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

Predicting & Quantifying Risk in Airport Capacity Profile Selection for Air Traffic Management*

James C. Jones, Richard DeLaura, Margo Pawlak, Seth Troxel, Ngaire Underhill
Air Traffic Control Systems Group
MIT Lincoln Laboratory
Lexington, MA 02420

***Abstract*— There is currently no data-driven approach widely used by air traffic managers and controllers to predict the capacity at airports. Instead controllers rely on rules-of-thumb to define the airport acceptance rate (AAR). As the approach is inherently subjective, it can lead to poor definition of Traffic Management Initiatives (TMIs) which rely on accurate airport capacity estimates and can lead to under-delivery or over-delivery of flights to airports. In this paper we propose a methodology for estimating airport capacity and capacity uncertainty based on the environmental conditions within the terminal and airport arrival routes and the projected arrival demand and aircraft spacing. To make these predictions we used a gradient tree boosting model in which the prediction model estimates are time-lagged and conditioned on the previous states. Additionally, estimates from previously predicted states are also used to condition the model based on the history of the predictor variables. The concept was validated against observations from historical data recorded at Newark Liberty Airport (EWR). The proposed method provides accurate prediction of airport capacity and produces a strong quantification of uncertainty in the form of a prediction interval. To explore the implications of applying information about the capacity uncertainty into planning in ground delay programs (GDPs), a stochastic integer programming model for GDP planning was created using the specific quantiles to define a constraint on airport capacity. This model allows the decision maker to make trades based on quantified levels of capacity deviation uncertainty. The results of a sensitivity analysis suggest that the decision maker may benefit from adopting a modest risk premium when planning GDPs.**

Keywords-Airport Capacity, Capacity Prediction, Ground Delay Programs, Capacity Uncertainty, Stochastic Programming

I. INTRODUCTION

The presence of weather represents an on-going challenge for air traffic managers and controllers and accounts for approximately 60-70% of all delays within the U.S. national airspace system [1]. Not only does it constrain available capacity and hence the ability to use resources within airspace, but the uncertainty associated with location and time of arrival of adverse weather can exacerbate the extent of the disruptions felt at both the strategic and tactical level. To deal with these disruptions, air traffic managers often control demand by imposing traffic management initiatives (TMIs) such as ground delay programs (GDPs), airspace flow programs (AFPs) and the new collaborative trajectory options programs (CTOPs) to limit the flow of traffic into the airports and airspace. At the tactical level, controllers resolve short-term capacity/demand imbalances through miles-in-trail (MIT) restrictions, time-based metering and airborne holding. To support these interventions near terminal airspace, air traffic managers and controllers rely on rules-of-thumb to set the expected capacity in the form of the airport acceptance rates (AARs) at the airport. These estimates are typically not data-driven and may over- or under-estimate the true capacity of the airport when planning TMIs. This estimation error can have a significant impact on the scope and duration of TMIs and lead to poorly designed programs.

In recent years, the level of complexity of the proposed control strategies available to plan TMI programs has increased dramatically. Several proposals have been made to apply speed control as a means of throttling demand to better match capacity in GDP planning [2-8]. Such practices have many benefits for the system including increased efficiency, reduced fuel consumption, improved predictability and flexibility as well as a more equitable allocation of resources to carriers. The state of fielded metering systems such as Time-Based Flow Management as well as the flight deck avionics has also matured to the point where they can be acceptably used to facilitate metering in trajectory-based operations [9-12].

Recently, these concepts have been studied and demonstrated on the NASA Integrated Demand Management (IDM) program [13][14]. The effort aims to preemptively resolve the imbalances between the demand from flights and the capacity at constrained resources by pre-conditioning traffic demand at the strategic level through the use of the Traffic Flow Management System (TFMS) to facilitate a more manageable arrival stream near the terminal. The demand was conditioned by using a CTOP to control flight arrival times to two boundaries: an inner circle Flow Constrained Area immediately surrounding the airport, making the program in

* This material is based upon work supported by the National Aeronautics and Space Administration under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration.

some ways similar to a GDP, as well as an outer Flow Evaluation Area at 400 NM where TBFM assumes control of the flights. The demand was also controlled at the tactical level with TBFM via extended metering initiated at 400 NM away from the airport. The complementary use of traffic flow management tools revealed many significant benefits including a reduction in the amount of ground delay assigned by TBFM, a drop in the amount of congestion in en route and terminal airspace following TBFM assignment and a more equitable distribution of delay across flights. The approach also reduced the extent of the double-delay penalty imposed by TBFM on short-haul flights previously identified in [15].

Another significant observation of the IDM program was that the capacity predictions issued by TFMS did not match the potential delivery rate of TBFM. This discrepancy resulted in an under-delivery of flights to the airport. Such an illustration suggests that the methodologies for predicting capacity in automated decision support systems must be enhanced to better facilitate interoperable system concepts in order to realize the full benefits of NextGen/SESAR. To support this need for an improved articulation of airport capacity both on the IDM program as well as in TMI decision making in general, researchers at MIT Lincoln Laboratory attempted to develop a new methodology for predicting AARs and quantifying uncertainty in the context of TMI planning.

A variety of approaches have been considered in the literature to address the inaccuracies in current capacity estimation practices. One line of research attempts to optimize the flight demand to match airport capacity through the use of scenario planning in ground delay programs [16][17]. Previous work in support of such goals has focused on generating scenarios for airport capacity by sampling from scenario-trees over time [18]. This work has evolved to utilize weather predictions by clustering day of planning Terminal Aerodrome Forecasts (TAFs) with historical forecasts [19] to generate similar capacity profiles to ones that have previously occured. Another line of research aims to develop decision support systems to enhance situational awareness and improve real-time decision making for air traffic controllers and managers working in the operational environment. Such aims have utilized a variety of weather products and methods to facilitate improved prediction. Tien et al employed the use of ensemble weather forecasts to predict AAR [20]. Kicinger et al proposed a model that incorporates the time-lagged High Resoluation Rapid Refresh (HRRR) forecast. The model attempted to calibrate out the inherent biases in the forecast by fitting forecast data products againts the observed weather [21].

An approach proposed by researchers at MIT Lincoln Laboratory combined statistical wind and traffic information with inputs from the air traffic control community to enhance the reliability of AAR predictions [22]. While the approach proved generally predictive of AAR levels, it missed a number of wind shift events that would have ultimately led to significantly different AAR estimates. Such events often drive the onset, duration and revision of both tactical and strategic intervention.

Along these lines others have proposed models that condition the resulting predictions on previously observed airport capacity levels [23-27]. These approaches have also been used to augment efforts in GDP prediction [28]. In addition to the direct estimation of capacity, others have proposed the use of discrete-choice models to predict runway configurations in support of the overall process of capacity estimation [29][30]. While the descriptive models presented in [20-27] fulfill the aim of providing the decision maker with useful information about the future state of the system, they do not effectively capture the uncertainty present in the information that is to be acted upon. On the other hand, the prescriptive models proposed in [16] and [17] leverage such information to produce planning decisions based on a set of generated profiles. The models are not, however, explicitly designed to directly map the proposed decision to a specific level of capacity profile uncertainty. In the operational environment it may be desirable to provide greater transparency about the assumptions associated with the uncertainty of decisions when presenting specific TMI strategies to the user. This philosophy has been instrumental to the development of decision support systems designed to facilitate the translation of convective weather to airspace capacity to support AFP planning in the en route phase of flight [31].

In this paper we propose a new methodology for predicting AARs and quantifying the uncertainty associated with capacity profile forecasts. The methods build upon the developments described in [23][24] adapting a time-lagged prediction model that incorporates the TAF weather forecasts, flight schedules and conditions its predictions on estimates of previous airport states. Unlike in other approaches to augment the model, an analysis of the environmental wind conditions along the airports' Standard Terminal Arrival Routes (STARs) was performed to produce a dataset that could supply additional predictor variables. Our methodology also yields a prediction interval to quantify the uncertainty in estimating future AAR values. To illustrate the applicability of our approach, the resulting quantiles are incorporated into a modified version of the GDP planning model presented in Ball et al [16] in the form of a chance constrained programming model. This model is then tested against different quantile constraints.

In Section II we describe the details of our modeling architecture and the features utilized. We also present a set of metrics developed to support an airport site adaptation. These metrics provide an additional set of prediction variables that were used to inform the model. In Section III we present some background information on stochastic integer programming for ground delay programs and present our proposed approach to better account for risk under current GDP planning practices. In Section IV we present a case study in which the proposed methods are validated against data from Newark Liberty Airport (EWR). The results demonstrate that our approach provides strong estimation of airport capacity and the associated capacity uncertainty. The section also describes how such profiles can be used to gauge the potential impact of assuming various levels of capacity uncertainty when designing GDPs.

II. METHODOLOGY

A. Model Architecture

A gradient tree boosting regression model was used to predict the airport acceptance rate based on inputs from terminal weather forecasts, traffic records, arrival route wind forecasts and the time of observation. This method is part of a larger class of non-parametric supervised learning methods known as decision trees. These methods aim to produce a set of predictions of a target variable by learning a set of decision rules derived from the input features and splitting the data categorically into several subsets based on the decision rules. Once the data is partitioned the model performs a regression on each subset to assess the fit.

The proposed method uses an ensemble of decision trees to predict the data and weight the predictions made by each tree to make a final estimate. To learn the parameters of the model, the method minimizes the cumulative loss over all points using a gradient descent algorithm [32]. A similar approach was used in [23] and [24] with other decision tree methods. While there is some latency associated with the predictions, the approach has strong tracking with the target observations. Unlike other methods, however, our model issues a prediction interval based on user defined quantile values.

The major sources of data for the model come from a set of airport and terminal weather metrics collected in the Airport System Performance Metrics (ASPM) and TAF data products, as well as a time-lagged HRRR forecast dataset that was processed to extract the relevant derived features relating to wind speed and encroachment along the arrival route between leading and following aircraft. The raw airport and terminal weather features as well as the site-specific derived features were fed into the model such that its inputs at each time step included the data for both the current time step and the previous time step. Prior predictions issued by the model were also used to supplement the feature set. A depiction of the process is shown in Figure 1.

A set of weather predictions was issued each hour of the period of interest. Thus, to make predictions that were viable for an 8-hour look-ahead, the model leveraged 8 different pairs of TAF and arrival route wind forecasts. Each one was valid for a specified look-ahead time period but issued prior to the time that it was used. For example at 7:00am a pair of TAF and arrival route weather forecasts could be used to make the initial prediction of the AAR at 8:00am, along with the other model features, provided both forecasts were issued prior to 7:00am. These derived predictions could then replace the observed AAR at 7:00am when used with the 2-hour forecast (the forecast for 9:00am taken prior to 7:00am). Similarly, when a prediction is made for the AAR 3 hours ahead of time, the prediction of the 2-hour AAR would replace the prediction of the 1-hour estimate as the relevant model input. By incorporating an estimate of the AAR for the prior hour rather than directly using the current AAR we help to improve the level of time correlation between the feature variables and the model predictions when we are predicting the AAR many hours into the future. This process continues to update the predictions until a final prediction for an 8-hour forecast can be issued. Figure 2 shows an illustration of the process.

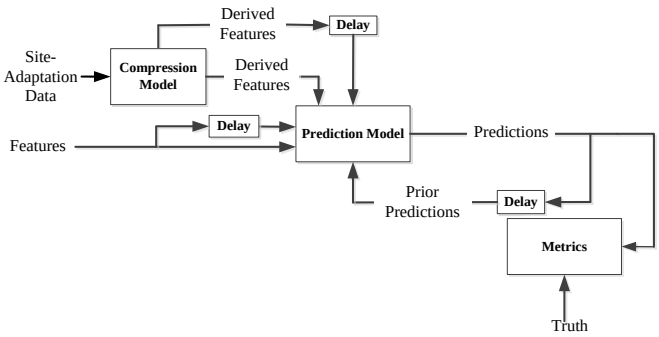

Figure 1. A representation of the data flow within the model.

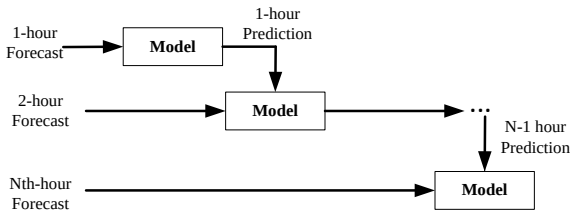

Figure 2. Weather forecast propagation to produce capacity estimates.

B. Model Features

The TAF, ASPM and HRRR datasets were all leveraged by the model to provide a useful set of predictors of the AAR. The raw data fields from ASPM and TAF were directly used while relevant information from the HRRR dataset were extracted and processed to provide additional predictors. These HRRR-based predictors and the process for deriving them are explained in the next section. The selected model features from the ASPM and TAF dataset include:

Wind Speed and Direction: Wind speed and direction are prominent factors in the determination of airport runway configuration. As airports are generally configured to allow flights to land into headwind conditions, the presence of strong tailwinds and/or crosswinds can have a significant impact on the operational status of individual runways. In airports with overflow runways, the presence of unfavorable headwinds or crosswinds can shut down the runway and significantly reduce the overall capacity of the airport.

Wind Gusts: Wind gusts can be quite hazardous to aircraft in the immediate vicinity. When they are present, additional spacing requirements may also be prescribed to limit the effect of encroachment between leading and trailing aircraft under such conditions. As such when present, wind gusts can have a significant effect on terminal airspace capacity.

Ceiling and Visibility: The meteorological conditions are captured in the parameters of visibility and cloud ceiling. The capacity of airports is typically highest when pilots can maintain visual separation from other aircraft. When the airport is operating under Visual Meteorological Conditions (VMC), the level of visibility is sufficiently high to allow pilots to maintain visual separation. On the other hand, when Instrument Meteorological Conditions (IMC) is present, pilots must rely

on their flight instruments on board the aircraft. The TAF data set contains information on the ceiling levels, state of the clouds and the range of visibility. These fields can be used to gauge the meteorological conditions at the airport.

Demand: In addition to the TAF data, the ASPM dataset contains information on the number of flights scheduled to arrive and depart each hour. These features were used in our model as a proxy for the airport demand in lieu of more recently available information on the status of the arrival and departure banks.

Time of Day: At almost any airport the volume of traffic will vary throughout the day. During night time hours, the traffic will subside and pick-up as the day resumes. To control for this effect we included the hour of the day in which the observation was recorded as a feature. As traffic volume is typically lighter over the weekend we also included the day of the week as a feature.

Estimated Features: In addition to the recorded measurements in the data sources, a set of estimated features was applied to enhance the predictive capability of the model. These variables included the previous airport acceptance rate, the previous presence of a VMC or IMC state and the previous airport departure rate (ADR). Previous studies have shown the AAR to be strongly dependent on its previous state [23][24]. In this study we sought to explore an additional set of predictors that may also influence the AAR. As AARs and ADRs are often tightly coupled, we hypothesized that a conditional dependence between the two was reasonable. The previous runway configuration was also explored as a predictor, however, it was not found to be a significant contributor in large part due to our use of the previous AAR variable as an estimator. Although meteorological information is present in the TAF data in some instances the fields for ceiling were not populated. To mitigate against the effect of missing data, we also included estimates of the previous meteorological conditions.

C. Adapting Site Specific Environmental Characteristics

There are a number of operationally relevant factors affecting the sequencing and spacing of flights when they reach the terminal area that can affect the airport's ability to accommodate incoming flights. In windy environments, path-based wind shear and wind gusts along the arrival route can cause a trailing aircraft to encroach upon a leading aircraft due to resulting ground speed differences. This phenomenon is known as compression. To ensure that flights maintain a safe distance at or beyond the required wake-vortex separation, air traffic controllers often assign an additional miles-in-trail buffer between aircraft. While assuring safe operation, this practice can have a significant effect on the number of flights that arrive at the airport. A site adaptation was performed at Newark Liberty Airport EWR to account for the impact of changes to the winds along the airport STAR paths. The airport has two parallel runways and an additional crossing runway primarily used to support overflow conditions. This layout has been actively used in operations from 2002 to the present. A map of the airport surface is shown in Figure 3.

The wind impact metrics are derived from weather forecast model winds sampled along the STARs and from strategically positioned 20 × 20 nautical mile "capture boxes" that capture characteristics of the winds at critical locations along the nominal arrival trajectories. These "capture boxes" were identified based on input from subject matter experts. The proposed characteristics are a set of headwinds and headwind differences along one or more trajectories that are related to ground speed differences among merging aircraft, compression, and difficulties in maintaining optimal spacing. Headwinds associated with each capture box are defined as the average value of all forecast or analysis grid points within the capture box. A diagram of the airport site adaptation under active runway 04 conditions is shown in Figure 4.

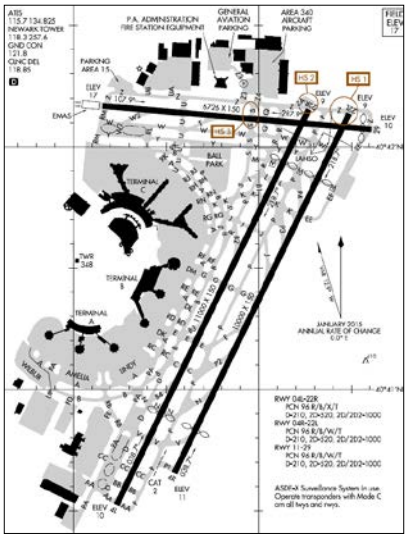

Figure 3. Layout of Newark Liberty Airport (EWR).

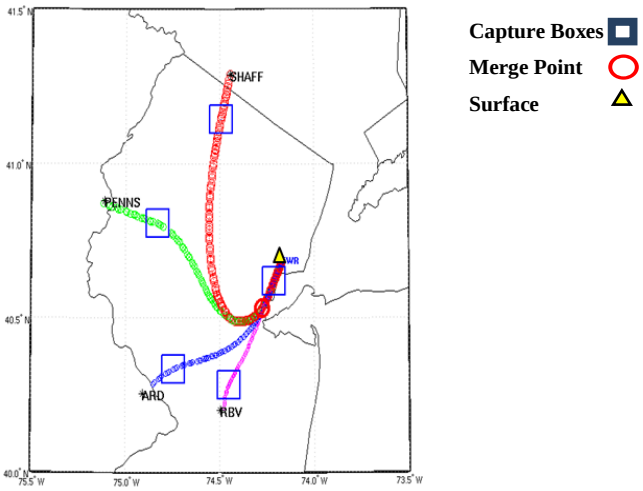

Figure 4. Nominal wind measurement locations for EWR Site Adaptation for Runway 4.

The metrics describing the effect of the near terminal winds on flights were derived from a set of wind forecasts which were generated using various numerical weather prediction models. These wind forecast sources included the HRRR, the Rapid Refresh model (RAP) and the Rapid Update Cycle model (RUC) datasets. A set of derived impact metrics were calculated from analysis fields with a model that analyzes wind measurements at various points along the

airport arrival streams. The wind impact metrics are defined as follows, with altitude ranges as configured for the EWR site adaptation:

DCB Headwinds: The Downstream Capture Box (DCB) identifies the segment of the arrival trajectory where the merged arrival streams are set up for final approach. There is a different DCB for each arrival runway approach. DCB altitudes for different approaches range between approximately 1.5 and 2.0 kft. Headwind measurements were taken along the appropriate STAR trajectory over the region of coverage encapsulated by the capture box.

DCB-to-surface headwind difference: The DCB-to-surface headwind difference measures the difference between forecast headwinds at the two locations. The metric relates to the likelihood of compression between the surface and 2.0 kft.

Mergepoint-to-DCB headwind difference: The merge points are where arrivals from different directions (north, west, south) merge into a single stream for preparation for final approach, and are associated with merge point capture boxes. Each arrival runway approach has a different merge point. Merge altitudes range between approximately 2.5 and 3.0 kft, so the merge-to-DCB headwind difference relates to the likelihood of compression roughly between 2 and 3 kft.

Headwind at TRACON entry capture box: The aircraft ground speed at TRACON entry, dependent in part on headwind, represents the initial condition for TRACON flow management. High tailwinds at TRACON entry and the resulting high ground speeds, may present significant challenges to TRACON controllers as they try to reduce aircraft ground speeds on final approach to acceptable levels. There is a TRACON entry capture box for each STAR/arrival runway combination. STAR entry capture boxes encompass altitudes ranging approximately between 5.0 and 7.0 kft.

Maximum merge headwind difference: Excessive differences in headwinds and the resulting differences in ground speed increase the difficulty of merging traffic from different STARs onto final approach. As such it is not unreasonable to expect controllers to increase the buffer spacing between aircraft thereby lowering the capacity of the airspace and reducing the AAR.

Maximum STAR-to-DCB difference: This metric is a rough measure of the possibility of compression in the approximate altitude range between 7.0 and 2.0 kft.

Maximum segment gain: Compression segments are defined as segments of the arrival trajectory along which the headwind increases monotonically. Compression segment headwind gain is the total increase in headwind from the beginning to the end of the segment. Compression segments may be defined for each STAR/arrival runway combination. This metric provides a rough measure of the possibility and severity of compression anywhere along the arrival trajectory. An example of a compression segment identified by the tool is shown in red in Figure 5. In this instance there is a monotonically increasing headwind toward the end of the arrival route.

Maximum compression segment headwind gain: This difference is analogous to the maximum merge headwind difference, and gives a sense of the potential difficulty of maintaining acceptable spacing while merging arrival streams.

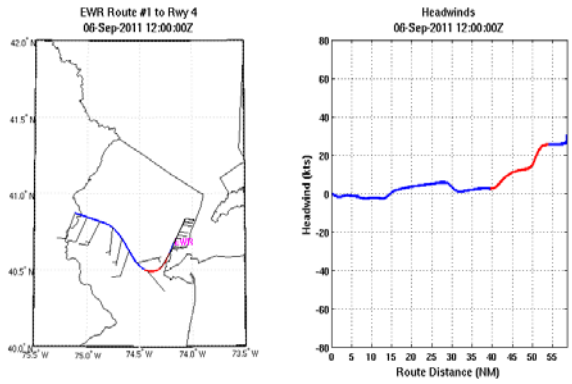

Figure 5. Nominal compression segment for EWR Site Adaptation for Runway 4.

III. Extensions to GDP Planning

In the previous section we described a model designed to predict the AAR and the uncertainty associated with the estimate. This in and of itself might provide useful information to the ATC community, however, one of its most promising potential applications is its extension to planning GDPs. Owing in large part to the persistent problem of delay a number of approaches have been proposed to help facilitate improved GDP planning. Such approaches vary from decision support tools that aim to provide better information to the decision maker to models that seek to optimize some objective given a set of constraints. Examples of such models proposed in [16] and [17] attempt to optimize the assignment of ground and air delays over a set of capacity scenarios. In order for these models to be effective the user needs to have some idea of what set of capacity profiles are appropriate for the operational environment. When the specification of capacity is accurate, the user can account for a wider range of possibilities and hedge against a range of uncertain futures. While several procedures have been proposed [18][19][23][24],[33] the task of generating such scenarios and incorporating them into a control strategy is not trivial and the resulting distributions may not be intuitive to many human decision makers. As such the practice has been slow to gain traction in field operations.

Since one of the primary goals of the paper is to provide a means of quantifying the impact of assuming a specific risk tolerance under current planning practices, we will not utilize such approaches for our evaluation. Instead we propose a different model that expresses the capacity bounds on the constraints using the quantiles of the capacity profile distribution. This model is shown below:

Parameters:

$T \equiv$The set of all time periods

$Q \equiv$The set of all scenarios

$X_t \equiv$The planned airport acceptance rate at time period t

$G_t \equiv$The number of flights held on the ground during time period t

$W_{tq} \equiv$The number of flights held in the air during time period t under scenario q

$D_t \equiv$The demand at time period t

$c_a \equiv$The cost of holding in the air

c_g≡The cost of holding on the ground
p_q≡The probability of scenario q
V_t≡The capacity at time period t
n_{tq}≡The number of unplanned flights arriving at time period t in scenario q

Demand-Based Scenario Model

$$\min \sum_{t\in T} c_g G_t + \sum_{t\in T, q\in Q} p_q c_a W_{tq} \quad (1)$$

$$s.t. \quad X_t + G_t - G_{t-1} = D_t \quad t \in \{1,...,|T|\} \quad (2)$$

$$X_t - W_{tq} + W_{t-1q} + n_{tq} \le V_t \quad \forall q \in Q, \forall t \in \{1,...|T|\} \quad (3)$$

$$G_o = 0 \quad (4)$$

$$W_{oq} = 0 \; \forall q \in Q \quad (5)$$

$$X_t, G_t \in Z_+ \forall t \in T, W_{tq} \in Z_+ \forall t \in T, \forall q \in Q \quad (6)$$

Equation (2) is a network flow queueing constraint that states that the demand at time t should be satisfied such that flights scheduled for take-off during that period take-off or be delayed on the ground. Equation (3) states that flights in the air should either be allowed to land or be delayed in the air until the next time period based on the available capacity. Equations (4) and (5) state that initially there are no flights held on the ground or in the air. Equation (6) says that all variables are positive integers.

Although the model bears some similarity to the one proposed in [16] its treatment of uncertainty is unique. Unlike the approaches shown in [16] and [17] where a set of sampled scenario capacities are used to represent the bounds on the number of aircraft that can arrive at the airport in any single time period, the capacity bound in our model is represented by a specified quantile chosen from the forecast of the capacity distribution. By using a quantile rather than a set of scenarios to define the capacity we are able to significantly reduce the number of constraints needed to account for the effect of capacity uncertainty.

Due in part to the reduction in the number of capacity bounds utilized, the model is also able to more readily account for the influence of demand uncertainty on arriving flights. This demand uncertainty represents the variation in arrivals due to pop-ups and schedule drift over time. The uncertainty is captured in equation (3) by using the parameter n_{tq} to define the number of unplanned flights arriving in each period. The objective of the problem is to minimize the expected total cost of air and ground delay using the capacity profile of the specified quantile while controlling for the demand uncertainty.

From the standpoint of capacity uncertainty, the approach described in the problem above corresponds to a chance constrained program in which the decision maker assumes a level of risk and optimizes based on that risk tolerance [34]. In this case the risk is quantified by the extent to which the observed capacity will exceed the assumed capacity. For example, if we chose the 85th percentile capacity, the true capacity would ideally exceed the measured capacity 15 percent of the time. This customization of user preference comes at a price as an inaccurate specification can lead to significant over- or under-delivery of flights. Given sufficient experience, however, the decision maker will gain a better sense of how to modulate and mitigate risk based on personal preferences.

While we presented this characterization of risk in the form of an integer program, the application need not be limited to such a format. One could notionally use such a metric in a fast-time simulation or real-time Human-in-the-Loop (HITL) experiment. In such a context the metric could define a goal delivery rate for the specified TMI. The decision maker could then adaptively manage the mismatch through revision of the program or the application of additional metering when the assumed capacity is violated. The metric could also be used for decision support in the operational environment. In this context, such a metric provides quantifiable goals to which traffic managers could aspire.

IV. RESULTS AND DISCUSSION

A. *Exprimental Description*

To conduct our studies we selected data collected at Newark Liberty Airport (EWR). The input data used fields from the ASPM, TAF and derived HRRR measurement metrics along each STAR trajectory. At each point the model issued predictions of the AAR, ADR and the presence of VMC or IMC weather. The model look-ahead horizon was set to a period of 8 hours consistent with strategic ATM. Thus eight TAF forecasts were used to make the predictions of the final AAR and ADR at each testing point. Each prediction was propagated forward 1 hour in time to make the ensuing prediction. The model was tested against data collected from January 1, 2014 to March 31, 2014 and trained with data from the last three months of 2013.

A python script was written to perform the analysis. The script leveraged the gradient tree boosting regression module in the scikit-learn package [35]. While the AAR is a discrete quantity the consequences of misclassifying an AAR with a much higher or lower value are more significant than misclassifying an AAR in close proximity to the prediction. To capture this relationship we argue the use of regression is more appropriate than classification in this context as it better accounts for the scope of error.

The model was trained and tested for two different situations. In the first, the decision maker wants to have some prediction of the AAR. In the second the decision maker would like to design a GDP. To tailor to the two circumstances we used different data sets to evaluate the model. For the AAR predictions we used the entire dataset over the period described above. In the use case to support GDP design, the model was tested and trained against days where GDPs actually occurred. While some might argue that it would be more appropriate to train against all of the days we emphasize that the objective of this study was not explicitly to predict GDPs but to design an effective capacity model to support GDP planning. Under these circumstances we assume that the decision maker knows that a GDP will occur because he/she is planning for it, however, the decision maker may need assistance in designing the structure of the program for implementation.

B. *Model Estimation Performance*

The resulting AAR predictions from the proposed approaches were aggregated to measure the performance over a 3 month span. In order to measure the spread of the data we calculated an RMSE score. The metric corresponds to the sample standard deviation between the predicted and observed measurements as shown in equation (7). A plot of the RMSE values for the total set of days as well as the GDP days is shown in Figure 6.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{x}_i - x_i\right)^2}{N}} \quad (7)$$

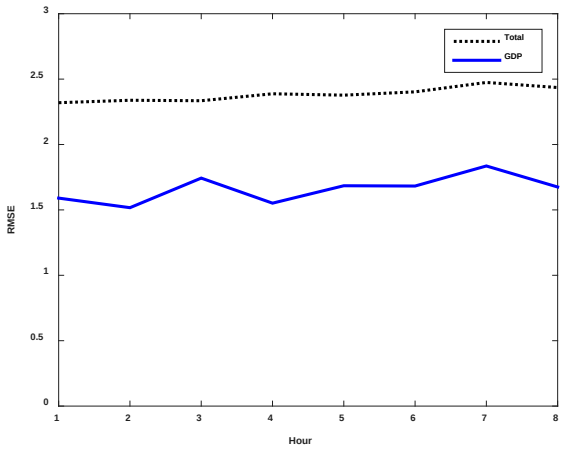

Figure 6. Evolution of RMSE score vs. the look-ahead horizon.

The trend of the plot suggests strong consistency across changes in the forecast horizon. This consistency is apparent in both the predictions for the total set of days and those where GDPs were implemented. The RMSE score for the total set of days is significantly higher than for the set of GDP days. This discrepancy is not entirely surprising as the total set of days incorporates a broader range of capacity profiles. We suspect that high variance between capacity on days in which ground stops were issued and days where the airport operated at its nominal rate signficantly contributes to raising variance relative to what was observed on the set of GDP days. When such tactical intervention occurs it injects high variability between the baseline AAR and the realized AAR.

Since the GDP models were trained on days in which GDPs were implemented, the model is only valid for use once the decision maker knows that a GDP is going to be implemented. Under these limitations, however, the decision maker can still act effectively once a decsion to call a GDP has been made. Notionally he/she could run the prediction model to design and estimate the capacity over the duration of the day. Alternatively, the model could be tuned to provide other metrics of interest such as the median value from the distirbution or other quantiles. In the manner, the model proves itself a vehicle not just for the estimation of the true AAR but also the model uncertainty.

C. *Quantifying Model Uncertainty*

A prediction interval was generated to assess the ability of the predicted quantiles to capture the target data. In order to evaluate the performance of the prediction interval two metrics were used: prediction interval coverage (PIC) and prediction interval width (PIW). The PIC measures the percentage of time that the target data lies between the two bounding quantiles that define the interval. An expression for the metric is shown below:

$$PIC = \frac{1}{N}\sum_i I\left(f_{q1}(\hat{x}_i) \le x_i \le f_{q2}(\hat{x}_i) \right) \quad (8)$$

In this expression $f_{q1}(\hat{x}_i)$ and $f_{q2}(\hat{x}_i)$ represent the two quantile bounds of the prediction interval while x_i is a random variable. The function I is an indicator function which equals 1 if the target sample lies between the two bounds and 0 if the target lies outside the bounds. Ideally, we would like to see a capture rate between the two quantile bounds equate to at least the difference between the percentiles they represent. Thus for an 80 percent prediction interval we would like the see that the bounds capture the true AAR at least 80% of the time. In many cases it is not possible, however, to achieve an exact correspondence between the range of the prediction intervals and the capture rate because much of the data is concentrated at the nominal AAR level. In these cases the interval will often cover more of the data than the stated size of the interval.

In addition to PIC, another metric known as prediction interval width is also used to evaluate the quality of the prediction interval. An expression for PIW is shown below:

$$PIW = f_{q1}(\hat{x}_i) - f_{q2}(\hat{x}_i) \quad (15)$$

The expression defines the width of the interval. When the width of the interval is sufficiently large, it may achieve strong coverage, however, the value of such information is quite trivial. Thus, in order to evaluate the quality of the prediction interval it is important to consider both PIC and PIW as a pair. The performance of our model with varying forecast horizons is shown for the two metrics in Figure 7 and Figure 8. For this experiment we set the prediction interval to 80%. As apparent in Figure 7 we see that the PIC achieves strong adherence to the target level for both models. In Figure 8 we see that the PIW is significantly smaller for the GDP case than for the set of all days. This was to be expected as the RMSE score for the GDP case was lower and the two typically exhibit strong correspondance. Similarly the PIC and PIW behave consistently over time. The fact the PIC remains at 80% suggests that the model may translate well to GDP planning. To provide some additional support for this conjecture we also ran the model to generate prediction intervals for 50%, 60%, 70%, 90% and 95%. The results of this parametric sweep are shown in TABLE I. The 50%, 60% and 70% all achieve capture rates well beyond their interval size. We attribute this phenomenon to the quantization of modes on GDP days. In most cases in the data the AAR is positioned at call rates of 32, 36, 38 and 40 aircraft per hour. We also suspect the lack of training days to be a potential contributor to the greater than expected coverage. Further validation against more historical data is needed, however, to confirm this hypothesis. In the

upper quantiles the model works fairly well at capturing the intended range of coverage. This observation suggests that at EWR decision making might be better suited to tailor capacity to the upper quantiles when facilitating GDPs.

TABLE I. AAR PREDICTION INTERVAL PEFORMANCE WITH VARIATION IN INTERVAL SIZE

Prediction Interval Size	*PIC*	*PIW*
50%	71.5%	2.05
60%	77.3 %	2.11
70%	80.2%	3.60
80%	81.5%	3.71
90%	86.5%	6.18
95%	94.2%	15.6

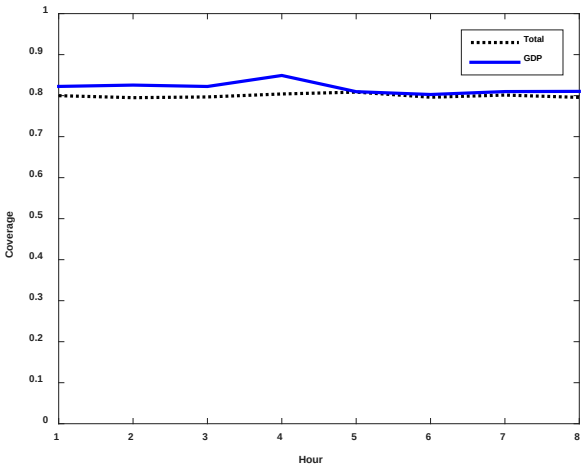

Figure 7. The evolution of AAR prediction interval coverage over a 8-hour look-ahead horizon.

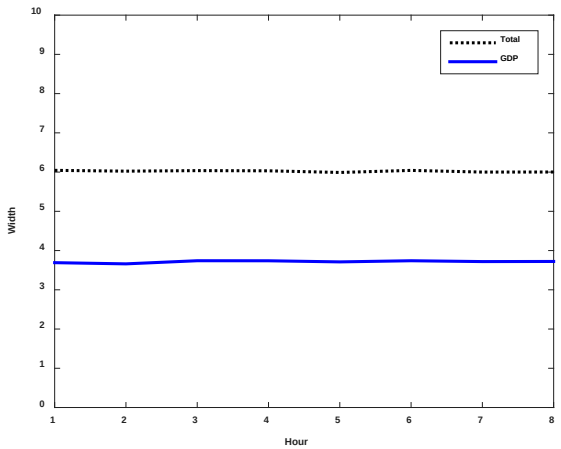

Figure 8. The evolution of AAR prediction interval width over a 8-hour look-ahead horizon.

D. Assessing the Impact on GDP Planning

The quantile bounds on the prediction intervals may not just provide a good indicator of the uncertainty within the data but could potentially be directly used to set the capacity profiles. In order to evaluate the operational impact of leveraging the profile associated with each percentile, the demand uncertainty-based scenario model proposed in section III was adapted over a set of GDPs. Eight GDP events were selected during the three month period over which the data was collected. NTML records were used to identify candidate GDPs. Profiles were constructed based on the evolution of the 75th, 80th, 90th, 95th, 97.5th and 99.5th quantiles. The model incorporated 100 sampled scenarios in which we assumed the demand could be perturbed by +/-2 flights every period in 15 minute intervals. As airports are typically capacity-constrained during the implementation period of a GDP, the percentile that served as the lower bound on each prediction interval was used to set the capacity. A plot of the variation in the expected cost of ground and airborne holding vs capacity quantile bound levels for the eight selected GDPs is shown in Figure 9.

The plots demonstrate an increasing but not monotonic trend in costs with percentile. This is not surprising as the larger percentiles correspond to reduced capacity which forces the system to take on more delay. In most cases the costs appears relatively stable between the 80^{th} and 90^{th} percentiles. This feature is largely reflective of the high level of coverage established for interval sizes of 70% or more, although in GDP8 the cost begins to increase sooner

To better understand the implications of adopting a specific prediction interval, the cumulative cost of assuming each capacity was calculated. This cost included the total expected cost of the objective function described in equation (7) as well as the cost of mismatch between the assumed capacity profile based on the shape of the corresponding quantile and the true capacity profile. This cost was averaged over the eight GDPs that occurred over the 3 month period of analysis. We performed 2 sets of trials. In the first, we assumed the cost of air vs ground delay was fixed at a ratio of 1.5:1 based on airline reported costs from 2015 [36]. In the second, we assumed a higher ratio of 2:1 based on rates seen in previous years. The results of our calculations are shown in Figure 10.

In both sets of calculations, the 80^{th} percentile solution provides the lowest cost. This result suggests that adopting a rate consistent with its capacity profile would yield the best balance of air and ground delay. Thus by employing a profile with a slightly higher tolerance for uncertainty the decision maker can develop an effective hedging strategy to minimize the effect of the potential costs imposed by a mismatch between the expected profile and the actual profile.

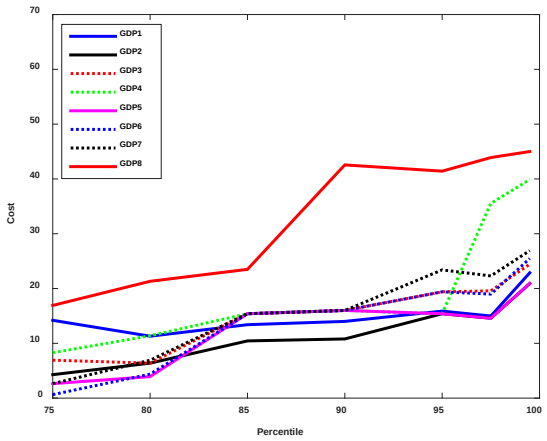

Figure 9. Variation in program costs with quantile capacities ignoring the cost of profile mismatch.

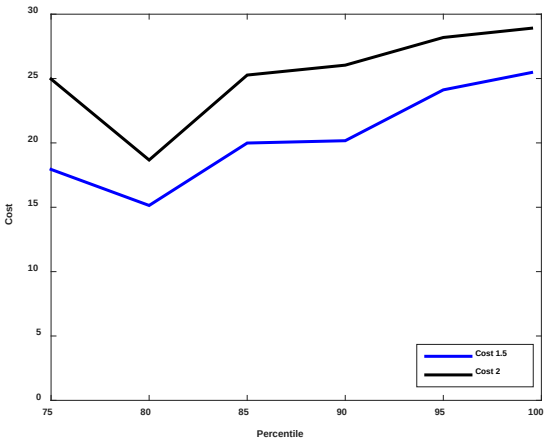

Figure 10. Average variation in program costs across all GDPs with quantile capacities including the cost of profile mismatch.

V. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a decision tree based model to predict the AAR during standard operations and GDP procedures. The method used weather forecast data, airline schedule information, the wind forecast relating to the environmental conditions on approach and information derived from previous states. The model was validated over a three month period and produced strong consistency over an extended forecast horizon. The approach yielded a set of metrics that could be used to quantify uncertainty in a number of contexts including real-time human-in-the-loop simulations, fast-time simulations, fielded decision support prototypes and integer programming models.

There are a number of applications that could leverage the concepts proposed in this paper. A decision support tool could be built to support upgrades to the Traffic Flow Management System and Time-Based Flow Management. In this context, these methods could be used to help traffic flow managers and controllers better understand the impact of assuming specific levels of risk. As an intermediate step, such a tool could be incorporated into HITL studies such as the IDM program to establish a set of constraints that provide subject matter experts with better situational awareness by applying various AAR quantile predictions to enforce capacity at the airport, thereby allowing them to focus on other aspects of the problem of interest.

The concept could also be extended to work with other weather-related decision support tools such as Traffic Flow Impact (TFI) [31]. In this context TFI could provide useful predictor variables to the model relating to changes in en route capacity. By mapping these changes to changes in the AAR the decision maker could gain a better sense of how en route flow constraints influence airport capacity.

Additionally, the stochastic integer programming model proposed in this paper could be modified and extended to account for more complicated dynamics in TMIs. Such approaches could achieve control over more resources, incorporate speed control/metering and be used to integrate arrival/departure coordination. When used in this context, the approach could provide an alternative basis for risk reduction in traffic flow management practices.

ACKNOWLEDGMENT

This material is based upon work supported by the National Aeronautics and Space Administration under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration. The authors would like to thank William Chan, Paul Lee and Nancy Smith for their guidance and support of this effort. The authors also wish to thank Jimmy Coschignano for several useful discussions.

REFERENCES

[1] http://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp?display=data&pn=1 [Accessed January 17, 2017].

[2] S. Grabbe, B. Sridhar, A. Mukherjee, and A. Morando, (2012) "Traffic Flow Management Impact on Delay and Fuel Consumption: an Atlanta Case Study," *Air Traffic Control Quarterly*, Vol, 20(3) pp. 203-224.

[3] L. Delgado, X. Prats, and B. Sridhar, (2013). Cruise speed reduction for ground delay programs: A case study for San Francisco International Airport arrivals. *Transportation Research Part C: Emerging Technologies, 36*, 83-96.

[4] L. Delgado and X. Prats (2012) "En-route speed reduction concept for absorbing air traffic flow management delays," *Journal of Aircraft*, 49,(1), 214–224.

[5] X. Prats, and M. Hansen, (2011) "Green delay programes, absorbing ATFM delay by flying at minimum fuel speed." in *9th USA/Europe Air Traffic Management Research and Development Seminar*, Berlin.

[6] L. Delgado, and X. Prats, (2013) "Effect of radii of exemption on ground delay programs with operating cost based cruise speed reduction case study: Chicago O'Hare International Airport," in *10th USA/Europe Air Traffic Management Research and Development Seminar* , Chicago.

[7] J.C. Jones and D.J. Lovell, (2014). "Methods for curbing the exemption bias in ground delay programs through speed control." in*T Transportation Research Record: Journal of the Transportation Research Board*, 2400,(1) 37-44.

[8] J. C. Jones, D. J. Lovell, and M. O. Ball, (2015). Combining Control by CTA and Dynamic Enroute Speed Adjustment to Improve Ground Delay Program Performance. In *11th USA/Europe ATM R&D Seminar (ATM 2015), Lisbon*.

[9] H. Swenson, J. E. Robinson III and J. Steve Winter, (2013). NASA's ATM Technology Demonstration-1: Moving NextGen Arrival Concepts from the Laboratory to the Operational NAS. *Journal of Air Traffic Control*, *55*(2), 27-37.

[10] T. Prevot, B. Baxley, T. Callantine, W. Johnson, L. Quon, J. Robinson, H. N. Swenson, (2012). NASA's ATM Technology Demonstration-1: Transitioning fuel efficient, high throughput arrival operations from simulation to reality. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero 2012), Brussels*..

[11] T. G. Reynolds, M. McPartland, T. Teller, and S. Troxel, (2015). Exploring Wind Information Requirements for Four Dimensional Trajectory-Based Operations. In *11th USA/Europe Air Traffic Management Research and Development Seminar*.

[12] C. Edwards, Y. Glina, J. Jones, M. McPartland, T. Reynolds, S. Troxel, (2016). Methods of Selecting Forecast Winds for Flight Management Systems to Support Four Dimensional Trajectory-Based Operations. In *16th AIAA Aviation Technology, Integration, and Operations Conference* (p. 4069).

[13] N.M. Smith, C. Brasil, P.U. Lee, N. Buckley, C. Gabriel, C.P. Mohlenbrink, F. Omar, B. Parke, C. Speridakos, and H.S. Yoo, (2016). Integrated demand management: Coordinating strategic and tactical flow scheduling operations. In *16th AIAA Aviation Technology, Integration, and Operations Conference* .

[14] H. S. Yoo, C. Mohlenbrink, C. Brasil, N. Buckley, A. Globus, N. M. Smith, and P. U. Lee, (2016). Required time of arrival as a control mechanism to mitigate uncertainty in arrival traffic demand management. In *Digital Avionics Systems Conference (DASC), IEEE/AIAA 35th* pp. 1-9.

[15] E.D. Antony, and P.U. Lee, (2016) "Analyzing Double Delays at Newark In *AIAA Aviation Technology, Integration, and Operations Conference*.

[16] M.O. Ball, R. Hoffman, A.R. Odoni, and R. Rifkin, (2003). A stochastic integer program with dual network structure and its application to the ground-holding problem. *Operations Research*, *51*(1), pp.167-171.

[17] A. Mukherjee, and M. Hansen, (2007). A dynamic stochastic model for the single airport ground holding problem. *Transportation Science*, *41*(4), pp.444-456.

[18] P. Liu, M. Hansen, and A. Mukherjee. (2008) Scenario-based air traffic flow management: From theory to practice. *Transportation Research Part B: Methodological*, 42(7):685–702.

[19] G. Buxi, M. Hansen, (2013) Generating day-of-operation probabilistic capacity scenarios from weather forecasts, *Transportation Research Part C: Emerging Technologies*, Volume 33, pp 153-166,

[20] S. L. Tien, C. Taylor, and C. Wanke. (2016). Using Ensemble Weather Forecasts for Predicting Airport Runway Configuration and Capacity. In *16th AIAA Aviation Technology, Integration, and Operations Conference* p. 3599.

[21] R. Kicinger, J.T. Chen, M. Steiner, and J. Pinto, (2016). Airport Capacity Prediction with Explicit Consideration of Weather Forecast Uncertainty. *Journal of Air Transportation*, *24*(12), pp.18-28.

[22] R. A. DeLaura, R. F. Ferris, F. M. Robasky, S. W. Troxel and N. K. Underhill, (2014) "Initial Assessment of Wind Forecasts for Airport Acceptance Rate (AAR) and Ground Delay Program (GDP) Planning," Project Report ATC-414, MIT Lincoln Laboratory,, Lexington, MA,

[23] C. Provan, L. Cook, and J. Cunningham (2011). A probabilistic airport capacity model for improved ground delay program planning. In Digital Aviation Systems Conference.

[24] J. Cunningham, L. Cook, and C. Provan, (2012) "The Utilization of Current Forecast Products in a Probabilistic Airport Capacity Model," AMS Annual Meeting, American Meteorological Soc., Boston, Paper 540.

[25] J. Cox, and M.J., Kochenderfer, (2015). Optimization Approaches to the Single Airport Ground-Holding Problem. *Journal of Guidance, Control, and Dynamics*, *38*(12), pp.2399-2406.

[26] Y. Wang, "Prediction of Weather Impacted Airport Capacity Using Ensemble Learning," (2011) *AIAA/IEEE Digital Avionics Systems Conference*, IEEE Publ.,Piscataway, NJ, pp. 2D6-1–2D6-11.

[27] Y. Wang, "Prediction of Weather Impacted Airport Capacity Using RUC-2 Forecast," (2012) *AIAA/IEEE Digital Avionics Systems Conference*, IEEE Publ.,Piscataway, NJ, pp. 1–22.

[28] M. Bloem, and N. Bambos, (2015). Ground Delay Program analytics with behavioral cloning and inverse reinforcement learning. *Journal of Aerospace Information Systems*, *12*(3), pp.299-313.

[29] J. Avery and H. Balakrishan, (2015) "Predicting Airport Runway Configurations: A Discrete-Choice Modeling Approach" in *The 11th U.S.A./Europe ATM R&D Seminar*, Lisbon..

[30] V. Ramanujam and H. Balakrishnan, (2015) "Data-Driven Modeling of the Airport Configuration Selection Process," in *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 490-499.

[31] M.P. Matthews, M.S. Veillette, J.C. Venuti, R.A. DeLaura, and J.K. Kuchar, (2016). Heterogeneous Convective Weather Forecast Translation into Airspace Permeability with Prediction Intervals. *Journal of Air Transportation*, pp.41-54.

[32] T. Hastie, R. Tibshirani,, and J. Friedman, (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed.* Springer–Verlag, NewYork, pp. 358-361.

[33] C. Taylor, T. Masek, C. Wanke, and S. Roy, (2015), June. Designing Traffic Flow Management Strategies Under Uncertainty. (2015) in *The 11th U.S.A./Europe ATM R&D Seminar*, Lisbon..

[34] A. Charnes, and W.W. Cooper, (1959). Chance-constrained programming. *Management science*, *6*(1), pp.73-79

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel,,B. Thirion, O. Grisel,M. Blondel,P. Prettenhofer, R. Weiss, V. Dubourg, „J. Vanderplas, A. Passos,D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-Learn: Machine Learning in Python," (Jan.–Dec. 2011) *Journal of Machine Learning Research*, Vol. 12, , pp. 2825–2830.

[36] http://airlines.org/data/per-minute-cost-of-delays-to-u-s-airlines/ [Accessed January 12, 2017]

AUTHOR BIOGRAPHIES

James C. Jones is Technical Staff in the Air Traffic Control Systems Group at MIT Lincoln Laboratory. He has a PhD in Civil and Environmental Engineering from the University of Maryland.

Richard DeLaura is Technical Staff in the Air Traffic Control Systems Group at MIT Lincoln Laboratory. He earned a BA in Chemistry and Physics from Harvard University when calculations were still done on slide rules.

Margo Pawlak is Assistant Staff in the Air Traffic Control Systems Group at MIT Lincoln Laboratory. She has a B.S. in mathematics from the University of New Hampshire.

Seth Troxel is a research meteorologist and aviation weather systems software engineer under sub-contract with the Air Traffic Control Systems Group at MIT Lincoln Laboratory.

Ngaire Underhill is Associate Staff in the Surveillance Systems Group at MIT Lincoln Laboratory.