



Research article

Data-driven optimization for airline service quality improvement under delay and cancellation risk

Jinho Cha¹ and Eun-Young Kwon^{2,*}

¹ Department of Computer Science, Gwinnett Technical College, Lawrenceville, GA, USA

² AI Industry Research Center, Kwangwoon University, Seoul, Korea

* **Correspondence:** Email: eykwon@kma.ac.kr, eykwonclass0524@gmail.com.

Abstract: Airline disruption prediction has operational value only when it improves resource-allocation decisions made before realized service outcomes are known. We develop an auditable forecast-to-optimization framework that separates service-quality measurement, ex-ante forecasting, constrained allocation, and ex-post evaluation. The service-harm construction uses a mutually exclusive decomposition so that cancellations are not counted both in the non-on-time component and again through the cancellation penalty. Bureau of Transportation Statistics flight-level records and U.S. Department of Transportation published measures are treated as linked reporting layers of the same operating system and are used for reconstruction-consistency diagnostics, not independent-source validation. On a locked 2025 carrier-month panel, the leakage-controlled forecast-based predict-then-optimize (Forecast-PTO) policy retains 97.95%–99.60% of hindsight-oracle value across budget fractions of 5%–30%. The main decision conclusion remains stable under heterogeneous relative costs and capacities, alternative response-curvature assumptions, and a finer carrier-airport-month analysis. An uncertainty-aware robust extension shows that protection against forecast disagreement can carry an ex-post decision cost rather than automatically improving realized value. The framework therefore distinguishes measurement consistency, forecast value, optimization value, and causal intervention effects not identified by observational service data.

Keywords: airline service quality; flight delays; flight cancellations; resource allocation; predictive-to-prescriptive analytics

Mathematics Subject Classification: 90B06, 90B15, 90C15, 90C90, 68T05

1. Introduction

Airline disruption is not merely a prediction problem. A delay or cancellation forecast becomes operationally valuable only when it changes a decision: which operating units should receive priority,

where limited service-improvement capacity should be directed, and how scarce planning resources should be distributed before disruption costs are fully realized. This distinction matters because airline operations are tightly coupled across aircraft, crews, airports, routes, and passengers; localized failures can propagate through the network and create service losses far beyond the original event [1–5]. The practical challenge is therefore not simply to predict where service degradation will occur, but to convert imperfect predictive information into defensible resource-allocation decisions under limited budgets and operational uncertainty. In this sense, the value of a predictive model is ultimately a decision-value question.

Yet this decision layer remains comparatively underdeveloped in airline analytics. Prediction studies primarily optimize statistical accuracy, whereas recovery studies typically optimize aircraft, crew, gate, or passenger schedules after irregular operations have already materialized [6–10]. Recent uncertainty-aware transportation models further demonstrate that predictive and operational uncertainty can be propagated into downstream decisions [11]. What remains less developed is the intermediate planning problem: How should information available *before* period t be converted into constrained service-quality allocations, and how should the induced decisions be evaluated without using realized period- t outcomes in the decision rule itself? This is not a minor empirical detail. If contemporaneous outcomes enter both the allocation score and the evaluation benchmark, near-oracle performance can be produced partly by construction. Likewise, strongly correlated operational and service-quality measures can appear incrementally informative even when they are different reporting layers of the same underlying system. These issues affect not only airline analytics but the credibility of predictive-to-prescriptive evaluation more generally. We address this gap by developing a leakage-controlled, uncertainty-aware forecast-to-optimization architecture in which measurement, forecasting, optimization, and hindsight evaluation have distinct and auditable roles.

We operationalize this architecture using two linked U.S. airline reporting layers. Bureau of Transportation Statistics (BTS) flight-level records provide detailed operational information on delays and cancellations, whereas Department of Transportation (DOT) Air Travel Consumer Reports provide official carrier-month service-quality outcomes. Because both sources describe the same underlying operating system, their association is treated as cross-layer measurement consistency rather than independent-source validation. We therefore use rate normalization, within-group ranking, operating-volume controls, and fixed-effect residualization before using the resulting service-harm measures in downstream decisions. A one-month-ahead forecasting layer then predicts carrier-month exposure-weighted service harm using only historically available information, and the resulting forecast is passed to a budget-constrained allocation model before realized month- t DOT harm is observed. This design makes it possible to measure the downstream decision cost of forecast imperfection rather than the mechanical similarity between a same-period score and its own evaluation target.

The empirical evidence shows that this distinction is consequential. On a locked 2025 panel of 120 carrier-month observations, the selected Gradient Boosting model provides strong point and within-month ranking accuracy. More importantly, when its forecast is used in the allocation model without access to realized current-month DOT harm, Forecast-PTO retains 97.95%–99.60% of hindsight-oracle value across budget fractions from 5% to 30%. The result is not confined to the simplest homogeneous allocation setting: The central decision conclusion remains stable under heterogeneous relative costs and capacities, explicit forecast uncertainty, alternative diminishing-return curvature, and a finer carrier-airport-month decision resolution. These tests therefore examine not whether one forecasting algorithm

dominates every alternative, but whether the forecast-to-decision principle survives realistic sources of implementation and model uncertainty.

The optimization layer is deliberately interpretable. Service-harm measures enter as priority or harm-at-risk weights rather than estimated causal treatment effects. The linear formulation yields a transparent cost-adjusted priority rule, the diminishing-return formulation produces a water-filling allocation structure, and the robust formulation propagates forecast disagreement through an exact max–min counterpart. This separation is important because observational service data can identify where harm is concentrated without identifying how much a specific intervention would causally reduce that harm. The framework therefore preserves prescriptive structure while avoiding causal-effect claims that the available public data cannot support.

The study makes four contributions with implications beyond the specific airline application. First, it reframes airline disruption analytics from a prediction problem into a decision-value problem: The relevant question is not only whether disruption harm can be forecast accurately, but whether forecasts improve constrained resource-allocation decisions. Second, it establishes an information-valid evaluation design for predictive-to-prescriptive analytics by separating ex-ante decision inputs from ex-post evaluation outcomes, thereby preventing same-period information leakage from being mistaken for prescriptive performance. Third, it integrates measurement, forecasting, and optimization within a single auditable architecture while explicitly distinguishing measurement consistency, forecast value, optimization value, and causal intervention effectiveness. Fourth, it shows how forecast uncertainty, implementation heterogeneity, and decision resolution can be examined within the same allocation architecture, providing a reproducible template for data-driven service systems in which predictive information must ultimately support scarce-resource decisions.

The interpretation remains deliberately bounded. Hindsight Oracle-DOT is used only as a normalization benchmark, contemporaneous DOT–BTS analyses are retrospective diagnostics, and the available public data do not identify the causal effectiveness of specific service interventions. Relative intervention costs and capacities are pre-period-anchored scenario quantities rather than observed accounting costs. Accordingly, the study does not claim to estimate realized treatment effects. Its contribution is to establish when and how predictive service information can be translated into credible resource-allocation decisions under explicit information and uncertainty constraints.

The remainder of the paper is organized as follows. Section 2 reviews flight-delay prediction, disruption management, service-quality measurement, predictive-to-prescriptive analytics, and uncertainty-aware optimization. Section 3 develops the allocation model, Section 4 presents its structural properties, and Section 5 describes the data and empirical design. Section 6 reports the computational evidence, Section 7 discusses managerial implications and limitations, and Section 8 concludes.

2. Literature review

In this section, we review four streams of literature that are closely related to the proposed framework: flight delay and cancellation prediction, delay propagation and airline disruption management, airline service-quality measurement, and predictive-to-prescriptive analytics. These streams have developed in parallel, but they remain only partially integrated. Prediction studies estimate the likelihood or magnitude of operational disruptions. Disruption-management studies formulate optimization models for recovering airline resources after irregular operations. Service-quality studies and official consumer reports

document the passenger-facing consequences of delays and cancellations. Predictive-to-prescriptive analytics provides the methodological bridge from predictive information to downstream optimization.

Our study is positioned at the intersection of these streams. We do not treat delay prediction as the final objective. Instead, historical and forecasted operational information supports ex-ante allocation, while official carrier-month service-quality outcomes provide the corresponding measurement and ex-post evaluation layer. The main distinction is therefore prescriptive: We ask how limited improvement resources should be allocated when disruption risk must be translated into a feasible decision before realized current-period service outcomes are known.

2.1. *Flight delay and cancellation prediction*

Flight delay prediction has become a central topic in intelligent aviation systems because delays affect airline operations, airport capacity, passenger satisfaction, and downstream schedule reliability. Several review studies summarize flight-delay prediction research and emphasize the growing role of data science, operational data integration, and machine learning in aviation delay analysis [12, 13]. Early studies relied on statistical and econometric models to identify operational, temporal, weather-related, and airport-level determinants of delay. Classical machine-learning approaches, including support vector machines and related classifiers, have also been used to model nonlinear delay patterns from operational and airport-level predictors [14]. More recent studies use increasingly rich data and learning architectures to capture nonlinear temporal, spatial, and operational patterns.

Hierarchical learning and recurrent neural-network models have been used to predict departure-delay occurrence and duration [15–17]. Graph-based spatial-temporal models and federated learning approaches further emphasize the network structure and distributed nature of delay prediction [18, 19]. Explainable multi-source learning frameworks show that integrating operational, airport, geographic, and weather information can improve interpretability as well as predictive performance [20]. These studies show that flight-level, airport-level, and network-level information can improve predictive performance relative to simple historical averages.

Recent work also emphasizes that delay prediction is not merely a statistical exercise. Delay predictive analytics for airport capacity management reviews simulation, queueing, and data-driven delay prediction methods and highlights the operational value of local delay forecasts for capacity-management decisions [6]. Similarly, flight arrival-time prediction has been embedded in gate-assignment optimization, where predicted arrival times are used to reduce potential gate conflicts under arrival uncertainty [7]. Machine learning-enhanced landing scheduling further demonstrates that probabilistic predictions can be incorporated into scheduling constraints under uncertainty [8]. These studies indicate a shift from standalone prediction toward operationally actionable prediction.

Despite this progress, much of the flight-delay prediction literature still evaluates models primarily through statistical performance measures such as accuracy, area under the curve (AUC), mean absolute error, or classification error. These measures are useful for assessing predictive quality, but they do not by themselves determine how limited improvement resources should be allocated across carrier-months. A flight-level model may identify high-risk flights, but the managerial question is different: Which carrier-month observations should receive scarce improvement resources? We address this gap by aggregating flight-level risk estimates into carrier-month measures and linking them to an explicit resource allocation model.

2.2. *Delay propagation and airline disruption management*

A second stream studies delay propagation and airline disruption management. Flight delays are rarely isolated events. Because aircraft, crews, gates, and passengers are connected through rotations, schedules, and network resources, a disruption in one part of the system can propagate to later flights and operating periods. Recent reviews of flight delay propagation emphasize data structure, network dynamics, propagation mechanisms, and opportunities for combining predictive and optimization methods [5]. Airline disruption-management reviews similarly categorize recovery models for aircraft, crews, passengers, and integrated resources, showing that disruption recovery remains a core operations research problem in aviation [21, 22].

Classical airline recovery models formulate aircraft, crew, and passenger recovery as large-scale network or integer optimization problems [9, 10, 23]. Recent optimization studies have extended this tradition by incorporating richer uncertainty, predictive information, and resource interactions. Robust aircraft maintenance routing with heterogeneous maintenance tasks shows that maintenance planning and routing decisions can materially affect delay risk and operational reliability [24]. In airline routing, predictive-prescriptive analytics has also been used to forecast primary delays and optimize day-ahead aircraft routing decisions, showing that operational predictions can improve downstream airline planning [25]. Machine learning-assisted aircraft and passenger recovery models use learning to reduce solution spaces and accelerate large-scale recovery optimization [26]. Related work on aircraft recovery and crew recovery combines supervised learning with mixed-integer optimization to obtain high-quality recovery decisions within short operational time limits [27, 28]. These studies provide important methodological precedents for combining learning and optimization in airline operations.

However, disruption-management papers usually focus on day-of-operations recovery after disruptions occur. Their primary objective is to repair disrupted schedules, reroute aircraft, recover crews, reassign passengers, or reduce operational recovery costs. Our setting is different. We focus on service-quality improvement at the carrier-month level. Rather than recovering a disrupted schedule in real time, we ask how predicted delay and cancellation risk, combined with official service-quality outcomes, can guide improvement-resource allocation across carrier-months. This distinction connects disruption analytics with service-quality planning and budget-constrained resource allocation.

2.3. *Airline service quality and official consumer-performance measures*

Airline service quality is strongly shaped by reliability outcomes such as on-time performance and cancellations. Delays and cancellations impose direct inconvenience on passengers and may affect missed connections, schedule uncertainty, waiting time, compensation costs, complaint behavior, and customer satisfaction. Empirical research on airport service quality shows that flight delays can significantly affect passengers' global satisfaction and that internally attributable delays may be especially salient in perceived service quality [29]. These findings support the use of delay and cancellation measures as service-harm proxies in service-quality analysis.

Official consumer-performance data provide an additional basis for evaluating airline service quality. The U.S. Department of Transportation's Air Travel Consumer Report is designed to provide consumers with information on airline service quality and includes indicators related to flight delays, cancellations, mishandled baggage, wheelchairs and scooters, oversales, consumer complaints, and related service outcomes [30]. In this study, we process DOT Air Travel Consumer Reports into a structured carrier-month

panel containing on-time arrival rates, cancellation rates, scheduled operations, canceled operations, and derived harm indices. This official panel provides an observed service-quality benchmark against which flight-level risk estimates can be interpreted.

The service-quality literature and official reporting systems motivate our outcome measures, but they do not solve the allocation problem. Ranking airlines by on-time performance or cancellation rate is informative, but ranking alone does not determine where improvement resources should be deployed. We therefore convert official performance measures into decision-relevant harm indices and embed those indices in a budget-constrained optimization model.

2.4. Predictive-to-prescriptive analytics and resource allocation

The predictive-to-prescriptive analytics literature provides the methodological foundation for converting risk estimates into operational decisions. Data-driven prescriptive analytics studies how historical observations can be used to choose decisions directly or to estimate quantities that feed downstream optimization models [31]. The smart predict-then-optimize framework further argues that prediction models should be evaluated not only by statistical error, but also by the quality of the decisions they induce in downstream optimization problems [32]. Related decision-focused learning approaches integrate prediction and optimization more directly by training models against task-based or downstream decision objectives [33]. Subsequent work develops risk guarantees and generalization bounds for end-to-end prediction-and-optimization processes [34,35]. These ideas are directly relevant to airline service operations because delay and cancellation predictions are valuable only when they improve downstream decisions.

Our study follows this decision-oriented perspective but does not train the predictive and optimization layers end-to-end. Instead, we use a deliberately modular forecast-then-optimize architecture. An upstream forecasting model produces a carrier-month service-harm score, and the downstream optimization model converts that score into a budget-feasible allocation. Forecast quality is therefore evaluated in two complementary ways: conventional point and rank diagnostics and, more importantly for the prescriptive task, the realized value of the allocation that the forecast induces.

Related transportation optimization research also shows how multiple sources of uncertainty can be propagated explicitly into downstream decisions rather than handled only through parameter sensitivity. For example, Tan et al. [11] study uncertainty-aware management of hydrogen-powered vehicles and renewable-energy units in coupled microgrid-transportation networks. Motivated by this distinction, the revised framework supplements deterministic Forecast-PTO with an explicit robust allocation model. Forecast uncertainty is represented by an ex-ante model-disagreement set constructed from seven implementable forecasting rules, and the resulting max-min problem is solved before realized current-month DOT harm is observed. The uncertainty radius is not interpreted as a calibrated probabilistic prediction interval; it is an ambiguity measure that permits a direct test of how explicit uncertainty protection changes the downstream allocation. Stochastic and distributionally robust formulations that jointly represent forecast uncertainty and intervention-response uncertainty remain extensions beyond the present study.

This distinction is central to the empirical design. The DOT-BTS reconstruction and contemporaneous allocation experiments are used primarily as measurement-consistency and structural diagnostics; they establish whether alternative operational representations preserve allocation-relevant ordering. A separate forecast-to-prescription experiment then removes the realized current-month DOT target

from the allocation score and evaluates the resulting decision against realized current-month harm. Consequently, near-oracle performance in the forecast experiment cannot arise mechanically from directly using the evaluation target in the allocation score. This design provides a more stringent test of the prediction-to-decision link while retaining a transparent separation between statistical forecasting and mathematical optimization.

2.5. Research gap and positioning

Table 1. Positioning of the proposed harm-aware allocation framework relative to adjacent research streams.

Research stream	Gap relative to harm-aware service-quality allocation
Flight delay and cancellation prediction [15–17]; [12, 19, 20]; [13, 14]	Predicts disruption occurrence, delay duration, or explanatory drivers, but does not determine how limited improvement resources should be allocated across carrier-month observations.
Airport capacity and operational prediction [6–8]; [18]	Uses predictive information to support capacity management, gate assignment, landing scheduling, or network-delay forecasting, but does not link official carrier-month service-quality outcomes to an explicit improvement-resource allocation model.
Delay propagation and disruption management [5, 21, 22]; [26–28]; [25]	Models delay propagation or repairs disrupted aircraft, crew, and passenger schedules after irregular operations, but does not address prospective service-quality improvement at the carrier-month level.
Robust airline planning and maintenance routing [24, 36]	Allocates buffers or constructs robust routes to reduce disruption exposure, but does not combine DOT official service-harm outcomes with reconstructed BTS operational risk for service-quality resource allocation.
Airline service-quality and consumer-performance studies [2, 29, 37]; [30]	Documents the passenger-facing consequences of delays, cancellations, and official service-performance outcomes, but does not formulate a downstream optimization model for allocating improvement resources.
Predictive-to-prescriptive analytics [31–33]; [34, 35]	Provides decision-aware learning and optimization principles, but is not specific to airline service-quality measurement, DOT-BTS data integration, or carrier-month service-harm prioritization.
Uncertainty-aware transportation optimization [11]	Propagates operational uncertainty into downstream optimization, but does not study official airline service-quality measurement, leakage-controlled carrier-month forecasting, or forecast-induced service-harm allocation.
Proposed framework	Combines BTS flight-level operational information, DOT-published carrier-month service-quality measures, exposure-weighted service-harm construction, forecast-based prioritization, heterogeneous feasible-set diagnostics, and budget-constrained optimization; downstream decision quality is evaluated under a leakage-controlled forecast-to-prescription design and an explicit max–min robust extension based on ex-ante forecast disagreement.

The reviewed literature reveals a clear gap. Flight-delay prediction studies have become increasingly sophisticated, but many remain prediction-centric. Disruption-management models provide strong optimization structures, but they often focus on recovery after disruptions rather than service-quality improvement planning. Service-quality research and DOT reporting systems document the consumer relevance of delays and cancellations, but they do not provide an explicit resource allocation model. Predictive-to-prescriptive analytics offers a general methodological bridge, but airline service-quality applications remain underdeveloped.

We address this gap through an integrated but auditable data-driven optimization framework. Flight-level operational records provide a micro-level basis for reconstructing and forecasting disruption-related service harm, while DOT reports provide the corresponding official carrier-month measurement layer. These quantities enter a budget-constrained optimization model whose decisions are evaluated by realized service-harm value rather than by prediction accuracy alone. The empirical design further separates retrospective measurement diagnostics from a leakage-controlled forecast-to-prescription evaluation, allowing us to determine whether forecast-induced priority rankings remain useful when converted into actual allocation decisions. A final robust layer propagates ex-ante forecast-model

disagreement through the allocation problem, so uncertainty is evaluated not only by sensitivity of parameters but also by its explicit effect on the optimization decision.

2.6. Summary comparison with prior research

Table 1 summarizes the positioning of the proposed framework relative to adjacent research streams. The key distinction is that our framework jointly incorporates flight-level risk estimation, official carrier-month service-quality measurement, exposure-weighted service-harm construction, and budget-constrained allocation optimization.

3. Model

In this section, we develop a harm-aware predictive-to-prescriptive framework for airline service-quality improvement. The framework links official carrier-month service-quality outcomes from DOT Air Travel Consumer Reports with flight-level operational risk measures reconstructed from BTS on-time performance records. The resulting service-harm measures are used as inputs to a budget-constrained resource allocation problem. The structural properties of the allocation model are analyzed in Section 4.

Let $\mathcal{A}$ denote the set of marketing carriers and let $\mathcal{T}$ denote the set of calendar months. The baseline observational unit is a carrier-month pair $(a, t) \in \mathcal{A} \times \mathcal{T}$. In the empirical implementation, the main sample contains 10 marketing carriers observed from January 2024 to December 2025. Observations from early 2026 are reserved for robustness analysis. We use two layers of measurement. The first layer is an official service-quality layer based on DOT reports. The second layer is a reconstructed operational-risk layer based on flight-level BTS records. The purpose of the model is to combine these two layers into a decision-relevant harm score and then allocate limited service-improvement resources accordingly.

3.1. Official service-harm measurement

We first define an official service-harm measure using DOT carrier-month service-quality reports. Let $OT_{a,t} \in [0, 100]$ denote the DOT-reported arrival on-time rate and let $CR_{a,t} \in [0, 100]$ denote the DOT-reported cancellation rate, both in percentage points. Because the complement $100 - OT_{a,t}$ contains canceled operations, adding a separate cancellation penalty to that complement would count cancellations twice. We therefore use a mutually exclusive decomposition. Define the non-cancellation, non-on-time service-loss component as

$$L_{a,t}^{nc} = 100 - OT_{a,t} - CR_{a,t}, \quad (1)$$

and define cancellation loss as

$$L_{a,t}^{can} = CR_{a,t}. \quad (2)$$

The first component contains non-canceled service failures represented by the DOT on-time table, including late and diverted operations to the extent that they enter the published non-on-time share; cancellations are represented only by $L_{a,t}^{can}$.

For cancellation-severity parameter $\lambda > 0$, the corrected DOT service-harm index is

$$H_{a,t}^{DOT}(\lambda) = L_{a,t}^{nc} + \lambda L_{a,t}^{can}. \quad (3)$$

Equivalently,

$$H_{a,t}^{DOT}(\lambda) = (100 - OT_{a,t}) + (\lambda - 1)CR_{a,t}. \quad (4)$$

At the baseline $\lambda = 2$, each cancellation receives total weight two, not three as would occur if $CR_{a,t}$ were added to an already cancellation-inclusive complement. The parameter λ is a transparent scenario weight rather than a monetary valuation or a causal estimate.

Let $S_{a,t}$ denote scheduled operations and $C_{a,t}$ canceled operations. The mutually exclusive exposure components are

$$E_{a,t}^{nc,DOT} = \frac{L_{a,t}^{nc}}{100} S_{a,t}, \quad (5)$$

and

$$E_{a,t}^{can,DOT}(\lambda) = \lambda C_{a,t}. \quad (6)$$

Thus,

$$E_{a,t}^{DOT}(\lambda) = E_{a,t}^{nc,DOT} + E_{a,t}^{can,DOT}(\lambda). \quad (7)$$

The rate-based measure is used for cross-layer consistency diagnostics, while the exposure-weighted measure is used when the decision objective should reflect the number of operations exposed to service loss.

3.2. Flight-level operational risk reconstruction

The BTS records describe the same underlying airline operations at flight level. We therefore use them as a linked reconstruction layer rather than as an independent external validation source. Let $\mathcal{I}_{a,t}$ denote the set of domestic marketed flights for carrier a in month t . Define

$$Y_i^{arr} = \mathbf{1}\{\text{flight } i \text{ arrives at least 15 minutes late}\}, \quad (8)$$

$$Y_i^{can} = \mathbf{1}\{\text{flight } i \text{ is canceled}\}, \quad Y_i^{div} = \mathbf{1}\{\text{flight } i \text{ is diverted}\}. \quad (9)$$

The arrival-delay indicator is evaluated only for non-canceled, non-diverted flights with valid arrival outcomes. Let $N_{a,t} = |\mathcal{I}_{a,t}|$, and define the event counts

$$D_{a,t} = \sum_{i \in \mathcal{I}_{a,t}^{arr}} Y_i^{arr}, \quad V_{a,t} = \sum_{i \in \mathcal{I}_{a,t}} Y_i^{div}, \quad C_{a,t} = \sum_{i \in \mathcal{I}_{a,t}} Y_i^{can}. \quad (10)$$

To match the corrected DOT decomposition on the scheduled-flight denominator, the BTS non-cancellation service-loss rate is

$$L_{a,t}^{nc,BTS} = 100 \frac{D_{a,t} + V_{a,t}}{N_{a,t}}, \quad (11)$$

and the BTS cancellation rate is

$$R_{a,t}^{can,BTS} = \frac{C_{a,t}}{N_{a,t}}. \quad (12)$$

The corrected BTS risk index is

$$R_{a,t}^{BTS}(\lambda) = L_{a,t}^{nc,BTS} + \lambda 100 R_{a,t}^{can,BTS}, \quad (13)$$

with baseline

$$R_{a,t}^{BTS}(2) = 100 \frac{D_{a,t} + V_{a,t} + 2C_{a,t}}{N_{a,t}}. \quad (14)$$

The corresponding exposure-weighted measure is

$$E_{a,t}^{BTS}(\lambda) = D_{a,t} + V_{a,t} + \lambda C_{a,t}. \quad (15)$$

Hence delayed operated flights, diversions, and cancellations enter as mutually exclusive event categories, and cancellations receive exactly the intended weight λ .

3.3. DOT–BTS alignment

The DOT and BTS measures serve different roles. The DOT measure $H_{a,t}^{DOT}(\lambda)$ is an official service-quality outcome, whereas $R_{a,t}^{BTS}(\lambda)$ is reconstructed from flight-level operational records. Before using reconstructed risk measures in a prescriptive allocation model, we examine whether the two measures are aligned at the carrier-month level.

We write the alignment relation as

$$H_{a,t}^{DOT}(\lambda) = R_{a,t}^{BTS}(\lambda) + \varepsilon_{a,t}, \quad (16)$$

where $\varepsilon_{a,t}$ captures differences due to report rounding, aggregation conventions, release timing, revisions, and scope differences between the two sources. We do not require exact equality. Instead, we evaluate alignment using absolute differences, correlations, and rank consistency.

For each carrier-month, define the alignment error

$$\Delta_{a,t}(\lambda) = H_{a,t}^{DOT}(\lambda) - R_{a,t}^{BTS}(\lambda), \quad (17)$$

and the absolute alignment error

$$A_{a,t}(\lambda) = |\Delta_{a,t}(\lambda)|. \quad (18)$$

A small value of $A_{a,t}(\lambda)$ indicates that the official DOT service-harm measure and the BTS reconstructed risk measure are internally consistent at the carrier-month level.

For the full sample, we summarize alignment using the mean absolute alignment error

$$\bar{A}(\lambda) = \frac{1}{|\mathcal{A}||\mathcal{T}|} \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} A_{a,t}(\lambda), \quad (19)$$

and the rank-consistency statistic

$$\rho^{rank}(\lambda) = \text{corr}_S \left(H_{a,t}^{DOT}(\lambda), R_{a,t}^{BTS}(\lambda) \right), \quad (20)$$

where corr_S denotes Spearman rank correlation across carrier-month observations. The rank statistic is particularly relevant for allocation because budget-constrained policies often depend more on priority ordering than on exact equality of harm scores.

A high exposure-weighted correlation is not, by itself, sufficient evidence of cross-layer informational complementarity because both exposure measures scale with the number of scheduled operations. Accordingly, the empirical analysis separates three questions: whether DOT and BTS agree in rate space, whether they remain aligned after removing common carrier-month operating scale, and whether they preserve within-month priority ordering. This distinction prevents a mechanical common-volume component from being interpreted as incremental decision information.

3.4. Harm-aware resource allocation

We now convert service-harm measures into resource allocation decisions. Let $x_{a,t} \geq 0$ denote a normalized service-improvement resource assigned to carrier-month (a, t) . The variable can represent generic improvement effort, recovery support, rebooking capacity, station-level support, schedule-reliability effort, or another service-quality intervention once a concrete implementation is specified. The present DOT–BTS panel, however, does not identify a unique intervention technology or a causal unit effect. Accordingly, $x_{a,t}$ is used as a planning variable, $c_{a,t} > 0$ as a unit-cost or relative coordination-cost coefficient, and $\bar{x}_{a,t}$ as the maximum feasible normalized intervention capacity. Let $B > 0$ denote the available budget.

The feasible allocation set is

$$\mathcal{X}(B) = \left\{ x \in \mathbb{R}_+^{|\mathcal{A}| \times |\mathcal{T}|} : \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} c_{a,t} x_{a,t} \leq B, \quad 0 \leq x_{a,t} \leq \bar{x}_{a,t} \right\}, \quad (21)$$

where the cost and capacity terms may be homogeneous for a benchmark experiment or heterogeneous when operationally motivated relative scenarios are imposed.

The linear harm-aware allocation model is

$$\max_{x \in \mathcal{X}(B)} \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} s_{a,t} x_{a,t}. \quad (22)$$

The quantity $s_{a,t}$ is a priority score, not an estimated treatment effect. It may be defined using official DOT harm, reconstructed BTS risk, or a blended measure. In the baseline retrospective implementation, we use exposure-weighted scores:

$$s_{a,t}^{DOT} = E_{a,t}^{DOT}(2), \quad s_{a,t}^{BTS} = E_{a,t}^{BTS}(2). \quad (23)$$

The DOT-based score emphasizes official reported service harm, while the BTS-based score emphasizes the flight-level operational reconstruction.

The retrospective predictive-to-prescriptive (PTO) diagnostic score is

$$s_{a,t}^{PTO} = \omega E_{a,t}^{DOT}(2) + (1 - \omega) E_{a,t}^{BTS}(2), \quad 0 \leq \omega \leq 1. \quad (24)$$

When $\omega = 1$, the policy relies only on DOT official service-harm measures. When $\omega = 0$, it relies only on BTS reconstructed operational risk. Intermediate values combine the two same-period measurement layers. The separate prospective Forecast-PTO policy in Section 3.7 replaces this contemporaneous score with an ex-ante forecast that excludes realized current-month DOT harm.

The linear model has a priority-index interpretation: When marginal response is constant, resources are assigned in nonincreasing order of the cost-adjusted score $s_{a,t}/c_{a,t}$, subject to the budget and upper-bound constraints. This structure provides an interpretable audit trail for the allocation decision. It ranks carrier-month opportunities by priority value per unit cost; it does not assert a causal mapping from one unit of planning resource to a measured reduction in realized service harm. Section 4 provides the formal priority result.

3.5. Priority-weighted intervention response under diminishing returns

The linear model assumes constant marginal response to the normalized planning resource. To represent the common operational idea that additional effort has declining marginal value, we separate the observed priority weight from an assumed intervention-response scenario. Specifically, let

$$G_{a,t}(x_{a,t}) = H_{a,t} r(x_{a,t}; \beta_{a,t}),$$

where $H_{a,t} \geq 0$ is the service-harm priority or harm-at-risk weight and $r(x; \beta)$ is a normalized response scenario satisfying $r(0; \beta) = 0$, $r_x(x; \beta) > 0$, and $r_{xx}(x; \beta) < 0$ over the relevant resource range. The baseline family is

$$r^{\text{exp}}(x; \beta) = 1 - \exp(-\beta x), \quad \beta > 0, \quad (25)$$

so that $G_{a,t}(x_{a,t}) = H_{a,t}(1 - \exp(-\beta_{a,t}x_{a,t}))$. The parameter $\beta_{a,t}$ is therefore a *scenario-response parameter*: it governs curvature and marginal-response decay but is not an empirically estimated causal-effect coefficient in the present data.

The diminishing-return allocation problem is

$$\max_{x \in \mathcal{X}(B)} \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} H_{a,t} (1 - \exp(-\beta_{a,t}x_{a,t})). \quad (26)$$

Its marginal scenario value is

$$\frac{\partial G_{a,t}(x_{a,t})}{\partial x_{a,t}} = H_{a,t} \beta_{a,t} \exp(-\beta_{a,t}x_{a,t}). \quad (27)$$

Thus, the model assigns high initial priority to carrier-months with large $H_{a,t}\beta_{a,t}$, while repeated allocation reduces marginal scenario value. This prevents complete concentration on one unit unless its cost-adjusted marginal priority remains dominant.

For interior allocations, the diminishing-return model equalizes cost-adjusted marginal scenario values:

$$\frac{H_{a,t} \beta_{a,t} \exp(-\beta_{a,t}x_{a,t})}{c_{a,t}} = \eta, \quad (28)$$

where $\eta \geq 0$ is the budget multiplier. This yields an interpretable water-filling rule over priority, scenario responsiveness, cost, and capacity. The condition is structural and should not be interpreted as an estimated causal marginal benefit. Section 4 gives the formal concavity and marginal-equalization result.

3.6. Benchmark allocation policies

To evaluate the value of the proposed harm-aware predictive-to-prescriptive policy, we compare it with several benchmark policies. All policies are evaluated using the same downstream decision-value metric.

The Oracle-DOT policy uses the evaluation target itself as the allocation score:

$$s_{a,t}^{\text{Oracle}} = E_{a,t}^{\text{DOT}}(2). \quad (29)$$

This policy provides an upper benchmark for DOT-evaluated allocation value.

The DOT-only policy uses official DOT exposure-weighted harm:

$$s_{a,t}^{\text{DOT-only}} = E_{a,t}^{\text{DOT}}(2). \quad (30)$$

In the deterministic DOT-evaluated setting, DOT-only and Oracle-DOT coincide. We retain both labels for interpretive clarity. Oracle-DOT denotes the evaluation upper benchmark used to normalize policy value, whereas DOT-only denotes the policy class that allocates resources using only official DOT service-harm measures. This distinction is useful when comparing the blended PTO rule with single-source allocation rules.

The BTS-only policy uses reconstructed BTS exposure-weighted operational risk:

$$s_{a,t}^{BTS-only} = E_{a,t}^{BTS}(2). \quad (31)$$

The proposed PTO policy uses the blended score in Equation (24). The baseline implementation uses $\omega = 0.5$, and the sensitivity analysis varies ω over the blending grid.

The equal-budget-share policy assigns an equal share of the available budget to each eligible carrier-month, with the resulting resource quantity adjusted by the corresponding unit-cost coefficient:

$$x_{a,t}^{EQ} = \frac{B}{|\mathcal{A}||\mathcal{T}|c_{a,t}}, \quad (32)$$

subject to feasibility and upper-bound constraints.

The delay-only policy ranks carrier-months using only the corrected non-cancellation exposure component:

$$s_{a,t}^{Delay} = E_{a,t}^{nc,DOT}. \quad (33)$$

The cancellation-only policy ranks carrier-months using only the corrected cancellation exposure component:

$$s_{a,t}^{Cancel} = E_{a,t}^{can,DOT}(2). \quad (34)$$

The volume-only policy allocates resources based on scheduled operations or flight volume:

$$s_{a,t}^{Volume} = N_{a,t}. \quad (35)$$

The historical-average policy uses carrier-level historical average harm:

$$s_{a,t}^{Hist} = \frac{1}{|\mathcal{T}_{train}(t)|} \sum_{\tau \in \mathcal{T}_{train}(t)} H_{a,\tau}^{DOT}(2), \quad \mathcal{T}_{train}(t) = \{\tau : \tau < t\}. \quad (36)$$

Thus, in prospective evaluation, the historical benchmark uses only periods strictly preceding the evaluation month.

The proposed policy solves either the linear allocation problem in Equation (22) or the diminishing-return problem in Equation (26). In the computational experiments, all benchmark policies are evaluated under the same realized-harm metric and feasible allocation set. In the contemporaneous experiments this comparison is retrospective; the separate forecast-to-prescription experiment below imposes strict information separation between the score used to allocate resources and the realized target used for evaluation.

3.7. Forecast-driven allocation and realized decision value

For prospective decision evaluation, the allocation must be determined without using the realized service-harm target that will subsequently be used to assess the decision. Let $\widehat{E}_{a,t}$ denote a forecast of exposure-weighted service harm for carrier a in evaluation month t . The corresponding nonnegative Forecast-PTO score is

$$s_{a,t}^F = \max\{0, \widehat{E}_{a,t}\}. \quad (3.1)$$

The realized current-month DOT exposure-weighted harm, $E_{a,t}^{DOT}(2)$, is excluded from $s_{a,t}^F$ and retained only for ex-post decision evaluation and construction of the hindsight oracle.

Under the diminishing-return formulation, the forecast-driven allocation for a budget B is

$$x^F(B) \in \arg \max_{x \in \mathcal{X}(B)} \sum_{a,t} s_{a,t}^F (1 - \exp(-\beta x_{a,t})), \quad (3.2)$$

where $\mathcal{X}(B)$ denotes the common budget and upper-bound feasible set. After the evaluation outcome is observed, the realized DOT value of an allocation x is

$$V_{DOT}^{real}(x) = \sum_{a,t} E_{a,t}^{DOT}(2) (1 - \exp(-\beta x_{a,t})). \quad (3.3)$$

The hindsight Oracle-DOT allocation is

$$x^{Oracle}(B) \in \arg \max_{x \in \mathcal{X}(B)} V_{DOT}^{real}(x), \quad (3.4)$$

which is not an implementable forecasting policy because it uses the realized evaluation target. Forecast-induced decision quality is summarized by the oracle ratio

$$OR^F(B) = \frac{V_{DOT}^{real}(x^F(B))}{V_{DOT}^{real}(x^{Oracle}(B))}, \quad (3.5)$$

and relative decision regret

$$Regret^F(B) = 1 - OR^F(B). \quad (3.6)$$

This separation between the forecast score and the realized evaluation target is the defining information constraint of the forecast-to-prescription experiment.

3.8. Uncertainty-aware robust forecast allocation

The deterministic Forecast-PTO formulation conditions the allocation on a single forecast score. To propagate forecast uncertainty explicitly into the downstream decision, let $s_{a,t}^{(m)}$ denote the implementable score produced by forecasting rule $m \in \mathcal{M}$ for carrier a and evaluation month t . The model set contains Gradient Boosting, Historical Average, Lag-1 DOT, Lag-1 BTS, Seasonal-12 DOT, Ridge, and Rolling-3 DOT. We summarize cross-model disagreement using the robust median absolute deviation

$$\widehat{\sigma}_{a,t} = 1.4826 \text{ median}_{m \in \mathcal{M}} |s_{a,t}^{(m)} - \text{median}_{j \in \mathcal{M}} s_{a,t}^{(j)}|. \quad (3.7)$$

The realized current-month DOT target is not used in $\widehat{\sigma}_{a,t}$. Consequently, this quantity is an ex-ante model-disagreement measure rather than a realized-error estimate or a probabilistically calibrated confidence interval.

Using the nonnegative Gradient Boosting Forecast-PTO score $s_{a,t}^F$ as the center, define the box uncertainty set

$$\mathcal{U}_{a,t}(\kappa) = \left[\max\{0, s_{a,t}^F - \kappa \widehat{\sigma}_{a,t}\}, s_{a,t}^F + \kappa \widehat{\sigma}_{a,t} \right], \quad \kappa \geq 0, \quad (3.8)$$

where κ controls conservatism. The uncertainty-aware allocation solves

$$\max_{x \in \mathcal{X}(B)} \min_{\widetilde{s}_{a,t} \in \mathcal{U}_{a,t}(\kappa)} \sum_{a,t} \widetilde{s}_{a,t} (1 - \exp(-\beta x_{a,t})). \quad (3.9)$$

For $\kappa = 0$, Equation (3.9) reduces exactly to deterministic Forecast-PTO. This nesting permits a controlled comparison of deterministic and robust forecast-conditioned decisions under the same response model, budget, and capacity constraints.

4. Theoretical analysis and structural insights

In this section, we analyze the structural properties of the harm-aware allocation model. The goal is not to introduce a new general-purpose optimization algorithm. Rather, we characterize how official service harm, reconstructed operational risk, cancellation penalties, budget constraints, intervention costs, and diminishing returns jointly shape the allocation decision. These results provide an interpretable bridge between the model primitives and the empirical allocation patterns reported in Section 6.

Throughout this section, let $k = (a, t)$ index a carrier-month pair, and let $\mathcal{K} = \mathcal{A} \times \mathcal{T}$. Let $H_k \geq 0$ denote the service-harm score assigned to carrier-month k . Depending on the implementation, H_k may be constructed from the DOT official harm measure, the BTS reconstructed risk measure, or the blended predictive-to-prescriptive score. Let $x_k \geq 0$ denote the resource assigned to k , let $c_k > 0$ denote the unit cost, and let $\bar{x}_k \in (0, \infty]$ denote an optional upper bound. The feasible allocation set is

$$\mathcal{X}(B) = \left\{ x \in \mathbb{R}_+^{|\mathcal{K}|} : \sum_{k \in \mathcal{K}} c_k x_k \leq B, \quad 0 \leq x_k \leq \bar{x}_k \quad \forall k \in \mathcal{K} \right\}. \quad (4.1)$$

We study two allocation models. The linear model maximizes

$$\max_{x \in \mathcal{X}(B)} \sum_{k \in \mathcal{K}} s_k x_k, \quad (4.2)$$

where $s_k \geq 0$ is a priority score. The diminishing-return model maximizes

$$\max_{x \in \mathcal{X}(B)} \sum_{k \in \mathcal{K}} G_k(x_k), \quad G_k(x_k) = H_k (1 - \exp(-\beta_k x_k)), \quad (4.3)$$

where $\beta_k > 0$ is the scenario-response curvature parameter.

4.1. Priority structure of the linear allocation model

We first characterize the linear allocation model. The model in Equation (4.2) is transparent because its allocation rule is fully determined by cost-adjusted priority scores. Define the priority index

$$\pi_k = \frac{s_k}{c_k}. \quad (4.4)$$

Proposition 4.1 (Priority-index rule). *Suppose $s_k \geq 0$, $c_k > 0$, and $0 < \bar{x}_k < \infty$ for all $k \in \mathcal{K}$. Then an optimal solution to Equation (4.2) is obtained by allocating resources in nonincreasing order of $\pi_k = s_k/c_k$, subject to the budget and upper-bound constraints.*

Proof. The problem is a continuous knapsack problem. Consider two carrier-months k and ℓ with $\pi_k > \pi_\ell$. If $x_\ell > 0$ and $x_k < \bar{x}_k$, shifting an infinitesimal amount of budget from ℓ to k changes the objective by

$$\pi_k dB - \pi_\ell dB > 0.$$

Hence, no optimal solution can allocate marginal budget to a lower-index carrier-month while a higher-index carrier-month remains unfilled. Repeating this exchange argument yields the stated priority-index allocation rule. $\square$

Proposition 4.1 is useful for interpretation. The linear model does not rank carrier-months by harm alone. It ranks intervention opportunities by priority-weighted scenario value per unit cost. This distinction is important when two carrier-months have similar harm but different intervention costs, or when a high-volume carrier-month has large aggregate harm but low marginal cost effectiveness.

4.2. Cancellation-penalty sensitivity

Under the corrected mutually exclusive construction,

$$H_{a,t}^{DOT}(\lambda) = L_{a,t}^{nc} + \lambda L_{a,t}^{can}, \quad (4.5)$$

and

$$R_{a,t}^{BTS}(\lambda) = L_{a,t}^{nc,BTS} + \lambda 100R_{a,t}^{can,BTS}. \quad (4.6)$$

Proposition 4.2 (Cancellation-penalty sensitivity). *For each carrier-month (a, t) ,*

$$\frac{\partial H_{a,t}^{DOT}(\lambda)}{\partial \lambda} = L_{a,t}^{can}, \quad \frac{\partial R_{a,t}^{BTS}(\lambda)}{\partial \lambda} = 100R_{a,t}^{can,BTS}.$$

Consequently, increasing λ weakly increases the relative priority of cancellation-intensive carrier-months without double counting cancellations in the non-cancellation component.

Proof. The derivative expressions follow directly from the corrected definitions. If two carrier-months differ in cancellation loss, the observation with the larger cancellation component receives the larger marginal increase as λ rises. The non-cancellation component is unaffected by λ . $\square$

The parameter λ is therefore an interpretable scenario weight. It is not estimated as a causal valuation of cancellation harm.

4.3. Concavity and diminishing marginal scenario value

We next analyze the diminishing-return improvement function. For each carrier-month k , the priority-weighted scenario-value function is

$$G_k(x_k) = H_k(1 - \exp(-\beta_k x_k)).$$

Its marginal scenario value is

$$MB_k(x_k) = G'_k(x_k) = H_k \beta_k \exp(-\beta_k x_k). \quad (4.7)$$

Proposition 4.3 (Monotonicity and concavity). *For each $k \in \mathcal{K}$, if $H_k \geq 0$ and $\beta_k > 0$, then $G_k(x_k)$ is nondecreasing and concave in x_k . Moreover, the marginal scenario value $MB_k(x_k)$ is increasing in H_k , increasing in β_k at $x_k = 0$, and decreasing in x_k .*

Proof. The first derivative is given in Equation (4.7) and is nonnegative. The second derivative is

$$G_k''(x_k) = -H_k\beta_k^2 \exp(-\beta_k x_k) \leq 0.$$

Thus, G_k is nondecreasing and concave. In addition,

$$\frac{\partial MB_k(x_k)}{\partial H_k} = \beta_k \exp(-\beta_k x_k) > 0,$$

and $MB_k(0) = H_k\beta_k$, which is increasing in β_k . Finally, $G_k''(x_k) \leq 0$ implies that $MB_k(x_k)$ is decreasing in x_k . $\square$

Proposition 4.3 formalizes the operational intuition that high-priority and highly scenario-responsive carrier-months should receive early attention, but that their marginal priority decreases as resources are allocated. This feature prevents the model from concentrating all resources on one carrier-month unless its cost-adjusted marginal benefit remains dominant.

4.4. Optimality conditions and water-filling structure

The diminishing-return model admits an interpretable marginal allocation rule. Let $x^*(B)$ denote an optimal solution to Equation (4.3). The cost-adjusted marginal scenario value of carrier-month k is

$$\Gamma_k(x_k) = \frac{H_k\beta_k \exp(-\beta_k x_k)}{c_k}. \quad (4.8)$$

Proposition 4.4 (Marginal scenario-value equalization). *Suppose x^* is an optimal solution to Equation (4.3). For every carrier-month k with $0 < x_k^* < \bar{x}_k$, there exists a multiplier $\eta \geq 0$ such that*

$$\Gamma_k(x_k^*) = \eta.$$

For carrier-months at the lower bound, $\Gamma_k(0) \leq \eta$; for carrier-months at the upper bound, $\Gamma_k(\bar{x}_k) \geq \eta$.

Proof. The result follows from the Karush–Kuhn–Tucker (KKT) conditions. The Lagrangian for the maximization problem can be written as

$$\mathcal{L}(x, \eta, u, v) = \sum_{k \in \mathcal{K}} G_k(x_k) - \eta \left(\sum_{k \in \mathcal{K}} c_k x_k - B \right) + \sum_{k \in \mathcal{K}} u_k x_k + \sum_{k \in \mathcal{K}} v_k (\bar{x}_k - x_k),$$

where $\eta, u_k, v_k \geq 0$. For an interior allocation, $u_k = v_k = 0$, and stationarity gives

$$G_k'(x_k^*) - \eta c_k = 0.$$

Dividing by c_k gives $\Gamma_k(x_k^*) = \eta$. The boundary inequalities follow from complementary slackness and stationarity. $\square$

This condition has a water-filling interpretation. Resources are allocated until all interior carrier-months have the same cost-adjusted marginal scenario value. Carrier-months with high initial marginal value enter the allocation earlier, but their priority declines as intervention resources are assigned.

The economic interpretation is deliberately conditional. The KKT condition characterizes optimal allocation under the selected monotone concave response scenario. Because the DOT–BTS panel does not observe intervention-outcome pairs, neither H_k nor β_k should be read as an estimated causal marginal effect. Empirical calibration of intervention-specific response functions requires additional intervention and subsequent-outcome data.

When there are no upper bounds, the optimal interior allocation has a closed form. If $x_k^* > 0$, then

$$\frac{H_k \beta_k \exp(-\beta_k x_k^*)}{c_k} = \eta,$$

so that

$$x_k^*(\eta) = \frac{1}{\beta_k} \log \left(\frac{H_k \beta_k}{\eta c_k} \right), \quad \text{whenever } \frac{H_k \beta_k}{c_k} > \eta. \quad (4.9)$$

With nonnegativity imposed, this can be written as

$$x_k^*(\eta) = \left[\frac{1}{\beta_k} \log \left(\frac{H_k \beta_k}{\eta c_k} \right) \right]_+, \quad (4.10)$$

where $[z]_+ = \max\{z, 0\}$, and η is chosen so that the budget constraint is satisfied.

Equation (4.10) shows explicitly how the allocation depends on the priority weight, scenario responsiveness, and cost. Higher H_k raises priority, while higher c_k lowers it. A larger β_k raises the initial marginal priority $H_k \beta_k / c_k$ but also accelerates marginal-response decay, so its effect on the final allocation level need not be monotone. The multiplier η represents the shadow price of the budget.

4.5. Budget monotonicity and shadow value

Let $V(B)$ denote the optimal value of the diminishing-return problem:

$$V(B) = \max_{x \in \mathcal{X}(B)} \sum_{k \in \mathcal{K}} G_k(x_k). \quad (4.11)$$

Proposition 4.5 (Budget monotonicity and concavity of the value function). *The value function $V(B)$ is nondecreasing and concave in B . Moreover, when V is differentiable, $V'(B) = \eta(B)$, where $\eta(B)$ is the optimal multiplier on the budget constraint.*

Proof. If $B_1 \leq B_2$, then $\mathcal{X}(B_1) \subseteq \mathcal{X}(B_2)$, so maximizing the same objective over the larger feasible set cannot reduce the optimal value. Hence, $V(B)$ is nondecreasing.

To show concavity, let $B_\theta = \theta B_1 + (1 - \theta) B_2$ for $\theta \in [0, 1]$, and let $x^1 \in \mathcal{X}(B_1)$ and $x^2 \in \mathcal{X}(B_2)$ be optimal allocations. Because $\mathcal{X}(B)$ is convex in (x, B) , the allocation $\theta x^1 + (1 - \theta) x^2$ is feasible for B_θ . By concavity of $\sum_k G_k(x_k)$,

$$V(B_\theta) \geq \sum_k G_k(\theta x_k^1 + (1 - \theta) x_k^2) \geq \theta V(B_1) + (1 - \theta) V(B_2).$$

Thus, $V(B)$ is concave. The derivative statement follows from the standard envelope theorem for concave resource allocation problems. $\square$

This result has an empirical implication. Increasing the budget should weakly improve attainable priority-weighted scenario value, but the incremental value of additional budget should decline once high-marginal-value opportunities have already been served. Therefore, budget sensitivity should be evaluated not only by total scenario value, but also by marginal scenario value per additional budget unit.

4.6. Linear allocation as a local approximation

The linear allocation model can be interpreted as a local approximation of the diminishing-return model. Around $x_k = 0$,

$$G_k(x_k) = H_k(1 - \exp(-\beta_k x_k)) = H_k \beta_k x_k + O(x_k^2).$$

Thus, at low allocation levels, the diminishing-return model behaves like a linear model with score

$$s_k = H_k \beta_k. \quad (4.12)$$

This connection clarifies the role of the two models. The linear model is appropriate when budgets are small or when interventions are interpreted as marginal improvements. The diminishing-return model becomes more relevant as budgets increase and repeated intervention in the same carrier-month generates declining marginal returns.

4.7. Information loss in benchmark allocation rules

The benchmark policies differ from the proposed allocation model in the information they use and the structural trade-offs they capture. The equal-allocation policy ignores heterogeneity in service harm, exposure, risk, responsiveness, and cost. It can be reasonable as a fairness-oriented administrative rule, but it does not target priority-weighted service-harm value.

The delay-only policy captures arrival reliability but ignores cancellation risk. It may under-prioritize carrier-months in which cancellations are the dominant service failure. The cancellation-only policy captures severe service failures but ignores broader delay exposure. It may over-prioritize low-volume carrier-months with high cancellation rates but limited total operational exposure. The volume-only policy captures scale but ignores service quality; it may allocate resources to large carrier-months even when their reliability is comparatively strong. The historical-average policy captures persistent carrier-level service problems but may fail to respond to month-specific disruptions.

The information used by the allocation rule differs between the retrospective and prospective settings. In the retrospective diagnostic, the PTO score combines the linked DOT and BTS measurement layers. In the prospective experiment, Forecast-PTO instead uses an ex-ante forecast constructed without realized current-month DOT harm. In both settings, the downstream optimization can incorporate exposure, feasibility constraints, heterogeneous costs and capacities when imposed, and diminishing returns. The method is therefore not merely a ranking rule; it is a resource-allocation rule whose priorities are determined by cost-adjusted marginal scenario value.

4.8. Robust counterpart under box forecast uncertainty

Proposition 4.6 (Exact lower-bound robust counterpart). *Suppose $x \in \mathcal{X}(B)$ satisfies $x_{a,t} \geq 0$ and the response term $r(x) = 1 - \exp(-\beta x)$ is nonnegative. For the componentwise box uncertainty set in*

Equation (3.8), the max–min problem in Equation (3.9) is equivalent to

$$\max_{x \in X(B)} \sum_{a,t} \underline{s}_{a,t}(\kappa) (1 - \exp(-\beta x_{a,t})), \quad (4.13)$$

where

$$\underline{s}_{a,t}(\kappa) = \max\{0, s_{a,t}^F - \kappa \widehat{\sigma}_{a,t}\}. \quad (4.14)$$

Proof. For any feasible allocation, every response coefficient $1 - \exp(-\beta x_{a,t})$ is nonnegative. The inner objective is therefore componentwise nondecreasing in each uncertain score $\widehat{s}_{a,t}$. Because the uncertainty set is rectangular, the minimizing value of each component is its lower endpoint $\underline{s}_{a,t}(\kappa)$. Substitution of these componentwise minima yields Equation (4.13). $\square$

The proposition shows that the robust model preserves the same water-filling solution structure as the deterministic diminishing-return allocation after replacing the center forecast score with its uncertainty-adjusted lower bound. The result concerns protection against the specified ambiguity set; it does not imply that a larger κ must improve realized ex-post value when the uncertainty radius is imperfectly aligned with future forecast error.

4.9. Implications for empirical implementation

The structural results guide the empirical analysis in four ways. First, Proposition 4.1 motivates reporting priority rankings and selected carrier-months under the linear model. Second, Proposition 4.2 motivates sensitivity analysis over the cancellation penalty λ . Third, Propositions 4.3 and 4.4 motivate comparing score-based allocation with the diminishing-return allocation model. Fourth, Proposition 4.5 motivates evaluating policy performance across multiple budget levels and reporting marginal gains from additional budget.

These results show how the model converts service-quality data into prescriptive allocation decisions. The empirical sections use these insights to evaluate how forecast-based service-harm priorities translate into budget-feasible decisions under transparent response scenarios and how those decisions compare with simpler benchmark rules.

5. Empirical design and diagnostics

In this section, we describe the empirical implementation of the proposed harm-aware predictive-to-prescriptive framework. The empirical design has eight objectives. First, we construct official carrier-month service-harm measures from DOT Air Travel Consumer Reports. Second, we reconstruct operational disruption risk from BTS flight-level on-time performance records and aggregate the reconstructed measures to the same carrier-month resolution. Third, we evaluate cross-layer DOT–BTS consistency and directly test whether the observed association survives rate normalization, operating-volume controls, within-group demeaning, fixed-effect residualization, and volume stratification. Fourth, we study retrospective allocation structure under cancellation-penalty, diminishing-return, and DOT–BTS blending sensitivity. Fifth, we test temporal persistence using one-month-ahead lagged information and a dominant-carrier exclusion check. Sixth, we construct an explicit ex-ante rolling forecasting layer. Seventh, we pass the forecast into a leakage-controlled allocation experiment and relax homogeneous

feasibility through pre-period-anchored relative-cost/capacity scenarios. Eighth, we propagate ex-ante forecast-model disagreement through an explicit max–min robust allocation model.

The empirical design therefore follows a measurement-audit-before-prescription logic. Because DOT and BTS are linked reporting layers of the same underlying operations, the first step is a reconstruction-consistency audit rather than an independent-source validation exercise. We then separate retrospective structural diagnostics from ex-ante forecasting and prospective allocation, and finally evaluate heterogeneity and explicit forecast uncertainty. This sequence separates measurement consistency from predictive and prescriptive performance and reduces the risk that policy comparisons are driven by common operating scale, same-period information, data-source inconsistencies, or a finely tuned parameter choice.

5.1. Data sources and unit of analysis

The empirical analysis combines two official U.S. aviation data sources. The first source is the DOT Air Travel Consumer Reports, which provide official service-quality outcomes at the carrier-month level [30]. These reports include arrival on-time rates, cancellation rates, scheduled operations, canceled operations, and related service-quality rankings. The second source is the BTS TranStats Airline On-Time Performance Data, specifically the Marketing Carrier On-Time Performance table, which provides flight-level operational records beginning in January 2018 [38]. These records include flight date, marketing carrier, operating carrier, origin and destination airports, origin and destination states, departure delay, arrival delay, cancellation status, diversion status, distance, and delay-cause variables.

The baseline unit of analysis is a carrier-month pair (a, t) , where $a \in \mathcal{A}$ indexes the marketing carrier and $t \in \mathcal{T}$ indexes the calendar month. The main sample covers January 2024 through December 2025 and contains 10 marketing carriers observed over 24 months, yielding 240 carrier-month observations. January and February 2026 are retained as a robustness sample, yielding 18 additional carrier-month observations after applying the same carrier-availability and reporting rules.

The BTS flight-level records are first standardized at the monthly file level and then aggregated to the carrier-month level. As a data-validation check, the September 2025 BTS extract contains 621,601 flight-level observations across the 10 marketing carriers retained in the main sample. In that monthly extract, the observed cancellation rate is approximately 0.005, and the arrival-delay rate of at least 15 minutes is approximately 0.163. This check confirms that the monthly BTS files contain the expected carrier coverage, airport-state information, cancellation indicators, and delay indicators before aggregation.

The carrier-month design is appropriate for two reasons. First, it matches the reporting frequency of official DOT service-quality outcomes. Second, it allows BTS flight-level records to be aggregated into operational-risk measures that are directly comparable to DOT service-harm measures. This common observational unit is essential because the downstream allocation model operates on carrier-month service-improvement opportunities rather than individual flights.

5.2. DOT service-harm construction

For each carrier-month (a, t) , the empirical analysis uses the mutually exclusive DOT decomposition from Section 3.1:

$$L_{a,t}^{nc} = 100 - OT_{a,t} - CR_{a,t}, \quad H_{a,t}^{DOT}(\lambda) = L_{a,t}^{nc} + \lambda CR_{a,t}.$$

At $\lambda = 2$,

$$H_{a,t}^{DOT}(2) = (100 - OT_{a,t}) + CR_{a,t}.$$

The corresponding exposure-weighted target is

$$E_{a,t}^{DOT}(2) = \frac{100 - OT_{a,t} - CR_{a,t}}{100} S_{a,t} + 2C_{a,t}.$$

Relative to a nonexclusive additive specification that would count cancellations inside the non-on-time share and again through the cancellation penalty, this decomposition removes exactly one cancellation-rate term. Across all 258 carrier-month observations, the numerical identity between that difference and $CR_{a,t}$ holds to machine precision (maximum absolute discrepancy 1.15×10^{-14}). The change is largest in high-cancellation months but preserves the broad carrier-month ordering: the mutually exclusive and nonexclusive specifications have Pearson correlation 0.9927 and Spearman correlation 0.9938 in the 2024–2025 main sample.

5.3. BTS flight-level risk reconstruction

The flight-level reconstruction is revised in parallel. For each carrier-month, we count delayed operated arrivals, diverted flights, and canceled flights as mutually exclusive categories. The corrected BTS rate and exposure measures are

$$R_{a,t}^{BTS}(2) = 100 \frac{D_{a,t} + V_{a,t} + 2C_{a,t}}{N_{a,t}},$$

and

$$E_{a,t}^{BTS}(2) = D_{a,t} + V_{a,t} + 2C_{a,t}.$$

This definition places the BTS reconstruction on the same event decomposition as the DOT measure. The empirical pipeline rebuilds all 26 monthly BTS feature files before forecasting so that every downstream analysis uses the same harm construction.

5.4. Domestic route filtering

The canonical BTS aggregation uses a domestic-route filter. A flight is retained if both its origin and destination state codes correspond to U.S. states, the District of Columbia, or U.S. territories included in BTS reporting. Formally, a flight i is retained if

$$s_i^{orig} \in S_{US} \quad \text{and} \quad s_i^{dest} \in S_{US},$$

where s_i^{orig} and s_i^{dest} are origin and destination state codes, and S_{US} denotes the domestic state-code set used in the analysis.

The filter serves two purposes. First, it aligns the flight-level BTS aggregation with the domestic service-quality environment represented in the DOT carrier-month reports. Second, it avoids mixing

domestic operational risk with international route structures that may have different reporting coverage, delay mechanisms, and cancellation patterns. In the processed data, the domestic filter removes only a small number of records per month while preserving the full set of reported marketing carriers in the main sample.

5.5. Carrier mapping and DOT–BTS merge

The DOT reports identify carriers by official report names, whereas BTS flight-level data identify marketing carriers using carrier codes. We therefore construct a deterministic carrier crosswalk before merging the two data sources. The crosswalk maps AA to American Airlines Network, AS to Alaska Airlines Network, B6 to JetBlue Airways, DL to Delta Air Lines Network, F9 to Frontier Airlines, G4 to Allegiant Air, HA to Hawaiian Airlines, NK to Spirit Airlines, UA to United Airlines Network, and WN to Southwest Airlines. Carrier names are standardized to a common uppercase format before merging in order to avoid mismatches caused by capitalization, spacing, or naming conventions.

The DOT and BTS panels are merged on the carrier-month key (a, t) . The resulting merged carrier-month panel contains both official DOT service-harm measures and BTS reconstructed operational-risk measures. The merge is exact in the main sample: Every DOT carrier-month observation has a corresponding BTS carrier-month observation, and every BTS carrier-month observation has a corresponding DOT record. This exact merge is important because the allocation experiments compare policies over a common feasible set of carrier-month units.

5.6. Alignment diagnostics

Before using reconstructed BTS risk measures in the prescriptive allocation experiments, we evaluate their alignment with official DOT service-quality outcomes. The alignment diagnostics answer the following question: Do flight-level BTS reconstructions recover the official DOT service-quality patterns at the carrier-month level?

For each carrier-month, we compute the difference between DOT harm and BTS risk:

$$\Delta_{a,t}^H = H_{a,t}^{DOT} - R_{a,t}^{BTS}, \quad (5.1)$$

and the corresponding absolute difference:

$$A_{a,t}^H = |\Delta_{a,t}^H|. \quad (5.2)$$

We compute analogous differences for arrival-delay rates, cancellation rates, and exposure-weighted measures.

The alignment analysis uses five diagnostic families. First, we report mean, median, and maximum absolute differences between DOT and BTS measures. Second, we compute Pearson and Spearman correlations between official DOT harm and reconstructed BTS risk. Pearson correlation evaluates linear agreement, whereas Spearman correlation evaluates rank-order agreement. Rank-order agreement is especially important because budget-constrained allocation decisions often depend on priority ordering rather than exact equality of scores. Third, we compare monthly carrier rankings under DOT harm and BTS risk, including top- q overlap for the highest-harm carrier-months. Fourth, we visualize the relationship between DOT harm and BTS risk using scatter plots and monthly trend comparisons.

Fifth, we explicitly test whether the high DOT–BTS association is explained only by common operating scale. Let $V_{a,t}$ denote scheduled operations and let $E_{a,t}^j$ denote exposure-weighted harm for reporting layer $j \in \{DOT, BTS\}$. We evaluate rate-based measures, exposure per 100 scheduled operations, within-carrier demeaning, within-month demeaning, and carrier–month two-way demeaning. We then estimate the layer-specific residualization model

$$E_{a,t}^j = \alpha_j + \theta_j \log V_{a,t} + \mu_a^j + \tau_t^j + u_{a,t}^j, \quad j \in \{DOT, BTS\}, \quad (5.3)$$

and compute the correlation between $u_{a,t}^{DOT}$ and $u_{a,t}^{BTS}$. We also report partial correlation controlling only for $\log V_{a,t}$, repeat the comparison within scheduled-operation quartiles, and use 1,000 month-block bootstrap replications with seed 20260824 for percentile uncertainty intervals. These checks directly separate common scale from cross-layer measurement alignment while preserving the month-level carrier structure of the data.

These diagnostics are not intended to establish independence between the two reporting layers, prove that they are identical, or demonstrate incremental information from combining them. Exact equality is neither expected nor required because DOT and BTS measures may differ in reporting conventions, timing, rounding, and aggregation rules. The purpose is more specific: to assess whether BTS flight-level reconstructed risk is sufficiently consistent with official DOT service harm, including after common volume effects are removed, to be used as an operational measurement and forecasting signal in the allocation framework.

5.7. Resource allocation experiment design

After establishing DOT–BTS alignment, we evaluate resource allocation policies. Each policy allocates a fixed budget B across carrier-month units. The empirical objective is to estimate how effectively different policies target high-harm service-quality observations.

We evaluate normalized budget fractions

$$b \in \{0.05, 0.10, 0.20, 0.30\}.$$

For a given budget fraction b , the corresponding feasible budget is

$$B(b) = b \sum_{a,t} c_{a,t} \bar{x}_{a,t}.$$

Under the homogeneous unit-cost and unit-upper-bound benchmark used to isolate score differences, this definition reduces to a fraction b of total allocable intervention capacity. We therefore report b as the budget fraction while retaining B for the actual optimization budget. Experiment 09 subsequently relaxes this assumption with carrier-specific relative costs and capacity limits anchored only to pre-2025 operating information. Within each experiment and scenario, all policies face the same feasible set so that performance differences are attributable to the priority signal and the optimization rule rather than to unequal budgets.

The proposed policy uses the blended score

$$s_{a,t}^{PTO} = \omega E_{a,t}^{DOT} + (1 - \omega) E_{a,t}^{BTS}, \quad (5.4)$$

where $\omega \in [0, 1]$. In the baseline implementation, we use $\omega = 0.5$, assigning equal weight to official DOT service harm and BTS flight-level reconstructed risk. Sensitivity analysis varies ω to examine

whether the results depend on the relative weight assigned to official outcomes versus reconstructed operational risk.

We evaluate two allocation forms. The first is a score-based linear allocation policy, which allocates resources to carrier-months with the highest cost-adjusted priority scores subject to budget and upper-bound constraints. The second is the diminishing-return allocation policy developed in Sections 3 and 4, where the priority-weighted scenario-response function is

$$G_{a,t}(x_{a,t}) = H_{a,t} (1 - \exp(-\beta x_{a,t})) . \quad (5.5)$$

For the baseline analysis, we use a common scenario-response parameter β across carrier-months. This parameter is not estimated as a causal effect; the empirical analysis uses it to examine the structural allocation implications of monotone diminishing marginal response.

5.8. Benchmark policies

We compare the proposed policy against five benchmark policies.

The equal-budget-share benchmark assigns equal shares of the available budget across eligible carrier-months, with resource quantities adjusted by unit cost. This benchmark represents a simple administrative rule that ignores heterogeneity in harm, exposure, and operational risk. The delay-only benchmark ranks carrier-months using the corrected non-cancellation DOT exposure component, while the cancellation-only benchmark uses the corrected cancellation DOT exposure component. The volume-only benchmark ranks carrier-months using scheduled operations or BTS flight counts. The historical-average benchmark ranks carriers using only DOT harm observed strictly before the evaluation month in prospective experiments.

These benchmarks represent common simplified decision rules. Equal allocation ignores heterogeneity in risk and harm. Delay-only and cancellation-only rules ignore the multidimensional structure of service quality. Volume-only allocation prioritizes large carriers regardless of service harm. Historical-average allocation captures persistent carrier-level differences but may fail to identify month-specific disruption patterns. The comparison therefore evaluates whether integrating official harm, flight-level risk, and exposure improves allocation quality relative to single-signal rules.

5.9. Evaluation metrics

We evaluate allocation policies using harm-capture and improvement-efficiency metrics. Let x denote the allocation produced by a policy. The first metric is captured rate-based harm:

$$\text{CapturedHarm}(x) = \sum_{a,t} H_{a,t} x_{a,t} . \quad (5.6)$$

The second metric is exposure-weighted captured harm:

$$\text{CapturedExposure}(x) = \sum_{a,t} E_{a,t} x_{a,t} . \quad (5.7)$$

The third metric is total priority-weighted scenario value under the diminishing-return model:

$$V(x) = \sum_{a,t} H_{a,t} (1 - \exp(-\beta x_{a,t})) . \quad (5.8)$$

The fourth metric is normalized improvement value (NIV):

$$\text{NIV}(x) = \frac{V(x)}{\max_{b \in \mathcal{B}} V(x^b)}, \quad (5.9)$$

where $\mathcal{B}$ denotes the set of benchmark and proposed policies.

We also report allocation concentration across carriers and months. This diagnostic helps identify whether a policy over-concentrates resources in a small number of carrier-months or distributes them across a broader set of service-risk targets. Concentration is reported using the share of allocated resources assigned to the top carrier, the top three carriers, and the top decile of carrier-months. These diagnostics complement the decision-value metrics by showing whether performance gains arise from focused targeting or broad allocation.

5.10. Forecast evaluation design

To establish a genuine predictive layer before prescriptive evaluation, we construct a temporally ordered carrier-month forecasting experiment with exposure-weighted DOT service harm at $\lambda = 2$ as the target. The locked evaluation panel contains 120 carrier-month observations covering 10 marketing carriers from January through December 2025. Forecasts are generated without using the realized current-month target as an input. We compare Gradient Boosting with Historical Average, Lag-1 DOT, Lag-1 BTS, Seasonal-12 DOT, Ridge, and Rolling-3 DOT benchmarks.

Forecast performance is evaluated in both value and ranking space. Point accuracy is summarized by normalized mean absolute error (NMAE), while Pearson correlation measures level association. Because the downstream allocation problem depends strongly on relative priority within each month, we additionally report monthly Spearman rank correlations. The primary forecasting model is selected using this combined accuracy-and-ranking perspective rather than on a single metric alone.

5.11. Leakage-controlled forecast-to-prescription evaluation

The forecast-to-prescription experiment uses the same locked 2025 panel and passes the primary forecast into the allocation model as the nonnegative score in Equation (3.1). Missing benchmark scores are not imputed from realized current-month DOT harm. The realized DOT target is used only after the allocation has been determined, for realized policy-value calculation and for the hindsight Oracle-DOT benchmark. This rule provides an explicit leakage audit for the complete forecast–optimization–evaluation chain.

We compare Forecast-PTO with Historical Average, Lag-1 DOT, Lag-1 BTS, Seasonal-12 DOT, Ridge, Rolling-3 DOT, Equal allocation, and the hindsight Oracle-DOT benchmark. The primary optimization model is the diminishing-return allocation model with $\beta = 1$, evaluated at $b \in \{0.05, 0.10, 0.20, 0.30\}$. The linear priority-index allocation is retained as a secondary diagnostic. Decision performance is evaluated using realized DOT allocation value, oracle ratio, relative decision regret, number of active carriers, and carrier-level concentration. To preserve the within-month cross-carrier decision structure, uncertainty is evaluated using 1,000 month-block bootstrap replications. We report bootstrap percentile uncertainty intervals rather than treating carrier-month rows as independent observations.

5.12. Heterogeneous relative-cost and capacity robustness

To exercise the heterogeneous-cost and capacity structure without introducing post-outcome information, we construct carrier-specific relative scenarios using only pre-2025 operating history. For carrier a , let m_a denote its 2024 mean scheduled operations and let v_a denote the coefficient of variation of monthly scheduled operations. We convert their percentile ranks to centered indices

$$u_a^S = 2\{\text{rank}(m_a) - 0.5\}, \quad u_a^V = 2\{\text{rank}(v_a) - 0.5\},$$

where the rank is scaled to $[0, 1]$. Larger schedule variability therefore implies a higher relative coordination-cost scenario, whereas larger operating scale implies a larger feasible intervention-capacity scenario.

For each heterogeneity scenario q , preliminary values are

$$c_a^{(q)} = 1 + \alpha_c^{(q)} u_a^V, \quad \bar{x}_a^{(q)} = 1 + \alpha_x^{(q)} u_a^S.$$

The Homogeneous, Mild, Moderate, and Strong scenarios use $(\alpha_c, \alpha_x) = (0, 0), (0.20, 0.25), (0.35, 0.40), (0.50, 0.60)$, respectively. Scenario-specific clipping bounds are $[0.80, 1.20]$ and $[0.75, 1.25]$ for Mild cost and capacity, $[0.65, 1.35]$ and $[0.60, 1.40]$ for Moderate, and $[0.50, 1.50]$ and $[0.40, 1.60]$ for Strong. Cost and capacity values are then normalized to mean one within each scenario. These quantities are normalized relative scenario values, not observed dollar costs or measured intervention limits.

The heterogeneous experiment retains the locked 2025 Forecast-PTO score and the same four budget fractions. For each month and scenario, the actual budget is

$$B_t(b) = b \sum_a c_a^{(q)} \bar{x}_a^{(q)}.$$

The primary model is the diminishing-return water-filling allocation with $\beta = 1$; realized current-month DOT harm is used only after the allocation is fixed, for realized value and the hindsight oracle. We designate Moderate as the primary heterogeneous scenario and retain Mild and Strong as robustness levels. A secondary linear diagnostic compares the theoretically implied cost-aware priority s_a/c_a with the cost-blind raw forecast ranking under exactly the same heterogeneous costs, capacities, and budget. Month-block bootstrap uncertainty again uses 1,000 replications with seed 20260824.

5.13. Ex-Ante model-disagreement robust allocation

The uncertainty-aware experiment uses the same locked 2025 panel of 120 carrier-month observations and the same Gradient Boosting Forecast-PTO center score as the leakage-controlled allocation experiment. For each carrier-month, $\widehat{\sigma}_{a,t}$ is calculated as 1.4826 times the cross-model median absolute deviation across the seven implementable forecasting rules defined in Section 3.8. All seven scores are available for all 120 evaluation observations. No realized current-month DOT outcome enters either the uncertainty radius or the allocation score.

We evaluate $\kappa \in \{0, 0.5, 1.0, 1.5, 2.0\}$, with $\kappa = 0$ serving as the exact deterministic Forecast-PTO reproduction and $\kappa = 1$ designated as the primary robust specification. To isolate forecast uncertainty from the cost/capacity heterogeneity examined separately, this experiment uses homogeneous normalized costs and capacities, $\beta = 1$, and budget fractions $b \in \{0.05, 0.10, 0.20, 0.30\}$. Realized current-month

DOT harm is revealed only after each allocation is fixed and is then used for realized value and the hindsight Oracle-DOT benchmark.

We report oracle ratio, relative regret, monthly lower-tail performance, allocation shifts relative to deterministic Forecast-PTO, and allocation breadth. Uncertainty in robust-minus-deterministic realized performance is assessed with 1,000 paired month-block bootstrap replications using seed 20260824. The same resampled months are used for both policies within each replication so that the reported difference isolates the allocation change.

5.14. *Carrier-airport-month finer-resolution robustness*

To address the aggregation concern directly, we add a finer-resolution robustness experiment on the carrier-airport-month unit. The experiment merges the DOT carrier-airport arrival panel with the BTS carrier-airport arrival-risk panel for the same 10 study carriers over January 2024–December 2025. After canonical matching and positive-volume filtering, the main sample contains 5,807 matched carrier-airport-month observations. The forecasting split uses 2,929 observations from 2024 for training and 2,878 observations from 2025 for out-of-sample evaluation.

Because DOT provides airport-specific arrival performance but not a direct airport-specific analogue of the cancellation-inclusive carrier-month harm target used in the main experiment, this exercise is intentionally arrival-side. It is therefore a finer-resolution robustness check rather than a full route-network recovery model. The forecasting target is DOT arrival-delay exposure. Features use only information available before month t : exact one-month lags, prior three-month means, expanding historical averages, lagged volume terms, and calendar seasonality at the carrier-airport pair level. The primary forecaster is a Gradient Boosting Regressor trained on the 2024 panel; Lag-1 DOT, Lag-1 BTS, Rolling-3 DOT, Historical Average, and Equal allocation serve as benchmarks.

Allocation follows the same diminishing-return water-filling structure with $\beta = 1$, homogeneous unit costs and capacities, and budget fractions 0.05, 0.10, 0.20, and 0.30. The primary inclusion rule retains cells with at least one reported arrival; thresholds of 100 and 500 reported arrivals provide thin-cell robustness checks. Realized current-month DOT arrival-delay exposure is excluded from score construction and is used only for ex-post policy evaluation and hindsight-oracle normalization. The experiment reports DOT–BTS alignment, forecast accuracy, monthly rank fidelity, realized oracle ratio, relative regret, mean active units, and threshold robustness.

5.15. *Empirical outputs*

The empirical analysis produces eleven complementary sets of evidence. First, DOT–BTS alignment diagnostics evaluate cross-layer consistency between DOT-published carrier-month service-quality measures and measures reconstructed from BTS flight-level records. Second, cancellation-penalty sensitivity tests how λ changes priority. Third, the contemporaneous allocation experiment examines whether alternative service-harm representations preserve allocation-relevant ordering under a common feasible set. Fourth, diminishing-return sensitivity evaluates the consequences of varying β . Fifth, blending-weight sensitivity evaluates whether retrospective PTO performance depends on a finely tuned ω . Sixth, the one-month-ahead lagged-information experiment, together with a dominant-carrier exclusion check, examines whether useful priority information persists after current-period DOT information is removed. Seventh, an explicit forecasting experiment evaluates point accuracy

and within-month ranking fidelity on the locked 2025 panel. Eighth, the leakage-controlled forecast-to-prescription experiment replaces the contemporaneous score with the forecast score and reserves realized current-month DOT harm exclusively for evaluation and hindsight-oracle construction. Ninth, the heterogeneous relative-cost/capacity experiment tests whether the principal Forecast-PTO decision results persist when the feasible set is carrier-specific and whether the cost-adjusted priority structure becomes decision-relevant as cost heterogeneity strengthens. Tenth, the ex-ante model-disagreement experiment propagates forecast uncertainty directly into a max–min allocation model and compares the resulting robust decisions with deterministic Forecast-PTO over a transparent conservatism grid. Eleventh, the carrier-airport-month robustness experiment tests whether the main forecast-to-allocation logic remains useful at a finer operational resolution using airport-specific arrival-side outcomes.

This sequence deliberately separates measurement consistency and structural diagnostics from the stricter question of whether forecast-induced decisions retain high realized value. It also prevents the near-oracle ratios in the same-period diagnostic experiments from being interpreted as prospective forecast performance.

6. Empirical results

This section reports the empirical evidence in a sequence that separates forecast performance, cross-layer measurement diagnostics, retrospective allocation behavior, and forecast-driven decision performance. We first report the locked 2025 forecasting experiment because it establishes the operational signal used later by Forecast-PTO. We then examine consistency between DOT-published carrier-month outcomes and BTS flight-level reconstructions, cancellation-penalty sensitivity, and retrospective allocation behavior under linear and diminishing-return formulations. These same-period experiments are interpreted as diagnostic evidence about the measurement and optimization layers rather than as prospective forecasting tests. We next examine blending-weight and lagged-information robustness. Finally, we conduct the leakage-controlled forecast-to-prescription experiment in which realized current-month DOT harm is withheld from the allocation score and used only to evaluate the decision after it has been determined.

6.1. Forecast accuracy and allocation-relevant ranking fidelity

Figure 1 summarizes the leakage-controlled one-month-ahead forecasting results on the locked 2025 panel. Gradient Boosting has the lowest normalized MAE, 0.178, followed by Historical Average (0.205), Lag-1 DOT (0.207), Lag-1 BTS (0.215), Seasonal-12 DOT (0.241), Ridge (0.245), and Rolling-3 DOT (0.264). For the selected Gradient Boosting forecast, the correlation with realized harm is 0.964 and the mean monthly Spearman rank correlation is 0.983 over the 120 carrier-month evaluation observations. Recent-history rules remain competitive in within-month ranking, so the forecasting evidence is interpreted jointly in point-error and ranking space rather than as universal dominance of one forecasting rule.

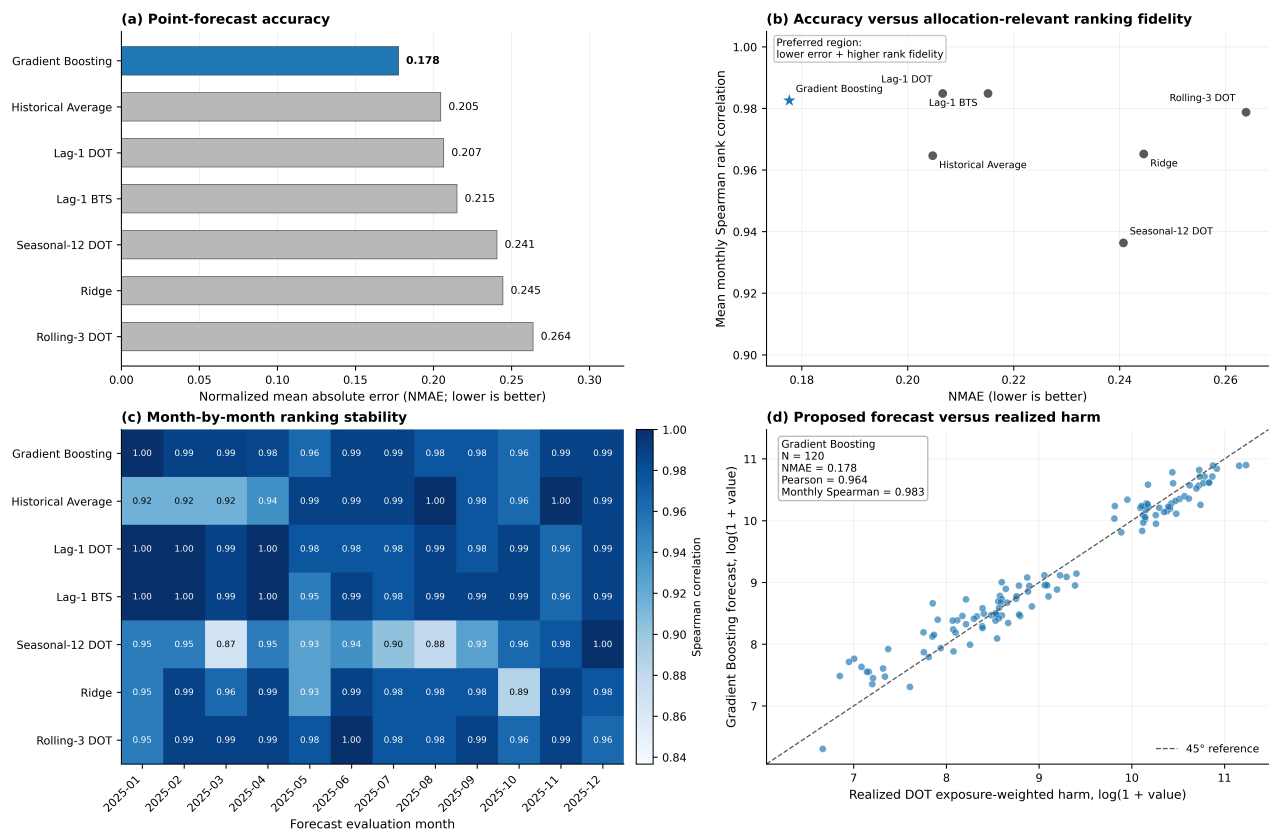


Figure 1. Leakage-controlled one-month-ahead forecast diagnostics on the locked 2025 carrier-month panel. (a) Point-forecast accuracy across the implementable forecasting rules. (b) Point accuracy versus allocation-relevant ranking fidelity. (c) Month-by-month ranking stability. (d) Selected Gradient Boosting forecast versus realized harm on the 120-observation evaluation panel (NMAE 0.178, Pearson 0.964, mean monthly Spearman 0.983). All forecast features are available before month t ; realized month- t DOT harm is excluded from the forecasting features and allocation score.

Table 2. Corrected DOT–BTS cross-layer consistency and scale-control audit.

Analysis	Pearson	Spearman
Corrected rate-based DOT–BTS	0.999983	0.999919
Corrected exposure-weighted DOT–BTS	0.999998	0.999985
Within-carrier exposure	1.0000	0.9999
Within-month exposure	1.0000	1.0000
Carrier+month demeaned exposure	1.0000	0.9999
Exposure controlling log operations	1.0000	1.0000
Exposure residual: volume+carrier FE+month FE	1.0000	0.9999

6.2. Corrected DOT–BTS cross-layer consistency

The corrected DOT and BTS measures are almost identical because they describe the same underlying operating events through linked reporting layers and now use the same mutually exclusive event decomposition. In the 2024–2025 main sample, the corrected rate-based Pearson and Spearman correlations are 0.999983 and 0.999919, respectively; the corresponding exposure correlations are 0.999998 and 0.999985. The correction itself is nontrivial in high-cancellation months: mean DOT harm falls by 1.4354 points, the median reduction is 1.00 point, and the maximum reduction is 11.9 points. Nevertheless, old-versus-corrected DOT harm retains Pearson 0.9927 and Spearman 0.9938, showing that the correction changes the level and cancellation weighting without overturning the broad priority ordering.

The volume-control audit leads to the same interpretation. Log scheduled operations explains approximately 75.66% of both corrected DOT and BTS exposure variation, yet within-carrier, within-month, two-way demeaned, partial-log-volume, and full residual correlations all remain essentially one. These statistics should not be read as independent validation. Rather, they show that the flight-level BTS reconstruction reproduces the DOT-published carrier-month quantities after the event definitions are aligned, and that the correspondence is not merely an artifact of common operating scale.

6.3. Cancellation-penalty sensitivity

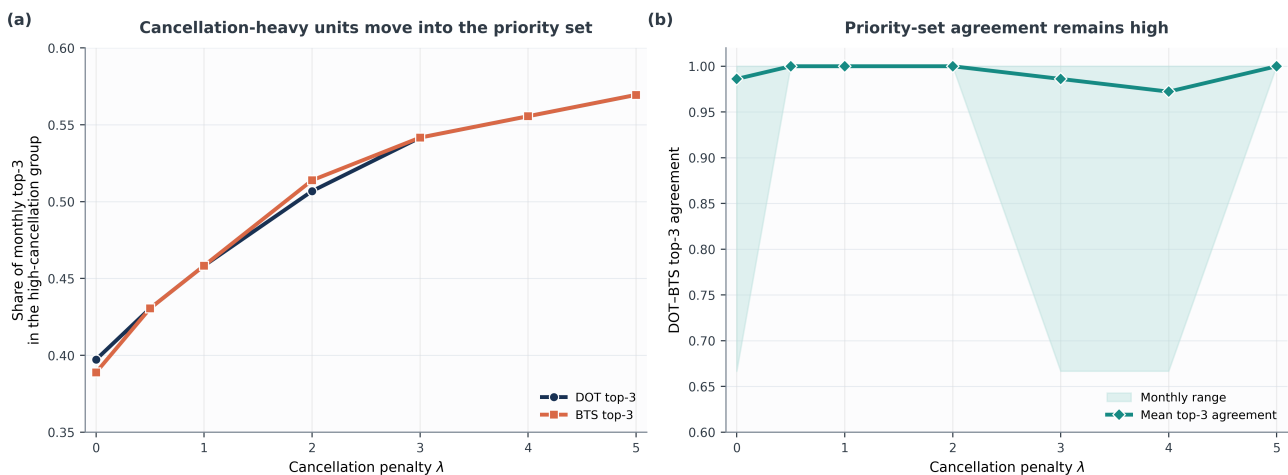


Figure 2. Cancellation-penalty sensitivity under the corrected mutually exclusive service-harm construction. (a) Share of monthly top-three priority observations belonging to the high-cancellation group as the cancellation penalty λ increases under the DOT and BTS measures. (b) Corresponding monthly DOT–BTS top-three priority-set agreement; the shaded region gives the monthly range. Larger cancellation penalties shift priority toward cancellation-heavy carrier-months while the linked DOT and BTS constructions continue to identify highly similar top-priority sets.

We next evaluate whether the corrected mutually exclusive harm construction is sensitive to the cancellation-severity parameter λ . The sensitivity analysis uses the balanced January 2024–

December 2025 main sample (240 carrier-month observations, 10 carriers, 24 months) and varies $\lambda \in \{0, 0.5, 1, 2, 3, 4, 5\}$. The high-cancellation group is defined from the monthly upper quartile of cancellation loss, and the priority diagnostic tracks its representation in the monthly top-three observations.

Figure 2 shows two complementary results. First, raising λ shifts priority toward cancellation-heavy carrier-months: the DOT high-cancellation share rises from 0.397 at $\lambda = 0$ to 0.569 at $\lambda = 5$, while the BTS share rises from 0.389 to 0.569. Second, the monthly DOT–BTS top-three sets remain highly consistent over the full penalty grid. Mean monthly agreement ranges from 0.972 to 1.000; the occasional monthly minimum of 0.667 corresponds to two of the three top-priority carriers being shared. Because DOT and BTS are linked reporting layers, this agreement is interpreted as reconstruction consistency, not independent validation.

Table 3. Corrected cancellation-penalty sensitivity summary.

λ	DOT high-cancel top-3 share	BTS high-cancel top-3 share	Mean top-3 agreement	Monthly minimum
0.0	0.397	0.389	0.986	0.667
0.5	0.431	0.431	1.000	1.000
1.0	0.458	0.458	1.000	1.000
2.0	0.507	0.514	1.000	1.000
3.0	0.542	0.542	0.986	0.667
4.0	0.556	0.556	0.972	0.667
5.0	0.569	0.569	1.000	1.000

6.4. Retrospective resource allocation consistency

We retain the same-period allocation comparison as a retrospective structural diagnostic. It is not used to establish prospective Forecast-PTO performance; the primary deployable evidence is the leakage-controlled month-by-month analysis in Section 6.8. The previous diagnostics established that DOT official service harm and BTS reconstructed operational risk are strongly aligned at the carrier-month level. In this subsection, we examine whether this alignment translates into improved resource allocation decisions under a limited budget.

We compare the proposed PTO policy against eight benchmark policies: DOT-only, BTS-only, equal allocation, delay-only, cancellation-only, volume-only, historical-average, and Oracle-DOT. The Oracle-DOT policy uses the evaluation target itself and therefore provides an upper benchmark for policies evaluated on DOT exposure-weighted harm.

It is important to emphasize that Oracle-DOT is not an implementable forecasting or prescriptive rule in the same informational sense as the PTO policy. Oracle-DOT uses the DOT exposure-weighted harm measure as the evaluation target and serves as a hindsight upper benchmark. In contrast, the proposed PTO policy combines the DOT carrier-month layer with the corresponding BTS flight-level reconstruction. Because these are linked aggregation layers of the same official reporting system, the near-oracle pattern is interpreted as retrospective priority consistency, not as independent external validation and not as evidence that the proposed policy has access to a prospective oracle.

All policies are evaluated on the same set of 240 carrier-month observations from January 2024 to December 2025, under the same budget and upper-bound constraints. The normalized budget fractions

are

$$b \in \{0.05, 0.10, 0.20, 0.30\},$$

with the corresponding actual optimization budget defined by $B(b)$. In the unit-cost implementation, these fractions correspond to 12, 24, 48, and 72 units of allocable intervention capacity, respectively.

The proposed PTO score is defined as

$$s_{a,t}^{PTO} = \omega E_{a,t}^{DOT} + (1 - \omega) E_{a,t}^{BTS}, \quad \omega = 0.5.$$

Thus, the policy assigns equal weight to official DOT exposure-weighted harm and BTS reconstructed exposure-weighted risk. The baseline allocation rule selects carrier-month units in descending order of the corresponding priority score. Because unit intervention costs are used in this experiment, the rule is equivalent to the priority-index allocation rule derived in Proposition 4.1.

For each policy p , let $x^p(B)$ denote the allocation generated under budget B . We evaluate policy performance using the diminishing-return value

$$V^{DOT}(x^p(B)) = \sum_{a,t} E_{a,t}^{DOT} (1 - \exp(-\beta x_{a,t}^p(B))), \quad \beta = 1.$$

We also report the ratio to the Oracle-DOT benchmark:

$$\text{OracleRatio}(p, B) = \frac{V^{DOT}(x^p(B))}{V^{DOT}(x^{\text{Oracle}}(B))}.$$

This metric measures the fraction of the oracle allocation value attained by each policy under the same budget.

Figure 3 summarizes the allocation-performance results. Figure 3(a) reports the allocation loss relative to the Oracle-DOT benchmark. The proposed PTO policy remains nearly indistinguishable from the Oracle-DOT allocation across all budget levels. Its oracle ratios are 0.9998, 0.9979, 1.0000, and 0.9999 for budgets 0.05, 0.10, 0.20, and 0.30, respectively. These results are interpreted only as a same-period priority-consistency diagnostic between closely aligned reporting layers.

These near-one ratios should not be interpreted as prospective predictive performance. Current-period DOT information enters the contemporaneous PTO score, while realized DOT harm also defines the hindsight evaluation benchmark. This experiment therefore establishes retrospective allocation consistency between closely aligned measurement signals. The forecast-to-prescription experiment in Section 6.8 removes this same-period target dependence from the allocation score.

Figure 3(b) compares policy values relative to the Oracle-DOT benchmark. The proposed PTO, DOT-only, BTS-only, and delay-only policies perform similarly because the earlier alignment diagnostics showed that DOT official harm and BTS reconstructed risk generate highly consistent priority rankings. In contrast, equal allocation performs poorly because it ignores heterogeneity in service harm and exposure. Volume-only allocation performs better than equal allocation but remains below the harm-aware policies because flight volume alone does not identify the most severe service-quality disruptions. Historical-average allocation also underperforms because it captures persistent carrier-level differences but responds less effectively to month-specific service-harm variation.

Figure 3(c) reports policy ranks by budget. The rank patterns show that harm-aware policies consistently occupy the highest positions, while benchmark rules that ignore either service harm,

exposure, or month-specific variation remain lower-ranked. Figure 3(d) reports the highest-harm carrier-months selected by the proposed PTO policy at $B = 0.10$. These selected units should be interpreted as high-impact intervention opportunities rather than as carrier-level quality rankings, because the harm score is exposure-weighted and therefore reflects both service-harm intensity and operational scale.

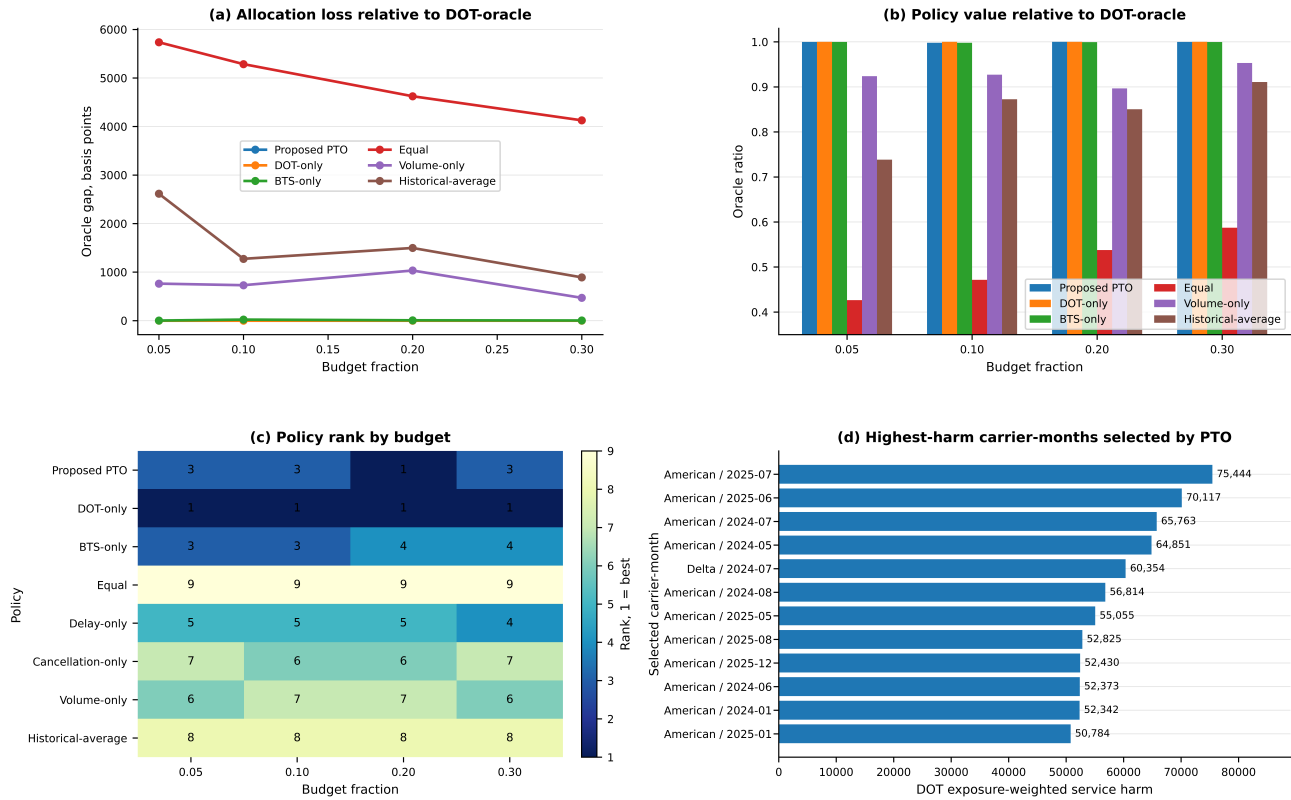


Figure 3. Resource allocation performance in Experiment 03. Figure 3(a) reports allocation loss relative to the Oracle-DOT benchmark. Figure 3(b) compares policy values relative to the Oracle-DOT benchmark across budget levels. Figure 3(c) reports policy ranks by budget. Figure 3(d) reports the highest-harm carrier-months selected by the proposed PTO policy at $B = 0.10$. The selected carrier-months are interpreted as high-impact intervention opportunities, not as carrier-level quality rankings, because the harm score is exposure-weighted and therefore reflects both service-harm intensity and operational scale.

Table 4 reports 95% month-block bootstrap percentile intervals computed using month-block bootstrap replications. DOT captured denotes captured exposure-weighted DOT service harm, while blended captured uses the average of DOT exposure-weighted harm and BTS exposure-weighted risk. DOT value denotes the diminishing-return policy value evaluated using DOT exposure-weighted harm. Oracle ratio is the ratio to the Oracle-DOT policy under the same budget. Top-3 share is the share of allocated resources assigned to the three largest carrier-level allocation recipients. Active units and active carriers measure the breadth of the allocation. The main sample contains 240 carrier-month observations from 10 marketing carriers over January 2024–December 2025.

The bootstrap percentile intervals show that the proposed PTO policy closely tracks the Oracle-DOT benchmark across the full budget range. At small budgets, the PTO policy identifies almost the same highest-harm carrier-months as the Oracle-DOT rule. At larger budgets, the percentile intervals remain concentrated near one, showing that the Oracle-DOT-tracking result is not driven by a single month or by a small number of observations.

The results are consistent with the structural analysis. Under the linear allocation model, the optimal rule is to allocate resources in descending order of cost-adjusted priority. Since the DOT and BTS measures are strongly aligned, the PTO score produces nearly the same high-priority carrier-month set as the Oracle-DOT rule. Hence, the proposed policy closely tracks the Oracle-DOT benchmark while retaining a flight-level operational interpretation through the BTS risk component.

The benchmark comparison also illustrates the value of harm-aware allocation. Equal allocation distributes resources broadly across all carrier-months and therefore fails to target the most severe service-harm observations. Volume-only allocation targets large carrier-months, but it does not account for whether those carrier-months have unusually high delay or cancellation harm. Historical-average allocation captures persistent carrier-level service differences, but it is less responsive to month-specific disruption patterns. Cancellation-only allocation improves targeting of severe cancellation-heavy observations, but it ignores the broader delay burden. The proposed PTO policy combines official service harm, reconstructed operational risk, and exposure, which explains its superior performance relative to these simplified rules.

Overall, the contemporaneous allocation experiment supports a narrower structural conclusion: Closely aligned DOT and BTS service-harm representations produce similar retrospective allocation priorities and outperform simplified rules that ignore key dimensions of harm or exposure. Prospective decision performance is assessed separately in the leakage-controlled forecast-to-prescription experiment.

6.5. Retrospective diminishing-return diagnostic and forecast-based response-curvature sensitivity

We retain a retrospective diminishing-return experiment as a structural diagnostic computed from the corrected mutually exclusive DOT-BTS harm construction. Figure 4 summarizes how the assumed scenario-response parameter β changes allocation breadth, concentration, and benchmark performance under the corrected same-period panel. Because this diagnostic uses contemporaneous service-harm information, it is not used as prospective predictive-to-prescriptive evidence. A separate prospective sensitivity analysis, Experiment 04R, uses the locked ex-ante Forecast-PTO score from Experiment 08 and varies only the curvature of the assumed diminishing-return response.

For Experiment 04R, the response family is

$$G_{a,t}(x_{a,t}) = \widehat{H}_{a,t|t-1}^F (1 - \exp(-\beta x_{a,t})),$$

where $\widehat{H}_{a,t|t-1}^F$ is fixed before month t using only pre-month information. Realized month- t DOT harm is revealed only after the allocation is fixed and is used for ex-post policy-value evaluation and hindsight-oracle normalization. The parameter β is not an estimated causal treatment-effect coefficient; it is a scenario-response curvature parameter controlling how quickly marginal value diminishes as normalized planning resources are allocated.

We evaluate

$$\beta \in \{0.25, 0.50, 1.00, 2.00\}, \quad B \in \{0.05, 0.10, 0.20, 0.30\}.$$

Table 4. 95% month-block bootstrap percentile intervals for resource allocation performance.

Budget	Policy	DOT captured	Blended captured	DOT value	Oracle ratio	Top-3 share	Active units	Active carriers
0.05	Proposed PTO	[627508.9, 780543.2]	[599464.3, 743662.1]	[396661.3, 493397.4]	[0.999, 1.000]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 3.0]
0.05	DOT-only	[627810.3, 780543.2]	[599464.3, 743662.1]	[396851.8, 493397.4]	[1.000, 1.000]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 3.0]
0.05	BTS-only	[627473.9, 780543.2]	[599464.3, 743662.1]	[396639.2, 493397.4]	[0.992, 1.000]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 3.0]
0.05	Equal	[175506.2, 217185.6]	[170115.3, 209347.8]	[117190.8, 211845.3]	[0.401, 0.459]	[0.300, 0.300]	[240.0, 240.0]	[10.0, 10.0]
0.05	Delay-only	[616003.0, 769961.4]	[591414.1, 735463.1]	[389388.2, 486708.4]	[0.968, 1.000]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 3.0]
0.05	Cancellation-only	[576998.7, 737466.8]	[545981.6, 700661.1]	[364732.7, 466167.9]	[0.879, 0.978]	[0.917, 1.000]	[12.0, 12.0]	[2.0, 4.0]
0.05	Volume-only	[551748.9, 734994.2]	[533425.2, 703623.5]	[348771.8, 464605.0]	[0.823, 0.979]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 1.0]
0.05	Historical-average	[435006.7, 656433.0]	[422189.6, 630711.9]	[274976.7, 414944.8]	[0.634, 0.898]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 1.0]
0.05	Oracle-DOT	[627810.3, 780543.2]	[599464.3, 743662.1]	[396851.8, 493397.4]	[1.000, 1.000]	[1.000, 1.000]	[12.0, 12.0]	[1.0, 3.0]
0.10	Proposed PTO	[1094821.9, 1366542.5]	[1052220.7, 1304338.6]	[692059.4, 863819.6]	[0.996, 1.000]	[0.917, 1.000]	[24.0, 24.0]	[3.0, 4.0]
0.10	DOT-only	[1094832.9, 1366542.5]	[1052213.7, 1304338.6]	[692066.4, 863819.6]	[1.000, 1.000]	[0.917, 1.000]	[24.0, 24.0]	[3.0, 4.0]
0.10	BTS-only	[1093551.6, 1366542.5]	[1051835.0, 1304338.6]	[691256.5, 863819.6]	[0.992, 1.000]	[0.875, 1.000]	[24.0, 24.0]	[2.0, 4.0]
0.10	Equal	[351012.4, 434371.2]	[340230.6, 418695.7]	[334032.5, 413358.9]	[0.452, 0.495]	[0.300, 0.300]	[240.0, 240.0]	[10.0, 10.0]
0.10	Delay-only	[1084349.3, 1360001.6]	[1045275.5, 1299646.9]	[685439.5, 859684.9]	[0.976, 1.000]	[0.875, 1.000]	[24.0, 24.0]	[3.0, 4.0]
0.10	Cancellation-only	[1021638.6, 1317094.5]	[969838.1, 1250910.1]	[645798.8, 832562.5]	[0.883, 0.981]	[0.792, 1.000]	[24.0, 24.0]	[3.0, 5.0]
0.10	Volume-only	[1017803.5, 1280990.8]	[982330.0, 1227759.2]	[643374.5, 809740.6]	[0.880, 0.966]	[1.000, 1.000]	[24.0, 24.0]	[1.0, 2.0]
0.10	Historical-average	[949375.0, 1264218.6]	[917187.0, 1212304.6]	[600119.5, 799138.6]	[0.797, 0.956]	[1.000, 1.000]	[24.0, 24.0]	[1.0, 2.0]
0.10	Oracle-DOT	[1094832.9, 1366542.5]	[1052213.7, 1304338.6]	[692066.4, 863819.6]	[1.000, 1.000]	[0.917, 1.000]	[24.0, 24.0]	[3.0, 4.0]
0.20	Proposed PTO	[1830583.3, 2318278.4]	[1760659.5, 2219828.0]	[1157149.4, 1465431.4]	[0.999, 1.000]	[0.812, 0.917]	[48.0, 48.0]	[4.0, 4.0]
0.20	DOT-only	[1830586.0, 2318278.4]	[1760659.5, 2219828.0]	[1157151.1, 1465431.4]	[1.000, 1.000]	[0.812, 0.917]	[48.0, 48.0]	[4.0, 4.0]
0.20	BTS-only	[1829884.5, 2316370.2]	[1760650.9, 2219281.2]	[1156707.6, 1464225.2]	[0.997, 1.000]	[0.812, 0.896]	[48.0, 48.0]	[4.0, 4.0]
0.20	Equal	[702024.8, 868742.4]	[680461.2, 837391.3]	[636277.5, 787381.4]	[0.522, 0.557]	[0.300, 0.300]	[240.0, 240.0]	[10.0, 10.0]
0.20	Delay-only	[1828702.0, 2303868.9]	[1758654.8, 2212291.8]	[1155960.1, 1456322.9]	[0.987, 1.000]	[0.812, 0.896]	[48.0, 48.0]	[4.0, 4.0]
0.20	Cancellation-only	[1741536.8, 2226586.2]	[1674652.8, 2122393.8]	[1100861.2, 1407470.9]	[0.931, 0.980]	[0.792, 0.917]	[48.0, 48.0]	[4.0, 6.0]
0.20	Volume-only	[1645160.5, 2127576.6]	[1589474.8, 2042500.0]	[1039939.8, 1344884.9]	[0.873, 0.935]	[0.938, 1.000]	[48.0, 48.0]	[3.0, 4.0]
0.20	Historical-average	[1549395.9, 2026383.5]	[1498693.6, 1948314.7]	[979405.0, 1280918.7]	[0.787, 0.909]	[0.875, 0.958]	[48.0, 48.0]	[4.0, 4.0]
0.20	Oracle-DOT	[1830586.0, 2318278.4]	[1760659.5, 2219828.0]	[1157151.1, 1465431.4]	[1.000, 1.000]	[0.812, 0.917]	[48.0, 48.0]	[4.0, 4.0]
0.30	Proposed PTO	[2417945.4, 3039750.3]	[2340200.0, 2921037.7]	[1528433.0, 1921488.6]	[1.000, 1.000]	[0.778, 0.833]	[72.0, 72.0]	[4.0, 4.0]
0.30	DOT-only	[2417945.4, 3039756.6]	[2340200.0, 2920979.8]	[1528433.0, 1921492.7]	[1.000, 1.000]	[0.778, 0.833]	[72.0, 72.0]	[4.0, 4.0]
0.30	BTS-only	[2417945.4, 3039750.3]	[2340200.0, 2921037.7]	[1528433.0, 1921488.6]	[0.999, 1.000]	[0.778, 0.833]	[72.0, 72.0]	[4.0, 4.0]
0.30	Equal	[1053037.2, 1303113.6]	[1020691.8, 1256087.0]	[909760.2, 1125811.0]	[0.578, 0.600]	[0.300, 0.300]	[240.0, 240.0]	[10.0, 10.0]
0.30	Delay-only	[2416952.3, 3038826.0]	[2339081.6, 2920911.7]	[1527805.2, 1920904.4]	[0.997, 1.000]	[0.778, 0.833]	[72.0, 72.0]	[4.0, 4.0]
0.30	Cancellation-only	[2241967.0, 2840295.2]	[2160690.1, 2718385.8]	[1417193.5, 1795409.0]	[0.910, 0.961]	[0.722, 0.833]	[72.0, 72.0]	[6.0, 8.0]
0.30	Volume-only	[2270091.3, 2908323.2]	[2200426.0, 2802377.4]	[1434971.4, 1838410.9]	[0.920, 0.973]	[0.806, 0.875]	[72.0, 72.0]	[4.0, 4.0]
0.30	Historical-average	[2197490.2, 2795555.8]	[2129162.6, 2692025.6]	[1389078.7, 1767128.3]	[0.874, 0.944]	[0.819, 0.903]	[72.0, 72.0]	[4.0, 4.0]
0.30	Oracle-DOT	[2417945.4, 3039756.6]	[2340200.0, 2920979.8]	[1528433.0, 1921492.7]	[1.000, 1.000]	[0.778, 0.833]	[72.0, 72.0]	[4.0, 4.0]

Uncertainty is summarized using 1,000 month-block bootstrap replications with seed 20260824. As an implementation audit, the $\beta = 1$ row reproduces the corrected Experiment 08 Forecast-PTO oracle ratios to within 4.54×10^{-7} .

Table 5 reports the leakage-controlled forecast-based sensitivity results. Across all 16 (β, B) cells, Forecast-PTO retains between 0.979467 and 0.997584 of hindsight-oracle value. The allocation geometry changes materially across the grid: mean active carriers range from 1.08 in the most concentrated cases to 4.92 at $\beta = 2$ and the 30% budget. We therefore use the grid only as a structural robustness analysis and do not select or calibrate β by ex-post performance.

Table 5. Leakage-controlled Forecast-PTO sensitivity to the scenario-response parameter β . Entries report aggregate realized value relative to the hindsight oracle, with 95% month-block bootstrap percentile intervals.

β	5% budget	10% budget	20% budget	30% budget
0.25	0.9951 [0.9855, 1.0000]	0.9899 [0.9717, 1.0000]	0.9834 [0.9743, 0.9907]	0.9881 [0.9801, 0.9936]
0.50	0.9899 [0.9717, 1.0000]	0.9795 [0.9636, 0.9939]	0.9864 [0.9789, 0.9928]	0.9929 [0.9892, 0.9959]
1.00	0.9795 [0.9636, 0.9939]	0.9833 [0.9733, 0.9914]	0.9908 [0.9854, 0.9950]	0.9960 [0.9938, 0.9979]
2.00	0.9833 [0.9733, 0.9914]	0.9908 [0.9854, 0.9950]	0.9964 [0.9943, 0.9981]	0.9976 [0.9965, 0.9985]

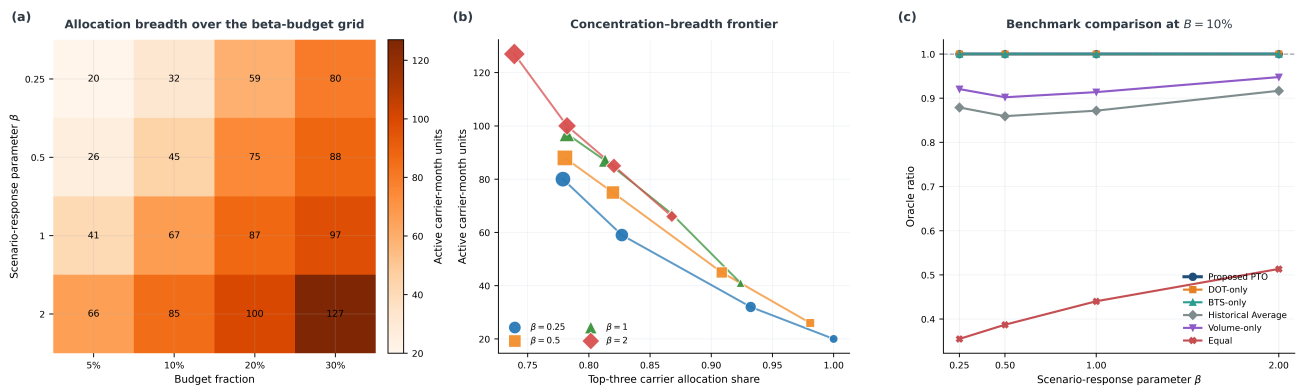


Figure 4. Corrected retrospective diminishing-return structural diagnostic under the mutually exclusive harm construction. (a) Active carrier-month units over the (β, B) grid, showing how allocation breadth changes under larger assumed β values and larger budgets. (b) Concentration–breadth frontier for the PTO policy, relating the top-three carrier allocation share to the number of active carrier-month units. (c) Oracle-relative benchmark performance at the 10% budget across the tested scenario-response values. This same-period analysis is used only to characterize allocation geometry and benchmark ordering; it is not evidence of prospective Forecast-PTO performance. The leakage-controlled forecast-based sensitivity analysis is reported separately in Figure 8.

In the corrected retrospective diagnostic, increasing β expands the active set substantially: active carrier-month units range from 20 at $(\beta, B) = (0.25, 0.05)$ to 127 at $(2, 0.30)$. Over the same grid, the top-three carrier allocation share falls from 1.000 to 0.739, documenting the expected tradeoff between concentration and allocation breadth. These quantities are descriptive properties of the assumed diminishing-return allocation geometry and are not interpreted as causal treatment effects.

Figure 8 visualizes the leakage-controlled forecast-based sensitivity results. Figure 4 has a different and deliberately narrower role: it is a corrected retrospective structural diagnostic. Rather than repeating near-one oracle-ratio surfaces, it emphasizes the aspects of the same-period allocation that actually vary with β .

6.6. Omega-blending sensitivity

We finally use the same-period DOT–BTS blend only as a retrospective diagnostic of whether the baseline blending choice $\omega = 0.5$ is finely tuned. Because the corrected DOT and BTS exposure measures are linked reconstructions and are nearly identical after harmonizing their event definitions, near-oracle ratios are not interpreted as predictive performance.

Figure 5 reports decision sensitivity directly rather than displaying near-one ratios. Across $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$ and budget fractions $\{0.05, 0.10, 0.20, 0.30\}$, the maximum shortfall from the DOT hindsight oracle is 2.290 parts per million. Relative to the baseline $\omega = 0.5$ allocation, the maximum fraction of total resource mass that must be reallocated is only 0.086%. The result therefore documents decision redundancy with respect to the same-period blending weight; it does not establish ex-ante forecast value.

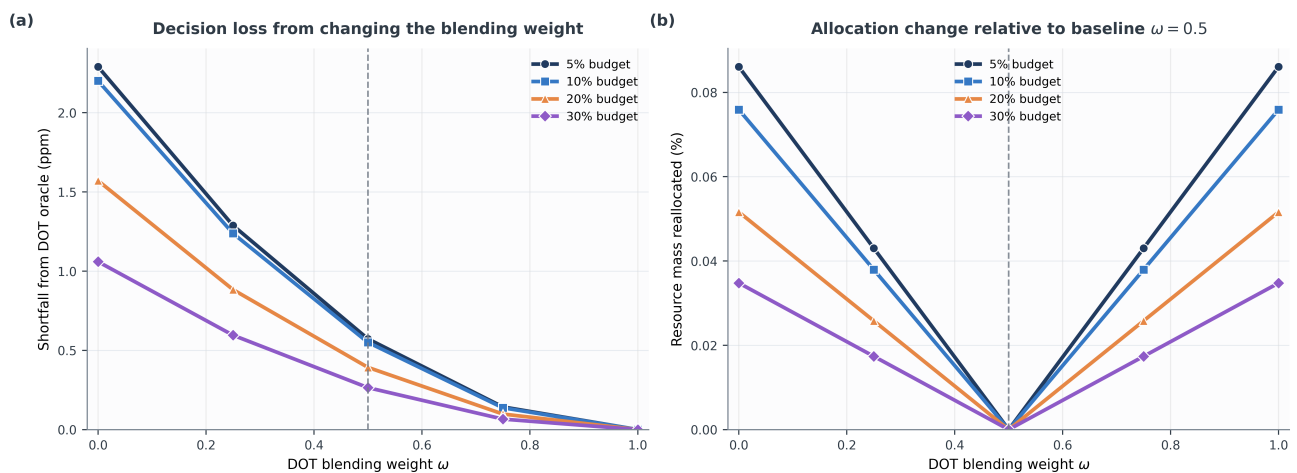


Figure 5. Retrospective sensitivity of the corrected PTO allocation to the DOT–BTS blending weight ω . (a) Shortfall from the DOT hindsight oracle in parts per million (ppm) as ω varies across budget levels. (b) Percentage of total resource mass reallocated relative to the baseline choice $\omega = 0.5$. Both quantities remain extremely small, documenting insensitivity of the retrospective allocation to the blending choice rather than prospective forecast performance.

6.7. Lagged-information and dominant-carrier robustness

The preceding same-period experiments are retained as retrospective structural diagnostics under a common feasible set. Because their near-oracle patterns can be partly driven by contemporaneous DOT information, we next conduct a one-month-ahead lagged-information robustness experiment. In this experiment, each allocation for month t uses only carrier-level information from month $t - 1$, while policy value is evaluated using current-month DOT exposure-weighted service harm. The resulting design separates the information used for allocation from the outcome used for evaluation.

The lagged-information experiment uses the main 2024–2025 carrier-month panel. After imposing the one-month lag restriction, the evaluation sample contains 230 carrier-month observations over 23 months. We compare lagged PTO, lagged DOT-only, lagged BTS-only, prior historical-average, lagged volume-only, equal allocation, and a current-month Oracle-DOT benchmark retained only as a hindsight upper bound. The proposed lagged PTO score is

$$S_{a,t}^{LagPTO} = 0.5E_{a,t-1}^{DOT} + 0.5E_{a,t-1}^{BTS}.$$

Thus, current-month DOT harm is not used in the lagged PTO allocation score.

The deterministic results show that lagged risk-based policies remain close to the current-month Oracle-DOT. At the 5% and 10% budget levels, several lagged policies coincide with the oracle because the same dominant high-harm carrier is selected in each month. We therefore do not interpret these low-budget oracle ratios as evidence of forecasting superiority. Rather, they indicate that service harm is highly concentrated and that the prior-month signals identify the same dominant priority carrier as the hindsight oracle. At the 20% and 30% budget levels, where the allocation expands beyond the single dominant carrier, the lagged PTO policy attains 0.967 and 0.995 of the current-month Oracle-DOT value, respectively, while equal allocation remains substantially lower.

Because the low-budget lagged result is strongly affected by the dominant carrier, we further conduct a conservative dominant-carrier exclusion check. The carrier most frequently selected by the low-budget Oracle-DOT allocation, American Airlines Network, is removed from the sample before re-solving the lagged allocation problem. This exclusion reduces the sample from 230 to 207 carrier-month observations and leaves 9 carriers in the evaluation set. The purpose is not to redefine the main empirical population, but to test whether the lagged PTO result is solely an artifact of one large carrier.

The dominant-carrier exclusion results further support this interpretation. Even after removing the dominant carrier, the lagged PTO policy attains 91.9%, 91.9%, 98.6%, and 99.9% of the Oracle-DOT value across the four budget levels. Equal allocation remains much lower, with oracle ratios of 0.265, 0.265, 0.193, and 0.427. These results show that the lagged-information finding is not driven solely by a single dominant carrier. At the same time, the allocations remain concentrated among the top three carriers, reflecting the empirical concentration of exposure-weighted service harm in the airline panel.

This robustness exercise clarifies the interpretation of the main allocation results. The proposed PTO policy should not be read as a claim that prior-month information perfectly forecasts current-month service harm. Instead, the evidence shows that lagged DOT–BTS operational signals preserve enough priority information to generate high-quality resource allocations, even under a conservative exclusion of the dominant carrier. This directly addresses two potential concerns: dependence on contemporaneous DOT scores and dependence on a single large carrier.

The lagged-information experiment therefore provides an intermediate test between contemporaneous diagnostic allocation and a genuine forecast-driven decision experiment. It shows that simple prior-

period information retains substantial allocation-relevant persistence, but it does not establish that an explicit forecasting model necessarily improves on every persistence benchmark. The next subsection imposes a stricter information separation: the allocation is generated from an explicit forecast score, while the realized current-month DOT target is withheld until the evaluation stage.

Table 6. Dominant-carrier exclusion robustness for the one-month-ahead lagged-information allocation experiment.

Budget	Policy	DOT value	Oracle ratio	Active carriers
0.05	Oracle-DOT	264862.3	1.000	3
0.05	Lagged BTS-only	245335.5	0.926	3
0.05	Lagged PTO	243412.6	0.919	3
0.05	Lagged volume-only	240161.7	0.907	1
0.05	Equal	70230.6	0.265	1
0.10	Oracle-DOT	433746.0	1.000	3
0.10	Lagged BTS-only	401768.4	0.926	3
0.10	Lagged PTO	398619.3	0.919	3
0.10	Lagged volume-only	393295.6	0.907	1
0.10	Equal	115011.6	0.265	1
0.20	Oracle-DOT	817255.1	1.000	3
0.20	Lagged PTO	805462.9	0.986	3
0.20	Lagged BTS-only	803606.9	0.983	3
0.20	Lagged DOT-only	794666.5	0.972	3
0.20	Equal	157434.8	0.193	2
0.30	Oracle-DOT	1160995.0	1.000	3
0.30	Lagged PTO	1159365.0	0.999	3
0.30	Lagged BTS-only	1158830.0	0.998	3
0.30	Lagged DOT-only	1157018.0	0.997	3
0.30	Equal	496238.3	0.427	3

*Notes: American Airlines Network, the carrier selected by the low-budget Oracle-DOT allocation, is excluded before rerunning the lagged-information experiment. Allocation scores use only prior-month carrier-level information, whereas policy value is evaluated using current-month DOT exposure-weighted service harm. Oracle-DOT is a hindsight upper benchmark.

6.8. Leakage-controlled forecast-to-prescription performance

The primary decision experiment allocates resources month by month using only information available before the evaluation month. Realized current-month DOT harm is reserved for ex-post value calculation and the hindsight Oracle-DOT normalization. The deterministic Forecast-PTO result is reproduced exactly in the downstream robust experiment at $\kappa = 0$ (maximum numerical difference 4.14×10^{-11}), providing an additional implementation audit.

Table 7. Forecast-PTO decision performance under the mutually exclusive harm target.

Budget	Oracle ratio	Relative regret	95% CI	Mean active carriers
0.05	0.9795	0.0205	[0.9636, 0.9939]	2.50
0.10	0.9833	0.0167	[0.9733, 0.9914]	3.67
0.20	0.9908	0.0092	[0.9854, 0.9950]	4.00
0.30	0.9960	0.0040	[0.9938, 0.9979]	4.00

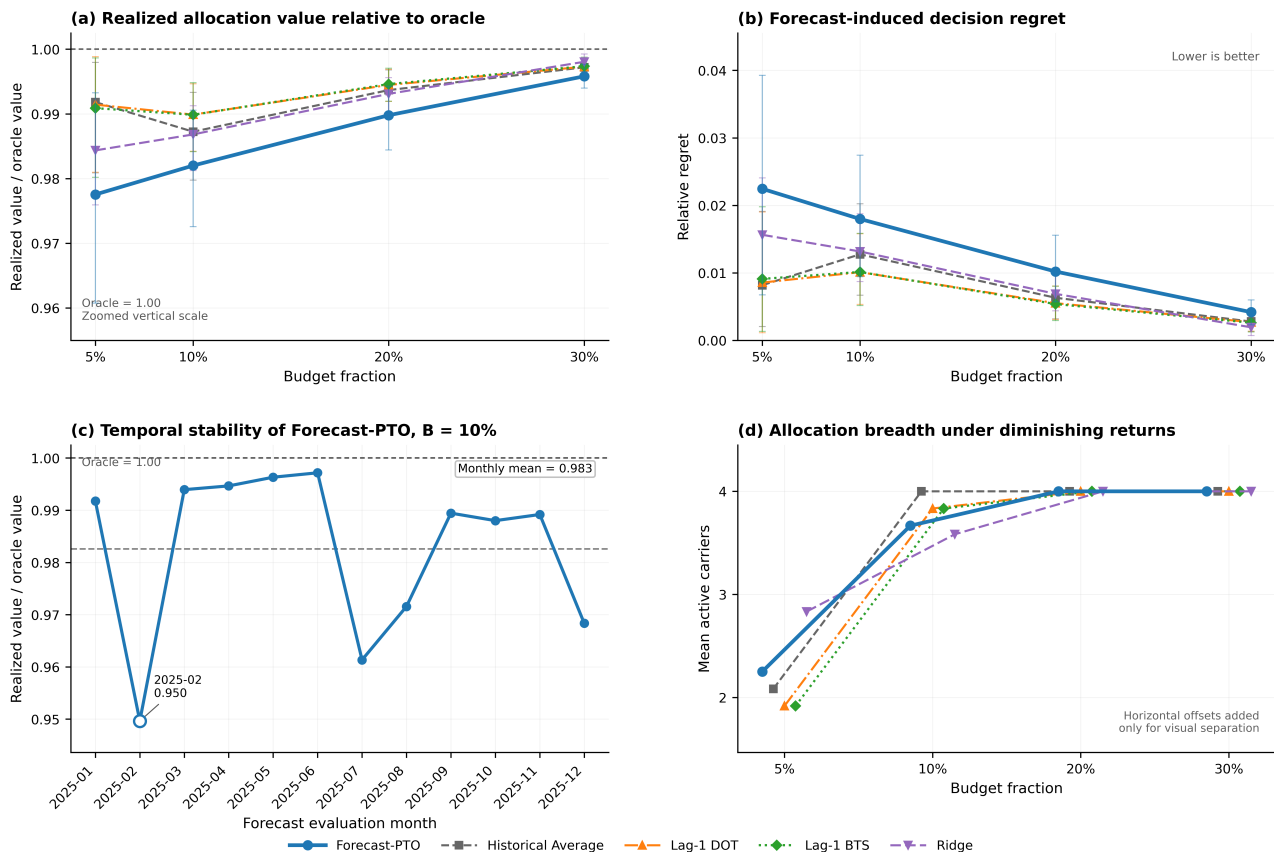


Figure 6. Corrected forecast-to-prescription decision results. (a) Oracle ratio across budget fractions for implementable forecast policies. (b) Decision regret relative to the hindsight oracle. (c) Month-by-month Forecast-PTO oracle ratio at the 10% budget. (d) Mean active carriers under the diminishing-return allocation. Realized current-month DOT harm is used only for ex-post evaluation and oracle normalization.

Forecast-PTO retains 0.9795, 0.9833, 0.9908, and 0.9960 of hindsight-oracle value at budget fractions 0.05, 0.10, 0.20, and 0.30, with relative regrets 0.0205, 0.0167, 0.0092, and 0.0040. The corresponding month-block bootstrap 95% intervals are [0.9636, 0.9939], [0.9733, 0.9914], [0.9854, 0.9950], and [0.9938, 0.9979]. The result is therefore robust to the corrected cancellation accounting. Forecast-PTO does not dominate every simple benchmark at every low budget: Lag-1 DOT, Lag-1 BTS, and Historical Average are slightly stronger in some sparse-budget cases. The relevant result is that a leakage-safe

ex-ante forecast retains most of the attainable decision value, not that a single forecast universally dominates persistence rules.

6.9. *Heterogeneous relative-cost and capacity robustness*

The heterogeneous experiment anchors relative cost and capacity scenarios only to pre-2025 operating-volume information. Under the primary Moderate scenario, Forecast-PTO attains oracle ratios 0.9815, 0.9825, 0.9918, and 0.9952, with 95% intervals [0.9671, 0.9945], [0.9737, 0.9900], [0.9875, 0.9954], and [0.9928, 0.9973]. Under Strong heterogeneity, the ratios are 0.9839, 0.9834, 0.9921, and 0.9954. The secondary linear diagnostic shows why cost adjustment matters when the feasible set becomes heterogeneous: at the 20% and 30% budgets, cost-aware priority improves the oracle ratio over cost-blind priority by 0.0183 and 0.0208 in the Moderate scenario and by 0.0316 and 0.0442 in the Strong scenario. These are normalized scenario costs and capacities, not observed airline accounting quantities.

Table 8. Forecast-PTO robustness under heterogeneous relative costs and capacity limits.

Scenario	B=0.05	B=0.10	B=0.20	B=0.30
Homogeneous	0.9795	0.9833	0.9908	0.9960
Mild	0.9800	0.9824	0.9914	0.9951
Moderate	0.9815	0.9825	0.9918	0.9952
Strong	0.9839	0.9834	0.9921	0.9954

6.10. *Uncertainty-aware robust allocation*

Forecast uncertainty is summarized by 1.4826 times the cross-model median absolute deviation across seven ex-ante forecasting rules. The uncertainty radius contains no realized current-month outcome. At the primary $\kappa = 1$ setting, robust allocation attains oracle ratios 0.9674, 0.9771, 0.9865, and 0.9939 across the four budgets. Relative to deterministic Forecast-PTO, the changes are -0.0121, -0.0062, -0.0043, and -0.0022. The paired bootstrap 95% intervals for those changes are [-0.0273, 0.0014], [-0.0139, 0.0002], [-0.0087, -0.0005], and [-0.0047, -0.0001]. Thus, the robust formulation is a transparent uncertainty-protection mechanism, but in this sample it carries a measurable ex-post decision cost; robustification should not be described as a uniform performance improvement.

6.11. *Carrier-airport-month finer-resolution robustness*

Experiment 12 addresses the spatial-aggregation concern by testing whether the forecast-to-allocation logic remains meaningful on carrier-airport-month observations. The finer-resolution DOT and BTS layers continue to track one another closely. Across 5,807 matched carrier-airport-month observations, the DOT–BTS arrival-delay loss-rate correlation is 0.9860 (Spearman 0.9876), and the corresponding exposure correlation is 0.9967 (Spearman 0.9986). Forecasting also remains viable at the finer unit. On the locked 2025 test panel of 2,878 carrier-airport-month observations, the Airport-GBR model attains MAE 87.12, RMSE 252.40, NMAE 0.2159, Pearson correlation 0.9659, and Spearman correlation 0.9744. Its mean monthly Spearman rank correlation is 0.9839, marginally above Lag-1 BTS (0.9831) and Lag-1 DOT (0.9830), and its mean top-10% overlap with the realized highest-harm set is 0.9318.

Thus, the finer-resolution experiment again shows that forecast point accuracy and allocation-relevant ranking fidelity are related but not identical.

Figure 7 reports the decision results. Under the primary threshold requiring at least one reported arrival, the Airport-GBR policy attains oracle ratios of 0.9821, 0.9920, 0.9953, and 0.9953 at 5%, 10%, 20%, and 30% budgets, corresponding to mean relative regrets of 0.0186, 0.0087, 0.0051, and 0.0050. Equal allocation remains far below all harm-aware policies, with oracle ratios of 0.1877, 0.2589, 0.3835, and 0.4909. At the most restrictive 5% budget, Lag-1 DOT is slightly better than Airport-GBR (0.9831 versus 0.9821), again showing that persistence-based rules remain serious benchmarks. At 10% and above, Airport-GBR remains among the strongest evaluated implementable rules in realized oracle ratio.

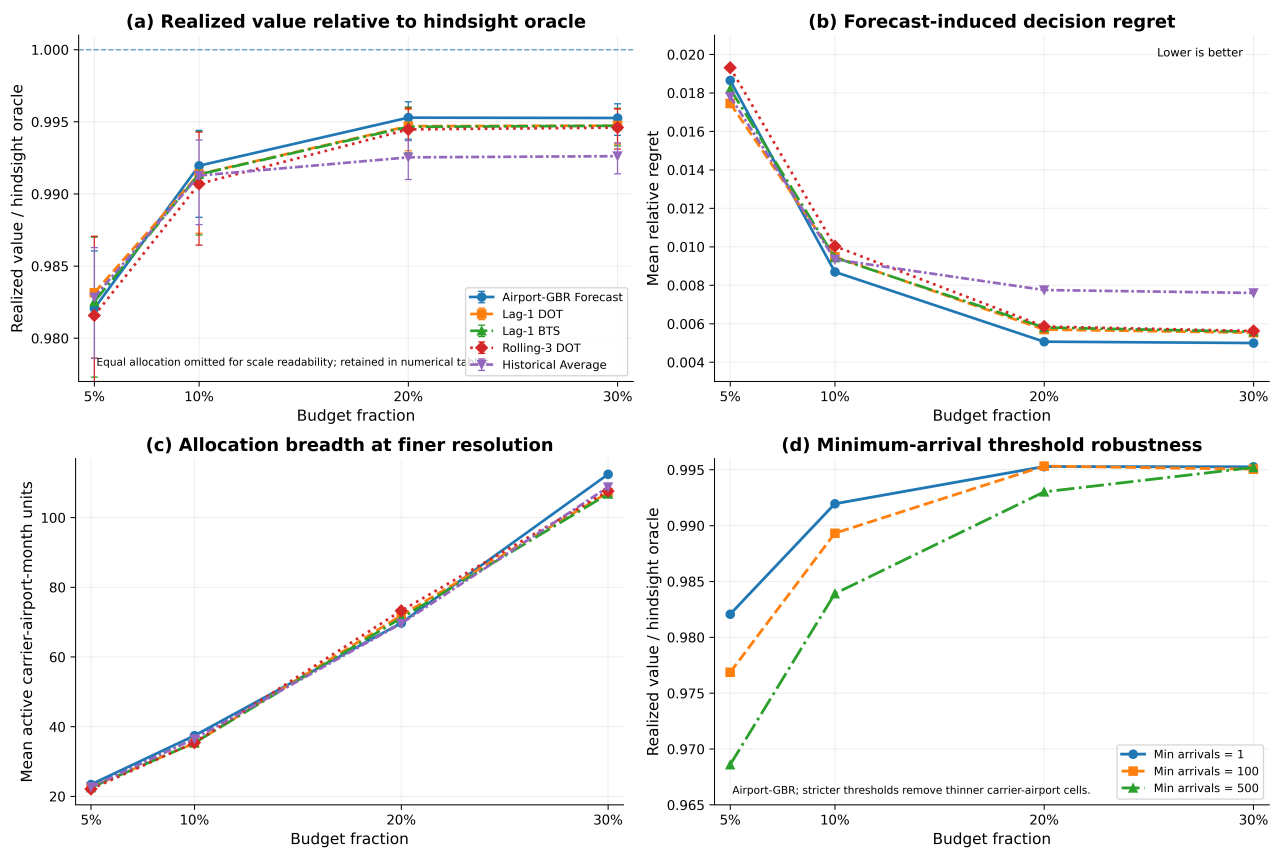


Figure 7. Carrier-airport-month finer-resolution robustness in Experiment 12. (a) reports realized policy value relative to the hindsight oracle across budget fractions for the main implementable policies; error bars are 95% month-block bootstrap percentile intervals. (b) reports the corresponding mean relative regret. (c) reports allocation breadth, measured as the mean number of active carrier-airport-month units. (d) reports threshold robustness for the Airport-GBR forecast policy when the minimum reported-arrival requirement is increased from 1 to 100 and 500. Equal allocation is omitted from (a) for scale readability but remains in the manuscript-ready tables. The experiment is arrival-side and does not impute airport-specific cancellation harm.

Figure 7(c) shows that the finer-resolution allocation broadens as the budget expands: the mean number of active carrier-airport-month units under Airport-GBR rises from 23.42 at 5% to 37.42, 69.75, and 112.42 at 10%, 20%, and 30%. Figure 7(d) shows that the result is robust to excluding thinner airport cells. When the minimum-arrival threshold increases from 1 to 100 and 500, the Airport-GBR oracle ratios become 0.9769/0.9686 at 5%, 0.9893/0.9839 at 10%, 0.9953/0.9930 at 20%, and 0.9950/0.9952 at 30%. Thus, stricter support thresholds reduce low-budget performance modestly but do not overturn the main conclusion.

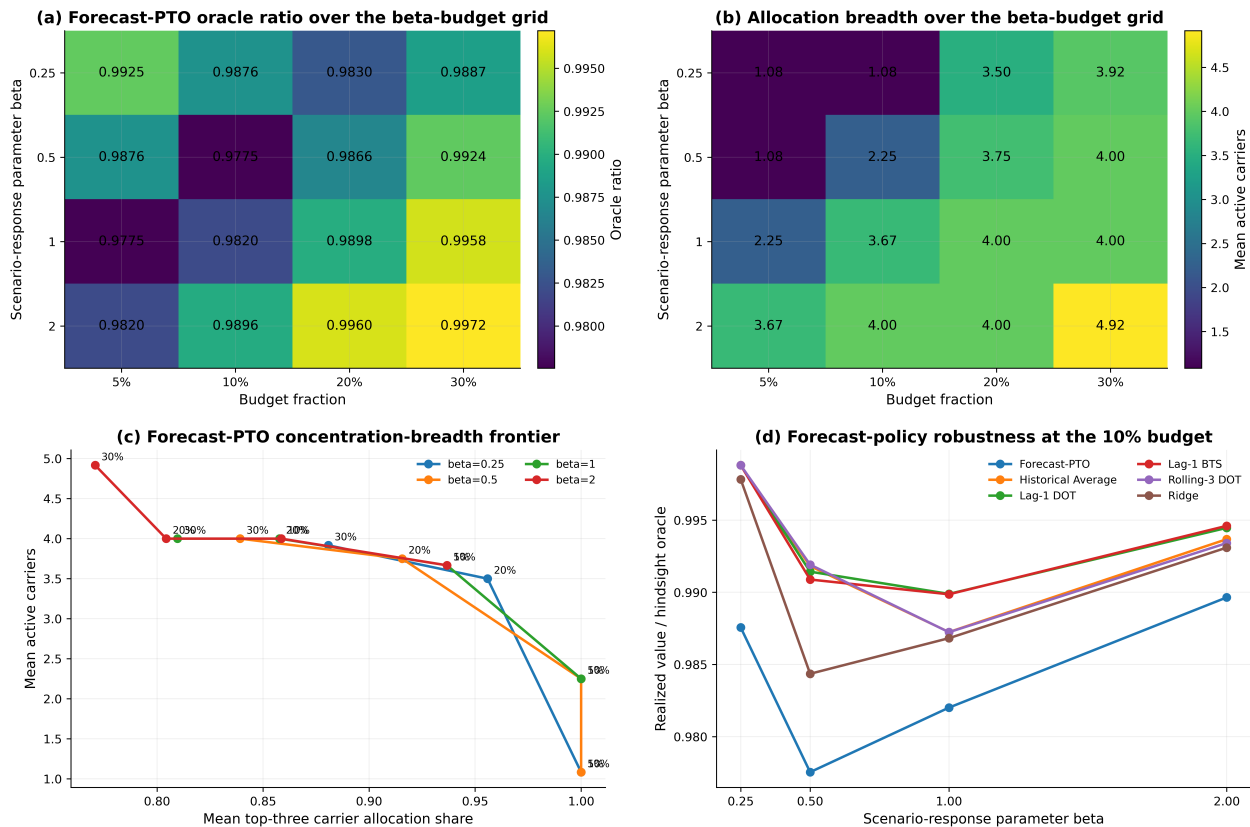


Figure 8. Leakage-controlled forecast-based response-curvature sensitivity in Experiment 04R.

(a) Forecast-PTO oracle ratio over the (β, B) grid. (b) Allocation breadth measured by mean active carriers. (c) Concentration–breadth frontier using mean top-three carrier allocation share and mean active carriers. (d) Implementable forecast-policy oracle ratios across β at the 10% budget. The allocation scores use only pre-month information; realized month- t DOT harm is reserved for ex-post evaluation and hindsight-oracle normalization. The parameter β is a scenario-response curvature parameter, not an estimated causal treatment effect.

This finer-resolution experiment does not transform the study into a full airline-network recovery model, because aircraft rotations, route spillovers, gate constraints, and airport-specific cancellation outcomes remain outside the observed target. Nevertheless, the finer-resolution analysis shows that even at the carrier-airport-month level, a leakage-controlled forecast-based prioritization rule continues to recover the large majority of hindsight-oracle value.

Forecast-based response-curvature sensitivity. Using the locked Experiment 08 panel, we vary $\beta \in \{0.25, 0.5, 1, 2\}$ over all four budget fractions. Across the 16 β -budget cells, Forecast-PTO oracle ratios range from 0.979467 to 0.997584. At $\beta = 1$, the four values reproduce the corrected Experiment 08 ratios to within 4.54×10^{-7} . Thus, the prospective decision conclusion is not an artifact of a single response-curvature setting, although allocation breadth changes materially with β . The parameter is used only as a structural scenario quantity and is not interpreted causally.

6.12. Summary of empirical findings

The empirical evidence supports five bounded conclusions. First, correcting the cancellation double count materially changes high-cancellation harm levels but preserves the broad priority ordering. Second, DOT-BTS near-equality is best understood as cross-layer reconstruction consistency, not independent-source validation. Third, leakage-controlled Forecast-PTO retains 97.95%–99.60% of hindsight-oracle value under the corrected target. Fourth, this decision result persists under heterogeneous feasible sets and response-curvature scenarios, while explicit robustification has a measurable decision cost. Fifth, the arrival-side finer-resolution experiment shows that most of the attainable priority value can also be preserved at carrier-airport-month resolution.

7. Managerial implications, limitations, and future research

The results support a decision-oriented interpretation of airline service-quality analytics. Prediction is useful only to the extent that it changes a resource-allocation decision under the information actually available at the decision time. In the mutually exclusive-harm analysis, Gradient Boosting has the lowest point forecast error, while simple lagged rules remain highly competitive in within-month ranking. Forecast-PTO nevertheless retains 97.95%–99.60% of hindsight-oracle value across the tested budgets. Additional forecasting complexity should therefore be justified by downstream decision value rather than by accuracy alone.

Budget scarcity magnifies the cost of ranking error. Forecast-PTO regret falls from 2.05% at the 5% budget to 0.40% at the 30% budget as the allocation broadens from an average of 2.50 to four active carriers. The diminishing-return model therefore changes not only the ordering of targets but also the breadth of the intervention portfolio.

Heterogeneous implementation conditions do not overturn the prospective result. Under the Moderate scenario, Forecast-PTO retains 98.15%–99.52% of oracle value. Cost adjustment becomes particularly relevant under Strong heterogeneity, where the cost-aware linear diagnostic improves the oracle ratio by 3.16 and 4.42 percentage points at the 20% and 30% budgets. These are normalized scenario quantities rather than observed intervention costs.

Explicit uncertainty protection should also be viewed as a design choice rather than costless insurance. At $\kappa = 1$, the robust policy remains high-value but performs slightly below deterministic Forecast-PTO, and the ex-post cost of robustification becomes statistically visible at the larger budgets. The ambiguity set is based on ex-ante model disagreement and has no nominal probabilistic coverage guarantee.

The corrected DOT-BTS results require especially careful interpretation. DOT-published measures and BTS flight-level records describe the same underlying U.S. airline operations. After harmonizing the event decomposition, their correlation is nearly one. This is useful as an audit of reconstruction and aggregation consistency, including after volume controls, but it is not evidence that two independent

data-generating systems validate one another or that contemporaneous blending creates incremental information. Prospective decision value comes from the forecast-to-prescription experiment, where the realized month- t target is withheld from the score.

Several limitations remain. Public data do not identify the causal effect of a specific staffing, scheduling, recovery, or customer-service intervention. The allocation variable is therefore a normalized planning resource, and the harm score is a priority measure rather than an estimated treatment effect. Likewise, β is a scenario-response curvature parameter and λ is a cancellation-severity weight, not a causal valuation. The heterogeneous cost/capacity profiles are pre-period-anchored scenarios rather than observed dollar costs or engineering limits. The primary prospective evaluation spans 12 months and 120 carrier-month observations, so longer histories and unusual network regimes remain important extensions.

The carrier-airport-month experiment reduces, but does not eliminate, spatial aggregation concerns. It is arrival-side only and does not model aircraft rotations, passenger connections, gate conflicts, cancellation recovery, or network spillovers. Future research can combine richer internal airline data with causal or quasi-experimental intervention estimates, longer forecast-error histories, route-network optimization, and stochastic or distributionally robust formulations calibrated to observed predictive and intervention uncertainty.

8. Conclusion

This study develops an auditable forecast-to-optimization framework for airline service-quality prioritization under delay and cancellation risk. The central question is not simply whether disruption-related harm can be measured or forecast, but whether information available before a decision period can be translated into a budget-feasible allocation that retains value when evaluated against outcomes observed later.

The service-harm construction ensures that cancellations are not counted once inside the non-on-time complement and again through the cancellation penalty. The corrected DOT measure separates non-cancellation service loss from cancellations, and the BTS reconstruction uses the parallel mutually exclusive categories of delayed operated flights, diversions, and cancellations. All carrier-month experiments that depend on this target were rerun from the corrected data pipeline.

The main prospective result is stable after this correction. On the locked 2025 panel, Forecast-PTO attains 97.95%, 98.33%, 99.08%, and 99.60% of hindsight-oracle value at budget fractions 5%, 10%, 20%, and 30%, respectively. The result persists under heterogeneous relative costs and capacities and across response-curvature scenarios. An uncertainty-aware robust extension remains high-value but shows that conservatism can reduce realized ex-post value, so robustness has an observable decision cost rather than being a free improvement.

The analysis also clarifies the role of DOT and BTS data. Their nearly perfect alignment under the corrected decomposition reflects linked reporting layers of the same airline operations. The alignment and volume-control exercises are therefore reconstruction-consistency audits, not independent validation and not evidence that contemporaneous data blending creates incremental predictive information. The strongest empirical evidence instead comes from the explicit information separation in the forecast-to-prescription experiment: current-month realized DOT harm is excluded from the decision score and used only for ex-post evaluation and hindsight-oracle normalization.

The framework remains intentionally bounded. It identifies priority under a specified response scenario; it does not estimate the causal effectiveness of a particular intervention. Within those limits, the study provides a reproducible example of how measurement, forecasting, optimization, uncertainty, and hindsight evaluation can be separated so that predictive-to-prescriptive claims remain auditable.

Author contributions

Jinho Cha: conceptualization, methodology, mathematical formulation, optimization model development, data curation and preprocessing, software, formal analysis, computational experiments, validation, visualization, interpretation of results, and writing—original draft and review/editing; Eun-Young Kwon: conceptualization, methodology review, interpretation of findings, manuscript structure, supervision, writing—review and editing, and corresponding-author responsibilities. Both authors reviewed and approved the submitted manuscript.

Use of Generative-AI tools declaration

Generative artificial intelligence tools were used only as limited assistive tools for language editing, formatting, and code-development/debugging support during manuscript preparation and revision. All data sources, model specifications, mathematical derivations, computational results, interpretations, and final scientific judgments were independently checked and approved by the authors. The authors take full responsibility for the content of this article.

Conflict of interest

The authors declare no conflicts of interest in this paper.

References

1. C. Barnhart, P. Belobaba, A. R. Odoni, Applications of operations research in the air transport industry, *Transp. Sci.*, **37** (2003), 368–391. <https://doi.org/10.1287/trsc.37.4.368.23276>
2. R. Britto, M. Dresner, A. Voltes, The impact of flight delays on passenger demand and societal welfare, *Transp. Res. Part E Logist. Transp. Rev.*, **48** (2012), 460–469. <https://doi.org/10.1016/j.tre.2011.10.009>
3. S. AhmadBeygi, A. Cohn, Y. Guan, P. Belobaba, Analysis of the potential for delay propagation in passenger airline networks, *J. Air Transp. Manag.*, **14** (2008), 221–236. <https://doi.org/10.1016/j.jairtraman.2008.04.010>
4. P. Fleurquin, J. J. Ramasco, V. M. Eguíluz, Systemic delay propagation in the US airport network, *Sci. Rep.*, **3** (2013), 1159. <https://doi.org/10.1038/srep01159>
5. C. Li, J. Mao, L. Li, J. Wu, L. Zhang, J. Zhu, Z. Pan, Flight delay propagation modeling: Data, methods, and future opportunities, *Transp. Res. Part E Logist. Transp. Rev.*, **185** (2024), 103525. <https://doi.org/10.1016/j.tre.2024.103525>

6. N. A. Ribeiro, J. Tay, W. Ng, S. Birolini, Delay predictive analytics for airport capacity management, *Transp. Res. Part C Emerg. Technol.*, **171** (2025), 104947. <https://doi.org/10.1016/j.trc.2024.104947>
7. F. Cao, T. Tang, Y. Gao, O. Michler, M. Schultz, Predicting flight arrival times with deep learning: A strategy for minimizing potential conflicts in gate assignment, *Transp. Res. Part C Emerg. Technol.*, **169** (2024), 104866. <https://doi.org/10.1016/j.trc.2024.104866>
8. Y. Pang, P. Zhao, J. Hu, Y. Liu, Machine learning-enhanced aircraft landing scheduling under uncertainties, *Transp. Res. Part C Emerg. Technol.*, **158** (2024), 104444. <https://doi.org/10.1016/j.trc.2023.104444>
9. J. M. Rosenberger, E. L. Johnson, G. L. Nemhauser, Rerouting aircraft for airline recovery, *Transp. Sci.*, **37** (2003), 408–421. <https://doi.org/10.1287/trsc.37.4.408.23271>
10. J. D. Petersen, G. Sölveling, J.-P. Clarke, E. L. Johnson, S. Shebalov, An optimization approach to airline integrated recovery, *Transp. Sci.*, **46** (2012), 482–500. <https://doi.org/10.1287/trsc.1120.0414>
11. B. Tan, Z. Liang, C. Y. Chung, H. Dong, X. Yin, S. K. Rayeem, J. Zhu, Uncertainty-Aware Approaches Toward Managing Hydrogen-Powered Vehicles and Renewable Energy Units in Microgrid–Transportation Networks, *IEEE Trans. Transp. Electr.*, (2026), In press. <https://doi.org/10.1109/TTE.2026.3712731>
12. A. Sternberg, J. Soares, D. Carvalho, E. Ogasawara, A Review on flight delay prediction, arXiv preprint, arXiv:1703.06118, 2017. <https://doi.org/10.48550/arXiv.1703.06118>
13. L. Carvalho, A. Sternberg, L. Maia Gonçalves, A. B. Cruz, J. A. Soares, D. Brandão, et al., On the relevance of data science for flight delay research: A systematic review, *Transp. Rev.*, **41** (2021), 499–528. <https://doi.org/10.1080/01441647.2020.1861123>
14. E. Esmaeilzadeh, S. Mokhtarimousavi, Machine learning approach for flight departure delay prediction and analysis, *Transp. Res. Rec.*, **2674** (2020), 145–159. <https://doi.org/10.1177/0361198120930014>
15. W. A. Khan, H.-L. Ma, S.-H. Chung, X. Wen, Hierarchical integrated machine learning model for predicting flight departure delays and duration in series, *Transp. Res. Part C Emerg. Technol.*, **129** (2021), 103225. <https://doi.org/10.1016/j.trc.2021.103225>
16. Q. Li, X. Guan, J. Liu, A CNN-LSTM framework for flight delay prediction, *Expert Syst. Appl.*, **227** (2023), 120287. <https://doi.org/10.1016/j.eswa.2023.120287>
17. M. Mamdouh, M. Ezzat, H. Hefny, Improving flight delays prediction by developing attention-based bidirectional LSTM network, *Expert Syst. Appl.*, **238** (2024), 121747. <https://doi.org/10.1016/j.eswa.2023.121747>
18. C. Li, X. Qi, Y. Yang, Z. Zeng, L. Zhang, J. Mao, FAST-CA: Fusion-based Adaptive Spatial–Temporal Learning with Coupled Attention for airport network delay propagation prediction, *Inf. Fusion*, **107** (2024), 102326. <https://doi.org/10.1016/j.inffus.2024.102326>
19. X. Shen, J. Chen, R. Yan, A spatial-temporal model for network-wide flight delay prediction based on federated learning, *Appl. Soft Comput.*, **154** (2024), 111380. <https://doi.org/10.1016/j.asoc.2024.111380>

20. J. Pineda-Jaramillo, C. Munoz, R. Mesa-Arango, C. Gonzalez-Calderon, A. Lange, Integrating multiple data sources for improved flight delay prediction using explainable machine learning, *Res. Transp. Bus. Manag.*, **56** (2024), 101161. <https://doi.org/10.1016/j.rtbm.2024.101161>
21. Y. Su, K. Xie, H. Wang, Z. Liang, W. A. Chaovalitwongse, P. M. Pardalos, Airline disruption management: A review of models and solution methods, *Engineering*, **7** (2021), 435–447. <https://doi.org/10.1016/j.eng.2020.08.021>
22. Y. Hu, S. Wang, S. Zhang, Z. Li, Review of optimization problems, models and methods for airline disruption management from 2010 to 2024, *Digit. Transp. Saf.*, **3** (2024), 246–263. <https://doi.org/10.48130/dts-0024-0022>
23. L. Cadarso, V. Vaze, Passenger-Centric Integrated Airline Schedule and Aircraft Recovery, *Transp. Sci.*, **57** (2023), 813–837. <https://doi.org/10.1287/trsc.2022.1174>
24. Q. Zhang, S.-H. Chung, H.-L. Ma, X. Sun, Robust aircraft maintenance routing with heterogeneous aircraft maintenance tasks, *Transp. Res. Part C Emerg. Technol.*, **160** (2024), 104518. <https://doi.org/10.1016/j.trc.2024.104518>
25. S. Birolini, A. Jacquillat, Day-ahead aircraft routing with data-driven primary delay predictions, *Eur. J. Oper. Res.*, **310** (2023), 379–396. <https://doi.org/10.1016/j.ejor.2023.02.035>
26. Y. Xu, Z. Shen, Z. Liu, X. Sun, Machine learning assisted dynamic copy generation approach for integrated aircraft and passenger recovery problem, *Transp. Res. Part C Emerg. Technol.*, **178** (2025), 105208. <https://doi.org/10.1016/j.trc.2025.105208>
27. N. Rashedi, N. Sankey, V. Vaze, K. Wei, A machine learning approach for solution space reduction in aircraft disruption recovery, *Eur. J. Oper. Res.*, **323** (2025), 297–308. <https://doi.org/10.1016/j.ejor.2024.11.025>
28. A. E. Hızır, C. Barnhart, V. Vaze, Large-scale airline crew recovery using mixed-integer optimization and supervised machine learning, *Transp. Sci.*, **60** (2026), 177–196. <https://doi.org/10.1287/trsc.2025.0105>
29. A. V. M. Oliveira, B. F. Oliveira, M. D. Vassallo, Airport service quality perception and flight delays: Examining the influence of psychosituational latent traits of respondents in passenger satisfaction surveys, *Res. Transp. Econ.*, **102** (2023), 101371. <https://doi.org/10.1016/j.retrec.2023.101371>
30. U.S. Department of Transportation, Air Travel Consumer Reports, Office of Aviation Consumer Protection, 2026. Available from: <https://www.transportation.gov/individuals/aviation-consumer-protection/air-travel-consumer-reports>.
31. D. Bertsimas, N. Kallus, From predictive to prescriptive analytics, *Manag. Sci.*, **66** (2020), 1025–1044. <https://doi.org/10.1287/mnsc.2018.3253>
32. A. N. Elmachtoub, P. Grigas, Smart “Predict, then Optimize”, *Manag. Sci.*, **68** (2022), 9–26. <https://doi.org/10.1287/mnsc.2020.3922>
33. B. Wilder, B. Dilkina, M. Tambe, Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization, *Proc. AAAI Conf. Artif. Intell.*, **33** (2019), 1658–1665. <https://doi.org/10.1609/aaai.v33i01.33011658>

34. N. Ho-Nguyen, F. Kılınç-Karzan, Risk guarantees for end-to-end prediction and optimization processes, *Manag. Sci.*, **68** (2022), 8680–8698. <https://doi.org/10.1287/mnsc.2022.4321>
35. O. El Balghiti, A. N. Elmachtoub, P. Grigas, A. Tewari, Generalization bounds in the predict-then-optimize framework, *Math. Oper. Res.*, **48** (2023), 2043–2065. <https://doi.org/10.1287/moor.2022.1330>
36. S. Lan, J.-P. Clarke, C. Barnhart, Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions, *Transp. Sci.*, **40** (2006), 15–28. <https://doi.org/10.1287/trsc.1050.0134>
37. C. Song, X. Ma, C. Ardizzone, J. Zhuang, The adverse impact of flight delays on passenger satisfaction: An innovative prediction model utilizing wide & deep learning, *J. Air Transp. Manag.*, **114** (2024), 102511. <https://doi.org/10.1016/j.jairtraman.2023.102511>
38. U.S. Bureau of Transportation Statistics, Marketing Carrier On-Time Performance Data, Beginning January 2018, TranStats Airline On-Time Performance Data, U.S. Department of Transportation, 2026. Available from: https://www.transtats.bts.gov/Tables.asp?QQ_VQ=EFD.



AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)