Predicting Peak Sector Occupancy with Two-Hour Convective Weather Forecasts

Shawn R. Wolfe [+]
NASA Ames Research Center, Moffett Field, CA, 94035

and

Deepak S. Kulkarni [+]
NASA Ames Research Center, Moffett Field, CA, 94035

An important function of traffic flow management ensuring the number of aircraft entering a sector does not exceed the amount that can be safely controlled by the sector controller. One factor that makes this task difficult is the uncertainty of the impact of convective weather, as both the weather forecast and the impact given specific weather is uncertain. In this investigation, we study this effect indirectly by exploring the relationship between convective weather forecasts and observed peak sector occupancy. Specifically, we measure how well the peak sector occupacy can be predicted using area-based and directional-based weather translation models. We also present a methodology for comparing weather translation models using a machine learning approach. Our results over the entire range of weather forecasts show that weather translation models produce better predictions than models without translation, but little significant difference is observed among the weather translation models. When restricted to heavy forecasted weather, however, some statistically significant differences were observed among the weather translation models.

Nomenclature

CIWS	= Corridor Integrated Weather System	*MAP*	= Monitor Alert Parameter
CWAM1	= Convective Weather Avoidance Model	*MWU*	= Mann-Whitney U
DFWCI	= Divided Forecast Weather Coverage	*RMS*	= Root Mean Squared
FWCI	= Forecast Weather Coverage Index	*WITI*	= Weather Impacted Traffic Index

I. Introduction

TRAFFIC flow management is concerned with balancing the demand for airspace resources with the capacity of the same resource. A specific example of a resource, and the domain of our study, is that of a high altitude sector. A simple model of the demand of the sector for some period of time is the number of aircraft that would occupy the sector during that time in the absence of constraints. In contrast, the analogous capacity of a sector for some period of time is the number of aircraft the sector controller can safely manage. Traffic flow management actions are not needed when demand is less than or equal to the capacity, but are needed when demand exceeds capacity (a demand/capacity imbalance). For the most part, capacity cannot be increased through traffic flow management, so instead actions are taken to reduce the demand, i.e., the number of aircraft that would occupy the sector in that period of time. It is important to choose traffic flow management mitigating actions that reduce demand by only the necessary amount: too little reduction puts undue workload on the sector controller, but too much reduction can create delays and reduce airspace efficiency.

Unfortunately, both sector demand and capacity are uncertain quantities, particularly given the fact that they must be estimated well enough in advance to enact the appropriate traffic management intiatives. Uncertainty in demand can come from uncertainty in transit times, delays elsewhere (ground or air), unscheduled traffic, and other traffic flow initiatives. Uncertainty in demand comes from the variability among the individual controllers, as well as the specifics of the traffic and current environmental factors, such as convective weather. The current operational

[+] Computer Scientist, Intelligent Systems Division, Mail Stop 269-2.

standard for modeling capacity is the monitor alert parameter (MAP)[1]; a constant number that defined separately for each sector and taken as a default estimate of sector capacity. As it is a constant, it is independent of situational factors such as controller variability and environmental conditions; traffic flow managers must alter the estimate of demand based on their experience and without the assistance of decision support tools.

Our ultimate goal is to be able to provide better models of both demand and capacity, validated on historical data. Unfortunately, this is difficult as neither demand nor capacity are directly observable. Demand could be measured as the number of occupying aircraft plus all those that were routed away from the constrained sector; however it is difficult to aggregate the latter number, and that would still miss aircraft proactively rerouted by the operators. Likewise, capacity can be measured by observing the aircraft count at the point the controller refuses to accept any new aircraft into the sector; however this is not a situation that should be intentionally created, and when it does happen, it is not necessarily easily captured or even frequent, as a goal of traffic flow management is to avoid such situations.

We use the aircraft peak count, defined as the observed instantaneous aircraft peak count in a sector over a fifteen minute period, to serve as a proxy of both capacity and demand. The observed aircraft count can be thought of as function of both demand and capacity. Assuming optimal operations, when demand is below capacity, the peak count is equal to demand, since the capacity is not a factor; when demand is above capacity, the peak count is equal to the capacity, since not all demand can be satisfied. As such, the peak count is not always an indicator of the actual capacity. We restrict our study to times when air traffic is normally reasonably high, with the expectation that this will increase the impact of capacity reductions on the observed peak count since demand will be closer to capacity under nominal operating conditions.

In this study, we attempt to build a model of how convective weather impacts capacity, ignoring any other potential impacts. We also do not try to model demand except in a rudimentary way; other research efforts addressed modeling demand (as captured by peak counts) in more detail[2]. Our main contributions are a comparison of the models, and a methodology for making such comparisons; the models themselves are largely derived from previous work. Our paper is organized as follows. In Section II we described previous efforts to model weather impacts on capacity, from which we largely draw from to build our models of weather impact capacity. In Section III we provde a formal defintion of all our models, along with two baselines. In Section IV, we describe the dataset used in our study, and describe several of its properties that expose some of the difficulties in observing the effect of weather on capacity. We present our experiment and results in Section V, and summarize our conclusions and opportunities for future work in Section VI.

II. Related Work

A simple model of weather-impacted capacity is simply the percentage of the sector (in terms of area or volume) that is free of convective weather that meets a certain threshold times the nominal capacity (often MAP). This concept has been given several names in the literature, for instance Weather Severity Index and Weather Avoidance Altitude Field coverage; we use Weather Coverage Index to refer to the volume of a sector occupied by qualifying weather. A prior simulation of delays due to weather included an analysis of historical data that showed reducing MAP by the Weather Severity Index (weather area) results in a reasonable upper bound observed peak sector counts.[3]

More complex weather-impact models measure aspects of convective weather in terms of some property meaningful from a traffic flow management perspective. In order to capture directional aspects of capacity, a scanning method was developed that runs "scan lines" along particular directions[4]. Scan lines that intersect convective weather meeting a particular criteria are seen as partially or totally unusuable; it is assumed that aircraft do not deviate from the scan line. From this, the total reduction in capacity in the direction can be estimated.

The MaxFlow/MinCut Theory has been used to model the available capacity under several air traffic control scenarios[5]. Contrasted with the scanning method, the restriction that aircraft must follow a straight line is loosened. Instead, the aircraft are modeled as entering and exiting the airspace at certain points; the width of the narrowest gap in their chosen transit determines the number of aircraft that can make the transit and thus an upper bound on the weather-impacted capacity.

The MinCut concept has been applied to a flow-based model of sector capacity as well[6]. Flows are represented as triplets of sectors – an origin sector, transit sector, and destination sector chain. The flow-based model models the capacity of the transit sector in terms of the restrictions of its flows; the capacity of the flow is the nominal peak flow reduced by the amount of constriction according to the min cut. An evaluation of this model using the largest three flows showed a better linear fit with the 95th percentile observed peak count than did area- or volume-based capacity models.

Several efforts have been made during the past few years to understand the connection between weather and delay. Of particular importance is the the Weather Impacted Traffic Index[7] (WITI). WITI captures the number of aircraft affected by weather at a given instant of time. Specifically, a grid is defined on the airspace, and the number of aircraft that fall within the same cell as severe weather are summed up to create the index. Studies[7,8] have established that an aggregate national WITI has strong correlation with national OPSNET delays.

III. Models

Our approach to estimating the weather-impacted peak sector count requires some forecast of the weather. Finer-grained forecasts potentially enable more precise predictions of weather impacts: location information could distinguish between scattered "popcorn" convective cells and a single convective mass, as well as the vertical position of the storm; differences in cell intensity may also distinguish between usable and impenetrable sections of the storm. On the other hand, excess resolution in the forecast is not particularly beneficial if it exceeds the level of accuracy the forecast. Probabilistic forecasts may ultimately lead to better predictions by presenting a range of possible weather scenarios, but introduce more complexity for the same reason, so we prefer deterministic forecasts for our initial study. Ultimately, any forecast model may be used, with a preference for more accurate and finer-grained models.

For this investigation, we have used the Convective Weather Avoidance Model[9,10] (CWAM1) for our weather forecasts. CWAM1 is itself a translation of the Corridor Integrated Weather System (CIWS), which produces two-hour forecasts (as well as shorter-term forecasts) every five minutes for the eastern corridor of the United States of America. CWAM1 is easier to use for our study than CIWS because it performs the first stage of translating weather forecasts into air traffic impacts. Of course, by selecting CWAM1, our study is influenced by whatever strengths or weaknesses it has. The goal of CWAM1 is to predict what percentage of traffic will avoid areas of convective weather. CWAM1 uses predictions of Vertical Integrated Liquid (a measure of the amount of liquid in a column of the storm) and echo tops (estimations of the height of the storm) from CIWS to create the predicted areas of avoidance. These areas of avoidance are represented as two-dimensional polygons every thousand feet in high altitude sectors with a resolution less than one nautical mile. A prior validation study[11] showed that the 80% avoidance polygons were reasonably accurate, so we use only these 80% avoidance polygons as our representation of forecasted weather.

As mentioned previously, our ultimate interest is in how weather affects the capacity of the sector, but this capacity is not easily observed. Instead, we use the observed peak count as a proxy for the sector capacity. Unlike capacity, observed peak count is also affected by the demand of the sector, so we also need to model this element as well. We use a simple model for demand; for a given sector and time of day, we use the average observed peak count over the study period when clear weather was forecasted. This demand model does not capture variation over the week or months, nor does it factor what has happened recently in this and other related sectors, so it is a somewhat crude model. In any case, the representation for weather and demand is translated into a single prediction for peak count, which we describe in Section V below.

We explored use of several translations of the CWAM1 model, largely inspired by related literature:

1) Forecast Weather Coverage Index

Our simplest translation model, the Forecast Weather Coverage Index (*FWCI*), uses only the volume of the forecasted convective activity in the sector and does not account for any other features, such as inter-sector cell location, flight patterns, etc. Specifically, it is the percentage of the volume of the sector that is forecasted to have convective weather that meets the 80% avoidance criterion. Recall that CWAM1 produces (two-dimensional) polygons for various flight levels in the sector. For a given sector s and time t, let F_s be the set of flight levels for sector s, $P_{f,t}$ be the set of CWAM1 polygons at flight level f at time t, and s to be the (two-dimensional) sector geometry. We compute *FWCI* as

$$FWCI_{s,t} = \frac{1}{|F_s|} \sum_{f \in F} \sum_{p \in P_{f,t}} \frac{|p \cap s|}{|s|} \tag{1}$$

where $|F_s|$ is the size of the set F_s (i.e., the total number of flight levels), $|p \cap s|$ is the area of the intersection of p and s, and $|s|$ is the area of sector s. Figure 1 shows a two-dimensional depiction of the FWCI calculation.

Figure 1. Example of Forecasted Weather Coverage Index.

2) Divided Forecast Weather Coverage Index

Similar in spirit to the WITI approach [7], we extend the *FWCI* model by superimposing an arbitrary grid to create the Divided Forecast Weather Coverage Index (*DFWCI*). We used 3 x 3 grid that covers the entire sector, with equal grid cell heights and width, extending from the bottom to the top of the sector (see Figure 2 for an example). This grid divides the sector into nine subsectors; though the grid cells are all equal in size, the subsectors may vary in volume, as the sector and the grid do not have the same geometry. Indeed, in some cases a subsector may be empty. *DFWCI* essentially repeats the *FWCI* on a smaller scale, resulting in nine estimates instead of one (a *FWCI* for each subsector). *DFWCI* is calculated as

$$DFWCI_{i,j,s,t} = \frac{1}{|F_s|} \sum_{f \in F_s} \sum_{p \in P_{f,t}} \frac{|p \cap s \cap c_{i,j}|}{|s \cap c_{i,j}|} \tag{2}$$

where $c_{i,j}$ is the geometry of the cell in the ith row and j^{th} column, and all other quantities are defined as in the *FWCI* calculation.

There are several motivations behind the *DFWCI* model. First, since the volume of airspace in lower in each *DFWCI* calculation (when compared to *FWCI*), it may better capture the impact on vital areas of the airspace. For instance, if the sector contains an important fix or crossing of streams of traffic, it may be easier to estimate the impact when evaluating the *DFWCI* estimate of the containing subsector than the *FWCI* estimate for the entire sector. Second, certain patterns among the cells may indicate meaningful weather structure in the sector. For instance, imagine an unbroken weather system occurring only in the middle column of subsectors. Such a pattern could indicate that the sector has very little capacity for East-West traffic; however, the *FWCI* for the entire sector would not reveal this pattern and thus give a less telling picture. Third, the variation in *DFWCI* estimates among the subsectors can provide an indication of the type of convective weather. Imagine the *DFWCI* estimates for the same sector from two different times that have an overall *FWCI* of 40%: one with a single convective cell, and one with scattered "popcorn" areas of convection. As stated, both would look the same in terms of *FWCI*, but would look very different in terms of DFWCI: the single convection case would have higher *DFWCI* estimates in some subsectors and little or zero *DFWCI* estimates in other subsectors, whereas in the popcorn case, the *DFWCI* estimates would be more even among the subsectors.

On the other hand, the *DFWCI* estimates may provide finer resolution than is necessary or advantageous. If the sector capacity is primarily a function of the weather in the sector as a whole and not sensitive to the location of weather within the sector, then the additional resolution of *DFWCI* would make it more difficult to capture the larger picture. Likewise, two-hour forecasts may not be sufficiently accurate at the level of the *DFWCI* grid, in which case the higher resolution does not supply any additional information.

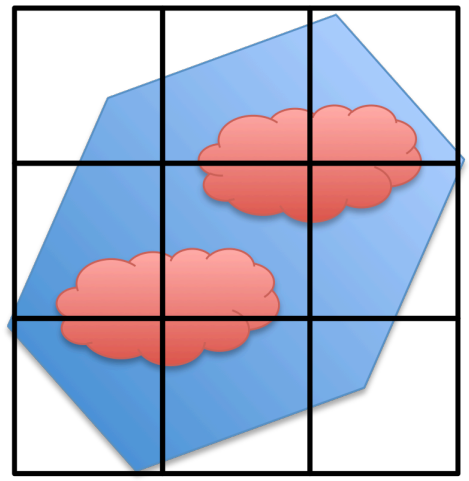

Figure 2. Example of Divided Forecast Weather Coverage Index.

3) Directional Models

Our third and final translation model is inspired by work of Klein et al.[4] and is manifested as three related but distinct models. Each variant defines the same set of parallel scan lines and uses some aspect of their intersection with the forecasted convective polygons as the relevant feature (see Figure 3). Scan lines are run at a spacing of approximately every five nautical miles in a given direction; they can also be thought of as scan planes as each line intersects every flight level. The scan lines are run in nine directions, measured as a clockwise offset from due North; 0º, 20º, 40º, 60º, 80º, 100º, 120º, 140º and 160º. (The information would be redundant if we extended it further, since the 180º lines would be the same as the 0º lines). Like *DFWCI* but unlike *FWCI*, this creates multiple features for the translation model. Intuitively, one can regard the scan lines as capturing the capacity in the given direction, with the unobstructed lines representing clear lines and the obstructed ones potentially losing some capacity. In the original translation model, the greatest weather intensity was used to determine the reduction, but this is not meaningful in our representation with only one level of forecasted weather intensity (the 80% avoidance region).

For each direction d, each scanning variant defines a set of set of scan lines L_d, in addition to flight levels F_s and CWAM1 polygons $P_{f,t}$ at flight level f at time t, as before. The definitions follow:

1. Countscan model.

The *countscan* model captures the number of scan lines that intersect some forecasted convective activity. For a given direction d, it is calculated as

$$countscan_{d,s,t} = \frac{1}{|F_s||L_d|} \sum_{f \in F_s} \sum_{p \in P_{f,t}} \sum_{l \in L_d} I_{<0}\left(|p \cap l|\right) \tag{3}$$

where $I_{<0}(\bullet)$ is an indicator function that returns 1 when its argument is greater than zero, 0 otherwise. The idea behind the *countscan* model is that flights approximately travel along the scan lines; when convective weather intersects a scan line, that path is presumed to be unusable. However, the *countscan* model assumes flights follow a straight path across the sector, which is not always the case.

2. Maxscan model.

The *maxscan* model captures the highest percentage of a scan line intersecting any forecasted convective activity. For a given direction d, it is calculated as

$$maxscan_{d,s,t} = \frac{1}{|F_s|} \sum_{f \in F_s} \sum_{p \in P_f} \max_{l \in L_d} \left(\frac{|p \cap l|}{|l|} \right). \tag{4}$$

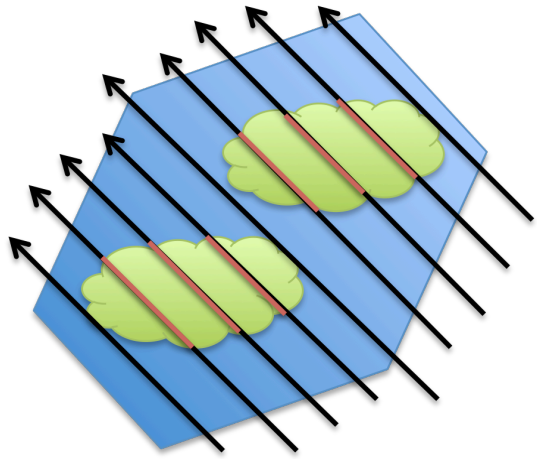

Figure 3. Example of Scanning Approach.

The *maxscan* model is not meant to capture the directional capacity in the given direction *d*, but rather the perpendicular capacity, i.e., the directional capacity in direction *d*+90º. In this conception, the scan lines are perpendicular to the flow and when they intersect convective weather, chokeholds are created. Thus, the maximum constriction captures the smallest chokehold the traffic must flow through, and is presumed to the limiting factor. This is our approximation of the MinCut concept[5]. However, the *maxscan* model assumes that flights travel along the full length of the sector (i.e., no clipping), which is not always the case.

3. Totalscan model.

The *totalscan* model captures the average scan line intersection with forecasted convective activity. For a given direction *d*, it is calculated as

$$totalscan_{d,s,t} = \frac{1}{|F_s||L_d|} \sum_{f \in F_s} \sum_{p \in P_f} \sum_{l \in L_d} \frac{|p \cap l|}{|l|}. \tag{5}$$

The idea behind the *totalscan* model is similar to the *countscan* model, but does not use a binary model of permeability. Instead, the percentage of the scan line intersecting with forecasted convective activity is used as a gradual indicator of loss of capacity. Instead of a percentage, the raw intersection length could be used, but this just produces an approximation of the WCI model. Like the *countscan* model, the *maxscan* model assumes flights follow a straight path across the sector, which is not always the case.

4. Mean models.

Finally, we also use models that average the directional capacity estimates to produce a single estimate. We refer to these translation models as *countscanmean*, *maxscanmean*, and *totalscanmean*.

4) Non-translation models

We define two additional models that have no explicit weather component to act as baselines for our translation models. This allows us to put the performance of the weather translation models into a broader context, which is particularly important as the peak count is also affected by demand; we need to isolate the effect of the weather translation model. The first model, *wxmean*, includes only the average peak count over all times for that sector in the dataset, which are all peak times (i.e., 13:00-23:00 EDT) when weather was forecasted two hours in advance. *Wxmean* for a sector *s* at time *t* is simply defined as

$$wxmean_{s,t} = \frac{1}{|T \cap W_s|} \sum_{x \in T \cap W_s} peak_{s,x}. \tag{6}$$

where T is the set of peak times, W_s is the set of times in our dataset[1] where some convective weather was forecast for sector s, and $peak_{s,t}$ is the recorded peak aircraft count for sector s and time x. Note that *wxmean* is not dependent on the specified time t. Ideally, a separate average would be computed for each time of day (e.g., for all 14:15 EDT time periods where weather was forecast), but we do not have enough data to compute means that are reliable estimates for the expected mean at that time. The *wxmean* model does allow for an impact from the presence of weather, but assumes that neither the properties of the weather (e.g., size, location) nor the time of day impact the peak count.

Our second baseline model, *clearmean*, captures only the demand component based on observations when no weather was forecast. For a given sector s and time t, we calculate *clearmean* as follows:

$$clearmean_{s,t} = \frac{1}{|H_t \setminus W_s|} \sum_{x \in H_t \setminus W_s} peak_{s,x} . \tag{7}$$

where H_t are all times at the same time of day as t (e.g., if t is 14:45 June 17 EDT, then H_t is all times at 14:45 EDT), W_s is defined as before, and \ denotes set difference. The *clearmean* model assumes that there are no weather related impacts on observed peak count, but time of day does matter.

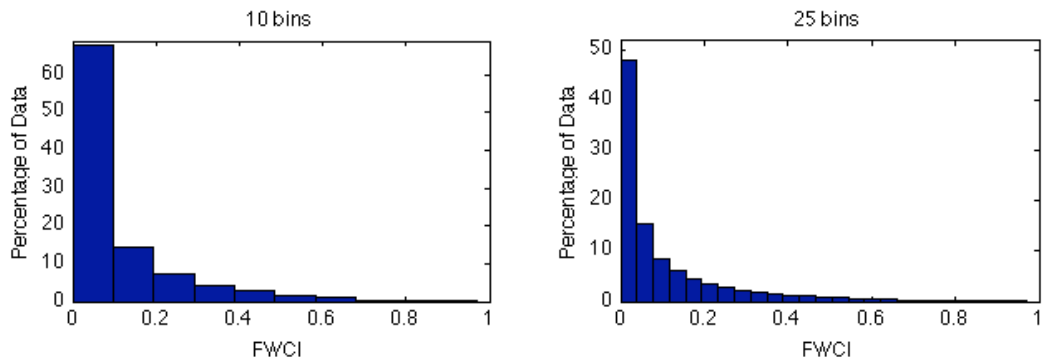

Figure 4. Distribution of weather severity as measured by *FWCI*.

IV. Data

Our dataset covers the 122 days from June 1, 2007 to September 31, 2007 for the high altitude sectors in ZOB (Cleveland Air Route Traffic Control Center), ZID (Indianapolis Air Route Traffic Control Center), and ZDC (Washington, D.C. Air Route Traffic Control Center). Convective activity was forecast for 17% of the peak times over this period. However, this varied by sector, from a low of 8% to a high of 37%, with a standard deviation of 6%.

Figure 4 shows the distribution of forecast weather, as measured by *FWCI*, across the entire dataset. Weather severity, when measured in this way, appears to follow an exponential distribution, with light weather occuring far more frequently than heavy weather. Indeed, nearly 70% of our forecast weather situations predicted less than 10% coverage of the volume of the sector, according to our CWAM1 80% avoidance level. As a result, it will be more difficult to characterize the effects of heavy weather than light weather. Furthermore, if light weather has relatively little impact on capacity, then we would expect to see little impact from weather overall as light weather dominates our dataset.

Figure 5 gives the distribution of observed peak count in the sectors over our study period of 13:00-23:00 EDT when clear weather was forecast. The distribution appears to be approximately normal, which leads us to the conclusion that peak count varies considerably, even when no weather is forecast for the sector. Indeed, it would appear that the normal situation in our dataset is 50% of the available capacity is being used for a given sector. Coupled with our earlier observation that most weather forecast is relatively light, it may be the weather impact on capacity may not be observable through peak count in most situations: if the weather tends to be light, and there tends to be significant unused capacity, then it may be that the resultant drop in capacity is not great enough to create a demand/capacity imbalance and hence unobservable through an analysis of peak aircraft counts.

[1] Here we use only data from the testing subset, as described in Section V.

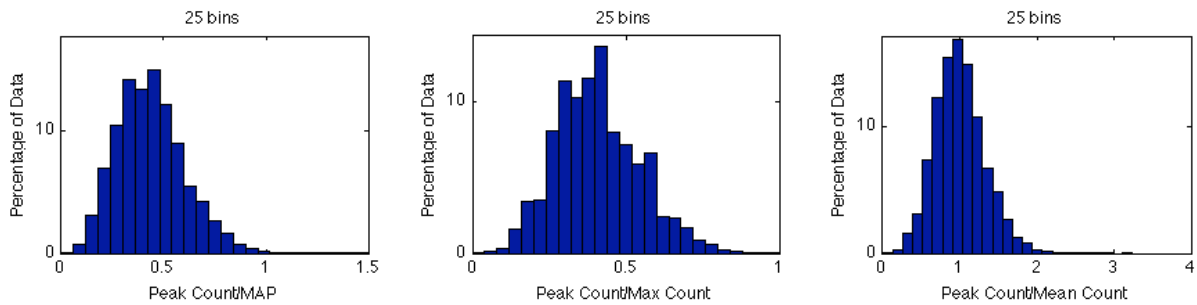

Figure 5. Distribution of Peak Count.

Overall, our analysis of the basic characteristics of the dataset reveals that relatively light forecasted weather and moderate amounts of traffic dominate the forecasted weather subset of the data. Observed peak counts, as measured in clear weather, showed noticeable variation within a given time of day and across the 13:00-23:00 time period, as shown in Figure 6. These results appeared to be fairly applicable to all sectors.

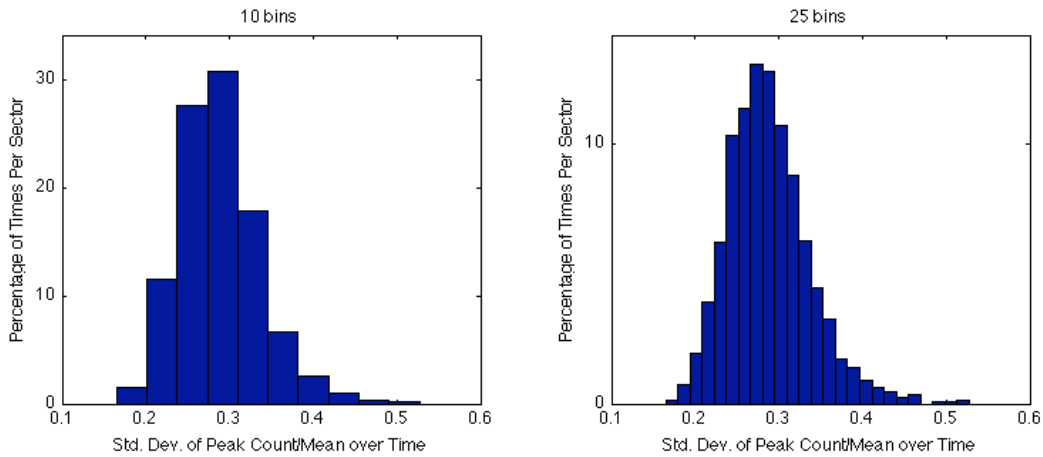

Figure 6. Distribution of Standard Deviation of Observed Peak Count/Mean over all sectors and times

V. Experimental Results

Rather than define *a priori* a model of capacity, or report the best fitting parameters retrospectively, we adopt a machine learning approach to build and evaluate our models based on the data. Model construction and evaluation occur in two separate phases on disjoint training and test subsets of the dataset[2]. This matches the likely use in practice, as models are created based on historical data and applied to incoming data that were not available at the time of model construction. Many machine learning algorithms exist; for this study, we used straightforward linear regression as implemented in the Weka[12] data mining software package. We use ten-fold cross-validation to decrease the chance of spurious results and make better using of the available data for training. Specifically, the dataset is split into ten approximately equal folds. For each fold, the remaining nine folds (approximately 90% of the data) is used as training data from which to infer the model, and the fold itself is used as testing data to evaluate the model. This means there are ten separate trials, each evaluated on separate (disjoint) subsets of the dataset, and all of the dataset is used for evaluation in some trial. Our results presented below are the aggregate results for the entire dataset over these ten trials.

We make two assumptions that lead us to certain choices in the representation and experimental setup. The first assumption is that despite the overall variability, there are more and less busy times of day during our 13:00-23:00 EDT study period. Therefore, it seems logical to include some representation of this time-varying demand component to enhance our weather translation models (but not the *wxmean* baseline). In our data representation, we combine the *clearmean* model with the translation models described in Section III.

[2] The *clearmean* model deviates somewhat from this as the training set is always the forecasted clear weather cases, and the peak count prediction is given by Eq. 7.

Our second assumption is that every sector may have different qualities, which would justify a different model for each one. For instance, different sectors have different geometries, which would affect the *DFWCI* representation. Also, different sectors may be dominated by different directional flows, or no particular flow, which would affect the directional models. Therefore, we perform linear regression on each sector separately. The downside of this decision is that there is less data available for each regression (since we are doing several disjoint regressions on the same data instead of one). The results we present are an aggregation of all the predictions from these separate models on their corresponding testing datasets.

Our experiment was performed only on sector/time pairs from which some convective weather was forecast. For each model, a prediction of the peak count was made, and the difference between the observed peak count and the predicted peak count was recorded as the resulting error. We calculated the following statistics from these set of errors:

- **Error Bias**: The overall error mean, indicating if the method is generally under- or over-estimating peak count.
- **Median Absolute Error**: The median absolute error. Roughly half of the errors are lower than the median, and roughly half are above.
- **Mean Absolute Error**
- **Root Mean Squared Error:** Linear regression minimizes the root mean squared (RMS) error on the training set. Since the error is squared, large errors result in a disproportionately larger score.

The median absolute, mean absolute and root mean squared errors are similar but not exactly the same. In some of our cases, a model will have better results in one error statistic and worse in another when compared to a different model.

In addition to these means, we also evaluate the statistical significance of the results by comparing the absolute errors of two models with a one-tailed version of the Mann-Whitney U Test[13] (MWU test, also called the Mann-Whitney-Wilcoxen test). The (usual) two-tailed MWU test compares two samples to determine if the evidence suggests the medians of the respective generating distributions are different; we use only one tail to evaluate if one (distributional) median is higher than the other. Though this is related to the absolute median statistic, it is not the same as comparing only the two observed medians, as the MWU test compares *all* points from the samples and not just the corresponding medians. As a result, occasionally the MWU test will assign a significance level that is in opposition to the observed medians. We use the MWU test to evaluate whether the distribution of errors from model *A* (listed in the first column) has a lower absolute error median than model *B* (listed in the first row). The cited level of significance indicates the fraction of times this statement would be false under identical conditions (rounded up); thus lower numbers indicate a stronger level of significance. 0.05 is commonly used as the cutoff of acceptance. Numbers above 0.5 represent that the statement is false under most identical conditions; in other words, most of the evidence contradicts the statement.

Table 1. Results over entire forecasted weather dataset.

	Bias	Median Abs. Error	Mean Abs. Error	RMS Error
Wxmean	0.000	1.531	1.865	2.379
Clearmean	-0.002	1.500	1.843	2.368
FWCI	0.000	1.418	1.723	2.211
DFWCI	-0.001	1.404	1.713	2.197
Countscan	0.000	1.416	1.729	2.217
Countscanmean	0.000	1.420	1.727	2.215
Maxscan	0.000	1.420	1.733	2.222
Maxscanmean	-0.001	1.424	1.736	2.225
Totalscan	0.000	1.411	1.712	2.194
Totalscanmean	0.000	1.414	1.720	2.206

Table 1 shows the results over all 28996 sector/time pairs that had at least some convective weather forecasted. There is almost no bias in the errors, indicating that there is no trend to consistently under- or over-predict peak count by a substantial amount among the various models. This is somewhat surprising, because this set includes the *clearmean* model, which predicts as if there is no forecasted weather, and still has only a very slight overestimate of

peak count. Within any given model, the mean absolute error is higher than the median absolute error, and the root mean squared error is higher still. This indicates that within the set of errors, less frequent but rather large errors occurred; in other words, each method made very inaccurate predictions at some point. Comparing among the models, *clearmean* has lower error statistics than *wxmean*; it also has significance under 0.05 when compared to *wxmean*, as shown in Table 2. This indicates that the time of day is an important factor in predicting the peak count, as *wxmean* has no time of day component. All the weather translation models have lower error statistics and a significant difference over both *wxmean* and *clearmean*. However, among the weather translation models, none have a significant difference at or below the 0.05 level with *FWCI*, though several have lower error statistics, and in fact only *totalscan* less than *maxscanmean* meets the 0.05 level of significance. None of the mean scanning models outperformed their non-mean scanning counterparts, though the differences were slight. Of all the models, *DFWCI* and *totalscan* had the lowest error statistics.

Table 2. Statistical significance over entire forecasted weather dataset.

	Wx-mean	*Clear-mean*	*FWCI*	*DFWCI*	*Count-scan*	*Count-scan-mean*	*Max-scan*	*Max-scan-mean*	*Total-scan*	*Total-scan-mean*
Wxmean	—	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Clearmean	**0.01**	—	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FWCI	**0.01**	**0.01**	—	0.72	0.31	0.37	0.20	0.15	0.75	0.55
DFWCI	**0.01**	**0.01**	0.29	—	0.14	0.19	0.08	0.06	0.54	0.33
Countscan	**0.01**	**0.01**	0.70	0.87	—	0.58	0.38	0.30	0.89	0.74
Countscanmean	**0.01**	**0.01**	0.64	0.82	0.43	—	0.31	0.24	0.85	0.68
Maxscan	**0.01**	**0.01**	0.81	0.93	0.63	0.70	—	0.43	0.94	0.84
Maxscanmean	**0.01**	**0.01**	0.86	0.95	0.71	0.77	0.58	—	0.96	0.88
Totalscan	**0.01**	**0.01**	0.26	0.47	0.12	0.16	0.07	**0.05**	—	0.30
Totalscanmean	**0.01**	**0.01**	0.46	0.68	0.27	0.33	0.17	0.13	0.71	—

As shown previously in Figure 4, the majority of forecasted weather in our dataset is relatively light. Assuming a linear relationship between weather severity and impact on capacity, it follows that even a perfect weather translation would have little effect on our error statistics. Therefore, we performed an analysis over the subset where *FWCI* was 0.5 or above (weather forecasted for at least 50% of the volume of the sector), with the hope that the effect of weather might be more observable. However, this comes at a price, as the size of this subset is only 848 sector/time of day pairs – less than 1/30th of the original dataset. This small size makes it harder to establish statistical significance and to have confidence in our mean statistics.

Table 3. Results over subset where *FWCI* ≥0.5.

	Bias	Median Abs. Error	Mean Abs. Error	RMS Error
Wxmean	-1.715	2.145	2.291	2.792
Clearmean	-1.741	2.100	2.337	2.865
FWCI	0.095	1.389	1.697	2.151
DFWCI	-0.011	1.383	1.651	2.096
Countscan	-0.545	1.484	1.768	2.217
Countscanmean	-0.591	1.521	1.790	2.245
Maxscan	-0.725	1.528	1.832	2.297
Maxscanmean	-0.857	1.576	1.859	2.329
Totalscan	0.055	1.352	1.624	2.077
Totalscanmean	0.098	1.359	1.683	2.134

Table 3 shows the results on this restricted dataset. Error biases are larger as the models were fit to the entire training dataset, not just *FWCI* ≥ 0.5, and apparently are not as well adjusted for this subset. *Wxmean* and *clearmean* have the largest magnitude of bias as they do not model the extent of the weather. Indeed, the *clearmean* model has lost its apparent edge over *wxmean*, as *clearmean* does not make any allowance for the impact of weather, which is larger in this subset. As with the entire dataset, the weather translation models have superior error statistics and a significant difference (as shown in Table 4) over both *wxmean* and *clearmean*. Actually, the error statistics are lower

for the weather translation models overall. This might seem surprising, but it is because the peak counts have generally decreased in this subset, and so the magnitude of variance has decreased as well. The comparative results are primarily the same, but in some cases stronger. The statistical significance between *FWCI* and the *maxscan* models is below the 0.05 level, but in favor of *FWCI* – indicating the deficiency of *maxscan* and *maxscanmean*. Likewise, *countscan* and *countscanmean* compare unfavorably to *FWCI*, though the 0.05 level of significance is not reached. Indeed, both *DFWCI*, *totalscan*, and *totalscanmean* meet the 0.05 level of significance against *countscan*, *countscanmean*, *maxscan*, and *maxscanmean*, calling their worth into question. Finally, *DFWCI* and *totalscan* again have the best statstics, but more in the favor of *totalscan*.

Table 4. Statistical significance over subset where *FWCI* ≥0.5.

	Wx-mean	*Clear-mean*	*FWCI*	*DFWCI*	*Count-scan*	*Count-scan-mean*	*Max-scan*	*Max-scan-mean*	*Total-scan*	*Total-scan-mean*
Wxmean	–	0.37	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Clearmean	0.64	–	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FWCI	**0.01**	**0.01**	–	0.75	0.13	0.09	**0.03**	**0.02**	0.90	0.58
DFWCI	**0.01**	**0.01**	0.26	–	**0.04**	**0.03**	**0.01**	**0.01**	0.71	0.33
Countscan	**0.01**	**0.01**	0.88	0.97	–	0.40	0.23	0.14	0.99	0.91
Countscanmean	**0.01**	**0.01**	0.92	0.98	0.61	–	0.30	0.20	1.00	0.94
Maxscan	**0.01**	**0.01**	0.98	1.00	0.78	0.71	–	0.37	1.00	0.99
Maxscanmean	**0.01**	**0.01**	0.99	1.00	0.87	0.81	0.64	–	1.00	1.00
Totalscan	**0.01**	**0.01**	0.11	0.30	**0.02**	**0.01**	**0.01**	**0.01**	–	0.15
Totalscanmean	**0.01**	**0.01**	0.43	0.68	0.10	0.07	**0.02**	**0.01**	0.86	–

VI. Conclusions and Future Work

In this paper, we evaluated several weather translation models in conjunction with a particular two-hour forecast product to predict the impact on sector capacity. However, capacity is not directly observable, so we used the observed peak aircraft count in the sector over a fifteen minute period. Other factors, such as demand, also affect the observed peak aircraft count, making analysis more difficult. To compenate, we incorporated a simple demand model into our representation based on observations when clear weather was forecast. We presented a methodology for comparing different weather translation models, by evaluating them on the same dataset, using a machine learning approach to infer the models, and using a statistical test to establish the significance of the results. Our use of the machine learning paradigm, when compared to simply finding the best linear fit, is needed to decrease the possibility of spurious results given our methods have a different number of features.

Our results show a statistically significant difference in the prediction of peak observed aircraft count of all of our weather translation models when compared to the simple demand model without weather translation, indicating that some impact of weather on capacity was captured in our weather translation models. In terms of statistical significance, little difference was observed in our weather translation models over the total dataset, though some variants were shown to be inferior when restricted to heavier weather conditions. Indeed, the simplest of the translation models (which measures only the percentage of the sector's volume with forecasted weather) performed reasonably well when compared with all other translation models. In terms of absolute numbers, the translation methods that preserved more detail outperformed those that aggregated separate features into an average, suggesting that it may be preferable to maintain multiple weather features.

More study is needed to differentiate between the various translation models. A better estimation of demand would eliminate some of the variability in the observed peak aircraft counts, making the differences in translation models easier to detect. Of all the translation models we used, *DFWCI*, which measures the forecasted weather volume in various subspaces of the sector, and *totalscan*, which measured the percentage of weather blockage in particular directions, showed the most promise. *DFWCI* could potentially be improved by choosing different divisions of the sector; the current subdivisions were completely arbitrary. *Totalscan* might be improved by weighting different segments within a direction unequally. However, all translation models in our study were dependent on the properties of the forecast product we used. Other forecast products, as well as shorter term forecasts or nowcasts, might reveal different properties of the translation models.

Acknowledgements

The authors would like to thank Shon Grabbe, Yao Wang, John Love, and Carl Russell for their contributions to the approach and ideas presented in this work.

References

[1]Federal Aviation Administration, "Facility Operation and Administration", edited by Services, System Operations (2008), Vol. Order JO 7210.3V.

[2]Chen, N. Y. and Sridhar, B., "Weather-Weighted Periodic Auto Regressive Models for Sector Demand Prediction", in *AIAA Guidance, Navigation, and Control Conference* (Chicago, Illinois, 2009).

[3]Krozel, J. and Doble, N. A., "Simulation of the National Airspace System in Inclement Weather", in *AIAA Modeling and Simulation Technologies Conference and Exhibit* (Hilton Head, SC, 2007).

[4]Klein, A., Cook, L., and Wood, B., "Airspace Availability Estimates for Traffic Flow Management Using the Scanning Method", in *27th Digital Avionics Systems Conference (DASC)* (St. Paul, MN, 2008).

[5]Krozel, J., Mitchell, J. S. B., Polishchuk, V. et al., "Capacity Estimation for Airspaces with Convective Weather Constraints ", in *AIAA Guidance, Navigation, and Control Conference* (Hilton Head, SC, 2007).

[6]Song, L., Wanke, C., Greenbaum, D. et al., "Methodologies for Estimating the Impact of Severe Weather on Airspace Capacity", in *8th AIAA 2008 ATIO Conference* (Anchorage, AK, 2008).

[7]Callaham, M. B., DeArmon, J. S., Cooper, A. M. et al., "Assessing NAS Performance: Normalizing for the Effects of Weather ", in *the 3rd USA/Europe Air Traffic Management Research and Development Symposium* (Sante Fe, NM, 2001).

[8]Chatterji, G. B. and Sridhar, B., "National Airspace System Delay Estimation Using Weather Weighted Traffic Counts", in *AIAA Guidance, Navigation and Control Conference* (San Francisco, CA, 2004).

[9]DeLaura, R. and Evans, J., "An Exploratory Study of Modeling Enroute Pilot Convective Storm Flight Deviation Behavior", in *12th Conference on Aviation, Range and Aerospace Meteorology* (Atlanta, GA, 2006).

[10]DeLaura, R., Robinson, M., Pawlak, M. et al., "Modeling Convective Weather Avoidance in Enroute Airspace ", in *13th Conference on Aviation, Range and Aerospace Meteorology* (New Orleans, LA, 2008).

[11]Chan, W. N., Refai, M., and DeLaura, R., "An Approach to Verify a Model for Translating Convective Weather Information to Air Traffic Management Impact", in *7th AIAA Aviation Technology, Integration and Operations Conference (ATIO)* (Belfast, Ireland, 2007).

[12]Witten, I. H. and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.

[13]Hogg, R. V., Craig, A., and McKean, J. W., Introduction to Mathematical Statistics, Sixth ed., Prentice Hall, 2004.