore, for each airport, we develop independent models for arrival and departure capacity predictions. We acknowledge that this is a simplifying assumption, as arrival and departure operations are inherently interdependent due to shared runway usage, taxiway conflicts, and gate constraints [60, 61]. In practice, high arrival volumes can reduce departure capacity and vice versa, particularly at airports with intersecting or closely spaced parallel runways. However, modeling these complex interdependencies would significantly increase model complexity and data requirements. This independence assumption allows for more tractable model development while still capturing the primary capacity dynamics for each operation type. Future work could explore joint arrival-departure capacity prediction models or post-processing techniques to ensure consistency between arrival and departure predictions.

3.1.1. Deriving actual capacities from throughput

Training and validating the prediction model requires observations of actual airport capacity values. However, a fundamental challenge in airport capacity prediction is that true operational capacity cannot be directly observed from historical data [62]. It is crucial to distinguish between “throughput” and “capacity”

in aviation operations: *throughput* represents the actual number of aircraft operations (arrivals or departures) processed during a given time period, while *capacity* represents the maximum number of operations the airport system can theoretically handle under specific conditions [63]. Historical throughput data reflects realized demand constrained by both capacity limits and actual demand levels—when demand is low, observed throughput will be below the airport’s true capacity. Conversely, capacity represents the system’s theoretical maximum processing rate, which may not be fully utilized due to demand variations, schedule spacing, or operational inefficiencies.

To address this fundamental measurement challenge, we employ a systematic rule-based methodology to estimate capacity by identifying *capacity-saturated* periods where observed throughput accurately reflects true operational capacity. During these time periods t , we assume that $\widehat{\text{capacity}}_t = \text{throughput}_t$, meaning the airport is operating at or very close to its maximum capacity limit and demand exceeds the available capacity.

Our approach leverages multiple operational indicators to comprehensively identify capacity-saturated conditions. We define three complementary criteria based on key airport performance metrics: throughput levels, demand-supply relationships, and delay characteristics. For each airport i at time t , we apply the following decision rule separately for departure and arrival operations:

$$\widehat{\text{capacity}}_{i,t} = \text{throughput}_{i,t} \iff (\text{Criterion 1}) \cup (\text{Criterion 2}) \cup (\text{Criterion 3}), \quad (1)$$

where the three criteria are defined as:

1. *Criterion 1*—Throughput-based criterion:

$$\text{Throughput Ratio} = \frac{\text{actual throughput}_{i,t}}{\text{threshold}_i} \geq 1, \quad (2)$$

where threshold_i refers to the value in the 90th percentile of the throughput distribution for airport i .

2. *Criterion 2*—Demand-based criterion:

$$\text{Throughput-Demand Ratio} = \frac{\text{actual throughput}_{i,t}}{\text{scheduled demand}_{i,t}} \leq \alpha, \quad (3)$$

where α is a threshold parameter (we use $\alpha = 0.8$ in our implementation) that identifies periods when the airport cannot fully satisfy scheduled demand. This indicates that meaningful demand pressure exists where scheduled flight demand exceeds the airport’s processing capacity, pushing the

system toward its operational limits. We note that we did not conduct a comprehensive sensitivity analysis on this parameter, which represents a limitation of our current approach that could be addressed in future work.

3. *Criterion 3*—Delay-based criterion:

$$(\text{Average delay}_{i,t} \geq 15 \text{ mins}) \cap (\text{Delayed flights}_{i,t} \geq 2), \quad (4)$$

where delayed flights are defined as those experiencing more than 5 minutes of delay relative to the schedule. The 15-minute average delay threshold is defined to reflect significant operational stress indicative of capacity saturation. The requirement for at least 2 delayed flights ensures statistical significance and avoids classification based on isolated incidents. We acknowledge that these threshold parameters (15 minutes for average delay, 2 flights for minimum count, and 5 minutes for individual flight delay classification) were selected based on operational judgment rather than comprehensive sensitivity analysis.

This multi-criteria framework ensures robust identification of capacity-saturated periods by requiring when any one of three criteria is satisfied as shown in (1). The throughput criterion captures high activity levels, the demand criterion ensures meaningful operational pressure, and the delay criterion confirms that the airport is experiencing capacity-related stress. Only data from time periods satisfying all criteria are used for model training and validation, maximizing the chances that our capacity estimates reflect true operational limits rather than demand-constrained throughput values.

While this approach provides a practical solution for capacity estimation, we acknowledge that it may underestimate true capacity during periods of severe weather or other operational disruptions when capacity limits change but our historical thresholds may not capture these variations. However, alternative approaches face similar challenges—FAA-published AARs and ADRs are conservative "called rates" set for traffic management rather than reflecting absolute limits. AARs and ADRs are frequently adjusted based on anticipated conditions and risk tolerance, potentially underestimating the airport's true maximum throughput capability under optimal conditions.

Therefore, in the absence of a standardized ground truth for airport capacities, we consider our approach to be a reasonable benchmark for identifying periods when airports are operating near their practical limits under prevailing conditions.

3.1.2. Distributional capacity prediction

Our model includes input feature engineering, a multilayer perceptron (MLP)-based prediction model, and output reformulation. The MLP architecture consists of three layers with 17, 32, and $MAX_z + 1$ neurons respectively, where MAX_z represents the historical maximum capacity of airport z . The network employs ReLU activation functions in the hidden layers, produces probabilistic outputs through a softmax activation function in the final layer, and is trained using cross-entropy loss to optimize the capacity distribution predictions.

We denote the predicted capacity distributions for arrivals and departures at airport z and time t by

$$\left(\hat{\xi}_t^{(z,a)}, \hat{p}_t^{(z,a)}\right) \quad \text{and} \quad \left(\hat{\xi}_t^{(z,g)}, \hat{p}_t^{(z,g)}\right),$$

respectively. We learn a mapping $\Psi^{z,\bullet}$ (with $\bullet \in \{a, g\}$) from inputs X_t^z to a predictive distribution:

$$\left(\hat{\xi}_t^{(z,\bullet)}, \hat{p}_t^{(z,\bullet)}\right) = \Psi^{z,\bullet}(X_t^z).$$

Rather than predicting a single point estimate of capacity, our model outputs a complete probability distribution over all possible capacity values. This distributional approach is essential for the subsequent distributionally robust optimization (DRO) framework, which requires the full uncertainty characterization of capacity predictions rather than point estimates. Given the inherent aleatoric uncertainty in airport operations due to factors such as individual air traffic manager decisions, aircraft performance variations, and dynamic operational constraints, capacity prediction naturally exhibits distributional characteristics that cannot be adequately captured by point predictions.

To enable this distributional prediction, we reformulate the capacity prediction as a multi-class classification problem. We use a categorical encoding where $\hat{p}_t^{(z,\bullet)}$ is represented as a probability vector over all possible capacity values. This process converts scalar capacity values into a probability distribution across the discrete capacity space. For example, as shown in Figure 1, instead of predicting a single capacity value of 2, the model outputs a probability distribution such as $[0.05, 0.10, 0.70, 0.10, 0.03, 0.02]_t$, where each element represents the probability that the corresponding capacity value occurs. The length of this vector corresponds to the range of capacity values observed at this airport historically ($MAX_z + 1$ possible values from 0 to MAX_z). During training, the ground truth capacity observations (derived from the capacity-saturated periods as described

in Section 3.1.1) are encoded using one-hot vectors. For instance, an observed capacity of 2 would be represented as the sparse vector $[0, 0, 1, 0, 0, 0]_t$. However, the model’s output during inference is a complete probability distribution over all capacity values, providing the distributional information required for DR-MAGHP.

Our model incorporates two categories of input features for X_t : runway configurations and meteorological information. The runway configuration features capture the operational setup of the airport, including active runway assignments and operational modes, which directly impact the airport’s capacity limits. The meteorological features are converted and vectorized, including seven weather variables: ceiling, visibility, vertically integrated liquid (VIL), temperature, dew points, wind direction, surface wind speeds, and one phenomenon indicator of adverse weather. The selection of these features is based on their studied impacts on airport capacity (e.g., see [64, 65]). By combining both runway configuration and weather information, the model can capture the primary operational and environmental factors that determine airport capacity under different conditions.

3.2. Distributionally robust MAGHP (DR-MAGHP)

Recall that we seek to make optimal ground delay allocation decisions that are *robust* with respect to the predicted airport capacity distribution (output of Section 3.1). To do so, we will solve the MAGHP under a “worst-case” capacity distribution, located within the Wasserstein ambiguity set centered at the predicted airport capacity distribution. Figure 2 presents a numerical example of various distributions—and distribution families—that lie within an ambiguity set of radius $\epsilon = 0.005$, centered at a Gaussian distribution $\mathcal{N}(20, 3)$. The parameters are chosen purely for illustrative purposes and carry no operational significance. Formally, let $\mathcal{M}(\Xi)$ be the space of all probability distributions \mathbb{Q} with support Ξ . The Wasserstein distance $d_w : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \rightarrow \mathbb{R}_{\geq 0}$ is the minimum transportation cost between distributions $\mathbb{Q}_1 \in \mathcal{M}(\Xi)$ and $\mathbb{Q}_2 \in \mathcal{M}(\Xi)$ [47], and is given explicitly as:

$$d_w(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\Pi \in \mathcal{D}_{\Pi}(\xi_1, \xi_2)} \int_{\Xi^2} \|\xi_1 - \xi_2\|_2 \Pi(d\xi_1, d\xi_2), \quad (5)$$

where Π is a joint distribution of random variables ξ_1 and ξ_2 with marginals \mathbb{Q}_1 and \mathbb{Q}_2 , respectively. We denote $\mathcal{D}_{\Pi}(\xi_1, \xi_2)$ as the set of all joint distributions on ξ_1 and ξ_2 with marginals \mathbb{Q}_1 and \mathbb{Q}_2 .

Let Z be the set of all airports. For airport z , we let M_z be the maximum capacity of airport z and $\{\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_{M_z}\}$ be the set of airport capacities for airport

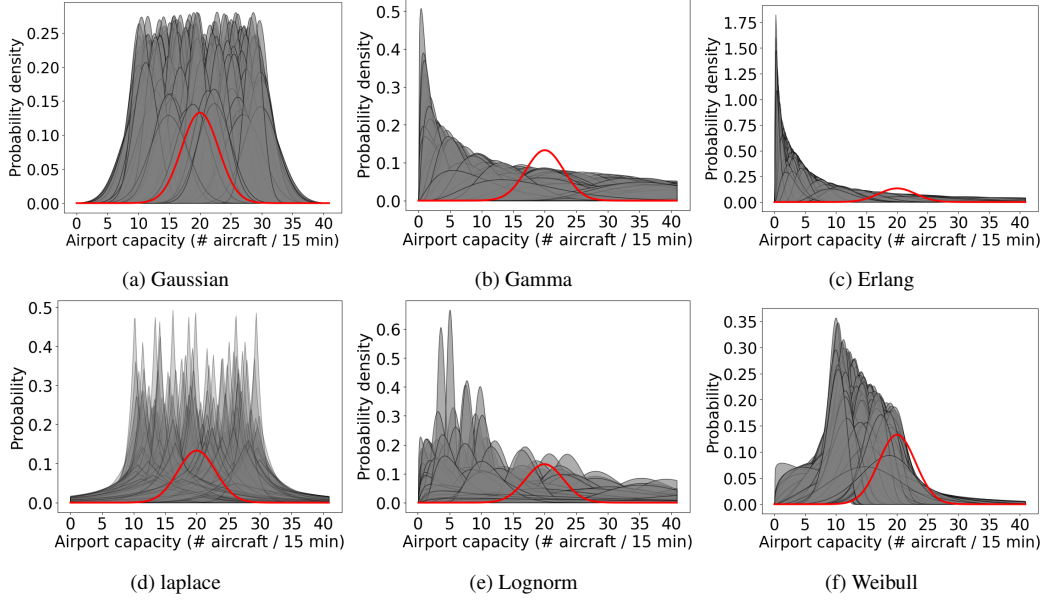


Figure 2: Example of distributions included in the Wasserstein ambiguity set for the empirical Gaussian distribution $\mathcal{N}(20, 3)$ with $\epsilon = 0.005$.

z with the corresponding estimated probabilities of occurrence $\{\hat{p}_1, \hat{p}_2, \dots, \hat{p}_{M_z}\}$ (i.e., this is precisely the predicted capacity distribution's probability mass function, or PMF, which is the output of the prediction step described in Section 3.1). The ambiguity set centered around a predicted capacity distribution \hat{P} with radius $\epsilon > 0$, denoted as $\mathcal{P}_\epsilon(\hat{P})$, is given by

$$\mathcal{P}_\epsilon(\hat{P}) := \left\{ \mathbb{Q} \in \mathcal{M}(\Xi) : d_w(\hat{P}, \mathbb{Q}) \leq \epsilon \right\}. \quad (6)$$

Note that to differentiate between arrival and departure capacities, we will use subscripts and superscripts with a and g , respectively, when we write out the full DR-MAGHP model. We refer readers to [66] for additional technical details on specifying the ambiguity set.

We denote the set of flights by F , $F^{(g)}(z)$, $F^{(a)}(z)$ as the set of flights departing from and landing at airport z , $D_z(t)$ and $R_z(t)$ as the departure and arrival capacity of airport $z \in Z$ at time t , respectively, and $\mathcal{C} = F \times F$ as the set of all flight pairs connected by the same aircraft (or tail), where $(f, f') \in \mathcal{C}$ denotes the preceding flight f and successive flight f' . The decision variables $u_{f,t}$ and $v_{f,t}$ are binary variables where $u_{f,t}$ equals one if flight f will depart at time t ;

similarly, $v_{f,t}$ equals one if flight f will land at time t . z_f^g and z_f^a represent flights scheduled to depart or land at airport z , respectively. d_f is the scheduled departure time of f and r_f is the scheduled arrival time of f . T_d^f is the set of available time periods for f to take off and T_a^f is the set of available time periods for f to land. g_f, a_f are ground holding delay and airborne delay, respectively, where $g_f = \sum_{t \in T_d^f} t u_{f,t} - d_f$ and $a_f = \sum_{t \in T_a^f} t v_{f,t} - r_f - g_f$. These definitions follow the standard formulation of the MAGHP [1] and we denote the deterministic MAGHP as DET-MAGHP. The formulation is given explicitly below:

$$\min_{u,v} \sum_{f=1}^F (C_f^g g_f + C_f^a a_f) \quad (7a)$$

$$\text{s. t.} \quad \sum_{f: z_f^g = z} u_{f,t} \leq D_z(t), \quad \forall z \in Z, t \in T, \quad (7b)$$

$$\sum_{f: z_f^a = z} v_{f,t} \leq R_z(t), \quad \forall z \in Z, t \in T, \quad (7c)$$

$$\sum_{t \in T_d^f} u_{f,t} = 1, \quad \forall f \in F, \quad (7d)$$

$$\sum_{t \in T_a^f} v_{f,t} = 1, \quad \forall f \in F, \quad (7e)$$

$$g_{f'} + a_{f'} - s_{f'} \leq g_f, \quad \forall (f, f') \in \mathcal{C}, \quad (7f)$$

$$a_f, g_f \geq 0, \quad \forall f \in F, \quad (7g)$$

$$u_{f,t}, v_{f,t} \in \{0, 1\}, \quad \forall f \in F, t \in T. \quad (7h)$$

The objective function in Equation (7a) is the sum of ground holding delay and airborne delay across all flights. Constraints (7b) and (7c) are airport departure and arrival capacity constraints, respectively. Constraints (7d) and (7e) ensure only one departure and arrival time slot is assigned to each flight, respectively, and (7f) enforces minimum ground turnaround times for connecting flights $(f, f') \in \mathcal{C}$. We incorporate predicted airport departure and arrival capacity distributions from Section 3.1, along with robustness guarantees, through a two-stage formulation. In the two-stage setting, in addition to the first stage decision variables $u_{f,t}$ and $v_{f,t}$, we introduce second stage decision variables $y_{z,t}^{(g)}$ and $y_{z,t}^{(a)}$. Decision variables $y_{z,t}^{(g)}$ denote the additional number of flights entering the departure queue at time period t , while $y_{z,t}^{(a)}$ represents the additional number of flights entering the arrival queue at the same time period. The second stage of the model

focuses on rescheduling the actual departure and arrival times of flights once the actual airport capacity distributions are realized. The objective of the second stage model is to minimize the number of flights joining either queues. We first formulate the two-stage stochastic MAGHP (SP-MAGHP) as follows:

$$\min_{u,v,g,a,s} \left\{ \sum_{f \in F} (C_g g_f + C_a a_f) + \sum_{z \in Z} \mathbb{E}_{\hat{P}_z^{(g)}} [Q_z^{(g)}(u, \xi_z^{(g)})] + \sum_{z \in Z} \mathbb{E}_{\hat{P}_z^{(a)}} [Q_z^{(a)}(v, \xi_z^{(a)})] \right\} \quad (8a)$$

$$\text{s.t.} \quad \sum_{t \in T} u_{f,t} = 1, \quad \forall f \in F, \quad (8b)$$

$$\sum_{t \in T} v_{f,t} = 1, \quad \forall f \in F, \quad (8c)$$

$$g_{f'} + a_{f'} - s_{f'} \leq a_f, \quad \forall (f, f') \in \mathcal{C}, \quad (8d)$$

$$u_{f,t} \in \{0, 1\}, \quad \forall f \in F, t \in T, \quad (8e)$$

$$g_f, a_f, s_f \geq 0, \quad \forall f \in F, \quad (8f)$$

where $Q^{(g)}(u, \xi^{(g)})$ is:

$$\min_{y^{(g)}} \sum_{t \in T} C_g (\xi_z^{(g)}) y_{z,t}^{(g)} (\xi_z^{(g)}) \quad (9a)$$

$$\text{s. t.} \quad \sum_{f: z_f^g = z} u_{f,t} \leq \xi_{z,t}^{(g)} + y_{z,t}^{(g)} (\xi_z^{(g)}), \quad (9b)$$

$$\begin{aligned} & \forall z \in Z, t \in T, \xi_z^{(g)} \in \Xi_z^{(g)}, \\ & y_{z,t}^{(g)} (\xi_z^{(g)}) \geq 0, \quad \forall z \in Z, t \in T, \xi_z^{(g)} \in \Xi_z^{(g)}, \end{aligned} \quad (9c)$$

and $Q^{(a)}(u, \xi^{(a)})$ shares a similar formulation:

$$\min_{y^{(a)}} \sum_{t \in T} C_a(\xi_z^{(a)}) y_{z,t}^{(a)}(\xi_z^{(a)}) \quad (10a)$$

$$\text{s. t. } \sum_{f: z_f^a = z} v_{f,t} \leq \xi_{z,t}^{(a)} + y_{z,t}^{(a)}(\xi_z^{(a)}), \quad (10b)$$

$$\begin{aligned} & \forall z \in Z, t \in T, \xi_z^{(a)} \in \Xi_z^{(a)}, \\ & y_{z,t}^{(a)}(\xi_z^{(a)}) \geq 0, \quad \forall z \in Z, t \in T, \xi_z^{(a)} \in \Xi_z^{(a)}. \end{aligned} \quad (10c)$$

We denote $\hat{P}_z^{(g)}$ and $\hat{P}_z^{(a)}$ as the predicted departure and arrival capacity distributions, respectively, obtained from the upstream model described in Section 3.1. For each airport z , let $\Xi_z^{(g)}$ and $\Xi_z^{(a)}$ represent the sets of possible supports (scenarios) for the departure and arrival distributions. We further denote $\xi_z^{(g)}$ and $\xi_z^{(a)}$ as specific scenarios within these sets. Finally, $y_{z,t}^{(g)}$ and $y_{z,t}^{(a)}$ correspond to the second-stage decision variables under each scenario at time t . The objective function (8a) includes the first-stage costs plus $\sum_{z \in Z} \mathbb{E}_{\hat{P}_z^{(g)}} \left[Q_z^{(g)}(u, \xi_z^{(g)}) \right]$ and $\sum_{z \in Z} \mathbb{E}_{\hat{P}_z^{(a)}} \left[Q_z^{(a)}(v, \xi_z^{(a)}) \right]$. Constraints (8b)-(8d) are the first stage assignment and coupling constraints inherited from the standard, deterministic MAGHP. The first constraint in the second stage minimization problem ensures that, even if the airport capacity is reduced, the number of departing (or arriving) flights at time t does not exceed the realized airport capacity, plus the total number of extra flights allowed to depart (or arrive) at time t , across all airports. When there is a drop in airport capacity, the two-stage model will optimally adjust delay allocations based on the weighting between airborne and ground delays (typically airborne delays are 1.2 to 3 times more expensive [67]). We also assume in this formulation that the capacity distributions across airports are mutually independent. Additionally, recall that we assume that the departure and arrival capacity distributions at each airport are considered independent of one another.

With the formulation of SP-MAGHP and the assumptions we have mentioned above, we then develop the distributionally robust MAGHP model (DR-MAGHP). For each airport z , we construct the Wasserstein ambiguity sets around its depar-

ture and arrival capacities, and the DR-MAGHP can be written as follows:

$$\min_{u,v} \left\{ \sum_{f \in F} (C_g g_f + C_a a_f) + \sum_{z \in Z} \max_{p \in \mathcal{P}_\epsilon(\hat{P}_z^{(g)})} \mathbb{E}_p [Q^{(g)}(u, \xi_z^{(g)})] \right. \\ \left. + \sum_{z \in Z} \max_{p \in \mathcal{P}_\epsilon(\hat{P}_z^{(a)})} \mathbb{E}_p [Q^{(a)}(v, \xi_z^{(a)})] \right\} \quad (11a)$$

$$\text{s. t.} \quad \sum_{t \in T_d^f} u_{f,t} = 1, \forall f \in F, \quad (11b)$$

$$\sum_{t \in T_a^f} v_{f,t} = 1, \forall f \in F, \quad (11c)$$

$$g_{f'} + a_{f'} - s_{f'} \leq a_f, \forall (f, f') \in \mathcal{C}. \quad (11d)$$

The objective function (11a) includes two inner maximization problems, which seek the worst-case distribution within each ambiguity set that maximizes the expected second stage cost. Constraints (11b)-(11d) and $Q^{(g)}, Q^{(a)}$ are the same from the formulation of SP-MAGHP in (10). $\mathcal{P}_\epsilon(\hat{P}_z^{(g)}), \mathcal{P}_\epsilon(\hat{P}_z^{(a)})$ are Wasserstein ambiguity sets of size $\epsilon > 0$ constructed based on capacity distribution predictions and p is an arbitrary distribution within each ambiguity set.

3.3. Scenario reduction and DR-MAGHP reformulation

Recall that we optimize ground holding decisions across a prediction horizon of 12 hours, subdivided into 48 time periods (with each unit time period of 15 minutes). From the airport capacity distribution prediction models (Section 3.1), we are given new predicted distributions of the arrival and departure capacities (and hence, associated Wasserstein ambiguity sets) at each time, across all airports. This leads to an exponential increase in the number of scenarios for the optimization model, resulting in severe computational intractability. The root of this challenge lies in the temporal dependency of decisions: The decision at any given time depends on the sequence of prior decisions. Specifically, if the maximum capacity at an airport within each 15-minute interval is M , and the planning horizon contains $|T|$ time periods (discretizations), the total number of scenarios becomes $M^{|T|}$. For instance, at Hartsfield-Jackson Atlanta International Airport (ATL) with a maximum hourly capacity of 35, the scenario space expands to 35^{48} , rendering a direct solution to be computationally intractable.

To address this, we apply a scenario reduction strategy aimed at reducing the scenario tree complexity and thus improving model tractability. Our strategy is to reduce discretization points used for capacities and time periods. Firstly, to reduce the number of discretizations $|T|$, we compute the pairwise Wasserstein distance between capacity distributions at consecutive time periods. Time periods with small Wasserstein distances—indicating similar capacity behavior—are grouped together. In contrast, a time step is identified as a change point if its Wasserstein distance (i.e., Equation (5) with ℓ_1 norm) is among the n largest values across the horizon. These change points divide the full time horizon into $n + 1$ intervals. For each interval, a representative capacity distribution is computed as the average (i.e., centroid) of the predicted distributions within that interval. The algorithm for this grouping procedure is given in Algorithm 1.

Algorithm 1 Clustering of Capacity Time Series via Similarity Measures

- 1: **Input:** Predicted PMFs $\{\hat{p}_t^{(z,a)}, \hat{p}_t^{(z,g)}\}_{t \in T}$ for each airport $z \in Z$; number of clusters n ; Wasserstein distance as the similarity measure $d_w(\cdot, \cdot)$
 - 2: **for** each $z \in Z$ **do**
 - 3: Compute Wasserstein distances: $w_t^{(z,a)} = d_w(\hat{p}_t^{(z,a)}, \hat{p}_{t-1}^{(z,a)})$, and similarly for departure capacity distributions
 - 4: Identify top n indices $C^{(z,\cdot)} := \arg \max_{S \subset \{1, \dots, |T|-1\}, |S|=n} \sum_{t \in S} w_t^{(z,\cdot)}$, then sort such that $c_1 < \dots < c_n$, and $c_1, \dots, c_n \in C^{(z,\cdot)}$
 - 5: Partition time: $T_1 = [0, c_1]$, $T_2 = [c_1 + 1, c_2]$, \dots , $T_{n+1} = [c_n + 1, |T| - 1]$
 - 6: **for** each interval $T_k^{(z,a)}$, $k \in \{1, 2, \dots, n + 1\}$ **do**
 - 7: $\bar{p}_k^{(z,a)} = \frac{1}{|T_k^{(z,a)}|} \sum_{t \in T_k^{(z,a)}} \hat{p}_t^{(z,a)}$
 - 8: **end for**
 - 9: **for** each interval $T_k^{(z,g)}$, $k \in \{1, 2, \dots, n + 1\}$ **do**
 - 10: $\bar{p}_k^{(z,g)} = \frac{1}{|T_k^{(z,g)}|} \sum_{t \in T_k^{(z,g)}} \hat{p}_t^{(z,g)}$
 - 11: **end for**
 - 12: **end for**
 - 13: **Output:** Clustered time intervals $T_k^{(z,\cdot)}$ and representative PMFs $\bar{p}_k^{(z,\cdot)}$ for each airport
-

To further reduce the complexity of the predicted capacity distributions, we apply K -means clustering to the support–probability pairs of each probability mass function (PMF) within the time intervals identified by the Wasserstein-based clus-

tering. For each interval, the original distribution is approximated by a reduced PMF with a fixed number of representative support points. Specifically, the K -means algorithm groups the support-probability pairs into K clusters. For each cluster, the new support value is computed as the probability-weighted average of the original supports, and the total probability mass is obtained by summing the probabilities of all points in the cluster. This compression scheme preserves the overall shape of the original distribution while significantly reducing its dimensionality, thereby facilitating tractable downstream optimization. The details for compressing the PMFs themselves are laid out in Algorithm 2.

Algorithm 2 PMF Clustering via K -Means

- 1: **Input:** For each airport $z \in Z$ and time interval k : representative PMF $(\bar{p}_k^{(z,\cdot)}, \bar{\xi}^{(z,\cdot)})$ with supports $\bar{\xi}_i \in \mathbb{Z}_{\geq 0}$ for each airport and clustered time interval k , and target cluster count K
 - 2: **for** each $z \in Z$ **do**
 - 3: **for** each interval k **do**
 - 4: Form dataset $\mathcal{D}_k^{(z,\cdot)} = \{(\bar{\xi}_i, \bar{p}_i)\}_{i=1}^{N_k}$
 - 5: Apply K -means on $\mathcal{D}_k^{(z,\cdot)}$ with K clusters [68]
 - 6: **for** each cluster $j = 1, \dots, K$ **do**
 - 7: Let \mathcal{I}_j be indices assigned to cluster j
 - 8: Compute cluster weight: $\bar{p}_{k,j}^{(z,\cdot)} = \sum_{i \in \mathcal{I}_j} \bar{p}_i$
 - 9: Compute cluster centroid: $\bar{\xi}_{k,j}^{(z,\cdot)} = \text{round} \left(\frac{\sum_{i \in \mathcal{I}_j} \bar{\xi}_i \bar{p}_i}{\sum_{i \in \mathcal{I}_j} \bar{p}_i} \right)$
 - 10: **end for**
 - 11: Define clustered PMF: $\hat{p}_k^{(z,\cdot)} = \{(\bar{\xi}_{k,j}^{(z,\cdot)}, \bar{p}_{k,j}^{(z,\cdot)})\}_{j=1}^K$
 - 12: **end for**
 - 13: **end for**
 - 14: **Output:** Clustered PMFs $\{\hat{p}_k^{(z,\cdot)}\}$ for all z and k
-

An illustration of the reduced scenario tree is shown in Figure 3, where $\bar{\xi}_{k,j}$ denotes the j^{th} probability mass center (i.e., representative support) of the k^{th} time cluster. The scenario tree captures the joint distribution of an airport's arrival or departure capacities across the planning horizon, with each edge representing a possible support of this joint distribution. For example, the leftmost path in Figure 3, given by the sequence $\{\bar{\xi}_{1,1}^{(z,\cdot)}, \bar{\xi}_{2,1}^{(z,\cdot)}, \bar{\xi}_{3,1}^{(z,\cdot)}\}$, corresponds to a scenario in which, at each of the three time clusters, the capacity takes the first representative

value of the respective PMF. The probability associated with each scenario in the reduced scenario tree is computed as the product of the marginal probabilities of the selected probability mass clusters at each time stage. Specifically, for a scenario represented by the sequence $\bar{\xi}^{(z,\cdot)} = \{\bar{\xi}_{1,j_1}^{(z,\cdot)}, \bar{\xi}_{2,j_2}^{(z,\cdot)}, \dots, \bar{\xi}_{K,j_K}^{(z,\cdot)}\}$, where K is the number of time clusters and \bar{p}_{k,j_k} denotes the probability assigned to the j_k^{th} mass cluster at time cluster k , the joint scenario probability is given by:

$$\widehat{\mathbb{P}}^{(z,\cdot)}(\bar{\xi}_{1,j_1}^{(z,\cdot)}, \dots, \bar{\xi}_{K,j_K}^{(z,\cdot)}) = \prod_{k=1}^K \bar{p}_{k,j_k}^{(z,\cdot)}.$$

This formulation assumes conditional independence across time clusters in the reduced scenario construction, such that the joint distribution can be expressed as a product of marginal distributions over time. Therefore, given the reduced scenario tree, we update the formulation of SP-MAGHP as

$$\min_{u,v,g,a,s} \left\{ \sum_{f \in F} (C_g g_f + C_a a_f) + \sum_{z \in Z} \mathbb{E}_{\widehat{\mathbb{P}}^{(z,g)}} [Q_z^{(g)}(u, \bar{\xi}^{(z,g)})] + \right. \quad (12a)$$

$$\left. \sum_{z \in Z} \mathbb{E}_{\widehat{\mathbb{P}}^{(z,a)}} [Q_z^{(a)}(v, \bar{\xi}^{(z,a)})] \right\} \quad (12b)$$

$$\text{s.t. Constraints (8b)-(8f),} \quad (12c)$$

where both the supports and probabilities are obtained from the joint distribution encoded by the reduced scenario tree, rather than from the original marginal capacity distributions. The second stage value function $Q_z^{(g)}(u, \bar{\xi}^{(z,g)})$ is reformulated as

$$\min_{y^{(g)}} \sum_{t \in T} C_g(\bar{\xi}^{(z,g)}) y_t^{(g)}(\bar{\xi}^{(z,g)}) \quad (13a)$$

$$\text{s.t. } \sum_{f: z_f^g = z} u_{f,t} \leq D_t(\bar{\xi}^{(z,g)}) + y_t^{(g)}(\bar{\xi}^{(z,g)}), \quad (13b)$$

$$\begin{aligned} \forall t \in T, \bar{\xi}^{(z,g)} &\in \Xi^{(z,g)}, \\ y_t^{(g)}(\bar{\xi}^{(z,g)}) &\geq 0, \\ \forall t \in T, \bar{\xi}^{(z,g)} &\in \Xi^{(z,g)}, \end{aligned} \quad (13c)$$

where we denote $D_t(\bar{\xi}^{(z,g)})$ as the capacity value assigned at time t under scenario $\bar{\xi}^{(z,g)}$. Then, $D_t(\bar{\xi}^{(z,g)}) = \bar{\xi}_{k,j_k}$ if $t \in T_k^{(z,g)}$, where $T_k^{(z,g)}$ is the set of time steps

associated with time cluster k , and $\bar{\xi}^{(z,g)} = \left(\bar{\xi}_{1,j_1}^{(z,g)}, \bar{\xi}_{2,j_2}^{(z,g)}, \dots, \bar{\xi}_{K,j_K}^{(z,g)} \right)$ is a scenario in the reduced scenario tree. We also note that $Q_z^{(a)}(u, \bar{\xi}^{(a)})$ is of a similar form, and for brevity we skip writing down its formulation in its entirety here. The formulation of DR-MAGHP (11) is updated accordingly as follow:

$$\begin{aligned} \min_{u,v} \quad & \sum_{f \in F} (C_g g_f + C_a a_f) \\ & + \sum_{z \in Z} \max_{p \in \mathcal{P}_\epsilon(\hat{\mathbb{P}}^{(z,g)})} \mathbb{E}_p [Q^{(g)}(u, \bar{\xi}^{(z,g)})] \\ & + \sum_{z \in Z} \max_{p \in \mathcal{P}_\epsilon(\hat{\mathbb{P}}^{(z,a)})} \mathbb{E}_p [Q^{(a)}(v, \bar{\xi}^{(z,a)})] \end{aligned} \quad (14a)$$

$$\text{s.t.} \quad \text{Constraints (11b)–(11d)}. \quad (14b)$$

The Wasserstein ambiguity sets appearing in the objective function is explicitly expressed as follows

$$\mathcal{P}_\epsilon \left(\hat{\mathbb{P}}^{(z,\cdot)} \right) := \left\{ \mathbb{Q} \in M \left(\Xi^{(z,\cdot)} \right) : d_w \left(\hat{\mathbb{P}}^{(z,\cdot)}, \mathbb{Q} \right) \leq \epsilon \right\}. \quad (15)$$

However, the formulation in (14) is generally computationally intractable, primarily due to the integral expression in (5), which leads to an infinite-dimensional optimization problem. Furthermore, the inherent min–max structure introduces additional complexity, rendering the problem even more challenging to solve directly. To address these challenges, we reformulate the DR-MAGHP in (14) by transforming the inner second-stage maximization problems into equivalent minimization problems, resulting in a semi-infinite program. We then apply discretization techniques to handle the continuous support of this semi-infinite formulation, ultimately yielding a deterministic equivalent representation of the DR-MAGHP. For a comprehensive derivation of the deterministic reformulation, we refer the reader to [66]. In this paper, we directly present the deterministic equivalent formulation of DR-MAGHP as follows:

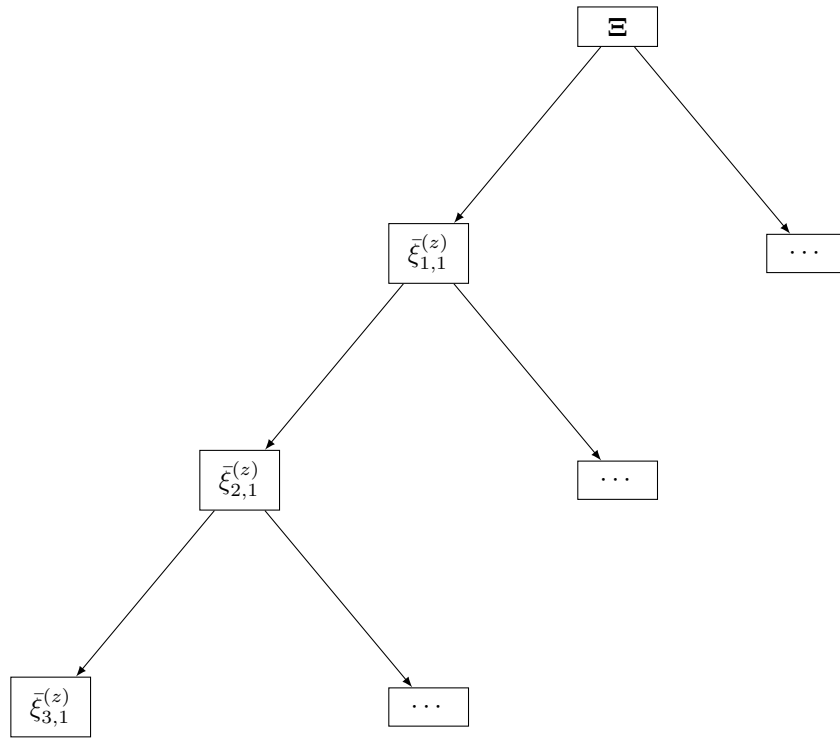


Figure 3: Illustration of the reduced scenario tree for capacity distributions. Each node $\bar{\xi}_{k,j}^{(z)}$ represents the j^{th} support value at stage k .

$$\min_{\alpha, \beta, \mathbf{y}, u, v} \left\{ \sum_{f \in F} \left(C_g g_f + C_a a_f \right) + \Phi^{(g)} + \Phi^{(a)} \right\} \quad (16a)$$

$$\text{s.t. } \alpha^{(z,g)} \left\| \bar{\xi}^{(z,g)} - \xi^{(z,g)} \right\|_2 + \beta^{(z,g)} \left(\bar{\xi}^{(z,g)} \right) \geq \sum_{t \in T} C_g \left(\bar{\xi}^{(z,g)} \right) y_t^{(g)} \left(\bar{\xi}^{(z,g)} \right),$$

$$\forall \bar{\xi}^{(z,g)} \in \Xi^{(z,g)} \forall z \in Z, \xi^{(z,g)} \in \Xi^{(z,g)}, \quad (16b)$$

$$\alpha^{(z,a)} \left\| \bar{\xi}^{(z,a)} - \xi^{(z,a)} \right\|_2 + \beta^{(z,a)} \left(\bar{\xi}^{(z,a)} \right) \geq \sum_{t \in T} C_a \left(\bar{\xi}^{(z,a)} \right) y_t^{(a)} \left(\bar{\xi}^{(z,a)} \right),$$

$$\forall \bar{\xi}^{(z,a)} \in \Xi^{(z,a)}, \forall z \in Z, \xi^{(z,a)} \in \Xi^{(z,a)}, \quad (16c)$$

$$\sum_{t \in T} u_{f,t} = 1,$$

$$\forall f \in F, \quad (16d)$$

$$\sum_{t \in T} v_{f,t} = 1,$$

$$\forall f \in F, \quad (16e)$$

$$D_t \left(\xi^{(z,g)} \right) + y_t^{(g)} \left(\xi^{(z,g)} \right) \geq \sum_{f \in F^{(g)}(z)} u_{f,t},$$

$$\forall t \in T, \xi^{(z,g)} \in \Xi^{(z,g)}, \quad (16f)$$

$$R_t \left(\xi^{(z,a)} \right) + y_t^{(a)} \left(\xi^{(z,a)} \right) \geq \sum_{f \in F^{(a)}(z)} v_{f,t},$$

$$\forall t \in T, \xi^{(z,a)} \in \Xi^{(z,a)}, \quad (16g)$$

$$y_0^{(\cdot)} \left(\bar{\xi}^{(z,\cdot)} \right) = 0, y_t^{(\cdot)} \left(\bar{\xi}^{(z,\cdot)} \right) \geq 0, \alpha^{(z,\cdot)} \geq 0,$$

$$\forall t \in T, \xi^{(z,\cdot)} \in \Xi^{(z,\cdot)} z \in Z. \quad (16h)$$

In the objective function (16a), $\Phi^{(g)}$ and $\Phi^{(a)}$ are substitutions for the dual objective function; explicitly, we have that:

$$\Phi^{(g)} = \epsilon^{(g)} \sum_{z \in Z} \alpha^{(z,g)} + \sum_{z \in Z} \sum_{\bar{\xi}^{(z,g)} \in \Xi^{(z,g)}} \widehat{\mathbb{P}}^{(z,g)}(\bar{\xi}^{(z,g)}) \beta^{(z,g)} \left(\bar{\xi}^{(z,g)} \right),$$

$$\Phi^{(a)} = \epsilon^{(a)} \sum_{z \in Z} \alpha^{(z,a)} + \sum_{z \in Z} \sum_{\bar{\xi}^{(z,a)} \in \Xi^{(z,a)}} \widehat{\mathbb{P}}^{(z,a)}(\bar{\xi}^{(z,a)}) \beta^{(z,a)} \left(\bar{\xi}^{(z,a)} \right).$$

α and β are dual variables for the two-stage dr-MAGHP, ϵ is the radius for the Wasserstein ambiguity set, and $\widehat{\mathbb{P}}^{(z,\cdot)}(\bar{\xi}^{(z,\cdot)})$ is the probability assigned to each

scenario within the empirical distribution. Constraints (16b) and (16c) are derived from the Wasserstein distance constraints of (12). (16d) and (16e) are the first stage assignment constraints, and (16f) and (16g) are the capacity constraints for each scenario of each distribution within the ambiguity set.

4. Numerical Experiments and Discussion

In the following numerical experiments, we evaluate the proposed integrated prediction and DRO framework as a whole. The upstream capacity prediction model generates probabilistic predictions, which are then used as input distributions for the downstream SP-MAGHP and DR-MAGHP. We first examine the quality of the predictive distributions and then assess how these predictions translate into prescriptive actions (i.e., air traffic management strategic rescheduling) through the optimization stage.

4.1. Data Description

We obtained airport throughput data from the US Department of Transportation’s Bureau of Transportation Statistics (BTS), and weather data from the US National Oceanic and Atmospheric Administration’s High-Resolution Rapid Refresh (HRRR) database. BTS provides detailed information for each flight, including the scheduled departure and arrival times, actual departure and arrival times, delay duration. Using these data points and the procedure described in Section 3.1.1, we estimate the capacity of each FAA Core 30 airport in our study (note that we can scale to a larger network of airports if needed). HRRR provides weather data on a $3 \text{ km} \times 3 \text{ km}$ grid covering all 50 US states, with a forecast horizon of up to 23 hours from the current hour. We collect data for the entire year of 2019. Each day is divided into 96 quarter-hour intervals, resulting in 35,040 time periods in total.

4.2. Prediction Model Setup and Evaluation

4.2.1. Experiment setup

We use data comprising 60 datasets in total from the 30 US airports, with each airport providing 2 datasets for arrival and departure operations. Each dataset contains approximately 13%-48% of time periods remaining after rule-based capacity estimation presented in (1), with an overall average of 25.7% for arrivals and 26.7% for departures. As a reminder, the rule-based approach identifies capacity-saturated periods when any one of three criteria is satisfied: (1) throughput-based

criterion (actual throughput \geq 90th percentile threshold), (2) demand-based criterion (throughput-demand ratio ≤ 0.8 , indicating unmet demand), and (3) delay-based criterion (average delay ≥ 15 minutes with at least 2 delayed flights). The exact number of remaining time periods for each airport across both arrival and departure operations is detailed in Appendix A.

For the train-validation-test split, we follow temporal order by using the first 10 weeks of each quarter for training, the 11th week for validation, and the final 12th week for testing. We normalize all numerical features using min-max normalization applied to the training set to prevent the model from being dominated by a subset of variables, with the same normalization parameters applied to both validation and test sets to ensure consistent scaling across all data splits. Hyperparameter tuning is performed via grid search, resulting in a learning rate of 0.0001, 300 epochs, and a batch size of 16.

4.2.2. Evaluation Metrics

We use four metrics to evaluate predictive model performance: *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, *Prediction Interval Coverage Probability (PICP)*, and *Mean Prediction Interval Width (MPIW)*. These metrics provide a comprehensive assessment of both point prediction accuracy and distributional prediction quality.

Point prediction results: The model predicts the capacity probability distribution for each airport i at time t . For point prediction evaluation, we select the capacity value with the highest predicted probability:

$$\hat{c}_{i^*,t} = \operatorname{argmax}_i p_{i,t}, \quad (17)$$

where $p_{i,t}$ represents the predicted probability for capacity value i at time t for a specific airport, and $\hat{c}_{i^*,t}$ is the predicted capacity value with the highest probability. The RMSE measures the prediction accuracy in terms of squared errors:

$$\text{RMSE} = \sqrt{\frac{1}{N_t} \sum_{t=1}^{N_t} (\hat{c}_{i^*,t} - c_t)^2}, \quad (18)$$

and the MAE provides a linear penalty for prediction errors:

$$\text{MAE} = \frac{1}{N_t} \sum_{t=1}^{N_t} |\hat{c}_{i^*,t} - c_t|, \quad (19)$$

where c_t is the actual observed capacity value at time t .

Distributional prediction quality: For evaluating the quality of predicted capacity distributions, we employ prediction interval-based metrics. The tolerance rate TR_t represents the prediction interval at time t under a 90% confidence level, defined as the set of capacity values whose cumulative probability reaches or exceeds 90%:

$$\sum_{i \in TR_t} p_{i,t} \geq 0.9. \quad (20)$$

The PICP measures how often the actual capacity falls within the predicted tolerance rate:

$$\text{PICP} = \frac{1}{N_t} \sum_{t=1}^{N_t} I(c_t \in TR_t), \quad (21)$$

where $I(\cdot)$ is the indicator function that equals 1 if the condition is true, and 0 otherwise. The MPIW quantifies the average width of the prediction intervals:

$$\text{MPIW} = \frac{1}{N_t} \sum_{t=1}^{N_t} |TR_t|, \quad (22)$$

where $|TR_t|$ represents the range or size of the tolerance rate.

The prediction model is optimized to minimize RMSE and MAE for point prediction accuracy, while simultaneously maximizing PICP and minimizing MPIW for distributional prediction quality. A well-calibrated model should achieve high PICP (ideally close to 0.9, corresponding to the 90% confidence level) while maintaining narrow prediction intervals (low MPIW).

4.2.3. Capacity Distribution Prediction Results

The capacity distribution prediction performance across the core 30 US airports, as shown in Tables 1 and 2, reveals significant variability in model accuracy and calibration quality, reflecting the diverse operational characteristics and complexity of different airport environments. The RMSE and MAE results demonstrate considerable variation in model performance across airports. The model achieves relatively strong prediction accuracy at medium or relatively smaller large hub airports such as MEM (arrival RMSE: 0.81, MAE: 0.45), MDW (arrival RMSE: 1.59, MAE: 1.11), and BWI (arrival RMSE: 1.73, MAE: 1.23). In contrast, major hub airports present substantial prediction challenges, with LAX showing the highest arrival prediction errors (RMSE: 8.55, MAE: 6.84), followed by DEN (RMSE: 7.93, MAE: 6.09) and CLT (RMSE: 6.99, MAE: 5.50). This

pattern suggests that airports with complex operational environments, multiple runway configurations, and high traffic variability inherently pose greater difficulties for capacity prediction models.

Additionally, departure capacity prediction exhibits a notably different performance pattern compared to arrivals. Several airports show significantly higher departure prediction errors, with DTW demonstrating particularly poor departure performance (RMSE: 10.04, MAE: 8.00) and IAH showing substantial challenges for both arrivals (RMSE: 5.66) and departures (RMSE: 10.41). This asymmetry between arrival and departure prediction accuracy likely reflects the greater complexity of departure operations, which are more susceptible to ground delays, pushback sequencing constraints, airline scheduling decisions, and air traffic control departure sequencing.

The PICP results reveal systematic calibration issues across the majority of airports, with most facilities achieving coverage probabilities substantially below the target 90% confidence level. The best-performing airports in terms of distributional prediction include DTW (arrival PICP: 77.89%, departure PICP: 79.34%) and IAD (departure PICP: 80.62%), though even these fall short of the ideal 90% coverage. Most airports exhibit PICP values in the 45-75% range, indicating significant underconfidence in the predicted distributions. The consistent underestimation of prediction intervals suggests that the current approach may not adequately capture all sources of capacity variability, including weather-related variations, operational disruptions, and seasonal traffic patterns. The MPIW results show that prediction interval width generally correlates with airport complexity and size. Major hub airports like DTW (departure MPIW: 19.87), CLT (departure MPIW: 17.71), and DEN (arrival MPIW: 14.13) exhibit the widest prediction intervals, reflecting the inherent uncertainty in predicting capacity at complex operational environments. Conversely, relatively smaller airports like MDW (arrival MPIW: 3.01) and TPA (arrival MPIW: 3.08) demonstrate narrower intervals, suggesting more predictable capacity patterns. The inherent prediction uncertainties observed across these diverse airport environments motivate the need for distributionally robust optimization procedures to effectively counter these prediction challenges in subsequent capacity allocation decisions.

Moreover, Figure 4 illustrates the model's capacity distribution prediction performance through two case studies from November 29, 2019. The probability heatmaps demonstrate that the model successfully captures temporal patterns of airport capacity variations, with higher probability concentrations during peak operational hours (06:00-22:00) and appropriate uncertainty quantification during transitional periods. The predicted distributions show concentrated probability

Index	Airport	RMSE (\downarrow)	MAE (\downarrow)	PICP (%) (\uparrow)	MPIW (\downarrow)
1	ATL	6.81	5.24	60.42	10.82
2	BOS	2.36	1.79	65.95	4.43
3	BWI	1.73	1.23	67.35	3.07
4	CLT	6.99	5.50	45.79	8.32
5	DCA	2.30	1.69	58.41	3.42
6	DEN	7.93	6.09	71.77	14.13
7	DFW	5.86	4.50	54.02	7.76
8	DTW	6.42	5.19	77.89	14.48
9	EWB	2.39	1.87	66.43	4.52
10	FLL	2.03	1.56	62.82	3.65
11	HNL	1.37	0.85	76.82	3.01
12	IAD	3.68	2.76	70.56	7.00
13	IAH	5.66	4.45	53.05	8.79
14	JFK	2.29	1.72	64.09	4.04
15	LAS	2.62	2.00	50.80	3.41
16	LAX	8.55	6.84	48.82	11.35
17	LGA	2.63	2.03	58.14	3.97
18	MCO	2.34	1.82	63.40	4.32
19	MDW	1.59	1.11	66.16	3.01
20	MEM	0.81	0.45	92.69	3.03
21	MIA	2.50	1.82	59.43	3.71
22	MSP	3.96	3.05	49.63	5.23
23	ORD	6.72	5.11	58.21	10.03
24	PHL	3.46	2.70	66.01	6.19
25	PHX	3.60	2.73	45.83	4.27
26	SAN	1.71	1.22	68.52	3.31
27	SEA	3.19	2.45	64.38	5.37
28	SFO	2.90	2.31	61.67	4.90
29	SLC	3.84	3.01	46.26	4.74
30	TPA	1.48	1.03	72.02	3.08

Note: Arrows indicate preferred direction: (\downarrow) lower is better, (\uparrow) higher is better. Airports follow the standard IATA codes. A complete list of airport names corresponding to IATA codes is provided in Appendix B.

Table 1: Arrival capacity distribution prediction performance for the core 30 US airports in the test set.

Index	Airport	RMSE (\downarrow)	MAE (\downarrow)	PICP (%) (\uparrow)	MPIW (\downarrow)
1	ATL	8.07	6.48	57.74	13.04
2	BOS	3.20	2.47	65.57	6.02
3	BWI	2.19	1.64	55.10	3.18
4	CLT	8.97	7.12	79.14	17.71
5	DCA	3.38	2.66	55.04	5.04
6	DEN	9.19	6.73	64.77	12.93
7	DFW	9.31	7.11	62.95	14.47
8	DTW	10.04	8.00	79.34	19.87
9	EWB	3.24	2.56	65.47	5.80
10	FLL	2.00	1.57	65.54	3.89
11	HNL	1.12	0.61	87.57	3.00
12	IAD	5.62	4.38	80.62	14.67
13	IAH	10.41	8.07	47.73	10.81
14	JFK	3.38	2.56	56.20	4.95
15	LAS	3.18	2.56	61.40	5.36
16	LAX	3.66	2.79	65.80	6.37
17	LGA	3.64	2.86	67.32	6.72
18	MCO	2.70	2.05	57.13	4.23
19	MDW	1.95	1.43	60.28	3.26
20	MEM	1.05	0.54	89.16	3.00
21	MIA	3.21	2.26	52.13	3.68
22	MSP	6.17	4.74	46.04	7.16
23	ORD	7.62	5.91	63.66	12.65
24	PHL	4.54	3.62	53.89	7.13
25	PHX	4.41	3.47	64.66	7.46
26	SAN	2.01	1.44	59.75	3.11
27	SEA	4.50	3.63	58.03	7.36
28	SFO	3.49	2.81	57.14	5.49
29	SLC	5.02	4.01	37.61	5.22
30	TPA	1.69	1.17	67.98	3.16

Note: Arrows indicate preferred direction: (\downarrow) lower is better, (\uparrow) higher is better. Airports follow the standard IATA codes. A complete list of airport names corresponding to IATA codes is provided in Appendix B.

Table 2: Departure capacity distribution prediction performance for the core 30 US airports in the test set.

mass around specific capacity values during stable operations, while becoming more dispersed during uncertain conditions. Comparing predicted distributions with actual throughput (black dots) and estimated capacity values (red triangles), the model shows reasonable alignment during morning and evening periods, but exhibits notable misalignment during midday operations where actual values frequently fall outside the high-probability regions. This visual evidence supports the quantitative findings of suboptimal PICP performance, indicating that while the model captures general capacity trends and temporal dependencies effectively, it requires enhanced uncertainty quantification to better account for operational disruptions and dynamic capacity-affecting factors for improved practical utility.

While the prediction results reveal certain limitations—such as coverage probabilities falling short of the nominal 90% level and occasional misalignments during disruptive periods—they nevertheless capture meaningful temporal patterns and distributional features of airport capacities. In the next section, we examine whether these imperfect yet informative predictions can still enhance downstream decision-making compared to purely deterministic approaches, and whether the distributionally robust DR-MAGHP can further mitigate the impact of prediction errors by providing more robust ground delay policies.

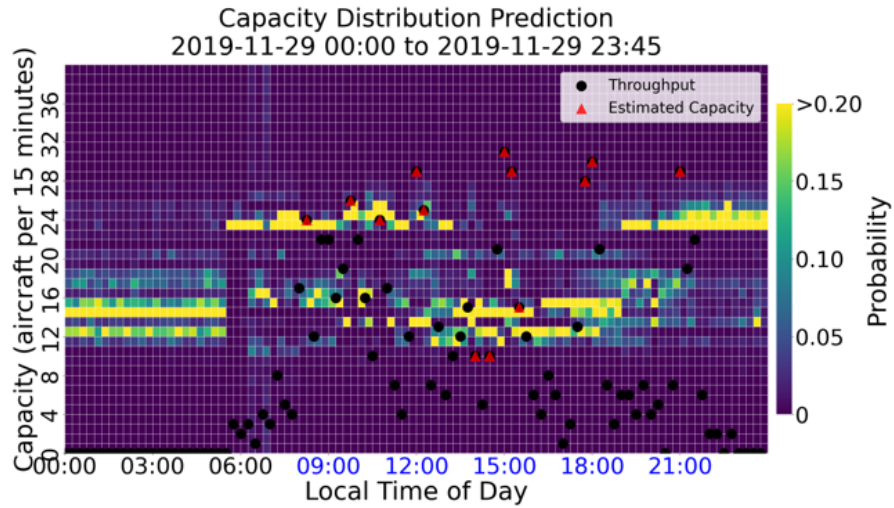
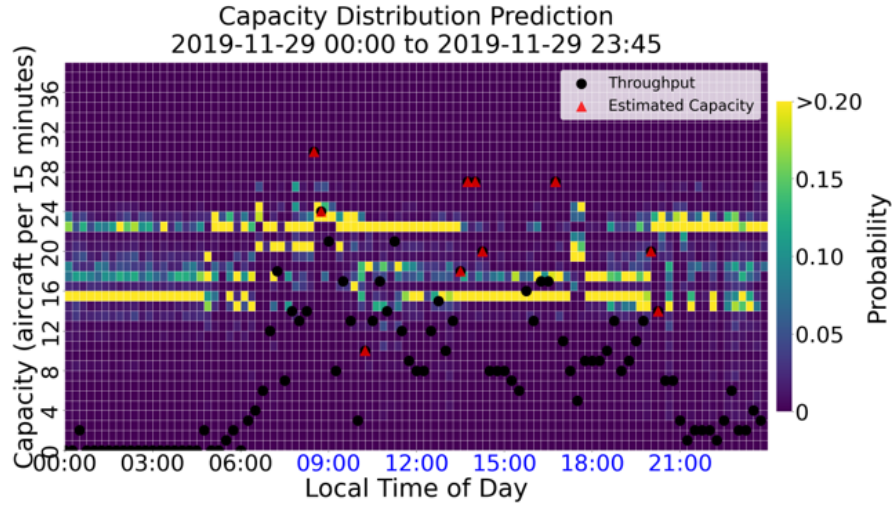


Figure 4: Capacity distribution prediction examples from November 29, 2019, showing predicted probability distributions (heatmap), actual throughput (black dots), and estimated capacity (red triangles) over a full operational day. The blue area from 9:00-21:00 is the 12-hour solution window of the DR-MAGHP.

4.3. Comparative Study of DET-MAGHP, SP-MAGHP and DR-MAGHP

Building on the predicted distributions from Section 3.1.2, we next evaluate how these predictions inform downstream decision-making. In particular, we compare the performance of prediction-driven models (SP-MAGHP and DR-MAGHP) against the deterministic baseline (DET-MAGHP) in generating ground delay policies. This comparison allows us to assess both (i) the extent to which upstream predictions improve decision quality relative to deterministic assumptions, and (ii) the additional robustness provided by DR-MAGHP when distribution shifts occur between predicted and realized capacities (e.g., realized capacities being systematically lower than what was predicted).

4.3.1. Experiment Setup

To ensure consistency with the methodology used for predicting capacity distributions, we also construct flight schedules using BTS flight data from the year 2019. Due to the computational complexity of solving the DET-MAGHP, SP-MAGHP, and DR-MAGHP models, our experiments focus on a selected set of representative days. To identify these days, we first compute the mean predicted capacity distribution for each airport and day over the full year. We then select the three dates exhibiting the greatest overestimation discrepancies, where predicted capacities were higher than the realized capacities. The selected dates used for evaluation are (in YYYY-MM-DD format): 2019-05-04, 2019-05-30, 2019-06-16, 2019-11-13, 2019-12-25, and 2019-12-26.

When computing airborne delays and enforcing connecting constraints, we restrict our attention to flights that depart from within the core 30 US airports. For flights originating from or arriving at out-of-network airports, we assume that those airports have unlimited arrival and departure capacities. As such, these flights are not subject to connecting constraints, and their airborne delays are calculated as the difference between scheduled and actual arrival times.

The planning horizon for the DR-MAGHP model spans 12 hours (from 09:00 to 21:00, expressed in the local time of each airport), discretized into 48 time periods ($|T| = 48$), each representing a 15-minute interval. We adopt a realistic minimum turnaround time of 45 minutes [1] and follow standard MAGHP formulations, including the use of an additional time period to accommodate excess flight volume [1]. In terms of cost modeling, we assume that the unit cost of airborne holding delay is three times higher than that of ground holding delay, thereby establishing a 3:1 cost ratio. Furthermore, the costs associated with second-stage departure and arrival delays are assumed to be equivalent to those of airborne holding. This assumption reflects two factors: (i) second-stage departure

delays typically involve aircraft waiting in taxiways or departure queues with engines running, leading to substantial fuel consumption[67]; and (ii) such delays disrupt gate assignment schedules, which in turn can generate additional airborne holding for arriving flights.

For the Wasserstein ambiguity sets, we impose a uniform radius ϵ across all airports. The distance between any two scenarios in the scenario tree is measured using the ℓ_2 norm. This metric is then used as the basis for constructing the Wasserstein ambiguity sets. Because the magnitude of inter-scenario distances is inherently scale-dependent, we normalize these distances prior to constructing the Wasserstein ambiguity sets. This procedure ensures that the ambiguity radius ϵ captures relative variations in the predicted capacity distributions, rather than being driven by the absolute scale of the underlying capacity values. Consequently, ϵ can be interpreted as a dimensionless measure of robustness, which facilitates consistent comparisons across airports and experimental settings with differing capacity ranges.

4.3.2. Evaluation Framework

We define the outputs from the upstream capacity distribution prediction procedure as predicted capacity distributions. We test the performance of the DR-MAGHP by comparing outcomes when the predicted capacity distributions, derived from the upstream prediction model, differ from the realized capacity distributions, which we adopt as testing distributions. Testing distributions of interest are the cases where realized capacity distributions are shifted to the left, i.e., the realized probabilistic capacities are lower than anticipated. We also solve the DET-MAGHP and SP-MAGHP to compare their delay costs with those of the DR-MAGHP when predicted capacity distributions are not accurate. We use the terms *in-sample performance* (ϕ_{IS}) and *out-of-sample performance* (ϕ_{OS}) to refer to the costs (i.e., optimal value in expectation) of DET-MAGHP, SP-MAGHP and DR-MAGHP when evaluated on predicted distributions and testing distributions respectively. Finally, we emphasize that these comparisons must be made on a day-by-day basis, just as unique GDP policies must be developed for each encountered NAS scenario.

4.3.3. Sensitivity Analysis Framework

A sensitivity analysis is conducted to assess the impact of discrepancies between predicted and realized capacity distributions on delay costs, as well as the ability of DR-MAGHP to mitigate these impacts. This analysis generates testing distributions at various levels of capacity reductions to compare the out-of-sample

Algorithm 3 Capacity Resampling Algorithm

- 1: **Input:** Predicted PMF $\{\hat{p}^{(1)}, \hat{p}^{(2)}, \dots, \hat{p}^{(|Z|)}\}$ of each airport's predicted capacity $\{\hat{\xi}^{(1)}, \hat{\xi}^{(2)}, \dots, \hat{\xi}^{(|Z|)}\}$ with mean values $\{\hat{\mu}^{(1)}, \hat{\mu}^{(2)}, \dots, \hat{\mu}^{(|Z|)}\}$, capacity reduction level r for all airports and a maximum variability rate of probability δ .
- 2: **for** each airport z **do**
- 3: Target mean $\mu^* = \hat{\mu}^{(z)} (1 - r)$
- 4: Update weights for each support $p^{(z)}$:

$$\begin{aligned}
p^{(z)} &= \arg \min_p \sum_{i=1}^{|\hat{\xi}^{(z)}|} p_i \hat{\xi}_i^{(z)} \\
\text{s.t.} \quad &\sum_{i=1}^{|\hat{\xi}^{(z)}|} p_i \hat{\xi}_i^{(z)} \geq \mu^*, \quad \sum_{i=1}^{|\hat{\xi}^{(z)}|} p_i = 1, \\
&p_i - \hat{p}_i^{(z)} \leq \delta \hat{p}_i^{(z)}, \quad -\delta \hat{p}_i^{(z)} \leq p_i - \hat{p}_i^{(z)}, \quad \forall i \in N.
\end{aligned} \tag{23}$$

- 5: **end for**
 - 6: Draw i.i.d samples $\tilde{\xi}^{(z)} \sim p^{(z)}, z = 1, 2, \dots, |Z|$
 - 7: **Output:** Reduced testing capacities for all airports $\{\tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \dots, \tilde{\xi}^{(|Z|)}\}$
-

Day	Reduction	Det.	Stoch.	DR	%↓ vs Det.	%↓ vs Stoch.	€*
2019-11-13	10%	1.6×10^5	9.4×10^4	9.4×10^4	42.5%	-0.1%	0.00
	20%	2.1×10^5	1.2×10^5	1.2×10^5	43.7%	3.4%	0.04
	30%	2.5×10^5	1.5×10^5	1.4×10^5	44.2%	7.3%	0.09
	40%	3.0×10^5	1.8×10^5	1.6×10^5	45.0%	10.6%	0.10
	50%	3.4×10^5	2.2×10^5	1.9×10^5	44.0%	12.8%	0.10
2019-12-25	10%	9.5×10^4	5.7×10^4	5.7×10^4	40.2%	0.1%	0.00
	20%	1.2×10^5	7.4×10^4	7.0×10^4	42.0%	4.8%	0.05
	30%	1.5×10^5	9.2×10^4	8.3×10^4	44.0%	10.3%	0.06
	40%	1.8×10^5	1.2×10^5	1.0×10^5	44.7%	14.1%	0.06
	50%	2.2×10^5	1.5×10^5	1.3×10^5	42.5%	14.9%	0.07
2019-12-26	10%	1.4×10^5	8.6×10^4	8.5×10^4	40.0%	0.8%	0.02
	20%	1.8×10^5	1.1×10^5	1.1×10^5	40.9%	5.2%	0.03
	30%	2.3×10^5	1.4×10^5	1.3×10^5	43.4%	10.7%	0.09
	40%	2.7×10^5	1.8×10^5	1.5×10^5	44.4%	14.2%	0.10
	50%	3.2×10^5	2.1×10^5	1.8×10^5	43.5%	15.6%	0.10
2019-05-04	10%	1.3×10^5	7.0×10^4	7.0×10^4	45.4%	0.8%	0.03
	20%	1.6×10^5	9.0×10^4	8.6×10^4	46.0%	5.1%	0.04
	30%	1.9×10^5	1.2×10^5	1.1×10^5	45.1%	7.8%	0.04
	40%	2.3×10^5	1.4×10^5	1.3×10^5	42.8%	9.0%	0.05
	50%	2.7×10^5	1.8×10^5	1.6×10^5	39.7%	10.0%	0.09
2019-05-30	10%	1.7×10^5	9.6×10^4	9.6×10^4	44.5%	0.5%	0.01
	20%	2.2×10^5	1.3×10^5	1.2×10^5	44.9%	4.4%	0.06
	30%	2.7×10^5	1.6×10^5	1.5×10^5	45.6%	10.0%	0.06
	40%	3.2×10^5	2.0×10^5	1.7×10^5	45.0%	13.1%	0.10
	50%	3.6×10^5	2.4×10^5	2.1×10^5	43.1%	14.2%	0.10
2019-06-16	10%	1.6×10^5	9.5×10^4	9.5×10^4	41.7%	-0.0%	0.00
	20%	2.1×10^5	1.2×10^5	1.1×10^5	44.7%	5.7%	0.05
	30%	2.6×10^5	1.6×10^5	1.4×10^5	46.1%	10.9%	0.06
	40%	3.0×10^5	1.9×10^5	1.7×10^5	45.0%	12.8%	0.10
	50%	3.4×10^5	2.3×10^5	2.0×10^5	42.2%	12.5%	0.10

Table 3: Out-of-sample cost comparison across different days and capacity reductions. Columns show deterministic (Det.), stochastic (Stoch.), and distributionally robust (DR) MAGHP. Bold values indicate the lowest (best) cost in each row.

performance of DET-MAGHP, SP-MAGHP, and DR-MAGHP. Ground delay policies generated by SP-MAGHP and DR-MAGHP are constructed using the full predicted capacity distributions, capturing the uncertainty in future capacities. In contrast, DET-MAGHP derives its policies from the single best-capacity scenario of the day, reflecting an optimistic assumption that airports will operate at their maximum predicted capacities.

4.3.4. Construction of Reduced Capacity Distributions

To sample from reduced capacity distributions at various reduction levels, we introduce a linear program in Equation (23) that performs valid adjustments to the PMFs to minimize the deviation between the current mean value and the targeted (reduced) mean value, while maintaining the probabilistic properties of the weights. We also introduce a parameter δ for the maximum variability rate to ensure a more uniform distribution of probability mass.

The details of the sampling procedure are presented in Algorithm 3. Given a general ground delay policy \tilde{x} , we generate 100 samples from the reduced capacity distributions from Algorithm 3 and compute the out-of-sample performance using the expression $\phi_{OS}(\tilde{x}) = \sum_{i=1}^{|\tilde{\xi}|} \widetilde{\phi_{OS}}(\tilde{x}, \tilde{\xi}_i) / |\tilde{\xi}|$, where $\widetilde{\phi_{OS}}(\tilde{x}, \tilde{\xi}_i)$ denotes the objective value of the proposed ground holding policy \tilde{x} under the i^{th} capacity sample $\tilde{\xi}_i$. We emphasize that $\widetilde{\phi_{OS}}$ corresponds to an optimization model structurally equivalent to SP-MAGHP, sharing the same objective function and constraint set. The key distinction lies in the fact that $\widetilde{\phi_{OS}}$ accepts a fixed first-stage ground delay decision \tilde{x} as input, and evaluates its second-stage performance under a specified realization of the capacity distribution.

4.3.5. In-Sample vs. Out-of-Sample Performance

We investigate the impact of erroneous capacity predictions, specifically cases in which realized capacities are reduced by 10% to 50% relative to predicted capacity distributions. To evaluate both in-sample and out-of-sample performance of DR-MAGHP under varying levels of distributional robustness, we conduct a sweep over the Wasserstein radius ϵ , ranging from 0 to 0.1. It is important to note that, since inter-scenario distances in the scenario tree are normalized, even relatively small values of ϵ can induce substantial changes in the objective function. Intuitively, normalization compresses the scale of distances so that scenarios appear closer together, which makes the Wasserstein ball “tighter” around the predicted distribution. As a result, a modest increase in ϵ is sufficient to admit a much larger set of probability distributions into the ambiguity set, thereby altering

the optimization landscape and, in turn, the optimal objective function value (i.e., ground delay policy cost).

We begin by examining the in-sample performance of SP-MAGHP and DR-MAGHP across all six selected evaluation dates, using the predicted capacity distributions. As illustrated in Figure D.5, both models yield identical objective values when $\epsilon = 0$. This is expected, since in this case the Wasserstein ambiguity set collapses to a trivial set with one element, which is exactly the predicted distribution. Solving DR-MAGHP over this trivial ambiguity set is equivalent to solving the original SP-MAGHP problem. As ϵ increases from 0 to 0.1, we observe a consistent rise in the objective values of DR-MAGHP across all test dates. This trend indicates that greater distributional robustness—i.e., larger ambiguity sets—leads to deteriorated in-sample performance when evaluated on the nominal predicted capacities. However, with regard to out-of-sample performance, when the realized capacities deviate only slightly from the predictions (i.e., a 10% reduction), a small ambiguity set (i.e., small ϵ) is sufficient for DR-MAGHP to moderately outperform SP-MAGHP on most evaluation dates, including on 2019-12-25, 2019-12-26, 2019-05-04, and 2019-05-30. However, for 2019-11-13 and 2019-06-16, the out-of-sample performance of SP-MAGHP is slightly better than that of DR-MAGHP across all tested values of ϵ .

4.3.6. *Effect of the Wasserstein Radius ϵ*

It is important to recognize that the out-of-sample performance of SP-MAGHP and DR-MAGHP may diverge, even when their in-sample objective values are identical under a Wasserstein radius of $\epsilon = 0$. This divergence arises from differences in the delay assignment policies produced by the two models, which, despite being theoretically equivalent at $\epsilon = 0$, can vary due to numerical tolerances and solver precision. As a result, matching in-sample performance does not necessarily imply identical out-of-sample behavior.

Furthermore, as the gap between realized and predicted capacity distributions widens, larger ambiguity sets (i.e., higher values of ϵ) are typically needed for DR-MAGHP to deliver improved out-of-sample performance. Table 3 summarizes the out-of-sample delay assignment costs associated with DET-MAGHP, SP-MAGHP, and DR-MAGHP across varying levels of capacity reductions, along with the corresponding optimal ambiguity set radius ϵ^* . Notably, Table 3 reveals a monotonic increase in the optimal ϵ^* as the severity of the capacity reduction intensifies. Additionally, we observe that the out-of-sample cost of DR-MAGHP tends to increase again once ϵ exceeds its optimal value ϵ^* . Although this trend may plateau at higher levels due to the experimental cap of $\epsilon = 0.1$, it underscores

the trade-off between robustness and conservativeness inherent in the design of distributionally robust policies.

4.3.7. Comparative Performance under Capacity Reductions

Table 3 further reveals that under substantial capacity reductions—specifically, when realized capacities are 30%, 40%, or 50% lower than the predicted values—the DR-MAGHP model achieves notable cost savings relative to even the SP-MAGHP, ranging from 7.30% to 10.85% at 30% reductions, 8.97% to 14.21% at 40% reductions, and 9.95% to 15.58% at 50% reductions. These reductions represent cases where the predictions are overly optimistic compared to the realized (testing) distributions. Moreover, the improvements are even more pronounced when compared with DET-MAGHP, exceeding 40% in all tested scenarios. These results demonstrate that robustified ground delay policies can yield meaningful cost reductions under adverse operational conditions.

To better understand this effect, we examine the results in Tables C.6 and C.7. These tables show that policies derived from DR-MAGHP consistently lead to lower second-stage departure and arrival delays, with the effect being more pronounced for departures. Because second-stage delays carry higher unit costs than first-stage delays, the conservative nature of DR-MAGHP—which may increase first-stage delay costs—ultimately reduces overall costs by mitigating more expensive downstream disruptions, particularly in scenarios with high degrees of disruptions across the NAS.

4.3.8. Key Findings

In summary, the numerical experiments provide three key insights: (1) leveraging airport capacity predictions enables data-driven models such as SP-MAGHP and DR-MAGHP to reduce operational costs; (2) when capacity predictions are expected to be reliable, SP-MAGHP may be preferable due to its cost efficiency, whereas in more uncertain conditions, DR-MAGHP provides more robust performance; and (3) the choice of the Wasserstein radius ϵ plays a critical role, balancing conservativeness and robustness in decision making.

5. Concluding Remarks

5.1. Summary

This paper proposed an integrated framework that combines learning-based prediction and Distributionally Robust Optimization (DRO) for Air Traffic Management (ATM), with a focus on flight rescheduling decision-making within the

context of Ground Delay Programs (GDP), given uncertainty in airport arrival and departure capacities. The upstream module utilizes a machine learning model to generate probabilistic airport capacity predictions across the planning horizon. These predictions are then used by the downstream DR-MAGHP model to construct Wasserstein ambiguity sets, enabling the generation of distributionally robust ground delay policies.

Our results show that prediction-based models such as SP-MAGHP and DR-MAGHP substantially reduce operational costs compared to the baseline DET-MAGHP model that does not incorporate predictive information. Notably, DR-MAGHP demonstrates superior performance under distribution shifts, highlighting its ability to better manage uncertainty in airport capacities. This work thus bridges data-driven prediction and robust optimization for strategic air traffic management decision-making, offering tractable reformulations and scalable scenario reduction techniques for real-world deployment. It also identifies the conditions under which DRO-based models outperform traditional stochastic optimization approaches.

5.2. *Limitations and Future Work*

While our approach demonstrates strong performance on data from the FAA Core 30 airports, it is built on several simplifying assumptions that may limit its fidelity to real-world operations. First, we assume independence between arrival and departure capacities at each airport, as well as independence across airports. In reality, operational dependencies are common: adverse weather or other disruptions often reduce both arrival and departure capacities simultaneously at a single airport, and severe regional weather phenomena can affect multiple neighboring airports at once. Second, we impose a uniform value of ϵ across all airports when constructing ambiguity sets. This assumption overlooks the heterogeneity of uncertainty across airports, as different airports face distinct traffic patterns, weather conditions, and operational environments that may warrant different ambiguity set sizes. Finally, due to limited access to historical airport capacity data, model evaluation was conducted using synthetically generated capacity reduction scenarios, which may not fully replicate the complexity of real-world disruptions.

Future research will focus on several extensions to enhance the realism and applicability of the proposed framework. First, we aim to model the correlation between arrival and departure capacities at individual airports. Second, we plan to incorporate inter-airport correlations to better capture the regional impacts of weather and operational disruptions. Third, we will explore assigning heterogeneous levels of uncertainty to different airports, allowing the ambiguity set size

to reflect local traffic patterns, weather variability, and operational characteristics, and investigate the trade-off between efficiency and robustness for each airport. Finally, we will evaluate the proposed DRO models on historical disruption days to assess their ability to mitigate the negative impacts of adverse events. Together, these directions are expected to strengthen the practical relevance of the framework for air traffic management.

Appendix A. Number of time periods remaining after rule-based capacity estimation for each airport.

Index	Airport	Arrival		Departure	
		Time Periods	%	Time Periods	%
1	ATL	12,816	36.6%	14,120	40.3%
2	BOS	10,023	28.6%	9,730	27.8%
3	BWI	6,758	19.3%	8,430	24.1%
4	CLT	9,503	27.1%	11,578	33.0%
5	DCA	9,065	25.9%	8,467	24.2%
6	DEN	13,004	37.1%	13,714	39.1%
7	DFW	12,548	35.8%	13,958	39.8%
8	DTW	7,556	21.6%	7,548	21.5%
9	EWR	12,279	35.0%	11,472	32.7%
10	FLL	6,848	19.5%	7,616	21.7%
11	HNL	5,370	15.3%	4,323	12.3%
12	IAD	4,981	14.2%	4,548	13.0%
13	IAH	8,288	23.7%	8,154	23.3%
14	JFK	7,813	22.3%	7,744	22.1%
15	LAS	10,060	28.7%	10,625	30.3%
16	LAX	14,202	40.5%	13,421	38.3%
17	LGA	11,891	33.9%	11,078	31.6%
18	MCO	8,673	24.8%	10,014	28.6%
19	MDW	6,365	18.2%	8,360	23.9%
20	MEM	7,391	21.1%	7,101	20.3%
21	MIA	5,394	15.4%	5,929	16.9%
22	MSP	7,192	20.5%	6,722	19.2%
23	ORD	16,395	46.8%	16,820	48.0%
24	PHL	8,173	23.3%	7,962	22.7%
25	PHX	9,721	27.7%	9,714	27.7%
26	SAN	6,989	19.9%	7,312	20.9%
27	SEA	9,784	27.9%	9,905	28.3%
28	SFO	12,682	36.2%	10,931	31.2%
29	SLC	6,059	17.3%	5,400	15.4%
30	TPA	5,737	16.4%	6,239	17.8%
Total		270,660	25.7%	280,376	26.7%

Table A.4: Number of time periods remaining after rule-based capacity estimation for each airport.

**Appendix B. IATA airport codes and corresponding airport names of the
core 30 US airports**

Index	IATA Code	Airport Name
1	ATL	Hartsfield-Jackson Atlanta International Airport
2	BOS	Logan International Airport
3	BWI	Baltimore/Washington International Airport
4	CLT	Charlotte Douglas International Airport
5	DCA	Ronald Reagan Washington National Airport
6	DEN	Denver International Airport
7	DFW	Dallas/Fort Worth International Airport
8	DTW	Detroit Metropolitan Wayne County Airport
9	EWB	Newark Liberty International Airport
10	FLL	Fort Lauderdale-Hollywood International Airport
11	HNL	Daniel K. Inouye International Airport
12	IAD	Washington Dulles International Airport
13	IAH	George Bush Intercontinental Airport
14	JFK	John F. Kennedy International Airport
15	LAS	McCarran International Airport
16	LAX	Los Angeles International Airport
17	LGA	LaGuardia Airport
18	MCO	Orlando International Airport
19	MDW	Chicago Midway International Airport
20	MEM	Memphis International Airport
21	MIA	Miami International Airport
22	MSP	Minneapolis-Saint Paul International Airport
23	ORD	Chicago O'Hare International Airport
24	PHL	Philadelphia International Airport
25	PHX	Phoenix Sky Harbor International Airport
26	SAN	San Diego International Airport
27	SEA	Seattle-Tacoma International Airport
28	SFO	San Francisco International Airport
29	SLC	Salt Lake City International Airport
30	TPA	Tampa International Airport

Table B.5: IATA airport codes and corresponding airport names.

Appendix C. Second stage delay comparison

Day	Reduction	Det.	Stoch.	DR	%↓ vs Det.	%↓ vs Stoch.
2019-11-13	10%	1.9×10^4	8.4×10^3	8.5×10^3	54.8%	-0.5%
	20%	2.4×10^4	1.3×10^4	1.1×10^4	54.6%	13.1%
	30%	2.9×10^4	1.7×10^4	1.3×10^4	54.7%	20.7%
	40%	3.4×10^4	2.0×10^4	1.6×10^4	52.4%	21.3%
	50%	3.9×10^4	2.5×10^4	2.0×10^4	48.5%	19.8%
2019-12-25	10%	1.1×10^4	4.4×10^3	4.3×10^3	58.6%	0.5%
	20%	1.4×10^4	6.9×10^3	5.3×10^3	61.9%	22.7%
	30%	1.7×10^4	9.2×10^3	7.0×10^3	59.4%	24.5%
	40%	2.1×10^4	1.3×10^4	9.5×10^3	55.2%	23.5%
	50%	2.5×10^4	1.6×10^4	1.2×10^4	51.7%	23.0%
2019-12-26	10%	1.4×10^4	6.9×10^3	6.5×10^3	54.7%	5.9%
	20%	1.9×10^4	1.1×10^4	9.1×10^3	52.5%	13.4%
	30%	2.4×10^4	1.4×10^4	1.1×10^4	54.5%	21.4%
	40%	2.9×10^4	1.8×10^4	1.4×10^4	52.2%	22.3%
	50%	3.4×10^4	2.2×10^4	1.7×10^4	48.3%	20.8%
2019-05-04	10%	1.3×10^4	6.1×10^3	5.5×10^3	58.8%	10.0%
	20%	1.7×10^4	8.8×10^3	7.5×10^3	56.2%	14.9%
	30%	2.2×10^4	1.3×10^4	1.0×10^4	52.7%	16.9%
	40%	2.7×10^4	1.7×10^4	1.4×10^4	49.3%	17.8%
	50%	3.2×10^4	2.1×10^4	1.6×10^4	48.7%	22.7%
2019-05-30	10%	1.8×10^4	8.9×10^3	8.4×10^3	53.1%	5.7%
	20%	2.4×10^4	1.3×10^4	1.1×10^4	55.5%	19.4%
	30%	2.9×10^4	1.7×10^4	1.4×10^4	52.7%	18.5%
	40%	3.5×10^4	2.1×10^4	1.7×10^4	52.3%	22.3%
	50%	4.0×10^4	2.6×10^4	2.1×10^4	47.9%	20.0%
2019-06-16	10%	1.9×10^4	8.8×10^3	8.8×10^3	53.6%	0.5%
	20%	2.4×10^4	1.3×10^4	1.1×10^4	54.6%	12.4%
	30%	3.0×10^4	1.7×10^4	1.4×10^4	52.5%	14.5%
	40%	3.5×10^4	2.1×10^4	1.7×10^4	52.3%	20.2%
	50%	4.0×10^4	2.5×10^4	2.1×10^4	47.9%	18.0%

Table C.6: Second-stage *arrival delay* comparison (in delay units) among deterministic (Det.), stochastic (Stoch.), and distributionally robust (DR) MAGHP across different days and capacity reductions. Bold values denote the lowest delay in each row.

Day	Reduction	Det.	Stoch.	DR	%↓ vs Det.	%↓ vs Stoch.
2019-11-13	10%	3.4×10^4	6.0×10^3	6.0×10^3	82.3%	-0.2%
	20%	4.4×10^4	1.1×10^4	7.9×10^3	81.9%	28.7%
	30%	5.4×10^4	1.8×10^4	8.1×10^3	85.0%	53.8%
	40%	6.4×10^4	2.4×10^4	1.2×10^4	81.1%	49.5%
	50%	7.4×10^4	3.2×10^4	1.8×10^4	76.2%	44.7%
2019-12-25	10%	2.0×10^4	4.7×10^3	4.7×10^3	76.9%	0.0%
	20%	2.5×10^4	7.7×10^3	3.8×10^3	85.1%	51.0%
	30%	3.1×10^4	1.2×10^4	5.4×10^3	82.6%	53.2%
	40%	3.8×10^4	1.7×10^4	8.8×10^3	77.0%	47.0%
	50%	4.7×10^4	2.4×10^4	1.4×10^4	71.0%	42.4%
2019-12-26	10%	3.1×10^4	6.8×10^3	6.1×10^3	80.6%	10.0%
	20%	4.0×10^4	1.3×10^4	9.9×10^3	75.3%	21.1%
	30%	5.0×10^4	1.9×10^4	7.0×10^3	86.0%	63.2%
	40%	6.0×10^4	2.6×10^4	1.1×10^4	81.2%	56.8%
	50%	7.0×10^4	3.4×10^4	1.7×10^4	76.0%	50.2%
2019-05-04	10%	2.7×10^4	4.8×10^3	3.6×10^3	86.8%	25.5%
	20%	3.4×10^4	8.8×10^3	6.1×10^3	81.9%	30.1%
	30%	4.1×10^4	1.3×10^4	1.0×10^4	75.2%	24.8%
	40%	4.8×10^4	1.9×10^4	1.3×10^4	72.5%	31.2%
	50%	5.6×10^4	2.7×10^4	1.8×10^4	68.8%	34.5%
2019-05-30	10%	3.7×10^4	6.6×10^3	6.4×10^3	82.7%	2.0%
	20%	4.7×10^4	1.3×10^4	5.8×10^3	87.8%	54.3%
	30%	5.8×10^4	2.1×10^4	1.1×10^4	81.3%	47.3%
	40%	6.8×10^4	2.9×10^4	1.5×10^4	78.3%	48.3%
	50%	7.9×10^4	3.8×10^4	2.2×10^4	72.4%	42.4%
2019-06-16	10%	3.3×10^4	6.8×10^3	6.9×10^3	78.9%	-0.4%
	20%	4.2×10^4	1.2×10^4	6.3×10^3	85.0%	47.2%
	30%	5.3×10^4	1.9×10^4	1.0×10^4	80.6%	46.6%
	40%	6.3×10^4	2.7×10^4	1.4×10^4	78.6%	49.8%
	50%	7.2×10^4	3.5×10^4	2.0×10^4	72.0%	41.6%

Table C.7: Second-stage *departure delay* comparison (in delay units) among deterministic (Det.), stochastic (Stoch.), and distributionally robust (DR) MAGHP across different days and capacity reductions. Bold values denote the lowest delay in each row.

Appendix D. *Out-of-sample* performance of SP-MAGHP and DR-MAGHP

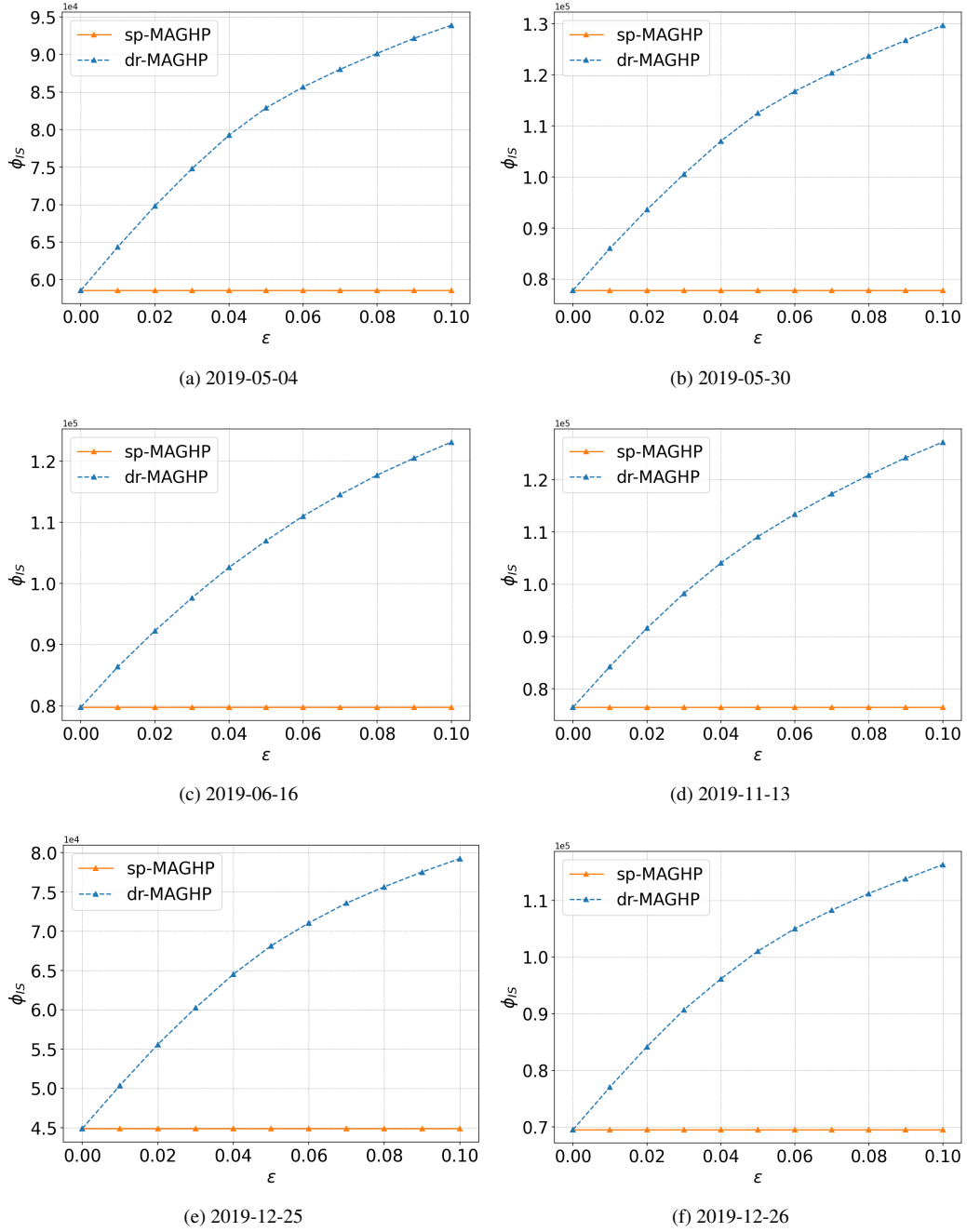


Figure D.5: *in-sample* performance of DR-MAGHP with different ϵ values under overestimation scenario on all testing dates.

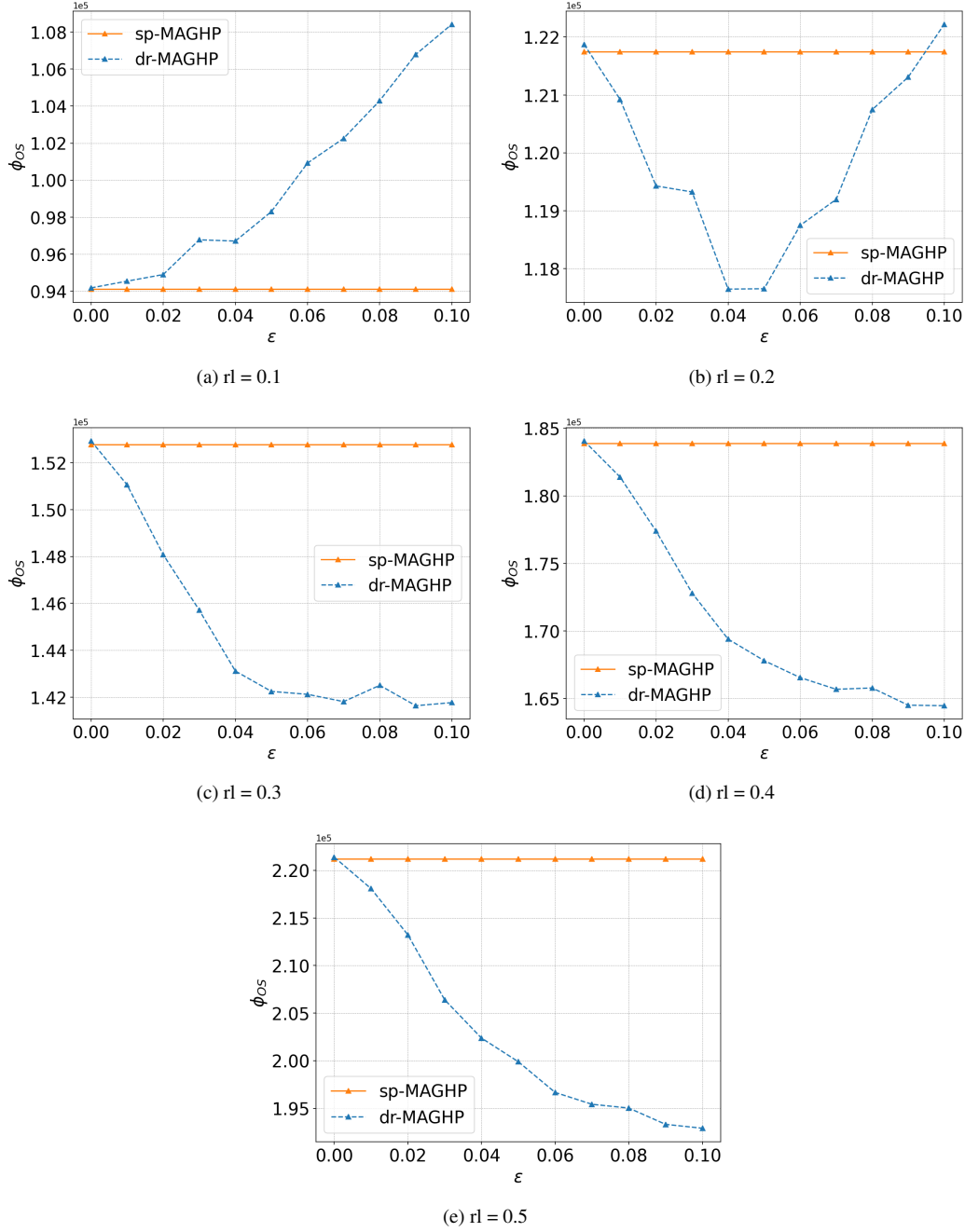


Figure D.6: *out-of-sample* performance of SP-MAGHP and DR-MAGHP with different ϵ values under overestimation scenario on 2019-11-13.

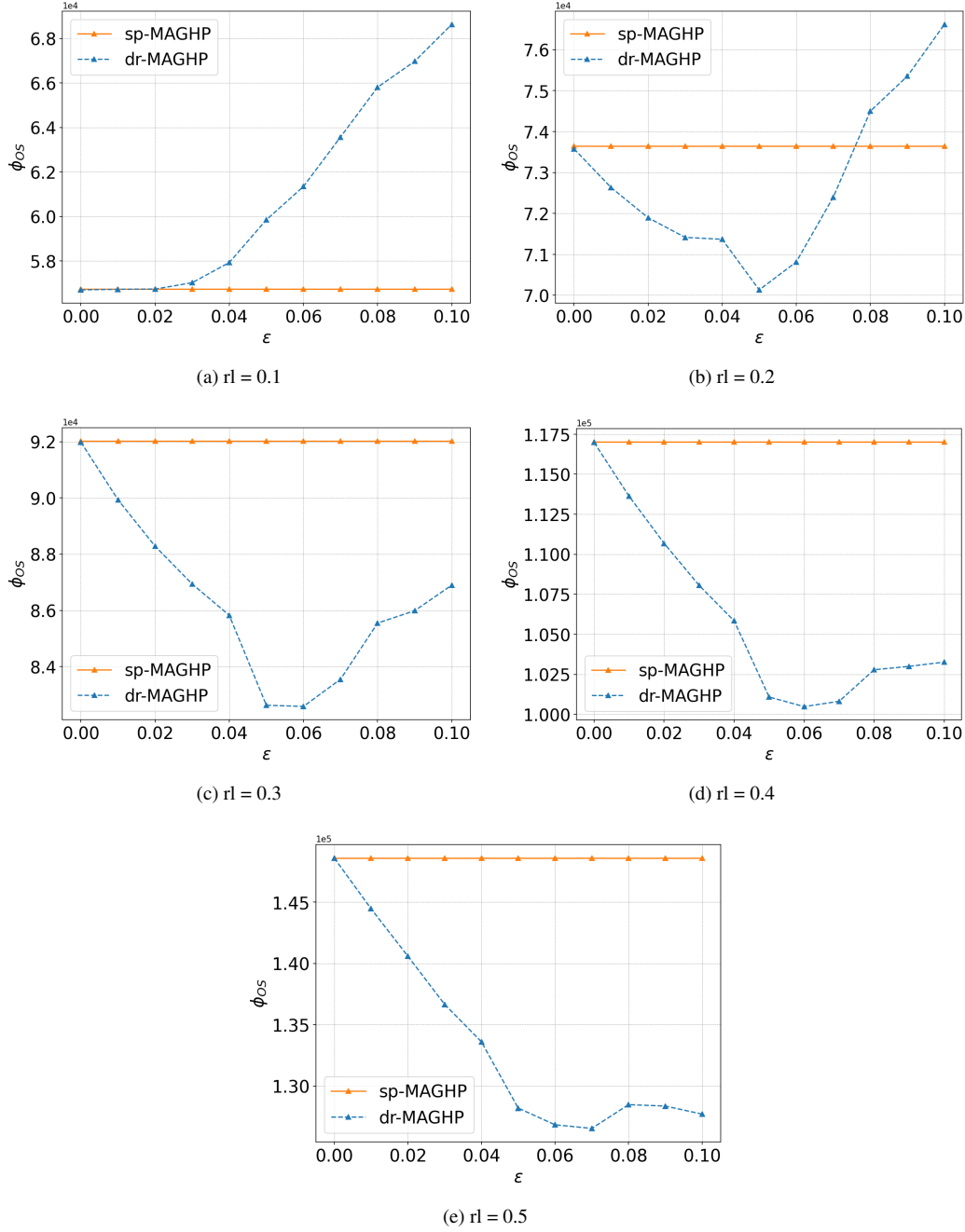


Figure D.7: *out-of-sample* performance of SP-MAGHP and DR-MAGHP with different ϵ values under overestimation scenario on 2019-12-25.

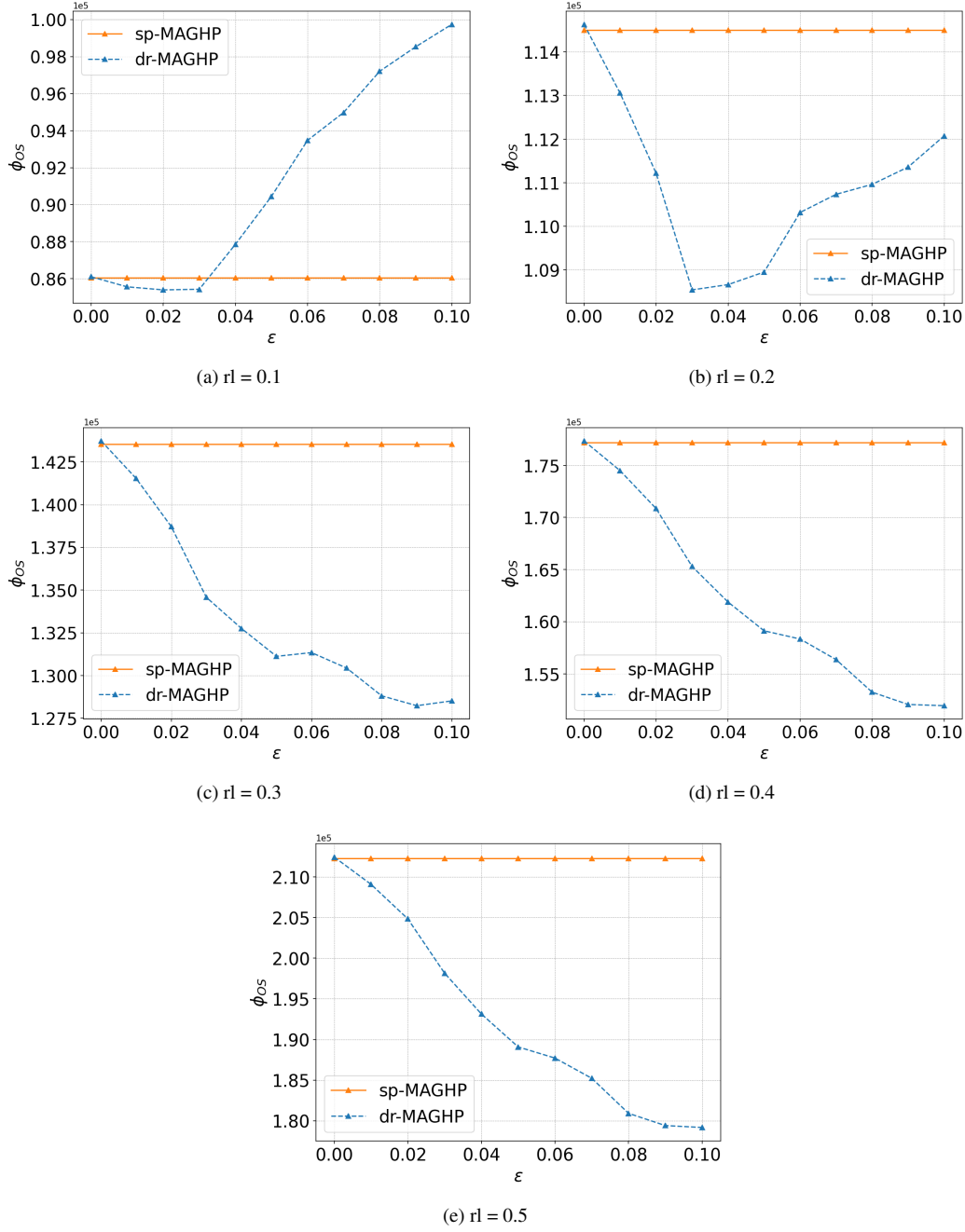


Figure D.8: *out-of-sample* performance of SP-MAGHP and DR-MAGHP with different ϵ values under overestimation scenario on 2019-12-26.

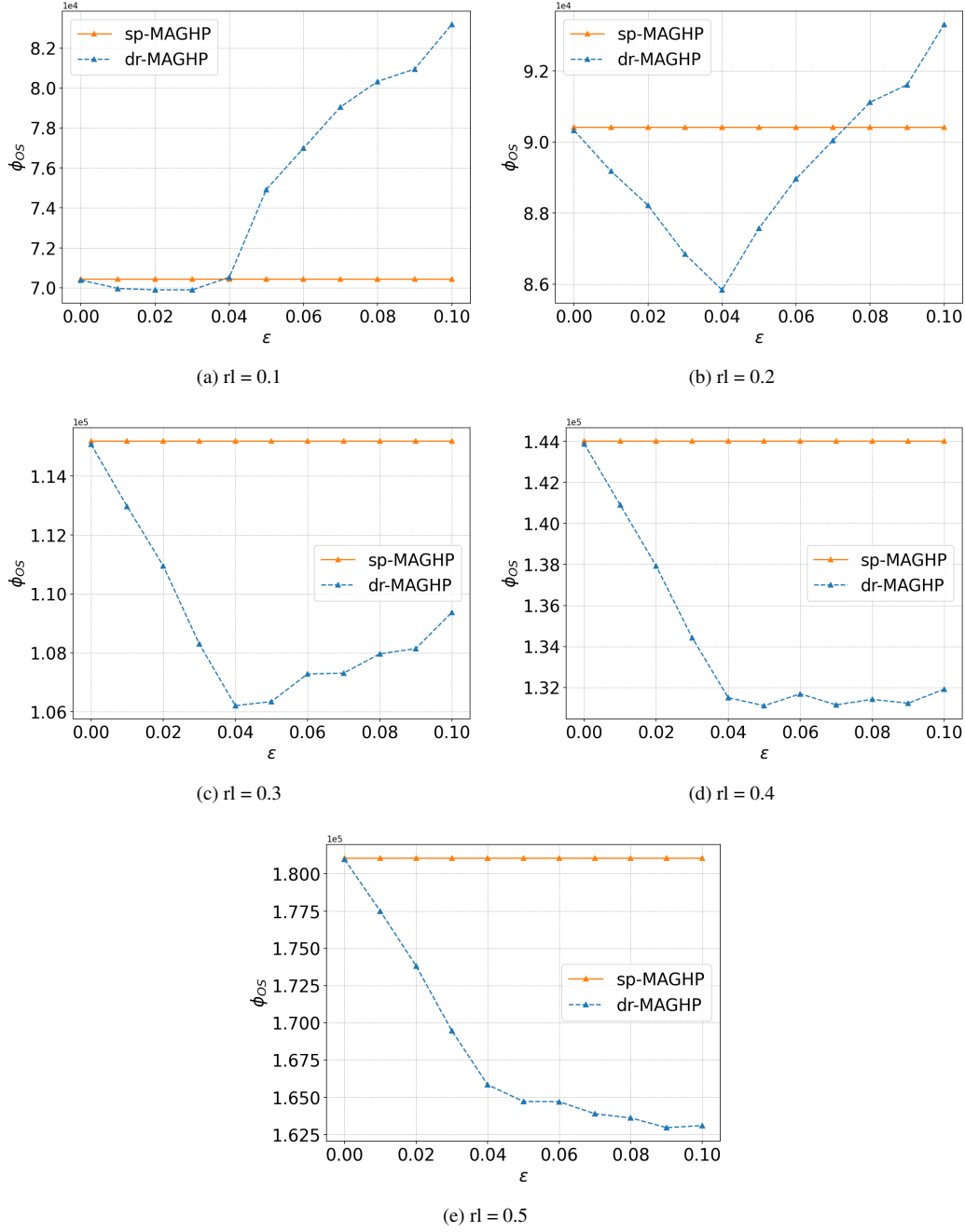


Figure D.9: *out-of-sample* performance of SP-MAGHP and DR-MAGHP with different ϵ values under overestimation scenario on 2019-05-04.

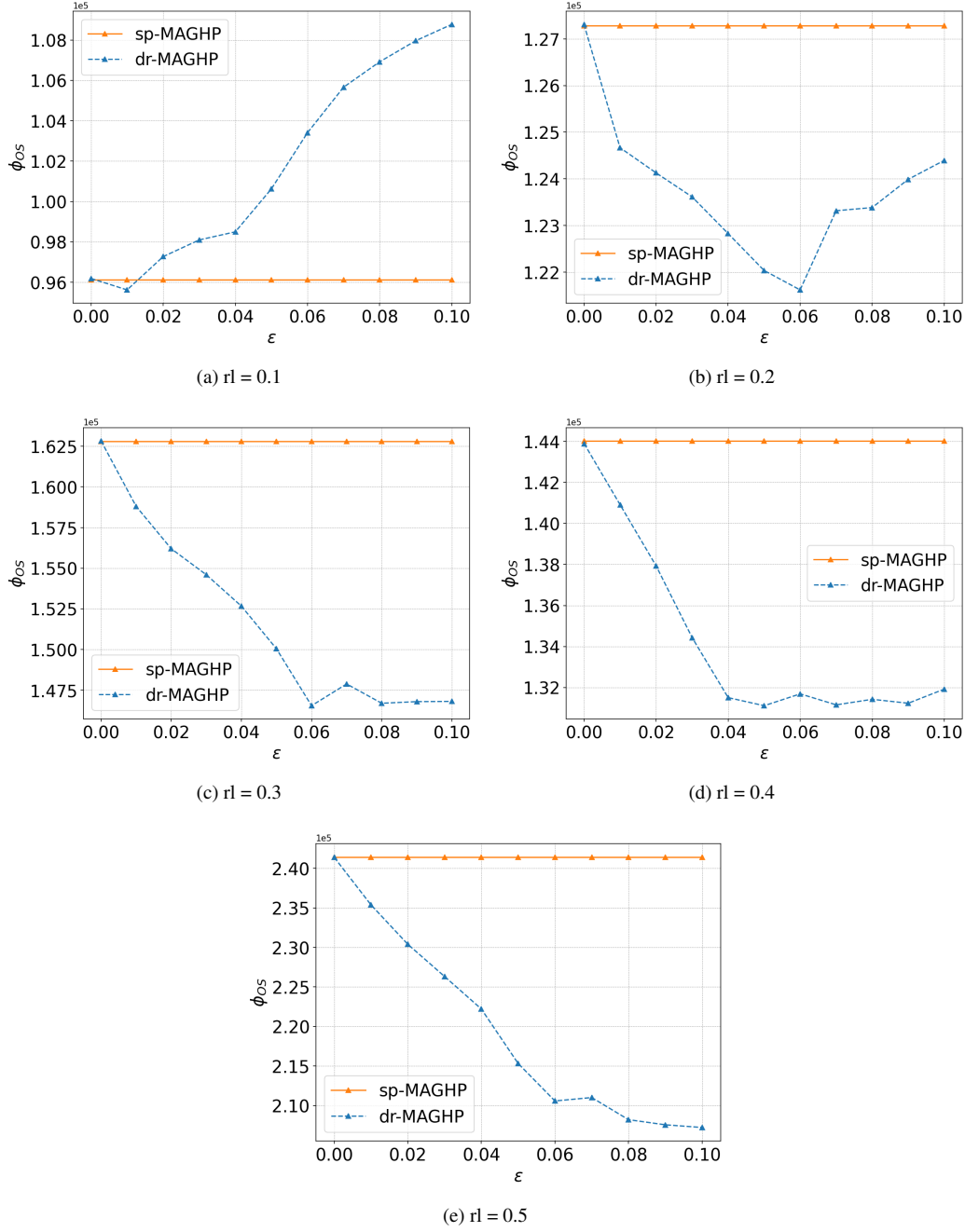


Figure D.10: *out-of-sample* performance of SP-MAGHP and DR-MAGHP with different ϵ values under overestimation scenario on 2019-05-30.

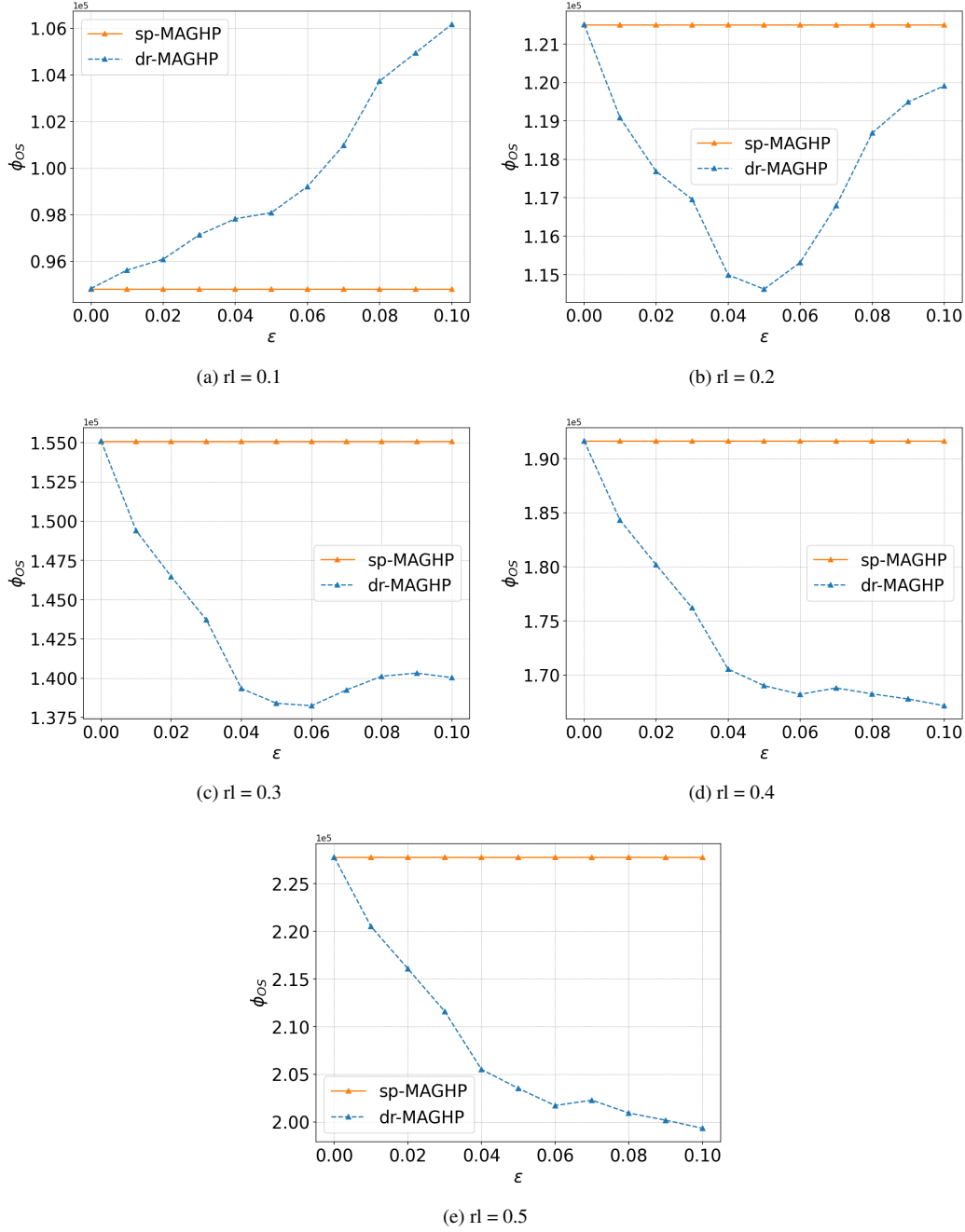


Figure D.11: *out-of-sample* performance of SP-MAGHP and DR-MAGHP with different ϵ values under overestimation scenario on 2019-06-16.

References

- [1] P. B. Vranas, D. J. Bertsimas, and A. R. Odoni, “The multi-airport ground-holding problem in air traffic control,” *Operations Research*, vol. 42, no. 2, pp. 249–261, 1994.
- [2] R. Kicing, J.-T. Chen, M. Steiner, and J. Pinto, “Airport capacity prediction with explicit consideration of weather forecast uncertainty,” *AIAA JAT*, vol. 24, no. 1, pp. 18–28, 2016.
- [3] D. Tittle, P. McCarthy, and Y. Xiao, “Airport runway capacity and economic development: A panel data analysis of metropolitan statistical areas,” *Economic Development Quarterly*, vol. 27, no. 3, pp. 230–239, 2013.
- [4] J. Avery and H. Balakrishnan, “Predicting airport runway configuration: A discrete-choice modeling approach,” in *Eleventh USA/Europe Air Traffic Management Research and Development Seminar*. Lisbon, Portugal: Federal Aviation Administration/EUROCONTROL, June 23-26 2015. [Online]. Available: http://www.atmseminarus.org/seminarContent/seminar11/presentations/509-Balakrishnan_0126150652-PresentationPDF-6-29-15.pdf
- [5] K. Vlachou, R. Sharma, and F. Wieland, “Simultaneous traffic management initiatives: The double delay problem,” in *2019 IEEE/AIAA 38th DASC*. IEEE, 2019, pp. 1–6.
- [6] FAA, “Initial Concept of Operations for an Info-Centric National Airspace System,” 2022, accessed: Feb 2024.
- [7] SESAR, “European ATM Master Plan 2025 Edition,” 2025, accessed: Aug 2025.
- [8] X. Zhu and L. Li, “Flight time prediction for fuel loading decisions with a deep learning approach,” *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103179, 2021.
- [9] X. Zhu, Y. Lin, Y. He, K.-L. Tsui, P. W. Chan, and L. Li, “Short-term nationwide airport throughput prediction with graph attention recurrent neural network,” *Frontiers in Artificial Intelligence*, vol. 5, p. 884485, 2022.
- [10] Z. Wang, C. Liao, X. Hang, L. Li, D. Delahaye, and M. Hansen, “Distribution prediction of strategic flight delays via machine learning methods,” *Sustainability*, vol. 14, no. 22, 2022.

- [11] L. Tamang, M. R. Bouadjenek, R. Dazeley, and S. Aryal, “Handling out-of-distribution data: A survey,” *IEEE Transactions on Knowledge and Data Engineering*, 2025.
- [12] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, “Can autonomous vehicles identify, recover from, and adapt to distribution shifts?” in *International Conference on Machine Learning*. PMLR, 2020, pp. 3145–3153.
- [13] Z. Yang, X. He, J. Zhang, J. Wu, X. Xin, J. Chen, and X. Wang, “A generic learning framework for sequential recommendation with distribution shifts,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 331–340.
- [14] X. Wang, Q. Kang, J. An, and M. Zhou, “Drifted twitter spam classification using multiscale detection test on kl divergence,” *IEEE Access*, vol. 7, pp. 108 384–108 394, 2019.
- [15] H. Wu, X. Zhu, S. Li, Y. Zhou, L. Li, and M. Z. Li, “Distributionally robust ground delay programs with learning-driven airport capacity predictions,” *arXiv preprint arXiv:2402.11415*, 2024.
- [16] R. Kicing, J. Krozel, M. Steiner, and J. Pinto, “Airport capacity prediction integrating ensemble weather forecasts,” in *Infotech Aerospace 2012*, 2012, p. 2493.
- [17] E. P. Gilbo, “Airport capacity: Representation, estimation, optimization,” *IEEE Transactions on Control Systems Technology*, vol. 1, no. 3, pp. 144–154, 1993.
- [18] S. Choi and Y. J. Kim, “Artificial neural network models for airport capacity prediction,” *Journal of Air Transport Management*, vol. 97, p. 102146, 2021.
- [19] S.-L. Tien, C. Taylor, E. Vargo, and C. Wanke, “Using ensemble weather forecasts for predicting airport arrival capacity,” *Journal of Air Transportation*, vol. 26, no. 3, pp. 123–132, 2018.
- [20] J. Cox and M. J. Kochenderfer, “Probabilistic airport acceptance rate prediction,” in *AIAA Modeling and Simulation Technologies Conference*, 2016, p. 0165.
- [21] Y. Pang, N. Xu, and Y. Liu, “Data-driven trajectory prediction with weather uncertainties: A bayesian deep learning approach,” *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103326, 2021.
- [22] M. Zoutendijk and M. Mitici, “Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem,” *Aerospace*, vol. 8, no. 6, p. 152, 2021.

- [23] A. Masaloni, S. Mulgund, L. Song, C. Wanke, and S. Zobell, "Using probabilistic demand predictions for traffic flow management decision support," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2004, p. 5231.
- [24] X. Zhang, J. Chen, H. Yang, M. Li, and Y. Wang, "Air traffic density prediction using bayesian ensemble graph attention network (began)," *Transportation Research Part C: Emerging Technologies*, vol. 152, p. 104140, 2023.
- [25] R. Dalmau, P. De Falco, M. Spak, and J. D. Rodriguez-Varela, "Probabilistic pretactical arrival and departure flight delay prediction with quantile regression," *Journal of Air Transportation*, vol. 32, no. 2, pp. 84–96, 2024.
- [26] T. Vandal, M. Livingston, C. Piho, and S. Zimmerman, "Prediction and uncertainty quantification of daily airport flight delays," in *Proceedings of The 4th International Conference on Predictive Applications and APIs*, ser. Proceedings of Machine Learning Research, vol. 82. PMLR, 2018, pp. 45–51. [Online]. Available: <https://proceedings.mlr.press/v82/vandal18a.html>
- [27] Y. Rodríguez and O. Díaz Olariaga, "Air traffic demand forecasting with a bayesian structural time series approach," *Periodica Polytechnica Transportation Engineering*, vol. 52, no. 1, pp. 75–85, 2024.
- [28] Y. Pang and Y. Liu, "Probabilistic aircraft trajectory prediction considering weather uncertainties using dropout as bayesian approximate variational inference," in *AIAA Scitech 2020 Forum*, 2020, p. 1413.
- [29] A. Lecchini Visintini, W. Glover, J. Lygeros, and J. Maciejowski, "Monte carlo optimization for conflict resolution in air traffic control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp. 470–482, 2006.
- [30] Y. Wang and Y. Zhang, "Prediction of runway configurations and airport acceptance rates for multi-airport system using gridded weather forecast," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103049, 2021.
- [31] A. R. Odoni, "The flow management problem in air traffic control," in *Flow control of congested networks*. Springer, 1987, pp. 269–288.
- [32] D. Bertsimas, G. Lulli, and A. Odoni, "The air traffic flow management problem: An integer optimization approach," in *International conference on integer programming and combinatorial optimization*. Springer, 2008, pp. 34–46.
- [33] D. Bertsimas and S. S. Patterson, "The traffic flow management rerouting problem in air traffic control: A dynamic network flow approach," *Transportation Science*, vol. 34, no. 3, pp. 239–255, 2000.

- [34] D. Sun and A. M. Bayen, “Multicommodity eulerian-lagrangian large-capacity cell transmission model for en route traffic,” *Journal of guidance, control, and dynamics*, vol. 31, no. 3, pp. 616–628, 2008.
- [35] M. O. Ball, R. Hoffman, A. R. Odoni, and R. Rifkin, “A stochastic integer program with dual network structure and its application to the ground-holding problem,” *Operations research*, vol. 51, no. 1, pp. 167–171, 2003.
- [36] J. Chen and D. Sun, “Stochastic ground-delay-program planning in a metroplex,” *Journal of Guidance, Control, and Dynamics*, vol. 41, no. 1, pp. 231–239, 2018.
- [37] C. Chin, M. Z. Li, K. Gopalakrishnan, and H. Balakrishnan, “Airport ground holding with hierarchical control objectives,” in *USA-Europe ATM R&D Seminar*, 2021.
- [38] C. N. Glover and M. O. Ball, “Stochastic optimization models for ground delay program planning with equity–efficiency tradeoffs,” *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 196–202, 2013.
- [39] A. Jacquillat, “Predictive and prescriptive analytics toward passenger-centric ground delay programs,” *Transportation Science*, vol. 56, no. 2, pp. 265–298, 2022.
- [40] Airports Council International (ACI) and International Air Transport Association (IATA) and Worldwide Airport Coordinators Group (WWACG), *Worldwide Airport Slot Guidelines (WASG), Edition 3, Effective 1 April 2024*, ACI / IATA / WWACG, Montreal / Geneva / Global, Sep. 2023, published jointly by ACI, IATA, and WWACG; available at <https://www.iata.org/en/programs/ops-infra/slots/slot-guidelines/> (accessed 09/15/2025).
- [41] P. Pellegrini, L. Castelli, and R. Pesenti, “Metaheuristic algorithms for the simultaneous slot allocation problem,” *IET Intelligent Transport Systems*, vol. 6, no. 4, pp. 453–462, 2012.
- [42] P. Pellegrini, T. Bolić, L. Castelli, and R. Pesenti, “Sosta: An effective model for the simultaneous optimisation of airport slot allocation,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 99, pp. 34–53, 2017.
- [43] M. O. Ball, A. S. Estes, M. Hansen, and Y. Liu, “Quantity-contingent auctions and allocation of airport slots,” *Transportation Science*, vol. 54, no. 4, pp. 858–881, 2020.
- [44] K. G. Zografos and Y. Jiang, “A bi-objective efficiency-fairness model for scheduling slots at congested airports,” *Transportation Research Part C: Emerging Technologies*, vol. 102, pp. 336–350, 2019.

- [45] N. A. Ribeiro, A. Jacquillat, and A. P. Antunes, “A large-scale neighborhood search approach to airport slot allocation,” *Transportation Science*, vol. 53, no. 6, pp. 1772–1797, 2019.
- [46] E. Delage and Y. Ye, “Distributionally robust optimization under moment uncertainty with application to data-driven problems,” *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.
- [47] P. M. Esfahani and D. Kuhn, “Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations,” *arXiv preprint arXiv:1505.05116*, 2015.
- [48] K. Kim, “Dual decomposition of two-stage distributionally robust mixed-integer programming under the wasserstein ambiguity set,” *Preprint manuscript*, 2020.
- [49] M. Cheramin, J. Cheng, R. Jiang, and K. Pan, “Computationally efficient approximations for distributionally robust optimization under moment and wasserstein ambiguity,” *INFORMS Journal on Computing*, vol. 34, no. 3, pp. 1768–1794, 2022.
- [50] R. Jiang, M. Ryu, and G. Xu, “Data-driven distributionally robust appointment scheduling over wasserstein balls,” *arXiv preprint arXiv:1907.03219*, 2019.
- [51] G. A. Hanasusanto and D. Kuhn, “Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls,” *Operations Research*, vol. 66, no. 3, pp. 849–869, 2018.
- [52] W. Wiesemann, D. Kuhn, and M. Sim, “Distributionally robust convex optimization,” *Operations research*, vol. 62, no. 6, pp. 1358–1376, 2014.
- [53] M. Ball, C. Barnhart, M. Dresner, M. Hansen, K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani, and B. Zou, “Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the united states,” Federal Aviation Administration, Washington, DC, Tech. Rep., 2010.
- [54] A. Mukherjee and M. Hansen, “Dynamic stochastic optimization model for air traffic flow management with en route and airport capacity constraints,” *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 61–73, 2012.
- [55] Y. Wang, M. Viscarra, L. Jin, and M. V. Bendarkar, “Airport capacity prediction with explicit consideration of terminal airspace constraints,” in *17th AIAA Aviation Technology, Integration, and Operations Conference*, 2017, p. 3428.

- [56] Federal Aviation Administration, “Flight schedule monitor user’s guide, version 13.0,” Technical Report, Feb 2016. [Online]. Available: https://tfmlearning.faa.gov/tfm-training/final_rel_13_tfms_fsm_users_guide_12686.pdf
- [57] —, “Airport arrival demand chart information briefing,” Technical Report, Jul 2017. [Online]. Available: https://tfmlearning.faa.gov/assets/media/tfm-training/TFMS_AADC_Information_Briefing_28Jul2017.pdf
- [58] B. Sridhar, K. S. Sheth, and S. Grabbe, “Modeling and optimization in traffic flow management,” *Proceedings of the IEEE*, vol. 96, no. 12, pp. 2060–2080, 2008.
- [59] Y. Liu, Y. Liu, M. Hansen, A. Pozdnukhov, and D. Zhang, “Using machine learning to analyze air traffic management actions: Ground delay program case study,” *Transportation Research Part E*, vol. 131, pp. 80–95, 2019.
- [60] E. P. Gilbo, “Airport capacity: representation, estimation, optimization,” *IEEE Transactions on Control Systems Technology*, vol. 5, no. 1, pp. 144–154, 1997.
- [61] I. Simaiakis, M. Sandberg, and H. Balakrishnan, “Balancing runway capacity and delay at airports through pushback rate control: A queuing network approach,” *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 378–395, 2014.
- [62] V. Ramanujam and H. Balakrishnan, “Estimation of arrival-departure capacity trade-offs in multi-airport systems,” in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC)*. IEEE, 2009, pp. 2534–2540.
- [63] R. de Neufville and A. R. Odoni, *Airport Systems: Planning, Design, and Management*, 2nd ed. New York: McGraw-Hill Education, 2013.
- [64] S. Allan, J. Beesley, J. Evans, and S. Gaddy, “Analysis of delay causality at newark international airport,” in *4th USA/Europe Air Traffic Management R&D Seminar*, 2001, pp. 1–11.
- [65] Z. Renhe, Q. Li, and R. Zhang, “Meteorological conditions for the persistent severe fog and haze event over eastern china in january 2013,” *SCES*, vol. 57, pp. 26–35, 2014.
- [66] H. Wu and M. Z. Li, “Distributionally robust airport ground holding problem under wasserstein ambiguity sets,” *arXiv preprint arXiv:2306.09836*, 2023.
- [67] Airports Council International – North America , “Aircraft Operating and Delay Cost per Enplanement,” 2014, accessed: Feb 2024.

- [68] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.