5.2 STATISTICAL FORECASTING OF CEILING FOR NEW YORK CITY AIRSPACE BASED ON ROUTINE SURFACE OBSERVATIONS[†]

Frank M. Robasky*
Massachusetts Institute of Technology Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

F. Wesley Wilson
National Center for Atmospheric Research
Boulder, CO 80301

1. INTRODUCTION

Air traffic in the United States is highly congested in its "Northeast Corridor", an area that roughly encompasses the airspace from Washington, DC to Boston. This region is frequently affected by low cloud ceiling and visibility conditions during the cool season, often in association with synoptic-scale low pressure systems. Operating under IFR (Instrument Flight Rules) for extended periods of time substantially reduces airport capacity and can cause significant delay at major airports. Anticipating transitions into and out of IFR ceiling and visibility conditions can mitigate air traffic disruption by allowing for appropriate upstream planning. For instance, an accurate forecast of the lifting of cloud ceiling out of IFR range would allow for the release of more planes upstream to take advantage of the anticipated increase in capacity.

The Federal Aviation Administration (FAA), through its Aviation Weather Research Program (AWRP), is currently sponsoring the Northeast Winter Ceiling and Visibility Project (NECV). Its purpose is to provide situational awareness of current ceiling and visibility conditions in the Northeast United States in a way tailored to the needs of air traffic control (ATC), as well as to bring a number of various but complimentary technologies to bear on providing automated 0-12 hour forecasts of upcoming conditions. Methodologies currently under development include numerical weather prediction (NWP) applications, 1-dimensional column modeling, tracking of aviation-impacting cloud, and statistical forecast models (Clark 2006). This presentation describes the development of statistical forecast models for major New York City airports.

The statistical forecast models use routine regional meteorological observations as predictors for future values of ceiling and visibility for selected locations. These predictors consist primarily of hourly surface observations, but upper air soundings and buoy data are available for use as well. The methodology for building the models is based on non-linear regression, with the nonlinearity entering in the spirit of Generalized Additive Models (Hastie and Tibshiriani 1990). Several innovations are introduced to aid in predictor selection and to enhance the skill and stability of the final models.

Statistical models such as these have been successfully developed and used recently in an operational setting for ATC. The recently completed San Francisco (SFO) Marine Stratus Initiative (also sponsored by AWRP) features a real-time display and forecast system, which contains as one of its components a regional statistical forecast model (Wilson 2004, Clark et al. 2005). The model uses hourly surface observations from the San Francisco Bay area along with the Oakland sounding to produce regular forecasts of stratus dissipation during the warm season. The performance of this model during two years (May – October) of real-time operations is given in Table 1. The context for the marine stratus model differs from that for NECV in several important ways. In SFO, warm season stratus dissipation is a diurnal phenomenon, governed primarily by mesoscale and radiative processes in conjunction with local topography. The NECV problem is more affected by synoptic dynamics, and less by the diurnal component.

This paper next provides a high-level summary of the methodology that has been developed to build these statistical forecast models followed by details of the initial NECV problem, including some discussion of the quality of the predictor data. Model accuracy can be improved by development over

[†]This work was sponsored by the Federal Aviation Administration under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

*Corresponding author address: Frank Robasky, MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420-9108; e-mail: frankr@ll.mit.edu

phenomenological partitions of the available cases; a method of partitioning the cases is described. The paper concludes with a discussion of near-term tasks.

Table 1.
Operational Performance Summary from the 2003-4 Warm Seasons of the SFO Regional Statistical Forecast Model.

Forecast Hour (UTC)	Number of Forecasts	Median Absolute Error (hrs)	Bias (hrs)
09	107	0.96	0.40
11	121	1.03	0.58
13	124	0.84	0.27
15	123	0.89	0.45
16	116	0.74	0.33
17	87	0.51	0.01
18	55	0.45	0.28

Forecasts are of stratus "burn-off". This typically occurs between 16-18 UTC, as is reflected in the smaller number of forecasts and the improved accuracy at these hours.

2. MODEL-BUILDING METHODOLOGY

Building meteorological statistical forecast models presents a number of challenges. There are a large number of potential predictors to choose from, far more than would be appropriate in any particular model. In the present case, there are at least 20 reliable observing stations within 200 km of New York City. If seven basic parameters are considered for each station (wind (u, v), temperature, dew point, pressure, ceiling, and visibility), that leads to 140 potential predictors. If past observations at various lags, either on their own or expressed as differences, are considered then the number increases exponentially. Many of these will be highly correlated with others (for example, the concurrent ceiling observations at JFK and LaGuardia). Some predictors have more value when considered in combination than when considered separately (e.g. pressures at two stations differenced and used as a crude gradient). An objective method of determining the most important predictors and their possible pairings is developed. Once these are determined, it remains to combine them into a single forecast model in a quasi-optimum way for the regression analysis, avoiding the pitfalls of over-training. A methodology has been developed which meets these challenges and provides other improvements. An earlier version of this process was used to develop the statistical models for the SFO system.

Once the training data have been subjected to quality control (to be described below) they are subjected to an automated model building process. The components of this process are:

- *Synergy:* A combination of predictors has a synergistic relationship when their combined predictive impact is significantly stronger than might be anticipated from their individual skills. This step automatically identifies such synergies from the full set of training data. It establishes their defining coefficients, and passes them along as single entities to subsequent steps. These synergies can be viewed as virtual sensors.

- *Pre-processed Correlation Enhancing Predictor Transformation (PreCEPT):* This step rescales each predictor to maximize its correlation to the response variable (predictand). A piecewise linear function is used for the rescaling. The effect is to highlight the ranges of the predictor that are associated with the most response in the predictand. An example from SFO is shown in Figure 1. This component introduces the non-linearity into the regression analysis.

- *Nulling:* This algorithm reduces information redundancies in the set of predictors by examining their multiple correlations with the predictand. Statistically, the process is designed to accept an additional small amount of bias error in exchange for a reduction in model complexity. This is accomplished iteratively, with the least important predictor being eliminated at each step. The output of the nulling process is a list of the most important predictors. In our application, this list has been limited to the 20.

- *Compare(h,k):* This process determines the forecast models with best performance, based on a combination of all-model comparison and cross-validation. The comparison step involves building all models with *h* or fewer predictors from a list of *k* predictors (the finalists from the Nulling analysis). A cross-validation exercise is conducted for each model candidate, in which a variety of model skill and stability measures are computed. The cross-validation exercise involves training the model candidate on a randomly selected 90% of the training data and evaluating its performance on the remaining 10% of the data. This process is repeated 100 times, thereby creating 100 models with the same list of predictors. Skill and stability parameters are computed from the aggregate of the model characteristics and evaluation data from these 100 models. The result of the Compare analysis is a list of the selected models, together with their stability and cross-validation properties, ranked according to their cross-validation R^2 skill (R is the correlation of the model prediction vector with the training response vector).

- *Selection:* This step is more interactive than the previous ones. It consists of a spreadsheet tool that allows the developer to select an ensemble of final forecast models from those models that have the best cross-validation R^2 skill. This is done by examining the other skill and stability metrics from the *Compare* step, and by examining ensemble model performance on the full training set. Diversity of the predictors in the ensemble is also examined to allow for robust real-time performance in the presence of missing observations. Currently a 10-member ensemble in envisioned for operational use.

While a great majority of this process is automated each step produces output that allows an easy audit of the results to examine their meteorological integrity.

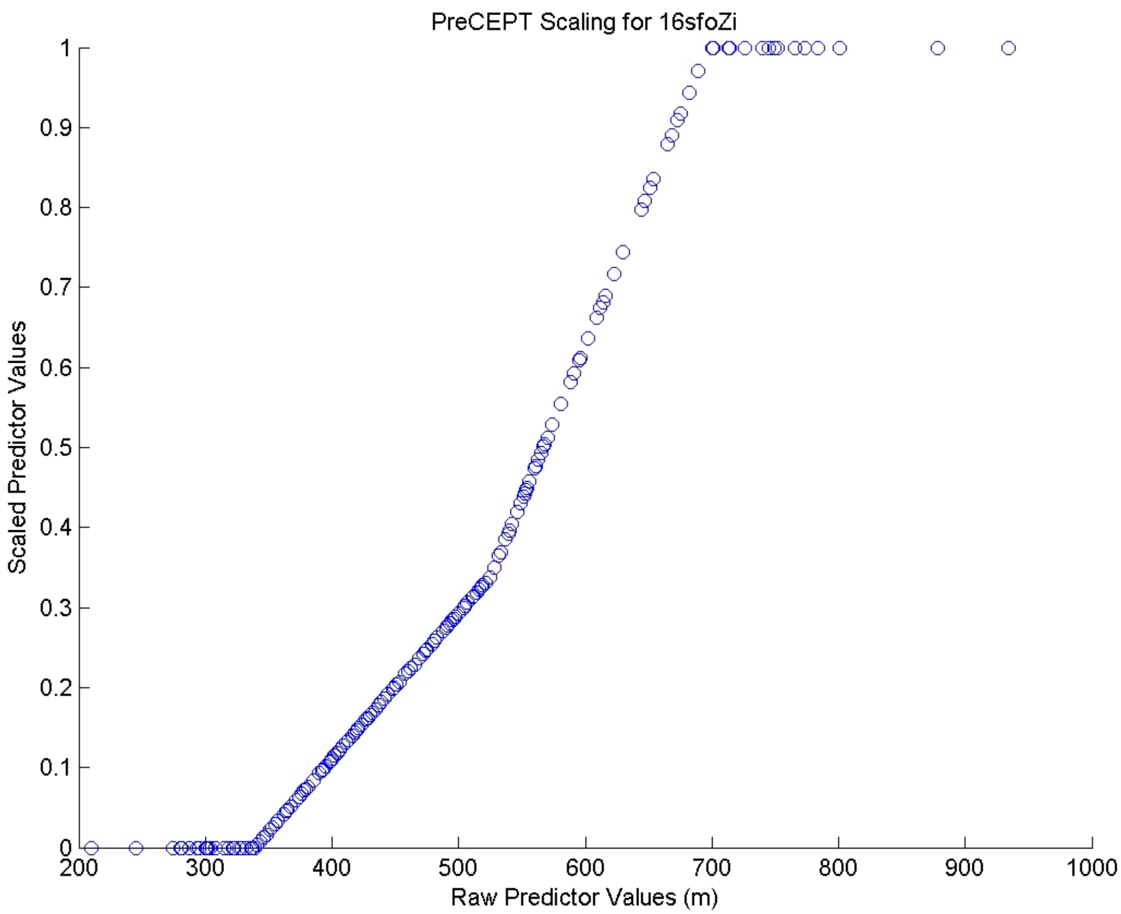

Figure 1. Scatter plot showing the PreCEPT scaling of 1600 UTC inversion height at the SFO airport. Predictand response for this predictor is seen to be limited to the 350 to 700 m range.

3. INITIAL FORECASTING APPROACH FOR NECV

The first modeling efforts focus on forecasting cloud ceiling at New York City's LaGuardia airport (LGA). Model equations are developed for situations where a ceiling of 2000 m or less is already in place. (The IFR threshold for ceiling is 1000 ft, or 305 m.) Separate models are developed for each forecast horizon from 1 – 12 hours, in hourly increments, with initial focus on the 1 – 3 hour horizons. The goal is to accurately predict transitions into and out of the IFR state.

The initial training data set is assembled from an archive of northeastern hourly surface observations covering the 1977-2004 period. The most reliable stations, defined by the ratio of hours for which an observation exists within 15 minutes of the top of the hour, are shown in Figure 2. The training set is limited to those hours where a ceiling of 2000 m or less is observed at LGA. If sampled hourly, this training set amounts to nearly 40,000 cases.

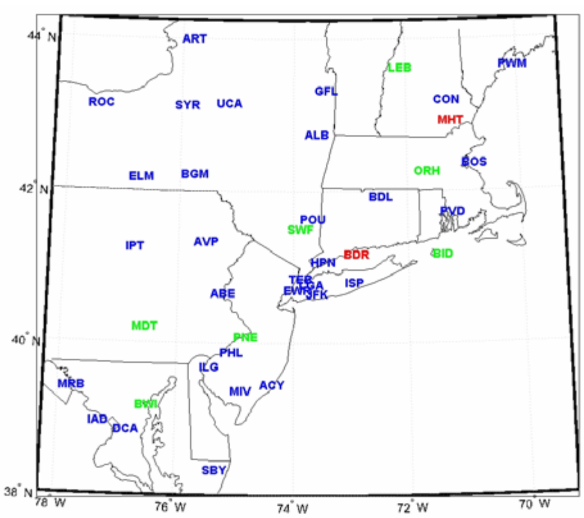

Figure 2. Most reliable NECV surface observation stations. Those in blue have an hourly reliability above 0.90, those in green from 0.80-0.90, and those in red from 0.75-0.80. The four stations crowded near New York City are TEB, EWR, LGA, and JFK.

Several obvious initial transformations of the raw observations are made to help in model development. Ceiling observations are capped at 3000 m. This is done primarily to minimize the discontinuity with respect to observations of unlimited or very high ceilings. Wind observations are decomposed into their u and v components. Dew point depression is calculated from the observations of air and dew point temperature.

4. DATA QUALITY ANALYSIS

The training dataset is subjected to a thorough data quality analysis (DQA), the main components of which are:

- *Simple bounds check:* Identify observations that are physically impossible as erroneous, through simple range checks.

- *Climatological bounds check:* Identify observations as erroneous that are climatologically unlikely. A climatology of each variable stratified by month and hour is constructed, limited to the constraints governing the construction of the training dataset (i.e. LGA ceilings of 2000 m or less). The observations are then subjected to a bounds check of the expected value for the given month/hour ± several standard deviations.

- *Parameter consistency check:* Any derived parameters (such as dew point depression) are also subjected to the simple and climatological bounds checks. If the derived parameter is identified as erroneous, then all contributing parameters are also identified as erroneous. This approach was introduced by Miller and Barth (2003).

- *Data filling:* Missing observations across many stations can quickly diminish the number of full cases available for training. For a first pass, a temporal fill of missing observations is implemented for gaps in the record of two hours or less by linear interpolation. Remaining missing observations are filled by their climatological value as described above.

With the exception of the data filling, which is applied only to the training data, the same DQA techniques are used for the training data and for implementation of the operational forecast models.

Our long period of record for training encompasses the operational shift from largely manual to automated surface observations (via ASOS). This was found to be reflected in the resolution with which cloud ceiling was observed. Prior to ASOS, ceilings greater than 2000 ft (610 m) were reported in 500 ft (152 m) increments. This discontinuity would be harmful to model development through the introduction of a bogus change in predictor variance across the training dataset. Simply ignoring the older data, however, would cause the loss of a large fraction of training cases (roughly from 1977-1985). To avoid this all ceilings above 2000 ft will be standardized to the nearest 500 ft increment across the entire time period.

5. CLUSTERING OF DEVELOPMENT CASES

In the SFO project separate models were developed based upon a phenomenological partitioning of the training data. This partitioning was guided by the desire to discriminate between typical cases where the models might be expected to perform well versus atypical cases where the models would likely be less accurate. In the case of the SFO regional statistical forecast model, this was based upon a crude parameterization of meridional pressure gradient. This indicated the likelihood of on- or off-shore surface flow, with onshore flow tending to prolong stratus in San Francisco Bay. The pressure difference used to represent this was a codification of a longstanding empirical rule used by operational forecasters.

The situation is more complex for the northeastern US. As a first attempt to ascertain what variables would be most useful to classify low ceiling cases at LGA, a clustering analysis was applied to a large number of candidate indicators, including the base observations along with their past values and time trends. Potential indicators were considered singly and in pairs. The analysis was attempted on many random samples of the training dataset taken at 3-hour increments, as it was computationally impractical to analyze the entire set. A hierarchical clustering algorithm was applied, and forced to terminate with 2 or 3 clusters. A metric of the quality of the clustering of these indicators was made by computing the mean Euclidian distance between the cluster centroids and multiplying this by the population ratio of the smallest cluster. A larger metric results from more evenly distributed yet well separated clusters. The indicator(s) were ranked by this metric. The most promising LGA indicators were wind direction and ceiling height, considered separately and as a pair.

A histogram of a typical sample of LGA surface wind directions separated into 3 clusters is shown in Figure 3. The first cluster encompasses the clear maximum of low ceiling observations that are associated with a northeast surface wind from the Atlantic Ocean via Long Island Sound. This is consistent with an independent climatology of low ceiling conditions for this area (Tardif 2006). A second cluster roughly consists of the remaining on-shore directions, and the third cluster corresponds to offshore flow. A sample histogram of LGA ceiling separated into 2 clusters is shown in Figure 4. This shows a separation of the lowest ceilings (anchored by a sample maximum at 100-300 m).

A clustering of the LGA ceiling-wind direction pair is shown in Figure 5. This two-cluster example combines characteristics of each separate indicator as mentioned above. One cluster includes the entire 2000 m depth at the key northeasterly wind direction, tailing to lower ceilings at the remaining on-shore directions. Some clustering of other samples broke this cluster into two based on high/low ceilings.

The initial forecast models are developed on the entire set of training data. This provides a baseline, to determine the benefit of imposing a clustering strategy. The development of a clustering strategy requires the development of a clustering algorithm that can be applied equally to the training data and in real-time for the operational implementation. In addition to improving model skill, the clustering strategy should not introduce product instability. This issue is being actively investigated.

6. SHORT TERM TASKS

Models for LGA ceiling are being developed using the processes described above. These will be developed for all cases, as a baseline, and also for some version of the “northeasterly flow” subset of cases. These models at their various forecast horizons will be evaluated on independent cases (consisting of one or more cool seasons kept out of the training set for this purpose). Statistical model performance will be evaluated against persistence, conditional climatology, NWP (extracted from the MM5 and/or RUC grids), and the operational Terminal Aerodrome Forecasts (TAF). Evaluation will specifically focus on the forecast horizons for which these models show the most skill. Based on this evaluation, regional statistical models will be deployed in realtime for the 2005-6 cool season and monitored by the developers along with the other components of the NECV operational display and forecast system. Long-term modeling work will focus on adding predictors taken from radiosonde, buoy, and NWP data, and on broadening the application to other key airports in the NECV domain.

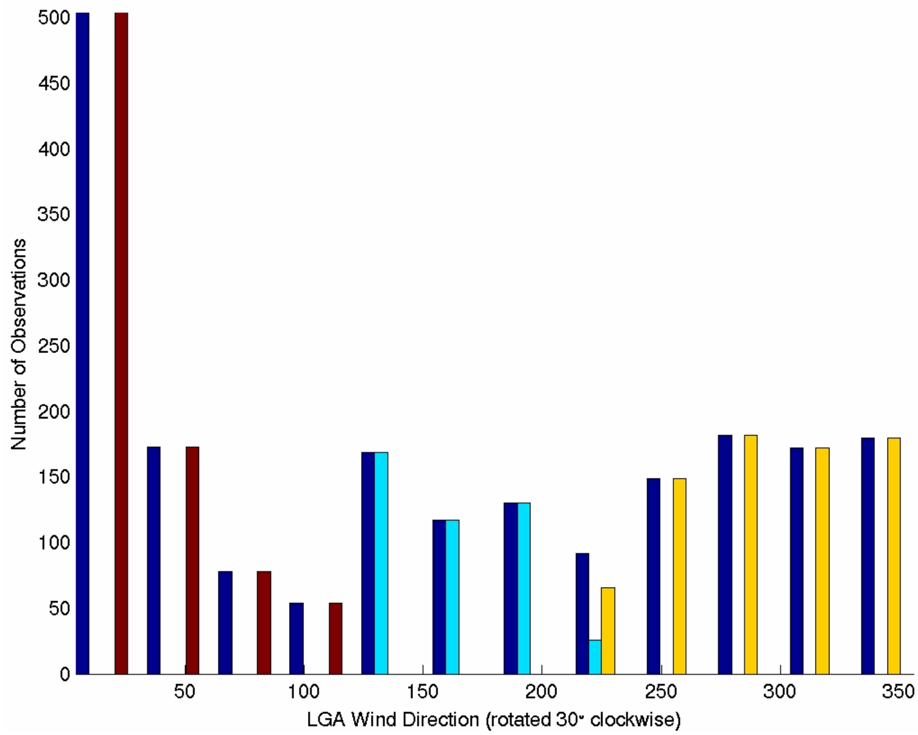

Figure 3. Sample historgam of LGA wind directions (rotated clockwise by 30°). Overall sample population is in dark blue, with a 3-way clustering shown by the other colored bars.

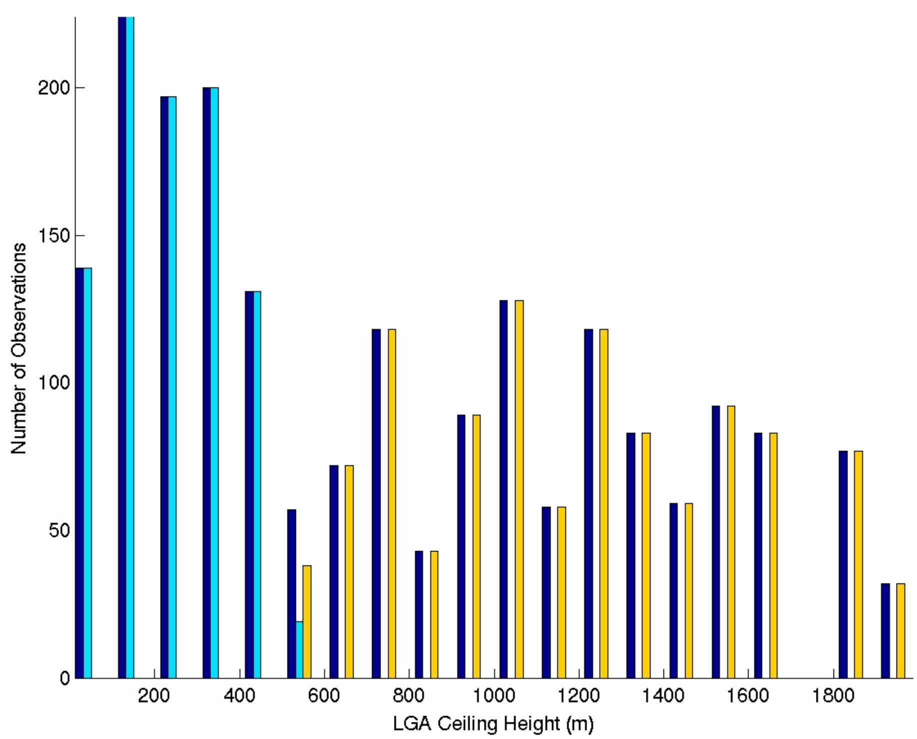

Figure 4. Sample histogram of LGA ceiling heights. Overall sample population is in dark blue, with a 2-way clustering shown by the other colored bars.

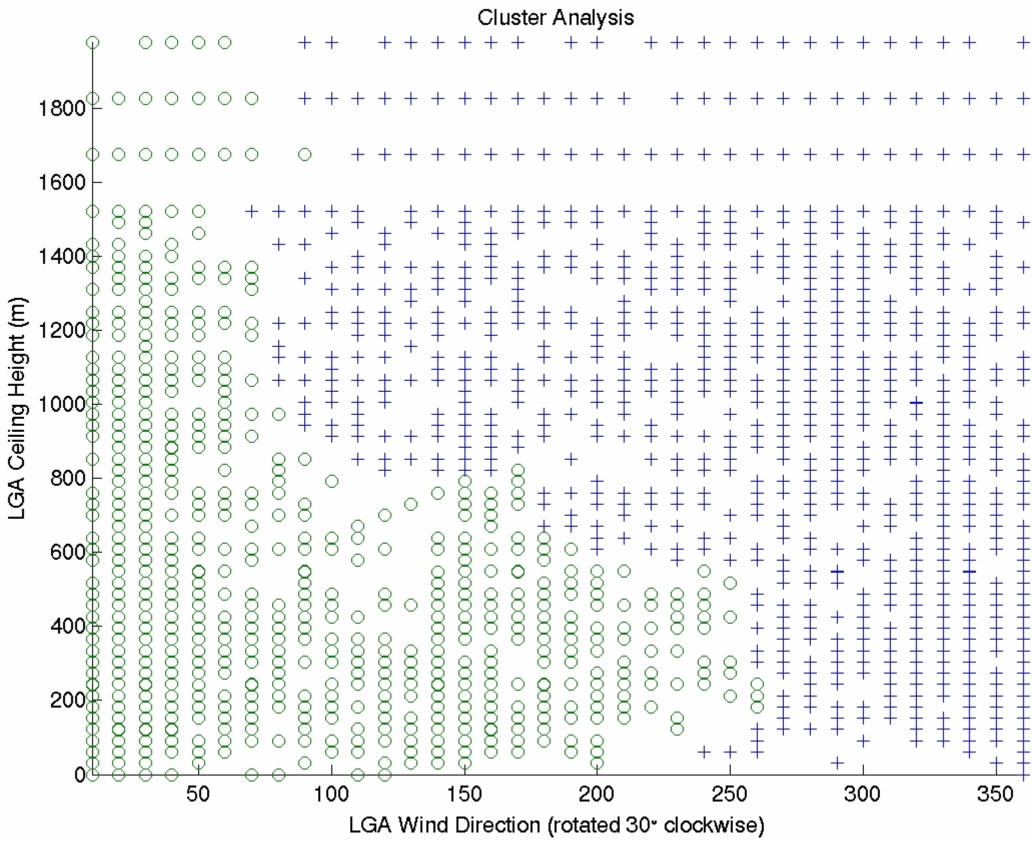

Figure 5. 2-way sample clustering of LGA wind direction (rotated clockwise by 30°) with LGA ceiling height.

7. REFERENCES

Clark, D. 2006: Terminal Ceiling and Visibility Product Development for Northeast U.S. Airports. *12th Conference on Aviation, Range, and Aerospace Meteorology*, Atlanta, GA, Amer. Meteor. Soc.

Clark, D.A., C.F. Ivaldi, F.M. Robasky, K. MacKenzie, R.G. Hallowell, F.W. Wilson, and D.M. Sinton, 2005: SFO Marine Stratus Forecast System Documentation, Massachusetts Institute of Technology, Lincoln Laboratory, Project Report No. ATC-319, Lexington, MA (in press).

Hastie, T.J. and R. J. Tibshiriani 1990: Generalized Additive Models, Chapman and Hall/CRC, Boca Raton, FL.

Miller, P.A. and M.F. Barth 2003: Ingest, integration, quality control, and distributions of observations from state transportation departments using MADIS. *19th International Conference on Interactive Information and Processing Systems*, Long Beach, CA, Amer. Meteor. Soc.

Tardif, R. 2006: A climatological study of low ceiling and fog associated with the occurrence of precipitation in the northeastern United States. *12th Conference on Aviation, Range, and Aerospace Meteorology*, Atlanta, GA, Amer. Meteor. Soc.

Wilson, F.W. 2004: An operational Marine Stratus forecast system for San Francisco International Airport. *11th Conference on Aviation, Range, and Aerospace Meteorology*, Hyannis, MA, Amer. Meteor. Soc.