Distributed Mechanisms for Determining NAS-Wide Service Level Expectations: Final Report

By

Michael Ball
Cynthia Barnhart
Mark Hansen
Lei Kang
Yi Liu
Prem Swaroop
Vikrant Vaze
Chiwei Yan

October 31, 2014

During the final year of the Service Level Expectation (SLE) project the NEXTOR-II team refined the models, completed development of the concept evaluation software and carried out a variety of user outreach activities. Specifically, much effort was devoted to supporting a human-in-the-loop (HITL) simulation. In addition to completing the software, the team produced a variety of tools to support the HITL participants. The team also separately reached out to flight operators to obtain both formal and informal feedback on system concepts and mechanisms.

We now provide an overview of the basic SLE components and both give some perspectives and updates on them and also indicate where appropriate detail and background information can be found. In many cases that background information exists in earlier project reports. Specifically, in this report we refer to the following reports for more project details: the Year 2 Project Report, the Year 3 Project Report, the HITL

Report and the Final Project Presentation, which is being delivered with this report. Also, please note the five Appendices of this report.

First, it should be noted that an intuitive description of all the concepts can be found in the SLE project white paper given in Chapter 1 of the year 3 report.

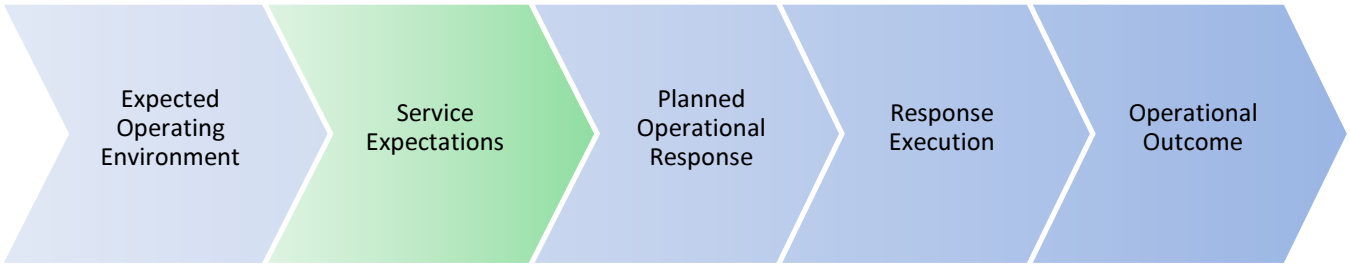


Figure 1: NextGen Operational Response Architecture

The mechanism generated by the SLE project is called COuNSEL: CONsensus Service Expectation Level setting. COuNSEL provides a solution to the Service Expectations step of the NextGen traffic management initiative (TMI) planning process given in Figure 1. We note that under NextGen operational concepts, the service expectations defined in this step should represent the consensus input of the flight operators. In fact, this is the primary mechanism for flight operators to provide strategic input into the planning of an operator response.

Under the COuNSEL architecture the specific output of the consensus service expectation process is a vector of performance metric goals. COuNSEL specifically has employed three performance categories: capacity, predictability and efficiency. Various prior documents, including the white paper, define the specific metrics used. 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 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.

This discussion immediately reveals two fundamental problems to be solved. First is defining the set of constraints that represents the feasible space of performance goal vectors for a given day/environment. Second, given this space of feasible vectors how does one define a consensus vector and what process should be used to find such a vector. The COuNSEL solution to these problem and the underlying research are discussed respectively in Sections 2 and 1.

COuNSEL could be applied in a number of different contexts. For example, it could potentially be applied in formulating a NAS-wide strategy for an entire day. Alternatively it might be applied to solve a specific regional problem, e.g. it could be applied to develop a strategy for a specific ground delay program (GDP). In each case there would be a certain set of impacted flights and flight operators. Since each such flight operator will be impacted by the resulting TMI to a different degree it makes sense that the impacted flight operators should have varying levels of influence in the ultimate COuNSEL recommendation. A weight is assigned to each flight operator to accomplish this and the topic of flight operator weights is discussed in Section 3. The related topic of the COuNSEL application context is treated in Section 4. A set of models was developed and experiments run to understand the relationship between the COuNSEL decisions and user costs. These provided the basis for both a benefits assessment and tools to support flight operator inputs into COuNSEL. This work is described in Section 5. The insights gained from various user outreach efforts is discussed in Section 6. COuNSEL is driven by the Majority Judgment voting mechanism. For COuNSEL to work well it is important for the users to vote “truthfully”. The topic of user voting behavior and user incentives is discussed in Section 7. Section 8 provides background on the concept evaluation software. Section 9 covers practical issues regarding next project steps and Section 10 discusses potential applications for the SLE concepts in air traffic management that do now fall within the specific architecture illustrated in Figure 1.

1. Basic Voting Mechanism and Handling a Very Large Space of Candidates

A fundamental question to ask is what is the definition of a consensus vector. The theory that underlies COuNSEL is the Majority Judgment voting procedure. This procedure has been developed and analyzed over the past several years. It can be used in a normal political election and specifically give a good solution to the challenge of picking a single winner among several competitive candidates (without the need for a runoff election). It also can be used in other ways, e.g. to judge athletic competitions. Its virtue lies in its resistance to "gaming": it generally encourages participates to vote in a straight-forward/truthful manner. Background on the method can be found in the book, Balinski and Laraki, 2011, *Majority Judgment: Measuring, Ranking, Electing*, MIT Press. A more comprehensive discussion of why Majority Judgment provides a good approach to defining and finding a consensus vector can be found Appendix I of the Year 3 Report.

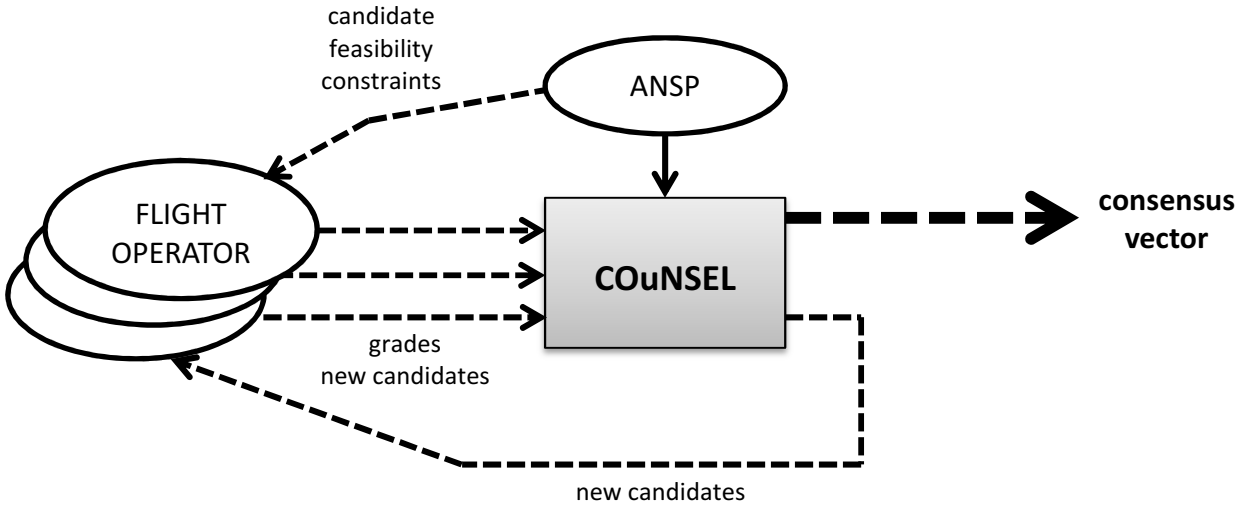


Figure 2: COuNSEL Architecture

It is the case, however, that Majority Judgment cannot be directly applied to the SLE problem. Specifically, there is a very large number (in fact infinite number) of candidates. A significant component of the SLE research involved developing an underlying theory and computational methods to deal with this challenge. Specifically, as Figure 2 illustrates, an iterative approach is used where candidate vectors are dynamically generated and multiple rounds of voting are employed. The theory and methods associated with this approach are described in Appendix I of the Year 3 project report.

2. Feasible Region of Performance Metric Goal Vectors

As discussed in the introduction, one would always want a goal vector of (1,1,1). However, on virtually all days, such perfect performance is not possible usually due to some instances of less than perfect weather. Other factors can also impact the feasible goal set including flight demand irregularities, equipment or infrastructure failures and the like. Thus the range of feasible goal vectors will depend on the conditions of the day and also the degree to which TMI parameters and airlines actions allow different performance criteria to be traded off. For example, GDPs can be planned so as to insure higher throughput/capacity by delaying start times, setting higher rates, etc. Other strategies might insure more predictability or efficiency. The SLE team has explored both analytic models and statistical analysis of historical TMIs to construct performance goal vector tradeoff spaces and the associated feasible region of goal vectors. Analytic models are described in Appendix III of the Year 3 report and the statistical approach is described under Topic 4 of the Final Project Presentation and Appendix I, II and III of this report.

3. Flight Operator Weights

Each application of COuNSEL will have an associated scope limiting the impacted flights and/or geographic region. At one extreme COuNSEL could be applied to generate strategic advice for planning a ground delay program into a particular airport. Such a GDP would have an associated start and end time and an associated destination airport. Thus, the impacted flights would include all of those flights to the designated destination airport whose expected arrival times were between the start and end time. At another extreme a NAS wide strategy could be sought so the set of impacted flights could be as large as all flights scheduled to pass through the US airspace within a particular 24 hour period.

In any such case, there will be a disparity of impact on each flight operator. For example, in the GDP case, a flight operator with a large presence at an airport could have a hundred or more impacted flights while another might have only a handful. In such cases, it seems reasonable and fair that the operator with the larger number of flights should have more influence over the strategy chosen. Flight operator influence can be varied by assigning unequal weights to each flight operator. These weights can be viewed as allowing each participant to have a number of "votes" greater than one. For example, flight operators designated for small influence could be given a weight of 1, others higher weights depending on the degree influence desired. The percentage of the total weight assigned to a flight operator would indicate the degree of influence. In particular, if a flight operator had more than 50% of the total weight then that flight operator would be able to insure its most preferred goal vector was the outcome of the COuNSEL process. Of course, most would consider this undesirable so a weighting scheme that assigned one operator a majority of weight would most likely not be considered.

The SLE team tested various approaches to assigning weights. All based the weights on some function of the number of impacted flights associated with each flight operator (so that more flights implied higher weight). The simplest scheme would assign weights in proportion to the number of flights – this could lead to one flight operator receiving a majority of the total weight or close to such a majority. Consequently, certain transformations were applied, e.g. functions that depended on the square root or log of the number of flights. While the team experimented with several ideas (see Appendix I of the Year 3 Report) this is certainly a topic that deserves further study. In particular, as discussed in the HITL report, there could be justification in considering weighting functions that do not depend on the number of impacted flights.

4. COuNSEL Application Context

As discussed in the previous section, there would always be a context and associated scope to which COuNSEL would be applied. The obvious context, which is consistent with today's TMI planning environment, would be to apply COuNSEL whenever a TMI is being considered. In this case, the broad parameters of the TMI would already be determined, e.g. consideration of a morning GDP for SFO airport. The output of COuNSEL would provide the FAA specialist planning the TMI with guidance on how to set the TMI parameters and could even influence the decision on whether to initiate the TMI at all. The longer term vision for COuNSEL is much broader. It is hoped that it would be able to play a role in setting a daily national or regional strategy for managing traffic.

The SLE team did some initial research on how COuNSEL might be applied in a hierarchical manner. That is, a natural vision for the application of COuNSEL would be to do hierarchical planning where COuNSEL would be applied to determine a NAS-wide plan and then regionally to plan specific regional or local TMIs. An overview of this work is included in Appendix IV of this report (this was provided in a previous project deliverable but was not included in a final report so it is included here).

5. Benefits Assessment and User Support Tools

The SLE developed several analytic and statistical models of flight operator behavior. The general goal of these models was to relate the three SLE performance metrics to GDP parameter settings and to flight-operator-specific performance. This work looked at historical flight operator actions and historical flight operator performance in the presence of GDPs. It also computed the SLE performance metrics for those GDPs. This allowed a relationship between the COuNSEL goal vector values and flight operator performance to be developed. Background on this research can be found in Appendix II of the Year 3 Report and also in Topic V of the Final Presentation. This work drove both a COuNSEL benefits assessments as well as the development of information and tools to support flight operator voting. This latter information was used to support the HITL.

6. User Acceptance and Practical Aspects

During the later stages of the project, the SLE team undertook efforts to reach out to the flight operator community. A presentation was made to the A4A ATC Council. Follow up meetings were held with certain flight operators and a survey was administered. Most recently in the summer of 2014, an HITL was held that included flight operator participation. The FAA command center also participated in the HITL. These various activities produced a significant body of material that provided feedback from the user community on COuNSEL. This feedback is analyzed and summarized in the HITL Report.

7. User Voting Behavior

The effective use of COuNSEL and more generally Majority Judgment requires that the users grade/vote in a truthful and consistent manner. Specifically, a key aspect of Majority Judgment relative to other voting methods is that by grading candidates rather than simply voting yes or no, the users provide "rich" information so that multiple candidates can be evaluated and compared. However, if users do not grade truthfully or consistently, then such rich information is not provided. For example, if a user just gave a high grade to one candidate and a zero to the others then the system would not work well. At the same time, a very significant appeal of Majority Judgment is that it is resistant to "strategic voting" and generally users are incentivized to grade in a manner consistent with their true valuation of the candidates. Of course, the particular context provided by COuNSEL has certain unique features so that the general properties of Majority Judgment might not apply well in this case. With this in mind the SLE team carried out various analyses to determine whether in fact flight operators using COuNSEL would be incentivized to grade in a manner consistent with their values. The main body of work in this area is provided in Appendix IV of the Year 3 Report. This work shows that it is very difficult for users to realize any gain by not grading in a way consistent with their internal valuations. The results of the HITL also provide insight on this issue. This is certainly an area that deserves additional research and investigation. Such research should involve simulations with real airline cost functions and also further human-in-the-loop simulations.

8. Concept Evaluation Software

Concept evaluation software was created in order to evaluate the various COuNSEL features and also to gain feedback from potential users. The software played a central role in the HITL. The software supports all the principal COuNSEL features as illustrated in

Figure 2. There are two types of users: the ANSP/FAA (one user) and the flight operator (multiple users).

The ANSP initiates any COuNSEL session (referred to as a poll). The key inputs required are: i) a list of flight operators, ii) flight operator weights and iii) a set of constraints defining the space of feasible performance vectors.

There are two flight operator functions: i) vector generation and ii) vector grading.

To start any iteration (execution of the loop illustrated in Figure 2), the ANSP may input a set of candidate vectors and then present these to the flight operators for grading. As an option, the ANSP could request that the flight operators input candidate vectors. These could augment any ANSP supplied vectors or serve as the only source of vectors. In either case, a set of candidate vectors is provided to the flight operators who then grade each candidate.

Once all grades are provided, the system computes the majority grade for each vector and determines the winner. The ANSP then has the option of either declaring the iteration winner the overall winner or starting a new iteration.

Instructions on the use of the software may be found both in the HITL report and also under Topic 6 of the Final Presentation.

It should be noted that this software does not contain all the capabilities developed by the SLE research team. Specifically, it does not contain any of the automatic vector generation models. It also does not contain methods for generating the constraints defining the feasible region of performance goal vectors. In this way, it retains a certain degree of flexibility and allows various research concepts to be evaluated.

9. Implementation Going Forward

The research carried out by the SLE team thus far has certainly given a proof of concept for the basic approach. It is also the case that the flight operator reaction has been quite positive. Specifically, the user community supports the basic premise of COuNSEL, namely that alternate TMI strategies can be characterized in terms of tradeoffs among performance goals. Furthermore, today, FAA specialists and airline operational control personnel, make these tradeoffs (sometimes implicitly) in formulating their plans. Of course, this does not mean that many of COuNSEL's details, including precise metric definitions, grade formats and the like do not need to be further evaluated. It is also the case, that the flight operators would need to develop grading strategies and internal support tools to make any system usable.

There are many software and concept refinements that would be required before COuNSEL be transitioned into use in an operational setting. No attempt will be made to

create an exhaustive list. However, the list below provides the most significant challenges the SLE team could identify at this time.

1) **Performance-Based TMI Planning:** Going back to Figure 1, it should be noted that COuNSEL provides a solution to a new step in the TMI planning/operational response architecture. In fact, the COuNSEL output does not directly provide a TMI plan or TMI parameters. The vector output can be viewed as strategic advice for use in TMI planning. Making use of this vector requires a performance-based TMI planning model. For example, it could be that the various tools used for GDP planning, including FSM (Flight Schedule Monitor, the GDP planning tool) might be modified to accept as input the vector output by COuNSEL and to use this in setting GDP parameters. There is currently some on-going NASA-sponsored research that may provide a solution to the general challenge of performance-based TMI planning. Of course, it is also the case that "more informal" solutions are possible. For example, it is possible that tables or policies could be created that converted vectors output by COuNSEL into strategic advice to be used by the traffic management specialist in creating the TMI plan in question. Appendix V of this report provides some concepts and models performance-based GDP design.
2) **Generating Constraints Defining Feasible Space of Performance Vectors:** In concept, the application of COuNSEL requires that the weather and demand conditions on a particular day-of-operations be converted into the set of constraints defining feasible performance goal vectors. The research to date provides analytic models that defines these constraints for a generic GDP and also describes an approach to the general problem. However, much more work is required to operationalize these ideas. It is also the case that if the COuNSEL approach was used in a broader context, e.g. to plan an overall NAS strategy, then even more novel ideas would be required. It may be that simple approximate models are possible that do not depend too heavily on the conditions of the day. Such models could potentially provide a near-term solution that could be put in place quickly, e.g. in a prototype setting.
3) **Vector Generation:** The SLE research provides a well-developed theory related to vector generation. However, there is still work to be done to apply this efficiently in various practical settings. It should be noted that experience with the HITL indicates that various approximate/practical approaches may work quite well, particularly in initial implementations. For example, a set of vectors could be generated ahead of time that provide broad coverage of the feasible region. Heuristic rules could be developed to choose from among these and vectors generated by flight operators to provide a supply of candidate vectors at each COuNSEL iteration.
4) **User Acceptance:**
   As discussed there is general flight operator acceptance of the concepts and approach underlying COuNSEL. However, much work will be required to obtain broader user acceptance most particularly of a specific implementation with specific operational concept and decision support tool. A key aspect of obtaining

such acceptance will be demonstrating value added provided to the flight operator community.

10. Other Uses

COuNSEL was developed to address a very specific function in the NextGen architecture given for generating an operational response. In this context, COuNSEL very specifically generates a performance goal vector to be input to a performance-based TMI planning model. However, viewed more broadly COuNSEL provides a mechanism for generating a consensus numeric vector that could potentially represent a range of planning decisions. For example, in another air traffic management (ATM) context, it might be reasonable to consider an architecture where the vector generated by COuNSEL represented specific parameters of ATM decision. This could be relevant in any case where an ANSP might wish to develop a plan based on consensus advice from flight operators. For example, one could imagine that the vector generated by COuNSEL directly represented GDP plan parameters, specifications related to a runway configuration change or parameters related to a strategic reroute plan. In these cases, the output of COuNSEL would more directly represent a planning decision and would not represent a strategic input into a second planning model. These ideas were discussed briefly during the HITL, but, of course, much further investigation would be required to make them a reality.

APPENDIX I: Generating Constraints to Define Feasible Space of Performance Vectors from a Set of Candidate Vectors

Piecewise linear approximation of a concave efficient frontier from given set of feasible points

Background

In a complex decision scenario, decision makers make expensive functional evaluations for several feasible data points. An efficient frontier over these functional evaluations is desired for a deeper understanding of the decision domain. For instance, insights may be gained by extrapolation over the regions that were not directly evaluated by the decision makers. These may lead to more functional evaluations, leading to an iterative learning process.

The efficient frontier can also be used to represent the feasible region in mathematical programs that may optimize some objective function. Piecewise linear inequalities that approximate the efficient frontier are thus of interest.

Problem

Without loss of generality, we assume that the efficient frontier is concave (similar to "output-based" Data Envelopment Analysis formulations), and all data points are non-negative. Suppose there are n points in the m dimensional non-negative real space $\Re_+^m$, represented in an $n \times m$ matrix $\mathbf{X}$. Any of these dimensions can represent the functional evaluation ("output"), while the others represent the underlying decision space ("input").

We seek to determine the set of inequalities $\mathbf{A}x \leq \mathbf{b}$ that define the concave efficient frontier over $\mathbf{X}$, where $\mathbf{A}$ is $k \times m$ matrix of coefficients, x is $m \times 1$ column vector $(x_1, x_2, \dots, x_m)^T$, and $\mathbf{b}$ is $k \times 1$ column vector of Right Hand Side (RHS) coefficients. The concavity and non-negativity assumptions imply that all the coefficients in $\mathbf{A}$ and $\mathbf{b}$ are non-negative.

Procedure

The concave efficient frontier is clearly a convex hull. However, the original set of data points in $\mathbf{X}$ have to be augmented to satisfy the concavity and non-negativity assumptions. Once the augmented set, denoted $\overline{\mathbf{X}}$, is determined, a half-space representation can be obtained using standard methods and software.

Phase 1: augmenting the original data

Three types of points need to be added.

Step 1: add the origin, namely (0,0,..) , or the set of points describing the smallest levels on each dimension.

Step 2: add the projects of maximum levels of each dimension on all the other dimensions. That is, add points like $(x_1^*, 0,0, \dots)$, $(0, x_2^*, 0 \dots)$ etc, where x_i^* is the maximum level for i-th dimension.

Step 3: add projections of each level of each dimension of each point on all the other dimensions. That is, add points like $\left(x_1^j, 0, x_3^j, \dots\right)$, $(x_1^j, x_2^j, 0, \dots)$ etc, where $\left(x_1^j, x_2^j, x_3^j, \dots\right)$ is the j-th point in the dataset.

Note that the third step needs to be done only for the vertices found after the initial two steps are taken to augment the original dataset. However, the resulting approximation is equivalent: the inequalities defined by non-vertex points will essentially be dominated by those at the vertex during the convex hull half-space representation phase. This is reflected in the following.

Alternate phase 1a: augmenting the original data

Step 1a: same as above Step 1

Step 2a: same as above Step 2

Step 3a: determine the vertex points from the augmented dataset so far, using standard software, one such software implementation is explained below.

Step 4a: add projections of each level of each dimension of each vertex on all the other dimensions. That is, add points like $(x_1^v, 0, x_3^v, \dots)$, $(x_1^v, x_2^v, 0, \dots)$ etc, where $(x_1^v, x_2^v, x_3^v, \dots)$ is the v-th vertex point in the augmented dataset.

The original Phase 1 is simpler to execute, as it avoids an interim step of finding the vertices. However, the following Phase 2 may be take longer than the alternate Phase 1a due to excessive points. This may be a concern for larger datasets.

Phase 2: generating the half-space representation over the augmented dataset

This can be done using standard software, see the following for one such implementation.

Using software to execute the procedure

We explain how qhull (freely available from www.qhull.org) can be used for the two tasks.

First, an input file, say input.txt has to be generated. Its first line is the number of dimensions; second line is number of points; followed by each point delimited with whitespace. Eg, for five points in three dimensions, the following input file has to be generated:

```
3
5
0.6077181 0.8483771 0.8937495
0.5901854 0.8812403 0.9323248
0.6933129 0.8370321 0.8420531
0.5197341 0.8546559 0.8597732
0.6639922 0.8170262 0.9143827
```

Following the Phase 1, this dataset will need to be augmented as below:

```
3
24
0.6077181 0.8483771 0.8937495
0.5901854 0.8812403 0.9323248
0.6933129 0.8370321 0.8420531
0.5197341 0.8546559 0.8597732
0.6639922 0.8170262 0.9143827
0 0 0
0.6933129 0 0
0 0.8812403 0
0 0 0.9323248
0.6077181 0.8483771 0
0.5901854 0.8812403 0
0.6933129 0.8370321 0
0.5197341 0.8546559 0
0.6639922 0.8170262 0
0.6077181 0 0.8937495
0.5901854 0 0.9323248
0.6933129 0 0.8420531
0.5197341 0 0.8597732
0.6639922 0 0.9143827
0 0.8483771 0.8937495
0 0.8812403 0.9323248
0 0.8370321 0.8420531
0 0.8546559 0.8597732
0 0.8170262 0.9143827
```

In the alternate Phase 1a, the dataset at the end of Step 2a would be:

```
3
9
0.6077181 0.8483771 0.8937495
0.5901854 0.8812403 0.9323248
0.6933129 0.8370321 0.8420531
0.5197341 0.8546559 0.8597732
0.6639922 0.8170262 0.9143827
0 0 0
0.6933129 0 0
0 0.8812403 0
0 0 0.9323248
```

Running the following command will result in vertices:

```
qhull Fx < input.txt
```

Its output is:

```
8
1
2
```

```
3
4
5
6
7
8
```

The first line shows the number of vertices, followed by the line numbers of points in the input file – starting from 0 – that form the vertices. In this example, all points except the very first one (0.6077181,0.8483771,0.8937495) – which would have been marked 0 – are vertices. The Step 4a would only use these eight points, and use this augmented set of 20 points as input for Phase 2:

```
3
20
0.5901854 0.8812403 0.9323248
0.6933129 0.8370321 0.8420531
0.5197341 0.8546559 0.8597732
0.6639922 0.8170262 0.9143827
0 0 0
0.6933129 0 0
0 0.8812403 0
0 0 0.9323248
0.5901854 0.8812403 0
0.6933129 0.8370321 0
0.5197341 0.8546559 0
0.6639922 0.8170262 0
0.5901854 0 0.9323248
0.6933129 0 0.8420531
0.5197341 0 0.8597732
0.6639922 0 0.9143827
0 0.8812403 0.9323248
0 0.8370321 0.8420531
0 0.8546559 0.8597732
0 0.8170262 0.9143827
```

Note that original Phase 1 had 24 points, while the alternate Phase 1a has only 20 points. In this toy example, this would not make any difference in Phase 2, but in very large settings, this may start to matter. However, the vertex determination step could also take some time.

Phase 2 of generating half-space representation uses a different parameter to the same program:

```
qhull n < input.txt
```

The output is as below:

```
4
10
0.6487738823205065 0.6239839949833628 0.4355874465860791 -1.338885695202095
0.9267486846573778      0 0.3756818806993756 -0.9588709103877315
    -0     -1     -0      0
0.236216022534486      0 0.9717005663773112 -1.045351783953538
    -0      1     -0 -0.8812403
     0     -0      1 -0.9323248
    -1     -0     -0      0
     0      0     -1     -0
     1     -0     -0 -0.6933129
0.3939998618229221 0.919110498734248     -0 -1.042490177687624
```

The first line is the number of coefficients (4), which is one more than the number of dimensions (3). Next line states the number of half-spaces (10). The following lines give the coefficients in the form of $\mathbf{A}x - \mathbf{b} \leq 0$. Thus, the last column has to be multiplied by -1 to obtain $\mathbf{b}$.

Three of these are just non-negativity constraints:

```
-0     -1     -0      0
-1     -0     -0      0
 0      0     -1     -0
```

Further, three are upper bounds for each dimension:

```
-0      1     -0 -0.8812403
 0     -0      1 -0.9323248
 1     -0     -0 -0.6933129
```

The remaining ones clearly have non-negative coefficients (the last column is negative, as explained above).

```
0.6487738823205065    0.6239839949833628    0.4355874465860791    -1.338885695202095
0.9267486846573778    0                     0.3756818806993756    -0.9588709103877315
0.236216022534486     0                     0.9717005663773112    -1.045351783953538
0.3939998618229221    0.919110498734248     -0                    -1.042490177687624
```

If there is a dominating point, that is, all its dimensions have the maximum values in the dataset, than the solution results in only the above $2m$ (6 in our case) constraints.

APPENDIX II: Generating Set of Candidate Performance Vectors

Feasible Performance Vector Generation

Introduction

At the beginning of this project, feasible performance vectors are generated using an analytical model based on continuous approximation and deterministic queueing theory. Last March, we proposed an alternative approach to generate the vectors based on historical precedent. Here, the procedure in this method will be discussed. The input to our analysis is weather forecast and demand forecast at the GDP decision time. The output from our analysis is feasible expected performance vectors given those forecasts. There are two steps in this procedure to generate feasible vectors for a given day:

Step 1. Identify similar historical days using weather forecast. The realized capacity scenarios (series of AARs) of these similar days will be used as possible actual capacity scenarios for this given day. Detail of this step can be found in the attached file, named as terminal weather forecast similarity_final (submitted to ICRAT 2014).

Step 2. Calculate the expected performance using the possible actual capacity scenarios from similar historical days, designed planned AARs, and demand forecast. Details regarding this step are presented below.

Performance metrics

In this analysis, we focus on three performance metrics: capacity utilization, efficiency and predictability.

1. Capacity utilization

This metric is defined to measure how much capacity is planned when the GDP is first implemented against the capacity under VMC condition. The equation is as follows:

$$\alpha_c = \frac{\sum_i Planned\ AAR_i}{\sum_i VMC\ AAR_i}$$

where, the AARs are for each quarter hour.

2. Efficiency

Efficiency is defined to measure how much delay has been transferred to the ground by the end of the program.

$$\alpha_e = \frac{\sum_j Ground\ delay_j}{\sum_j Total\ delay_j}$$

where, delays are first estimated for each flight j and then summed up. Delays are calculated based on the planned rates, the scenario-dependent actual rates, and the demand.

3. Predictability

Predictability is defined to measure the amount of unexpected delay in addition to the delay planned in the first GDP with respect to the amount of realized delay at the end. When the realized delay is less than planned delay for all the flights, then predictability is valued as 1.

$$\alpha_P = 1 - \frac{\sum_j max\{0, Realized\ delay_j - Planned\ delay_j\}}{\sum_j Realized\ delay_j}$$

where, delays are first estimated for each flight j and then aggregated. As before delays are calculated based on the planned rates, the scenario-dependent actual rates, and the demand.

Feasible Performance Vector Generation

Each performance metric, α, is a function of three terms:

- $\vec{D}$: scheduled demand, a series of quarter-hour arrival demand
- $\overrightarrow{AAR_P}$: airport acceptance rate planned in the initial GDP, a series of quarter-hour arrival capacity rates.
- $\overrightarrow{AAR_A}$: actual airport acceptance rate, a series of quarter-hour arrival capacity rates

For a given demand—$\vec{D}$ and planned AARs—$\overrightarrow{AAR_P}$, the expectation of the performance metric is then expressed as:

$$\bar{\alpha}\left(\vec{D}, \overrightarrow{AAR_P}\right) = \frac{\sum_s \alpha\left(\vec{D}, \overrightarrow{AAR_P}, \overrightarrow{AAR_{A,s}}\right)}{n_s}$$

where, s is the number of actual AAR scenarios—$\overrightarrow{AAR_{A,s}}$, which are identified through similarity analysis as in Step 1. Before, the set of $\overrightarrow{AAR_{A,s}}$ is used as the set for $\overrightarrow{AAR_P}$. Since $\overrightarrow{AAR_{A,s}}$ is from similar days, there is not much variability in it. As a result, $\overrightarrow{AAR_P}$ and the expectation of the performance metrics does not have much variability neither.

Current plan: $\overrightarrow{AAR_{A,s}}$ will still be identified from similarity analysis, whereas a different approach (under construction) will be used to generate $\overrightarrow{AAR_P}$. The rough idea is to find a lower bound and an upper bound of AARs for each quarter-hour and generate $\overrightarrow{AAR_P}$

based on this bounds. For instance, an extreme conservative $\overrightarrow{AAR_P}$ would be the series of lower bounds for each quarter-hour.

APPENDIX III: Assessing Terminal Weather Forecast Similarity for Strategic Air Traffic Management

Assessing Terminal Weather Forecast Similarity for Strategic Air Traffic Management

Yi Liu*, Michael Seelhorst, Alexey Pozdnukhov, Mark Hansen
Institution of Transportation Studies
University of California, Berkeley, CA 94720
liuyisha@berkeley.edu

Michael O. Ball
Robert H. Smith School of Business
University of Maryland, College Park, MD 20742

***Abstract*—in this paper, we propose a semi-supervised learning algorithm to assess similarity in weather forecast for strategic air traffic management. The distance metric between weather forecasts is supervised by pre-defined similarity and dissimilarity relationships. The distance metric considers the difference in each weather variable and also the interaction between two weather variables' differences. Using the proposed algorithm, two case studies are performed at Newark Liberty International Airport (EWR), where historically similar days in 2011 are identified for two given days-of-operation in 2012. The results show that similar weather forecasts could lead to very different airport acceptance rate and runway configuration outcomes in terms of capacity profiles and selections of runway configuration. Since differences in different weather phenomena are weighted differently in the distance metric, the algorithm could produce similar days to a given day which have considerable differences in some unimportant weather phenomena.**

Keywords-Air Traffic Management, Decision Making, Similar Days, Data Mining, Terminal Weather Forecast

I. INTRODUCTION

Air traffic managers today are typically limited to personal experience to make Traffic Flow Management (TFM) decisions [1, 2]. These decisions include whether or not there is a need for Traffic Management Initiatives (TMIs), such as Ground Delay Program (GDP) and Airspace Flow Program (AFP), and how TMIs should be planned when they are considered necessary. Managers with different experiences or different preferences may create different TMI plans for the same situation. This unpredictability in decision creates uncertainty for National Airspace System (NAS) users and may hinder them from taking effective proactive actions. To address this issue, systematic approaches should be developed to better inform and assist managers in TFM decision making.

One way of achieving this goal is to provide capacity profiles based on similar historical days to TFM decision makers. As illustrated in Fig. 1, two pieces of information are needed for traffic managers to make a TFM decision: a demand profile, which is known on a given day-of-operation; and a set of capacity profiles, which are to some degree uncertain. There is considerable research literature concerning the use of capacity profiles in the TMI planning process [3-7] and, to a lesser extent, the generation of these capacity profiles.

Reference [8] classifies historical capacity profiles into a small number of nominal scenarios for a given airport by using *K*-means clustering, which does not incorporate weather forecast. Reference [9] generates different capacity profiles for San Francisco International (SFO) airport for a given day by using empirical distributions of weather forecast error. Reference [10] creates capacity profiles also for SFO by using cumulative distribution functions of fog clearance time estimated from historical data. In [9] and [10], the source for weather information is exclusive to SFO and the models simplify the capacity profile by assuming an Airport Acceptance Rate (AAR) of 30 arrivals per hour before the fog clearance and 60 arrivals per hour afterwards. Reference [11] overcomes these limitations by basing the analysis on publically accessible Terminal Aerodrome Forecasts (TAFs) and developing possible capacity profiles from historical capacity scenarios. In this work, they find similar historical days to a given day using *K*-means clustering or Dynamic Time Warping (DTW) based on TAF.

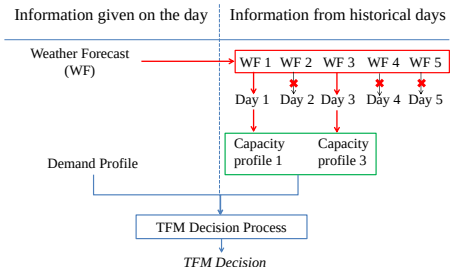


Figure 1. Flowchart of Traffic Flow Management Decision Making

In our work, we borrow the logic of generating capacity profiles in [11], which is shown in the upper part of Fig 1. First, similar historical days are identified to a given day by assessing the similarity between the weather forecasts. Then the AAR time series of the historically similar days are used as the candidate capacity profiles.

As mentioned, in [11], two ways to identify similar days to a given day are proposed: TAF clustering and DTW. In the former, the day is classified into one of several TAF clusters which has the shortest Euclidean distance between its centroid and the given day's TAF. The days in that cluster are then the similar days. In the DTW approach, the similarity of a given day TAF to all the historical TAFs is measured and the probability of a capacity profiles is inversely proportional to the degree of similarity. In both methods, the weather phenomena, such as visibility and ceiling, are weighted the

same in the distance metrics. Moreover, while there are many plausible ways of defining the distance metrics (Euclidean, city block, etc.), the choice of distance metric employed in this work is rather arbitrary.

This research addresses these issues by applying a semi-supervised learning algorithm to measure similarity between weather forecasts, where the distance metric is automatically learnt from similarity/dissimilarity relationships pre-defined by the users. The metric captures the differences in each weather phenomenon and also the interactions between different weather phenomena. The weights of the squared and the interaction terms are estimated based on the pre-defined similarity/dissimilarity relationships.

We describe the proposed methodology in Section II. In Section III, we describe the data that is used in this analysis. In Section IV, we present results from our case studies at Newark Liberty International Airport (EWR) airport. Finally, we conclude the paper in Section V.

II. METHODOLOGY

In this analysis, we apply a semi-supervised algorithm to identify similar days to a given day by learning distance metric [12]. The proposed algorithm consists of two steps: First, we learn the distance metric between **hourly** weather forecasts. Second, we identify similar days by selecting the days with small total distances in the weather forecast, summing over the hourly distances. In Sections A and B, we will elaborate on the two steps, respectively.

A. Learning Distance Metrics

In this section, we will explain how we generate the distance metric between hourly weather forecasts. Consider learning a distance metric of the form:

$$d_A(WF_i, WF_j) = \|WF_i - WF_j\|_A = \sqrt{(WF_i - WF_j)^T \cdot A \cdot (WF_i - WF_j)} \quad (1)$$

where, $WF_i \in \mathbb{R}^n$ is the weather forecast vector for hour *i*; *n* is the dimension of the vector. For this analysis, we use the following pieces of weather information found in the forecasts: ceiling, visibility, with speed and direction, as well as the presence of thunderstorm and snow. In the general case, *A* is a full matrix with diagonal and off-diagonal distance coefficients. The diagonal coefficients are the weights of the squares of the differences in each weather variable. The off-diagonal coefficients are the weights of the interaction term between two weather variable differences. If we set $A = I$, then the expression becomes the Euclidean distance. If we restrict *A* to be diagonal, this then corresponds to learning a metric in which the different weather attributes are given different weights without any interaction between them. In all cases, the required constraint on *A* is that it must be positive semi-definite, $A \succcurlyeq 0$. This ensures d_A to be a metric—satisfying non-negativity and the triangle inequality.

The key step in this approach is finding the *A* matrix and associated distance metric. We learn a distance metric that respects predetermined similarity between hours. Suppose we know that certain pairs of the hourly WF_i's are similar:

$$S: (WF_i, WF_j) \in S, \text{ if } \mathrm{WF_i} \text{ and } \mathrm{WF_j} \text{ are similar} \quad (2)$$

We can then learn a distance metric d_A respecting this so that similar hours end up close to each other. This is achieved by solving the following optimization problem:

$$\min_A \sum_{(WF_i, WF_j) \in S} \|WF_i - WF_j\|_A^2 \quad (3)$$

$$\text{s.t.} \sum_{(WF_i, WF_j) \in D} \|WF_i - WF_j\|_A \geq 1, \quad (4)$$

$$A \succcurlyeq 0. \quad (5)$$

where, the objective is to minimize the squared distance between the pairs of points in *S*. Since this is trivially solved with $A = 0$, we add the constraint (5) to ensure that *A* does not collapse the dataset into a single point. Here, *D* can be a set of pairs of hourly forecasts known to be dissimilar if such information is available; otherwise, it can be the complement of *S*. The objective function is linear in the parameters *A*, and both of the constraints can be verified to be convex. Thus, the optimization problem is convex, which enables us to derive efficient, local-minima-free algorithms to solve it. We consider this algorithm as semi-supervised since it learns a data transformation guided by user-defined *S* and *D* matrices as opposed to a fully supervised predictive model targeted at forecasting targets from sample input-output pairs.

The question that remains unanswered is how we define similarity and dissimilarity sets. In other words, we need to find the pairs of hours that belong to the sets *S* and *D*. The ultimate goal of identifying weather forecast similarity is to assist in air traffic management decision-making. Towards this goal, we define two hours as similar if all of the following conditions hold:

- The runway configuration is the same;
- Both hours have the same Meteorological Conditions (MC), either Instrument MC (IMC) or Visual MC (VMC);
- The absolute difference in actual AARs is smaller than S_{AAR}, where S_{AAR} is the AAR similarity threshold.

On the contrary, we define two hours to be dissimilar if they have different runway configurations, MCs and the differences in AARs are larger than the dissimilarity threshold, Di_{AAR}.

There is certainly more than one way to generate *S* and *D*. More research is needed to explore good definitions of the two sets. The definition of these sets is a place where feedback from TFM decision makers could be incorporated to tailor the similarity assessment to their need. In this paper, we focus on assessing terminal weather forecast similarity for airport capacity profile generation. But the methodology we use has generality and can be applied in other contexts. For instance, similarity in en route weather could be assessed with *S* and *D* defined using en route network characteristics such as traffic density.

B. Identifying Similar Days

In this section, using the learnt distance metric between hourly weather forecasts from Section A, we will describe how similar days are identified for a given day. Assume the concerned time horizon is from hour T_s to hour T_e. One way of defining the distance of weather forecast between two days is

to sum up the squares of the hourly distances with the time horizon:

$$D_{J,K} = \sum_{i=T_s}^{T_e} \left\| WF_{J,i} - WF_{K,i} \right\|_A^2 \quad (6)$$

where, $D_{J,K}$ is the total distance between day j and day k; $WF_{J,i}$ is the weather forecast in hour i on day J; $WF_{K,i}$ is the weather forecast in hour i on day k. For a given day, similar days are then days with small total distances. The user can decide on the number of similar days, *Ns*, to look at.

Once historically similar days are identified following the proposed mechanism, the time series of the actual runway configuration, MCs, and AARs on these days may then be provided to traffic managers as references for decision making.

III. DATA

The analysis uses data from two sources: Aviation System Performance Metrics (ASPM) and Terminal Aerodrome Forecast (TAF). ASPM report provides hourly data on runway configuration, MC and AAR.

Historical TAFs provides hourly data for terminal weather forecasts. The authors developed a Matlab script to download the historical TAF information from www.ogimet.com. Each TAF was read in as a text file. Afterwards, a parser written in Matlab was used to convert the text files to user-friendly numerical data. Weather variables created based on the TAF data are listed below.

First, we have two indicator variables representing the presence of thunderstorms—*TS*,and snow—*Sn*. Visibility in the forecasts ranges from 0.25 miles to 7 miles. We expect the effect of the visibility to be much stronger as it approaches zero, also when the value is below 4 as opposed to above 4, where 4 is the threshold for visual approach conditions. We use two variables for visibility, including a natural log transform—*Vis1* = log(visibility) and a discontinuity to capture the nonlinear effects—*Vis2* = max(0,log(visibility/4)). Ceiling in the forecasts ranges from 100 feet to 25,000 feet. Similar to visibility, two variables are defined for ceiling: *Ceil1* = log(ceiling) and *Ceil2* = max (0,log(ceiling/3000)), where 3000 feet is the threshold of visual approach condition. The TAFs contain both wind speed and direction. We include three variables to capture this information. When wind direction is specified (some observations have variable wind direction), we decompose the wind speed into two components: speed from the North to the South and speed from the East to the West. The direction of the wind blowing from north to south and from east to west is considered positive. When the wind direction is variable, we consider wind by its absolute speed only. This gives us three wind variables:

- *Ws*: absolute wind speed. It equals to the reported wind speed when wind direction is unspecified, and zero when wind direction is specified
- *WN*: component of wind speed from the North to the South. It equals to zero if wind direction is unspecified
- *WE*: component of wind speed from the East to the West. It equals to zero if wind direction is unspecified

The weather forecast vector, *WF*, is then a 9-dimensional vector:

$$WF = [TS; Sn; Vis1; Vis2; Ceil1; Ceil2; Ws; WN; WE] \quad (7)$$

For each day, a new TAF is updated every two to three hours. Therefore, we should determine which forecast to use before assessing weather forecast similarity between a given day and the historical days. Since the purpose of identifying similar days is to assist in traffic management decision-making, one way to select the TAF is to refer to the traffic management decision time. At the decision time, we compare the most recently issued TAF on the given day to the most recent TAFs that were available at this time for the historical days.

Each TAF forecasts weather for 24 hours from the forecast issuance time. The number of hours that will be compared to identify similar days is determined by the planning horizon.

For the case study, we pulled ASPM and TAF hourly data for EWR for years 2011 and 2012. TAF data is missing for 2011 November. We thus removed the observations for November from the ASPM data as well.

IV. EWR CASE STUDIES

In this section, we apply the proposed methodology to terminal weather forecast similarity assessment at EWR. This airport is selected since it is one of the most congested airports in the US and has many GDPs. Two case studies are performed for two selected days-of-operation: September 20, 2012 and June 8, 2012. There was no GDP on September 20, 2012 and there was a GDP on June 8, 2012 due to wind. The GDP was planned to start at 4:47 pm and end at 9 pm but actually ended at 5:43 pm.

We first use the ASPM and TAF data from 2011 to generate the *A* matrix. Then, using the learnt distance metric, we identify five similar days for each of the two days. When defining the *S* and *D* sets, we set the thresholds for similarity and dissimilarity in AARs as 1 and 7 arrivals per hour respectively. The results for the A matrices and similar days are shown and discussed in the following two subsections.

A. Similar Days for Sep 20, 2012

There was no GDP on 9/20/2012. Many of the EWR GDPs start around noon (without specification, time is assumed as local time) and planned for around 10 hours on average [13]. Accordingly, we set the time horizon of the analysis from noon to midnight. All the analysis results are then based on hourly data between noon and midnight from ASPM and TAF, including estimation of the *A* matrix. Moreover, the most recent TAF that was available at noon is referenced for values of the weather variables.

The estimation results for the *A* matrix are shown in Table I, containing values of diagonal and off-diagonal elements. The diagonal entries are distance coefficients (also referred to as weights) for the squares of the differences in each weather variable. The off-diagonal entries are distance coefficients for the interaction terms of two weather variables' differences.

To explain the interpretation of the *A* matrix, we offer a simple example. For a pair of hours, assume all the other weather conditions are the same but there could be differences

in thunderstorm and snow conditions. Then the distance metric is reduced to $61.15\cdot(\Delta TS)^2 + 7.05\cdot(\Delta Sn)^2 + 2\times4.78\cdot\Delta TS\cdot\Delta Sn$, where ΔTS and ΔSn are the thunderstorm and snow differences between these two hours. Since thunderstorm happens more often in the summer season and it snows only in the winter, the interaction term $\Delta TS\cdot\Delta Sn$ almost always takes the value -1 or 0. Therefore, the distance metric can take three values:

- 61.15 if one hour has thunderstorm, the other hour does not have thunderstorm, and they have the same snow condition (presumably no snow);
- 7.05 if one hour has snow, the other hour does not have snow, and they have the same thunderstorm condition (presumably no thunderstorm);
- $61.15+7.05-2\times4.78$ if they have different thunderstorm and snow conditions.

These indicate that there is a large difference between a thunderstorm hour and a non-thunderstorm hour, whereas there is less difference between a snow hour and a non-snow hour. If one hour has thunderstorm and the other hour has snow, the distance between these two hours is smaller than two hours with different thunderstorm conditions but the same snow conditions. This indicates how the interaction term affects the distance between 2 hours in which snow and thunderstorm conditions are both different.

TABLE I. ESTIMATION RESULTS ON A MATRIX

Variables	*TS*	*Sn*	*Vis2*	*Vis1*	*Ceil2*	*Ceil1*	*Ws*	*WN*	*WE*
TS	61.15[a]	4.78	4.14	15.52	-11.74	12.30	2.22	-0.47	-0.56
Sn	4.78	7.05	20.28	-7.67	-3.80	3.84	-0.10	-0.07	-0.10
Vis2	4.14	20.28	66.55	-17.37	-29.90	30.09	2.91	-0.75	-0.81
Vis1	15.52	-7.67	-17.37	39.05	-40.69	41.07	8.46	-1.35	-1.31
Ceil2	-11.74	-3.80	-29.90	-40.69	87.12	-87.76	-15.24	2.64	2.61
Ceil1	12.30	3.84	30.09	41.07	-87.76	88.41	15.35	-2.66	-2.63
Ws	2.22	-0.10	2.91	8.46	-15.24	15.35	2.76	-0.47	-0.46
WN	-0.47	-0.07	-0.75	-1.35	2.64	-2.66	-0.47	0.08	0.08
WE	-0.56	-0.10	-0.81	-1.31	2.61	-2.63	-0.46	0.08	0.08

a. All the weights are scaled by a factor of 1000.

The diagonal coefficients are the weights of the squared terms of each weather variable's differences. But they cannot be interpreted as the relative importance of the squared differences to the distance since the values of the weather variable differences vary. In order to compare the contributions of each weather phenomenon to the distance, we create five hypothetical *A* matrices for the five types of weather phenomena based on the full *A* matrix: thunderstorm, snow, visibility, ceiling and wind. The hypothetical *A* matrix for a weather phenomenon is created by keeping the diagonal distance coefficient(s) of the weather variables belonging to this weather phenomenon and, where applicable, their interaction coefficients, and replacing the rest of the entries with zero. Following this, there will be 1 non-zero element for thunderstorm and snow respectively, 4 non-zero elements for visibility and ceiling respectively, and 9 non-zero elements for wind, as highlighted in boxes in Table I. Hypothetical distances between the hourly TAFs are then estimated for each pair of hours selected for this analysis using the five hypothetical *A* matrices. For each weather phenomenon, these hypothetical distances could also be estimated by using the full *A* matrix and assuming no differences in the variables representing the other weather phenomena. Now, we can compare the contributions of different phenomena to the distance by comparing the values of these hypothetical hourly distances.

In Table II, we present summary statistics for the hypothetical hourly distances for the five weather phenomena. The median value is the highest for wind, followed by ceiling, whereas the rest of the median values are zero. These results indicate that typically wind differences account for most of the distance and ceiling differences are responsible for the rest. This is mainly because wind differences and ceiling differences are common whereas more than half of the pairs of hours have the same values for thunderstorm, snow and visibility variables, according to the percentages of non-zeroes. If we only consider non-zero observations for each weather phenomena, the average distances due to thunderstorm differences are the largest on average. This indicates that two hours would be viewed as the most different when one has thunderstorm and the other one does not. The average distances due to visibility and ceiling differences are similar in magnitude, and much higher than the distance due to wind differences. Out of the five differences, wind differences have the smallest average impact when considering only the non-zero values. As a result, even though wind difference is the most common difference, the average distance induced by this difference across both zero and non-zero observations is smaller than the average distance generated by the visibility differences and ceiling differences (as shown in Row 1 in Table II), which contribute significantly to the distance metric when they occur, and they do a fair amount of the time.

TABLE II. STATISTICS OF HYPOTHETICAL HOURLY DISTANCES

Statistics	Thunderstorm	Snow	Visibility	Ceiling	Wind
Mean	3.5[a]	2	47	94.6	24.5
Median	0	0	0	3.4	20.1
Max	247.3	84	491.5	809.9	320.8
% non-zero obs.	1.43	2.41	31.3	71.6	99.6
Mean of non-zero obs.	247.3	84	150.3	132.1	24.6
Median of non-zero obs.	247.3	84	149	5.5	20.2
Std. of non-zero obs.	0	0	103.4	190.9	21.8

a. All the weights are scaled by a factor of 1000.

In order to visualize the similarity between each hour, we have applied a metric Multi-Dimensional Scaling (MDS) [14] to the pair-wise distance matrix. An MDS plot provides a distance-preserving visualization of the data such that the pairwise distances in 9-dimensional space are reproduced in 2 dimensions with minimum distortion. It helps analyzing the effect of the metric transform learned by the algorithm and the scaling applied to original variables. An example of the MDS plot is presented in Fig. 2 for *Ceil1*—log(ceiling), where the color scale corresponds to the values of the original *Ceil1* variable. As shown, the ceiling value roughly increases from the right to the left. Hours with similar ceiling values are clustered to some degree. We have considered similar plots for all the variables in MDS projection. They are not shown here but the information is conveyed in Fig. 2. The plot of visibility has similar trend as ceiling, where the value increase from the right to the left. Hours with thunderstorms and snow locate at

the bottom and on the top, respectively, which are far away from the majority of the points. This indicates that a pair of hour with different thunderstorm or snow conditions is more different from that with different ceiling or visibility conditions. In the plots of wind variables, there is no obvious trend in the spatial distribution of hours with different wind speeds and wind components. This indicates that the contribution of wind difference to the hourly distance is similar in the range of wind speed we have here. The cluster of good weather conditions is enlarged in the plot. They are hours with high ceiling, high visibility, no thunderstorm, and no snow.

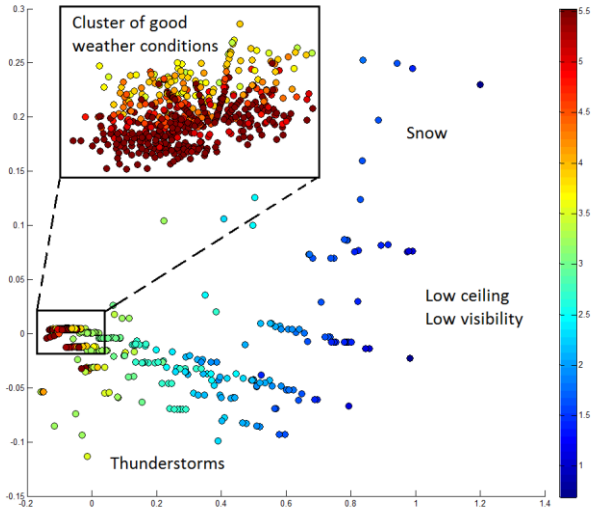


Figure 2. Metric Multi-Dimensional Scaling Results

Using (3), the similar days from 2011 were identified for Sep 20, 2012, based on the estimated *A* matrix and values of the TAF variables. The five similar days are selected as the five days with the smallest total hourly distances in TAF compared to the given day. All the hours were VMC except for the hour from 5 pm to 6pm on Sep 20, 2012. The actual AARs, selected runway configurations and the GDP decisions of the similar days and the given day are summarized in Table III. The first 4 similar days all share the same runway configuration with the given day for a considerable duration, whereas 12/3/2011 has a totally different configuration. The AARs of the first two similar days are similar to the given day, especially in terms of sum of AARs. The AARs of the last three similar days are smaller than those of the given day for about half of the time, with the largest hourly difference as 10 arrivals per hour.

The forecasted ceiling, wind speed (*Ws*) and Wind direction (*Wdir*) for the given day and similar days are summarized in Table IV. There were no thunderstorms or snow and visibility was 7 mile for all the hours. The weather conditions are very similar for these days except for wind. The wind directions in the similar days are very different from those on the given day, except 12/3/2011. The overall wind speed is generally small for all days. This indicates that the difference in wind direction is not very important in determining similarity when wind speed is not high. The weather conditions on 12/3/2011 are similar to the given day. However, the AAR and runway configuration of 12/3/2011 and 6/8/2012 are different according to Table III. This indicates that capacity profiles and selection of runway configuration could be very different given similar terminal weather forecasts. This could be a result of inherent uncertainty in weather forecasts, or non-weather factors (such as demand and facility outages) that also influence the runway configuration and AAR.

TABLE III. ACTUAL OBSERVATIONS ON 09/20/2012 AND ITS HISTORICALLY SIMILAR DAYS

	Given Day 9/20/2012 No GDP		Similar Day 1 9/16/2011 No GDP		Similar Day 2 7/27/2011 No GDP	
Hour (GMT[b])	**AAR**	**Rwy Conf.**	AAR	Rwy Conf.	AAR	Rwy Conf.
9/20/16Z	**46**	**4R, 11|4L**	46	4R, 11|4L	48	4R, 11|4L
9/20/17Z	**46**	**4R, 11|4L**	46	4R, 11|4L	48	4R, 11|4L
9/20/18Z	**46**	**4R, 11|4L**	46	4R, 11|4L	48	4R, 11|4L
9/20/19Z	**46**	**4R, 11|4L**	46	4R, 11|4L	48	4R, 11|4L
9/20/20Z	**41**	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
9/20/21Z	**38**[a]	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
9/20/22Z	**38**	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
9/20/23Z	**46**	**4R, 11|4L**	46	4R, 11|4L	38	4R | 4L
9/21/00Z	**44**	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
9/21/01Z	**38**	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
9/21/02Z	**38**	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
9/21/03Z	**38**	**4R | 4L**	46	4R, 11|4L	38	4R | 4L
	Similar Day 3 10/30/2011 GDP		Similar Day 4 7/9/2011 No GDP		Similar Day 5 12/3/2011 No GDP	
Hour (GMT)	AAR	Rwy Conf.	AAR	Rwy Conf.	AAR	Rwy Conf.
9/20/16Z	36	4R | 4L	42	4R | 4L	38	22L | 22R
9/20/17Z	36	4R | 4L	48	4R, 11 | 4L	38	22L | 22R
9/20/18Z	38	4R | 4L	41	4R | 4L	38	22L | 22R
9/20/19Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/20/20Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/20/21Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/20/22Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/20/23Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/21/00Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/21/01Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/21/02Z	38	4R | 4L	38	4R | 4L	38	22L | 22R
9/21/03Z	38	4R | 4L	38	4R | 4L	38	22L | 22R

a. This hour was IMC where the rest were all VMC.

b. GMT time is 4 hours ahead of local time during daylight saving and 5 hours ahead otherwise.

On similar day 3, a GDP was planned from 3 pm to 10 pm local time. There were no GDPs for the other 4 similar days. Therefore, if the demand profile for the given day is similar to that from the historically similar days, the similarity analysis would suggest no GDP on the given day, which was the actual TFM decision.

TABLE IV. TAF FOR 09/20/2012 AND ITS HISTORICALLY SIMILAR DAYS

	Given Day 9/20/2012			Similar Day 1 9/16/2011			Similar Day 2 7/27/2011		
Hour (GMT)	**Ceiling**	***Ws***	***Wdir***	Ceiling	*Ws*	*Wdir*	Ceiling	*Ws*	*Wdir*
9/20/16Z	**40**	**9**	**100**	250	12	340	250	11	340
9/20/17Z	**40**	**9**	**100**	250	12	340	250	11	340
9/20/18Z	**40**	**9**	**100**	250	9	320	250	11	340
9/20/19Z	**250**	**8**	**140**	250	9	320	250	10	310
9/20/20Z	**250**	**8**	**140**	250	9	320	250	10	310
9/20/21Z	**250**	**8**	**140**	250	9	320	250	10	310
9/20/22Z	**250**	**8**	**140**	250	9	320	250	8	300
9/20/23Z	**250**	**8**	**140**	250	9	320	250	8	300
9/21/00Z	**250**	**5**	**110**	200	7	330	250	8	300
9/21/01Z	**250**	**5**	**110**	200	7	330	250	4	300
9/21/02Z	**250**	**5**	**110**	200	7	330	250	4	300
9/21/03Z	**250**	**5**	**110**	200	7	330	250	4	300

	Similar Day 3 10/30/2011			Similar Day 4 7/9/2011			Similar Day 5 12/3/2011		
Hour (GMT)	Ceiling	*Ws*	*Wdir*	Ceiling	*Ws*	*Wdir*	Ceiling	*Ws*	*Wdir*
9/20/16Z	250	14	330	250	13	330	250	7	30
9/20/17Z	250	14	330	250	13	330	250	7	30
9/20/18Z	250	14	330	250	13	320	250	6	140
9/20/19Z	250	14	330	250	13	320	250	6	140
9/20/20Z	250	13	320	250	13	320	250	6	140
9/20/21Z	250	13	320	250	13	320	250	6	140
9/20/22Z	250	13	320	250	13	320	250	6	140
9/20/23Z	250	13	320	250	13	320	250	6	140
9/21/00Z	250	13	320	250	13	320	250	6	140
9/21/01Z	250	6	320	250	8	340	250	6	140
9/21/02Z	250	6	320	250	8	340	250	6	140
9/21/03Z	250	6	320	250	8	340	250	6	140

B. Similar Days for June 8, 2012

There was a GDP planned on 6/8/2012 from 4:47 pm to 9 pm due to wind. The program was actually ended earlier at 5:43 pm. To find similar days for assisting in decision-making, we set the time horizon as 4 pm to 10 pm. All the analysis results are then based on hourly data on this time horizon from ASPM and TAF, including estimation of the *A* matrix. The most recent TAF that was available at this time is referenced for values of the weather variables. It is worth mentioning that the *A* matrix is different from before because we are using a different set of pairs of hours and different TAFs. Usually, the most recent TAFs prior to noon and 4 pm are issued around 10 am and 1:30 pm, respectively. The results for this case study are shown in Tables V to VIII. There were no thunderstorms or snow and visibility was 7 miles for all the days.

As shown in Tables V and VI, compared to the previous case, snow and ceiling differences are weighted more where the other three differences are weighted less.

TABLE V. ESTIMATION RESULTS ON A MATRIX

Variables	***TS***	***Sn***	***Vis2***	***Vis1***	***Ceil2***	***Ceil1***	***Ws***	***WN***	***WE***
TS	19.46	32.41	8.61	16.16	-53.73	53.02	5.42	-2.12	-1.46
Sn	32.41	54.03	13.18	27.53	-88.80	87.64	8.99	-3.51	-2.41
Vis2	8.61	13.18	27.74	-5.05	-34.19	33.42	2.92	-1.24	-0.84
Vis1	16.16	27.53	-5.05	19.81	-38.65	38.30	4.21	-1.59	-1.10
Ceil2	-53.73	-88.80	-34.19	-38.65	155.17	-153.03	-15.27	6.04	4.14
Ceil1	53.02	87.64	33.42	38.30	-153.03	150.93	15.07	-5.96	-4.08
Ws	5.42	8.99	2.92	4.21	-15.27	15.07	1.52	-0.60	-0.41
WN	-2.12	-3.51	-1.24	-1.59	6.04	-5.96	-0.60	0.24	0.16
WE	-1.46	-2.41	-0.84	-1.10	4.14	-4.08	-0.41	0.16	0.11

a. All the weights are scaled by a factor of 1000.

TABLE VI. STATISTICS OF HYPOTHETICAL HOURLY DISTANCES

Statistics	**Thunderstorm**	**Snow**	**Visibility**	**Ceiling**	**Wind**
Mean	5.9	7.7	43.6	128.7	34.5
Median	0	0	0	7.7	28.6
Max	139.5	232.4	458	894.5	278.9
% non-zero obs.	4.19	3.33	35.8	74.4	99.5
Mean of non-zero obs.	139.5	232.4	121.7	173	34.6
Median of non-zero obs.	139.5	232.4	108.3	10.7	28.3
Std. of non-zero obs.	0	0	85.3	247.2	27.3

a. All the weights are scaled by a factor of 1000.

The capacity profiles and the runway configurations are similar between the given day and the first four similar days, as shown in Table VII. But again, differences are observed in the weather forecasts. The AARs and runway configurations are not very similar for the given day and similar day 5, neither are the weather forecasts. Specifically, an additional arrival runway (Runway 11) was used for much of the time on similar day 5, which greatly increased the arrival capacity. Further study is required to determine if this might have been predicted based on the weather forecast.

Out of the five similar days, only 8/1/2011 had a GDP. If the demand profiles are similar between 6/8/2012 and the similar days, then the analysis suggests no GDP for this day. Although there was a GDP for the given day, it was cancelled less than 1 hour after its implementation. The suggestion would have been helpful since it appears a GDP was not necessary on the given day.

TABLE VII. ACTUAL OBSERVATIONS ON 06/08/2012 AND ITS HISTORICALLY SIMILAR DAYS

	Given Day 6/8/2012 GDP		Similar Day 1 8/1/2011 GDP		Similar Day 2 12/30/2011 No GDP	
Hour (GMT)	**AAR**	**Rwy Conf.**	AAR	Rwy Conf.	AAR	Rwy Conf.
6/8/21Z	**39**	**22L | 22R**	32	22L | 22R	38	22L | 22R
6/8/22Z	**38**	**22L | 22R**	32	22L | 22R	38	22L | 22R
6/8/23Z	**38**	**22L | 22R**	35	22L | 22R	38	22L | 22R
6/9/00Z	**38**	**22L | 22R**	35	22L | 22R	38	22L | 22R
6/9/01Z	**38**	**22L | 22R**	35	22L | 22R	42	22L | 22R
6/8/21Z	**38**	**22L | 22R**	35	22L | 22R	42	22L | 22R

	Similar Day 3 7/23/2011 No GDP		Similar Day 4 12/31/2011 No GDP		Similar Day 5 5/1/2011 No GDP	
Hour (GMT)	AAR	Rwy Conf.	AAR	Rwy Conf.	AAR	Rwy Conf.
6/8/21Z	38	22L | 22R	38	22L | 22R	52	11, 22L | 22R
6/8/22Z	38	22L | 22R	46	11, 22L | 22R	52	11, 22L | 22R
6/8/23Z	38	22L | 22R	38	11, 22L | 22R	52	11, 22L | 22R
6/9/00Z	38	22L | 22R	38	22L | 22R	49	22L | 22R
6/9/01Z	38	22L | 22R	38	22L | 22R	38	22L | 22R
6/8/21Z	38	22L | 22R	38	22L | 22R	38	22L | 22R

TABLE VIII. TAF FOR 06/08/2012 AND ITS HISTORICALLY SIMILAR DAYS

	Given Day 6/8/2012			Similar Day 1 8/1/2011			Similar Day 2 12/30/2011		
Hour (GMT)	**Ceiling**	***Ws***	***Wdir***	Ceiling	*Ws*	*Wdir*	Ceiling	*Ws*	*Wdir*
6/8/21Z	**250**	**12**	**270**	250	10	260	250	6	190
6/8/22Z	**250**	**12**	**270**	250	10	260	250	6	190
6/8/23Z	**250**	**12**	**270**	250	10	260	250	6	190
6/9/00Z	**250**	**12**	**270**	250	10	260	250	6	190
6/9/01Z	**150**	**12**	**300**	250	8	290	250	5	160
	Similar Day 3 7/23/2011			Similar Day 4 12/31/2011			Similar Day 5 5/1/2011		
Hour (GMT)	Ceiling	*Ws*	*Wdir*	Ceiling	*Ws*	*Wdir*	Ceiling	*Ws*	*Wdir*
6/8/21Z	150	12	280	40	9	290	200	10	160
6/8/22Z	150	12	280	40	9	290	200	10	160
6/8/23Z	150	12	280	40	9	290	200	10	160
6/9/00Z	150	12	280	40	9	290	200	10	160
6/9/01Z	250	10	280	250	10	290	150	7	160

V. CONCLUSIONS AND FUTURE WORK

In this work, we propose a semi-supervised algorithm for assessing weather forecast similarity for air traffic management. The distance metric between hourly TAFs is automatically learnt from similarity and dissimilarity relationships pre-defined by comparing the actual outcomes (AAR, runway configuration and MC) for the hours. Distance coefficients for the squared differences in the weather variables and the coefficients for the interaction between two weather variables' differences are estimated. Then distance between two days is calculated as the sum of the squared hourly distances over a given time horizon. Finally, the degree of similarity of a historical day to a given day is inversely proportional to the distance between them.

Using the proposed algorithm, we perform two case studies at EWR, where historically similar days from 2011 are identified for 9/20/2012 and 6/8/2012 respectively. The distance metric shows that wind difference is the most common weather difference at EWR. However, on average, ceiling and visibility differences contribute most to the hourly TAF distances. A day with severe thunderstorms or heavy snow has the largest TAF distance compared to a good weather day.

Comparing the AAR and runway Configuration outcomes, and TAFs of the given day to the similar days, we learn that similar weather forecasts can lead to different outcomes, which demonstrate that the uncertainty in TFM might be unavoidable given the inherent uncertainty in weather forecast. It is also observed that wind direction could be very different for similar days when wind speed is small. This indicates that difference in the wind direction may not be important in determining the outcomes, when the wind speed is below a certain threshold. More work should be done to find this threshold.

Summarizing out two case studies, there was no GDP on 9/20/2012, and for its five historically similar days, only one day had GDP. There was a GDP planned for five hours on 6/8/2012 but was cancelled four hours earlier. Out of the historically similar days for 6/8/2012, only one day had GDP. Assuming similar demand profiles, then the analysis would suggest no GDP in either case based on the past decisions, and this would indeed have probably been the better course of action given the early cancellation on 6/8/2012.

This work is still at its preliminary research stage. In the on-going work, we are improving the current method and exploring other approaches to similarity assessment under the same idea of learning distance metric. One alternative way to supervise is defining similarity/dissimilarity relationship based on the realized performance such as delay cost, rather than actual AAR, runway configuration and MC. In this case, the distance metric will be learned for a pair of days instead of a pair of hours. Moreover, this will change the flowchart shown in Fig. 1 because demand profile will also be used in calculating delay cost. As a result, similarity will be defined based on weather forecast and also demand.

REFERENCES

[1] S. R. Wolfe, J. L. Rios, "A method for using historical ground delay programs to inform day-of-operation programs," AIAA Gidance, Navigation, and Control Conference, 2011.

[2] S. Grabbe, B. Sridhar, A. Mukherjee, "Similar days in the NAS: an airport perspective," Aviation Technology, Integration, and Operations Conference, 2013.

[3] O. Richetta, A. R. Odoni, "Dynamic solution to the ground holding problem in air traffic control," Transportation Research Part A, Vol. 28, pp. 167-185, 1994.

[4] A. Mukherjee, M. Hansen, "A dynamic stochastic model for the single airport ground holding problem," Transportation Science, Vol. 41, pp. 444-456, 2007.

[5] B. M. Ball, R. Hoffman, A. Mukherjee, "Ground delay Program planning under uncertainty based on the ration-by-distance principle," Transportaion Science, Vol. 44, pp. 1-14, 2010.

[6] H. D. Sherali, J. M. Hill, M. V. McCrea, A. A. Trani, "Integrating slot exchange, safety, capacity, and equity mechanisms within an airspace flow program," Transportation Science, Vol. 45, pp. 271-284, 2011.

[7] Y. Liu, M. Hansen, "Incorportating predictability into cost optimization for ground delay program," unpublished.

[8] P. B. Liu, M. Hansen, A. Mukherjee, "Scenario-based air traffic management: from theory to practice," Transportation Research Part B, Vol. 42, pp. 685-702, 2008

[9] L. S. Cook, B. Wood, "A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing," Proceedings of 8th USA/Europe Air Traffic Management R&D Seminar, 2009.

[10] A. Mukjerjee, M. Hansen, S. Grabbe, "Ground delay program planning under uncertainty in airport capacity," Transportation Planning and Technology, Vol 35., pp. 611-628, 2012.

[11] G. Buxi, M. Hansen, "Generating day-of-operation probabilistic capacity scenarios from weather forecasts," Transportation Research Part C, Vol. 33, pp. 153-166, 2013

[12] P.X. Eric, A.Y. Ng, M.I. Jordan, S. Russel, "Distance metric learning, with appliation to clustering with side-informtion," Advances in Neural Information Processing Systems, Vol. 15, , pp. 505-512, 2002.

[13] Y. Liu, M. Hansen, "Evaluation of the performance of ground delay program," Transportation Research Record, Vol. 2400, pp. 54, 2014.

[14] J. B., Kruskal, M. Wish, "Multidensional scaling," Sage Univerity Paper Series on Quantitative Application in the Social Science, pp. 7-11, 1978

APPENDIX IV: Architectures for Hierarchical Application of COuNSEL for Strategic Operational Planning

Distributed Mechanisms for Determining NAS-Wide Service Level Expectations:

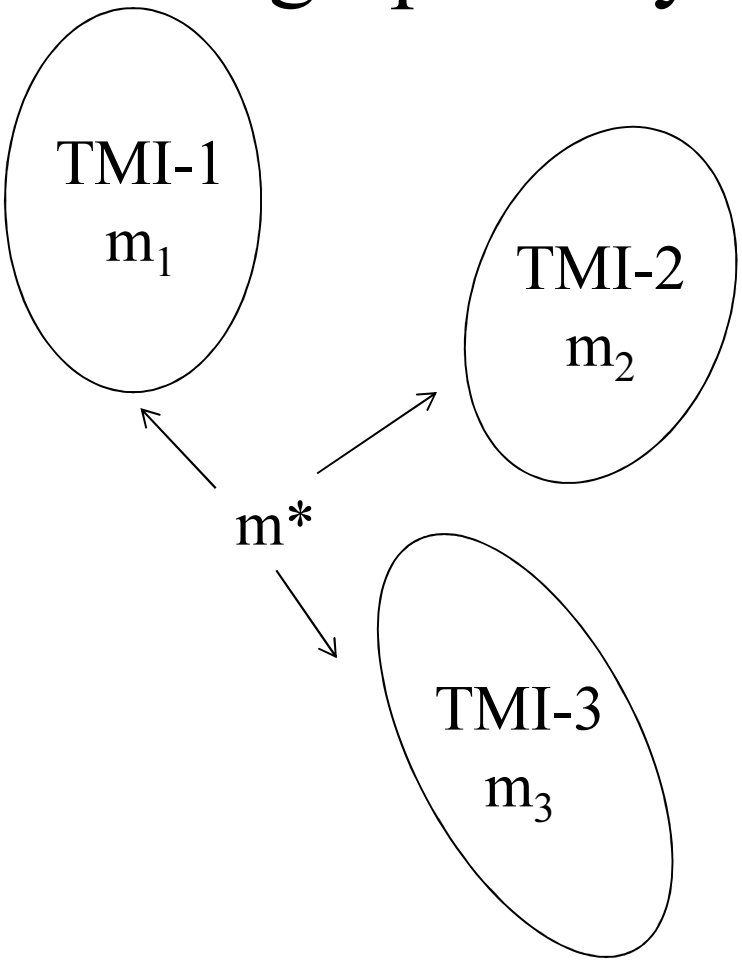
Discussion of Models for Application of CONSEL over Geographically Dispersed Problem Areas

Faculty: Michael Ball, Cindy Barnhart, Mark Hansen, Vikrant Vaze

Students and Post-Doc's: Yi Liu, Prem Swaroop, Chiwei Yan

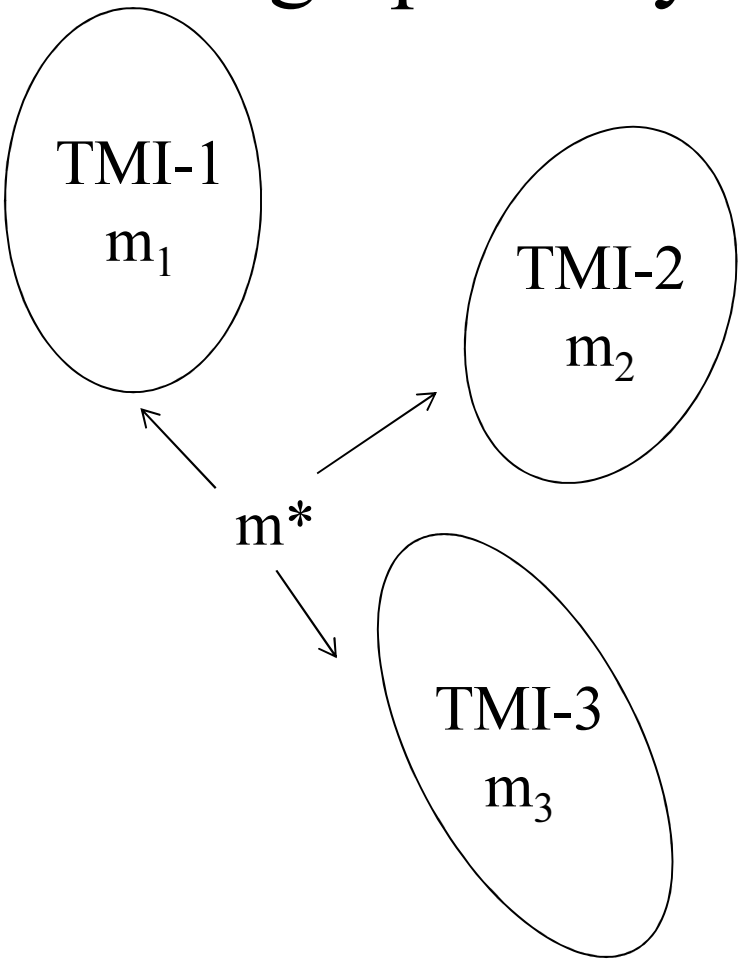
February 4, 2013

Service Level Expectation Setting over Geographically Dispersed Set of TMIs



- As is typical (today) there may be several TMIs running simultaneously.
- Each in concept should have its own metric vectors: m_1, m_2, m_3 to guide its design.
- There could also be an overall vector m* governing the NAS-wide (or region-wide) TMI strategy.
- Several control architectures are possible relative to relationship between m* and m_1, m_2, m_3 and how CONSEL would be applied.

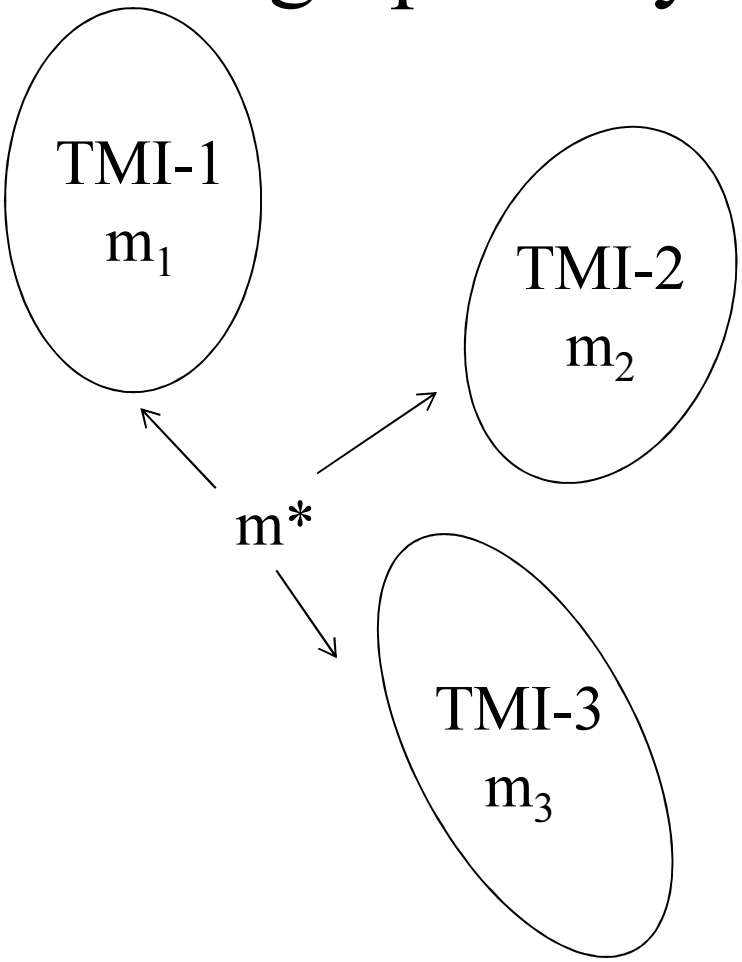
Service Level Expectation Setting over Geographically Dispersed Set of TMIs



Architecture 1:

- CONSEL is applied to determine m*.
- In a second step an optimization model is applied to determine m_1, m_2, and m_3
- These are set taking into account flight operator weights and the way in which flight operator weights vary by TMI, e.g. if DAL had the highest percentage of operations in TMI-1 then m_1 would tend favor DAL's most preferred vector.

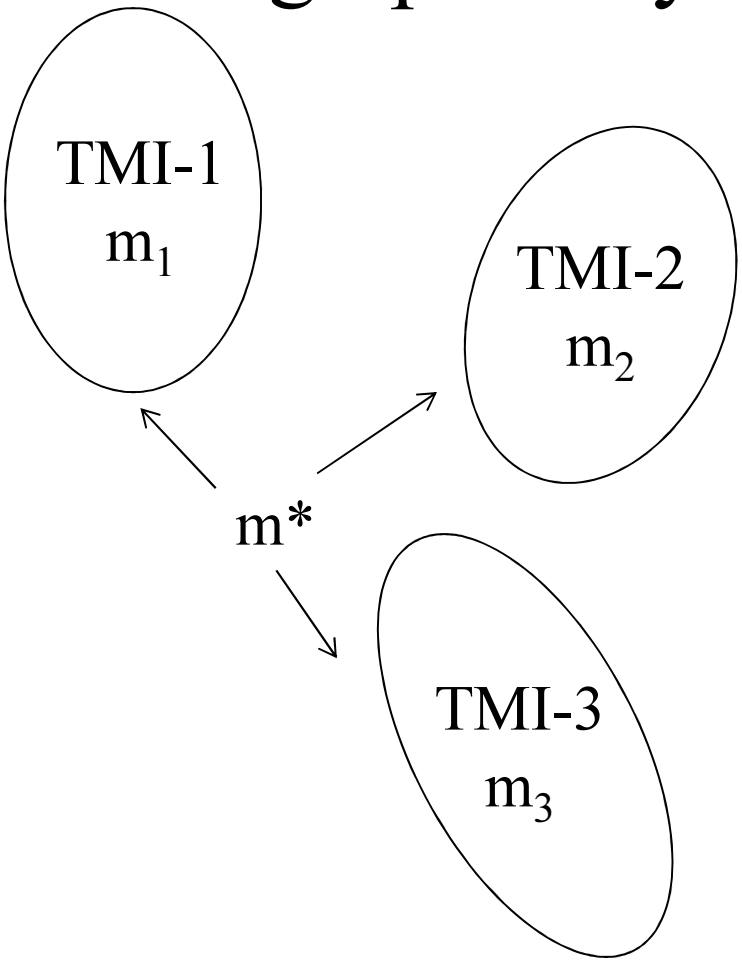
Service Level Expectation Setting over Geographically Dispersed Set of TMIs



Architecture 2:

- CONSEL is applied to determine m_1, m_2, and m_3. m* does not exist.
- However, the processes for determining m_1, m_2, and m_3 are not independent.
- The "winning" vector for a specific TMI might be adjusted to better balance the performance achieved by each flight operator, e.g. if a specific flight operator received poor performance in TMI-1 then that flight operator might be favored in TMI-2.

Service Level Expectation Setting over Geographically Dispersed Set of TMIs



Architecture 3:

- CONSEL is applied to determine m* and in a second step: m_1, m_2, and m_3.
- However, the processes for determining m_1, m_2, and m_3 are not independent and they also take into account m*.
- The "winning" vector for the specific TMI's is determined to seek balance as in architecture 2 but also these vectors are set in a way that the overall performance is close to m*.

APPENDIX V: Service Level Expectation Based Ground Delay Program Design

SERVICE LEVEL EXPECTATION BASED GROUND DELAY PROGRAM DESIGN

Yi Liu, Lei Kang, Robert Hoffman, Mark Hansen

1. INTRODUCTION

Air traffic congestion frequently occurs in the national airspace system (NAS) due to adverse weather, high demand or other disturbances. When congestion is foreseen, traffic management initiatives (TMIs), such as ground delay program (GDP) and airspace flow program (AFP) are called to balance demand with capacity. In the current air traffic management system, TMI decisions are discussed between Federal Aviation Administration (FAA) traffic specialists and flight operator personnel in the form of strategic planning telecons. The telecons allow flight operators to interact with managers and express their opinions on flow management strategies. The interactions are legitimate and desirable given that that they allow the NAS users who are impacted by various FAA decisions to help the FAA understand their priorities and the impact of FAA actions *(1)*. At the same time, since the interaction are verbal and the input focuses on TMI parameters rather than service level expectations (SLE)—expected system performances from the TMI designs, the decision-making process can be ad hoc and subjective. In light of these, Ball et al. (*2*) proposes a SLE-oriented mechanism to consider input of all involved flight operators in a systematic way and generate an output that can represent the consensus of these flight operators *(1, 2)*. The TMI planning process under this mechanism is illustrated in Figure 1.

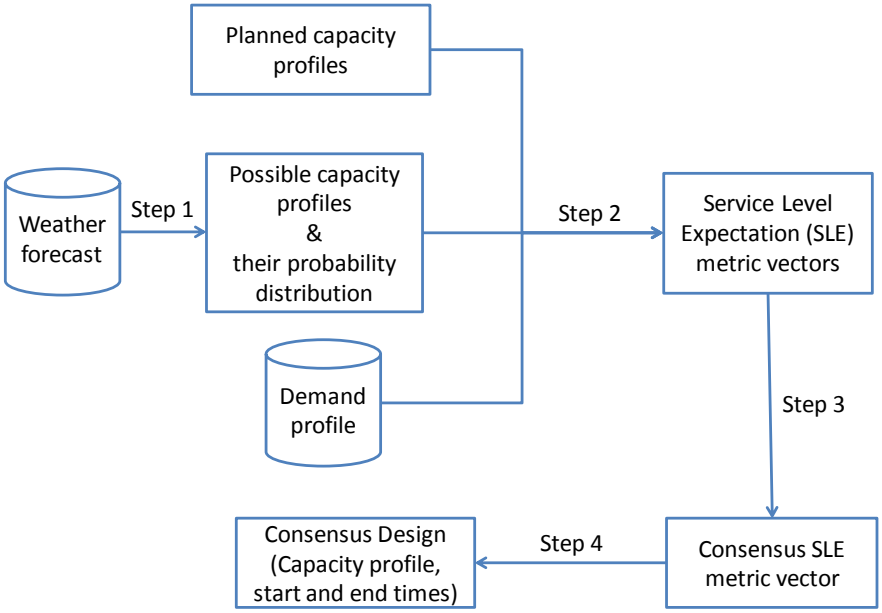


FIGURE 1: Proposed Traffic Management Initiative Planning Process

The proposed TMI planning process consists of four steps. The first step is to generate capacity profiles for the concerned time horizon in the future based on weather forecast. A capacity profile is defined as a time series of capacity rates. Due to uncertainty in weather forecast, there will be a set of capacity profiles that are possible to realize. Work has been performed to identify these capacity profiles and their probability distribution, mainly in the context of GDP planning *(3, 4, 5)*. Step 2 generates a set of SLE metric vectors using the identified capacity profile distribution. Here, we consider demand is known with certainty at the TMI decision time. When designing a TMI, we make a plan on the capacity profile. The planned capacity profile, together with the given demand, can determine other TMI plan parameters, such as TMI start and end times. For flights affected in the TMI horizon, we will assign them plans that are different from schedules. For instance, controlled times of arrivals/departures are assigned to flights scheduled to arrive between start time and end time in the GDP, where the controlled times are usually later than scheduled times of arrival/departure (STAs/STDs). According to the flight schedules and new plans, we can estimate the value of the SLE metric vector for the TMI design. Multiple criteria

can be applied for evaluating TMI performance at the system level (*6, 7*), such as efficiency and predictability, and thus the service level metric vector are multi-dimensional. Because of uncertainty, at the planning stage, we do not know the actual capacity profile that will realize at the end of the day or the actual performance that a TMI plan will lead to. Instead, we can estimate the expectation of the performance by considering the probabilistic distribution of the actual capacity profiles. In other words, we can quantify the SLE metric that a TMI plan is associated with. Different flight operators may have different preferences on TMI plans depending on their airline cost models (*8, 9*) or utilization functions (*10*). As a result, different flight operators will prefer different SLE metric vectors. Step 3 is then performed to find one consensus SLE metric vector out of all the submitted vectors by operators. Swaroop and Ball (*11*) have proposed a voting mechanism to reach consensus in an equitable and confidential way. The last step—Step 4—in the planning process is to identify the consensus TMI plan that provides expected service at the consensus SLE level. One possible way is using Step 2 to making a look-up table between TMI plans and SLE metric vectors, and selecting the plan with corresponding SLE vector closest to the consensus vector as the consensus plan.

The proposed process changes the focus of the discussion between traffic specialists and flight operators while planning a TMI. Instead of discussing on TMI parameters, the discussion will be on SLE metric vectors. It is then of interest to study the SLE metrics and their relationships with TMI plans. In this paper, we define SLE metrics and present a model linking program designs to these metrics for GDP, which is one of the most common TMIs in the United States. GDPs are implemented when there is imbalance between arrival demand and arrival capacity, usually due to adverse weather. The motivation of this program is to transform airborne delay in the terminal space of the arrival airport to ground delay at the departure airports. In the literature, most of studies have focused on minimizing the expected delay cost when designing GDPs (*12, 13, 14, 15*). Different from the cost optimization work, we are not aimed in designing an optimal GDP but demonstrating the relationship between GDP parameters and SLE metrics. Liu and Hansen (*10*) studies this relationship using continuous approximation and deterministic queueing models based on a small set of key GDP parameters. Here, we develop the models based on flight schedules and capacity profiles.

The remainder of the paper is organized as following. In section 2, we will describe the data used in this analysis. In section 3, we will introduce our GDP design model and GDP SLE metrics. In section 4, we will illustrate our methods with two case studies, one for Newark Liberty International Airport (EWR) and one for San Francisco International Airport (SFO). Finally, we conclude our paper in section 5.

2. DATA

The data source required in the methodology is the Metron Aviation Flight Scheduler Analyzer (FSA). The FAS data is based on aggregate demand lists (ADL) of the traffic flow management system. The data provides two types of information for historical GDPs: GDP parameter information and individual flight information. GDP parameter information includes historical GDP decisions for all GDP plans. After the initial implementation, the program may be revised with updated parameters. The ADL data records the planned capacity profiles, issuance times, start times, and end times of the initial GDP plan and the subsequence plans if applicable. The planned capacity profile provides the airport acceptance rates (AARs) for each 15-minute interval between GDP start time and end time. When a GDP is activated at a destination airport, usually not all the arriving flights are delayed. Flights, that are geographically further than the GDP scope or have departure times close to the GDP plan issuance time, are exempted and assigned no delay in the GDP. The ADL data also keeps the record of exemption criteria for the historical GDPs. At the individual flight level, information is available for the scheduled/controlled times of departure and arrival for each flight. When there is a GDP revision, these times are updated accordingly. The ADL data records these times for each GDP plan. More information regarding ADL data can be found in the paper by Liu and Hansen (*6*). The ADL data at SFO and EWR airport in year 2011 is pulled. This includes 177 GDPs at SFO and 155 at EWR.

3. METHODOLOGY

In this section, we will introduce our methodology on generating SLE metric vectors for GDP based on flight schedule, planned capacity profiles and possible actual capacity profiles and their probabilistic distribution. For each planned capacity profile, the SLE metric is estimated as a weighted average of the realized system performances over all the possible capacity profiles that may realize:

$$\overline{M_i^k} = \sum_{j=1}^{J} p_j \cdot M_{i,j}^k$$

where, $\overline{M_i^k}$ is SLE metric for performance goal k with planned capacity profile i; p_j is the probability that the actual capacity profile is profile j and there are J possible capacity profiles; $M_{i,j}^k$ is the realized performance for performance goal k if capacity profile i is planned and capacity profile j is the actual capacity profile. The planned capacity profile may be selected by referring to the possible actual capacity profiles but this is not a necessity. Following this, the SLE metric vector for plan i is written as ($\overline{M_i^1}$,…, $\overline{M_i^K}$), where K is the dimension and equal to the number of performance goals considered.

In order to quantify the SLE metric vector, we need to measure the realized system performances for each pair of planned and actual capacity profiles. In section 3.1, we discuss the system performance metrics that are considered in GDP design. In section 3.2, we describe out algorithm in estimating these metrics and generating the corresponding GDP parameters. In the current algorithm, we focus on expected performance from the initial GDP plan and do not look at the impact from GDP revisions on the performance. After the initial implementation, GDP may end earlier or later than planned depending on actual weather. If weather condition is better than expected, actual capacity is sufficient for executing the initial GDP plan and flights all arrive at their initial CTAs. If weather condition is worse than expected, because capacity was overestimated, some flights will arrive later than their initial CTAs and the unplanned extra delays are taken as airborne delay. More discussion can be found in section 2.

3.1 System Performance Criteria and Metrics

In this section, we present criteria and associated metrics for evaluating GDP plans. Three performance goals are considered: capacity utilization, efficiency and predictability. The criterion for efficiency is the same as in an early work by the authors (*6*). The criteria for the other two goals are defined differently, but the concepts are similar. As in the early work, the criteria are defined to be dimensionless, simple and robust.

Capacity utilization is used to measure how much arrival capacity is utilized in the GDP plan. When implementing a GDP, AARs are planned for each 15-minute interval between GDP start time and GDP end time. Controlled times of arrival (CTAs) are assigned to GDP affected flights according to the planned AARs. The metric of capacity utilization is defined as the ratio of the total number of assigned arrival time slots to the sum of slots that could have been assigned assuming visual meteorological condition (VMC) capacity and infinite demand, over the GDP horizon. The metric is written as

$$M_{i,j}^1 = \alpha_{cu,i,j} = \frac{N_{P,i}}{N_{VMC,i}}$$

where, $\alpha_{cu,i,j}$ is the capacity utilization metric with planned capacity profile i and actual capacity profile j; $N_{P,i}$ is the count of planned arrivals between GDP planned start time and end time when capacity profile i is planned; $N_{VMC,i}$ is the count of arrivals that could have been landed assuming VMC capacity and infinite demand during the same period. Under this definition, the capacity utilization for a given plan is independent from the capacity profile that will realize, and thus the expectation of capacity utilization (SLE metric for capacity utilization) is the same as the value estimated from this metric. The metric

encourage us to maximize throughput and avoid underestimating capacity. As the metric increases, the plan is more optimistic and there is a larger chance that the GDP will end later than planned.

Efficiency is defined referring to the motivation of GDP: transforming airborne delay to cheaper ground delay. There should be no airborne delay is the most efficient GDP. However, given uncertainty in the planning process, airborne delay may be unavoidable in the case of a late weather clearance. The metric is defined the ratio of total realized ground delay to total realized total delay and written as

$$M_{i,j}^2 = \alpha_{e,i,j} = \frac{\sum_k GD_{i,j,k}}{\sum_k TD_{i,j,k}}$$

where, $\alpha_{e,i,j}$ is the efficiency metric with planned capacity profile i and actual capacity profile j; $GD_{i,j,k}$ is the ground delay incurred by flight k for the same pair of capacity profiles; $TD_{i,j,k}$ is the total delay incurred by flight k, equal to realized ground delay plus realized airborne delay. Different from capacity utilization metric, GDP efficiency metric depends on the actual capacity profile and its expectation (SLE metric for efficiency) is the weighted average of efficiency over all possible actual capacity profiles. Also, efficiency degrades if we are optimistic in planning capacity rates, even though this is benefiting capacity utilization. This happens because the efficiency metric aims at transforming more delay to the ground but not minimizing the total delay cost. As a result, efficiency increases when we are more conservative. In the case of a late weather clearance, total delay is larger than ground delay for some flights. Moreover, the actual GDP end time is later than planned and more flights will be delayed. In the estimation of the efficiency metric, we involve all the flights that are delayed in the GDP.

Predictability is defined to capture the accuracy in estimating capacity rates. In the strategic planning telecons, most of the debate is on setting capacity rates. On one hand, we want to make sure available capacity will be effectively utilized. On the other hand, we also appreciate the accuracy of the guess on capacity rates. The former is considered in the capacity utilization and the latter is considered by predictability metric. Using the planned capacity profile, we identify planned GDP start time and end time for the given demand. For the same delay, a hypothetical planned start and end times can be identified assuming one of the possible capacity profile were planned. Define a time horizon with start time as the earlier one out of the two start times and end time as the later one. For all the 15-minute intervals on this time horizon, we calculate the ratio of the minimum of the planned capacity rates and the capacity rates that may realize to the maximum of the two, and then sum these ratios over the time horizon to obtain the value of predictability. Mathematically, the predictability metric is written as

$$M_{i,j}^3 = \alpha_{p,i,j} = \frac{1}{T}\sum_{t=1}^{T} \frac{min(PAAR_{i,t}, AAAR_{j,t})}{max(PAAR_{i,t}, AAAR_{j,t})}$$

where, $\alpha_{p,i,j}$ is the predictability metric with planned capacity profile i and actual capacity profile j; t is the index for the 15-minute interval and T is the total number of intervals; $PAAR_{i,t}$ is the planned airport acceptance rate for interval t given plan capacity profile as i; $AAAR_{j,t}$ is the actual airport acceptance rate for interval t when the actual capacity profile is j. Similar to efficiency metric, predictability of a GDP plan depends on the capacity profile that turns out to be and its expectation (SLE metric for predictability) is the weighted average of predictability over all possible actual capacity profiles. Predictability increases as the accuracy of AAR estimates. Where an aggressive/conservative decision benefits capacity utilization/efficiency, predictability is penalized in either way.

3.2 GDP Design

3.2.1 GDP Planned Start Time and End Time

To estimate the metrics, we first need to identify GDP affected flights, which are the flights scheduled to arrive on GDP time horizon. GDP start time and end time define the beginning and end of the predicted

capacity-demand imbalance. Following this definition, GDP should start when arrival demand starts to exceed capacity and delay starts to develop in the system; GDP should end when delay vanishes and there is no excess demand. As shown in Table 1, the historical observations match the definition for start time reasonably but not for end time. Table 1 summarizes descriptive statistics for historical GDPs at SFO and EWR in 2011. As mentioned earlier, GDPs can be revised after the initial implementation. The statistics presented are for the initial GDP plans. Six GDP characteristics are interested: start queue length—the length of arrival queue at historical GDP start time; first flight delay—delay for the flight scheduled to arrive at start time; end queue length—the length of arrival queue at historical GDP end time; last flight delay—delay for the flight scheduled to arrive at end time; max queue length—the maximum length of arrival queue in the system for a historical GDP; total delayed flights—the number of flights that were scheduled to arrive between start time and end time.

TABLE 1: Descriptive Statistics of the Initial Plans of GDPs at SFO and EWR in 2011

	SFO						EWR					
	Start queue length	First flight delay (min)	End queue length	Last flight delay (min)	Max queue length	Total GDP flights	Start queue length	First flight delay (min)	End queue length	Last flight delay (min)	Max queue length	Total GDP flights
Min	0	0	0	0	13	68	0	0	0	0	11	160
1st quartile	1	1.6	5	6.6	24	157	1	1.9	5	7.8	27	332
Median	2	3.8	13	18	30	192	3	5	13	21.3	35	387
Mean	3	6.2	17	28.6	35	230	5	9.5	23	45.9	44.5	379
3rd quartile	4	7.5	22	32.4	39	247	6	10.6	28	48.8	50	435
95th percentile	8	15	32	67	52	454	10	19	48	90.7	68	470
Max	24	52.5	88	194.9	102	582	39	117	221	825.5*	228	545

* Only two GDPs have last flight delays at this magnitude. Weather was severe on these two days and capacity was low at 16 arrivals per hour for a long period.

On average, at GDP start time, the queue length is 3 for SFO GDPs and 5 for EWR, and the flight delay is 6.2 minutes for SFO GDPs and 9.5 minutes for EWR. These numbers are much larger at GDP end time. Also, there are more cases where the queue length and flight delay are considerably large at GDP end time, according to the 3rd quartiles and 95th percentiles. Theoretically, the start time and end time should be deterministic when demand is known and capacity profile is given. However, uncertainty in the weather forecast may lead to differences between the realized values and planned values for these times. This could have been considered by the traffic specialists when they made the decisions. Compared to end time, start time is closer to the decision time, and can be decided with more confidence. Setting end time is more difficult because it is hours in the future. There are two possible approaches to addressing the uncertainty: set an earlier end time, then extend the program if it is needed; set a longer end time, then cancel the program if needed. The former method benefits capacity utilization, whereas the latter method has less risk in efficiency. The former method may also lose control of the program because there are not enough flights on the ground to absorb delay. The choice depends on the traffic specialist running the program. Either way, the planned end time is more like a placeholder. In addition, there may be ambiguity in understanding GDP end time. GDP end time may also be defined as the capacity recovery time instead of delay clearance time. The difference between these two times could be ignorable if there is not much queue before capacity recovery time. However, delay clearance time should always be later than capacity recovery time because it takes time for delay to vanish after capacity increases. Finally, there could be man-made deviation in the GDP end time data because conventionally a GDP can only end by 15-minute period at 14, 29, 44 or 59 minutes past the hour (*12*). In our data, we observe that the end time for the initial GDP plans is always 59 minutes past the hour. As a result, historical start time can be referred to

for calibrating our criteria for GDP start time but historical end time should not be used for the calibration for the end time.

The statistics of max queue length and total GDP flights indicate that GDPs are planned only when a considerable number of flights would be delayed in the air otherwise. Because of fluctuation in the demand, there are occasions when several flights were delayed with small delays but then delay cleared. GDPs are not necessary for these occasions. Similarly, we cannot simply define a GDP start time as the time when there is 3 (or 8, 95th percentile) delayed flights in the system, or the time when some flight has 6.2 minute (or 15 minute, 95th percentile) delay. To address this, we look at average delay over a certain number of flights. If the average delay over n_t flights exceeds d_t minutes, where n_t is the flight count threshold and d_t is the delay threshold, then the GDP start time is defined as the scheduled arrival time of the first flight in the n_t flights. There are many possible pairs of the thresholds. We select the pair which yields the smallest squared difference between estimated start time and actual start time on average. The objective is formulated as

$$\min_{n_t,d_t} \frac{\sum_{h=1}^{H}(\hat{S}_{n_t,d_t,h} - S_h)^2}{H}$$

where, $\hat{S}_h$ is the estimated start time for GDP h with count threshold and delay thresholds as n_t and d_t respectively; S_h is the actual start time; H is the total number of GDPs. Because there is too much noise in the historical GDP end time data, the thresholds for end time cannot be identified in the same way. We use the same pair of thresholds for end time and identify it using mirror logic: if the average delay over n_t flights drops below d_t minutes, then the end time is defined as the scheduled arrival time of the last flight in the n_t flights.

3.2.2 Arrival Time Slot Assignment

After we define the GDP time horizon and identify the GDP affected flights, we need algorithm to assign time slots to these flights so that we can estimate the performance metrics. Moreover, the assigned time slots are required for calculating average delay for identify start/end times. Two approaches are developed for this purpose, which are referred to as a flexible approach and a fixed approach respectively. The input is scheduled arrival times and planned capacity rates. In the flexible approach, we employ the idea of deterministic queueing model, where the CTA for each flight is calculated as

$$CTA_k = max(CTA_{k-1} + IA_k, STA_k)$$

where, CTA_k is the arrival time slot for flight k and CTA_{k-1} is the arrival time slot for the flights before flight k; IA_k is the inter-arrival time between flight k and the flight before; STA_k is the scheduled arrival time for flight k. The inter-arrival time is determined by the capacity rate. The capacity rate is for each 15-minute interval, and thus the inter-arrival time between flights for each interval is equal to 15 divided by the capacity rate. To be consistent with the ration-by-schedule rule, the flights are sorted by their STAs before assigning time slots. Flight delays are then calculated as the differences between CTAs and STAs.

In the fixed approach, we first generate arrival time slots based on the capacity rate for each 15-minute interval and then assign them to the affected flights according to ration-by-schedule. The inter-arrival time between flights is equal to 15 divided by the capacity rate. For each 15-minute interval, the first time slot is the beginning of the interval, the second time slot is the second time slot plus the inter-arrival time, and the other time slots are assigned accordingly until the end of the interval. The first time slot for the next 15-minute interval is the beginning of the next interval and thus the time slots in the successive intervals are independent. At first thought, the fixed approach seems to be naive but it can be proved to be a very reasonable method for the purpose of estimating SLE metric vectors. Due to limited space, we are not elaborating the reason here.

The slot assignment algorithm is used for locating the CTAs in the initial GDP plan and also the actual arrival times. When actual capacity profile is better than planned, we stick to the plan and the actual arrival times are the same as the assigned CTAs. GDP start and end as planned. When actual capacity profile is worse than planned, we use the initial CTAs as demand and re-assign time slots using the actual capacity rates. The update time slots are the actual arrival times. Since we do not consider revision, the extra flight delay from the revision is taken as airborne delay. GDP will end later than planned when delay vanishes.

We do not consider exemption in identifying GDP start and end times but assigning time slots. When there is exemption, we first assign CTAs to the exempt flights by taking their STAs as the only demand. Then, we assign CTAs to non-exempt flights use the remaining capacity. This is easier for the fixed approach where we simply assign the rest of time slots to the non-exempt flights. It is more complicated in the flexible approach. From the GDP start time, we assign time slots to non-exempt flights using the flexible approach equation. If the assigned time slot is within ±inter-arrival time of a time slot already assigned to an exempt flight, then this time slot will be moved to an inter-arrival time after the assigned slot to the exempt flight.

3.3 Summary

The SLE based GDP design process is summarized in Figure 2. For each GDP design, the SLE metrics are the expectations of the performance metrics. Exemption criteria are set as an exogenous decision variable. Different capacity profiles can be planned and generate different GDP time horizons and different sets of time slots. Using our algorithm, a mapping table can be constructed between GDP designs and their SLE metric vectors. For each selected SLE metric vector, we are able to identify the corresponding GDP design.

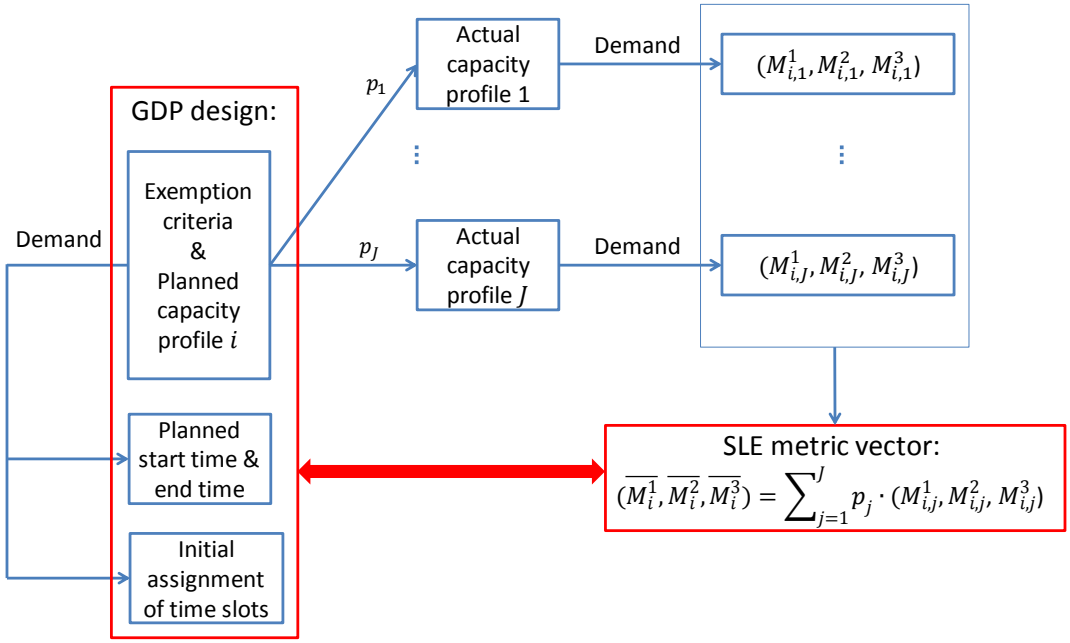


FIGURE 2: Service Level Expectation based GDP Design

4. CASE STUDIES

The information of demand and the actual capacity profile distribution is summarized in Figure 3 for both case studies. Hourly observations are shown but the analysis is performed at quarter-hour level. The EWR demand is from May 23, 2011 and SFO demand is from Aug 29, 2011. For EWR case,we perform analysis to generate the possible capacity profiles by assessing weather forecast similarity between the

given day (May 23, 2011) and historical days. The method is described in an earlier work (*4*) and not presented here. All the possible capacity profiles are equally likely to realize. The capacity profile distribution for the SFO case is borrowed from the work by Mukherjee and Hansen (*5*). The capacity rate is assumed to be 30 arrivals per hour before fog clears and 60 afterwards. So it is sufficient to show the plot of probability mass function of fog clearance time. As seen in the right-bottom plot, the fog is most likely to clear between 10 am and noon. Pilot studies show that there is little difference in SLE metric vectors estimated by the flexible and fixed approaches. We thus perform case studies using the fixed approach since the algorithm is more efficient.

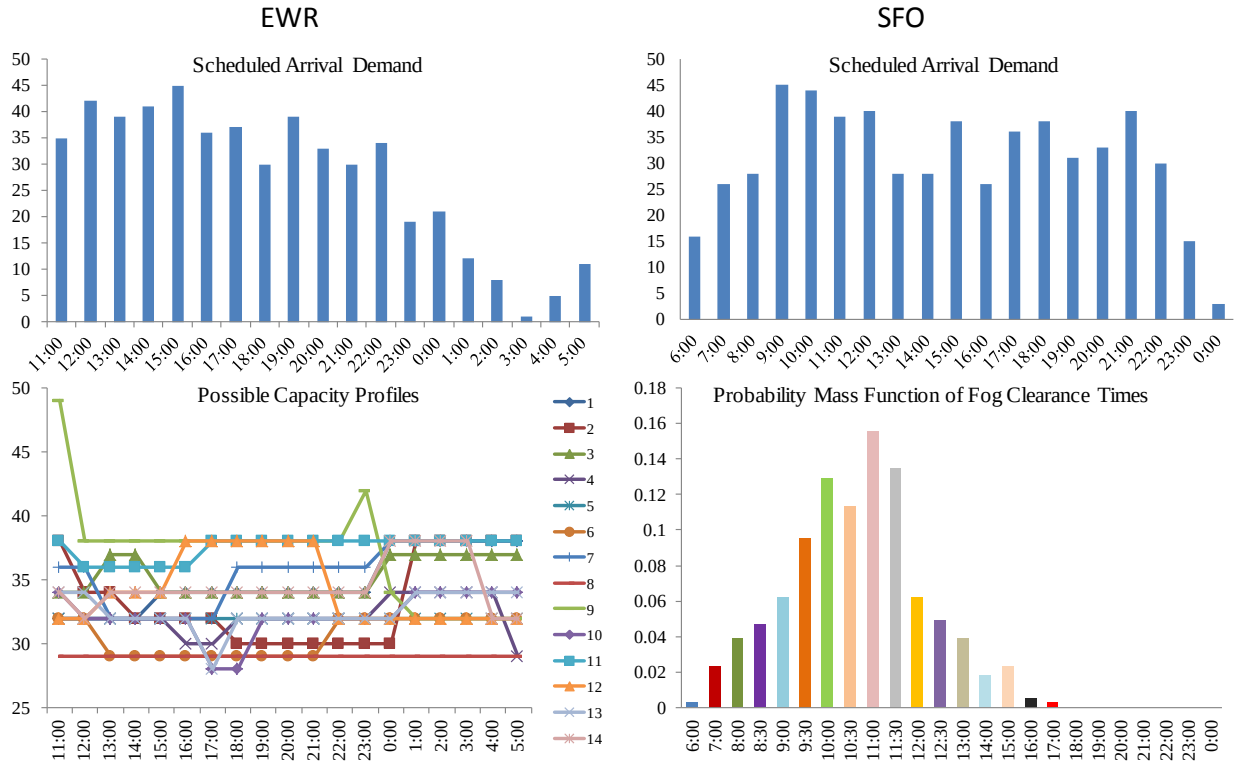


FIGURE 3: Demand and Possible Capacity Profiles in the Case Studies

4.1 EWR Case Study Results

The count and delay thresholds for identifying GDP start time are 15 flights and 14-minute delay. The delay is more than the mean first flight delay in Table 1 for EWR. When queue starts to develop, flight delay increases quadratically before capacity recovers and thus average delay over delayed flights is more than the first flight delay.

GDP designs and their corresponding SLE metrics are summarized in Table 2. GDP issuance time, when the GDP decisions are made and reported, is set as 10 am. Flights are exempted if their scheduled departure times are earlier than 10 am. Each possible capacity profile in Figure 3 is selected as a planned profile. In total, we have 14 planned capacity profiles.

As in the upper-left plot in Figure 3, demand level is high between 11 am and 11 pm. For the high demand period, capacity profiles 9 and 11 have the highest capacity rates and shortest GDP durations if they are selected as planned profiles. It is noticed that capacity rates drops after midnight in the profile 9. However, the demand level is low so no GDP is needed. The worst capacity profile is 8, with 29 arrivals per hour until 5 am. If this is the case, then a 17-hour long GDP is needed. Other cases are in between and most designs are giving long GDPs. This is because we picked a day with severe weather. On this historical day, GDP was planned from noon to midnight. From top to the bottom, capacity profiles become worse, which decreases capacity utilization expectation but usually benefits efficiency

expectation. As we become more conservative, capacity is not effectively utilized but it is more likely that actual weather will be better than expected. As a result, there is a lower chance of airborne delay and efficiency increases. Predictability is higher for 'moderate' capacity profiles, which are located in the middle of the bottom-left plot in Figure 1, such as 4 and 5. Out of the three, efficiency performance has the largest variability.

TABLE 2: GDP Designs and SLE Metrics, EWR Case Study

GDP design					SLE metrics		
Planned capacity profile	exemption ratio	planned start time	planned end time	Planned GDP duration (hr)	Capacity utilization	efficiency	predictability
9	0.09	2:15 PM	8:52 PM	397	0.726	0.267	0.880
11	0.12	12:44 PM	9:31 PM	527	0.716	0.434	0.891
12	0.12	12:03 PM	11:46 PM	703	0.691	0.591	0.914
3	0.10	12:39 PM	12:36 AM + 1 day	717	0.669	0.636	0.920
7	0.10	12:42 PM	12:41 AM + 1 day	719	0.663	0.750	0.915
14	0.11	12:21 PM	1:12 AM + 1 day	771	0.652	0.752	0.925
1	0.11	12:03 PM	1:48 AM + 1 day	825	0.636	0.817	0.927
5	0.11	12:03 PM	2:52 AM + 1 day	889	0.605	0.922	0.933
4	0.11	12:03 PM	2:52 AM + 1 day	889	0.605	0.929	0.932
2	0.09	12:35 PM	2:52 AM + 1 day	857	0.596	0.904	0.913
13	0.09	12:35 PM	2:52 AM + 1 day	857	0.596	0.911	0.931
6	0.11	12:03 PM	3:21 AM + 1 day	918	0.587	0.982	0.902
10	0.10	12:21 PM	3:21 AM + 1 day	900	0.581	0.937	0.928
8	0.13	11:16 AM	4:28 AM + 1 day	1032	0.552	0.992	0.869

4.2 SFO Case Study Results

The count and delay thresholds for identifying GDP start time are 11 flights and 11-minute delay. GDP issuance time is set as 6 am in the morning. In this case, we consider two sets of exemptions: set 1, flights are exempt if their departure times are within 45 minutes of the GDP issuance time or their departure airports are farther than 1000 miles away; set 2, flights are exempt if their departure times are within 45 minutes or their departure airports are farther than 1600 miles away. SLE metrics are estimated for planned clearance times between 6 am and 5 pm every 30 minutes. Results are summarized in Table 3. Where the results under exemption set 2 are put in the parenthesis. We observe that the differences in GDP planned start and end times or values of SLE metrics for the same planned capacity profile with different exemption scopes are trivial. This observation was first found before by Liu and Hansen using an analytical GDP no-revision model (*10*). For the following discussion, we just look at the values with exemption scope as 1000 miles.

A GDP is not called for the planned clearance time earlier than 10 am. Demand exceeds 30 per hour after 9 am, as shown in the top-right plot in Figure 3. If fog burns off and arrival capacity increases to 60 per hour before 10 am, the average delay per flight over 11 flights is less than 11 minutes and no GDP is needed.

When a GDP is called, the planned start time is the same and the planned end time depends on the planned fog clearance time. At SFO, the marine cloud layer develops overnight and delay starts to develop when morning traffic hits the airport. As a result, the planned start time is the same since demand profile is the same.

As we become more conservative in the clearance time, interestingly, capacity utilization expectation first increases then decreases. Compare scenario 2 to 1, clearance time is 30 minutes later but

the end time is 72 minutes later. The time ratio of high capacity duration to GDP duration for scenario 1 is about 0.24 (30/123) and it is about 0.37 (72/195) for scenario 2. Because of this, capacity utilization increases. The difference between clearance time and end time does not change much for different clearance times. However, the duration of low capacity level increases significantly with clearance time. As a result, capacity utilization decreases with clearance time for most of the cases. Similar to the EWR case, efficiency expectation increases when we become more conservative. Predictability expectation is larger for the scenarios with larger probabilities.

TABLE 3: GDP Designs and SLE Metrics, SFO Case Study

GDP design						SLE metrics		
Planned capacity profile	exemption ratio	Planned clearance time	planned start time	planned end time	Planned GDP duration (hr)	Capacity utilization	efficiency	predictability
	—	9:30 AM	—	—	—	—	—	—
1	0.51 (0.39)*	10:00 AM	8:27 AM (8:48 AM)	10:30 AM (10:28 AM)	123 (100)	0.618 (0.670)	0.424 (0.489)	0.860 (0.841)
2	0.57 (0.48)	10:30 AM	8:27 AM (8:48 AM)	11:42 AM (11:42 AM)	195 (174)	0.657 (0.690)	0.634 (0.673)	0.878 (0.868)
3	0.54 (0.46)	11:00 AM	8:27 AM (8:48 AM)	12:10 PM (12:08 PM)	223 (200)	0.673 (0.700)	0.743 (0.776)	0.878 (0.869)
4	0.52 (0.44)	11:30 AM	8:27 AM (8:48 AM)	1:19 PM (1:12 PM)	292 (264)	0.634 (0.659)	0.825 (0.861)	0.882 (0.869)
5	0.52 (0.45)	12:00 PM	8:27 AM (8:48 AM)	1:36 PM (1:34 PM)	309 (286)	0.631 (0.650)	0.874 (0.904)	0.859 (0.850)
6	0.52 (0.44)	12:30 PM	8:27 AM (8:48 AM)	1:58 PM (1:52 PM)	331 (304)	0.626 (0.645)	0.905 (0.930)	0.833 (0.822)
7	0.51 (0.43)	1:00 PM	8:27 AM (8:48 AM)	2:30 PM (2:24 PM)	363 (336)	0.612 (0.634)	0.925 (0.947)	0.820 (0.801)
8	0.51 (0.43)	1:30 PM	8:27 AM (8:48 AM)	2:57 PM (2:52 PM0	390 (364)	0.600 (0.613)	0.936 (0.955)	0.794 (0.781)
9	0.49 (0.41)	2:00 PM	8:27 AM (8:48 AM)	3:38 PM (3:32 PM)	431 (404)	0.603 (0.619)	0.946 (0.963)	0.784 (0.772)
10	0.47 (0.39)	2:30 PM	8:27 AM (8:48 AM)	3:59 PM (3:58 PM)	452 (430)	0.606 (0.614)	0.954 (0.969)	0.761 (0.748)
11	0.47 (0.39)	3:00 PM	8:27 AM (8:48 AM)	4:47 PM (4:45 PM)	500 (477)	0.598 (0.604)	0.961 (0.974)	0.761 (0.750)
12	0.47 (0.39)	3:30 PM	8:27 AM (8:48 AM)	5:08 PM (5:03 PM0	521 (495)	0.589 (0.598)	0.964 (0.976)	0.741 (0.729)
13	0.47 (0.38)	4:00 PM	8:27 AM (8:48 AM)	5:50 PM (5:41 PM)	563 (533)	0.592 (0.604)	0.967 (0.978)	0.736 (0.717)
14	0.46 (0.37)	4:30 PM	8:27 AM (8:48 AM)	6:10 PM (6:09 PM)	583 (561)	0.592 (0.595)	0.969 (0.980)	0.718 (0.706)
15	0.46 (0.38)	5:00 PM	8:27 AM (8:48 AM)	7:11 PM (7:09 PM)	644 (621)	0.589 (0.594)	0.972 (0.981)	0.721 (0.710)

* The values in the parentheses and above are for exemption scope as 1600 miles and 1000 miles respectively.

5. CONCLUSIONS

In this paper, we propose a methodology to connect GDP designs to their service level expectation (SLE) metric vectors. This enables us to inform flight operators of the performances that they can expect from their selected GDP designs. The methodology consists of three steps: GDP start and end times identification, flight arrival time slot assignment and SLE metrics calculation.

Moving average algorithm is used in finding GDP start and end times. When the average delay over n_t flights exceeds d_t minutes, where n_t is the flight count threshold and d_t is the delay threshold, then the GDP start time is defined as the scheduled arrival time of the first flight in the n_t flights. An unconstrained optimization problem is formulated to identify these thresholds. GDP end time is identified using mirror logic with the same thresholds.

Two approaches are presented for assigning arrival time slot: a flexible approach and a fixed approach. The flexible approach is based on deterministic queueing model and more precise. The fixed approach is an approximation and more efficiency computationally. Exemption can be considered in both approaches.

Three performance goals are currently considered for evaluating GDP performances: capacity utilization, efficiency and predictability. Their criteria and associated metrics are defined in section 3.1. For each GDP design, the possible realized performance is estimated when one of the possible capacity profiles is the actual one. Then, the SLE metrics—expectations of the performances—are calculated as the average over all the possible capacity profiles considering their possibilities.

Finally, two case studies are performed to illustrate the idea of SLE based GDP design: one at EWR and one at SFO. It is observed that capacity utilization expectation usually decreases with a more conservative plan, which on the opposite benefits efficiency expectation. Predictability expectation is larger when the selected planned capacity profile is very likely or moderate. In the SFO case, we consider two different exemption scopes. It is found that impact of exemption scope on SLE metric vectors is trivial when GDP revision is not considered.

References

(1) Ball, M.O., C. Barnhart, A. Evans, M. Hansen, Y. Liu, P. Swaroop, and V. Vaze. *Distributed Mechanisms for Determining NAS-Wide Service Level Expectations: Year 1 report*, 2011.

(2) Ball, M.O. *Distributed Mechanisms for Determining NAS-Wide Service Level Expectations: Concept Description*, White Paper, 2013.

(3) Gurkaran, B., M. Hansen. Generating day-of-operation probabilistic capacity scenarios from weather forecasts. *Transportation Research Part C*, Vol. 33, 2013, pp. 153-166.

(4) Liu, Y., M. Seelhorst, A. Pozdnukhov, M. Hansen, M. Ball. Assessing Terminal Weather Forecast Similarity for Strategic Air Traffic Management. 6th International Conference on Research in Air Transportation, 2014, Istanbul, Turkey.

(5) Mukherjee, A., M. Hansen, S. Grabbe. Ground delay program planning under uncertainty in airport capacity. *Transportation Planning and Technology*, Vol. 35, 2012, pp. 611-628.

(6) Liu Y., and M. Hansen. Evaluation of the Performance of Ground Delay Programs. Transportation Research Record: Journal of the Transportation Research Board, No. 2400, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 54-64.

(7) Bradford, S., D. Knorr, and D. Liang. Performance Measures for Future Architecture. Presented at 3rd USA–Europe Air Traffic Management.

(8) Vaze V., C Yan, C. Barnhart, M.O. Ball, and P. Swaroop. Mechanism Design for Setting the Parameters of Traffic Management Initiatives. 20th Conference of the International Federation of Operational Research Societies (IFORS), 2014, Barcelona, Spain.

(9) Yan C., C. Barnhart, and V. Vaze. Is it a good Ground Delay Program? -- Let the airlines decide! 2013 INFORMS Annual Meeting in Minneapolis, Minnesota, October 6-9.

(10) Liu, Y., and M. Hansen. Ground Delay Program Decision-Making Using Multiple Criteria: A Single Airport Case. Presented at 10th USA-Europe Air Traffic Management Research & Development Seminar, Chicago, Ill., 2013.

(11) Swaroop P., and M.O. Ball. Consensus-Building Mechanism for Setting Service Expectations in Air Traffic Flow Management. Transportation Research Record: Journal of the Transportation Research Board, No. 2325, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 87-96.

(12) Cook, L. S., and B. Wood. A Model for Determining Ground Delay Program Parameters Using a Probabilistic Forecast of Stratus Clearing. Presented at 8th USA-Europe Air Traffic Management Research & Development Seminar, Napa, Calif., 2009.

(13) Ball, M. O., R. Hoffman, and A. Mukherjee. Ground Delay Program Planning Under Uncertainty Based on the Ration-By-Distance Principle. Transportation Science, Vol. 44, No. 1, 2010, pp. 1-14.

(14) Mukherjee, A., and M. Hansen. A Dynamic Stochastic Model for the Single Airport Ground Holding Problem. Transportation Science, Vol. 41, No. 4, 2007, pp. 444-456.

(15) Richeta O., A. Odoni. Dynamic Solution to the Ground-holding Problem in Air Traffic Control. Transportation Research Part A. Vol. 28, pp. 167-185. 1994.