

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 is 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 occupancy can be predicted using area-based and directional-based weather models. We also present a methodology for comparing weather models using a machine learning approach. When the forecast is for light weather, the impact of weather is presumably minimal and little difference is observed between models in our evaluation. In contrast, when heavy weather is forecast, the weather models outperform those without a weather component and also have statistically significant differences among each other.

Nomenclature

<i>CIWS</i>	= Corridor Integrated Weather System	<i>FWCI</i>	= Forecast Weather Coverage Index
<i>CWAMI</i>	= Convective Weather Avoidance Model	<i>MAP</i>	= Monitor Alert Parameter
<i>DFWCI</i>	= Divided Forecast Weather Coverage Index	<i>WITI</i>	= 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 initiatives. 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 standard for modeling capacity is the monitor alert parameter (MAP)¹; a constant number that defined separately for

⁺ Computer Scientist, Intelligent Systems Division, Mail Stop 269-2, Non-member.

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 frequent enough to easily support analysis, as a goal of traffic flow management is to avoid such situations.

We use the *aircraft peak count*, defined as the observed highest instantaneous count of aircraft 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, this 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. In particular, we do not try to represent demand except in a rudimentary way; other research efforts have addressed predicting demand (as captured by peak counts) in more detail². Our main contributions are a comparison of the weather models, and a methodology for making such comparisons; the weather 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 provide a formal definition of all our models. 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. In Section V we describe our machine learning methodology and data representation, along with two baselines. We present our experiment and results in Section VI, and summarize our conclusions and opportunities for future work in Section VII.

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 percentage of 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 directions⁴. Scan lines that intersect convective weather meeting a particular criteria are seen as partially or totally unusable; 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 scenarios⁵. 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 well⁶. 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 Weather Impacted Traffic Index⁷ (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.

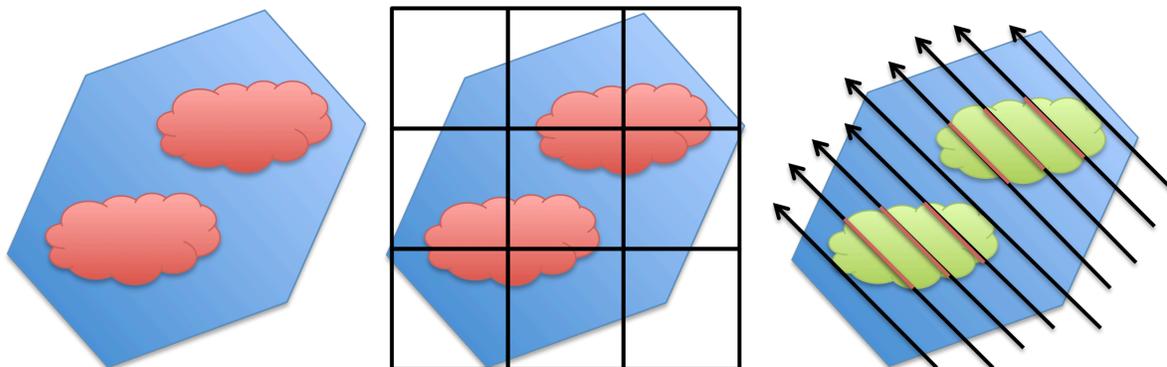


Figure 1. Example of Weather Models: *FWCI*, *DFWCI*, and scanning approaches

III. Weather 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 probably 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 study¹¹ showed that the 80% avoidance polygons were reasonably accurate, so we use only these 80% avoidance polygons as our representation of forecasted weather.

We explored use of several models derived from the CWAM1 representation, largely inspired by related literature:

A. Forecast Weather Coverage Index

Our simplest weather 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 intra-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

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

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 . Fig. 1 shows a two-dimensional depiction of the FWCI calculation in the leftmost picture.

B. Divided Forecast Weather Coverage Index

Similar in spirit to the WITI approach⁷, we extend the $FWCI$ model by superimposing an arbitrary grid to create the Divided Forecast Weather Coverage Index ($DFWCI$). We used 3×3 grid that covers the entire sector, with equal grid cell heights and width, extending from the bottom to the top of the sector (see Fig. 1 for an example, middle picture). 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

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

where $c_{i,j}$ is the geometry of the cell in the i^{th} 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 is 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.

C. Directional Models

Our third and final weather 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 Fig. 1, rightmost picture). 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; 0° , 20° , 40° , 60° , 80° , 100° , 120° , 140° and 160° . (The information would be redundant if we extended it further, since the 180° lines would be the same as the 0° lines). Like $DFWCI$ but unlike $FWCI$, this creates multiple features for the weather 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 Klein et al.’s original weather model, the greatest weather intensity encountered on the scan line 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

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

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

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

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 $d+90^\circ$. 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 be the limiting factor. This is our approximation of the MinCut concept⁵. 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

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

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 weather models as *countscanmean*, *maxscanmean*, and *totalscanmean*.

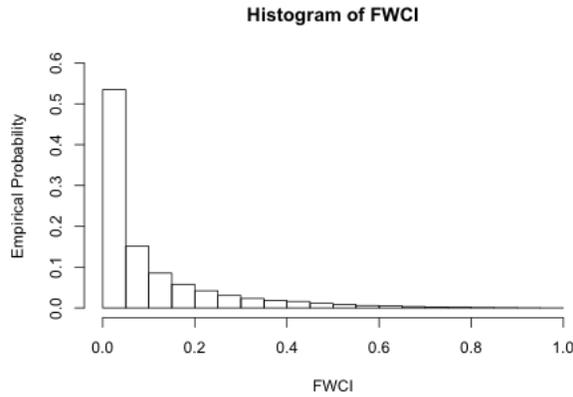


Figure 2. Distribution of weather severity as measured by *FWCI*.

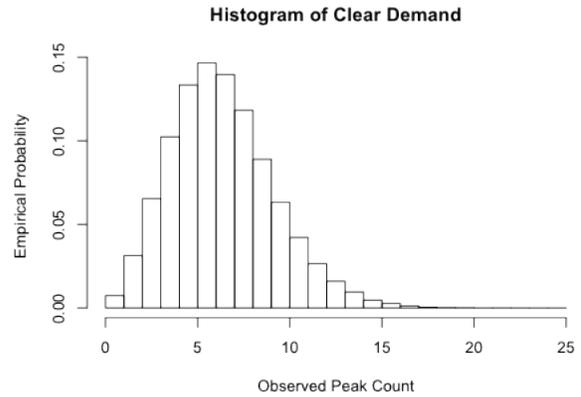


Figure 3. Distribution of peak count during forecasted clear weather.

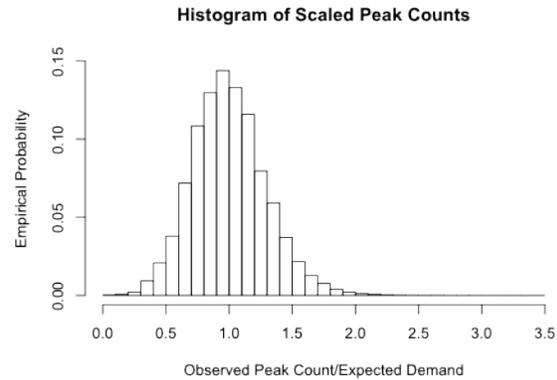
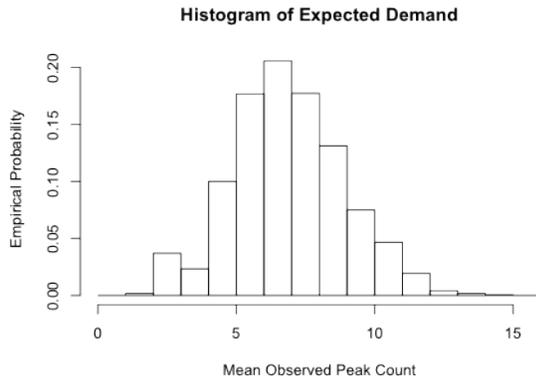


Figure 4. Distribution of expected demand and scaled peak counts during forecasted clear weather.

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), for 44 sectors in all. Since the observable effect of weather on peak counts is likely to be negligible when the traffic is light, we restricted our study to high demand periods, which we defined as 13:00-23:00 EDT. Convective activity was forecast for 17% of the fifteen-minute segments over this period. However, this varied by sector, from a low of 8% to a high of 37%, with a standard deviation of 6%.

Fig. 2 shows the distribution of forecast weather, as measured by *FWCI*, across the entire dataset. Weather severity, when measured in this way, appears to approximately follow an exponential distribution, with light weather occurring 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 capture the effects of heavy weather than light weather as there is less data available. 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.

Even though we restrict our study only to times when convective weather was forecast, an examination of peak counts when no weather is forecast can characterize how demand typically fluctuates. Fig. 3 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, and peak count varies considerably, even when no weather is forecast for the sector. Indeed, it would appear that more often than not, less than 50% of the available capacity for a given sector is being utilized in our dataset. Coupled with our earlier observation that most weather forecast is relatively light (see Fig. 2), it may be that 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.

Fig. 4 shows that a range of traffic is expected for a given sector and time during our study period. Apparently, though the 13:00-23:00 EDT time period we used eliminates many of the less busy times, traffic can be expected to be light even without weather for several sectors at particular times. It is likely that it will be more difficult to observe the effect of convective weather in such instances, as little of the available capacity is used anyway. However, as Fig. 4 also shows, this variability in expected demand does not fully explain the overall variation observed in Fig. 3. Indeed, when we rescale *a priori* the peak counts by dividing them by the expected demand, we see that the overall distribution shape is similar to Fig. 3, with only a little reduction in variance. We conclude that there is significant variation in demand in our dataset, and our representation of demand as dependent only on sector and time of day does not substantially reduce uncertainty.

V. Methodology and Data Representation

We use a machine learning methodology to build and evaluate our models, in lieu of defining them *a priori* or reporting the best fitting parameters over the complete dataset. An *a priori* approach is independent of the data and only as good as the researchers' intuition; given our lack of strong beliefs it is not an appropriate choice for this study. On the other hand, reporting the best fitting parameters for a selected model can be overly influenced by random patterns in the dataset. This results in a condition known as *overfitting*, where the resulting model fits the data used to build the model better than new data. The machine learning methodology strikes a balance between the extremes, by dividing the dataset into a *training* set, which is used to choose model parameters, and a *testing* set, which is used to evaluate the fit of the model. Though this alone does not prevent overfitting (as the selected parameters may still fit the training data better than the testing data), it does prevent overfitting from skewing the evaluation of the model. It also provides a more realistic expectation of how the models would perform in deployment, as they would be evaluating current conditions rather than exactly the 2007 data from our dataset. Overfitting is somewhat more of a concern in our study than it might be otherwise, because the different models do not all have the same number of independent variables (for instance, *FWCI* has one variable to characterize the forecast, whereas *DFWCI* has nine). Models will tend to produce as good or better fits on the data as more independent variables are added, even if the new variables have little or no explanatory power (because of chance correlations on limited data). However, by using the train and test framework, this spurious advantage is eliminated and all models are put on equal footing.

Many machine learning algorithms exist, each with their own assumptions about the data and the form of the model to be induced. We used linear regression in this study as it is easily understood and though we do not know that the true relationship is linear, it is presumably monotonic. We used the implementation of linear regression in the Weka¹² data mining software package, which by default uses Tikhonov regularization and attribute selection (as strategies to reduce overfitting), though these options had negligible effect in our study. Cross-validation with ten folds was used to provide a more accurate evaluation of performance than with a single testing and training set. In ten-fold cross validation, the dataset is randomly partitioned into ten disjoint subsets. Each subset is set aside once to be used as the testing data, with the remaining nine folds (approximately 90% of the data) used as training data from which to infer 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 at some point. We provide error statistics over this entire set as testing data; the fit of the model on the training (i.e., model-building) data is *not* reported.

Our study uses the observed fifteen-minute peak aircraft count as a proxy for the sector capacity, but the peak count is also affected by other factors, in particular sector demand. To capture this effect, we incorporate the expected demand into our models 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 (see Section IV and Fig. 4 for a characterization). This expected demand is calculated as

$$D_{s,t} = \frac{1}{|H_t \setminus W_s|} \sum_{x \in H_t \setminus W_s} peak_{s,x}, \quad (6)$$

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 the set of times in our dataset where some convective weather was forecast for sector s , $peak_{s,t}$ is the recorded peak aircraft count for sector s at time x , and \setminus denotes set difference. This demand representation does not capture variation over the week or months, nor does it factor what has happened earlier in this and other related sectors, so it is a somewhat crude model. We incorporate this into the model in two ways. The first way is that we rescaled the dependent variable (peak count) as follows:

$$Y_{s,t} = \frac{peak_{s,t}}{D_{s,t}}, \quad (7)$$

effectively making it a proportion of expected demand (as defined in Eq. 1) rather than aircraft count. The second way is that we included the expected demand (as defined in Eq. 1) as an additional independent variable. This makes our complete data representation for each weather model as 2-10 independent variables (unscaled expected demand, plus 1-9 forecasted variables, as described in Sec. III) and the observed peak scaled by expected demand (as defined in Eq. 2) as the sole dependent variable.

In addition, we define two models that have no explicit weather component to act as baselines. This allows us to put the performance of the weather models into a broader context, which is particularly important as the peak count is also affected by demand. The first model, *wxmean*, has no independent variables. As such, it can estimate the overall impact of weather, but cannot differentiate between forecasted weather cases nor the time of day (though this is indirectly included through the rescaling of observed peak count). The resulting prediction is simply the mean of the dependent variables, defined as

$$wxmean_{s,t} = \frac{1}{|T_s|} \sum_{x \in T_s} peak_{s,x}. \quad (8)$$

where T is the set of peak times in the testing set (see Sec. V) for this sector, $peak_{s,t}$ is defined as in Eq. 1. Note that *wxmean* is independent of the specified time t . If the forecast of convective weather (vs. a forecast of clear weather) does provide useful information regarding observed peak counts, but our particular models are not capturing useful characteristics, the *wxmean* model should perform as well or better.

Our second baseline model, *clearmean*, is based on the assumption that forecasted weather cases are not different than the clear forecasted weather, i.e., the weather forecast has no relation to observed peak counts. None of the data with forecasted weather (which is our entire training/test dataset) is used to build the *clearmean* model. Instead, *clearmean* predicts the expected demand for that sector and time of day based on observations from *clear* forecasted weather. As we normalized the target variable by this expectation, this is constant model: specifically, for a given sector s and time t , the *clearmean* is:

$$clearmean_{s,t} = 1. \quad (9)$$

If the forecast does not provide any useful information regarding observed peak counts, *clearmean* should do as well or potentially better than the other models, as its mean estimate is created with more data.

Finally, we assume 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). Also, due to space limitations we aggregate the results over the various sectors, rather than reporting results for all 44 sectors separately, which we chose to do through a weighted average. Two common weighted averaging approaches are the *microaverage* and *macroaverage*. Given a collection of sets C (with m sets total), and a function $f()$ defined on those sets, we define the microaverage as:

$$average_{micro} = \frac{1}{\sum_{j=1}^m |C_j|} \sum_{j=1}^m [|C_j| \times f(C_j)] \quad (10)$$

The microaverage weighs each function result by the size of the set from which it was computed, so results from larger sets are weighted higher. When the sets consist of numbers and the function $f()$ is the average of those numbers, as is the case for our error statistics in the following section, the microaverage simply produces the average over the combined set without respect to the original set membership.

In contrast, the macroaverage weighs each function result equally, as follows:

$$average_{macro} = \frac{1}{\sum_{j=1}^m (1)} \sum_{j=1}^m [1 \times f(C_j)] \quad (11)$$

When all of the sets are of the same size, the microaverage and macroaverage produce the same average. This is not true in our case, because although each sector has the same defined period of high demand (13:00-23:00 EDT), the frequency of forecasted weather varied. In a machine learning application when the sets are not of the same size, all things being equal, the microaveraged error will tend to be lower than the macroaveraged error because sets with less training have a lower weight.

VI. Experimental Results

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. As shown in Fig. 2, our dataset overwhelmingly consists of light forecasted weather (as measured by *FWCI*), where we expect little affect from the weather. Coupled with the much larger expected variation from demand uncertainty (see Section IV), there may be little to distinguish an accurate weather model from an inaccurate one when evaluated over the entire dataset. To compensate somewhat, we primarily focus on the cases where heavy weather was forecast, which we define to be when $FWCI \geq 0.5$. Presumably differences in weather models will be more apparent in this region, though we note that again this subset is skewed to the lighter end of the heavy weather cases (i.e., there are far more cases in $0.5 \leq FWCI < 0.6$ than there are in $0.9 \leq FWCI < 1.0$), and we may have introduced some bias by using *FWCI* as the selection criterion. We calculated the following statistics from the set of errors:

- **Error Bias:** The overall error mean, indicating if the model is generally under- or over-estimating peak count.
- **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.

Table 1 provides the mean error statistics over the 848 cases where $FWCI \geq 0.5$. The baselines *wxmean* and *clearmean* clearly have the worst results on this subset. This supports the theory that peak counts do differ when heavy weather is forecast, and that properties of those forecast have some predictive power. *wxmean* and *clearmean* also have a large negative error bias, indicating that they overestimate the available capacity under these conditions overall. Of the weather models, the count- and max- scan variants tend to do worse than the others. This would suggest that those scan features are not as informative as the *totalscan* feature or even simple *FWCI*. There is also a general trend for slightly poorer results for models that aggregate features; namely, the mean models do worse than their non-mean counterparts and *FWCI* is outperformed by *DFWCI*. Some discrimination seems to be gained from the additional detail, though the difference is slight. Over all the models, *DFWCI* and *totalscan* have the lowest error statistics.

In contrast, Table 2 provides the mean error statistics over the 28148 where $FWCI < 0.5$. Much of the difference between the results of various models has dissipated for these lighter forecasted weather cases. Again, this is consistent with our assumption that demand uncertainty overwhelms the affect of light forecasted weather. The error

biases exhibited by *wxmean* and *clearmean* have dropped considerably. This makes sense, as the models were built over the entire dataset (all 28996 instances, including both heavy and light forecasted weather but not clear weather forecasts), and since this is dominated by the light weather cases, the bias should be lessened in this region*. Nonetheless, *clearmean* still has somewhat worse error statistics, further invalidating its assumption that forecasted weather does not affect peak counts. On the other hand, *wxmean* performs only slightly worse than the other models, and the remaining models appear nearly identical in their error statistics.

The mean error statistics give an overall view of how the various models compare. For a more detailed view, we look at a comparison against the other models for the best and worst performing models, again on the $FWCI \geq 0.5$ subset, by plotting the difference in scaled absolute errors (absolute error/expected demand) on the same observed instance. Observing the difference is useful because it factors out the inherent variability that comes from demand uncertainty. Since we are observing absolute scaled error, a difference of 1.0 would be quite large; for instance, if the expected demand was 10 aircraft, it would mean having an error of 10 aircraft *more* than that of the other model.

Table 1. Results over subset where $FWCI < 0.5$

	Mean Bias		Mean Med. AE		Mean Mean AE		Root Mean MSE	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
<i>Wxmean</i>	0.008	0.008	0.216	0.220	0.258	0.263	0.328	0.334
<i>Clearmean</i>	0.019	0.022	0.241	0.252	0.289	0.302	0.377	0.399
<i>FWCI</i>	0.000	-0.001	0.212	0.215	0.251	0.255	0.319	0.323
<i>DFWCI</i>	0.000	0.000	0.211	0.215	0.250	0.253	0.318	0.321
<i>Countscan</i>	0.002	0.002	0.213	0.216	0.252	0.255	0.320	0.323
<i>Countscanmean</i>	0.003	0.002	0.211	0.214	0.251	0.255	0.319	0.323
<i>Maxscan</i>	0.003	0.003	0.212	0.216	0.252	0.256	0.320	0.324
<i>Maxscanmean</i>	0.004	0.004	0.214	0.217	0.252	0.256	0.320	0.324
<i>Totalscan</i>	0.000	0.000	0.212	0.215	0.250	0.253	0.317	0.321
<i>Totalscanmean</i>	0.000	-0.001	0.212	0.215	0.251	0.254	0.318	0.322

Table 2. Results over subset where $FWCI \geq 0.5$

	Mean Bias		Mean Med. AE		Mean Mean AE		Root Mean MSE	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
<i>Wxmean</i>	-0.265	-0.287	0.338	0.359	0.357	0.376	0.425	0.439
<i>Clearmean</i>	-0.280	-0.307	0.362	0.393	0.380	0.405	0.456	0.476
<i>FWCI</i>	0.012	0.047	0.214	0.219	0.252	0.253	0.319	0.325
<i>DFWCI</i>	-0.002	0.015	0.214	0.212	0.245	0.238	0.310	0.301
<i>Countscan</i>	-0.080	-0.068	0.243	0.232	0.264	0.257	0.327	0.319
<i>Countscanmean</i>	-0.087	-0.071	0.245	0.238	0.268	0.260	0.331	0.325
<i>Maxscan</i>	-0.107	-0.101	0.246	0.242	0.274	0.270	0.338	0.335
<i>Maxscanmean</i>	-0.127	-0.116	0.254	0.248	0.279	0.271	0.343	0.337
<i>Totalscan</i>	0.008	0.029	0.208	0.203	0.242	0.231	0.308	0.300
<i>Totalscanmean</i>	0.013	0.048	0.213	0.217	0.250	0.251	0.316	0.321

* This is not always guaranteed by linear regression but is when the distribution of errors is approximately normal, which is the case for the complete dataset.

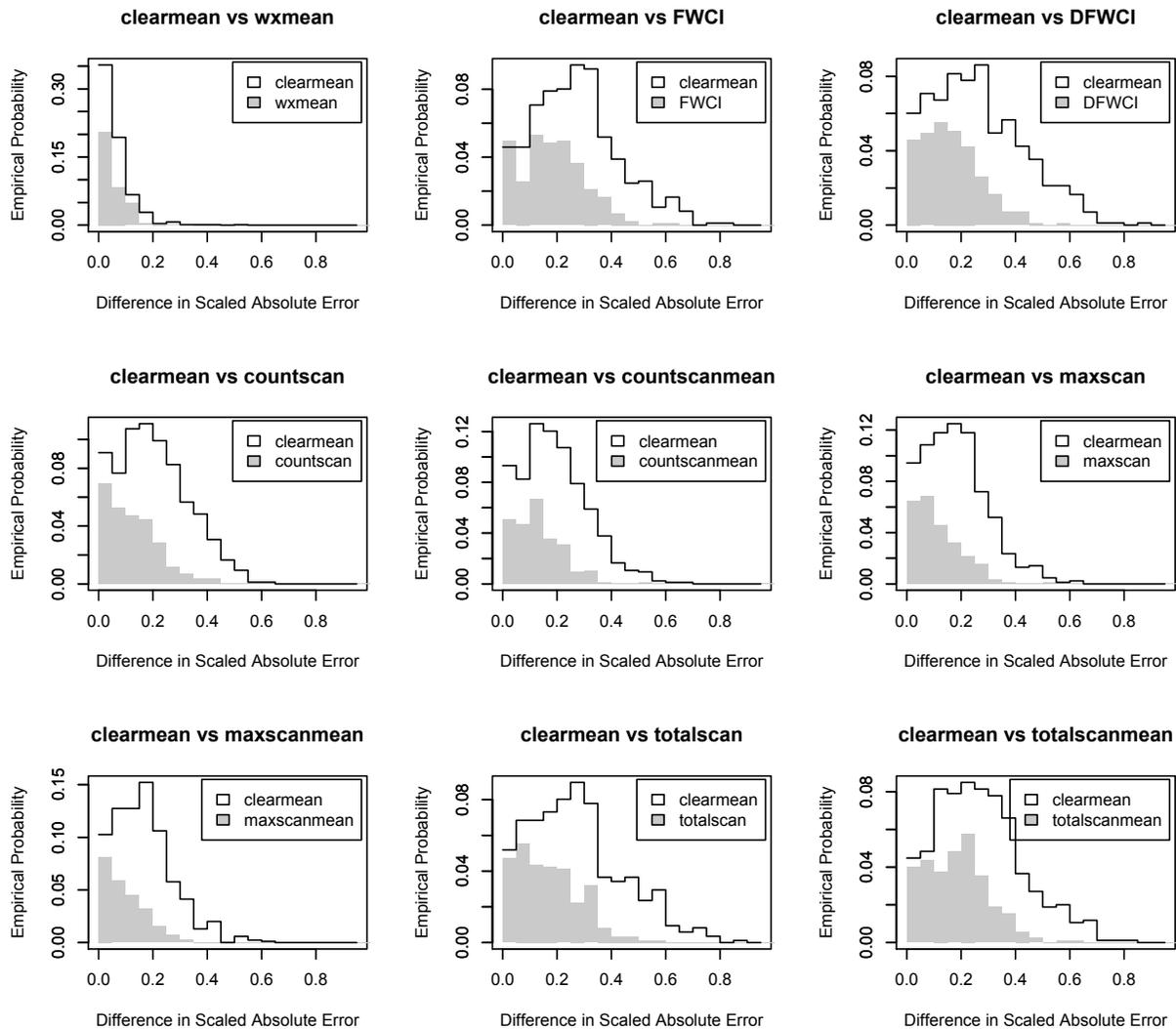


Figure 5. Observed distribution of error differences for *clearmean* when $FWCI \geq 0.5$

Fig. 5 shows how the *clearmean* model performed relative to the other models in our experiment. Each plot is a histogram folded over onto the positive range only, to make it easier to compare frequencies. The white boxed region shows how often we observed *clearmean* performing worse than the compared model. For example, the *clearmean* vs *FWCI* plot shows that *clearmean* had a scaled absolute error between 0.6 and 0.65 more than *FWCI* roughly 2% of the time. The shaded region shows the opposite condition: how often we observed the compared model performing worse than *clearmean*. For the same chart, the observed percentage of *FWCI* having a scaled absolute error between 0.6 and 0.65 more than *clearmean* is represented by only a thin line, perhaps 0.1%; the observed empirical probability of *FWCI* having a scaled error between 0.1 and 0.15 is about 5%. Overall, the white boxed area is the empirical probability of the featured model being outperformed by the compared model, and the shaded area is the empirical probability of the compared model outperforming the featured model, with relative errors increasing from left to right. Since charts are probabilities, the sum of the area under the two regions (boxed and shaded) equals one.

Fig. 5 makes a compelling graphical case for the inferiority of *clearmean* to all other models. Indeed, *clearmean* was observed to be slightly less likely to have scaled error between 0.0 and 0.1 greater than *FWCI* than the converse; for all other models and scaled error differences, *clearmean* was as or more likely to be off by the given amount. The difference between *clearmean* and *wxmean* is rather small, which is unsurprising as both models are insensitive to characteristics of the forecasted convective weather. The difference between *clearmean* and the other models is

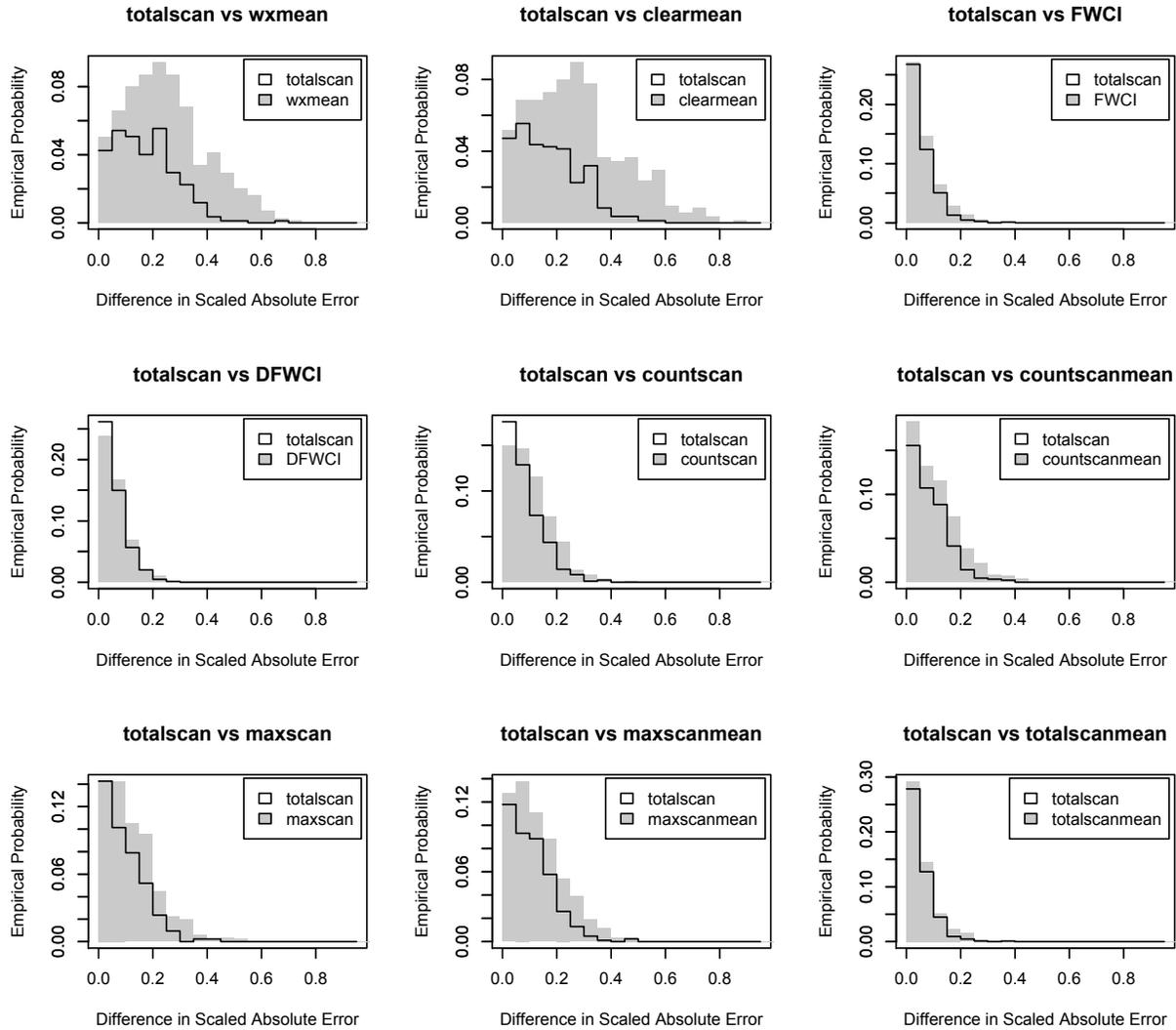


Figure 6. Observed distribution of error differences for *totalscan* when $FWCI \geq 0.5$.

more striking in the other cases, each with a fairly wide spread of outcomes. This can be explained by the fact that *clearmean* is a constant model; it is often wrong, but will have a better prediction when the weather impact is less than expected.

Fig. 6 shows the relative performance of *totalscan* with respect to the other models. *Totalscan* was observed to be more likely to produce a scaled error between 0.0 and 0.05 greater than *DFWCI* and *countscan* than the converse; for all other models and scaled error differences, *totalscan* was as or less likely to be off by the given amount. The difference between *totalscan* and the baselines (*wxmean* and *clearmean*) is striking; it is more subtle for the forecast-dependent models. Indeed, the difference between *totalscan* and *FWCI*, *DFWCI* and *totalscanmean* is challenging to visually assimilate at this resolution of our charts.

Since the probabilities given are based on observation, they may be perturbed by random chance and not exactly reflect the true probabilities. Unfortunately, the distribution of errors in this subset is clearly *not* normal, which violates a requirement of many well-known statistical tests, for instance the paired t-test for a comparison of means (for instance, those in Table 2). Instead, we investigate if one model can be expected to give a lower error than another model on a given weather situation (without being concerned about the size of any difference). This corresponds to comparing the area of the boxed and shaded regions in Fig. 5 and Fig. 6. Specifically, given two models to compare, M_0 and M_A , our null hypothesis H_0 is that the probability of M_0 producing a lower or equal error to M_A is 50% or greater; our alternate hypothesis H_A is the complement, namely that the probability of M_0 producing a greater error than M_A exceeds 50%. With the samples paired, we perform our test without differentiating which

sector the samples came from. This would support a decision-maker who wanted to choose the same model for all sectors that has the lower error most of the time (again, without concern error differences or overall sums).

Table 3 gives the p-values (rounded up to the nearest hundredth) under the null hypothesis H_0 over the forecasted heavy weather subset, with model M_A indexing the rows and model M_0 indexing the columns. The p-value gives the likelihood of observing our data if the null hypothesis is true, so higher numbers give support for H_0 and lower numbers cast doubt on H_0 . Standard statistical procedure is to accept the complementary alternate hypothesis H_A if a certain threshold against the null hypothesis H_0 is reached[†]. 0.05 is commonly used as the threshold for statistical significance, and all such p-values are marked in bold. At this level of significance, we would reject H_0 and accept its complement, which is exactly H_A ; in our decision-making example, the decision-maker who seeks lower error most of the time should choose M_A instead of M_0 . We see that every other model reaches the desired level of statistical significance vs. *clearmean* (by reading down *clearmean*'s column), and so we can reject *clearmean* as a superior model over any other model (in terms of this comparison). Likewise, *totalscan* has a statistically significant difference over every model with the exception of *DFWCI*, so we can also reject all but *DFWCI* in favor of *totalscan*. There is no statistically significant difference between *DFWCI* and *totalscan* according to this test, so we would need to use a different test to prefer one over the other, or simply choose one arbitrarily.

Table 3. P-values under H_0 over subset where $FWCI \geq 0.5$, with M_A by row and M_0 by column.

	<i>Wx-mean</i>	<i>Clear-mean</i>	<i>FWCI</i>	<i>DFWCI</i>	<i>Count-scan</i>	<i>Count-scan-mean</i>	<i>Max-scan</i>	<i>Max-scan-mean</i>	<i>Total-scan</i>	<i>Total-scan-mean</i>
<i>Wxmean</i>	—	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>Clearmean</i>	1.00	—	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>FWCI</i>	0.01	0.01	—	0.87	0.01	0.01	0.01	0.01	0.99	0.96
<i>DFWCI</i>	0.01	0.01	0.14	—	0.01	0.01	0.01	0.01	0.64	0.10
<i>Countscan</i>	0.01	0.01	1.00	1.00	—	0.02	0.01	0.01	1.00	1.00
<i>Countscanmean</i>	0.01	0.01	1.00	1.00	0.99	—	0.01	0.01	1.00	1.00
<i>Maxscan</i>	0.01	0.01	1.00	1.00	1.00	1.00	—	0.01	1.00	1.00
<i>Maxscanmean</i>	0.01	0.01	1.00	1.00	1.00	1.00	1.00	—	1.00	1.00
<i>Totalscan</i>	0.01	0.01	0.02	0.37	0.01	0.01	0.01	0.01	—	0.03
<i>Totalscanmean</i>	0.01	0.01	0.05	0.91	0.01	0.01	0.01	0.01	0.98	—

VII. Conclusions and Future Work

In this paper, we evaluated several weather 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 compensate, 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 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, as compared to simply finding the best parametric fit over the dataset, decreased the possibility of spurious results given our models have a different number of features.

One issue revealed in our analysis is that observed peak counts vary widely when no weather is forecast, suggesting significant variability exists even when there are no weather impacts. Another issue is that forecasted light weather is far more common than forecasted heavy weather in our dataset, but it is presumably more difficult to observe weather impacts in the former case. To compensate, we primarily restricted our analysis to the forecasted heavy weather cases and expected a large amount of residual variability from non-weather sources. Nonetheless, we were able to observe numerical and statistically significant differences in the prediction of peak observed aircraft count of all of our weather models when compared to our weather-insensitive baselines on forecasted heavy weather, indicating that some impact of weather on capacity was captured in our weather models. Among the weather models, *DFWCI* and *totalscan* outperformed the others and were essentially tied in terms of performance. In

[†] However, when the threshold is not reached, it is not necessarily correct to accept H_0 .

addition, there was a tendency for models that preserved more detail to outperform 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 weather models. A better estimation of demand would eliminate some of the variability in the observed peak aircraft counts, making the differences in weather models easier to detect. Of all the weather 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 weather 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 weather models.

Acknowledgments

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

References

- ¹Federal Aviation Administration, "Facility Operation and Administration", edited by Services, System Operations (2008), Vol. Order JO 7210.3V.
- ²Chen, N. Y. and Sridhar, B., "Weather-Weighted Periodic Auto Regressive Models for Sector Demand Prediction," *AIAA Guidance, Navigation, and Control Conference*, 2009.
- ³Krozel, J. and Doble, N. A., "Simulation of the National Airspace System in Inclement Weather," *AIAA Modeling and Simulation Technologies Conference and Exhibit*, 2007.
- ⁴Klein, A., Cook, L., and Wood, B., "Airspace Availability Estimates for Traffic Flow Management Using the Scanning Method," *27th Digital Avionics Systems Conference (DASC)*, 2008.
- ⁵Krozel, J., Mitchell, J. S. B., Polishchuk, V., and Prete, J., "Capacity Estimation for Airspaces with Convective Weather Constraints" *AIAA Guidance, Navigation, and Control Conference*, 2007.
- ⁶Song, L., Wanke, C., Greenbaum, D., Zobell, S., and Jackson, C., "Methodologies for Estimating the Impact of Severe Weather on Airspace Capacity," *8th AIAA 2008 ATIO Conference*, 2008.
- ⁷Callaham, M. B., DeArmon, J. S., Cooper, A. M. et al., "Assessing NAS Performance: Normalizing for the Effects of Weather," *The 3rd USA/Europe Air Traffic Management Research and Development Symposium*, 2001.
- ⁸Chatterji, G. B. and Sridhar, B., "National Airspace System Delay Estimation Using Weather Weighted Traffic Counts," *AIAA Guidance, Navigation and Control Conference*, 2004.
- ⁹DeLaura, R. and Evans, J., "An Exploratory Study of Modeling Enroute Pilot Convective Storm Flight Deviation Behavior," *12th Conference on Aviation, Range and Aerospace Meteorology*, 2006.
- ¹⁰DeLaura, R., Robinson, M., Pawlak, M., and Evans, J., "Modeling Convective Weather Avoidance in Enroute Airspace" *13th Conference on Aviation, Range and Aerospace Meteorology*, 2008.
- ¹¹Chan, W. N., Refai, M., and DeLaura, R., "An Approach to Verify a Model for Translating Convective Weather Information to Air Traffic Management Impact," *7th AIAA Aviation Technology, Integration and Operations Conference (ATIO)*, 2007.
- ¹²Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco, 2005.