Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017)

Service Level Expectation Setting for Air Traffic Flow Management: Practical Challenges and Benefits Assessment

Michael Ball, Prem Swaroop
Robert H. Smith School of Business and
Institute for Systems Research
University of Maryland
College Park, MD 20742
Email: mball@rhsmith.umd.edu

Cynthia Barnhart, Chiwei Yan
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139

Mark Hansen, Lei Kang, Yi Liu
Department of Civil and Environmental Engineering
University of California
Berkeley, CA 94720

Vikrant Vaze
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755

***Abstract*—This paper describes a mechanism for determining consensus service level expectations to be used in designing air traffic management initiatives (TMIs). Our approach, which employs the Majority Judgment voting mechanism, enables those flight operators impacted by a potential TMI to provide service level preference information to an air navigation service provider (ANSP) that initiates the TMI. The output of the process is a numeric vector that specifies performance goals for the TMI enabling the ANSP to tradeoff competing performance criteria, when designing the TMI. Earlier work has described key components of the overall approach. This paper gives a comprehensive view and also gives the results of a fast-time simulation benefits assessment and a human-in-the-loop simulation.**

Keywords: Air Traffic Flow Management, Collaborative Decision Making, Majority Judgment

I. Introduction

Over the past 15 or more years, collaborative air traffic management (CATM) has become a fundamental principle underlying all new air traffic management (ATM) system development both in the U.S. and Europe. Its origins go back to the deployment of Collaborative Decision Making (CDM) information exchange and resource allocation mechanisms for planning and controlling ground delay programs (GDPs) in the U.S. [1]–[3] and similar information exchange and distribution mechanisms focused on flows into, through and out of an airport in Europe (A-CDM) [4]. In the U.S., GDP decision support tools evolved and became more sophisticated and the underlying GDP ideas were transferred to the enroute environment with the development of airspace flow programs. A variety of other tools, based on CATM paradigms, have been developed and adopted or are on their way to adoption both in the U.S. and Europe.

It is probably safe to say that the bulk of the CATM-based research in the U.S. has focused on tools and processes to support very specific ATM operational decisions, e.g. assigning ground delay to a specific flight during a GDP. At the same time, there is a very important strategic planning aspect to the daily execution of ATM. Specifically, FAA traffic managers consult with airline/flight operator operational personnel at both the local and national levels in planning operational strategies for the day. These take the form of strategic planning telecoms (SPTs). To be sure, the SPTs should be considered a very important part of the general trend toward the widespread use of CATM. They perform a very legitimate, and even vital, function in the overall traffic management process. Specifically, flight operators have key information not known by the FAA, including air carrier business objectives and economic tradeoffs and the status of aircraft and personnel, just to name a few. However, while CATM initiatives have produced a host of innovations specific to how traffic management initiatives (TMIs) are planned and controlled, very little innovation has been directed toward the operation of SPTs [4]. While this per se may not necessarily be bad, there are several concerns and issues related to SPTs and more generally strategic planning on the day-of-operations that merit research attention:

1) The SPTs are free form and highly unstructured and so, at times, can devote an inordinate amount of time to unimportant topics.
2) Due to their free-form nature, the SPTs do not attempt to assign priority to the various flight operators based on objective measures. Thus, the more persistent and/or "loudest" flight operators tend to have the most influence.

3) The operational concept for the Next Generation Air Transportation System (NextGen) calls for a performance-based ATM system. One embodiment of this concept calls for the separation of strategic ATM planning into: i) service level expectation setting; and ii) planning of an operational response [5]. Flight operator input should be provided and the air navigation service provider (ANSP) should then optimize based on the output of service level expectation setting step. Today's SPT's generally do not discuss performance expectations, focusing instead on specific TMI parameters.

The proposed system – COuNSEL, CONsensus Service Expectation Level setting – described in this paper addresses the service level expectation (SLE) setting problem and, in so doing, seeks to eliminate the deficiencies discussed above. The initial COuNSEL concepts were introduced in [6], the mathematical models underlying a key component (candidate vector generation) are provided in [7]. For other related work, see [8] and [9]. In this paper, we provide an overview of the concept and processes, and give the results of human in the loop experiments and benefits assessment based on fast-time simulation.

II. Background and Literature Review

Our work on COuNSEL started with a very broad-based concept development in which several alternative conceptual approaches were considered. For example, an initial proposal viewed the problem from the perspective of investments in performance categories and portfolio selection. This concept and others were viewed as inferior to those generally based on the notions of voting and multi-criteria optimization. A key criterion that led to the chosen approach was to seek a mechanism in which the various parties (flight operators) were incentivized to provide inputs consistent with their actual business goals and to not seek to strategically "game" the process.

Voting, in particular, and social choice in general, is concerned with aggregating evaluations over a multitude of voters, in ways such that the final outcome has appeal to a large cross-section of the decision-makers. After investigating various alternatives, we chose the recently developed Majority Judgment (MJ) procedure [10]. It is defined as a social decision function. It involves grading – instead of preference rankings – of each candidate, by all voters, in a common language. It is a natural, rich preference elicitation method, already being practiced in spirit in many contests and juries around the world, as well as a few political elections. It has many good properties; among them, high resistance to strategic voting.

Aside from voting methods, the past work most relevant to this paper is from the literature of multi-criteria decision making (MCDM), especially the group version of it. The decision-making framework in the general MCDM involves a decision maker evaluating a set of candidates on multiple criteria or attributes. [11] categorizes MCDM problems into two streams based on the characteristics of the feasible space: (1) multiple criteria discrete alternative problems where sets of alternatives typically consist of a modest number of choices, e.g., choosing the location for a new airport, selecting a computer network, electing a political leader, and identifying which nuclear power plant to decommission; and (2) multiple criteria optimization problems where feasible sets of alternatives are usually defined by systems of equations and inequalities, e.g., engineering component design, portfolio selection, capital budgeting and R&D project selection. The focus of this paper falls into the second category with additional consideration of group decision making. We note in particular the Analytic Hierarchy Process (AHP), a multi-criteria decision making [12] approach, which has been extended to the group setting, e.g. [13]–[15]. It relies on pairwise comparisons over a set of alternatives, eliciting preference rankings on several criteria organized in a hierarchy on a nine-point scale. The group version of AHP usually consists of the following three approaches: (1) consensus, (2) vote or compromise, and (3) geometric mean of the individuals' judgments [16]. Consensus refers to the achievement of a consensus of group participants in jointly conducting an AHP. If a consensus cannot be reached, the group may then choose to vote on a decision. If a consensus cannot be achieved and the group is unwilling to vote or to compromise, then a geometric average of the individuals' scores can be calculated.

There is perhaps a philosophical difference in our approach to group decision making using MJ and one approach found in the literature on consensus building typically using AHP and/or fuzzy systems. The consensus literature typically assumes that participants do not have precise knowledge of their own preferences or value functions. Preferences and values are quantified through an iterative process. In the group setting there is typically a goal of achieving consistency among the preferences of the various participants. In fact, there is a literature on consistency metrics and consistency-improving processes, e.g. [14], [17], [18]. In our MJ-based approach we address applications where it may be possible to precisely estimate value (grading) functions and we make no attempt to resolve any inconsistencies among such functions. In strategic air traffic management applications, it is possible (at least conceptually) to relate a flight operator's grading function to the expected financial performance that flight operator would receive under candidate vectors in question. In this application, it is also the case that the participants are competitors. So, in many cases, there would be little or no incentives for the participants to cooperate or to seek a consensus. The underlying philosophy is that the participants have agreed to abide by the MJ winner criterion and that the general incentive-compatibility properties of MJ will ensure that participants provide inputs leading to results with the anticipated properties.

A multi-criteria decision analysis based approach was adopted in a strategic decision making context by Eurocontrol [19]. Similar to our setting, the problem involved the ANSP and the airlines collaboratively arriving at a common decision for selecting operational improvements. Further, the decision was subject to constraints like safety and environ-

mental impact, and was expected to improve on objectives like predictability and efficiency. However, unlike our problem that seeks to evaluate at a day-of-operations level, Eurocontrol was faced with a one-time strategic decision. In the parlance of multi-criteria decision making, they were faced with an evaluation decision problem, while we are handling a design decision problem.

III. System Concepts and Operation

The basic mode of operation for COuNSEL is fairly straight-forward, however, it is quite different from the SPTs because its basic output is different. As discussed above, the system seeks to set service level expectations. Another process, not the subject of this paper, takes the further step of converting the service level expectations into planned TMIs. Note that SPTs talk directly in terms of TMIs, e.g. discussing which TMIs should be run and which parameter setting should be used. The SLE problem can be viewed as one of setting constraints or guidelines to be used in determining those TMIs and their parameters. A basic question then is what form should the output of a SLE setting process take.

The TMI planning process is viewed as a design problem that requires performance goals. The output of COuNSEL is a set of such goals. In the follow-on step, traffic management specialists carrying out the design process are faced with decisions that require trading off one performance criterion with another. The performance goals provide the designers with the necessary information to do this. As discussed above, this second step is not discussed in this paper: only the first goal-setting step is addressed.

a) Performance Metrics: Before describing the exact nature of the output, it is worthwhile to consider an appropriate set of performance criteria. The global ATM community working through International Civil Aviation Organization has agreed upon a set of eleven service expectation categories [20]. These were considered carefully in the context of the SLE setting problem and a set of three was identified to be the most relevant to the TMI design and control problem. These are discussed below.

Capacity measures the number of flight operations that the overall system or constituent subsystems can process safely over a specified time-period. In the context of a GDP, an important capacity metric is the number of arrivals that can be accepted by an airport per hour. Capacity is perhaps the most visible and important performance category as it directly relates to flight delays and more generally the ability of an air carrier to maintain its schedule.

Predictability has multiple interpretations depending on the time frame in question. For planning-specific TMIs, predictability refers to the degree to which flight operators know in advance resources available to them and, more generally, the intentions and planned actions of the ANSP. An ANSP could increase predictability by announcing farther in advance its intention to carry out specific TMIs, and giving earlier indications of the ground delays assigned to flights, earlier announcements of the open/closed status of airways, etc.

Efficiency refers to the cost-effectiveness of individual flight operations from the perspective of the flight operator. During a GDP, a policy that leads to high amounts of airborne holding would be less efficient than one that converted that airborne delay into less costly ground delay.

TMI design strategies very often trade off these performance criteria either explicitly or implicitly. For example, one GDP strategy might limit the amount of assigned ground delay, when compared to others. Such a strategy would tend to send a larger number of flights to the airport earlier in hopes that the weather would clear earlier than expected or that a slightly higher than planned acceptance rate could be accommodated. Such a strategy on the average would lead to higher rates of arrival throughput, increasing the capacity metric, but larger amounts of airborne holding, decreasing the efficiency metric [21]. Another strategy might announce and implement early in the day certain TMI actions, such as ground delays and reroutes. These would provide ample time for the flight operators to plan for the day's operations, allowing them, for example, to cancel strategic flights and to take early steps to re-accommodate passengers. On the other hand, such a strategy would likely impose unnecessary ground delays or reroutes. Thus, it would tend to have a higher level of predictability but lower levels of capacity and efficiency.

The SLE setting problem is to provide guidance to Traffic Flow Management specialists on how to trade off TMI performance in the three performance categories given above. The approach chosen to do this involves choosing a specific metric for each of the three categories and specifying a goal for each of those metrics. Thus, the output of COuNSEL is a vector of size three that contains a goal for each of the metrics. The metrics chosen are normalized to be between 0 and 1, with 1 being the best possible value and 0 the worst. One can view a value of 1 as indicating the best performance level for that performance category on a perfect-weather day. Of course, a very simplistic solution to this goal setting problem would be to choose a goal of 1 for each metric. However, a vector of three 1's provides little insight or tradeoff guidance. Rather one should view the process as starting with an assessment of the weather and traffic conditions. This in turn implies constraints on the set of feasible goal vectors. For example, it would generally be the case that on a poor weather day, it would be impossible to achieve a vector of three 1's. In general, the constraints implied by the day's conditions would generate an *efficient frontier* of possible vector values. Conceptually any such vector could be achieved on the day given an appropriate TMI. In fact, the choice between these vectors represents the choice among TMI strategies and provides exactly the tradeoff information that is sought. For example, suppose that the SLE vector was ordered as follows:

(capacity metric, predictability metric, efficiency metric)

Consider the following possible vectors chosen from the efficient frontier:

A: (.95, .90, .91), B: (.90, .94, .89), C: (.97, .87, .89)

Suppose a particular flight operator had a very heavy emphasis on capacity. That flight operator when given the choice

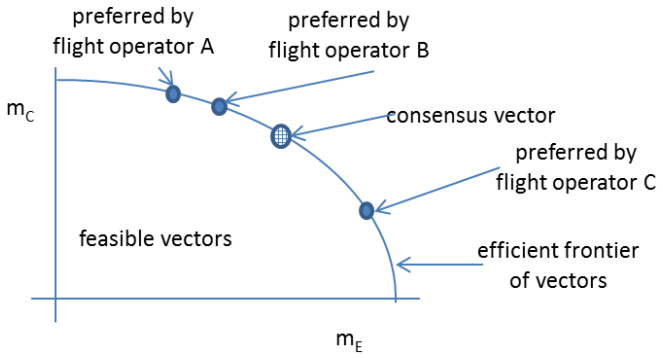

Fig. 1. Efficient Frontier of Performance Metric Vectors.

between A and B might choose A, indicating a willingness to increase capacity and to a less extent efficiency, while sacrificing predictability. That flight operator might further be given the choice between A and C and choose C again in order to increase capacity while further sacrificing predictability and efficiency. In this way, by choosing a particular vector, a flight operator is forced to make key performance tradeoffs.

Given this choice of three performance categories one is still left with the problem of choosing three specific metrics. Obviously, the choice of specific metrics is very fundamental and a key driver to the effectiveness of the system. However, the development and/or choice of metrics is not a focus of the research activity summarized here and so specific metric definitions will not be provided in this paper. Henceforth, it is assumed that metrics for capacity (C), predictability (P) and efficiency (E) have been provided. The output of COuNSEL is a vector of metric values: (m_C, m_P, m_E), where each of m_C, m_P, and m_E is between 0 and 1.

The output vector represents goals for the metric values that the ANSP should seek on the day in question. The "feasible" values for (m_C, m_P, m_E) depend on the conditions of the day so that on poorer weather days, the possible values will tend to be lower (closer to 0) than on better weather days.

b) Feasible Metric Vectors : COuNSEL seeks to generate a consensus among the flight operators. As such, an iterative process is required where each flight operator evaluates and compares possible vectors. Flight operators are also given the opportunity to generate candidate vectors. Fig. 1 illustrates the domain of feasible vectors, flight operator preferences and a consensus vector, in the case where there are two (rather than three) metrics.

On any given day of operations there would be many feasible vectors. However, the flight operators and the ANSP should only consider vectors on the efficient frontier. These dominate the others in the sense that for a vector on the efficient frontier, it is not possible to increase one metric value without decreasing another. Generally, it is the case that each flight operator would have a preferred vector. The consensus vector would tend to represent a compromise among the vectors preferred by each flight operator.

	G1	G2	G3	G4	G5	G6	G7
Vector 1	55	60	76	78	88	90	95
Vector 2	50	59	65	70	70	85	91
Vector 3	60	60	70	75	84	87	89

Fig. 2. Illustration of Majority Judgment.

c) Choosing a Consensus Vector: As discussed in Section II, we have chosen to use Majority Judgment (MJ) as the basis for defining and choosing a consensus vector. It can be viewed as a voting mechanism because it can, and has been, used to run elections in which a single candidate is chosen from a list of several without the need for a runoff election. It more properly is defined as a social decision function. It involves grading – instead of preference rankings – of each candidate, by all voters, using a common language. This viewpoint illustrates its broader usefulness, e.g. to combine the scores from several judges in sporting competitions. It has many good properties; among them, high resistance to strategic voting/grading. In fact, there is a rich theory, which supports its strong properties as a voting and consensus grading mechanism.

In our application, we wish to choose a consensus metric vector. We illustrate the use of MJ by assuming we have a small set of candidate vectors we wish to choose among. Each flight operator is asked to assign a grade to each candidate vector. A grade is a value between 0 and 100, 100 being the best possible and 0 the worst. The flight operators/voters are free to interpret and assign grades as they see fit. However, in concept, grades should vary in proportion to the value, or inverse of cost, that a vector brings to the flight operator. MJ takes the grades given by the voters, and produces *a Majority-grade* of each candidate as an output. The Majority-grade of a candidate is the highest grade *approved by an absolute majority of the voters.* In case of an odd number of voters, it is the median of the grades; if there are even number of voters, then it is the lower middlemost of the grades.

Fig. 2 illustrates the MJ mechanism using three candidate vectors together with the grades assigned by seven flight operators/voters. The assigned grades are ordered from lowest to highest so that the majority grade for each is the one that appears in the fourth column. Note that a given column will usually not correspond to the same flight operator so that a given flight operator might have assigned the grade in column 6 to the 1^{st} vector, the grade in column 2 to the 2^{nd} and the grade in column 4 to the 3^{rd}. Thus, the majority grades for vectors 1, 2 and 3 are 78, 70 and 75 respectively and the winning vector is vector 1. It is possible for ties to occur and there are a set of tie-breaking rules, which will not be discussed here. Our example also assumed that each flight operator had a single vote. In a typical ATM application, it makes sense to allow the voting weight of flight operators to vary based on their number of involved operations or some

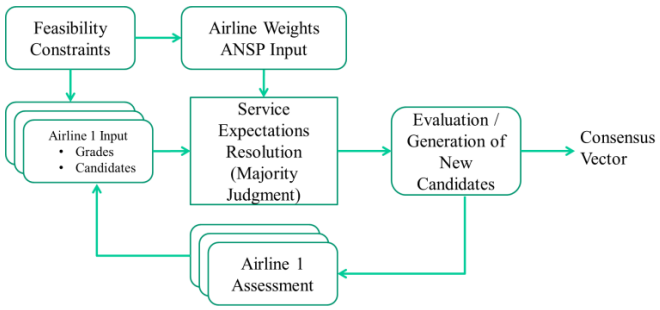

Fig. 3. COuNSEL Architecture.

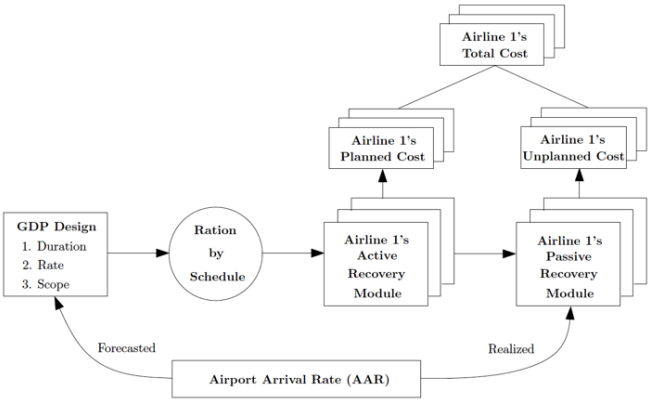

Fig. 4. Benefits Assessment Process.

other metric. Such weights can be viewed as giving each flight operator more or fewer votes to cast. In our experiments so far, weights have increased with the number of operations but the increase is at a less than linear rate.

d) System Architecture and Candidate Generation: The example and discussion of the previous section assumed that MJ was executed over a small set of candidate vectors. This is potentially problematic since, as has been discussed, a candidate vector could be any one drawn from a (continuous) space of feasible vectors, i.e. the set of candidates is in concept infinite. Fig. 3 provides an architecture of the overall system. Note that there is an iterative loop in which candidates are generated, candidates are graded by flight operators and a winner is chosen based on the MJ criterion. The loop might be executed a few times until a final winner is chosen. There are multiple options for candidate generation. A simple one is simply to ask the flight operators for their preferred candidate. In fact, we have developed an underlying theory that automatically generates candidates by estimating flight operator preferences based on their voting history. Underlying this theory is an optimization model that computes the MJ winner over the entire (infinite) set of candidate vectors, given explicit knowledge of the flight operator preference functions [7].

In our context, MJ proceeds by asking flight operators to grade candidate metric vectors until one is found that represents a consensus. The "evaluation" of a metric vector on the part of a flight operator involves assigning a "grade" to the vector. The flight operators will be asked to grade many vectors and given the opportunity to generate new vectors, including providing their most preferred vector. Over time, it is expected that flight operators will develop formal approaches for grading vectors and generating candidate vectors. We foresee eventually automated systems for grading. Automatic generation of candidate vectors by the flight operators is also possible although the system can operate without operator-generated candidate vectors.

Since the COuNSEL process produces the equivalent of a consensus strategic plan on the day-of-operation, it is important for it to have very fast response time. As discussed above, in the long run, it should be the case that all processes both on the ANSP and flight operator sides should be automated so that the entire process, including multiple iterations, should be completed in a matter of minutes, if not seconds. In the initial stages of implementation, there may be human involvement in the grading so that response time could be slower. However, a total response time of less than 30 minutes should be possible.

IV. Benefit Mechanisms and Benefits Assessment

To evaluate the benefits of applying COuNSEL to TMI design compared with the existing approach, we developed an FAA/airline integrated simulation platform to facilitate our analysis. For our benefits assessment, we focus on GDP as a representative example of TMIs. The overall benefits assessment process is depicted in Fig. 4. Note that this fast-time simulation-based framework does not assess the full functionality of COuNSEL: it evaluates a simplified situation where the number of candidates is only a handful. The airlines grade all the candidates in a single round and then the MJ winner is identified. In the following Section V, we will discuss the result of a human-in-the-loop (HITL) experiment with actual flight operators and FAA air traffic controllers where the iterative candidate generation is adopted.

The simulator contains a GDP design module which allows setting appropriate values for different GDP parameters such as planned start time, planned end time, program rate (the arrival rate during the duration of the GDP), and program scope (the set of departure airports affected by the GDP). Once we fix the GDP design, the FAA Ration by Schedule (RBS) module assigns ground delays to impacted inbound flights. Based on this anticipated delay information, each airline runs its recovery module, marked as Active Recovery Module in Fig. 4, to reduce adverse impacts of delays and disruptions through recovery operations such as flight cancellations, flight re-timings, aircraft swaps, etc. This Active Recovery Module employs a mixed-integer linear programming model to compute the optimal recovery operations. We refer readers to [22] for the details and solution methods for the model. Due to intra-day airport capacity uncertainty, these recovery operations may not get executed as planned. The active recovery module is based on the assumption that the capacity forecast at the time of designing the GDP is accurate. However, as the day progresses, the actual realized weather conditions might

be different than initially anticipated, thus rendering some of the planned recovery operations infeasible. For instance, if it turns out that the forecast underestimated the extent of bad weather then some flights would have additional airborne delays beyond those forecasted at the time of GDP design. This may cause some of the aircraft and passenger connections, which were initially deemed feasible in the recovery plan, to get disrupted. In such cases, airlines may require using additional recovery actions to get their schedules back on track. We model this step in the Passive Recovery Module. We call it "passive" because such disruptions caused by inaccurate delay information often require urgent fixes, and hence the airlines usually don't have enough time to come up with a sophisticated new active recovery plan. Such plans need to be simple in nature, and thus can be less effective compared to the recovery plans developed by the Active Recovery Module. Due to this reason, we model the Passive Recovery Module such that it simply propagates all the delay to downstream flights if the aircraft connection in the original recovery plan becomes infeasible due to delays to the previous flight. The Passive Recovery Module also re-accommodates the passengers if their itineraries are disrupted due to these additional unplanned delays to their schedules. The details of the logic in the Passive Recovery Module can also be found in [22] .

Using this evaluation framework described in Fig. 4, we simulate airlines' responses to different GDP designs and evaluate their resultant delay costs. As a case study, we use actual flight schedules for a representative day in the summer of 2007 at San Francisco International Airport (SFO). 10 domestic airlines are involved. We assume that SFO operates at its Visual Flight Rules (VFR) capacity level before the bad weather period starts; operates at its Instrument Flight Rules (IFR) capacity level during the bad weather period; and then it returns to its VFR capacity level once the bad weather clears up. We set up 14 candidate GDPs. The planned duration of each GDP, the difference between planned start and end times, is varied from 3 through 9.5 hours in steps of one half-hour each (i.e., 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, and 9.5 hours). Note that for simplicity, we do not transform these 14 candidate GDP designs into 14 performance metrics since we only conduct a single round of grading of candidates. We model the actual capacity reduction duration as a discrete uniform distribution. Hence, a GDP with a planned duration of 3-hours is considered to be a highly aggressive design in the sense that with high probability, the airport capacity is over-estimated by the GDP design. On the other hand, a GDP with a planned duration of 9.5 hours is highly conservative because with a high probability, capacity is under-estimated by the GDP design. Our goal here is to compare the airport-wide delay costs under the GDP design suggested by COuNSEL with that suggested by the existing approach. Here, existing approach refers to centralized GDP decision-making methods such as [23], and [24], just to name a few. The common objective of these methods is to minimize the sum of expected airborne delay costs and ground delay costs for all the flights heading into the GDP-affected airport. We call this cost the

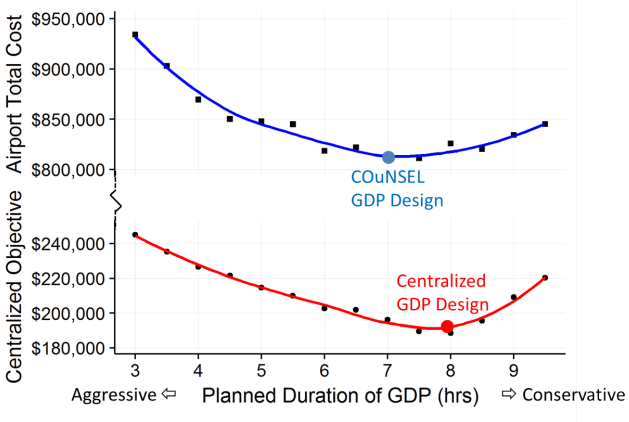

Fig. 5. Airport-wide Total Cost and Centralized Cost Value under 14 Candidate GDP designs.

centralized cost value. Thus, the GDP design out of the 14 candidates with the least centralized cost value serves as our baseline. In this case, it is the one with planned duration of 8 hours (see red line in Fig. 5, where we plot the centralized cost value as a function of the planned duration of the GDP).

To apply COuNSEL in this case study, we first calculate each airline's expected delay costs under all 14 candidate GDP designs using the recovery module. We then use a linear transformation to convert all the airlines' delay costs into grades with a maximum grade of 100 and a minimum grade of 0. The weight of an airline is calculated based on the number of impacted operations it has during the GDP. Since the airline with the largest number of impacted operations (United Airlines) has over half of the total operations, we should not directly use the number of operations as the weight if we want to avoid United Airlines dictating the outcome. Hence we use a power transformation to fix the largest airline's (United Airline's) proportion of total weight at 40%. Suppose o_i is the number of operations of airline i and o_{max} is the number of operations of the largest airline. Under this weighting scheme, we would choose α such that $\frac{o_{max}^{\alpha}}{\sum_{i \in N} o_i^{\alpha}} = 0.4$, where N is the set of airlines involved in the GDP, and o_i^{α} is the weight of airline i which we use in COuNSEL. With these weights and grades, the majority winner GDP design turns out to be the one with planned duration of 7 hours. Interestingly, this coincides with the design which has the least airport-wide total delay cost as shown by the blue line in Fig. 5. The airport-wide total delay cost is calculated by summing over all involved airlines' delay costs. Thus, the GDP design with planned duration of 8 hours is the most preferred design according to the existing centralized approach, with the total airport-wide cost of $825,808. The COuNSEL design, also the one with the lowest airport-wide cost, has a planned duration of 7 hours and with the total cost of $809,297, which is 2.0% less than the centralized design. Note that since each minute of airborne delay is usually much more expensive than each minute of ground delay (roughly in a 3:1 ratio), the centralized decision-making approach is very conservative regardless of what kind

Airlines	*Centralized Design*	*COuNSEL Design*
US Airways	22.46%	15.18%
Frontier	68.16%	52.05%
Northwest	64.92%	62.15%
Continental & ExpressJet	23.24%	19.17%
Delta	12.70%	4.10%
American & American Eagle	13.43%	12.80%
Alaska	12.71%	5.65%
JetBlue	24.67%	1.72%
United & SkyWest	7.27%	10.00%
AirTran	0.00%	7.78%
Mean	24.96%	19.06%
Standard Deviation	22.00%	19.78%

Fig. 6. Percentage Cost Increase over Each Airline's Most Preferred Design.

of airline composition the airport under consideration has. Thus, the delay cost reduction could be even larger if in certain airports at certain periods of time, the majority of the airlines prefer aggressive GDP designs.

An interesting additional analysis is summarized in Fig. 6 that shows how much each airline is worse-off compared to its most preferred design under the centralized design and the COuNSEL design. The average percentage cost increase under the COuNSEL design is 19.06%, which is almost 6% smaller than that under the centralized design. The standard deviation under the COuNSEL design is 19.78%, which is also about 2% smaller than that under the centralized design. These results suggest that by incorporating airlines' preferences, the COuNSEL design also produces more equitable GDP designs in terms of distributing delay costs.

In summary, we find in this specific case study at the SFO airport that the system is better-off in both reducing costs and enhancing inter-airline equity if operated under a slightly more aggressive approach than what the existing centralized approaches would suggest. This benefit gained from the COuNSEL design is due to its ability to incorporate not only the trade-offs between airborne and ground delays, but also most importantly, airlines' diverse preferences over GDP designs due to their different business objectives and operating characteristics.

V. Results of HITL and Implementation Challenges

The COuNSEL system was the subject of a human-in-the-loop (HITL) simulation. The goals of the HITL were to familiarize flight operators and FAA analysts with the system and SLE concepts underlying it, obtain feedback from flight operators on both the general concept and the SLE software, assess the ability of flight operators to provide inputs required for the SLE, and identify high-priority areas for future research.

Held on July 10, 2014, the HITL included participants from American Airlines, United Airlines, Delta, United Parcel Service (representing air cargo operators), the FAA, and the university research team. After several familiarization briefings, there were three HITL sessions, two involving problems at EWR and on at SFO. The day concluded with a structured feedback session. Also, after conclusion of the HITL, FAA and flight operator participants were given a web-based survey to gauge their feedback in a more systematic manner.

A given scenario was specified by a decision time, at which a demand forecast and a weather forecast for the remainder of the day were provided. The demand and weather were such as to warrant a GDP, which was to be planned using COuNSEL. The specific planning process varied from session to session. For example, in the first EWR session, FAA initially submitted candidate SL performance vectors, from which a winner was selected using the Majority Judgement algorithm. Next, each airline submitted two candidates, from which the FAA selected several which, along with the first round winner, were graded in the second round. Finally the winner of that round, along with some new candidates entered by the FAA, were graded in a third round.

A major effort in conducting the HITL was to construct the performance goal tradeoff curves in a plausible manner consistent with real world conditions. Note that the performance goals and associated metrics are expectations over different capacity scenarios. At a high level, we employed the following process:

1) For a given weather forecast, we identified three historical days with a similar forecast and found the arrival capacity profile for each of these days.
2) We then devised 71 possible GDPs which, given, the demand and capacity profiles, would give a broad but reasonable range of performance in terms of throughput, efficiency, and predictability
3) We then assessed the throughput, efficiency, and predictability performance of each GDP under each capacity profile.
4) Finally, assuming that each capacity profile is equally likely, we determined the expected performance of each GDP.
5) We now had 71 performance vectors, from which we constructed continuous tradeoff curves by taking the convex hull.

Airline participants were provided with various aids to assist them in their grading and candidate generation. One such aid was a set of curves showing the tradeoffs between the three performance goals of capacity utilization, efficiency, and predictability. Second, individual airlines were presented with estimates of the expected airborne delay, ground delay, passenger delay, and cancellations from different performance vectors, and estimates of the associated monetary costs. These were based on the benefit models described in Section IV. These aids were provided to compensate for participants' lack of experience with COuNSEL, and also to mimic the proprietary decision support tools that might eventually be developed to support individual flight operator grading and candidate generation.

User feedback was generally encouraging. Airline HITL participants mentioned four valuable and useful features. The first is that COuNSEL would allow each flight operator to specify a corporate policy, established at the executive level, for the relative importance of the difference performance goals. A second, related, benefit is that the system would make explicit the heretofore implicit performance tradeoffs faced in TMI decision making. Third, participants lauded the recognition of predictability as an important performance goal. Last, participants stated that COuNSEL would enable a more systematic and efficient decision making process.

Participants (as well as COuNSEL developers) recognized the challenge of translating the consensus performance vector into a TMI plan. This led to discussion of how COuNSEL might be used without a formal translation step. Possibilities identified included using the system to output TMI parameters directly, or having its output performance goals be used as input to FAA subjective decision making.

If COuNSEL assumed the latter role—systematically finding consensus advice for decisions that would remain subjective—participants noted that it could be used in a broader set of decision contexts than planning GDPs or AFPs. Specific examples included planning transcontinental routes and developing broader regional strategies, for example for the NY Metroplex or South Florida.

Other major outcomes of the post-HITL discussion included:

1) Concern that COuNSEL outcomes for certain airports and types of problems might be controlled by the same coalition of flight operators in each problem instance.
2) Need to recognize the extra burden in using COuNSEL in light of the high workload of airline operations personnel on high delay days.
3) Need to think carefully about how to assign operator weights, including basing them on factors other than the number of flights (for example, also consider to cost of delay for different flights)
4) Concern about cases where several performance vectors have consensus grades that are very close, in which case some other "near-tie-breaking" criteria might be employed by FAA.
5) Concern that other stakeholders, including sub-carriers, general aviation, and airports, be given an appropriate role in the process.

The post-HITL survey included four sets of closed-ended questions. While administered to both flight operator and FAA participants, we focus here on the results for the former group (n=4). The first questions concerned the importance of different COuNSEL features. Flight operator participants considered the most important features to be that "TMI decisions are tied explicitly to performance vectors" and that "Flight operators provide structured input to the TMI planning process." The least important feature was that "Less time and effort are required for the TMI planning telecons." From the second set of questions, which gauged participant agreement with a set of statements about COuNSEL, we learned that participants generally agreed that the performance goals used in COuNSEL are appropriate. The third question set asked participants to compare COuNSEL with the current TMI planning process. Respondents believed that COuNSEL was between somewhat better and much better in most respects. The final set of questions pertained to the COuNSEL software. From these questions we learned that participants considered it easy to submit grades and candidate goal vectors, but were less enthusiastic about the clarity with which information was presented and how the various screens and layouts were structured.

VI. Conclusions and Next Steps

We recognize that COuNSEL represents a somewhat radical departure from existing approaches to TMI planning. It provides a solution to a problem (service level expectation setting) that today is not solved or is solved implicitly or informally. Yet, the explicit definition of this problem as an important step in future TMI planning is clearly laid out in NextGen documents. Moreover, our discussions with experts and the HITL results clearly show that flight operator and ANSP traffic management specialists recognize that TMI design requires performance tradeoffs to be made. The implication is that these tradeoffs today are being made in an ad hoc, implicit or subjective manner. The results of this paper show that COuNSEL provides an effective solution to the service level expectation setting problem and it can provide measurable benefits to flight operators.

Additional work is required before COuNSEL could be used in practice. First, TMI planning tools are required to make use of the output from COuNSEL. The underlying problem to be solved can be called performance-based TMI planning. In fact, there is on-going research in this general area [9]. It is also the case that it is conceptually possible to develop heuristic rules or high-level tools to convert the COuNSEL performance goals into parameter settings for existing TMI planning tools. There is other research on COuNSEL itself and the auxiliary processes required to make it work in practice. More work is required to assist the flight operators in providing COuNSEL inputs. For the HITL, certain tools and information resources were generated. These must be refined based on feedback from the HITL, better modeling of airline operations and monitoring of flight operator use. It is important to further investigate the potential for "gaming". For example, it seems important to prohibit flight operator collusion. More broadly there needs to be a set of rules governing flight operator use of COuNSEL. Otherwise, further refinements of various COuNSEL components certainly should be pursued.

Acknowledgment

This work was supported in part by the FAA through the NEXTOR-II Consortium.

References

[1] M. Wambsganss, "Collaborative decision making through dynamic information transfer," *Air Traffic Control Quarterly*, vol. 4, no. 2, pp. 109–125, 1996.

[2] M. O. Ball, R. L. Hoffman, D. Knorr, J. Wetherly, and M. Wambsganss, "Assessing the benefits of collaborative decision making in air traffic management," *Progress In Astronautics and Aeronautics*, vol. 193, pp. 239–252, 2001.
[3] K. Chang, K. Howard, R. Oiesen, L. Shisler, M. Tanino, and M. C. Wambsganss, "Enhancements to the faa ground-delay program under collaborative decision making," *Interfaces*, vol. 31, no. 1, pp. 57–76, 2001.
[4] E. G. Modrego, M. Igaru, M. Dalichampt, and R. Lane, "Airport CDM network impact assessment," in *Proceedings of the Eighth USA/Europe Air Traffic Management Research and Development Seminar*, 2009.
[5] Joint Planning and Development Office, "Concept of operations for the next generation air transportation system, v. 2.0," Tech. Rep., 2007.
[6] P. Swaroop and M. Ball, "Consensus building mechanism for setting service expectations in air traffic management," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2325, pp. 87–96, 2012.
[7] C. Yan, P. Swaroop, M. O. Ball, C. Barnhart, and V. Vaze, "Applying majority judgment over a polyhedral candidate space," 2016, available at SSRN: https://ssrn.com/abstract=2746568.
[8] Y. Liu, M. Seelhorst, A. Pozdnukhov, M. Hansen, and M. O. Ball, "Assessing terminal weather forecast similarity for strategic air traffic management," in *International Conference on Research in Air Transportation*, 2014.
[9] L. Kang, Y. Liu, R. Hoffman, and M. Hansen, "Ground delay program decision-making based on utility maximization," 2017, submitted to Transportation Research Part E: Logistics and Transportation Review.
[10] M. L. Balinski and R. Laraki, *Majority judgment: measuring, ranking, and electing*. MIT Press, 2010.
[11] J. Wallenius, J. S. Dyer, P. C. Fishburn, R. E. Steuer, S. Zionts, and K. Deb, "Multiple criteria decision making, multiattribute utility theory: recent accomplishments and what lies ahead," *Management science*, vol. 54, no. 7, pp. 1336–1349, 2008.
[12] T. L. Saaty and L. G. Vargas, *Models, methods, concepts & applications of the analytic hierarchy process*. Springer Science & Business Media, 2012, vol. 175.
[13] J. Aczel and T. Saaty, "Procedures for synthesizing ratio judgements," *Journal of Mathematical Psychology*, vol. 27, pp. 93–102, 1983.
[14] Z. Wu and J. Xu, "A consistency and consensus model for group decision making with multiplicative preference relations," *Decision Support Systems*, vol. 52, pp. 757–767, 2012.
[15] N. Bryson, "Group decision-making and the analytic hierarchy process: exploring the consensus-relevant information content," *Computers & Operations Research*, vol. 23, pp. 27–35, 1996.
[16] R. F. Dyer and E. H. Forman, "Group decision support with the analytic hierarchy process," *Decision support systems*, vol. 8, no. 2, pp. 99–124, 1992.
[17] Z. Xu and X. Cai, "Group consensus algorithms based on preferences relations," *Information Sciences*, vol. 181, pp. 150–162, 2011.
[18] E. Herrera-Viedma, L. Martinez, F. Mata, and F. Chiclana, "A consensus support system model for group decision-making problems with multigranular linguistic preference relations," *IEEE Transactions on Fuzzy Systems*, vol. 13, pp. 644–658, 2005.
[19] Y. Grushka-Cockayne, B. D. Reyck, and Z. Degraeve, "An integrated decision-making approach for improving european air traffic management," *Management Science*, vol. 54, no. 8, pp. 1395–1409, 2008.
[20] International Civil Aviation Organization, "Global air traffic management operational concept," in *Document 9854*, 2005.
[21] Y. Liu and M. Hansen, "Ground delay program decision-making using multiple criteria: a single airport case," in *10th USA/Europe Air Traffic Management R&D Seminar*, 2013.
[22] C. Yan, V. Vaze, and C. Barnhart, "Airline-driven ground delay programs: A benefits assessment," 2017, working Paper.
[23] O. Richetta and A. R. Odoni, "Solving optimally the static ground-holding policy problem in air traffic control," *Transportation Science*, vol. 27, no. 3, pp. 228–238, 1993.
[24] A. Mukherjee and M. Hansen, "A dynamic stochastic model for the single airport ground holding problem," *Transportation Science*, vol. 41, no. 4, pp. 444–456, 2007.

Michael Ball holds the Dean's Chair in Management Science in the Robert H. Smith School of Business at the University of Maryland. He also has a joint appointment within the Institute for Systems Research in the Clark School of Engineering and is co-Director of NEXTOR-II, an FAA consortium in aviation operations research. Dr. Ball received his Ph.D. in Operations Research in 1977 from Cornell University.

Cynthia Barnhart is the Ford Professor of Engineering and Chancellor at the Massachusetts Institute of Technology. Her research interests include applications of optimization to networked systems, particularly transportation systems. She is a member of the National Academy of Engineering.

Mark Hansen is a Professor of Civil and Environmental Engineering at UC Berkeley, and a co-director of the National Center of Excellence for Aviation Operations Research.

Lei Kang is a Ph.D. candidate of the Institute of Transportation Studies at University of California, Berkeley. He is currently a member of the Committee on Airfield and Airspace Capacity and Delay, Transportation Research Board. His research interests are in the application of statistical methods and machine learning techniques to air traffic management and airline fuel loading decisions.

Yi Liu is a research scientist at Amazon. Her research interests are in air traffic management, air transportation system performance, airline economics, machine learning and data-driven decision making. Dr. Liu received her Ph.D. in transportation in 2016 from University of California, Berkeley.

Prem Swaroop is Principal Data Scientist at Altisource. His research and practical endeavors are at the cross-section of large-scale machine learning and optimization. Dr. Swaroop received his Ph.D. from the University of Maryland, with joint appointments at the department of Operations Management / Management Science in the Robert H. Smith School of Business, and the Institute for Systems Research in the Clark School of Engineering.

Vikrant Vaze is an Assistant Professor at the Thayer School of Engineering at Dartmouth College. His research focuses on improving the planning, management and operations of large-scale, multi-stakeholder systems such as transportation and healthcare using game theory, optimization and data analytics. He received an MS in Transportation, another MS in Operations Research and a PhD in Systems, all from the Massachusetts Institute of Technology.

Chiwei Yan is currently a PhD candidate in Operations Research at Massachusetts Institute of Technology. His research interests are in large-scale optimization, robust optimization with applications in air transportation, urban transportation and logistics systems and revenue management. His research is recognized by several awards including a First Prize in the Anna Valicek Best Paper Award from AGIFORS, a Best Presentation Award from INFORMS Aviation Application Section and First Prize in the Problems Solving Competition from INFORMS Railway Application Section.