En Route Flight Time Prediction Under Convective Weather Events

Guodong Zhu* Chris Matthews† and Peng Wei‡
Iowa State University, Ames, IA, 50011, U.S.A.
Matt Lorch§ and Subhashish Chakravarty¶
Rockwell Collins, Cedar Rapids, IA, 52402, U.S.A.

Flight en route time is expected to be longer when traffic demand exceeds airspace capacities. When demand is high, flights can often be subjected to rerouting, Miles in Trail (MIT) and vectoring, etc. During summer months, convective weather, which could severely affect the capacities of major air routes and airports, accounts for the most significant share of weather-relate delays. Flight en route time prediction could help airline dispatchers and traffic management coordinators make strategic plans, and alert travelers of potential delays. This paper aims to use machine learning techniques to model traffic volume and especially the effect of convective weather within 100 miles of great circle line between city pairs on the en route time. Several best off-the-shelf algorithms including LightGBM, XGBoost and Random Forest are tested and compared. These algorithms are trained and validated on four data sources: historical flight punctuality data, NEXRAD level III data, surface weather observing data and wind aloft data.

I. Introduction

Estimated Time En route (ETE) is one of key time-related parameters in airline operation. The others parameters include Estimated Taxi-In Time, Estimated Taxi-Out Time and Estimated Turn-round Time, etc. The common way to calculate ETE is based primarily on flight route/profile and aircraft performance information [1]. Only limited weather information like head/tail wind speed at cruise altitude is explicitly considered [2]. It is well known that weather is a major cause of delay in the National Airspace System (NAS). During summer months, especially April to July, most of the delays are due to convective weather, low ceiling and visibility, airport winds, etc [3]. Traffic volume also plays an important role in affecting en route time. Even in a clear weather day, inbound flights to the banked hubs are expected to have a longer en route time at the morning and afternoon peak hours. This paper is an endeavor to improve the pre-departure en route flight time prediction by employing machine learning techniques on weather and traffic information.

Many research works on predicting flight delays have been done. Choi, etc used Airline On-time Performance dataset and Integrated Surface Database to predict weather affected airline delays [4]. Using the same data sources, Kim, etc built Long Short Term Memory RNN model to predict the sequences of flight delays and claimed to get the best accuracy result [5]. Both papers focused on classification application. Rebollo and Balakrishnan employed random forest to predict aggregate delays at airport level and considered the delay state at other airports when making predictions [6]. Callaham, Klein, Sridhar, etc have been developing various Weather Impacted Traffic Index (WITI) models and using the models to predict delay at airport and system level [7] [8] [9]. There are three more relevant regression work to our paper. Glina, etc use Quantile Regression Forest to generate near arrival Estimated Time of Arrival (ETA) prediction and associated prediction interval widths for individual flights [10]. For flights 60 NM away from DFW, the Mean Absolution Error (MAE) is around 75 seconds. No weather data is used in [10]. Kern, etc use METAR data

*Graduate Student, Aerospace Engineering Department, Iowa State University, Ames, IA, AIAA Student Member.
†Undergraduate Student, Computer Science Department, Iowa State University, Ames, IA.
‡Assistant Professor, Aerospace Engineering Department, Iowa State University, Ames, IA, AIAA Senior Member.
§Senior Engineering Manager, Advanced Technology Center, Rockwell Collins, Cedar Rapids, IA.
¶Principal Systems Engineer, Advanced Technology Center, Rockwell Collins, Cedar Rapids, IA.

and scheduled traffic data to enhance flight ETA prediction generated by Enhanced Traffic Management System (ETMS), achieving MAE reduction of 42.7%[2]. GE Aviation proposed "Flight Quest Phase I" in 2012 to solicit solutions for gate and runway ETA prediction of en route flights. The Root Mean Square Error for runway ETA prediction obtained by the winning team is 3.2 minutes. Detailed flight data (flight history/plan/track), weather data (METAR/TAF/AIRSIGMET) and FAA advisories are provided to the participating teams[11]. Recently Sternberg, etc wrote a literature review on flight delay prediction problem from the aspects of data, method, problem and scope. Related researches like delay propagation are also covered[12]. As a matter of fact, because flight delay is so prevalent and affecting people's everyday life, predicting flight delays is a popular project topic in many machine learning classes and competitions. The contribution of this paper include: we use machine learning methods to do en route time prediction at pre-departure phase, and we are use more public available weather data than the existing work and we have a detailed introduction for data processing and comparison for a variety of algorithms.

The paper is structured as follows: in the first chapter, several high-level convective weather products, terminal and wind aloft prediction and weather translation models are briefly introduced. The considerations of selecting the three hub airports are also listed. The goal of the first chapter is not just to highlight the up-to-date aviation related weather technologies, but more importantly, it points to the sources of potential features that future versions of machine learning models can use. In chapter 2, the procedures of acquiring, cleaning and joining flight and meteorology data sets are presented in detail. The reasons for selecting these features are explained. In chapter 3, several commonly recognized best machine algorithms are tested and compared. Important analyses like feature importance ranking, correlation analysis are performed. It needs to be pointed out that there are many other important en route time related factors which are not currently considered in this paper, for example, filed route and pilot behavior modeling. Chapter 4 lists these features, if included in the data set, that could likely improve model performance. Chapter 5 is a summary of the main points of this report.

II. Major Convective Weather Products

1. Corridor Integrated Weather System (CIWS)

Developed by Massachusetts Institute of Technology (MIT) Lincoln Laboratory (LL), Corridor Integrated Weather System (CIWS) is a three-dimensional weather visualization, analysis and forecast product, and is widely used in many Air Traffic Management (ATM) decision support tools[13] [14] [15]. The goal of CIWS is to improve tactical ATM decision making by providing accurate short-term (0-2 hours) convective forecasts of Precipitation [Vertically Integrated Liquid (VIL)] and Echo Tops. Improved forecast accuracy enables air traffic controller to use otherwise lost capacities like the over-the-top route. NEXRAD raw data is the major data source of CIWS. There are many other data sources like Airport Surveillance Radar (ASR) that used by CIWS to provide different products or do quality check against NEXRAD. CIWS is not publicly available. More description about CIWS can be found in [16][17].

A related product which is specially designed for terminal weather forecast is Integrated Terminal Weather System (ITWS). The products of ITWS include wind shear and microburst detection and prediction, detail wind information, etc which are critical to airport operation[18] [19].

2. National Convective Weather Forecast Product (NCWF)

National Convective Weather Forecast Product is designed and implemented by National Center for Atmospheric Research (NCAR). The main features of NCWF include: visualization of the current convective hazards and 1-hour extrapolation forecasts of thunderstorm hazard locations; the diagnostic analysis combines NEXRAD with vendor-provided echo top mosaics and cloud-to-ground lightening; the Convective Hazard Detection field is depicted based on a 6 level intensity scale; Target user of NCWF are airline dispatch, general aviation and Federal Aviation Administration (FAA) Traffic Management Units (TMU). Since NCWF is designed to focus on large areas of convective activity, small convective cell complexes, which can be of concern for congested airspace, do not have NCWF forecasts[20]. More introductions and limitations of NCWF can be found in [21].

3. Collaborative Convective Weather Forecast Product (CCFP)

The Collaborative Convective Forecast Product (CCFP) was jointly developed by the FAA, NWS and the airline industry. It is a strategic forecast (2 to 8 hours in the future, updated every 2 hours) for intense convective activity based on many different inputs (NWS, Center Weather Service Units (CWSUs), airline meteorologists, etc). CCFP is designed for traffic management personnel (operation managers and flow managers) to mitigate the influence of severe convective events. It is not intended to forecast all thunderstorm activity, but only these with tops above 25,000 ft and 25% or greater coverage. More information about CCFP can be found in [22][23].

4. High-Resolution Rapid Refresh (HRRR) Data

HRRR model, which covers the contiguous United States, is a subset and derived from the Rapid Refresh (RAP) model. HRRR model is designed for aviation and nowcasting. It has the highest spatial resolution (3 km) as well as high temporal resolution (1 hour). Many weather datasets, including NEXRAD data, Radiosonde data, METAR data and Satelite winds, etc and aircraft data are assimilated by HRRR model. Because of its higher spatial and temporal resolution, HRRR model could capture the development, structure and movement of thunderstorm more accurately.

The full list of HRRR model products can be found on the website of Earth System Research Laboratory [24]. The last two days of HRRR output can be downloaded from NOMADS server. Unfortunately, there is no official website to archive the historical HRRR data. University of Utah runs an unofficial archive and only data after June 15, 2016 are available [25].

A. Other Relevant Weather Forecast Products

1. Terminal Aerodrome Forecast (TAF)

TAF is the international standard code format for terminal forecasts issued for airports. TAFs are issued 4 times a day at 6 hour intervals and are generally applied to 24 or 30 hour period. The forecast includes forecasted wind speed, wind direction, visibility, ceiling, type of precipitation (i.e. snow, rain, etc.) and/or weather phenomenon [26].

2. Winds and Temperatures Aloft Forecast

Winds and Temperatures Aloft Forecast (FBs) is generated by National Centers for Environmental Prediction(NCEP) from runs of the North American Mesoscale (NAM) forecast model. The forecast is issued four times a day for the following levels: 3, 6, 9, 12, 18, 24, 30, 34, 39 thousand feet. The forecast is useful for pilot in determining the most favorable altitude abased on winds and direction of flight, identifying areas of possible aircraft icing and temperature inversion and predicting turbulence by observing abrupt changes in wind direction and speed at different attitude [27].

Wind aloft prediction is of fundamental importance to enable Next Generation Air Transportation System (NextGen) concepts like Four-Dimensional Trajectory Based Operations (4D-TBO). A detailed introduction to more wind prediction models and requirements can be found in [28].

B. Weather Translation Models

Use the definition from Krozel's paper [29], weather translation model transform weather forecast data into characterizations of weather-related NAS constraints, threshold events (e.g. VFR to MVFT) and expected resource performance (e.g. capacity and permeability of a NAS element). There are two representative weather translation models: Convective Weather Avoidance Model (CWAM) and Convective Weather Avoidance Polygon (CWAP). Developed by MIT Lincoln lab using detailed weather observations and aircraft trajectories, CWAM models the probability that a pilot will change the route to avoid convective weather. To predict the most likely weather-avoiding trajectories, CWAP is further developed by MIT Lincoln to define the boundary of a no-fly zone. Weather translation techniques are also frequently used to determine the physical acceptance limit of Flow Constraints Area (FCA), which is the upper bound for the actually acceptance rate.

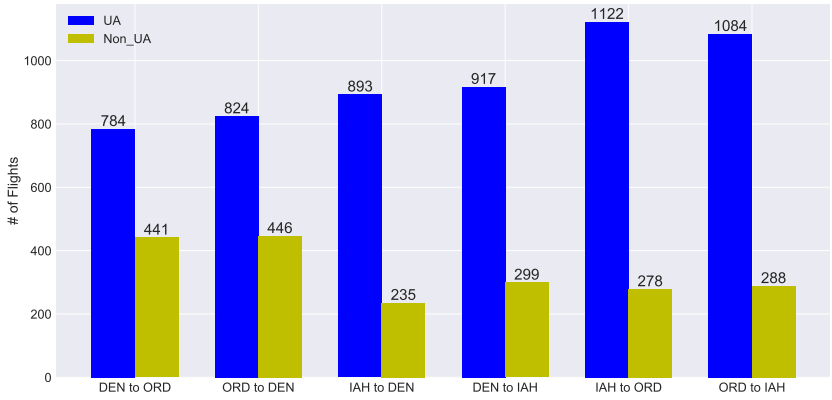


Figure 1: Number of UA and Non-UA Flights by City Pair May 2013-2015 (AOTP Data)

III. Data Acquisition, Cleaning and Manipulation

Three hubs of United Airlines (UA) IAH, DEN and ORD are selected for this paper. The intention is we will only focus on flights belonging to a single airline company to minimize the impact of airline behavior. As shown in Fig 1, UA flights account for 2/3 or more total traffic between three city pairs. Also these three cities are among top-six weather-delayed U.S. major airports. The main weather issue of IAH is thunderstorms and fog. The problem of DEN result from snowstorms, thunderstorms, hail and tornadoes. ORD is suffer from snow, low clouds/fog, wind, thunderstorms [30]. Convective weather can be blamed for the on-time performance of all three airports in the summer times.

Nine months data from four publicly accessible datasets are used in this paper. Bureau of Transportation Statistics Airline On Time Performance dataset (AOTP), which is the historical flight punctuality data (no flight route information) [31]; Automated Surface Observing System (ASOS) data, which is used to evaluate airport surface weather condition [32]; The Next Generation Weather Radar (NEXRAD) level III data, which consists numerous meteorological products, including several products that are used to predict and track convective weather [33]; Integrated Global Radiosonde Archive (IGRA) data, which comes from radiosonde and pilot balloon observations, is used to obtain the wind direction and speed at cruise altitude [34]. The reason to consider these three weather datasets are: if the convective activities block the historical flight route, the flight needs to take reroute and hence the en route time will usually increase; degenerated weather condition at the destination airport will force some flights to take air delay; wind aloft will influence the ground speed, thus affects the en route time.

In this chapter, we will talk about how to obtain, clean these datasets and discuss what the useful features are. After getting the target and the features needed, we will join them into one table to be used in training and analyzing machine learning algorithms.

A. Airline On Time Performance (AOTP) Data

AOTP data is the on-time arrival data for non-stop domestic flights by major air carriers since 1987. It includes features such as origin and destination airports, carrier ID, OOOI times (gate out, wheel off, wheel on, gate in), airtime (en route time) and causes of delay, etc. This data set is easy to access and is already in clean and well organized csv files. We use Selenium and and Chrome driver to automatically download AOTP data in batches.

The columns we choose are

1. **Flight date, CRSDepTime, wheel off time**. Flight date is local scheduled departure date (scheduled gate out date). Flight data and wheel off time determine when did the flight leave the origin airport and therefore what weather condition the flight was going to subject to.

2. **Carrier, flight number**. En route time is also largely affected aircraft type, the information AOTP doesn't provide and ideally can be mapped from flight date and flight number. Here we use carrier as a feature to capture its choice choosing certain fleet types to fly city-pairs.

3. **Origin and destination airports**. We are only interested in flights that serve between three a city-pairs.

4. **Airtime**. This is the target of the machine learning algorithm.

We use the number of hourly actual landing flight as a feature to model the traffic demand, because it is easy to obtain from AOTP data. In reality, in clear weather day the number of scheduled and actual landing flights is very close; in a bad weather day, the number of actual landing flights can be estimated from Airport Acceptance Rates (AAR), which is available from Flight Schedule Monitor (FSM), etc. Another option is to use the schedule demand data, as it is the true demand. Scheduled demand information can be downloaded from the FAA Aviation System Performance Metrics (ASPM) Dataset [35].

B. Automated Surface Observing System (ASOS) Data

ASOS data are generated from the automatic observation of runway visual range, wind speed and direction, etc around ASOS sites. There are three different temporal resolution ASOS data: 1-minute, 5-minute and hourly. ASOS data, if reported hourly, is compiled into Integrated Surface Database (ISD). Because ISD data is integrated from many data sources, thus has more columns than ASOS data. We use rnoaa package to get ISD data from three ASOS stations near DEN, ORD and IAH. The data field definitions and values for missing fields can be found in INTEGRATED SURFACE DATA documentation [36]. Linear interpolation results are used to to replace missing values. Data is saved as csv files.

We select the following columns

1. **Data and time**. They are used to interpolate and joint the data sets.

2. **Wind direction and speed**. They will affect the flight landing.

3. **Ceiling height**. Low ceiling height will necessitate instrument flight rules.

4. **Visibility**. Low visibility will affect airport surface operation and therefore throughput.

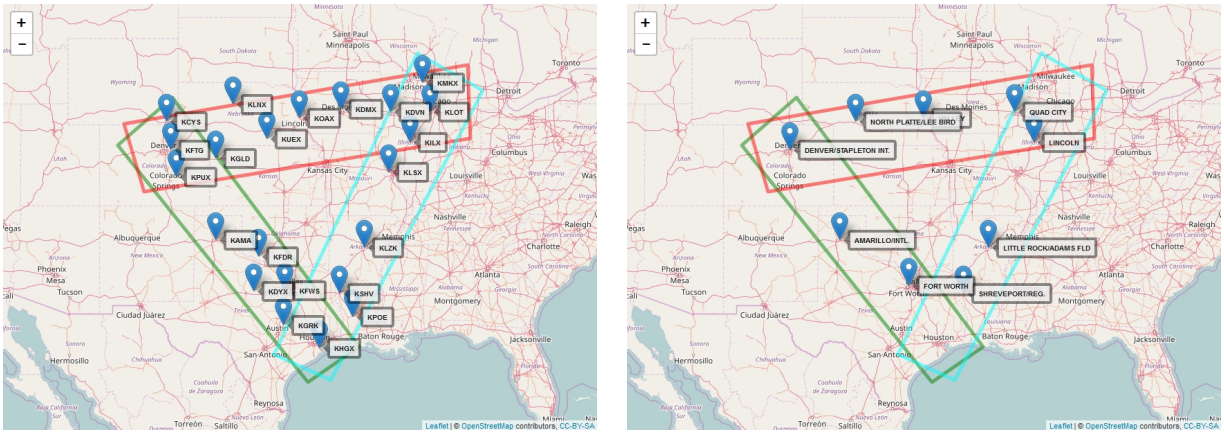
5. **Temperature, dew point, air pressure**. They are important meteorological parameters that related precipitation.

C. NEXRAD Level III Data

NEXRAD level III data are derived products from Level-II base data (reflectivity, mean radial velocity, and spectrum width, etc). NEXRAD Level III has over 75 products including storm tracking information, hail index, storm structure, etc. We use two convective weather related products: Digital Vertically Integrated Liquid (DVIL), which is effective indicator of storm and heavy rain fall, and Enhanced Echo Tops (EET), which is the radar indicated top of an area of precipitation.

There are 22 NEXRAD stations located in these three 200 miles wide corridors, shown in Figure2a. NEXRAD data is in a unique binary format. We use a Java API from Unidata to convert Level III data into NetCDF files, then use Casssandra database on Amazon AWS to further process these files.

NEXRAD data is stored by data products in matrices, with two axes indexed by azimuth and range respectively. The location of the station, azimuth angle and range distance uniquely determine a cell in the airspace. NEXRAD data can not be directly used by machine learning model since there are too many cells for each station and too many data products. We want to extract a single feature to represent the severity of convective clouds blocking jet routes. The method used here (Algorithm 1) is similar to the scanning method in the WITI algorithm [37]. We create a set of parallel flight lines as surrogates for jet routes, find the cells that could possible block those lines, check if any of those flight lines intersect the blocking cells, and then calculated the percentage of blocked lines or *blocked ratio*. A blocking cell is defined as any cell where $\texttt{Flight Altitude - EET} \leq \texttt{4000}$ *and* $\texttt{DVIL} \geq \texttt{3.5}$. This criterion is a research result of MIT Lincoln Lab [38].



(a) 22 NEXRAD Stations

(b) 9 IGRA Stations

Figure 2: Locations of NEXRAD and Wind Stations

Algorithm 1 Blocked Ratio Calculation

1: Take flight corridor (OD city pair) and a timestamp as input, and get the list of stations (precalculated) that intersect the corridor
2: For stations in list, convert, read and process the closest radar products within 10 mins of the timestamp. Processing involves finding the locations of potential blocking cells
3: Creating evenly spaced lines (225) parallel to the original flight line (OD Great Circle)
4: For each flight lines, iterate through potential blocking cells to check intersection
5: Blocked ratio is the number of blocked lines divided by the total number of lines

1. Verification of Blocked Ratio Calculation

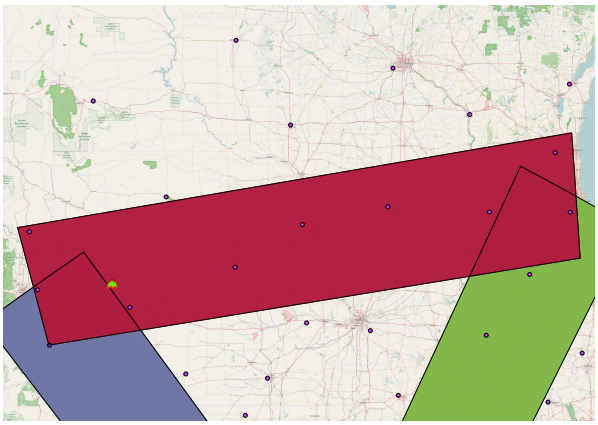
To verify the blocked ratio calculation, we visualize the blocked cells and the flight lines using QGIS, and compare the result from NOAA's Weather and Climate Toolkit (WCT). The logic is: if the blocked area overlaps with severe weather affected region shown in WCT, we have confidence the weather products mosaic and visualization was done correctly; if the percentage of flight lines intersecting the blocked regions is the same with the blocked ratio, we are assured that ratio computation is right.

Three timestamps are selected in May 1, 2015, which is a bad weather day. In Figure 3a, the green region near Denver is the blocked area. We export the shapefiles from WCT, create a single shapefile for that time stamp, filter out the non-blocking cells and show the result as red boundary. We can see the red boundary matches perfectly with the blocked area. In Figure 3b, it can be seen that in total 7 out 225 flight lines are blocked in DEN-IAH Corridor, so the blocked ratio is 0.031, which is the same as the result from Python code. Figure 4 shows the comparison between the blocked area and Echo Tops picture in CIWS. We find the weather affected regions matches well.

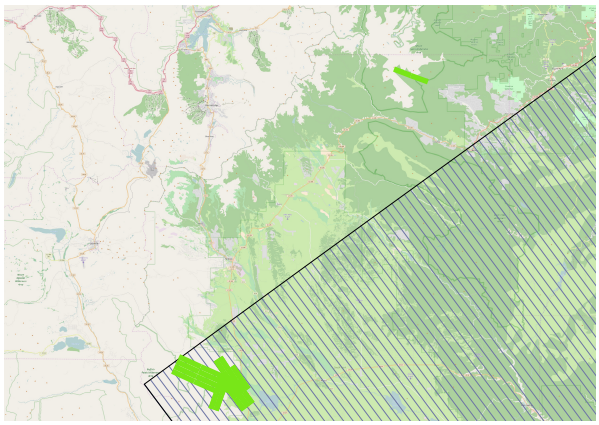
D. Integrated Global Radiosonde Archive (IGRA) Data

Reported from radiosonde and pilot balloon observations twice a day, IGRA data consists of pressure, temperature, relative humidity, dew point depression, wind direction and speed, etc at different pressure and geopotential height. There are 9 IGRA stations in these three corridors, shown in Figure 2b. We assume the cruise altitudes of all flights are 35000 ft. We only interested in the wind direction and speed at the cruise altitude. To study the effect of large-scale variations in winds aloft, in each corridor 3 stations are selected and wind is decomposed into head/tail wind and crosswind assuming flight flying great circle. IGRA data is in condensed radiosonde format. We use Python to parse, clean the data file and save the wind information at the cruise altitude as csv file.

Figure 5 shows the head and cross wind for flights from DEN to ORD. The fluctuations of wind speed are fairly consistent for the three stations, which verifies our understanding that different from surface wind,

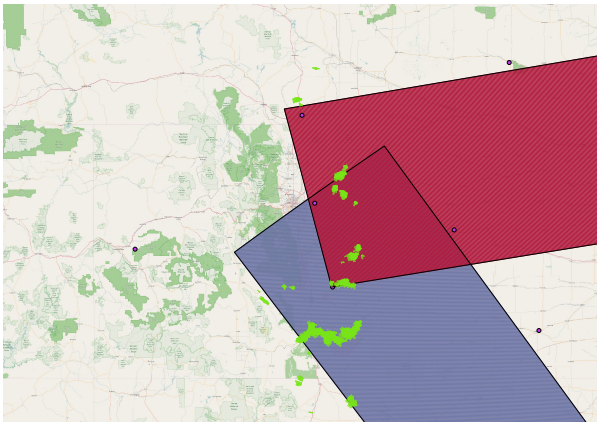


(a) Blocked Area 3am 05/01/2015 GMT

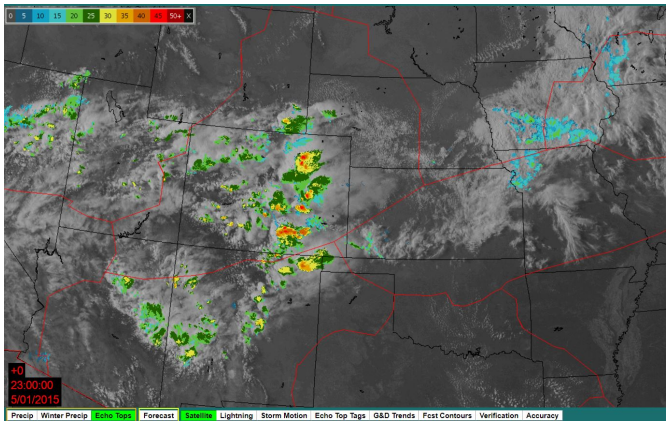


(b) Blocked Area 5pm 05/01/2015 GMT

Figure 3: Blocked Area Matches with WCT Results



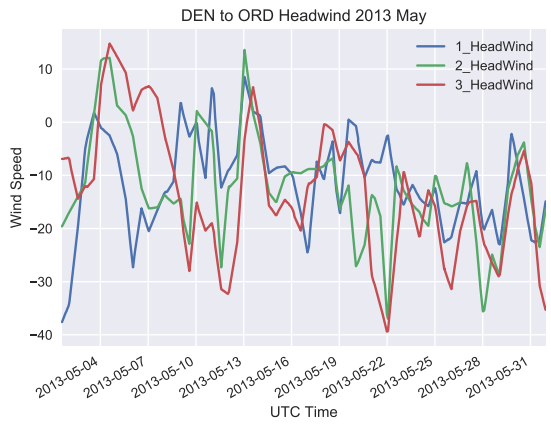
(a) Blocked Area 11pm 05/01/2015 GMT



(b) Echo Tops 11pm 05/01/2015 GMT (CIWS)

Figure 4: Blocked Area Comparison with Severe Weather Region Shown in CIWS

wind aloft has similar wind direction even across a large region.



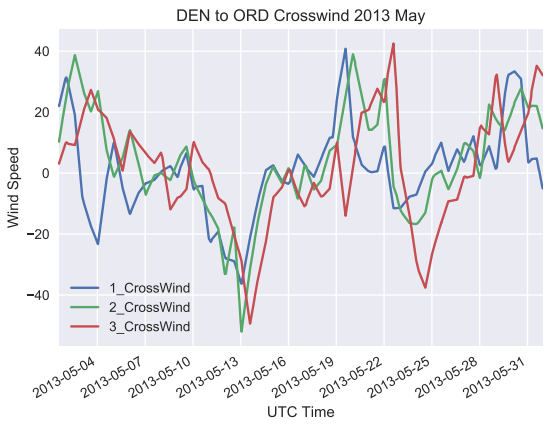


Figure 5: Wind Information for Flights from Den to ORD in May 2013

E. Joining Datasets

It needs to be pointed out that the three aforementioned weather products: NEXRAD, ASOS and IGRA are weather observation data, rather than weather forecast data. However, because we are predicting the en route time of short-haul flights, and short-term convective weather is reasonably accurate, we could use these products to get a sense how much benefit machine learning and weather products at this resolution could bring us. To develop more realistic prototype, weather and traffic products introduced in chapter 1 and 4 can be considered.

AOTP data is in local time. From May to July, which are the three months under study, ORD and IAH use Central Daylight Time (CDT UTC-5 hours) and DEN is in Central Standard Time (CST UTC -6 hours). All three weather products are recorded using UTC time. ASOS data is employed to capture the impact of landing condition on the en route time. When joining ASOS with AOTP data, for every flight we use the ASOS information that is 2 hours ahead of its departure time. NEXRAD and wind aloft data are used to represent the en route weather condition. When joining these two weather products with AOTP, we use the information that is 1 hour ahead of a flight's departure time. In total there are 15 features: 1 (actual hourly landing flights) + 7 (ASOS columns) + 1 (NEXRAD blocked ratio) + 6 (head/tail and cross wind at three stations).

IV. Methods and Results

A. Preprocessing and Some Discussions

We first standardize the dataset, which is beneficial for the numerical stability and is a requirement for some algorithms like Support Vector Machine(SVM) with RBF kernel. Fig 6 shows the box plot for the en route times between three city pairs. It can be clearly seen that even for the same city pair, the average en route times are distinctly different. We use one hot encoding to transform city pair information to six boolean columns. It needs to be pointed that there is a time dimension in the date. A natural question is should we treat it as a time series prediction problem or an ordinary machine learning problem where data points are considered independent. We regard the time component in the data relevant but not as important as in applications like stock market prediction. In this report we do not consider the problem as time series problem. More discussions are left to Chapter V. 80% of data is used as training set and 20% is for test set. We use 5-fold cross validation to pick the best hyper-parameter combination for each algorithm. Python scikit-learn package is used to implement the algorithm, unless stated otherwise.

Root Mean Square Error (RMSE), which represents sample standard deviation of the prediction error, is an important metric we care about. Another index we pay attention to is the percentage of data whose prediction error is within 5 minutes and 15 minutes. R^2 is a commonly used metric in regression. We don't put too much emphasis on R^2 because the criterion for a good R^2 differs for different applications.

B. Generalized Linear Models

Because generalized linear models are easy to implement and interpret, therefore we use it as a benchmark for other methods. The other goal to use generalized linear models, specifically Lasso [39], is to identify the factors that are mostly likely to cause flight delays. The objective function of Lasso take the following form:

$$\frac{1}{2n}\|y - Xw\|^2 + \alpha\|w\|_1$$

Larger α results in smaller coefficients. Bias-variance tradeoff can be made by adjusting this hyper-parameter. Using cross validation, α is picked as 0.001.

From Fig.7a we can see blocked ratio has the largest positive value, which shows the value of including NEXRAD data. After normalization, blocked ratio 1 corresponds to 2.60, which is estimated to cause 8.0 minutes extra en route time. Headwinds are all associated with positive coefficients, which is consistent with our command knowledge that headwind increases en route time. Because the average en route time between IAH and ORD is the largest among three city pairs, hence categorical variable ORD_IAH and IAH_ORD have positive correlation with en route time. The positive sign of AAR, negative signs of visibility distance and ceiling height are all in line with our knowledge.

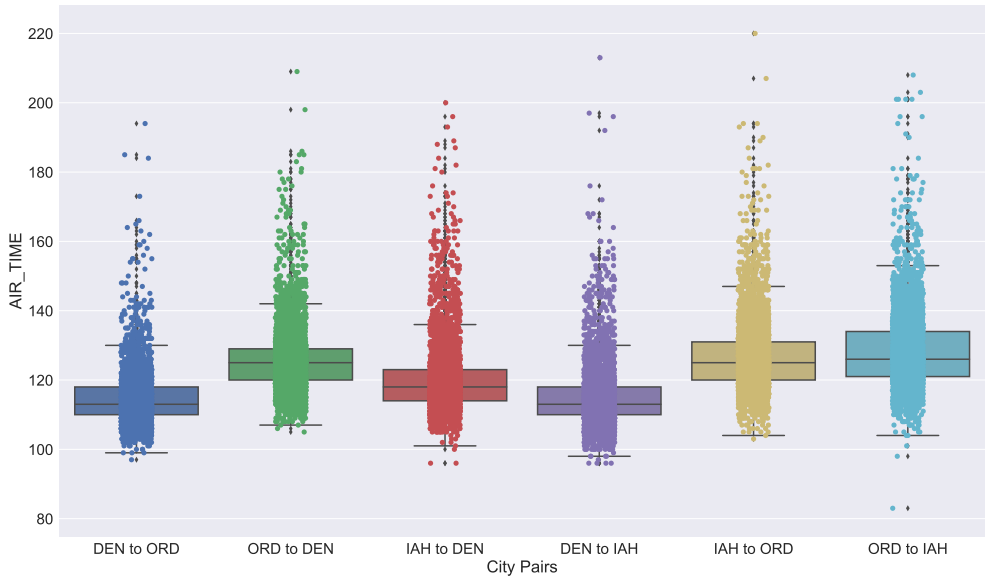
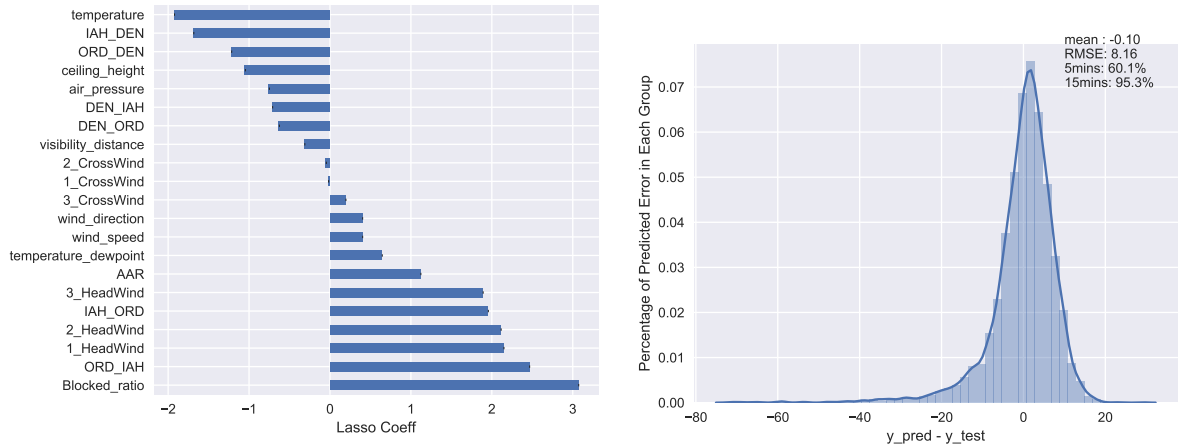


Figure 6: En Route Time Box Plots



(a) Lasso Coefficients

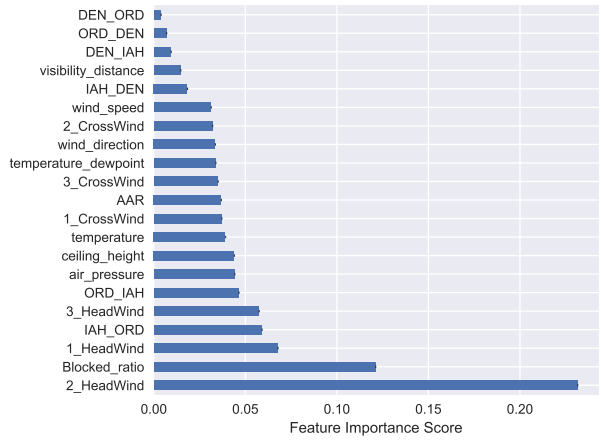
(b) Histogram of Prediction Error

Figure 7: Generalized Linear Model Lasso Results

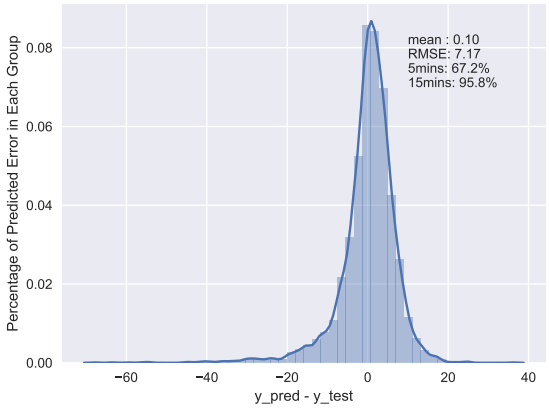
C. Random Forest Regression

Using trees as building blocks, Random Forest (RF) is a widely used ensemble learning method that fixes overfitting problem of a single decision tree. Random Forest is an extension to Bagging method. By only considering a subset of random features in each candidate split, random forest decorrelates decision trees and therefore improves the generalization performance [40].

There are three main hyperparameters to tune in random forest: number of trees, maximum features considered when making split and maximum depth. Random forest doesn't suffer from overfitting as the number of trees increase, so we pick a sufficiently large number 1000 as the tree number. Maximum number of features controls the process of decorrelation and maximum depth affects the complexity of a single tree. Based on the cross validation results, maximum feature number is selected as 5, which is the square root of total number of features. Maximum depth is 18. We use the default values for the other hyperparameters. The result of random forest is shown in Figure 8.



(a) Random Forest Feature Importance Ranking



(b) Histogram of Prediction Error

Figure 8: Random Forest Results

D. Boosting Tree

Different from random forest which build trees independently on bootstrapped datasets, boosting grows trees sequentially, with each tree fitted on a modified version of the original dataset. There are two state-of-the-art boosting trees algorithms: XGBoost [41] and LightGBM [42]. Apart from speed advantage, XGBoost uses a more regularized model formulation to control overfitting, thus generally has slightly better performance compared with original version of gradient boosting tree. Same as XGBoost, LightGBM is designed to be distributed and efficient. By changing from depth-wise to best-first tree growth and changing the way to handle categorical variable, developers of LightGBM claim their algorithm also has accuracy improvement in comparison with existing boosting frameworks.

Three major tuning hyperparameters in boosting is number of trees, learning rate and maximum depth (in LightGBM, number of leaves plays the same role). These hyperparameters are determined (Table 1) using cross validation for both methods.

In many application, data rather than algorithm is the deciding factor for success. As shown in the classical paper [43], for a given problem with very large dataset, very different algorithms perform almost the same. We draw the learning curves for our three best performing algorithms in Figure11. It can be seen that there is large gap between training and cross validation curves and as the number of training example increases, there is a notable increase in the model performance. These are the clear evidences for a high variance problem and increase the data size will help improve the result.

One disadvantage of feature importance ranking is that it can not tell us whether an importance feature positively or negatively influence the target. By normalizes the influences of other predictors, Partial Dependence Plot (PDP) gives a visualization of the marginal effect of a given variable (or multiple variables) over the target [44]. Current version of scikit-learn only support PDP for gradient boosting. From Figure13 we can clearly see that lower ceiling height, shorter visibility distance, high head wind and high blocked ratio contribute to longer en route time. Larger AAR also results in longer en route time, but the trend is not significant.

Models	# of Trees	Learning Rate	Maximum Depth/# of Leaves
XGBoost	1000	0.01	10
LightGBM	1000	0.01	186

Table 1: Boosting Methods Selected Hyperparameters

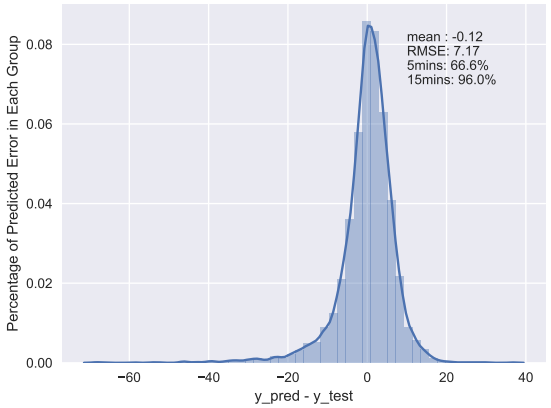


Figure 9: Histogram of XGB Prediction Error

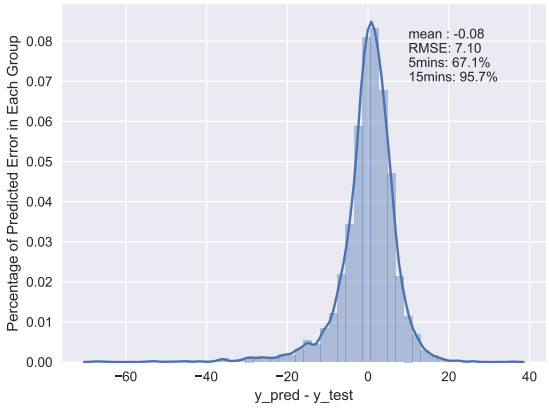


Figure 10: LightGBM Prediction Error

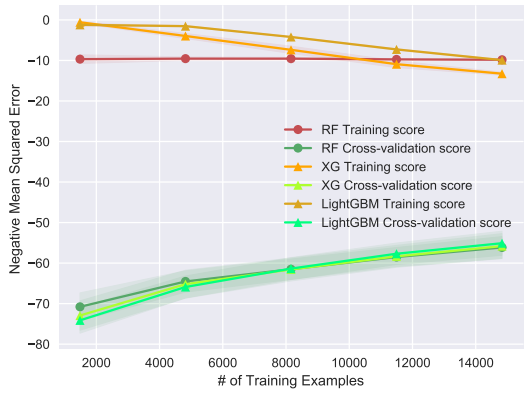


Figure 11: RF and XGBoost Learning Curves

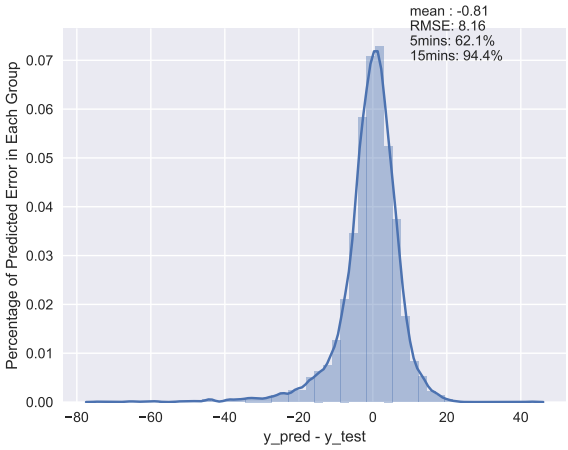


Figure 12: Histogram of NN Prediction Error

E. Neural Networks

As one of most commonly used machine learning algorithms, neural networks is well known for its powerful approximation ability and its black box nature, heavy computation requirement and proneness to overfitting[45]. In this paper, neural networks is implemented using Keras with Tensorflow backend and compared with other methods. The major tuning hyperparameters include number of hidden layers, number of hidden unites and batch size. The hyperparameters we use is single hidden layer, 45 hidden units and batch size 40. The result is shown in Figure12.

F. Other Machine Learning Methods

Two nonparametric methods: K-Nearest-Neighbor (KNN) and Support Vector Regression (SVR) are also tested. For KNN, two most important hyperparameters we can optimize are the number of neighbors and weighting metrics. From cross validation result we choose the number of neighbors as 10 and the distance based weighting strategy. For SVR, we choose RBF kernel. The penalty parameter and kernel coefficient, which are used to balance bias and variance, are selected as 500 and 0.01 respectively.

Summary for the performance results of all 7 algorithms are listed in Table2. Three tree ensemble methods are the best performers, which are anticipated since they have won numerous machine learning competitions in recent years. Apart from easy implementation and very few tuning parameters, KNN gives us the best

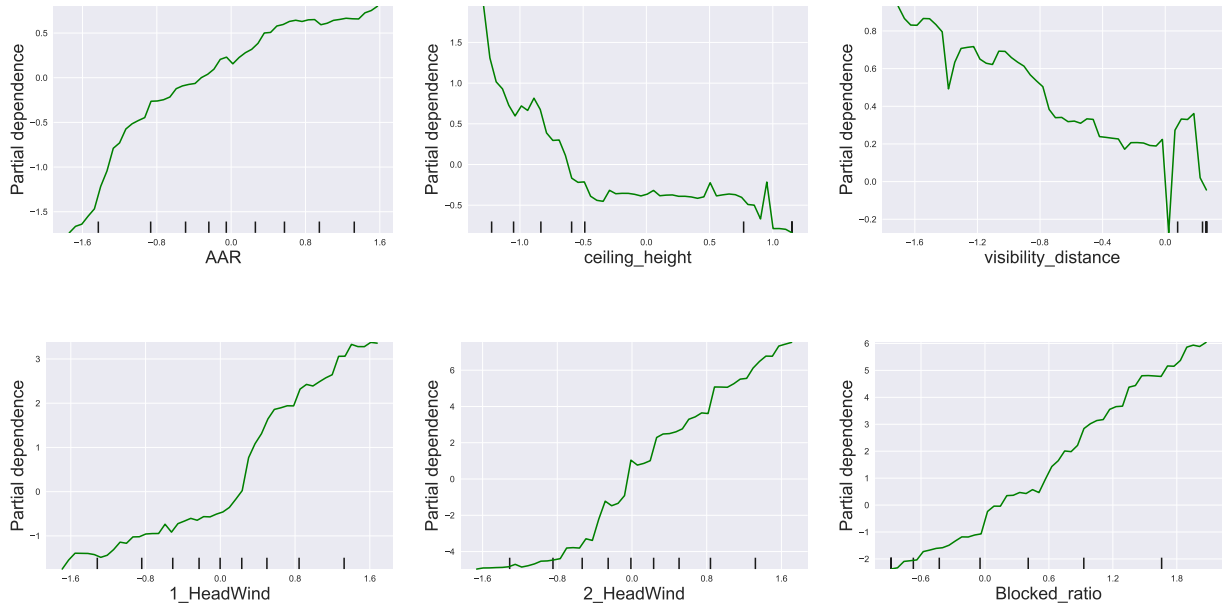


Figure 13: Partial Dependence Plot for Selected Features

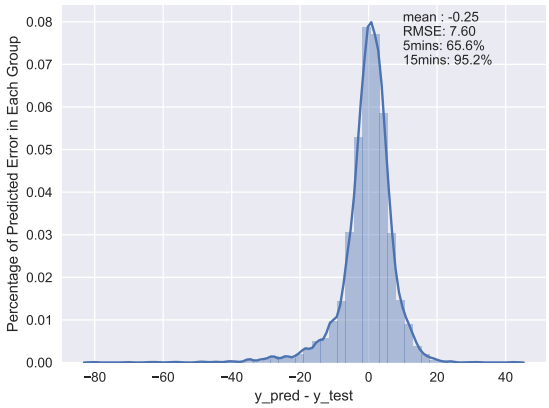


Figure 14: Histogram of KNN Prediction Error

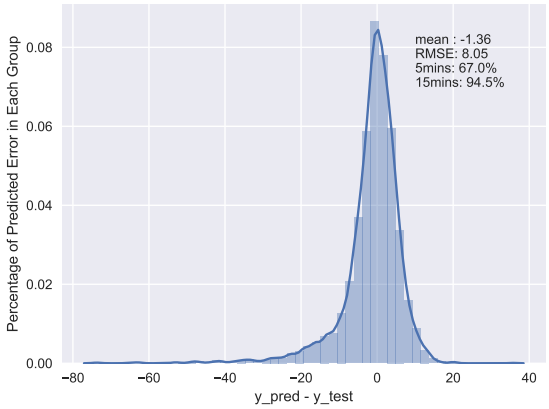


Figure 15: Histogram of SVR Prediction Error

result among all non-ensemble models using the least amount of time among all 7 models.

V. Future Works

On the model side, potential improvement could be achieved if we treat our problem as a time series predictive problem. By explicitly considering the time dimension rather than regard the data as independent and identically distributed, we can hope the temporal effects like arrival queuing can be better captured. Stratified random sampling based cross validation will be replaced by time series cross validation. Dedicated methods like recurrent neural network will be used.

Once a model is selected, in general there are two ways to improve the model performance. One way is feature engineering, which is to use domain knowledge to introduce or create more relevant features. The other way is gather more data, as it could mitigate overfitting for high capacity models. In this chapter, we focus on feature engineering and we will talk about where can we get better pertinent features to model

MODELS	RMSE	R^2	5/15MINS PCT	RMSE BY AIR ROUTE					
				ORD DEN	DEN ORD	IAH DEN	DEN IAH	IAH ORD	ORD IAH
GLM Lasso	8.16	0.48	60.1%/95.3%	7.29	6.91	9.66	7.79	8.84	8.36
RF	7.17	0.60	67.2%/95.8%	6.55	6.31	8.37	7.08	7.86	6.79
XGBoost	7.17	0.60	66.6%/96.0%	6.58	6.27	8.32	6.76	7.95	6.76
LightGBM	7.10	0.61	67.1%/95.7%	6.53	6.14	8.20	7.03	7.85	6.80
NN	8.16	0.48	62.1%/94.4%	7.89	6.73	9.49	7.83	8.75	8.19
KNN	7.60	0.55	65.6%/95.2%	6.88	6.60	8.73	7.49	8.40	7.44
SVR	8.05	0.49	67.0%/94.5%	7.19	6.59	9.43	7.90	8.85	8.23

Table 2: Summary of Machine Learning Results

traffic and weather situation.

A. Traffic Related Features

Different aircraft have different performances. For example, among the three typical short to medium range aircraft series A320, B737 classic and M80, A320 has a higher cruise speed than B737 classic family and M80. Adding fleet type information to the dataset is anticipated to improve the prediction. To the best of our knowledge, there is no publicly available dataset that could map historical flight number aircraft type information.

Another relevant feature is airport flow configuration, which is mainly driven by wind direction. For example ORD has two primary directional traffic flows: east and west. It is expected flight from DEN to ORD will have different en route time for different configurations.

Even between the same city pair, there are several commonly filed routes. The reasons behind routing decisions include severe weather avoidance and arrival fix offload. A flight route sometimes can be modified several times before departure. One related work on predicting pre-departure routing decisions was recently done by Arneson, etc [46]. Like aircraft type information, historical flight route information is not publicly available and can be purchased.

We have some sample data for flight from ORD to DFW. Other conditions being equal (clear-weather-day, same flow direction, similar landing time), the average en route of M80 is 115.3 minutes, B737 114.5 minuets, A320 114.2 minuets. For the same aircraft type B737, a flight will spend around 4 minutes longer in north flow than in south flow. We have some national playbook route information between major airports. Table 3 shows some commonly flew routes between IAH and DEN. We can see that the difference of nominal Expected En route Time (ETE) can be as large as 30 minutes.

NAME	TYPE	FREQUENCY	ORIGIN	DEST	NMI	ETE SECONDS
HOLIDAY_GULF_ROUTES	PLAY	6	IAH	DEN	1239	9360
DEN_GCK_1_MODIFIED	PLAY	4	IAH	DEN	969.32	7500
TEXAS_TO_DEN	PLAY	4	IAH	DEN	1112.93	8460
TX_TO_DEN	PLAY	1	IAH	DEN	881.99	6900

Table 3: DEN to IAH Route Options Example

To model the network effect in the NAS on the en route time, similar to the authors did in [6], we could include the average delay information of all three airports at the time of prediction (or even include other major hub airports) as model features.

Traffic Flow Management (TFM) actions like Mile-in-Trail (MIT) can also have an impact on the en route time, as some flights may be instructed to speed down or do vectoring to maintain required space. Historical TFM information can be acquired from National Traffic Management Log (NTML).

The number of air-hold flights at a given time near an airport is another useful feature, since it is directly related to the queueing effect. This information can be ordered from commercial vendors.

B. Weather Related Features

One apparent drawback of blocked ratio is that it doesn't have *depth* information. A giant convective cloud that totally blocked all the arrival fixes and runways may have a blocked ratio less than 1. Whereas several dispersed convective clouds with big gaps in between can possibly block all the flight lines. Part of the reason we adopt this scanning-based approach is that we don't have routing and position information. If we could get this information from Aircraft Situation Display (ASDI) data, it would be very valuable to check the relative location of the filed route and blocked area and determined the extra miles that needs to be flown.

Higher temporal and spatial resolution products can help to increase the prediction performance. For example, we can use CIWS/HRRR as weather forecast source for terminal and en route meteorological conditions, and use HRRR for wind aloft prediction. Weather translation models, if available, can also be used and is expected to bring great benefit to the model performance.

VI. Conclusions

This is the first work that use machine learning methods to do pre-departure en route time regression at the flight level. Several state-of-the-art machine learning models have been trained and tested on nine months publicly available traffic and weather data. The RMSEs of the best performing models are around 7.1 minutes. We pointed out the future work that can be done to improve the model performance. Among these we think gathering more data, adding aircraft type, flow configuration and filed route information will be great help to improve the prediction.

VII. Acknowledgements

The authors from Iowa State University would like to gratefully acknowledge Heather Arneson at NASA Ames Research Center, Scot Campbell and Richard DeLaura at MIT Lincoln Laboratory and Kevin Kronfeld at Rockwell Collins for their helpful discussions on processing the weather data.

References

[1] Murphy, J., Reisman, R., Clayton, J., and Wright, R., "Physics-based and parametric trajectory prediction performance comparison for traffic flow management," *AIAA Guidance, Navigation and Control*, 2003.

[2] Kern, C. S., de Medeiros, I. P., and Yoneyama, T., "Data-driven aircraft estimated time of arrival prediction," *Systems Conference (SysCon), 2015 9th Annual IEEE International*, IEEE, 2015, pp. 727–733.

[3] Federal Aviation Administration, "FAQ: Weather Delay," `https://www.faa.gov/nextgen/programs/weather/faq/#faq1`.

[4] Choi, S., Kim, Y. J., Briceno, S., and Mavris, D., "Prediction of weather-induced airline delays based on machine learning algorithms," *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, IEEE, 2016, pp. 1–6.

[5] Kim, Y. J., Choi, S., Briceno, S., and Mavris, D., "A deep learning approach to flight delay prediction," *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, IEEE, 2016, pp. 1–6.

[6] Rebollo, J. J. and Balakrishnan, H., "Characterization and prediction of air traffic delays," *Transportation research part C: Emerging technologies*, Vol. 44, 2014, pp. 231–241.

[7] Callaham, M., DeArmon, J., Cooper, A., Goodfriend, J., Moch-Mooney, D., and Solomos, G., "Assessing NAS performance: Normalizing for the effects of weather," *4th USA/Europe Air Traffic Management R&D Symposium*, 2001.

[8] Klein, A., "NAS/ATM performance indexes," *Proc. USA/Eur. Air Traffic Manage. R&D Symp*, 2007.

[9] Sridhar, B. and Chen, N. Y., "Short-term national airspace system delay prediction using weather impacted traffic index," *Journal of guidance, control, and dynamics*, Vol. 32, No. 2, 2009, pp. 657.

[10] Glina, Y., Jordan, R., and Ishutkina, M., "A Tree-Based Ensemble Method for the Prediction and Uncertainty Quantification of Aircraft Landing Times," *American Meteorological Society–10th Conference on Aritificial Intelligence Applications to Environmental Science, New Orleans, LA*, 2012.

[11] GE Aviation, "GE Flight Quest Phase 1," https://www.kaggle.com/c/flight#ge-product.

[12] Sternberg, A., Soares, J., Carvalho, D., and Ogasawara, E., "A Review on Flight Delay Prediction," *ArXiv e-prints*, March 2017.

[13] MIT Lincoln Lab, "Integrated Terminal Weather System (ITWS)," https://ll.mit.edu/mission/aviation/faawxsystems/itws.html.

[14] McNally, D., Gong, C., and Sahlman, S., "Integration of dynamic weather routes automation with air/ground data communications," *Integrated Communication, Navigation, and Surveillance Conference (ICNS), 2015*, IEEE, 2015, pp. D4–1.

[15] DeLaura, R., "Route Availability Planning Tool," Tech. rep., MIT Lincoln Laboratory, 2012.

[16] Evans, J. E. and Ducot, E. R., "Corridor integrated weather system," *Lincoln Laboratory Journal*, Vol. 16, No. 1, 2006, pp. 59.

[17] Klingle-Wilson, D. and Evans, J., "Description of the Corridor Integrated Weather System (CIWS) Weather Products," *Project Report ATC-317, MIT Lincoln Laboratory, Lexington, MA*, 2005.

[18] Cole, R. E. and Wilson, F. W., "The integrated terminal weather system terminal winds product," *The Lincoln Laboratory Journal*, Vol. 7, No. 2, 1994, pp. 475–502.

[19] MIT Lincoln Lab, "Corridor Integrated Weather System (CIWS)," https://www.ll.mit.edu/mission/aviation/faawxsystems/ciws.html.

[20] Evans, J., Carusone, K., Wolfson, M., Crowe, B., Meyer, D., and Klingle-Wilson, D., "The corridor integrated weather system (CIWS)," *10th Conference on Aviation, Range, and Aerospace Meteorology*, 2001.

[21] National Oceanic and Atmospheric Administration, "National Convective Weather Forecast Overview," https://www.aviationweather.gov/products/ncwf/info.

[22] National Oceanic and Atmospheric Administration, "Collaborative Convective Forecast Product Product Description Document," http://products.weather.gov/PDD/CCFP_PDD_Enhancement.pdf.

[23] Federal Aviation Administration, "Collaborative Convective Forecast Product (CCFP) Supports Collaborative Decision Making (CDM)," http://cdm.fly.faa.gov/?page_id=24.

[24] National Oceanic and Atmospheric Administration, "The High-Resolution Rapid Refresh (HRRR)," http://ruc.noaa.gov/hrrr/.

[25] Brian Blaylock, "HRRR Archive at the University of Utah," http://hrrr.chpc.utah.edu/.

[26] National Weather Service, "Terminal Area Forecast (TAF)," https://www.aviationweather.gov/taf/help?page=over.

[27] CFI Notebook, "Winds and Temperatures Aloft," http://www.cfinotebook.net/notebook/weather-and-atmosphere/winds-and-temperatures-aloft.

[28] Reynolds, T., Glina, Y., Troxel, S., and McPartland, M., "Wind Information Requirements for NextGen Applications Phase 1: 4D-Trajectory Based Operations (4D-TBO)," Tech. rep., FAA-ATC-399, Massachusetts Institute of Technology Lincoln Laboratory, 2013.

[29] Krozel, J., Kicinger, R., and Andrews, M., "Classification of Weather Translation Models for NextGen," Tech. rep., American Meteorological Society, 2011.

[30] Sean Breslin, "10 Most Weather-Delayed U.S. Major Airports," https://weather.com/travel/commuter-conditions/news/most-weather-delayed-airports-us-2016.

[31] United States Department of Transportation, "On-Time Performance," http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time.

[32] National Oceanic and Atmospheric Administration, "Automated Surface Observing System (ASOS) Data," https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/automated-surface-observing-system-asos.

[33] National Oceanic and Atmospheric Administration, "Next Generation Weather Radar (NEXRAD) Data," https://www.ncdc.noaa.gov/data-access/radar-data/nexrad.

[34] National Oceanic and Atmospheric Administration, "Integrated Global Radiosonde Archive (IGRA)," https://www.ncdc.noaa.gov/data-access/weather-balloon/integrated-global-radiosonde-archive.

[35] Federal Aviation Administration, "Aviation System Performance Metrics (ASPM)," https://aspm.faa.gov/, Accessed: 2016-10-28.

[36] National Oceanic and Atmospheric Administration National Climatic Data Center, "Federal Climate Complex Data Documentation for Integrated Surface Data," ftp://ftp.ncdc.noaa.gov/pub/data/noaa/ish-format-document.pdf.

[37] Klein, A., MacPhail, T., Kavoussi, S., Hickman, D., Phaneuf, M., Lee, R. S., and Simenauer, D., "NAS weather index: Quantifying impact of actual and forecast en-route and surface weather on air traffic," *89th AMS Annual Meeting*, 2009.

[38] Rich DeLaura, M., "An Exploratory Study of Modeling En Route Pilot Convective Storm Flight Deviation Behavior," *12th Conference on Aviation Range and Aerospace Meteorology*, 2006.

[39] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267–288.

[40] Breiman, L., "Random forests," *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32.

[41] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 2016, pp. 785–794.

[42] Microsoft, "LightGBM, Light Gradient Boosting Machine," https://github.com/Microsoft/LightGBM.

[43] Banko, M. and Brill, E., "Scaling to very very large corpora for natural language disambiguation," *Proceedings of the 39th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2001, pp. 26–33.

[44] Friedman, J., Hastie, T., and Tibshirani, R., *The elements of statistical learning*, Springer series in statistics New York, 2nd ed., 2016.

[45] Tu, J. V., "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of clinical epidemiology*, Vol. 49, No. 11, 1996, pp. 1225–1231.

[46] Arneson, H., Bombelli, A., Segarra-Torne, A., and Tse, E., "Analysis of convective weather impact on pre-departure routing of flights from Fort Worth Center to New York Center," 2017.