Available online at http://docs.lib.purdue.edu/jate

JATE

Journal of Aviation Technology and Engineering 12:2 (2023) 53–68

Impact of Weather Factors on Airport Arrival Rates: Application of Machine Learning in Air Transportation

Robert W. Maxson[1], Dothang Truong[2], and Woojin Choi[2]

[1]NOAA Aviation Weather Center
[2]Embry-Riddle Aeronautical University

Abstract

Weather is responsible for approximately 70% of air transportation delays in the National Airspace System, and delays resulting from convective weather alone cost airlines and passengers millions of dollars each year due to delays that could be avoided. This research sought to establish relationships between environmental variables and airport efficiency estimates by data mining archived weather and airport performance data at ten geographically and climatologically different airports. Several meaningful relationships were discovered from six out of ten airports using various machine learning methods within an overarching data mining protocol, and the developed models were tested using historical data.

Keywords: data mining, airport arrival rate, flight delay, weather, machine learning

I. Introduction

The Federal Aviation Administration (FAA, 2015) outlines the major causes of delays in the National Airspace System (NAS). These sources of delay (by the percentage of total delay) are attributed to weather (69%), traffic volume (19%), equipment failures (e.g., navigation, communications, surveillance equipment; 1%), runway unavailability (6%), and other miscellaneous causes (5%). As documented by a review of NAS performance data collected over six years (from 2008 to 2013), adverse weather is the single largest cause of NAS delays, accounting for almost 70% of all delays (Sheth et al., 2015).

Delays generate enormous costs to both the flying public and airlines. In an FAA-sponsored National Center of Excellence for Aviation Operations Research (NEXTOR) report, Ball et al. (2010) estimated the total cost of flight delays in 2007 was $32.9 billion. This estimate combined the direct costs borne by airlines and passengers as well as the more subtle indirect costs that ripple through the U.S. economy resulting from flight delays. In 2014, flight delay costs were estimated to be $25 billion for U.S. air carriers by AviationFigure (2015). As weather is responsible for the majority of flight delays in the NAS (Sheth et al., 2015), a great deal of effort has been spent trying to predict and estimate the effects of weather on the NAS.

https://doi.org/10.7771/2159-6670.1285

The key components necessary to enhance airspace efficiencies are accurate weather prediction and correctly converting these anticipated environmental conditions into expected impacts on scheduled traffic flows. A key metric in translating weather conditions and other impacts affecting air traffic flows at each major terminal is the aircraft arrival rate (AAR). Per the FAA (2016), the AAR is "a dynamic parameter specifying the number of arrival aircraft that an airport, in conjunction with terminal airspace, can accept under specific conditions throughout a consecutive sixty (60) minute period" (sec. 10-7-3). FAA tactical operations managers along with terminal facility managers establish primary airport runway configurations and associated AARs on at least a yearly basis for each facility or as required (e.g., as a result of airport construction or terminal airspace redesign). The AAR establishes maximum airport capacity as a function of aircraft separation (miles-in-trail) on approach to the runway as determined by aircraft approach speeds. Based on a simple equation, average aircraft approach speeds (in knots) are divided by the desired miles-in-trail aircraft separation distance (with fractional remainders from this division conservatively rounded down to the nearest whole number).

It is fortunate that both the FAA and the National Oceanic and Atmospheric Administration (NOAA) have maintained historical databases that can be applied to better understand how these variable relationships may contribute to AAR values. Most notably, the FAA has assembled a comprehensive set of NAS performance and weather data over the last decade. For the most part, this information has been used in hindsight to assess previous day, week, month, and year airspace performance statistics to reactively improve airspace efficiency problems. While this information is useful, what is needed are predictive tools that can assess the impacts of weather-based NAS constraints before they occur.

Previous research has set the stage to create these tools. A great deal of this effort has been spent establishing the relationships between various input variables and airport arrival rates or runway configurations using evolving modeling approaches and statistical tools, e.g., support vector machines (Smith, 2008), bagging decision trees (Wang, 2011), Bayesian networks (Laskey et al., 2012), and logistic regression (Dahl et al., 2013). More recently, Hughes (2016) examined NAS performance data and NOAA National Centers for Environmental Information (NCEI, formerly the National Climate Data Center) data archives using data mining techniques to better understand how external constraints, such as weather, alter airport and terminal operational efficiencies. Explored in this study was the potential use these data have in understanding how the airspace system responds to flow constraints, and if correctly interpreted, how this knowledge can be used to predict future NAS reaction and performance by applying numerical predictive weather guidance.

Previous research has been encouraging, but the results have been difficult to apply operationally. Further, the actual impact of weather on operations is often complicated by the accuracy of forecasts issued by the National Weather Service (NWS), traffic metering inconsistencies, and scheduled airspace loadings. Therefore, the present study asked two fundamental questions:

- First, can data mining methods be used to discover significant relationships between various meteorological variable inputs and airport efficiencies recorded in the FAA and NCEI databases?
- Second, what factors can then be used as inputs to estimate AARs?

The outcomes resulting from the first question fed directly into the second question. Any consistencies in modeling results were noted across the ten airports selected.

As a result, this research sought to translate predictive weather guidance into NAS performance impact. Foundational to this study was the use of data mining techniques to detect patterns in the behavior of the airspace system through its airport terminals as they react to changing weather conditions and traffic demands. With an airport's response to various weather conditions better understood, arrival rates could then potentially be estimated with some degree of skill (perhaps out to several days) using predictive numerical weather guidance. The ability of national airspace managers to set realistic airport arrival rates during the early planning phases of NAS operations could enhance airport efficiencies, lower operational costs, and improve flight safety.

II. Methodology

2.1. Data Collection

In this study, two datasets were collected from two different sources and consolidated into one final dataset. The Aviation System Performance Metrics (ASPM) database is the FAA Operations and Performance data that consists of airport performance statistics and limited weather variables archived at 15-minute and hourly intervals. In order to add additional weather parameters to the analysis, meteorological hourly station data from NOAA NCEI were collected for the same airports. This dataset was merged with the ASPM data to increase the number of environmental variables (e.g., precipitation type and amount). Additionally, the NWS provided Localized Aviation MOS (Model Output Statistics) Program (LAMP) as a third source of data that supplied predictive numerical weather guidance.

Ten specific airports were chosen for the data mining: (a) Hartsfield-Jackson Atlanta International Airport (ATL), (b) Los Angeles International Airport (LAX), (c) O'Hare International Airport (ORD), (d) Dallas/Fort Worth International Airport (DFW), (e) John F. Kennedy International Airport (JFK), (f) Denver International Airport (DEN), (g) San Francisco International Airport (SFO), (h) Charlotte-Douglas International Airport (CLT), (i) LaGuardia Airport (LGA), and (j) Newark Liberty International Airport (EWR).

For each of the ten airports, a two-year sample of 15-minute-interval ASPM performance metrics and weather observations were extracted from the FAA database in 2014 and 2015. This sampling resulted in 70,080 observations (rows of data) with 83 variables within each observation (or row) for each of the ten airports selected. This sample size is large enough for the data mining purpose in this study.

2.2. Demographics

All ten airports selected for this study are part of the FAA's "Core 30" and are located in major metropolitan areas that see exceptionally high passenger and/or air cargo demands. Some of the airports are capacity constrained by physical airport layout or by geographical location and associated weather and climate conditions. A summary of the airport demographics is provided in Table 1.

2.3. Variables

Tables 2 and 3 present variables and descriptions from two different datasets. The number of available weather variables increases from the 15-minute, to the hourly, and then hourly merged datasets. The hourly merged dataset encompasses all the weather variables contained in the 15-minute and hourly data and adds weather variables beyond those two datasets.

The 15-minute (quarterly hour) data contain a simple set of weather data. These are CEILING (measured in hundreds of feet), TEMP (or temperature, measured in degrees Fahrenheit), VISIBLE (or visibility, measured in statue miles), WIND_ANGLE (or wind angle, measured in degrees), and WND_SPED (or wind speed, measured in knots). A categorical variable, MC (meteorological conditions) completes the weather variables contained in the 15-minute dataset and reports if the terminal weather conditions were IFR (I) or VFR (V).

The hourly data introduces three new variables beyond those contained in the 15-minute datasets. These are NEARBYTS which counts the number of thunderstorms detected by nearby ASOS stations within 50 miles of the terminal, SEVERITY which assesses local weather impacts on airport operations, and WTHR_TYPE which describes weather conditions impacting traffic flow.

Finally, the hourly merged dataset joins the hourly FAA ASPM data with the near-hourly NCEI meteorological station data, adding both redundant and new weather variables into the modeling analyses. The two datasets are not perfectly time matched, and the NCEI data times needed to be advanced or retarded in time to synchronize the variables to the nearest hour, as well as to adjust the GMT times to local time to match the FAA ASPM data formats. As an example, CEILING is found in both the ASPM and NCEI (as CLG, or ceiling) datasets, but unlimited ceilings are reported as the numeric character 999 in the ASPM data, while unlimited ceilings in the NCEI data are reported as 722, making the two datasets appear to be more different than they actually are.

2.4. Data Analysis Procedure

As previously stated, this study used all available 2014 and 2015 ASPM records to train and validate each model created and 2016 ASPM records to then score these models. Decision tree (DT), linear regression (REG), and neural network (NN) models were created using combined 2014 and 2015 ASPM data sampled. Cases between midnight and 0600 were removed (per Dhal et al., 2013, and others) to eliminate periods of light airport traffic demands in the model analyses. These are the data used for reporting the results of this study.

The combined 2014 and 2015 datasets were partitioned 60% and 40% respectively to train and validate the performance of all of the models. The 2016 data were

Table 1
Airport demographics summary.

Airport	Number of runways	Arrival/departure configs	Max. AAR	Min. AAR	Passenger enplanements (millions)	Cargo moved (metric tons)
ATL	10	17	132	18	50.5	1,200,000
CLT	8	13	92	35	21.5	211,944
DEN	12	19	152	32	28.2	646,566
DFW	14	7	120	30	31.3	1,800,000
EWR	6	9	48	16	19.9	1,300,000
JFK	8	12	60	26	29.2	1,500,000
LAX	8	10	80	12	39.6	3,100,000
LGA	4	11	40	24	14.7	7,586
ORD	16	11	114	32	37.5	4,200,000
SFO	8	19	54	25	25.7	590,110

Note. 2016 data provided by FAA (2017a, 2017b).

Table 2
FAA ASPM variable definitions.

Name	Level	Definition
ARR_RATE	Interval	Airport-supplied arrival rate for capacity
CEILING	Interval	Ceiling measure in hundreds of feet
MC	Nominal	Meteorological conditions (IFR or VFR)
NEARBYTS	Interval	Number of nearby thunderstorms within 50 miles per ASOS
N_CEILING	Interval	Nearby ceilings within 50 miles per ASOS
SEVERITY	Interval	Assessed weather impact by category
TEMP	Nominal	Temperature (°F)
VISIBLE	Interval	Visibility in nautical miles
WIND	Interval	Wind impact categories (airport specific)
WND_ANGL	Nominal	Wind direction (degrees from magnetic north)
WND_SPED	Nominal	Wind speed (knots)
WTHR_TYPE	Nominal	Predominant weather categorized by type

Note. Bolded variables are contained in the hourly ASPM, but not in the 15-minute dataset.

Table 3
NCEI meteorological station data variable definitions.

Name	Level	Definition
ALT	Nominal	Altimeter setting
A.W.	Nominal	Auto-observed present weather
CLG	Nominal	Ceiling (hundreds of feet)
DEWP	Nominal	Dew point (°F)
DIR	Nominal	Wind direction in 36 compass degrees 990 is variable
GUS	Nominal	Wind gust (MPH)
H	Nominal	High cloud type
L	Nominal	Low cloud type
M	Nominal	Middle cloud type
MAX	Nominal	Maximum temperature (°F)
MIN	Nominal	Minimum temperature (°F)
M.W.	Nominal	Manually observed present weather
PCP01	Nominal	One-hour liquid precipitation (inches to nearest 100th)
PCP06	Nominal	Six-hour liquid precipitation (inches to nearest 100th)
PCP24	Nominal	24-hour liquid precipitation (inches to nearest 100th)
PCPXX	Nominal	3- or 24-hour liquid precipitation (inches to nearest 100th)
S.D.	Nominal	Snow depth (inches)
SKC	Nominal	Sky cover (by octal)
SLP	Nominal	Sea level pressure (millibars to nearest tenth)
SPD	Nominal	Wind speed (MPH)
STP	Nominal	Station pressure (millibars to nearest tenth)
TEMP	Nominal	Temperature (°F)
VSB	Nominal	Visibility (statute miles to nearest tenth)
W	Nominal	Past weather indicator

then used to score the models by using the Score node within the SAS® EM™. The 2016 scored data results yielded predicted arrival rates that were then compared with the actual arrival rates observed that year. Finally, as a demonstration, a "present-day" case was run using NWS 24-hour predictive weather guidance to predict future AARs, and this estimate was then compared with the actual arrival rate observed in hindsight. A summary of the data analysis is shown in Figure 1.

Three different analyses were conducted for each airport to produce more meaningful results, as follows.

The first analysis used the entire two-year (2014–2015), 15-minute interval ASPM data. The second analysis used the two-year FAA ASPM dataset extracted at hourly intervals, allowing comparison of the results at each airport using different sampling rates with several additional meteorological variables. In the third analysis, merging FAA ASPM and hourly NOAA NCEI surface meteorological data added even more weather information variables (beyond those found in the ASPM data) into the model. In all three analyses, the AAR is the target variable, and date, hour, and weather variables are the predictors.

The performance of each model (decision tree, linear regression, and neural network) was assessed for each airport. The goal was to create a predictive system where estimated input variables could then forecast airport efficiency. The 2014–2015 15-minute and hourly ASPM datasets, as well as the hourly merged ASPM and surface

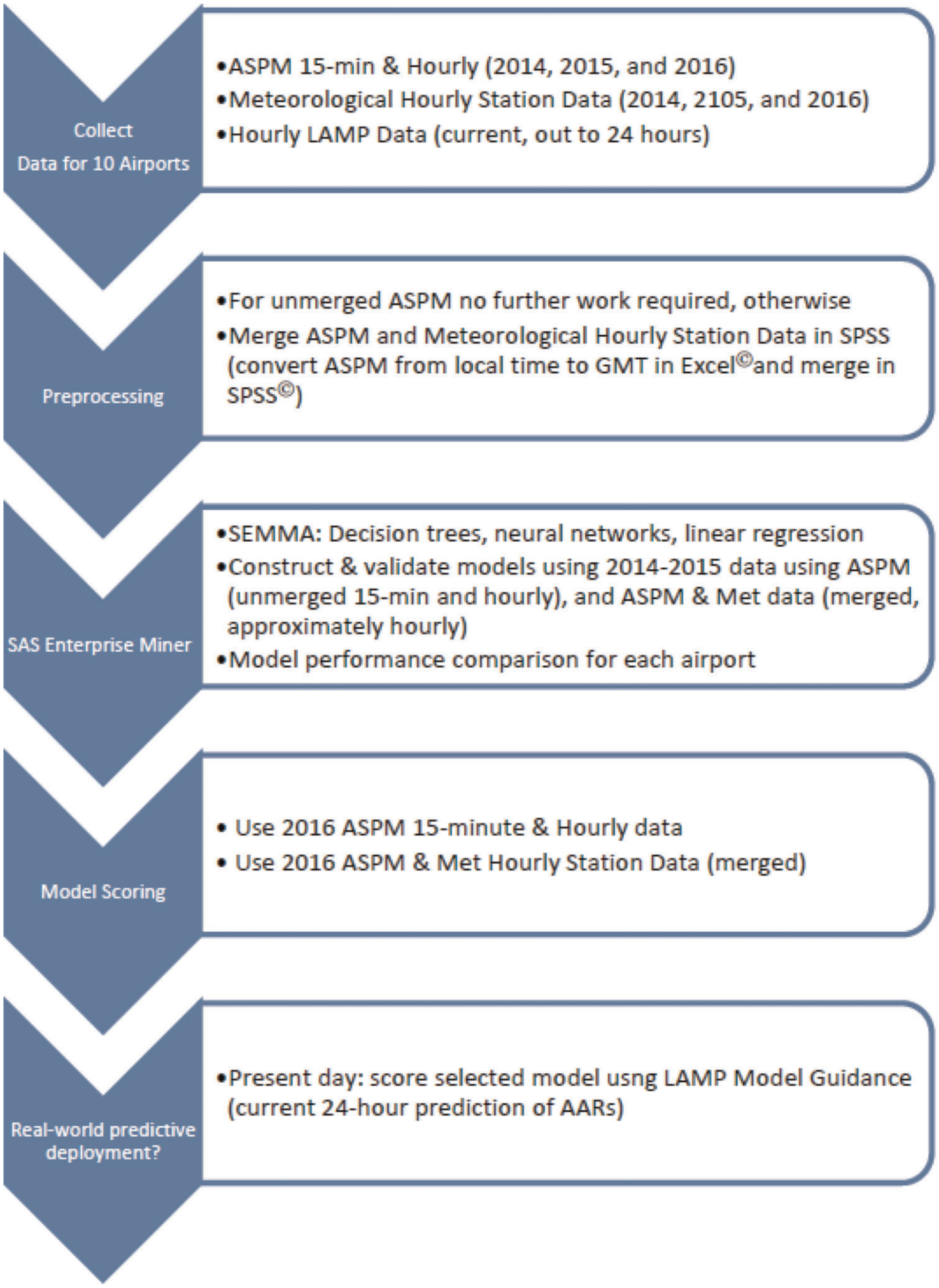
Collect
Data for 10 Airports
•ASPM 15-min & Hourly (2014, 2015, and 2016)
•Meteorological Hourly Station Data (2014, 2105, and 2016)
•Hourly LAMP Data (current, out to 24 hours)
Preprocessing
•For unmerged ASPM no further work required, otherwise
•Merge ASPM and Meteorological Hourly Station Data in SPSS (convert ASPM from local time to GMT in Excel©and merge in SPSS©)
SAS Enterprise Miner
•SEMMA: Decision trees, neural networks, linear regression
•Construct & validate models using 2014-2015 data using ASPM (unmerged 15-min and hourly), and ASPM & Met data (merged, approximately hourly)
•Model performance comparison for each airport
Model Scoring
• Use 2016 ASPM 15-minute & Hourly data
• Use 2016 ASPM & Met Hourly Station Data (merged)
Real-world predictive deployment?
•Present day: score selected model usng LAMP Model Guidance (current 24-hour prediction of AARs)

Figure 1. Data analysis schematic.

meteorological weather datasets, were used to create and validate the models, and these models were then scored using actual 2016 observed weather and airport AARs.

Three datasets were assembled for each of the 10 selected airports: (a) a 15-minute ASPM dataset with a limited number of meteorological variables, (b) an hourly dataset that essentially takes the information contained from the 15-minute ASPM dataset at the top of each hour and introduces several more meteorological variables not contained in the 15-minute data, and (c) a merged dataset containing the hourly ASPM data and NCEI meteorological station data that introduce even more weather variables (beyond the hourly ASPM) into the model decision-making process. As a result, 90 models were trained, validated, and scored (ten selected airports using three different datasets using three models per dataset).

III. Results

3.1. Model Comparison

To directly compare the models, the square roots of the validated model average square errors, or ASE, were compared. The ASE is the sum of all squared errors (SSE) divided by the number of cases (*N*); hence, ASE = SSE/*N*. ASE allows comparison of model performance in both linear and nonlinear models. In the cases of the 15-minute models, to account for a full hourly error, the square root of the ASE was multiplied by four. Using this method, the lowest value found amongst the nine validated models constructed for each airport determined the best model. These results are presented in Table 4, and the bolded text indicates the best single model selected for scoring using the fresh 2016 data for each airport. The 2016 scored results are presented in the Scoring section.

Of the ten best airport models selected, four used the hourly data, four used the hourly merged data, and two used the 15-minute data. Seven models were decision tree models, while the remaining three were neural network models. While the linear regression models performed comparatively well, none were selected for scoring using this process. In general, all the validated model square root ASEs were very close in value for each airport studied.

3.2. Variable Importance

Variable importance was identified in the splitting decisions made by the decision tree models. In examining the 15-minute variable importance, there is little similarity of variable importance between airports, but it was found that ceilings and temperatures are of slightly more importance than visibilities and wind speeds (Table 5). Of more interest is how the variables are added to the decision processes. Several changes or replacements of variable importance between the 15-minute and hourly datasets (as shown in Table 6) are noteworthy within each airport.

The first is that the weather impact variable SEV, or severity, has displaced other variables found in the 15-minute data as a top-five variable in five out of the ten airports (it actually occurs as a top-eight or better variable in all ten airports). NBTS, or nearby thunderstorms, also moves into the top five most important variables for ATL, CLT, DEN, and DFW and becomes the sixth most

Table 4
Comparison of square root of validated 2014–2015 model ASE.

Airport	Model type	Square root of 15-minute data ASE	Square root of hourly data ASE	Square root of merged data ASE
ATL	DT	8.776	8.051	**7.937**
	REG	9.208	8.365	8.287
	NN	10.431	18.460	8.205
CLT	DT	9.429	13.009	9.608
	REG	9.611	12.931	9.807
	NN	**9.426**	12.997	9.661
DEN	DT	17.919	17.497	17.447
	REG	18.082	17.449	17.544
	NN	17.856	17.398	**17.287**
DFW	DT	13.744	13.566	**13.521**
	REG	13.975	13.635	14.249
	NN	13.836	13.566	13.918
EWR	DT	**3.810**	3.868	3.823
	REG	3.954	3.862	9.210
	NN	5.296	3.882	3.813
JFK	DT	8.977	9.070	**8.792**
	REG	9.053	9.064	9.057
	NN	9.979	10.347	9.095
LAX	DT	14.298	**8.101**	8.125
	REG	14.761	8.131	8.550
	NN	14.766	8.755	8.162
LGA	DT	4.870	**4.497**	4.694
	REG	4.956	4.680	4.843
	NN	5.295	6.108	4.717
ORD	DT	12.276	11.772	11.896
	REG	12.484	11.917	13.384
	NN	12.592	**11.762**	12.900
SFO	DT	6.537	**5.904**	5.904
	REG	6.761	6.166	6.214
	NN	6.771	5.920	5.929

Note. Decision tree (DT), regression (REG), and neural network (NN). Bold/underlined indicates the best model selected overall by airport based on the square root of ASE. The square root of 15-minute data ASE multiplied by four to account for a full hour of potential error.

Table 5
15-minute data decision tree variable importance.

Airport	1st Var	2nd Var	3rd Var	4th Var	5th Var
ATL	MC	TEMP	CEIL	VIS	ALH
CLT	ALH	MC	CEIL	VIS	TEMP
DEN	CEIL	TEMP	VIS	ALH	WND_S
DFW	TEMP	MC	ALH	VIS	WND_A
EWR	VIS	TEMP	ALH	WND_S	CEIL
JFK	MC	ALH	TEMP	WND_A	CEIL
LAX	ALH	CEIL	WND_A	TEMP	VIS
LGA	WND_A	TEMP	CEIL	VIS	WND_S
ORD	WND_A	TEMP	CEIL	VIS	WND_S
SFO	ALH	CEIL	WND_A	VIS	WND_S

Note. ALH is adjusted local hour, CEIL is ceiling, MC is met condition, TEMP is temperature, VIS is visibility, WND_A is wind angle, and WND_S is wind speed. Importance compares within each airport for the three datasets.

Table 6
Hourly data decision tree variable importance.

Airport	1st Var	2nd Var	3rd VAR	4th VAR	5th VAR
ATL	MC	TEMP	VIS	NBTS	CEIL
CLT	MC	CEIL	SEV	WND_A	NBTS
DEN	CEIL	TEMP	VIS	NBTS	WIND
DFW	MC	TEMP	ALH	NBTS	SEV
EWR	CEIL	TEMP	ALH	WIND	VIS
JFK	MC	CEIL	WND_A	VIS	TEMP
LAX	ALH	CEIL	WIND	VIS	SEV
LGA	WND_A	SEV	CEIL	TEMP	WX_TYP
ORD	WND_A	SEV	CEIL	TEMP	WX_TYP
SFO	ALH	CEIL	WND_A	SEV	VIS

Note. ALH is adjusted local hour, CEIL is ceiling, MC is met condition, NBTS is nearby thunderstorms, SEV is severity, TEMP is temperature, VIS is visibility, WND_A is wind angle, WIND is wind speed, WND_S is wind speed, and WX_TYP is weather type.

important variable (not shown) for LGA and ORD. Curiously, out of nine total weather variables examined in the hourly data, NBTS was not selected at any level of importance for EWR, JFK, or LGA. Nor was NBTS of interest for LAX or SFO, but this is understandable given that the west coast maritime climate patterns prevalent at these airports inhibit the growth of thunderstorms. WX_TYP, or weather type, a descriptor of various types of weather, creeps into the top five as the fifth most important variable for LGA and ORD. It also is used by DEN (7th), CLT (8th), DFW (8th), JFK (8th), and SFO (10th).

Finally, WIND has replaced WND_S (or wind speed) at EWR (4th) and LAX (3rd) as the top five variables of importance. Recall that the WIND variable appears to have been created to account for wind speed and direction as a combined impact variable, but for each airport studied, it simply mimics the wind speed variable (shown in the descriptive statistics as WND_SPED). Therefore, these two variables are considered to be indistinguishable in this study.

Examining the hourly merged data (Table 7), the combination of the FAA ASPM data with the NCEI meteorological data is evident as several meteorological data not found in the ASPM 15-minute or hourly data have become variables that fall within the top five of importance. Most notable among these is DEWP, or dew point, which is listed for ATL, DEN, and DFW. In addition, added as new variables are A.W., or auto-observed present weather, and GUS, or gusts. Several of the NCEI meteorological variables have replaced essentially the same meteorological variables already found in the FAA ASPM data, and these are TEMP_1 (that mimics TEMP) and VSB (that mimics VIS). However, it should be noted these sister variables may not contain exactly the same values due to the rounding of the NCEI data to the nearest hour used in merging these data. That is, the merger between the ASPM and NCEI datasets may not be precisely time-synchronized. In any case, if there are differences, the values for these variables are very close and follow the same trends within the two individual dataset time series. Other variables can be found in the 14 variables contained in the hourly merged data. These are ALT (altimeter), CLG (mimics CEIL, or ceiling), DIR (mimics WND_A, or wind angle), PCP01 (amount of last hourly precipitation as liquid water in inches), PCP06 (amount of last six-hour precipitation as liquid water in inches), and SKC, or sky conditions.

3.3. Model Reliability and Validity

Model reliability begins with the data collected to build the models, followed by the construction of the models themselves and the quality of data subsequently collected to evaluate the models. In general, the ASPM data were found to be of very high quality with nearly no missing values. Problems were discovered with outliers; for example, the 2016 15-minute DEN data reported impossible AARs of 800 for 47 cases (out of 26,352 cases scored when the nighttime cases were removed) that are clearly not possible with a published AAR maximum of 152 per FAA

Table 7
Hourly merged data decision tree variable importance.

Airport	1st Var	2nd Var	3rd VAR	4th VAR	5th VAR
ATL	MC	DEWP	VIS	NBTS	CEIL
CLT	ALH	MC	CEIL	SEV	NBTS
DEN	CEIL	DEWP	ALH	VSB	AW
DFW	MC	DEWP	ALH	TEMP_1	AW
EWR	CEIL	TEMP_1	ALH	SPD	VSB
JFK	MC	ALH	CEIL	WND_A	TEMP_1
LAX	ALH	CEIL	VSB	WIND	VIS
LGA	DIR	AW	CEIL	WIND	WND_A
ORD	DIR	AW	CEIL	WIND	WND_A
SFO	ALH	CEIL	SEV	GUS	VIS

Note. ALH is adjusted local hour, A.W. is auto-observed weather, CEIL is ceiling, DEWP is dew point, DIR is wind direction, GUS is gust, MC is met condition, NBTS is nearby thunderstorms, SEV is severity, TEMP_1 is temperature, VIS and VSB are visibility, WND_A is wind angle, WIND is wind speed, and WND_S is wind speed.

Operational Information System. Therefore, these 47 cases were list-wise removed, and the model was scored again.

The NCEI meteorological station data also undergo a great deal of scrutiny but may suffer from missing or misleading variable values due to ASOS sensor errors or station data recording errors. However, the additional NCEI information was simply appended to the hourly FAA ASPM data to expand the potential reach of the weather variables contained in the NCEI database to those already included in the FAA ASPM hourly datasets in the model analyses. In addition to adding fresh weather variables to each analysis, these data mergers for each airport created redundant variables, e.g., Wind_ANGL (wind angle, FAA ASPM data) and DIR (wind direction, NCEI meteorological station data) found in both datasets. In building the hourly merged data models, all the weather variables from both the ASPM and NCEI were used. The time-match merging of the FAA ASPM and NCEI data offered the opportunity to compare common variables contained in both datasets, such as ceiling, wind speed, and visibility. For the most part, even if the rounded hourly time-merger of the ASPM and NCEI data was not perfect, across the 10 airports considered (except for CLT, where the hourly merged validated model results were greatly improved over the hourly data models), the output results were extremely close when comparing the hourly and hourly merged model validation ASE results. This indicates the added meteorological variables contained in the NCEI data did not degrade the results found in the less meteorologically comprehensive models constructed with the hourly ASPM or 15-minute data.

As Kulkarni et al. (2013) noted, three different modeling methods yielding such similar outcomes lends credence to the reliability of this data mining approach. Three distinctly different models, namely decision trees, neural networks, and linear regression, were tested with strikingly similar validated ASEs regardless of the model used. These results confirm the observations of Kulkarni et al.

Per Tufféry (2011), model validity should be established by using an "out of date" testing dataset. This was accomplished by using fresh 2016 data to score the selected best model for each airport. It is also of note that the 2016 datasets used to score the models were of roughly the same size as the 2014–2015 training and validation sets that pulled from 60% and 40% of the two-year population, respectively.

3.4. Scoring

All of the models were scored using a full year's worth of 2016 ASPM or combined ASPM and NCEI merged data. SAS® EM™ provides a scoring node that was used to predict the 2016 AARs using the weather inputs from the three datasets. The models were scored using the best model for each airport with the 2016 data, with 2400 to 0600 cases removed. For brevity, the "best" (ATL) and "worst" (JFK) airports are presented here as examples. The results for ATL are depicted in Table 8 and for JFK in Table 9. In each table, within the footnote, the model chosen to score is labeled (DT, NN, REG) and reflects the results noted in Table 4 between the actual AAR observed in 2016 and the predicted AAR estimated by the model. Histograms are used to present the model fit graphically. The histograms indicate the difference between the actual airport AAR observed and the values estimated by SAS® EM™, as well as error residuals, separately, and are presented as Figures 2 and 3 (ATL) and Figures 4 and 5 (JFK).

For the histograms, a perfect score would place the actual versus predicted AAR differences at zero for all cases considered. Thus, the larger the actual and model estimate AAR differences are, the larger the spread by cases become and tend to flatten the histograms as shown for each airport. In addition, large horizontal displacements from the origin on the *X*-axis indicated the likely presence of outliers in the scored data inputs. Subsequent model and data input reevaluations were warranted if the difference spread tended to exceed the maximum AAR, as presented in Table 1.

3.5. Numerical Weather Model Prediction of AAR Demonstration

With the models and modeling strategies established, it was desirable to test the efficacy of using basic weather variables to estimate the AARs *a priori*. For this effort, NWS numerical weather data estimates were reformatted into the FAA 15-minute ASPM data formats so that the models created could be used in a true predictive sense to test if a 24-hour forecast of weather parameters from NWS can yield useful estimates of FAA airport arrival rates.

As an example, NWS LAMP output data were obtained and reformatted to be accepted into the SAS® EM™ frameworks established within the 15-minute modeling

Table 8
ATL observed versus predicted AAR difference in scored 2016 data.

	ATL DT 15 min	**ATL DT hourly**	**ATL DT merged**
Mean	0.981	3.692	3.067
Standard error	0.012	0.083	0.083
Median	1.033	4.126	3.178
Mode	2.663	4.126	3.178
Standard deviation	1.942	6.702	6.717
Sample variance	3.773	44.916	45.121
Kurtosis	11.569	9.072	8.933
Skewness	−2.116	−1.592	−1.463
Range	35.802	131.191	131.191
Minimum	−26.585	−84.980	−84.980
Maximum	9.217	46.211	46.211
Sum	25842.290	24319.620	20189.520
Count	26352	6588	6584

Note. Hourly merged DT model selected from the nine-model suite for scoring with 2016 data.

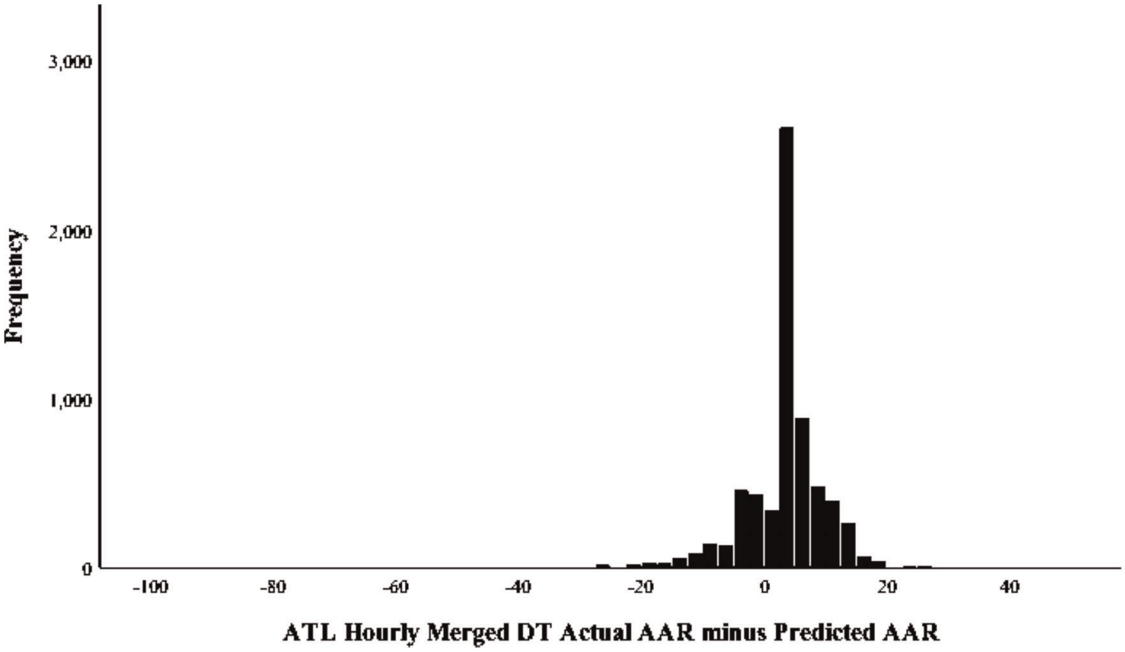

Figure 2. Difference between ATL actual and predicted AAR in scored 2016 data.

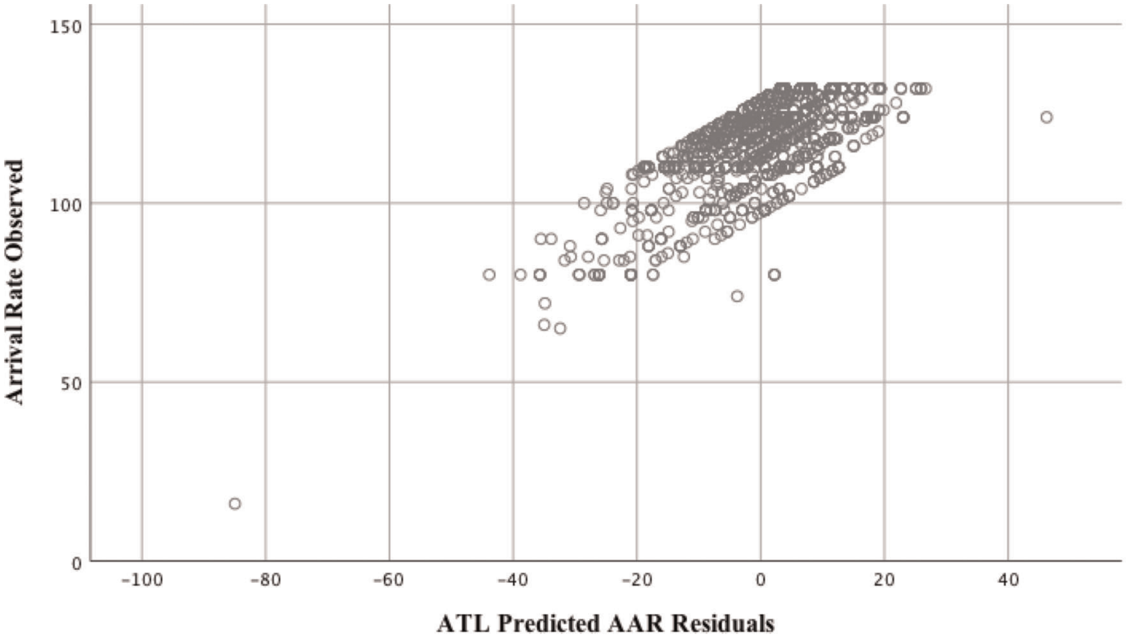

Figure 3. Observed ATL arrival rates versus predicted AAR residuals.

format. The 15-minute ASPM data contain the fewest number of weather variables of the three variable sets used in this study but generally had favorable ASEs in the train and validation model output results and also did reasonably well when scored. As a result, these data are ideal for a simple scoring test in assessing airport AARs using NWS LAMP weather guidance. Variables that needed to be reformatted or created from the LAMP data into ASPM format include WIND_ANGLE, WIND_SPED, CEILING, VISIBILTY, ALH, GMT_YMDHM, and MC. With the LAMP model output limited to 24 hours, a dataset was collected on November 15, 2017, with a valid forecast period beginning at 1700 GMT on November 16 and running through 1700 GMT on November 17. These data were then reformatted to represent ASPM variables, scored within the SAS® EM™, and were subsequently compared to the actual AARs observed and recorded in the FAA ASPM database on November 18. Compared to the datasets used to train and validate the models, the NWS 24-hour datasets are very small. Nonetheless, the initial test results were encouraging. Actual airport arrival rates minus the predicted airport arrival rates for a 15-minute decision

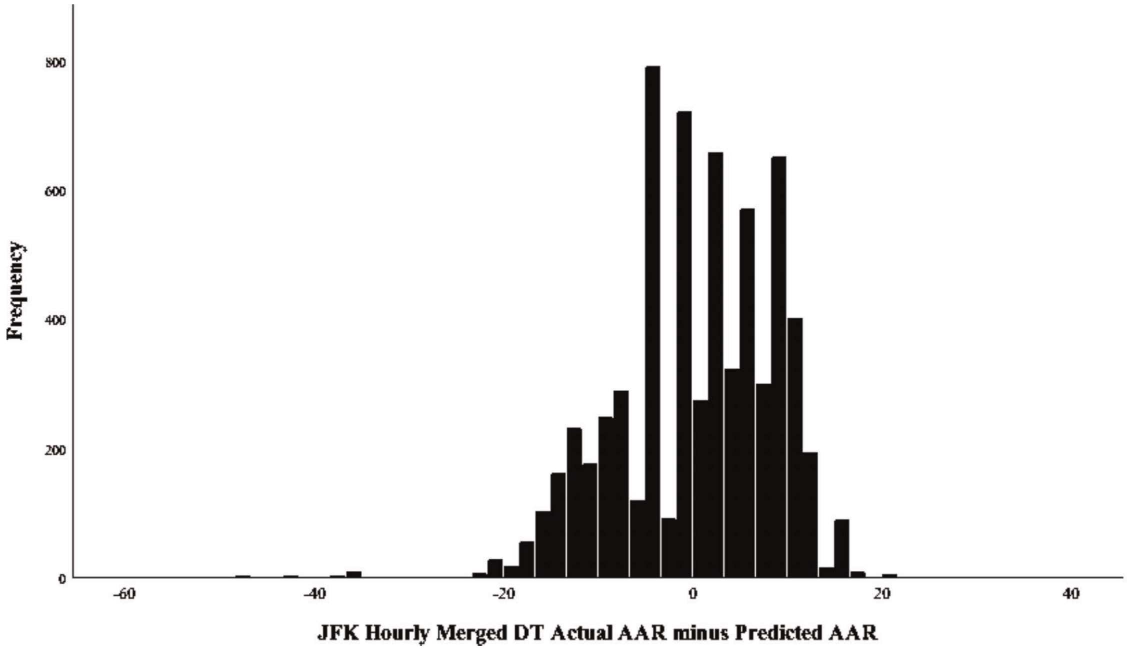

Figure 4. Difference between JFK actual and predicted AAR in scored 2016 data.

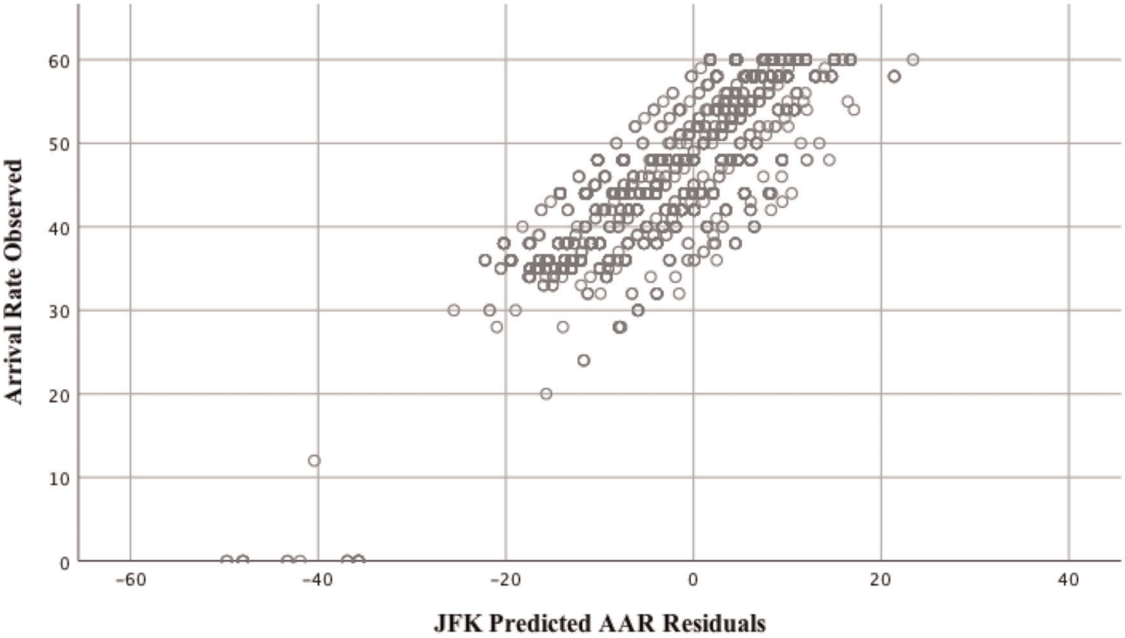

Figure 5. Observed JFK arrival rates versus predicted AAR residuals.

tree model at LGA are presented in Table 10. A histogram showing the differences between the actual and predicted AARs (by frequency of cases) is presented in Figure 6.

The date chosen for the collection of these data was happenstance due to the timing of this research. November 17, 2017, was a blustery day at LGA with winds gusting to 36 mph mid-morning, VFR conditions, and no precipitation throughout the 24-hour period. The relevance of this demonstration is that NWS predictive weather model guidance can be potentially applied *a priori* to estimate airport arrival rates in a 24-hour cycle. A positive observed versus predicted AAR difference represents an underestimated arrival capacity at LGA, while the opposite (negative) difference marks an overestimation of airport capacity based on weather input variables and local time.

IV. Discussion

The intent of this research was to objectively examine the usefulness of applying weather information predictively

Table 9
JFK observed versus predicted AAR difference in scored 2016 data.

	JFK DT 15 min	JFK REG hourly	JFK DT merged
Mean	0.123	0.325	0.394
Standard error	0.014	0.107	0.104
Median	0.281	1.003	0.726
Mode	1.835	9.351	1.780
Standard deviation	2.216	8.679	8.397
Sample variance	4.909	75.326	70.502
Kurtosis	1.473	1.064	1.207
Skewness	−0.735	−0.748	−0.688
Range	18.510	69.865	73.138
Minimum	−13.266	−50.576	−49.738
Maximum	5.244	19.289	23.400
Sum	3250.260	2137.990	2591.280
Count	26352	6588	6584

Note. Hourly merged DT model selected from the nine-model suite for scoring with 2016 data.

Table 10
LGA observed versus predicted AAR in scored 20171116 data.

Statistic	LGA LAMP 24 hour
Mean	0.856
Standard error	0.096
Median	0.583
Mode	0.583
Standard deviation	0.789
Sample variance	0.623
Kurtosis	0.282
Skewness	0.092
Range	3.623
Minimum	−1.417
Maximum	2.206
Sum	57.340
Count	67

to estimate AARs. To achieve this goal, a closer examination of model performance at each airport was required. Of the 90 models created for 10 different airports, the best for each airport (using the square root of ASE from the 2014–2015 validated data) were directly compared by scoring these models using fresh 2016 data. As an approximate estimate of acceptable model error, an arbitrary threshold was set at 10% of an airport’s maximum AAR. Recalling the maximum arrival rates for each airport are contained in Table 1, this result implies that the maximum acceptable error (absolute value of observed minus predicted AAR) for an AAR prediction at DEN would be 15.2 (or 15), while at LGA the threshold for acceptable model performance would be an AAR predictive error of four. Additionally, simple line plots of the observed AAR minus predicted AAR versus actual AAR are presented for each airport, so a visual depiction and interpretation of model performance can be more easily understood. A model with little difference between observed versus predicted AARs would have a near-zero error for all cases. These results are summarized in Table 11 and are based on the percentage of cases that fall within the arbitrarily set acceptable threshold of plus or minus 10% of the airport’s maximum AAR. Further, the models were ranked overall from 1 to 10 (best to worst) by comparing the 2014–2015 model validation results. Additionally, the type of model considered the best performer for each airport and the dataset used to create the model are included in the table.

What follows now is a brief discussion of the scoring results for the single model selected for ATL and JFK using the 2016 datasets. Based on Table 11, ATL can be considered to be a good modeling result, while the JFK modeling effort is deemed to be poor.

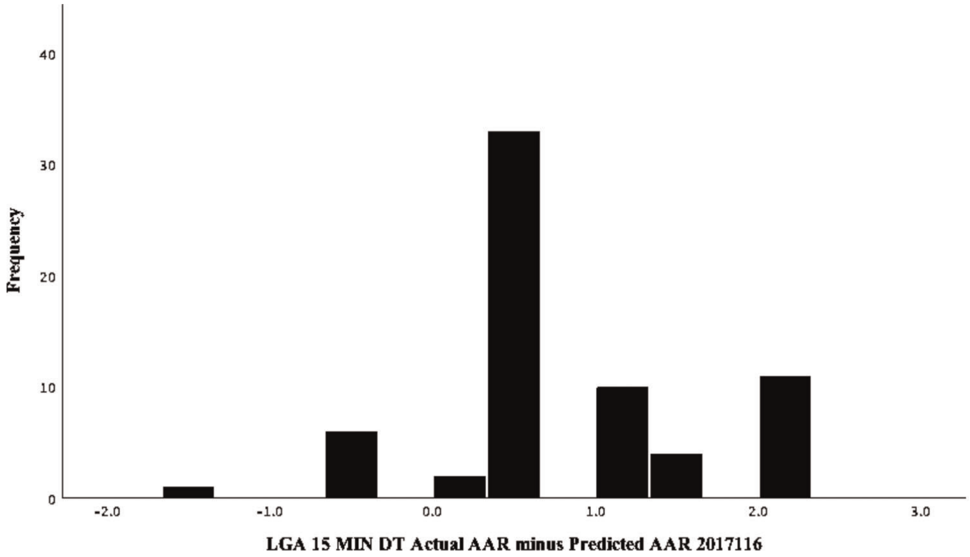

Figure 6. LGA difference in observed versus predicted AAR 20171116 data.

Table 11
Model performance summary and rankings.

Airport	Percentage of cases within 10% of maximum AAR[a]	Validated model ranking based on squared root of ASE[b]	Model type	Dataset used
LAX	91.7	5	Decision tree	Hourly merged
ATL	91.6	4	Decision tree	Hourly merged
EWR	87.2	1	Decision tree	15-minute
LGA	68.3	2	Decision tree	Hourly
SFO	68.0	3	Decision tree	Hourly
DFW	65.3	9	Decision tree	Hourly merged
DEN	60.6	10	Neural network	Hourly merged
CLT	59.3	7	Neural network	15-minute
ORD	45.1	8	Neural network	Hourly
JFK	44.4	6	Decision tree	Hourly merged

[a]Based on scoring results using 2016 data. [b]Based on model validation using withheld 2014–2015 data.

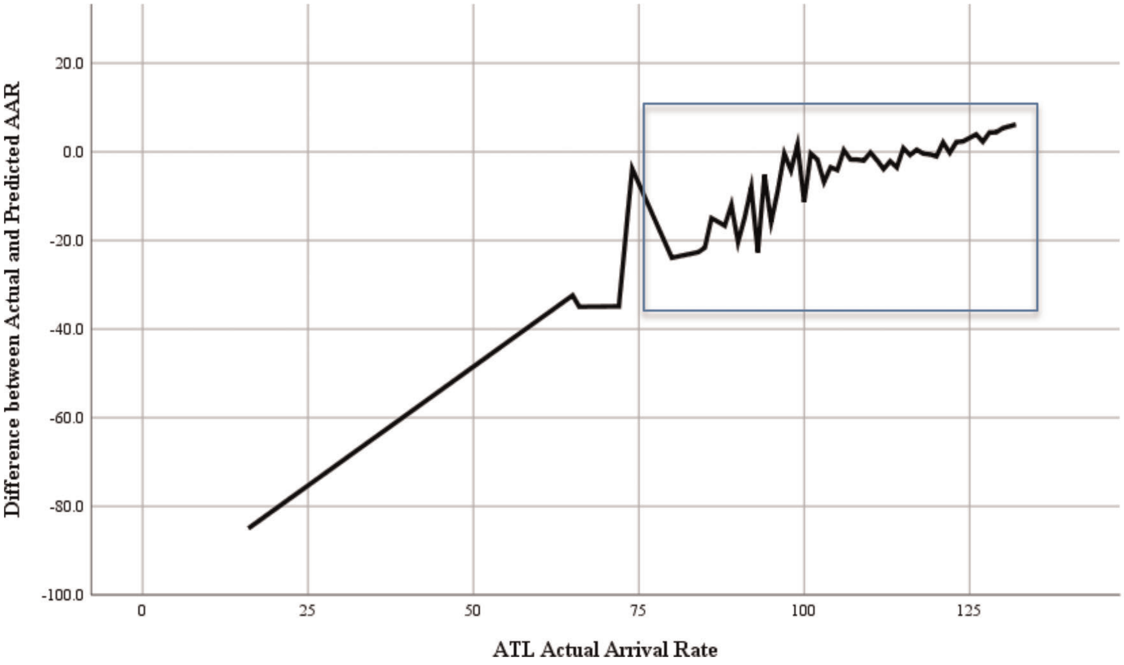

Figure 7. ATL actual and predicted difference versus actual AAR.

4.1. Hartsfield-Jackson Atlanta International Airport

ATL has a maximum arrival rate of 132, so an acceptable error based on 10% of the maximum AAR is an absolute value of the observed minus predicted AAR of 13. These results were derived from the decision tree model using the merged hourly ASPM and meteorological station data. This model was selected as the best model based on model validation using data withheld from the 2014–2015 data. Figure 7 shows the line graph of the difference between the actual and predicted AAR plotted against the actual AAR. The highlighted area of the graph is of interest and depicts the residuals (difference between the actual and predicted AAR) when the AAR is roughly above 80. Examining the variable importance for Atlanta using this dataset, the top five variables ranked by order of importance in supporting the model decision making were: (1) meteorological conditions (IMC versus VMC), (2) dew point, (3) visibility, (4) nearby thunderstorms, and (5) ceiling.

At first glance, the model performed poorly when actual AARs were low, likely due to the presence of adverse weather or when other capacity-limiting factors were encountered, such as a closed runway. This can be seen as an over-forecast of airport capacity where the differences between the actual and predicted AARs are negative, and the over-forecasts are observed at the lower left-hand section of the figure. However, further scrutiny of the data revealed that of the 6,584 cases scored using the 2016 hourly merged data, there were only five cases where the actual AAR fell below 80. Recall an AAR represents the number of aircraft an airport can accept in 60 minutes based on its physical runway configuration, weather conditions, and other factors and is measured in whole numbers.

Therefore, the output was replotted for ATL with the five cases where the AARs fell below 80 and are not shown by limiting the range of the *X*-axis and are presented in Figure 8. This is simply an expansion of the highlighted portion of Figure 7. If the useable error limit (again, arbitrarily set) is a positive or negative AAR difference of 13, acceptable model performance may be seen at actual AARs of roughly 105 or higher. An actual AAR of 105 or higher accounts for all but 296 cases scored using the 2016 data: 6,288 of the 6,584 cases, or 95.5% of the total cases analyzed.

In fact, 91.6% of all the 2016 cases studied had an absolute observed minus predicted AAR error of less than 13, and over half the cases had an AAR error of less than 4. However, even in the replotted graph presented in Figure 8, the decision tree model struggles with 296 cases with AARs below 105. Again, an over-forecast of airport capacity is seen at lower AARs, and a slight under-forecast of airport capacity is noted as the actual AAR climbs to its 132 maximum.

4.2. New York-John F. Kennedy Airport

JFK has a maximum arrival rate of 60, so an acceptable error based on 10% of the maximum AAR is a value of the absolute values of the observed minus predicted AAR of 6. Figure 9 shows the line graph of the difference between the actual and predicted AAR plotted against the actual AAR. These results were derived from the decision tree model using the merged hourly ASPM and meteorological station data. This model was selected as an underperforming model based on model validation using data withheld from the 2014–2015 data. Examining the variable importance for JFK using this dataset, the top five variables ranked by order of importance in supporting the model decision making were: (1) meteorological conditions (IMC versus VMC), (2) adjusted local hour, (3) ceiling, (4) temperature, and (5) wind speed.

As with some of the other models, difficulty with over-forecasting airport capacity occurred at the lower spectrum of AARs. Looking at the graph and associated data, there are 236 cases out of 8,780 where the observed AAR was less than a negative 35, and the model struggles to correctly map these outlying events. Further, only 44.4% percent of the 2016 cases scored fell within the plus or minus 6 AAR error thresholds for this decision tree. The steepness of the line's curve suggests that this decision tree model only performs well between AARs of 38 and 57.

V. Conclusions and Recommendations

This study sought to examine detailed historical NAS airport performance archives as well as environmental data to see if there are meaningful signals in these data that could gainfully apply machine learning predictively. Decision tree, neural network, and linear regression models were created and validated for 10 geographically dispersed airports with different arrival capacities using comprehensive FAA ASPM airport performance and NOAA NCEI 2014–2015 environmental datasets. The "best" models, based on the squared root of the validated model ASE, were scored using a full year's worth of data (2016) with the same formats as those used to previously create and validate the models. While many variables were available

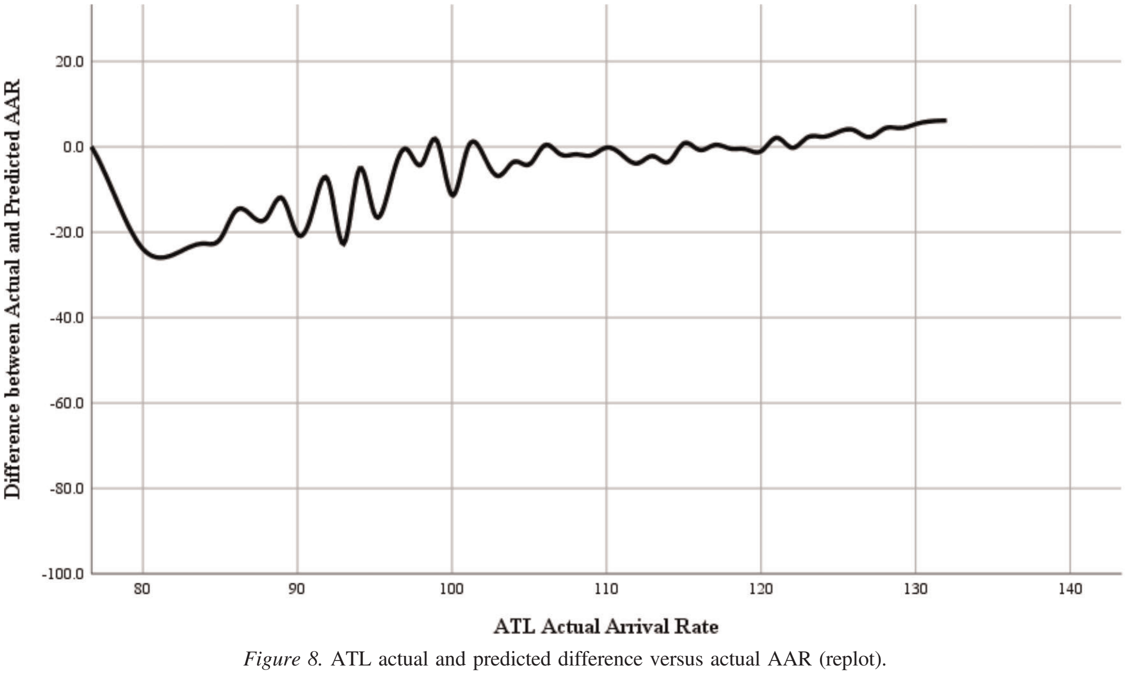

Figure 8. ATL actual and predicted difference versus actual AAR (replot).

to apply to the prediction of AARs in these datasets, ultimately, it was decided to only use weather variables in estimating airport arrival rates, as the goal of this research was to determine if NWS predictive weather model guidance could potentially be fed into the created models to estimate key airport AARs. It was hoped this effort could support FAA national airspace managers to estimate NAS capacity *a priori* to more efficiently regulate air traffic flows in weather-constrained airspace.

Using only weather variables to create, validate, and score the models, the results were mixed but positive. Based on this approach, three airports, ATL, EWR, and LAX, all exhibited superior 2014–2015 validated model performance as well as when scored using the 2016 data. All three of these "best" airport models were placed within the top five of the ten airport models created and validated, and all were decision tree models. Interestingly, while the top three models after scoring were all decision tree models, each employed a different type of data: LAX using hourly, ATL using hourly merged, and EWR using 15-minute datasets.

Seven of the "best" ten airport models selected gravitated toward decision trees, while the remaining three airport models settled on neural network models, with linear regression models failing to be selected for any airport as a "best" model overall—regardless of the dataset selected. This is likely due to the nonlinear relationships between the predictors and target variable; nonetheless, the regressions performed surprisingly well and perhaps reflect the power of using this modeling approach with a very large number of cases (over 15,000 to build and verify and over 6,000 cases to score the models). The large number of cases used appears to overpower the need to meet the basic parametric linear regression constraints required to ensure the selected sample is an unbiased representation of the population being estimated. In this study, all of the available cases were applied, and a linear regression was selected specifically to estimate a continuous AAR variable. While linear regression was not selected as a "best" model for any of the ten airports, Table 4 shows how favorably the linear regression modeling technique behaved (for the most part) when compared to the nonlinear decision tree and neural network models ultimately selected as the "best" models.

Model performance was, for the most part, remarkably consistent across each of the three model types created (decision tree, neural network, or linear regression models), and all model types used three different training and validation datasets. Based on the 2016 scored data, at least three airport model and dataset combinations, ATL (DT, hourly merged), EWR (DT, 15-minute), and LAX (DT, hourly), demonstrate a predictive capability that could potentially be deployed operationally. LGA and SFO, with validated model rankings (based on the squared root of the ASE) at ranking two and three, respectively, were somewhat disappointing when the 2016 scored model results were reviewed. However, the top five models, based on the model validation squared root of ASE, were also within the top five models based on the scored 2016 data—but the ranking orders were shuffled.

Based on the performance parameters used in this study, why do the EWR and LGA models rise to the top of the three New York airports while JFK has far less successful results, given these three airport locations experience nearly the same weather conditions? Considering the models tested and scored, the simple answer is the weather-based variables affect each airport model differently as meaningful predictors in capturing AAR variability, and this performance is relative to other non-weather inputs that also play roles in determining the AAR. The three airport models ranked the importance of weather variable inputs differently, and only EWR included visibility within its top five input variables—as its highest-ranked variable of importance. Also, EWR has a known weather constraint based on its Runway 11 crosswind component that significantly lowers its AAR when this runway becomes unavailable. Additionally, better results were expected at SFO due to the marine stratus conditions that have such a large impact on its AARs, but the results when evaluating the 2016 scored data were marginal (only 68.0% of all cases falling within 10% of the maximum AAR). More research is needed here.

Based on the 2016 scored model results, the 15-minute data only supported two "best" case models (EWR and CLT), while the hourly data supported four models (LAX, LGA, ORD, and SFO), and the hourly merged data supported the remaining four "best" models (ATL, DEN, DFW, and JFK). Within eight of the ten airports studied, the hourly and hourly merged datasets outperformed the 15-minute data, indicating the additional weather variables contained in the hourly and hourly merged data improved model performance overall.

The greater the complexity of the observed weather variables used to create the predictive AAR models, the greater the level of effort needed to approximate these same variables from the upstream feeding NWS predictive numerical weather models, raising the level of difficulty in deploying models constructed with more weather variables. Also, the number of forecast hours contained in the NWS weather models depends on the model selected. Some models run out to 80 hours and beyond, while higher temporal resolution models with shorter time steps (e.g., 15 minutes) cover a relatively shorter overall period of time (e.g., 24 hours). So, in designing a deployable predictive system, the underlying weather model used as input to the AAR-estimating model should match the predictive weather model's native time steps, spatial resolution, and extractable parameters. A weather-based predictive AAR model that cannot be easily supported by an underlying weather prediction model is not useful.

It is recommended that one, or all three of the "best" models created here be experimentally deployed for

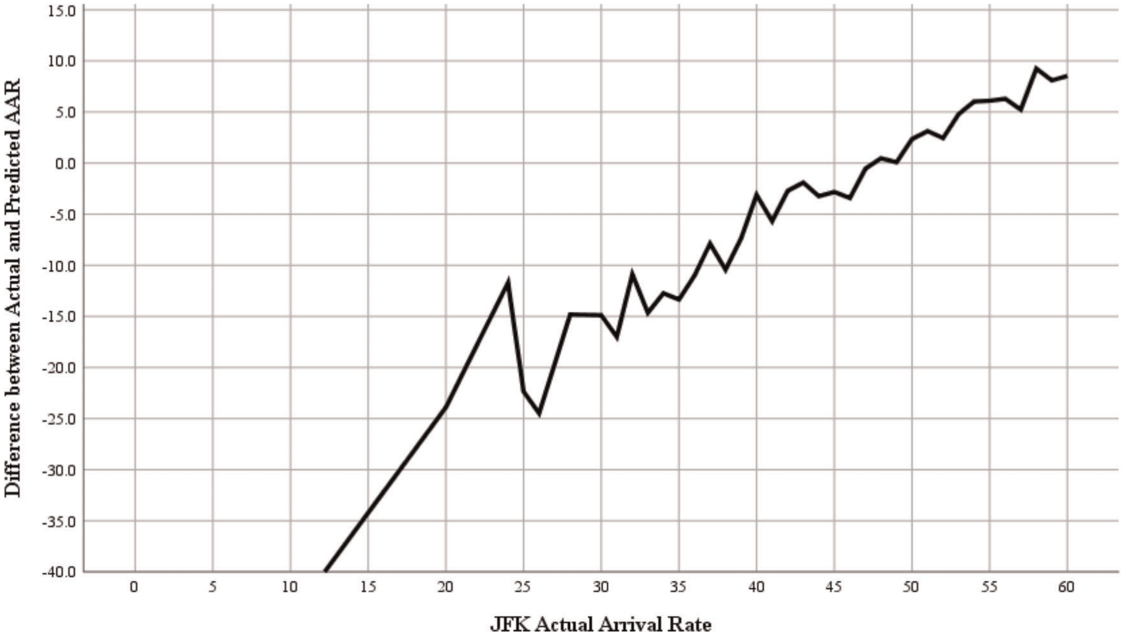

Figure 9. JFK actual and predicted difference versus actual AAR.

continuous observation. While not the highest-ranked model per the evaluation criteria used in this study, EWR as a 15-minute model employs relatively simple weather variable inputs that can be estimated and autonomously produced from NWS LAMP numerical weather model guidance. Using the computer code generated in this study by the SAS® EM™ decision tree model, the constant output of AARs fed by automated NWS meteorological weather input data could be monitored for accuracy for a lengthy period. The inspection of a prototype EWR predictive system will thoroughly examine the operational efficacy of this modeling approach and can also identify the strengths and weaknesses inherent with this design. Long-term observation and evaluation of such a system would shed a great deal of light on the positive and negative aspects of this machine learning modeling approach.

Compliance with Ethical Standards

Funding: There was no funding for this study.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

AviationFigure. (2015). *Worldwide flight delays cost airlines US$25B in 2014*. Retrieved from http://www.aviationfigure.com/worldwide-flight-delays-cost-airlines-us25b-in-2014/

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., …, Zou, B. (2010). *Total delay impact study: A comprehensive assessment of the costs and impacts of flight delay in the United States*. Technical report, National Center of Excellence for Aviation Operations Research (NEXTOR). Retrieved from https://www.isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

Dhal, R., Roy, S., Taylor, C. P., & Wanke, C. R. (2013). Forecasting weather-impacted airport capacities for flow contingency management: Advanced methods and integration. *AIAA Aviation Technology, Integration, and Operations Conference*, Los Angeles, CA. https://doi.org/10.2514/6.2013-4356

Federal Aviation Administration. (2013). *Types of delay*. Retrieved from http://aspmhelp.faa.gov/index.php/Types_of_Delay

Federal Aviation Administration. (2015). *FAQ: Weather delay*. Retrieved from http://www.faa.gov/nextgen/programs/weather/faq/

Federal Aviation Administration. (2016). *Facility operation and administration* (Change 1 ed.). Retrieved from https://www.faa.gov/documentLibrary/media/Order/7210.3Z_-_Change_1.pdf

Federal Aviation Administration. (2017a). *Measuring the performance of airports*. Retrieved from https://www.faa.gov/NextGen/snapshots/airport/

Federal Aviation Administration. (2017b). *National airspace system status*. Retrieved from https://www.fly.faa.gov/ois/

Hughes, D. (2016). *The FAA eyes big data possibilities*. Retrieved from http://www.atca.org/Big-Data

Kulkarni, D., Wang, Y., & Sridhar, B. (2013). Data mining for understanding and improving decision-making affecting ground delay programs. Paper presented at the *32nd IEEE/AIAA Digital Avionics Systems Conference (DASC)*, 5B1-1-5B1-8. Retrieved from https://www.google.com/search?q=predicting+airport+arrival+rates&ie=utf-8&oe=utf-8

Laskey, K. B., Xu, N., & Chen, C. (2012). *Propagation of delays in the national airspace system*. Retrieved from http://erau.summon.serialssolutions.com/

Sheth, K., McNally, D., Morando, A., Clymer, A., Somersall, P., & Shih, F. (2015). Assessment of a national airspace system airborne rerouting tool. *Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015)*. Lisbon, Spain. Retrieved from http://www.aviationsystemsdivision.arc.nasa.gov/publications/2015/ATM2015_Sheth.pdf

Smith, D. A. (2008). *Decision support tool for predicting aircraft arrival rates from weather forecasts*. Retrieved from http://erau.summon.serialssolutions.com/

Tufféry, S. (2011). *Data mining and statistics for decision making* [Data Mining et Statistique Decisionnelle]. John Wiley & Sons, Ltd.

Wang, Y. X. (2011). Prediction of weather impacted airport capacity using ensemble learning. *IEEE/AIAA 30th Digital Avionics Systems Conference*. Seattle, WA. Retrieved from http://erau.summon.serialssolutions.com/