Airline-Driven Performance-Based Air Traffic Management: Game Theoretic Models and Multicriteria Evaluation

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Evans, Antony et al. “Airline-Driven Performance-Based Air Traffic Management: Game Theoretic Models and Multicriteria Evaluation.” Transportation Science 50,1 (February 2016): 180–203 © 2016 INFORMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1287/trsc.2014.0543">http://dx.doi.org/10.1287/trsc.2014.0543</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute for Operations Research and the Management Sciences (INFORMS)</td>
</tr>
<tr>
<td>Version</td>
<td>Author’s final manuscript</td>
</tr>
<tr>
<td>Accessed</td>
<td>Tue Mar 12 19:38:15 EDT 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/112157">http://hdl.handle.net/1721.1/112157</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
Airline-Driven Performance-Based Air Traffic Management: Game Theoretic Models and Multi-Criteria Evaluation

Antony Evans\textsuperscript{a}, Vikrant Vaze\textsuperscript{b} and Cynthia Barnhart\textsuperscript{c}

\textsuperscript{a}UCL Energy Institute, University College London
\textsuperscript{b}Thayer School of Engineering, Dartmouth College
\textsuperscript{c}Department of Civil and Environmental Engineering, Massachusetts Institute of Technology

Abstract

Defining Air Traffic Management as the tools, procedures and systems employed to ensure safe and efficient operation of air transportation systems, an important objective of future air traffic management systems is to support airline business objectives, subject to ensuring safety and security. Under the current model for designing air traffic management initiatives, the central authority overseeing and regulating air traffic management in a region makes trade-offs between specified performance criteria. The research presented in this paper aims instead to allow the airline community to set performance goals and thus make trade-offs between different performance criteria directly, before specific air traffic management strategies are determined. We propose several approaches for collecting inputs from airlines in a systematic way and for combining these airline inputs into implementable air traffic management initiatives. These include variants of averaging, voting and ranking mechanisms. We also propose multiple criteria for evaluating the effectiveness of each approach, including Pareto optimality, airline profitability, system optimality, equity, and truthfulness of airline inputs. We apply a game-theoretic approach to examine the potential for strategic (gaming) behavior by airlines. We offer a broad evaluation of each approach, first by providing some theoretical insights, and then by simulating each of the approaches for a generic system using Monte-Carlo methods, sampling values for input parameters from a wide range. We also provide an indication of how the approaches might perform in a real system by simulating ground delay programs at two airports in the New York City area. We first apply a simplified model that simulates the process of selecting only planned end times of a ground delay program, using Monte-Carlo methods. Next, we apply a more detailed model that simulates the process of selecting planned end times and reduced airport arrival rates. Finally, we characterize the effectiveness of each of the considered approaches on the proposed criteria and identify the most desirable approaches. We conclude that voting schemes, which score highly on all criteria (including airline profitability, system optimality and equity), represent the most promising approaches (among those considered) to elicit airline preferences, thereby allowing the central authority to design air traffic management initiatives that optimize system performance while respecting the objectives of airlines.

Keywords

Performance-based Air Traffic Management, Airline preferences, Game theory, Nash equilibrium.
1. Introduction

Air Navigation Service Providers (ANSPs), such as the Federal Aviation Administration (FAA) in the United States and EUROCONTROL in Europe, are the central authorities responsible for safe and efficient operation of our air transportation systems. In order to ensure these goals, ANSPs employ various tools, procedures and systems, which together are termed Air Traffic Management (ATM). ATM systems in the U.S. and Europe are currently poised for a major overhaul, under projects titled Next Generation Air Transportation System (NextGen) in the U.S., and Single European Sky ATM Research (SESAR) in Europe. An important objective of future ATM systems as envisioned by the FAA is supporting the business objectives of airlines, subject to ensuring safety and security (JPDO, 2007; FAA, 2011). In addition to safety and security, airlines value many different operational aspects of the air transportation system, such as capacity, efficiency, flexibility, predictability etc. Better availability of sufficient capacity in the various components of the system reduces or eliminates congestion related delays. Greater efficiency in resource utilization translates into reduced operating costs. Greater flexibility in scheduling operations enables airlines to make appropriate changes closer to departure times, as their needs evolve with time. Better predictability, which refers to the reliability of the system to deliver on planned performance, leads to more certainty about future operations, which in turn helps airlines plan better. Different airlines might value these different performance criteria differently.

An ANSP may support airlines’ business objectives by designing Traffic Management Initiatives (TMIs) in such a way as to maximize a single performance goal or some pre-defined combined measure based on multiple performance criteria, subject to ensuring safety and security. Here, performance goal refers to the quantified value, based on some defined metric, of a performance criterion. However, an ANSP cannot typically maximize all performance goals simultaneously, and must identify an appropriate trade-off between them. For example, consider a Ground Delay Program (GDP), a common TMI implemented by the FAA to control the flow of aircraft into an airport by delaying flights destined for that airport at their respective origin airports. A GDP is typically implemented for a period of time when increased aircraft spacing is considered necessary between landing aircraft, to ensure safety, and is often associated with adverse weather. However, weather forecasts are uncertain, so the point when conditions improve and additional spacing is no longer necessary is usually difficult to predict. Setting a GDP end time to be optimistically early maximizes the airport capacity and therefore throughput, because inbound flights are not delayed at their departure gates any longer than necessary. Therefore, no matter when conditions improve, and the airport capacity can be returned to the normal level, there are aircraft positioned to land. However, last minute extensions may have to be made to the GDP if the adverse weather continues longer than was forecast, potentially requiring airborne holding. So an early GDP end time would be at the expense of predictability (in addition to safety concerns and fuel costs associated with airborne holding). To maximize predictability, the GDP end time should be set conservatively late, allowing airlines to be confident that the GDP would not be extended. But this would be at the expense of throughput, as capacity might be underutilized if conditions were to improve earlier than the set GDP end time. There is therefore a trade-off between throughput and predictability.

Different airlines might have different preferences for prioritizing throughput over predictability. For example, for an airline operating a frequent shuttle service with low load factors, which allows easy
rebooking of passengers and easy reassignment of aircraft, some throughput reduction is not as detrimental as operating an unpredictable schedule. On the other hand, for an airline with lower frequency and higher load factors, for which delay recovery is difficult, high throughput may be preferred to predictability so that the airline does not have to cancel flights. The primary motivation for our research is to investigate various approaches for ANSPs to determine the trade-off between performance criteria, based on inputs (i.e., preferences) from airlines. We apply our research specifically to the case of GDPs.

In the existing literature, supporting airline preferences has typically been studied at the level of individual flight trajectories, through trajectory-based initiatives. In such initiatives, airlines are, for example, given authority to modify their own flight trajectories in a far- and mid-term time horizon to avoid an identified constraint, such as a region of airspace with high traffic or a region impacted by weather (e.g., Garcia-Chico et al., 2008). Alternatively, airlines are given the opportunity to provide the ANSP with multiple prioritized flight trajectories, individual flight priorities, or route priorities (e.g., Sheth and Gutierrez-Nolasco, 2008). In this paper, we consider accommodating airline preferences at the more aggregate level, shifting the focus from flight trajectories to overall system performance. To the best of the authors’ knowledge, ours is the first study that addresses the challenge of supporting airline preferences at a system level. Such a performance-based ATM system would be capable of making trade-offs between different performance criteria, such as capacity, efficiency, flexibility, predictability etc., at the system level, and could account for the system-level performance preferences of different airlines. The chosen system performance objectives would then serve as the basis for deciding on specific parameters, such as the length, scope and magnitude of TMIs, that include GDPs, Ground Stops, Miles-in-Trail (MIT) restrictions, traffic re-routes, etc.

The performance of the U.S. National Airspace System (NAS) is typically measured by the number of delayed flights and by the length, scope and magnitude of TMIs. The length of a GDP, Ground Stop, MIT restriction or re-route typically refers to the planned duration of the initiative. In the case of a GDP, Ground Stop, MIT restriction or re-route, scope refers to the subset of flights impacted by the initiative. For example, in the cases of a GDP or Ground Stop, scope refers to the set of flights delayed on departure as a consequence of the initiative. This set is comprised of all flights whose destination is the constrained airport and whose origins are within a specified maximum distance from the constrained airport. The magnitude of a GDP refers to the specified airport arrival rate (AAR) at the destination airport, which in the case of a Ground Stop is zero. The magnitude of a MIT restriction is the actual in-trail spacing required of the traffic, while the magnitude of a re-route can be considered to be how far from the planned flight trajectory the reroute takes the traffic. These values, however, do not represent the performance of the system per se, but are rather indicators of aspects of the system performance (FAA, 2011). They are also inadequate to describe differences in system-level preferences and requirements of different airlines. It is therefore important to identify what the performance criteria of airlines are, and to describe them in quantifiable terms. The International Civil Aviation Organization (ICAO, 2005) lists performance criteria, or the “expectations of the ATM community”, as follows: access and equity, capacity, cost-effectiveness, efficiency, environment, flexibility, global interoperability, participation by the ATM Community, predictability, safety, and security. In this paper, we focus directly on a subset of these performance criteria rather than dealing with indicators of aspects of performance, as has been done traditionally. The reader is referred to Liu and Hansen (2012) for an example of how performance goal vectors can be expressed as a function of these indicators of aspects.
of performance. A performance goal vector refers to a vector with individual components that are the values or goals for specific performance criteria, such as capacity, predictability etc. In this paper, we make use of the expressions of capacity and predictability developed by Liu and Hansen (2012) when analyzing specific TMIs in Section 5.

Given the differences in the valuation of these performance criteria by different airlines, an approach or a mechanism is needed to reconcile their competing preferences. In contrast, in the existing ATM system, the ANSP has sole responsibility for determining these trade-offs when designing TMIs. For example, the trade-off of throughput and predictability is determined in designing a GDP with the ANSP selecting the GDP end time, as described above, as well as scope and magnitude. The research presented in this paper aims instead to allow the airline community to influence TMI design by providing preferences in advance of the TMI design and implementation. In so doing, the ANSP can then design TMIs that capture airline preferences in the most effective way. A primary aim of this research is to design and assess various candidate mechanisms for this process, and demonstrate their applicability through general experiments and specific, real-world motivated case studies.

Consistent with the FAA objective of supporting airlines’ business objectives, a large amount of research has been focused on formulating and solving the problem of system optimality in air transportation (e.g., Odoni and Bianco, 1987; Bertsimas and Stock-Patterson, 1998, 2000; Lulli and Odoni, 2007). In effect, these studies describe ways in which airlines and the ANSP might attempt to maximize total airline profits and minimize system cost by trading off performance goals, without compromising safety and security. The ICAO performance criteria that are most likely to be traded in such a scenario are capacity, efficiency, predictability and flexibility. Because each airline may prefer a different trade-off, it becomes important to ensure equity in how each airline’s preferences are combined to set the final system-wide performance goals. It is noted that, as described by Bertsimas et al. (2011), the trade-off that minimizes system cost might not, in fact, be equitable, depending on how equity is defined. This is further complicated by the fact that when an airline requests certain performance goals, the request might not, in fact, represent the airline’s true preferences. In other words, an airline might not be truthful about its preferences, and might behave strategically, in effect gaming the system, requesting performance goals different from its true preference in order to draw the finally selected system performance goals closer to its desired outcome. This can have significant consequences for equity, because, while it might appear that a solution is equitable based on the submitted preferences, it could be far from it. Furthermore, if airlines game severely (requesting solutions far from their true preferences), the ANSP is provided with an inaccurate picture of the airlines’ preferences, and therefore of how well it is serving the airlines. Therefore, in this paper, in addition to airline profitability and system optimality, we also use equity and truthfulness to assess the effectiveness of any mechanism under consideration.

As mentioned earlier, due to safety and security concerns, not all of the relevant system-wide performance criteria, such as capacity, efficiency, predictability and flexibility, can be simultaneously maximized. This is especially true in cases of high traffic and/or adverse weather. Thus, there is a tradeoff between these different system-wide performance goals. An increase in one performance goal beyond a certain level necessarily requires a reduction in another. We define the trade space as the set of all combinations of
system-wide performance goals that are feasible, subject to ensuring safety and security. Note that the trade space as well as the performance goals are defined at the system level and all else being equal, each airline is assumed to prefer a higher value of a system performance criterion at least as much as a lower value of the same performance criterion. A subset of the boundary of the trade space is Pareto efficient, in that, for any point A in this Pareto efficient subset of boundary points, there does not exist another point B in the trade space such that the value of each performance criterion at point B is at least as high as the value of that performance criterion at point A and the value of at least one performance criterion at point B is strictly greater than the value of that performance criterion at point A. We define this subset of the boundary of the trade space as the Pareto frontier.

2. Contributions

In this paper, a number of contributions are made to performance-based ATM research:

- This is the first study that investigates various approaches to allow the airline community to set the system-level performance goals of the ATM system, and thus make trade-offs between performance criteria directly. We do this using a rigorous game-theoretic approach, which identifies the potential for gaming by airlines.

- We propose several approaches for collecting inputs from airlines in a systematic way and for combining these airline inputs into implementable TMIs. These include an approach which takes a weighted average of the airline-preferred performance goals; an approach that pushes the weighted average of the airline-preferred performance goals out to the Pareto frontier; an approach that makes a weighted random choice of airline-preferred performance goals; an approach that allows airlines to rank preferred performance goals; and an approach in which airlines vote on preferred performance goals.

- We propose multiple criteria for evaluating the effectiveness of each approach to performance-based ATM, including Pareto optimality, airline profitability, system optimality, equity, and truthfulness of airline preferences.

- By first performing a theoretical analysis and then simulating each of the approaches for a generic system using Monte-Carlo methods (sampling values for input parameters from a wide range), we offer a broad evaluation of each approach to performance-based ATM.

- We also apply the approaches to more realistic cases in which we simulate GDPs at Newark Liberty International airport (EWR) and at LaGuardia airport (LGA), both in the New York City area. Two models are run: one simplified GDP model that simulates decisions regarding the planned GDP end time, and is applied using Monte-Carlo methods; and a second more detailed model that simulates decisions regarding both planned GDP end time and GDP magnitude, and is run for a single GDP case at each of EWR and LGA, respectively. The results of these simulations provide an indication of how the approaches would perform in a real system, and how the results differ from those for the generic experiments.
Finally, we characterize the effectiveness of each of the considered approaches on the proposed criteria and identify the most desirable approach accordingly. Taking a weighted average of the user preferred performance goal vectors, making a weighted random choice of the user preferred performance goal vectors, and voting on ANSP provided candidate performance goal vectors were all found to be reasonable candidates for practical implementation. However, the voting scheme shows particular promise, scoring highly on all criteria.

3. Framework

Figure 1 provides a high level view of the process to be investigated.

![Figure 1. Architecture for process by which performance goals are set](image)

The ultimate output of the process is a set of system-wide performance goals (upper right-hand box in Figure 1) that would be used by the ANSP to set specific TMIs. A possible form of these performance goals is described in Section 5. This set of performance goals could be for a single TMI within a single resource, such as the GDP at an airport described in the example in Section 1. Alternatively, the set of performance goals could be for multiple initiatives with multiple resources, or for the national airspace system as a whole. The process may start with an initial set of candidate performance goals suggested by the ANSP, or directly with inputs from each airline (set of boxes on upper left in Figure 1), which would be the performance goals preferred by each airline. A performance goal resolution process would then take the inputs from all airlines, confirm the feasibility of each input, and identify a set of system-wide performance goals by combining these different inputs in some way. The ANSP can then provide individual airline feedback, which would be a description of how the system-wide performance goals translate into changes in each specific airline’s operational plan, e.g. delays to individual flights scheduled to arrive at an airport under a GDP. This allows each airline to assess the impact of the system-wide performance goals on its operational performance, such as propagated delays, passenger and crew schedule disruptions, additional fuel burn, etc., through an airline assessment process. In this process, each airline considers the feedback and determines what adjustments to make to its input in order to influence the system-wide performance goals in such a way as to improve its own operational performance. This feedback loop can be executed several times until an equilibrium is reached, where no airline can unilaterally adjust its own inputs to produce a “better” set of system-wide performance goals.
performance goals (better being in terms of the performance objectives of that airline). This represents a pure strategy Nash equilibrium, a concept commonly used in game-theoretic literature to model situations with multiple, interacting, autonomous decision makers. We will use this concept to model the outcome of this iterative process.

The airline inputs and the performance goal resolution process may take a number of different forms. In this paper, different forms are studied in order to identify which has the best characteristics for setting system-wide performance goals. Each is described in detail in Sections 3.1 and 3.2 respectively. This is followed by a description of the metrics used to evaluate all approaches in Section 3.3.

3.1 Form of Airline Inputs

In this paper we analyze two forms of airline inputs:

1. A preferred performance goal vector – Each airline \( k \) simply specifies its preferred performance goal vector \( I_k \). This airline input is most applicable to a continuous trade space. In the example of a GDP described in Section 5.2, this input would take the form of each airline’s preferred trade-off between capacity and predictability, using some pre-defined metrics. This trade-off could be input in the form of the parameters of the GDP (such as GDP end time and AAR), or it could be input in the form of capacity and predictability metrics that are calculated from these parameters. For example, a metric describing capacity could be the ratio of expected throughput, given known uncertainty in the GDP end time, to the maximum throughput that would be possible with perfect information. Similarly a metric for predictability could be the ratio of the expected flight delay assuming the GDP were to end as planned, to the expected delay given known uncertainty in the GDP end time. These metrics are described in more detail in Section 3.3.

2. Votes or rankings on a set of candidate performance goal vectors – Each airline \( k \) specifies its preferences in the form of votes or rankings \( I_{kp} \) for each of the \( P \) candidate performance goal vectors \( G_p; p \in \{1, 2, ..., P\} \). This airline input is most applicable for a discrete trade space, in which only a finite set of candidate performance goal vectors are valid or are under consideration. In the GDP example described in Section 5.2, this input would take the form of either a vote or a ranking, from each airline, for each of the candidate vectors.

3.2 Performance Goal Resolution Approaches

The ANSP determines the system-wide performance goal vector \( (G^*) \) by combining all airline inputs according to a defined resolution approach. Five different approaches are analyzed in this paper. For each approach described below and each airline \( k \), the weights \( (w_k) \) are proportional to some non-decreasing function of the number of operations of \( k \) impacted by the initiative.

1. Taking a weighted average of all airline-preferred performance goal vectors – This is a simple and intuitive way of combining continuous-valued airline inputs. After each airline \( k \) has specified its preferred performance goal vector \( I_k \), a weighted average performance goal vector is calculated. Mathematically, this can be represented as:
\[ G^* = \sum_k w_k I_k \]  

In the case of the GDP example, if specific parameters of the GDP representing capacity and predictability are traded off, such as GDP end time \( T \) and AAR \( C \) (specifying the duration and magnitude of the GDP), this equation is represented by (2), with the inputs from each airline \( k \) represented by \( T_k \) and \( C_k \), and the system-wide solution represented by \( T^* \) and \( C^* \).

\[ \{T^*, C^*\} = \{ \sum_k w_k T_k, \sum_k w_k C_k \} \]

The iterative framework described above is applied, allowing airlines to modify their preferred performance goal inputs according to inputs from other airlines. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 5.

A sample outcome of such an iterative process is illustrated in Figure 2 for the case of two airlines and two performance criteria. In this example, the Pareto frontier is represented by an arc of a circle centered at the origin, and concave increasing quadratic payoff functions are assumed for each airline. Truthful solutions for both airlines are shown (as red and blue circles). A line of constant payoff for a particular airline and payoff value is defined as the set of all performance goal vectors corresponding to that payoff value for that airline. In Figure 2 it represents the range of trade-offs between the simulated performance goals that would result in identical payoff to the airline. As we move towards the upper right of the figure, increasing the values of both performance goals, airline payoff increases. Therefore, the truthful solutions at which each airline maximizes its payoff lie where the lines of constant payoff are tangent to the Pareto frontier, as shown. Airline inputs at equilibrium are shown respectively by red and blue \( \times \)s for the two airlines. These are the results of the aforementioned iterative process where each airline maximizes its payoff, given the input of the other airline. In the case of this sample instance, these points differ from the truthful solutions, because each airline is able to increase its payoff from the system-wide performance goal vector by gaming. As can be seen, each airline attempts to “pull” the system-wide solution towards its truthful solution. The final system-wide performance goal vector, calculated as a weighted average of goal vectors input by each airline, is also shown on the figure (black \( \times \)). This is an interior point, and therefore not Pareto optimal.
2. Taking a weighted average of all airline-preferred performance goal vectors (as in 1), and pushing the result out to the Pareto frontier – This approach is similar to that presented above but avoids the issue of the combined output vector being in the interior of the trade space, and therefore not Pareto-optimal. In the case of a convex trade space where this Pareto frontier is non-linear, a weighted average of airline-preferred inputs might not fall on the Pareto frontier itself. After each airline \( k \) has specified its preferred performance goal vector \( I_k \), a weighted average performance goal vector is calculated in the same way as in approach 1, above. This is then shifted out to the Pareto frontier. This shift can be done in many different ways. One reasonable approach is to do this in such a way as to maintain the same ratios of the values of each performance goal, generating a new system-wide performance goal vector that is on the Pareto frontier. This approach is also most applicable for a continuous trade space, for which the airline inputs are preferred performance goal vectors. Mathematically, this can be represented as:

\[
G^* \in \text{ParetoFrontier} \quad \text{such that:} \quad G^*_m / G^*_n = G'_m / G'_n \quad \text{for all} \quad m, n
\]

where

\[
G' = \sum_k w_k I_k
\]

Note that this point, \( G^* \), on the Pareto frontier will always be unique because if there are two such points with the same ratio of the values of each performance goal then one of the two points will have a lower value of each performance goal compared to that for the other point and hence the former will not lie on a Pareto frontier.

In the case of the GDP example described above, this equation would be represented as follows:

\[
T^* = f(C^*) \quad \text{such that:} \quad T^* / C^* = T' / C' \\
\text{where:} \quad [T', C'] = [\Sigma_k w_k T_k, \Sigma_k w_k C_k]
\]
As in approach 1, the iterative framework described above is applied, allowing airlines to modify their preferred performance goal inputs according to inputs from other airlines. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 5.

A sample outcome of such an iterative process is illustrated in Figure 3 for the case of two airlines and two performance criteria. Comparing Figure 3 to Figure 2, it is immediately clear that, by pushing the system-wide solution to the Pareto frontier, there is more gaming from both airlines. Airline inputs at equilibrium are shown respectively by red and blue ×s for the two airlines. These are the results of the aforementioned iterative process where each airline maximizes its payoff, given the input of the other airline. In Figure 3a, a case is shown in which the airline inputs (red and blue ×s) for both airlines fall at corner points. In Figure 3b, a case is shown in which only one of the airlines input is at a corner point. The other airline does not request a corner point because it is able to “pull” the system-wide solution to coincide with its truthful solution without moving to the corner point.

**Figure 3.** Sample result applying weighted average of all user preferred performance goal vectors, pushed to Pareto frontier: (a) both airlines request corner point solutions, (b) only one airline requests a corner point solution.

3. **Making a weighted random selection of the airline-preferred performance goal vectors** – The main motivation for considering this approach is that it eliminates strategic gaming behavior by the airlines, as we will see later in Section 6 and in Appendix B. After each airline $k$ has specified its preferred performance goal vector $I_k$, one of the airline-preferred performance goal vectors $I_k$ is randomly selected for $G^*$. The probability of each airline-preferred performance goal vector being selected is proportional to its weight $w_k$. This approach is applicable for both continuous and discrete trade spaces, for which the airline inputs are preferred performance goal vectors. Mathematically, this can be represented as:

$$G^* = I_j$$

(5)
where \( \Pr(j = k) = w_k \) for all \( k \)

In the case of the GDP example, this equation would be represented as follows:

\[
[T^*, C^*] = [T^j, C^j]
\]

where \( \Pr(j = k) = w_k \) for all \( k \)

A sample result is illustrated for the case of two airlines and two performance criteria in Figure 4. As shown, the airline inputs and truthful solutions coincide. The reason for this is that airlines are not incentivized in any way to submit a different, non-truthful input because, if their solution is not chosen, their input does not affect the chosen solution in any way. Thus, they are incentivized to submit truthful solutions irrespective of how the probabilities are defined to randomly choose one of the airline inputs. This also means that the iterative framework described above is not necessary, as there is no incentive for any airline to change its preferred performance goal inputs based on other airline inputs. The system-wide goal vector \((G^*)\) is also always Pareto optimal because each user input is Pareto optimal.

![Figure 4](image-url)

**Figure 4. Sample result applying weighted random selection of the user preferred performance goal vectors.**

One disadvantage of this approach is that it does not account for the fact that the payoff of any chosen solution may vary significantly across airlines. The chosen solution may therefore have highly disproportionate impacts on each airline. A solution may exist that has the lowest overall impact on all airlines, but is not the most preferred solution for any of the airlines.

4. **Ranking the candidate performance goal vectors based on airline preferences** – After each airline \( k \) has specified its preferences for each of the \( P \) performance goal vectors \( G_p \) in the form of descending ranks \( I_{kp} \) from \( P \) to 1 (\( P \) being the rank of that airline’s most preferred performance goal vector, and 1 the rank of its least preferred performance goal vector), the combined rank for each vector is calculated as a weighted sum of individual ranks assigned by different airlines to that vector. The
performance goal vector with the greatest combined rank is assigned to be the system-wide performance goal vector $G^*$. This approach is most applicable for a discrete trade space, in which only a candidate set of performance goal vectors is valid. Mathematically, this can be represented as:

$$G^* = G_q$$

where $q$ is such that $R_q = \max ( R_1, R_2, ..., R_P )$

and $R_p = \sum_k w_k I_{kp}$ for all $p \in \{1, 2, ..., P\}$

In the case of the GDP example described above, this equation would be represented as follows:

$$[T^*, C^*] = [T_q, C_q]$$

where $q$ is such that $V_q = \max ( V_1, V_2, ..., V_P )$

and $V_p = \sum_k w_k I_{kp}$ for all $p \in \{1, 2, ..., P\}$

As in approaches 1 and 2, the iterative framework described above is applied, allowing airlines to modify their rankings according to rankings from other airlines. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 5.

Ranking is intuitive to understand and use, and unlike the first three approaches discussed above allows the airline to input its relative preferences for all the candidate performance goal vectors, instead of just specifying its single most preferred performance goal vector. Furthermore, an airline does not require exact knowledge of the payoffs of each performance goal vector to submit an input to the ranking mechanism. Instead airlines are only required to have a good idea of their comparative preference of one vector over the others. However, ranking is not devoid of drawbacks; the most significant is that the approach frequently does not converge, nor is a solution necessarily unique. Convergence is not guaranteed because airline rankings can alternate between two different rankings from iteration to iteration. In the cases run in this paper, this is a significant problem, with as little as 8% of runs converging (in the simplified GDP case at LGA). Convergence is highly dependent on the input parameters, and therefore different cases perform very differently (in contrast to the LGA case, 73% of runs converged in the simplified GDP case at EWR). In contrast, while convergence is not guaranteed for any of the other approaches (with the exception of taking a weighted random choice of user preferred performance goal vectors), convergence is significantly better under these other approaches than under ranking, as shown in Section 5.

5. Voting on the candidate performance goal vectors based on airline preferences – After each airline $k$ has specified its preferences for each of the $P$ performance goal vectors $G_p$ in the form of $I_{kp}$ votes, the weighted sum of votes is calculated. This approach differs from ranking in that airlines can apply varying numbers of votes to each performance goal vector, according to their preferences, instead of just rank order. The total number of votes that can be assigned by an airline across different performance goal vectors is the same for each airline. The performance goal vector with the highest weighted sum of votes is assigned to be the system-wide performance goal vector $G^*$. This approach is most applicable for a discrete trade space, in which only a candidate set of performance goal vectors is valid. Mathematically, the approach can be represented as:

$$G^* = G_q$$

where $q$ is such that $V_q = \max ( V_1, V_2, ..., V_P )$
and \[ V_p = \sum_k w_k I_{kp} \text{ for all } p \in \{1, 2, ..., P\} \]

In the case of the GDP example described above, this equation would be represented as follows:

\[ [T^*, C^*] = [T_q, C_q] \]

where \( q \) is such that \[ V_q = \max(V_1, V_2, ..., V_P) \]

and \[ V_p = \sum_k w_k I_{kp} \text{ for all } p \in \{1, 2, ..., P\} \]

A very general voting framework is considered, in which each airline has a fixed maximum number of votes that it can distribute across available options (i.e., a form of range voting). We set this fixed number to 100. The airline may assign all its votes to its highest preference, or may distribute the votes across multiple options. The value of each airline’s vote in determining the system-wide performance goal vector is proportional to that airline’s weight. As in approaches 1, 2 and 4, the iterative framework described above is applied, allowing airlines to modify their votes according to what other airlines have voted. The process is continued for a set number of iterations, or until convergence to an equilibrium, as described in Section 5.

In the voting framework simulated, airlines are not required to allocate all their votes at any time, and can increase their votes from iteration to iteration. Only integer votes are considered. In order to ensure convergence, an airline is not permitted to reduce its vote for any candidate vector between iterations. An airline can only increase its vote, or maintain it at the same level.

As with ranking, voting allows airlines to input their relative preferences for all the candidate performance goal vectors, instead of just specifying their single most preferred performance goal vector. Unlike ranking, however, voting allows airlines to apply different values to different preferred performance goal vectors, beyond simply providing the rank order.

### 3.3 Characterization of Mechanism Performance Metrics

In order to evaluate different approaches to combine airline preferences to set system-wide performance goals, it is important to compare how each approach performs relative to the goals of each airline. A number of metrics are defined for this purpose: Pareto optimality, airline profitability, system optimality, equity and truthfulness. These are described and defined below. For the experiments modeling the generic initiative and the simplified GDP, each of these metrics is calculated for each run of a Monte-Carlo simulation, described in Section 5. A simple average is taken across different runs to estimate the expected value of each metric. In all cases, the metrics are designed to vary from 0 to 1, with larger values being better.

1. **Pareto Optimality** – Defined as how close the final system-wide solution is, on average, to the Pareto frontier. This metric provides a general indication of whether or not the selected system performance goals make maximum use of the available resources. It is defined as follows:

\[ \text{ParetoOpt} = \frac{a}{b} \]  

where \( a \) and \( b \) are as defined in Figure 5a. \( a \) is the distance of the system-wide solution from the origin and \( b \) is the length of the vector from the origin to the Pareto-frontier, which passes through
the system-wide solution. The metric provides an indication of how far from the origin the system-wide solution is compared to how far it would be if on the Pareto frontier with the same ratio of the values of individual performance goals. The metric equals 1 when the system-wide solution is in fact Pareto optimal for every Monte-Carlo run.

![Figure 5](image)

**Figure 5. Parameters for defining metrics for (a) Pareto optimality, and (b) truthfulness.**

2. **Airline Profitability** – Defined as the normalized difference between each airline’s maximum payoff and its payoff applying the system-wide solution. Each airline’s maximum payoff is the payoff obtained if we maximize that airline’s payoff function over the trade space. The metric is averaged over all airlines, and provides an indication of how close each airline’s profit is, at the system-wide solution, to its maximum possible profit over the trade space. It is defined as follows.

\[
\text{AirlineOpt} = \frac{\sum_{k=1}^{K} \left( 1 - \frac{P_{M_k} - P^*_k}{\max(|P_{M_k}|,|P^*_k|)} \right)}{K}.
\tag{12}
\]

where \(P^*_k\) represents the payoff for airline \(k\) applying the system-wide solution \(G^*\), \(P^M_k\) the maximum payoff for airline \(k\), and \(K\) is the number of airlines. Note that the payoff can be negative, i.e., a cost. In order to ensure that the metric is meaningful (i.e. varying from 0 to 1) in this case, we define the denominator in equation (12) to be the larger of the absolute value of the maximum payoff and the absolute value of the payoff at the system-wide solution. We subtract the metric from 1 to ensure that the larger values of the metric are to be considered better.

3. **System Optimality** – Defined as the difference between the total payoff across all airlines for the system optimal solution and the total payoff across all airlines applying the system-wide solution, normalized by the system optimal total payoff. The system optimal total payoff is calculated by maximizing the sum of airline payoff functions over the trade space. This metric provides an indication of how closely the ANSP goal of maximum system “effectiveness” is achieved. It is defined as follows.

\[
\text{SystemOpt} = \frac{\text{System Optimal Total Payoff} - \text{System-wide Total Payoff}}{\text{System Optimal Total Payoff}}.
\]
\[
SysOpt = 1 - \frac{\sum_{k=1}^{K} p_{SysOpt_k} - \sum_{k=1}^{K} P^*_k}{\max(\{\sum_{k=1}^{K} p_{SysOpt_k} | \sum_{k=1}^{K} P^*_k\})}. 
\]

where \( P^*_k \) represents the payoff for airline \( k \) applying the system-wide solution \( G^* \), and \( p_{SysOpt_k} \) the payoff for airline \( k \) at the point of system optimality (maximum total payoff across all airlines). Again, we define the denominator to ensure that the metric has meaningful value (i.e. varying from 0 to 1) even when payoffs are negative (costs). Also, we subtract the metric from 1 to ensure that the larger values of the metric are considered to be better.

4. **Equity** – Equity or fairness in resource allocation problems, in which some scarce resources must be allocated among multiple players by a central decision maker, has been extensively studied in social sciences, welfare economics and engineering. However, because of the multiple interpretations of concepts of fairness, and the different characteristics of different problems, no single criterion is universally accepted. For the purposes of this paper, we use one of the most prominent concepts in the literature: the max-min concept of fairness. This is one of the two concepts that Bertsimas et al. (2011) consider most applicable to air transportation (the other is the proportional concept of fairness). The max-min concept of fairness is a generalization of Rawlsian justice (Rawls, 1971) and the Kalai-Smorodinsky solution to the two-player game (Kalai & Smorodinsky, 1975). It maximizes the minimum (normalized) utility level that all players derive. In the context of this work, we denote this as \( P^{Fair} \), the point of maximum minimum-payoff, or the “Fair” solution, which we calculate by solving a separate optimization problem in which the minimum payoff across all airlines is maximized over the trade space. If we define \( e_k \) as the normalized change in payoff for airline \( k \) at the system-wide solution, and \( f_k \) as the normalized change in payoff for airline \( k \) at the “Fair” solution, our equity metric is defined as the ratio of the minimum value of \((1 - e_k)\) across all airlines to the minimum value of \((1 - f_k)\) across all airlines, as shown in equation (14). Subtracting the normalized change in payoff from 1 ensures that: (a) the metric is higher for a more equitable (as defined by the max-min fairness concept) strategic solution than for a less equitable strategic solution; and (b) the maximum possible value of the metric is 1, which is consistent with the definitions all our other performance metrics.

\[
Equity = \frac{\min_k (1 - e_k)}{\min_k (1 - f_k)},
\]

where \( e_k = \frac{P^*_k - p_{M_k}}{\max(\{P^*_k | P^{Fair}_k\})} \) and \( f_k = \frac{P^*_k - p_{Fair}_k}{\max(\{P^*_k | P^{Fair}_k\})} \). 

\( P^*_k \) represents the payoff for airline \( k \) applying the system-wide solution \( G^* \), and \( P^{Fair}_k \) the payoff for airline \( k \) at the point of maximum minimum-payoff, that is, at the “Fair” solution. Again, our definition of the denominator ensures that the metric always takes values between 0 and 1.

5. **Truthfulness** – Defined as how close a solution submitted as an input by an airline (also known as the airline’s “strategic” solution) is, on average, to the true preference of the airline (the airline’s “truthful” solution). This “truthful” solution is the “maximum-payoff” solution referred to in metric 2 above. This provides an indication of the degree to which the airline is gaming the system. Truthfulness is not of value in itself, unlike the other metrics, but does provide an indication of whether the airline inputs are close to their true preferences. This is important because the larger the
extent of gaming, the less likely it is that a fair mechanism can be implemented. The metric is defined as follows.

\[
Truth = \frac{\sum_{k=1}^{K} \max \left(1 - \frac{c_k}{d_k}, 0\right)}{K}. \quad (15)
\]

where \(c_k\) and \(d_k\) are as defined in Figure 5b for airline \(k\), and \(K\) is the number of airlines. \(d_k\) is the distance of an airline’s truthful solution from the origin. \(c_k\) is the distance from the truthful solution to the strategic solution of an airline. Subtracting the ratio \(c_k/d_k\) from 1 ensures that: (a) our truthfulness metric is higher for a strategic solution closer to the true solution than for a strategic solution farther from the true solution; and (b) the maximum possible value of the metric is 1, which is consistent with the definitions of all the other performance metrics. It is noted that because \(c_k\) can be greater than \(d_k\), we take a maximum of the numerator with 0 to ensure that the metric remains in the range from 0 to 1. (Note that in almost 100% of the cases in our experiments, the ratio \(c_k/d_k\) is less than or equal to 1.)

4. Theoretical Insights

In this section, we provide some insights into the theoretical aspects of each approach. We attempt to evaluate the five approaches for performance goal resolution (as described in Section 3.2) based on five mechanism performance metrics, namely Pareto optimality, system optimality, airline profitability, equity, and truthfulness (as described in Section 3.3). Unless explicitly stated otherwise, our theoretical analysis in this section assumes concave non-decreasing payoff functions and a convex trade space.

First, we note that none of these five approaches, except for the Weighted Random Choice approach, is completely immune to manipulation by players. Voting and ranking approaches have a long and notorious history of results about their potential manipulability, starting with Arrow (1951) who famously stated that “when voters have three or more distinct alternatives, no voting system can convert the ranked preferences of individuals into a community-wide (complete and transitive) ranking while also meeting a certain set of criteria, namely: unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives.” This result and the subsequent body of research (notably including Gibbard 1973 and Satterthwaite 1975) shows that our voting and ranking approaches are not completely immune to manipulation, and we cannot guarantee their system optimality and truthfulness, in general.

For the Weighted Average approach, Appendix A provides some guarantees of pure strategy equilibrium existence, uniqueness, truthfulness, and the convergence of the best response dynamic. Most of these results are only applicable for the case of linear payoff functions, which is a special case of the concave payoff functions that we have assumed in this paper (described in Section 5). Appendix B states, and proves, three propositions related to the necessary and sufficient conditions for the truthfulness of the Weighted Random Choice approach and the Weighted Average approach. Propositions 2 and 3 in Appendix B prove that the truthfulness of the Weighted Average approach under quadratic payoff functions and arc-shaped and parabolic Pareto frontiers can be disproved unless some very restrictive conditions are met. A similarly restrictive result can be proved for the piecewise linear case. It involves many more cases than the first two, and as a result is both lengthy and relatively less informative. Hence we decided to exclude this proof from
this paper. Instead, we motivate the issues with the truthfulness of the Weighted Average approach for the piecewise linear Pareto frontier using the following simple example. Consider a case of two equally weighted airline players. If their respective truthful solutions lie at two different points on the same line segment of the Pareto frontier, then each will be incentivized to move its strategic solution away from other’s solution in order to ‘pull’ the resultant system-wide solution closer to the respective truthful solutions. So the points, on the same line segment of the Pareto frontier, that are on the side opposite to the other’s solution will be more attractive to the airline compared to it’s truthful solution. The Weighted Average Pushed to Pareto Frontier approach (approach 2) is typically even more prone to manipulation than the Weighted Average approach, as we will see in Section 5. The exact conditions for truthfulness are more difficult to prove in this case. Figure 3 provides some intuition towards this end. Finally, Proposition 1 in Appendix B proves that the Weighted Random Choice approach is always guaranteed to yield a truthful solution. However, it is easy to see that unless each player’s truthful solution is identical to the system optimal solution, the system-wide solution will not be system optimal.

Note that all of these aforementioned results focus purely on the truthfulness and system optimality properties in an absolute sense. These results do not eliminate the possibility of these approaches yielding a system-wide solution that is relatively close to the Pareto frontier, the system-optimal solution, the most profitable solution, the most equitable solution, and/or the truthful solution. In the next section, we investigate these issues in detail. Through Monte-Carlo simulation, we evaluate and compare the five approaches in terms of their relative closeness to these idealized solution points. Here, the closeness is as defined by the five metrics in Section 3.3.

5. Computational Experimental Setup

In order to gain a better understanding of how effective the approaches described in Section 3 are in setting system-wide performance goals, each approach is simulated, first for a generic TMI, and then for a specific TMI, a Ground Delay Program (GDP), at each of EWR and LGA airports. This allows us to derive general results, which cover various types of initiatives, and specific results, which provide an indication of likely results for a real-world case. In order to simplify the analysis, trade-offs are simulated between only two performance criteria, for a small number of airlines (between 3 and 4). This small number of airlines is reasonable given that the number of airlines with greater than 5% of operations at EWR and LGA, two of the busiest airports in the U.S., is 3 and 4, respectively (FAA, 2012). The forms of the trade space, Pareto frontier and airline payoff functions are described for each initiative in the following sections, followed by a description of the simulation methodology.

5.1 Generic Traffic Management Initiative

Trade Space and Pareto Frontier

For a generic TMI, we define a convex trade space, with the Pareto frontier defined by one of three alternative functions: an arc, a parabola, and a piecewise-linear function. The functional form of each of these is shown below as a function of $G_1$ and $G_2$, performance goals for two performance criteria (e.g., capacity and predictability). The alternative forms of the Pareto frontier, illustrated in Figure 6, are:
1. An arc: \[ G_1^2 + G_2^2 = 1 \] (16)

2. A parabola: \[ G_2 = aG_1^2 + bG_1 + c, \] (17)
   where \( a = -1/(G_{1TP} - 1)^2; \) \( b = -2aG_{1TP}; \) and \( c = -a - b. \)

\( G_{1TP} \) represents the value of performance goal 1 at the turning point of the parabola, sampled from a uniform distribution between 0 and 1. Defining the parabola in this way results in a Pareto frontier that more closely resembles the Pareto frontiers in the more realistic GDP scenarios described below. Note that only the right half of the parabola forms the Pareto frontier. The upper portion from the turning point to the \( G_2 \) axis is formed by a horizontal straight line, as shown in Figure 6b.

3. A piecewise-linear function: \[ G_2 = -m_1 G_1 + 1, \] and \[ G_2 = -m_2 G_1 + m_2 \] (18) (19)

\( m_1 \) is sampled from a uniform distribution between 0 and 1, while \( m_2 \) is sampled from a uniform distribution between 1 and \( \infty \). In our experiments, the Pareto frontier is described by only two lines, as shown in Figure 6c. In theory, any number of lines can be used.

![Figure 6. Alternative Pareto frontiers describing the trade space for the generic TMI: a) an arc, b) a parabola, and c) a piecewise-linear function.](image)

**Airline Payoff Functions**

The airline payoff functions are defined by concave increasing functions in these performance goals, rather than linear functions, because of the small buffers and redundancies built into airline schedules to reduce the impact of performance goal reductions of smaller magnitudes. Beyond a certain threshold, decreases in, e.g., capacity, lead to faster than linear increases in passenger re-accommodation costs, crew delay and reserve crew costs, airline recovery costs, etc. For the generic TMI, the payoff function for airline \( k \) is defined as the sum of quadratic functions of each of the two performance goals \( G_1 \) and \( G_2 \), as follows.

\[ P_k = \sum_g (a_{g,k}(G_g^*)^2 + b_{g,k}(G_g^*) + c_{g,k}), \quad \forall k \] (20)

with \( a_{g,k} \leq 0 \) and \( b_{g,k} > -2a_{g,k} \).
The linear case is a special case of this function setting all $a_{g,k}$ values to 0. Note that the additive constant $c_{g,k}$ is later dropped from this expression without any loss of generality because it is inconsequential to any of the analysis. The parameters defining each airline’s payoff function, $a_{g,k}$ and $b_{g,k}$, are sampled from uniform distributions from -1 to 0 (in the case of $a_{g,k}$) and from 0 to 2 (in the case of $b_{g,k}$, ensuring that $b_{g,k} > -2a_{g,k}$ so that the payoff is non-decreasing in each performance goal). It is noted that payoff functions do not in fact need to be additive functions of $G^*_g$, and may include coupling between different performance goals. The simpler additive function is, however, retained for this paper.

5.2 Ground Delay Program

Trade Space and Pareto Frontier

For the specific TMI, we consider a GDP under capacity uncertainty, in which we trade-off capacity and predictability. As described earlier, a GDP typically has three decision variables: duration, scope and magnitude. We apply two different models of a GDP. In the first, a simplified, computationally efficient model is applied in which we consider only duration, while in the second, we apply a more detailed model and consider both GDP duration and magnitude. In all cases we ignore the impact of GDP modification in response to updated information. The first model is run using Monte-Carlo methods, making use of the expressions for the expected values of capacity and predictability metrics developed by Liu and Hansen (2012). The second model is run for two specific GDP scenarios, one at each of EWR and LGA respectively, making use of expressions for the capacity and predictability metrics described in Appendix C. In the absence of closed form expressions for their expected values, we resort to numerical integration. A detailed Monte-Carlo simulation of the second model is not within the scope of this paper, but is considered a useful next step in this research. The second model should be treated as an example of how our modeling framework is easily extendable to more complex forms of ATM initiatives and its results further validate our main conclusions, as shown later in Section 5.

For the simplified model of the GDP, we utilize metrics for capacity and predictability derived for a single airport by Liu and Hansen (2012), assuming a constant scheduled arrival demand rate, $\lambda$. When the GDP is initiated, the AAR is reduced from a known constant high level, $C_H$, which is assumed to be greater than $\lambda$, to a known constant low level, $C_L$, which is lower than $\lambda$. The planned duration of the GDP is $T$, at which time the AAR is expected to return to $C_H$. However, due to errors in prediction, the AAR may return to $C_H$ at a different time, $\tau$. When the GDP is initiated, $T$ is set but $\tau$ is unknown, and assumed to be uniformly distributed between $t_{min}$ and $t_{max}$. Conceptually, if $T$ is set close to $t_{min}$, $\tau$ is likely to be larger than $T$, and the GDP ends late. In this case, capacity will be more fully utilized but there will be less predictability. Alternatively, if $T$ is set close to $t_{max}$, then $\tau$ is likely to be smaller than $T$, and the GDP ends early. In this case, capacity will be underutilized and unnecessary delay will result. However, the delay is predictable. Thus, for this specific TMI, the only input from the airline is $T$, the planned duration of the GDP. Mathematical representations for capacity utilization and predictability, as developed by Liu and Hansen (2012), are presented below.
Capacity Utilization, $\alpha_c$, is defined as the ratio of realized throughput, from the beginning of the GDP until the time when there is no more delay, to the maximum possible throughput with perfect information, were the airlines able to take advantage of the increase in AAR at time $\tau$. $\alpha_c$ varies from 0 to 1, and is shown by Liu and Hansen (2012) to have the expected value shown below, which we use to define our metric for the performance goal of capacity $G_c$:

$$G_c = E[\alpha_c] = \frac{t_{max} - T}{t_{max} - t_{min}} + \frac{a}{t_{max} - t_{min}} \cdot \log \left( \frac{B + cT}{b + cT} \right),$$

where $a = \lambda \frac{C_H}{C_H - \lambda} T$, $b = C_H \frac{C_H - C_L}{C_H - \lambda} T$, and $c = C_L - C_H$.

Predictability, $\alpha_p$, is defined as the ratio of expected flight delay, assuming the GDP ends at the planned time $T$, to the total realized delay, i.e., the delay actually incurred given the early or late increase in AAR at $\tau$. Again, $\alpha_p$ varies from 0 to 1 (given that we ignore GDP modifications in response to updated information), and is shown by Liu and Hansen (2012) to have the expected value shown below, which we use to define our metric for the performance goal of predictability $G_p$:

$$G_p = E[\alpha_p] = \frac{1}{t_{max} - t_{min}} \cdot \left( -\frac{\tau^2}{t_{max}} + 2T - t_{min} \right).$$

For the more detailed model of a GDP, we calculate expected values of the metrics for capacity and predictability through numerical integration for a single airport using a similar approach to that used by Liu and Hansen (2012), but allowing for a varying scheduled arrival demand rate $\lambda(t)$, and planned and actual values of GDP AAR equal to $C_L$ and $C_H$, respectively. When the GDP is initiated, a planned AAR, $C_L$, is specified, which is lower than $\lambda$. This represents the expected AAR for the duration of the GDP. As in the simplified model, the planned duration of the GDP is $T$, at which time the airport capacity is expected to return to the known $C_H$, which is greater than $\lambda$. Due to errors in prediction, the actual AAR ($C_l$) for the GDP may be different than that planned. Similarly, the AAR returns to $C_H$ at a time $\tau$, which may be different than that planned. When the GDP is initiated, $C_L$ and $T$ are set but $C_l$ and $\tau$ are unknown. They are assumed to be uniformly distributed between $C_{L,min}$ and $C_{L,max}$, and $t_{min}$ and $t_{max}$, respectively. Conceptually, if $C_L$ is set close to $C_{L,min}$, $C_l$ is likely to be larger than $C_L$, and the rate at which aircraft arrive at the airport will be lower than the available capacity. In this case, capacity will be underutilized and unnecessary delay will result. However, the delay is predictable. Alternatively, if $C_L$ is set close to $C_{L,max}$, then $C_l$ is likely to be smaller than $C_L$, and arriving aircraft will be required to hold because they will arrive at a faster rate than the airport can accommodate. In this case, the available capacity will be more fully utilized but there will be less predictability. As in the simplified model, if $T$ is set close to $t_{min}$, $\tau$ is likely to be larger than $T$, and the GDP ends late. In this case, capacity will be more fully utilized but there will be less predictability. Alternatively, if $T$ is set close to $t_{max}$, then $\tau$ is likely to be smaller than $T$, and the GDP ends early. In this case, capacity will be underutilized but predictability will be high.

As in the simplified model, Capacity Utilization, $\alpha_c$, is defined as the ratio of realized throughput, from the beginning of the GDP until the time when there is no more delay, to the maximum possible throughput with perfect information, were the airlines able to take advantage of the actual GDP AAR $C_l$, and the increase in AAR at time $\tau$. The value of $\alpha_c$ varies from 0 to 1. Predictability, $\alpha_p$, is defined as the ratio of expected flight delay, assuming the planned GDP AAR $C_L$ and the planned GDP end time $T$, to the total realized delay, i.e., the delay actually incurred given the actual GDP AAR $C_l$ and the early or late increase in AAR at $\tau$. Again,
the value of \( \alpha_p \) varies from 0 to 1 (given that we ignore GDP modifications in response to updated information). Expressions for \( \alpha_c \) and \( \alpha_p \) are presented in Appendix C. Based on these expressions, the expected value of the performance goals for capacity \( G_c \) and predictability \( G_p \), can be calculated through numerical integration assuming uniform distributions of \( C_l \) between \( C_{L,\min} \) and \( C_{L,\max} \), and \( \tau \) between \( t_{\min} \) and \( t_{\max} \). For numerical integration, we divide the range of possible values of \( C_l \) into 60 discrete points (steps of 0.1 aircraft/hour, across 6 aircraft/hour) and the range of possible values of \( \tau \) into 60 discrete points (steps of 2 minutes, across 2 hours). Given the specification of performance goals for capacity and predictability, the trade space is identified by calculating \( G_c \) and \( G_p \) across the range of planned \( C_l \) values from \( C_{L,\min} \) to \( C_{L,\max} \), and \( \tau \) values from \( t_{\min} \) and \( t_{\max} \), using data from the FAA Aviation Systems Performance Metrics (ASPM) database (FAA, 2012) and GDP duration data made available by Metron Aviation\textsuperscript{®}, described in detail in Section 5.3 below. The Pareto frontier is then identified for this trade space by discretizing the range of \( G_p \) values into a set number of points (100), and extracting the maximum corresponding value of \( G_c \) in each case. The relationship between the capacity and predictability metrics on the Pareto frontier was found to be very close to parabolic. A parabolic function is therefore fitted to the resulting maximum values of \( G_c \) for each given value of \( G_p \), of the form:

\[
G_c = aG_p^2 + bG_p + c \tag{23}
\]

We used the aforementioned 100 points on the Pareto frontier to estimate 3 parameter values (\( a \), \( b \), and \( c \)). Note that this model described in equation (23) is linear in parameters and hence the parameters can be estimated using an ordinary least squares (OLS) estimator. The estimation was performed using the standard MATLAB function for linear regression. Using data from the ASPM database (FAA, 2012), Form 41 database (DOT, 2012), and GDP duration data made available by Metron Aviation\textsuperscript{®}, described in detail in Section 5.3, the estimated parameters and fit performance at EWR and LGA are shown in Table 1. The parabolic nature of the relationship is confirmed by the consistently high \( R^2 \) values shown.

<table>
<thead>
<tr>
<th>Airport</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>EWR</td>
<td>-0.485</td>
<td>0.663</td>
<td>0.776</td>
<td>0.91</td>
</tr>
<tr>
<td>LGA</td>
<td>-1.086</td>
<td>1.626</td>
<td>0.395</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 1. Estimated Pareto frontier parameters for the detailed GDP model

Airline Payoff Functions

For the simplified GDP model, airline payoff functions are defined based on the cost of the flight delays incurred as a result of the GDP (payoff is therefore negative), similar to the way in which Liu and Hansen (2012) define a metric for efficiency. We assume that all the flights depart their respective origin airports under the assumption that the GDP ends at exactly the planned GDP end time \( T \). If the actual GDP end time, \( \tau \), is before or equal to \( T \), flights incur a delay on the ground equal to the difference between the actual departure time based on the planned GDP end time and the originally scheduled departure time. We assume that this delay is incurred at a specific ground delay cost of \( Cost_D \) dollars/minute, which varies by airline. If, however, the GDP is extended, and the GDP end time is after \( T \), airborne delay is incurred in addition to
ground delay (assuming the flight has already departed by the time the GDP is extended). The cost of incurring delay in airborne holding is generally significantly higher than that of holding on the ground. We define the ratio of the airborne delay cost per minute to the ground delay cost per minute as $k$, which is typically greater than 1. This parameter $k$ is also different for different airlines. The values of parameters $Cost_D$ and $k$ for each airline are obtained from the Form 41 database (DOT, 2012). Using Liu and Hansen’s (2012) simplified model of a GDP, we derive the airline payoff as a function of $\tau$ as follows:

$$Payoff(\tau) = \begin{cases} \frac{1}{2}Cost_D \left(\frac{c_H-c_L}{c_H-\lambda}\right) \left(T^2 + k\tau^2 - kT^2, \text{if } \tau > T\right), \\ T^2, \text{if } \tau \leq T \end{cases}$$

(24)

Given our assumed uniform distribution of $\tau$ between $t_{min}$ and $t_{max}$, we derive the expected value of the payoff as follows:

$$P_k = E[Payoff] = -\frac{1}{2}Cost_D \left(\frac{c_H-c_L}{c_H-\lambda}\right) \left(\frac{T^2(t_{max}-t_{min}^3-kT^2)}{t_{max}^3-t_{min}} + \frac{k^2}{kT^3} \right)$$

(25)

The derivations of equations (24) and (25) are shown in Appendix D. Yi and Hansen (2012) apply a special case of these equations with $k=2$ in the derivation of their efficiency metric. In this paper, however, we model parameters $Cost_D$ and $k$ differently for different airlines, as described in Section 5.3 below.

For the more detailed GDP model, airline payoff functions are also defined based on the cost of the flight delays incurred as a result of the GDP, as in the simplified model, but this flight delay is calculated as a function of both GDP end time and AAR. We assume that all the flights take off from their respective origin airports assuming that the GDP AAR is exactly as planned, $C_a$, and that the GDP will end at exactly the planned GDP end time $T$. If the actual GDP AAR, $C_a$ is greater than or equal to $C_L$, flights are assumed to be delayed on the ground in order to meet the planned GDP AAR $C_L$. Similarly, if the actual GDP end time, $T$, is before or equal to $T$, flights also incur a delay on the ground in order to meet the planned GDP end time $T$. As in the simplified model, we assume that ground delay is incurred at a specific ground delay cost of $Cost_D$ dollars/minute, which is different for different airlines. If, however, the actual GDP AAR, $C_a$ is less than $C_L$, airborne delay is incurred (in addition to ground delay), because flights arrive at the airport at a rate higher than the actual AAR. Similarly, if the GDP is extended, and the GDP end time is after $T$, airborne delay is incurred (in addition to ground delay). As in the simplified model, we define the ratio of the airborne delay cost per minute to the ground delay cost per minute as $k$, which is typically greater than 1. This parameter $k$ is also different for different airlines. As before, the values of parameters $Cost_D$ and $k$ for each airline are obtained from the Form 41 database (DOT, 2012). Expressions describing airline payoff as a function of $C_L$ and $T$ are presented in Appendix C. Based on these expressions, the expected value of airline payoff can be calculated numerically assuming uniform distributions of $C_L$ between $C_L_{min}$ and $C_L_{max}$, and $T$ between $t_{min}$ and $t_{max}$.

Given the specification of airline payoff as well as performance goals for capacity and predictability, airline payoff functions are identified by calculating airline payoff, $G_c$ and $G_p$ across a range of $C_L$ values from $C_L_{min}$ to $C_L_{max}$, and $T$ values from $t_{min}$ and $t_{max}$. The relationship between the performance goals and the airline payoffs was found to be very close to linear. As a result, a linear function was fitted to the resulting airline payoff as a function of $G_c$ and $G_p$, of the form:

$$P_k = b_{c,k} * G_c + b_{p,k} * G_p + \text{Constant}_k \quad \forall k$$

(26)
The parameters were estimated using an OLS estimator implemented in MATLAB. Using data from the ASPM database (FAA, 2012), Form 41 database (DOT, 2012), and GDP duration data made available by Metron Aviation\textsuperscript{®}, described in detail in Section 5.3, the estimated parameters and fit performance for each airline at EWR and LGA are shown in Table 2. The values of $\text{Constant}_k$ are not shown because they are inconsequential to the analysis. The linear nature of the relationship is confirmed by the consistently high $R^2$ values.

<table>
<thead>
<tr>
<th>Airport</th>
<th>Airline</th>
<th>Capacity $b_{i,k}$</th>
<th>Predictability $b_{p,k}$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EWR</td>
<td>Delta</td>
<td>1,209</td>
<td>355</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>US Airways</td>
<td>1,550</td>
<td>287</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>United</td>
<td>1,528</td>
<td>343</td>
<td>0.93</td>
</tr>
<tr>
<td>LGA</td>
<td>Delta</td>
<td>1,154</td>
<td>621</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>US Airways</td>
<td>1,436</td>
<td>495</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>United</td>
<td>1,430</td>
<td>596</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>1,554</td>
<td>608</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 2. Estimated payoff-function parameters for the detailed GDP model

5.3 Simulation Methodology

The methodology for simulating the generic initiative and the simplified GDP model is to run a Monte-Carlo simulation, sampling values for key parameters for each simulation run from appropriate distributions, and solving for the system optimal, most equitable, truthful, and strategic airline-preferred performance goal vectors in each case. Each of these problems is formulated as a mixed integer optimization problem with linear or convex objective function and linear and/or convex constraints. The detailed mathematical formulations of these optimization problems are provided in Appendix E. For solving the optimization problems for all approaches except ranking and voting we use CVX, a MATLAB-based modeling system for convex optimization. For ranking and voting we use CPLEX, a commercial optimization software package sold by IBM ILOG. The mechanism performance metrics described in Section 3.3 are then calculated in each case for the performance goal vectors and the corresponding airline payoffs. The metrics are then averaged over all simulation runs for comparison. Other aggregate statistics, such as standard deviation, minimum and maximum are also calculated. This is done for each of the performance goal resolution approaches described in Section 3.2, allowing each averaged metric to be compared across the different approaches. In contrast, the detailed GDP model is run for only one GDP at each of EWR and LGA, simulating in greater detail how airlines would behave under those specific conditions. Because of the significant burden associated with the numerical integration process in calculating expected values for each run, the use of Monte-Carlo methods on the detailed GDP model for repeated computation of these values is considered to be beyond the scope for this paper. Again, we solve for the system optimal, most equitable, truthful, and strategic airline-preferred performance goal vectors in each case, before calculating the mechanism performance metrics described in Section 3.3 for each of the performance goal resolution approaches described in Section 3.2.

As described in Section 3.3, the system optimal solution is found by identifying the performance goal vector that, if applied, maximizes the total payoff across all airlines combined. The most equitable solution is found
by identifying the performance goal vector that, if applied, maximizes the minimum payoff of any airline. Each airline’s truthful solution is found by identifying the performance goal vector that, if applied directly as the chosen system-wide performance goal vector, maximizes the individual payoff of that airline. Finally, each airline’s strategic solution is found by simulating a myopic best-response game between airlines. This best-response game is solved iteratively, and is myopic in the following sense: In each iteration, each airline’s input is determined by identifying the performance goal vector that, if combined with the performance goal vectors input by all other airlines in the previous iteration using the given performance goal resolution approach, maximizes the payoff for that airline. This applies to both a continuous trade space and a discrete trade space. In the latter case, the discrete performance goal vectors are ranked or voted upon by each airline in such a way that the winning performance goal vector maximizes the payoff for that airline, given the rankings or votes of all other airlines in the previous iteration. The performance goal vectors of all airlines are combined based on the different approaches described in Section 3.2. Each airline’s input is re-optimized at every iteration, based on the corresponding inputs of all other airlines. The game is simulated until convergence, but stopped after 100 iterations if the stopping criterion is not met by then. Convergence is said to be achieved when the difference in each of the airline preferred performance goal vectors in successive iterations is less than $1 \times 10^{-6}$ (for the case of a continuous trade space), or when each of the airline preferred performance goal vectors in successive iterations are the same (for the discrete case). Note that each performance goal has a value between 0 and 1. If convergence is not reached in 100 iterations, the results of that simulated game are not included in the calculation of the metrics described in Section 3.3. For each approach, the percentage of runs that converged to equilibrium before 100 iterations is provided in Section 6. The mathematical formulations for each of the optimizations described above are presented in Appendix E.

Applying Monte-Carlo methods for the generic initiative and the simplified GDP model, the simulations are run 1,000 times, randomly sampling for the key parameters in each case. In the generic initiative, the parameters sampled include those defining the Pareto frontier and each airline’s payoff function, described in Section 5.1, and the airline weights ($w_k$). This last parameter, which represents the relative numbers of impacted flights of different airlines, is sampled from a uniform distribution from 0 to 1 for each airline, and then divided by the sum of sampled values across all airlines, to ensure that the sum of airline weights is 1.0.

In the case of ranking and voting, 5 candidate performance goal vectors are specified in each run, which the airlines must rank or vote on. The 5 discrete candidate performance goal vectors are generated using the following process: The capacity metric value ($G_c$) for each candidate vector is assumed to (1) vary randomly according to a uniform distribution from 0 to 1, and (2) to fall on the Pareto frontier. This automatically and uniquely determines the predictability metric value ($G_p$). Note that we also ran an experiment with the 5 discrete candidate performance goal vectors being distributed evenly, rather than randomly, along the Pareto frontier, with the results changing by less than 2% of the values shown in Section 6 in all cases.

For both the simplified and detailed GDP models, we simulate operations at both EWR airport and at LGA airport. This allows us to compare the performance of the different approaches applied to real airports, and also allows us to compare how each approach performs for different types of airports. EWR was chosen as an example of an airport that is dominated by a single airline: United Airlines. LGA was chosen as an
example of an airport with no single dominant carrier. Both these airports are among the busiest and the most congested airports in the U.S. (Barnhart et al., 2012). In each scenario, we simulate all airlines that operate at least 5% of the operations at that airport. For EWR, this includes only United Airlines, Delta Air Lines and US Airways; and for LGA, it includes United Airlines, Delta Air Lines, US Airways, and American Airlines (FAA, 2012).

For the simplified GDP model, we analyze historical data for 2011 (specifically, ASPM data (FAA, 2012) and Form 41 data (DOT, 2012)) to define distributions from which to sample values for each parameter in the Monte-Carlo simulation. The parameters, and the type of distributions from which they are sampled, are described in Table 3. In each case, average values from the Operational Evolution Partnership (OEP) 35 airports (FAA, 2012) are also shown for comparison.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution Assumed</th>
<th>Airline</th>
<th>EWR</th>
<th>LGA</th>
<th>OEP35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of airline operations to total - ( w_k )</td>
<td>None (Fixed)</td>
<td>American</td>
<td>-</td>
<td>0.207</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delta</td>
<td>0.067</td>
<td>0.272</td>
<td>0 to 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>United</td>
<td>0.886</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>US Airways</td>
<td>0.112</td>
<td>0.282</td>
<td></td>
</tr>
<tr>
<td>Avg. scheduled arrival demand rate - ( \lambda ) [ac/hr]</td>
<td>Poisson</td>
<td>All</td>
<td>33.7</td>
<td>35.1</td>
<td>40.0</td>
</tr>
<tr>
<td>Avg. AAR during GDP - ( C_L ) [ac/hr] (Std. dev.)</td>
<td>Normal</td>
<td>All</td>
<td>36.2 (3.32)</td>
<td>33.3 (4.44)</td>
<td>45.0 (20.9)</td>
</tr>
<tr>
<td>Avg. AAR not under GDP - ( C_H ) [ac/hr] (Std. dev.)</td>
<td>Normal</td>
<td>All</td>
<td>39.8 (6.40)</td>
<td>34.6 (9.9)</td>
<td>63.2 (27.2)</td>
</tr>
<tr>
<td>Avg. GDP duration [hrs] (Std. dev.)</td>
<td>Normal</td>
<td>All</td>
<td>6.95 (2.90)</td>
<td>8.01 (3.87)</td>
<td>6.14 (3.57)</td>
</tr>
<tr>
<td>Avg. difference between max and min GDP durations [hrs] (Std dev.)</td>
<td>Normal</td>
<td>All</td>
<td>1.59 (1.35)</td>
<td>1.62 (1.21)</td>
<td>2.03 (1.77)</td>
</tr>
<tr>
<td>Avg. airline flight operating cost during airport holding - ( k \times Cost_D ) [US$/hr] (Std. dev.)</td>
<td>Normal</td>
<td>American</td>
<td>$6,592 ($3,770)</td>
<td>$7,968</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delta</td>
<td>$5,365 ($1,435)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>United</td>
<td>$5,777 ($4,115)</td>
<td>($5,997)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>US Airways</td>
<td>$5,211 ($2,798)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. airline flight operating cost on ground - ( Cost_D ) [US$/hr] (Std. dev.)</td>
<td>Normal</td>
<td>American</td>
<td>$2,819 ($3,845)</td>
<td>$4,061</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delta</td>
<td>$2,074 ($922)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>United</td>
<td>$2,592 ($2,195)</td>
<td>($4,184)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>US Airways</td>
<td>$2,607 ($1,343)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of airline cost during holding and on the ground - ( k )</td>
<td>Normal</td>
<td>American</td>
<td>2.27 (0.28)</td>
<td>2.32 (1.40)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delta</td>
<td>2.79 (1.75)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>United</td>
<td>2.36 (2.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>US Airways</td>
<td>2.11 (0.53)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. GDP parameter distributions

Based on equations (27) and (28) below, the maximum and minimum duration of the GDP (\( t_{max} \) and \( t_{min} \)), are calculated by sampling the average GDP duration \( t_{avg} \), and the difference between the maximum and minimum durations \( t_{diff} \). The distributions of these two factors (\( t_{avg} \) and \( t_{diff} \)) are described in Table 3. These distributions are based on historical GDP durations and historically observed differences between the initial GDP duration and the actual GDP duration, respectively. This data was made available by Metron Aviation®.
\[ t_{\text{max}} = t_{\text{avg}} + \frac{t_{\text{diff}}}{2} \]  
\[ t_{\text{min}} = t_{\text{avg}} - \frac{t_{\text{diff}}}{2} \]  

(27)  
(28)

In the case of ranking and voting, 5 candidate performance goal vectors are specified in each run, which the airlines must rank or vote on. \( T \) values are sampled for each candidate vector from a uniform distribution between \( t_{\text{min}} \) and \( t_{\text{max}} \). The capacity and predictability metric values \( (G_c \text{ and } G_p) \) for each candidate vector are then calculated by substituting each sampled value of \( T \) in equations (21) and (22). Across all approaches and all runs of the simplified model, the minimum values for \( G_c \) and \( G_p \) at EWR are found to be 0.71 and 0.03, respectively, while at LGA, they are found to be 0.75 and 0.10, respectively.

We use historical data for the two simulated days to define each parameter, as follows.

- The ratio of operations of each airline to the total \( (w_k) \) and the airline delay costs \( (k \times Cost_D \text{ and } Cost_D) \) are as shown in Table 3.
- The scheduled arrival demand rate \( (\lambda) \), which varies with time, is specified per hour according to Figure 7.
- The AAR after the GDP \( (C_H) \) is assumed to be 36 aircraft/hour and 35 aircraft/hour at EWR and LGA, respectively, based on the specified AARs in Figure 7.
- The minimum AAR of the GDP \( (C_{L, \text{min}}) \) is assumed to be 22 aircraft/hour and 12 aircraft/hour at EWR and LGA, respectively, while the maximum AAR of the GDP \( (C_{L, \text{max}}) \) is assumed to be 28 aircraft/hour and 18 aircraft/hour, respectively. These compare to the actual GDP AARs in Figure 7 of 25 aircraft/hour and 15 aircraft/hour, respectively.
- The minimum duration of the GDP \( (t_{\text{min}}) \) is assumed to be 3 hours and 2 hours at EWR and LGA, respectively, while the maximum duration of the GDP \( (t_{\text{max}}) \) is assumed to be 5 hours and 4 hours, respectively. These compare to the actual durations in Figure 7 of 4 hours and 3 hours, respectively.
- In the case of ranking and voting, 5 candidate performance goal vectors are specified in each run, which the airlines must rank or vote on. The capacity metric value \( (G_c) \) for each candidate vector is assumed to be uniformly distributed between the minimum and maximum values of \( G_c \). These are calculated based on the ranges of planned \( C_L \) values from \( C_{L, \text{min}} \) to \( C_{L, \text{max}} \) and \( T \) values from \( t_{\text{min}} \) and \( t_{\text{max}} \) described above. The corresponding value of the predictability metric is then uniquely determined by solving equation (23) that characterizes the Pareto frontier. At EWR, the minimum values for \( G_c \) and \( G_p \) are found to be 0.91 and 0.55, respectively, while at LGA, they are found to be 0.92 and 0.74, respectively.
6. Simulation Analysis and Results

In this section, we describe the results of our analysis. We evaluate the five approaches for performance goal resolution (as described in Section 3.2) based on five mechanism performance metrics, namely Pareto optimality, system optimality, airline profitability, equity, and truthfulness (as described in Section 3.3).

Each of the approaches described in Section 3.2 is simulated for the generic initiative and for the simplified and detailed models of a GDP at EWR and LGA, for a small number of competing airlines (three for the generic initiative and GDP at EWR, four for the GDP at LGA), trading off two performance criteria. Monte-Carlo methods are applied in the cases of the generic initiative and the simplified model of the GDP to simulate varying conditions by sampling values from distributions for each of the parameters, as described in Section 5.3. The more detailed model of the GDP is run for two sample GDP cases: July 25 2011 at EWR, and July 8 2011 at LGA. Results are presented for each model in the following sub-sections.

6.1 Generic Traffic Management Initiative

The results for the generic TMI are shown in Table 4. Each of the metrics described in Section 3.3 is calculated for each of the performance goal resolution approaches described in Section 3.2. In each case, the simulations are run with three different types of Pareto frontiers, as described in Section 5.1: an arc, a parabola, and a piecewise-linear function. The airline payoff functions are as described in Section 5.1. The average metric values across all 1000 Monte-Carlo runs are presented, along with the standard deviations (in parentheses), and maximum and minimum values (in square brackets). The percentage of runs that converge and the average number of iterations to convergence, when convergence occurs, are also presented in each case.
Table 4. Generic TMI: Comparison of average metrics (with standard deviations in parentheses) [and maximum and minimum values in square brackets] for all approaches.

For each of the three Pareto frontier shapes, all the simulation experiments converge under approaches 3 and 5 within 100 iterations. More than 99% of runs converge under approach 1, while between 69% and 91% of runs converge under approach 2, depending on the Pareto frontier shape. This means that pushing the weighted average of user inputs out to the Pareto frontier has a negative impact on convergence, and may make its practical application challenging, at least in the form simulated here. However, under approach 4
(i.e., ranking), only between 21% and 25% of the simulation runs converge, depending on the Pareto frontier shape. Given that the ranking approach converges in only a small fraction of the simulated cases, it seems unlikely to be useful for practical implementation in the form simulated here because a stable solution can rarely be achieved. Convergence can be improved to some extent by adjusting the values assigned to the different ranks, although this has not been investigated in detail in this paper.

In all the cases that converge, the average number of iterations required to reach convergence is significantly lower (averaging 6 iterations) than the maximum number of 100 iterations permitted by the model. Approach 3 always converges in a single iteration, by definition, while approach 5 (i.e., voting) has the poorest performance, averaging 10 iterations to convergence, which is still reasonable for practical implementation.

The first metric, describing how close the system-wide solution is to the Pareto frontier, shows that Pareto optimality is achieved by all approaches with the exception of approach 1, which takes a weighted average of the user preferred performance goal vectors. This is expected, as this is the only approach that is not specifically designed to achieve Pareto optimality. Even in this approach, however, the metric is consistently high – average values are between 0.95 and 0.98 (with standard deviations between 0.03 and 0.05), while the lowest value achieved in the Monte-Carlo simulation is 0.68 (piecewise-linear). This is because the preferred solutions input by each airline, from which the weighted average is calculated, are always Pareto optimal.

The second, third and fourth metrics, describing how close the system-wide solution is to maximizing each airline’s payoff; how close the system-wide solution is to system optimality; and how equitable the system-wide solution is, all show similar results. The metrics are generally lowest for approach 4, which applies ranking. This approach also shows the highest variability across different Monte-Carlo runs, and the lowest (or near lowest) minimum metric values (0.24, 0.24 and 0.09 for metrics two, three and four, respectively). Each of these three metrics is generally highest for approach 5, which applies voting, with approach 2, which pushes the weighted average of the user preferred performance goal vectors to the Pareto frontier, close behind. Approach 2 also shows the lowest metric variability and highest minimum values (0.57, 0.76 and 0.16 for metrics two, three and four, respectively) of all the approaches. Approach 1 and approach 3 (a weighted random choice of the user preferred performance goal vectors) show results that are slightly lower than those for approaches 2 and 5, and with slightly greater variability (although this pattern does not hold as strongly for the fourth metric). The results are also generally similar across the different Pareto frontiers, with the piecewise-linear results typically slightly lower than for the arc or parabola. In many cases the piecewise-linear Pareto frontier leads to the system-wide solution falling at the intersection of the lines (or as close to it as possible in the case of discrete options), which is generally farther from the airline-optimal, system-optimal, and most-equitable solutions than for the arc or parabola.

The fifth metric describing the truthfulness of the airline inputs is particularly low (between 0.41 and 0.83) for approach 2. This is a recognized concern about approach 2, as discussed in Section 3.2. In contrast, the metric is highest (1.00) for approach 3. This is expected as this approach is specifically designed to prevent gaming. Also notable, however, is that the metric is also high for approach 5 (between 0.99 and 1.00), while it is lower for approach 4 (between 0.69 and 0.79), and to a lesser extent for approach 1 (between 0.85 and 0.95). The variability of the truthfulness metric is greatest for approach 4 (between 0.17 and 0.21), and lowest for approach 3 (0) and approach 5 (between 0.02 and 0.03). The minimum value of truthfulness is
lowest for approaches 2 and 4 (0.00), highest for approach 3 (1.00), and second highest for approach 5 (0.62). The metric also shows some variation across the different Pareto frontiers, with the piecewise-linear results typically slightly lower than for the arc or parabola. This is because gaming leads each airline to request solutions at the vertices of the Pareto frontier, as illustrated in Figure 3 (or as close to the vertices as possible in the case of discrete options), which are generally farther from their truthful solutions than in the arc and parabola cases. Furthermore, the truthfulness metric has a higher value for the parabolic frontier than for the arc-shaped frontier, which is consistent with our theoretical results (Propositions 2 and 3 in Appendix B) that show that the necessary and sufficient conditions for truthfulness are less restrictive for the parabolic frontier than for the arc-shaped frontier for approach 1.

Additionally, the joint distribution of the different mechanism performance metrics across the simulation runs provides some other interesting observations. It is noted that, when the truthfulness metric is at its minimum, the other metrics are all generally high. This is the case across all approaches, except approach 4 (ranking). Also notable, is that when the other metrics are at their minimums, the metric for truthfulness is generally high. This is the case across all approaches except approaches 2 and 4, and approach 1 applying a piecewise-linear Pareto frontier. This is not to say, however, that the metric values are either high for truthfulness, or for the other metrics, but not for both. Indeed, results are observed that are high across all metrics, particularly under voting. A further observation is that the minimum metric values very rarely coincide with the minimum values of other metrics, except under ranking and voting, where the trade space is a discrete set of performance goal vectors, and even then it is relatively rare. In the instances when the minimum metrics do coincide under ranking and voting, it is metrics 2 and 4 that coincide, describing how close the system-wide solution is to maximizing each airline’s payoff, and how equitable the system-wide solution is. It seems reasonable that if one of these metrics were low, the other would also likely be. In these cases the metric for truthfulness is high, while the metric for system optimality is neither particularly high, nor particularly low.

In summary, based on the results from this generic initiative, we can derive a number of practical insights. First, we note that amongst the five approaches for performance goal resolution, all but approaches 2 (pushing the weighted average of user inputs out to the Pareto frontier) and 4 (ranking) converge to equilibrium in almost all the simulation runs. Ranking, particularly, frequently runs into convergence issues, and is thus unsuitable for practical implementation, at least in this form. Approach 2 may also have challenges in practical application because of issues with convergence. It is worth noting here that voting (approach 5) could potentially have run into convergence issues had it not been for the constraint that prohibits airlines from reducing their votes for any of the candidate solutions in subsequent iterations. Of the four approaches 1, 2, 3, and 5, approach 2 performs substantially worse than all the other approaches in terms of truthfulness. Therefore, even though approach 2 shows reasonable levels of performance on some of the other metrics, the submitted preferences by the airlines are unlikely to have much meaning. Approaches 1, 3 and 5, therefore, seem to be reasonable candidates for practical implementation. There is still, however, quite a large variation in performance across these approaches, as displayed in Table 4. Overall, voting out-perform approaches 1 and 3, with none of the average values of the metrics being below 0.88 across all the three shapes of the Pareto frontier. These conclusions based on the generic initiative set the stage for further evaluation of these approaches on GDP initiatives at EWR and LGA.
6.2 Ground Delay Program

Simplified Model of a GDP

Table 5 shows the results for the simplified model of the GDP. Each of the metrics described in Section 3.3 is calculated in each case, for each of the performance goal resolution approaches described in Section 3.2. The Pareto frontiers and airline payoff functions are simulated as described in Section 5.2. The average metric values across all 1000 Monte-Carlo runs are shown, along with the standard deviations (in parentheses), and maximum and minimum values (in square brackets). The percentage of runs that converge and the average number of iterations to convergence, when convergence occurs, are also presented in each case.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Airport</th>
<th>% Conv.</th>
<th>Avg. It.</th>
<th>Pareto Optimality</th>
<th>Airline Profitability</th>
<th>System Optimality</th>
<th>Equity</th>
<th>Truthfulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Weighted Avg.</td>
<td>EWR</td>
<td>100%</td>
<td>7</td>
<td>1.00 (0.00)</td>
<td>0.97 (0.03)</td>
<td>0.99 (0.02)</td>
<td>0.96 (0.05)</td>
<td>0.92 (0.06)</td>
</tr>
<tr>
<td></td>
<td>LGA</td>
<td>100%</td>
<td>6</td>
<td>1.00 (0.00)</td>
<td>0.90 (0.11)</td>
<td>0.90 (0.13)</td>
<td>0.82 (0.16)</td>
<td>0.95 (0.05)</td>
</tr>
<tr>
<td>2. Weighted Avg. Pushed to Pareto</td>
<td>EWR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LGA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3. Weighted Random Choice</td>
<td>EWR</td>
<td>100%</td>
<td>1</td>
<td>1.00 (0.00)</td>
<td>0.97 (0.04)</td>
<td>0.99 (0.03)</td>
<td>0.96 (0.06)</td>
<td>1.00 (0.00)</td>
</tr>
<tr>
<td></td>
<td>LGA</td>
<td>100%</td>
<td>1</td>
<td>1.00 (0.00)</td>
<td>0.98 (0.03)</td>
<td>0.99 (0.02)</td>
<td>0.96 (0.06)</td>
<td>1.00 (0.00)</td>
</tr>
<tr>
<td>4. Ranking</td>
<td>EWR</td>
<td>73%</td>
<td>3</td>
<td>1.00 (0.00)</td>
<td>0.95 (0.06)</td>
<td>0.95 (0.08)</td>
<td>0.93 (0.10)</td>
<td>0.95 (0.05)</td>
</tr>
<tr>
<td></td>
<td>LGA</td>
<td>8%</td>
<td>12</td>
<td>1.00 (0.00)</td>
<td>0.98 (0.04)</td>
<td>0.98 (0.04)</td>
<td>0.98 (0.06)</td>
<td>0.98 (0.03)</td>
</tr>
<tr>
<td>5. Voting</td>
<td>EWR</td>
<td>100%</td>
<td>7</td>
<td>1.00 (0.00)</td>
<td>0.98 (0.04)</td>
<td>1.00 (0.01)</td>
<td>0.98 (0.07)</td>
<td>1.00 (0.00)</td>
</tr>
<tr>
<td></td>
<td>LGA</td>
<td>100%</td>
<td>10</td>
<td>1.00 (0.00)</td>
<td>0.99 (0.02)</td>
<td>1.00 (0.01)</td>
<td>0.98 (0.05)</td>
<td>1.00 (0.00)</td>
</tr>
</tbody>
</table>

Table 5. Simplified GDP model: Comparison of average metrics (with standard deviations in parentheses) [and maximum and minimum values in square brackets] for all approaches.

It is noted that no results are presented in Table 5 for approach 2, which pushes the weighted average of the user preferred performance goal vectors out to the Pareto frontier. This is because, for this model, approach 2 is no different from approach 1, which just takes the weighted average of the user preferred performance goal vectors. This is because the simplified GDP scenario is defined for a single ANSP decision variable, that is, the planned GDP end time, $T$. Performance metrics for capacity and predictability are calculated as a
function of this single decision variable, as described in Section 5.2. Therefore, for each planned GDP end
time, there are unique values for both the capacity and predictability performance goals. The Pareto frontier
represents the corresponding values for these two performance goals at all possible planned GDP end times.
The entire trade space therefore lies on the Pareto frontier, and no interior points are feasible. For this reason,
a more detailed GDP model is also run, modeling multiple decision variables, as described above, although
only for a single GDP case at each of EWR and LGA.

For both airports, all the simulation experiments converge under approaches 1, 3 and 5 within 100 iterations.
However, under approach 4 (ranking), only 8% of the simulation runs converge at LGA. This result
reinforces our conclusion that ranking is unlikely to yield stable outcomes for practical implementation, at
least in the form simulated here. At EWR, 73% of simulation runs converge under ranking, which is
significantly better, but still not as reliable as the other approaches. The reason for this improved
performance is the highly skewed distribution of operations across airlines at EWR, with United operating
89% of flights. However, for an initiative involving a more even distribution of operations across airlines,
like at LGA, ranking seems to be unsuitable for practical implementation in its current form.

In all the cases that converge, the average number of iterations required to reach convergence is significantly
lower (averaging 6 iterations) than the maximum number of 100 iterations permitted by the model. Approach
3 always converges in a single iteration, by definition, while approach 5 (i.e., voting) has the poorest
performance, averaging 9 iterations to convergence, which is still reasonable for practical implementation.

The first metric shows that all approaches simulated are Pareto optimal at EWR and LGA, as was the case
for the generic initiative in Table 4. All other metrics, with the exception of those under approach 1 at LGA,
are high (between 0.92 and 1.00), have relatively low variability (between 0 and 0.10), and have minimum
values above 0.45. They are also very similar to the results presented for the generic initiative in Table 4,
with the exception of somewhat higher values of metrics 2, 3, 4 and 5 in approach 4 (between 0.93 to 0.99, in
comparison to values between 0.69 and 0.90 in Table 4). The similarities between the results presented here
and in Table 4 suggest that the conclusions drawn for the generic initiative are indeed applicable to more
realistic cases, as simulated here. Particularly, the best performing approach across all metrics at both EWR
and LGA is approach 5, (with the average of all metrics for both airports consistently at least 0.98), with
approach 3 not far behind (with the average of all metrics for both airports consistently at least 0.96).

The results for EWR and LGA differ relatively little. In approach 1, airline profitability, system optimality
and equity are all somewhat higher at EWR, where a single airline is dominant, than at LGA, where no
airline is dominant. For all other approaches, the results differ very little across the two airports. This
suggests that only approach 1 is likely to be significantly affected by the dominance of an airline at the
airport. (Note that approach 2, not considered here, is also likely to be affected by airline dominance.)

The overall results from the simplified GDP scenarios at EWR and LGA therefore reinforce our conclusions
from the generic initiative. Once again, ranking in its current form can be eliminated from consideration
because of convergence issues for the GDP scenario at LGA. Approaches 3 and 5 (and to a lesser extent
approach 1) seem to be reasonable candidates for practical implementation. However, based on these results,
voting (approach 5) performs the best, with none of the metrics values being below 0.98 for either the EWR or the LGA scenario.

The joint distribution of the different mechanism performance metrics across the simulation runs provides very similar observations to those for the generic initiative. When the truthfulness metric is at its minimum, the other metrics are all generally high, and when the other metrics are at their minimums, the metric for truthfulness is generally high. This is the case across all approaches except under approach 1 at EWR. Again, this is not to say that metric values are either high for truthfulness, or for the other metrics, but not for both. Indeed, results are observed that are high across all metrics, particularly under voting. The minimum metric values rarely coincide with the minimum values of other metrics, but unlike for the generic initiative, they do sometimes coincide under all approaches, and it can be any of metrics 2, 3 or 4 that coincide.

**More Detailed Model of a GDP**

The results applying the more detailed model of the GDP are shown in Table 6. Each of the metrics described in Section 3.3 is calculated in each case, for each of the performance goal resolution approaches described in Section 3.2. The Pareto frontiers and airline payoff functions simulated are as estimated in Section 5.2. The results are presented with a higher number of significant figures than for the other models in order to distinguish differences in the metric values. Note that we present only a single value per metric, rather than average, standard deviation, minimum, and maximum as for the other models, because we only do this evaluation for a single GDP case rather than using a Monte-Carlo simulation approach. The numbers of iterations to convergence, when convergence occurs, are also presented in each case.

Comparing the results in Table 6 to the results in Table 5, it is immediately clear that the values of all the metrics are closer to unity than modeled using the simplified GDP model, hence the requirement for increased significant figures in the table. This is because of the size of the feasible region in the detailed GDP model. As described in Section 5.3, while the metrics $G_c$ and $G_p$ can each nominally vary from 0 to 1, because of the variation in GDP end time $T$ from $t_{min}$ to $t_{max}$ and GDP magnitude $C_L$ from $C_{L_{min}}$ to $C_{L_{max}}$, $G_c$ and $G_p$ only vary from 0.91 to 1 and 0.55 to 1, respectively, at EWR, and from 0.92 to 1 and 0.74 to 1, respectively, at LGA. In the generic initiative, across the 1000 Monte-Carlo simulation runs, $G_c$ and $G_p$ both vary from 0 to 1, while in the simplified GDP, across the 1000 Monte-Carlo simulation runs, $G_c$ and $G_p$ vary from 0.71 to 1 and 0.03 to 1, respectively at EWR, and from 0.75 to 1 and 0.10 to 1, respectively at LGA. Hence the range over which the different solutions can be found in the detailed GDP is reduced relative to the previous results. All the solutions also lie on or close to the Pareto frontier, combining to make most of the metric values close to unity. Also, note that this is a single computational experiment and not a set of 1000 Monte-Carlo simulation runs, as in the case of the generic and the simplified GDP initiatives. Therefore, the low variation in the metric values is not surprising.
Table 6: Detailed model of GDP at EWR (July 25 2011) and LGA (July 8 2011): Comparison of metrics for all approaches.

As for the simplified GDP model, we find that the detailed GDP model converges under approaches 1, 3 and 5 at both EWR and LGA. Also similar to previous results, the detailed GDP model does not converge under approach 4 at either EWR or LGA, reinforcing our conclusion that ranking is unlikely to yield stable outcomes for practical implementation, at least in the form simulated here. Approach 2, which was not run for the simplified GDP model, converges at LGA, but not at EWR. Convergence is not guaranteed in this approach and a significant number of cases of non-convergence are observed modeling the generic initiative under this approach (19% of runs), shown in Table 4. While convergence is observed at LGA, this result suggests that approach 2 may also not be suitable for practical implementation, at least in the form simulated here. In the cases that do converge, convergence is fast – generally in 2 iterations, except under approach 2 at LGA, where it is in 3 iterations. This suggests that the airlines have similar preferences, and that a high number of iterations would not be required in practical implementation.

As in Table 4, all approaches are Pareto optimal except approach 1, as expected. However, because of where the airline preferred solutions lie on the Pareto frontier in this case, even approach 1 is almost Pareto optimal. All other metrics are very high. Particularly, truthfulness is at a maximum under all approaches except approach 2, where it is still high (0.99). Approach 2 has a recognized issue with truthfulness, as discussed in Section 3.2, so it is expected that this metric would be lower under approach 2 than the other approaches. Similar to the results for the generic initiative and the simplified GDP, the best performing approach is 5

1 This approach failed to converge at this airport.
(voting), with all metric values consistently above 0.9997, and approach 3 is close behind with all metric values consistently above 0.9991. Similar to the results for the simplified GDP, the results for EWR and LGA differ very little, with the exception of the differences in convergence under approach 2.

The overall results from the detailed GDP scenarios reinforce our conclusions from the generic initiative and the simplified GDP scenarios. Approaches 2 and 4, in their current form, can be eliminated because of convergence issues. Approaches 1, 3 and 5 seem to be reasonable candidates for practical implementation, but voting, approach 5, consistently performs the best.

7. Conclusions

In this paper a number of approaches are considered within the context of an air traffic management system for setting system-wide performance goals based on preferred trade-offs as expressed by airlines. This is the first study to investigate how system-wide performance goals can be set based on airline inputs. We investigate this by evaluating differing approaches for doing so using a rigorous game-theoretic method, which identifies the potential for gaming in each approach. Each approach is evaluated based on a number of proposed criteria representing stakeholder objectives, including Pareto optimality, airline profitability, system optimality, equity and truthfulness of airline preferences. By simulating each of the approaches using Monte-Carlo methods, sampling values for input parameters from representative distributions, we offer a broad evaluation of each approach to performance-based ATM. A generic TMI is simulated, as well as a simplified model of a ground delay program at Newark Liberty International airport and LaGuardia airport. Finally, a more detailed model of a ground delay program is also run, for a single case at each of Newark Liberty International airport and LaGuardia airport, confirming the validity of the previous Monte-Carlo results.

The results presented in Section 6 suggest that taking a weighted average of the user preferred performance goal vectors (approach 1), making a weighted random choice of the user preferred performance goal vectors (approach 3), or voting on ANSP provided candidate performance goal vectors (approach 5) may all be reasonable candidates for practical implementation. However, approach 1 has somewhat inferior performance for the simplified GDP case, while approach 3 might suffer from lack of stakeholder buy-in because the process of combining airline preferences involves a non-deterministic (randomized) step. Overall, voting shows the highest promise. Our results based on multi-criteria evaluation suggest that this approach has the most potential to satisfy the objectives of all stakeholders.

While the results presented here suggest that voting is likely to be the best way to satisfy the objectives of all stakeholders in setting performance goals, different voting schemes have not been examined in detail. It is therefore recommended that future research include the detailed analysis of specific voting schemes, including instant run-off voting (Lewyn, 2012; Robb, 2012) and majority judgment voting (Balinski and Laraki, 2007; 2011), with the goal of developing implementable airline-driven approaches for performance-based air traffic management that will benefit both airlines and ANSPs.
Acknowledgements

This work was funded through the National Center of Excellence for Aviation Operations Research (NEXTOR), a consortium of 8 universities contracted by the FAA to provide research support for a wide variety of aviation issues. Their support is gratefully acknowledged. The authors would also like to thank colleagues on the Service Expectations project (Distributed Mechanisms for Determining NAS-Wide Service Level Expectations), including Rich Jehlen of the FAA, Mike Ball and Prem Swaroop of the University of Maryland, and Mark Hansen and Yi Liu of the University of California, Berkeley for helpful discussions, and information on the GDP models.

References

- DOT (Department of Transportation), 2012. Air Carrier Financial Reports (Form 41 Financial Data), Schedule P52, U.S. Department of Transport, Research and Innovative Technology Administration, Bureau of Transportation Statistics, Washington DC, USA.
• Robb, D.M. 2012. The Effect of Instant Runoff Voting on Democracy, PhD Thesis, Department of Political Science, University of California, Irvine, CA.