2.3 Forecast Confidence Measures for Deterministic Storm-scale Aviation Forecasts *

Mark. S. Veillette, Haig Iskenderian, and Mike M. Matthews
MIT Lincoln Laboratory, Lexington, MA

1. INTRODUCTION

Deterministic storm-scale weather forecasts, such as those generated by the FAA's 0-8 hour CoSPA system (Wolfson et al. 2008; Pinto et al. 2009; Iskenderian et al., 2011) and shown in Figure 1, are highly valuable to aviation traffic managers. They provide forecasted characteristics of storm structure, strength, orientation, and coverage that are helpful for strategic planning purposes in the National Airspace System (NAS). However, these deterministic weather forecasts contain inherent uncertainty that varies with the general weather scenario at the forecast issue time, the predicted storm type, and the forecast time horizon. This uncertainty can cause changes in the forecast from update to update, thereby eroding user confidence and ultimately reducing the forecast's effectiveness in the decision-making process. Deterministic forecasts generally lack objective measures of this uncertainty, making it difficult for users of the forecast to know *a priori* how much weight to give the forecast in their decision making process.

Forecast confidence generated by a human is often based on heuristic principles that can be influenced by a forecaster's past experience. This heuristic approach can result in flaws in decision making due to personal biases (Gibbons et al., 2013). For example, a recent poor forecast that resulted in a costly decision to operations might decrease a forecaster's faith in future forecasts from the forecast system, even if that recent bad case is not representative of typical system performance. A goal of this work is to develop an automated forecast confidence metric to help increase the operational utility of the 0-8 hour deterministic forecasts. This confidence metric uses historical forecast performance to provide a probabilistic prediction of forecast performance.

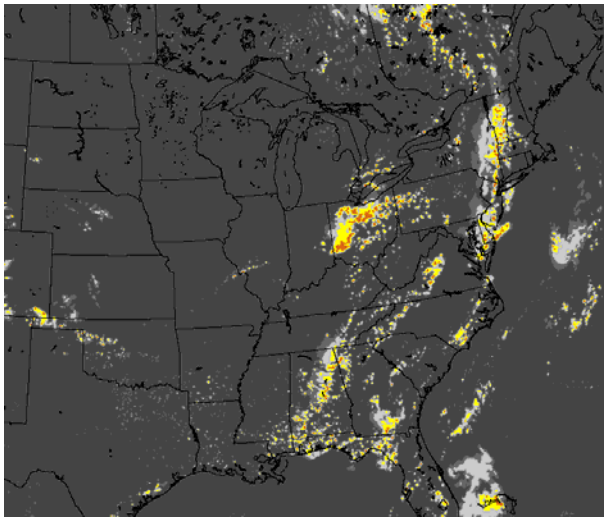


Figure 1: Eight hour CoSPA precipitation forecast valid 20 UTC 10 July 2013.

When designing a forecast confidence metric, it is important to consider the users of the forecast product as well as the decisions influenced by the forecast. The focus of this effort is aimed at strategic (2-8 hour lead time) enroute decision-making in the NAS. This concept is notably different from the forecast confidence currently available for the 0-2 hour Corridor Integrated Weather System (CIWS; Wolfson and Clark 2006) forecast. The 2-8 hour forecast confidence will be focused on

*This work was sponsored by the Federal Aviation Administration under Air Force Contract No. FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.
†*Corresponding author address*: Mark S. Veillette, MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420-9185; e-mail: mark.veillette@ll.mit.edu

how accurately a forecast is expected to depict availability of enroute airspace associated with key Flow Constrained Areas (FCAs) in the NAS, rather than recent forecast performance based on the precise location and intensity of individual storm cells in near-airport regions, which is the basis of the CIWS scores.

The method described here uses characteristics of the current and historical weather forecasts, such as spatial scale, intensity, weather type, orientation, permeability, and run-to-run variability of the forecasts in a statistical model to provide a measure of confidence for forecasted aircraft blockage associated with various FCAs. The results from the method, which will also be presented, provide the user with a measure of forecast confidence in several blockage categories (none, low, medium, and high) associated with each FCA. This method is directed toward helping strategic planners in the NAS by providing them with an objective measure of forecast confidence in terms of FCA route blockage.

2. FORECAST CONFIDENCE DEFINITION

Before describing a forecast confidence algorithm, it is first necessary to provide a concrete definition of "forecast confidence" from which an algorithm can be based. Below, forecast confidence is defined in general terms, and later this theory is applied to a particular application in Air Traffic Management.

To define forecast confidence for a deterministic storm scale forecast, we require the following three pieces of information:

1) In what region is the forecast confidence being provided? This question seeks to identify the domain over which confidence is to be computed. This domain can represent a geographic region, or a specific air traffic resource (e.g., a terminal area or an FCA).

2) Confidence in what? This question seeks to define a measure of forecast skill through which confidence is provided. This measure is represented by a scoring function $S(F,T)$, which measures the skill of a forecast F over the specified domain when provided the observed truth (T).

3) What information is available to compute confidence? This question seeks to identify all information available at forecast issue time that could be leveraged to calculate forecast confidence. This information includes components of the current forecast, and all previous forecasts and including validation when it is available. This information is denoted by $I(F)$, which represents a numerical vector of features extracted from current and previous forecasts. This information should be relevant to how the forecast is expected to perform in the context of the scoring function $S(F,T)$.

Once these have questions have been addressed, forecast confidence is defined as the expectation of forecast skill, conditioned on information available to the forecast system at forecast issue time:

$$\text{Confidence} = \langle S(F,T) | \, I(F) \rangle \qquad (1)$$

This expression of forecast skill is to be estimated at forecast issue time, when the truth that would verify the forecast is unknown. The skill function can be any numeric function which quantifies skill, however it is helpful to consider the case where S is binary ("right" or "wrong"). In this case, the forecast confidence can be interpreted as the probability that the forecast is "correct" (i.e. $\text{Prob}[S(F,T) = 1]$).

The expectation of forecast skill in Equation (1) is computed using historical forecast performance. By conditioning on $I(F)$, this definition predicts the skill of a current forecast F by averaging skill over "similar"

historical forecasts (i.e. historical forecasts with similar $I(F)$'s). With this in mind, estimation of (1) can be framed as a supervised learning problem with a training set of the form $(F_i, T_i), i = 1, \dots, N$ containing historical forecasts F_i and their corresponding observations T_i. The expectation in (1) can be estimated using machine learning methods. The best choice of learning algorithm will vary depending on the scoring function and the complexity of $I(F)$. Below we give an example for a specific choice of $S(F,T)$ and $I(F)$.

The general framework described above can be applied to a number of "classical" skill functions: e.g. CSI, FSS, Forecast Bias, etc.; however it is important in practice that confidence scores are easily interpreted by users. If the meaning of the confidence score translates poorly to the application of the forecast, the confidence score will be of little help to users, even if an algorithm can perfectly predict a forecast's skill. It is for this reason we consider an "operationally relevant" scoring function which is meant to score a forecast based on air traffic impact and whose meaning is well defined to Air Traffic Management (ATM).

3. FORECAST CONFIDENCE CONCEPT FOR FCA BLOCKAGE

3.1 Route Blockage Algorithm

The concept of forecast confidence developed in this work looks at the impact of convective weather on airspace flow across FCAs. To develop this concept, a route blockage algorithm was applied to measure weather impact on air traffic flow. This route blockage algorithm is a slight variation of the algorithm described in (DeLaura et al. 2011). The algorithm described in this paper creates equally spaced parallel routes aligned transverse to a given FCA. Impact is assessed by considering two main factors along each route: (1) how long an aircraft is expected to encounter hazardous weather, and (2) the maximum intensity of the weather encountered measured by the Convective Weather Avoidance Field (CWAM) (DeLaura, et al. 2008). By combining impact over all routes, the algorithm estimates the percentage of an FCA which is blocked due to weather. Using this blockage percentage, the weather scenario in the vicinity of the FCA is classified into one of four impact categories: No Impact (0% blockage), Low Impact (1-15% blockage), Medium Impact (16-50% blockage), and High Impact (51-100% blockage). These are initial blockage ranges and we expect to refine them based on user feedback. Figure 2 provides examples of weather scenarios from each impact category.

The route blockage algorithm can be applied to radar mosaics (as pictured in Figure 2), or it can be applied to CoSPA 2-8 hour forecasts (as shown in Fig. 1) to generate a forecast of weather impact across an FCA. When applied to the forecast, the route blockage algorithm can be used to score a forecast in terms of how well the forecast estimates the weather impact on air traffic. By doing so, this method measures forecast skill in a way that is related to ATM decision making (e.g. setting flow rates within an FCA for Airspace Flow Programs).

Let C_F be the forecasted FCA impact category of a given forecast, and let C_T be the observed impact category at forecast valid time. We define forecast skill using the function

$$S(F,T) = \mathbf{1}(C_F == C_T) \quad (2)$$

which is equal to 1 if and only if a forecast correctly predicts the weather impact on flow within an FCA. Because S is binary, forecast confidence (which is defined as an expectation of forecast skill in (1)) can be interpreted as the probability the forecast correctly predicts FCA impact.

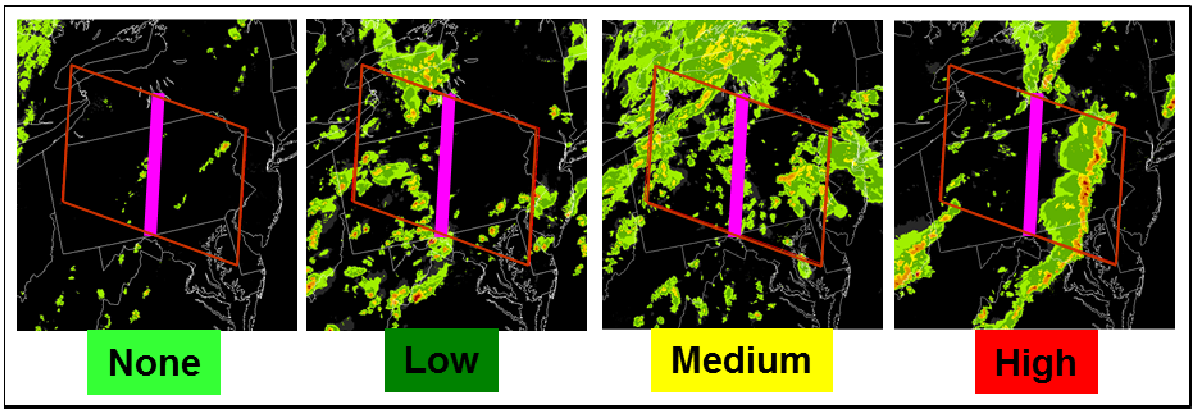


Figure 2: Example of blockage categories associated with flow across the FCA shown by the pink line. The relevant weather used to calculate the blockage is within the red polygon.

3.2 Forecast Features

Forecast information, $I(F)$, represents the input to the forecast confidence model. The input is composed of a fixed number of numerical forecast features extracted from the current CoSPA forecast and a corresponding time-lagged ensemble. The forecast features used in this study include forecast intensity and scale of forecasted VIL, blockage features computed from the route blockage algorithm, features extracted from a three member time-lagged ensemble of forecasts, and auxiliary features such as time of day and geographic location. For this study, $I(F)$ consisted of 57 numerical features taken from the most recent forecast and an associated time-lagged ensemble.

3.3 Algorithm Training

The forecast confidence model was trained using six months of data from the summers of 2012 and 2013. To target times of highest convective weather impact, hourly forecasts out to 8 hours valid between 12 UTC and 23 UTC were used for training. Data were collected for the nine FCA regions shown in Figure 3. The FCA regions were chosen to represent major flow structures in the Eastern US and the sizes of the FCAs were chosen to be comparable to commonly used FCAs by ATM. For each forecast F_i in the dataset, the vector of forecast features $x_i = I(F_i)$ is computed and saved. In addition, the impact category of observation T_i corresponding to the valid time of the forecast is encoded as an integer: $y_i = 1$ (No impact) through 4 (High impact).

The forecast confidence model developed here is made of two parts, a classifier to assign "scores" to each impact category, followed by a calibration which maps these scores to probabilities. Forecast confidence is given by the probability of the forecasted category (which is equivalent to the definition (1) for this choice of skill function). As an added benefit, this methodology also provides probabilities of other blockage categories (even though these categories are not explicitly forecasted in the current deterministic forecast). Providing the probabilities of all impact categories, in addition to the confidence in the forecasted category, can help users better assess uncertainty in weather impacts.

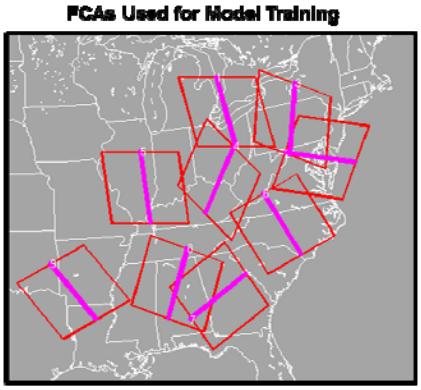


Figure 3: Locations of FCAs and the corresponding regions used to train the forecast confidence model.

The classifier was trained using a variation of the AdaBoost.M2 algorithm (Freund and Schapire, 1996) known as RUSBoost (Seiffert, 2010). Since the training set is largely dominated by 'None' or 'Low' impact events, RUSBoost uses a random undersampling technique to balance the distribution of impact categories in the training set prior to each boosting iteration. Given $I(F)$, the classifier outputs scores associated to each impact category, where the largest score corresponds to the most likely class.

Next, a calibration is performed to map classifier scores to reliable probabilities (i.e. probabilities that reflect the actual frequency of observations given $I(F)$). For this step, a Support Vector Machine with Platt Scaling (Platt, 1999) was trained using the set of classifier scores generated by the RUSBoost classifier. The training set used for calibration was constructed using 10-fold cross validation procedure where classifier scores are generated using subsets of days from the original training set, similar to the method in (Wolpert, 1992).

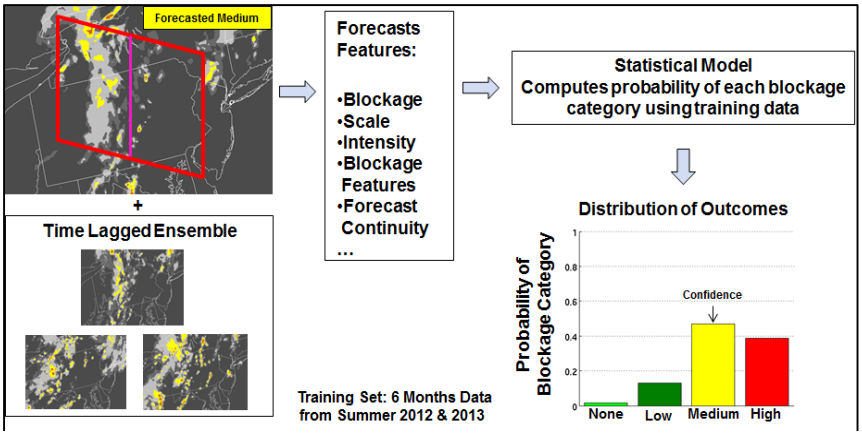


Figure 4: Steps in the forecast confidence algorithm. The forecasted blockage category is input with consistency information from a time-lagged ensemble and other forecast features into a statistical model that computes the probability of each blockage category from the training data.

Figure 4 shows the steps of the forecast confidence algorithm and the probabilities it generates. When a CoSPA forecast is issued, the impact category predicted across an FCA is measured using the route blockage algorithm. The forecast confidence algorithm extracts information from the issued forecast, and uses the confidence model described earlier to estimate the probability that the forecasted impact category will occur, as well as the probabilities of the other impact categories. The probability of the forecasted impact category is the forecast confidence.

3.3 Accuracy of Forecast Confidence

Forecast confidence skill is assessed quantitatively by comparing the probabilistic outputs of the model to observed frequency of each impact category. Reliability diagrams provide a way to visualize how accurately the probabilities describe the actual occurrence of each category (None, Low, Medium and High). Figure 5 shows reliability diagrams

computed for forecast confidence at 2, 4, 6 and 8 hours. These results show that the probabilities generated by the confidence algorithm accurately describe the occurrence of each impact category, especially for the operationally-important None and High cases. The algorithm is less reliable with the Low and Medium impact categories. Reliability diagrams for the other hours (3, 5, and 7) show similar results (not shown).

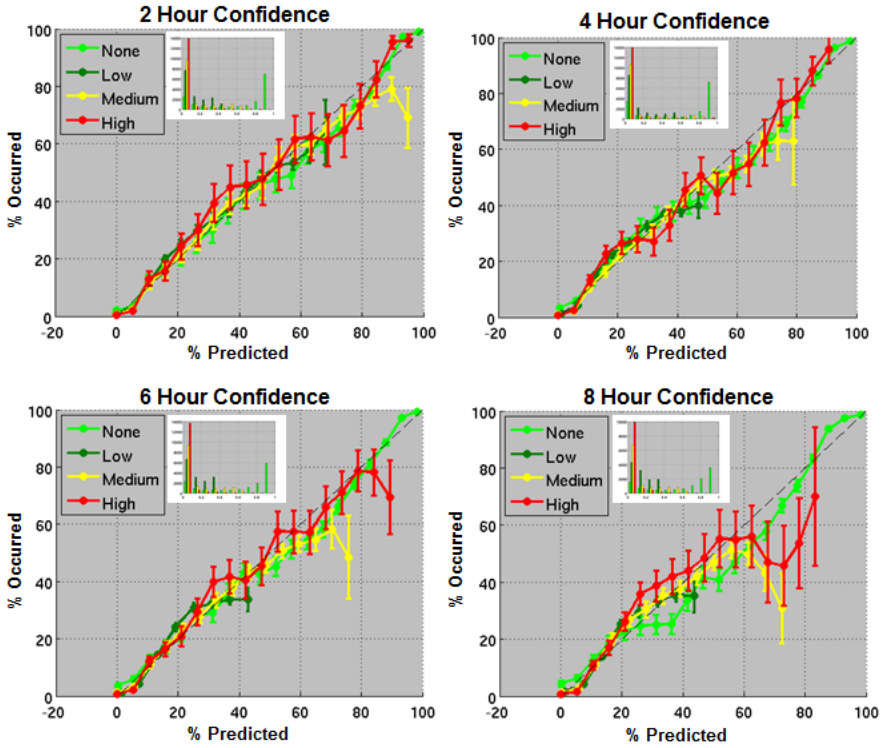


Figure 5: Reliability diagrams for forecast confidence at 2, 4, 6 and 8 hours. These plots show the predicted probability generated by the forecast confidence algorithm along the x-axis, and the associated rate of occurrence along the y-axis separately for each impact category (None, Low, Medium and High). Ideally, all curves would line up on the dashed line. These graphs show that the algorithm is able to generate accurate probabilities for the None (light green) and High (red) categories. The algorithm does not perform as well with the transitional categories. Low (Dark Green) and Medium (yellow) past 2 hours as those reliability curves fall away from the dash line for larger probabilities.

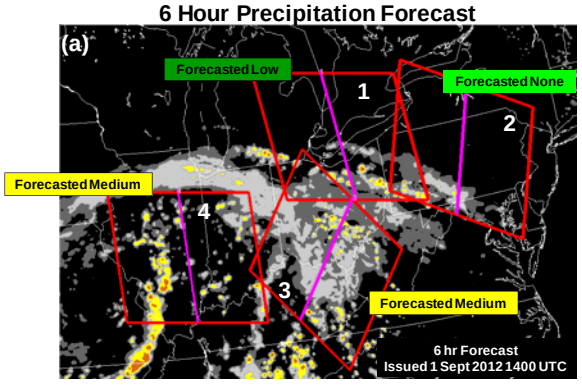
4. FORECAST CONFIDENCE CASE STUDY

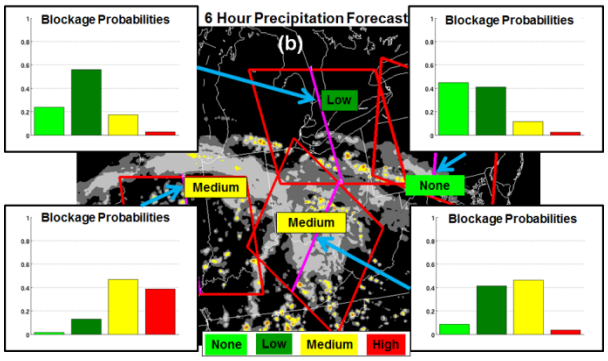
To demonstrate the forecast confidence algorithm, Figure 6 shows a CoSPA forecast with weather impacting enroute flows. The images in Figure 6a show the 6 hour VIL forecast issued on 1400 UTC 1 September 2012. The forecast confidence algorithm was run on four FCAs (shown in Fig. 6a) positioned over the Northeast containing a number of east-west routes. The label above each of the four boxes provides the forecasted blockage category. From the

deterministic forecast it appears that there will be significant weather impacts along the southern routes, but how confident should a user be in this 6 hour forecast?

Figure 6b shows the probabilities estimated for the four impact categories in each FCA. The most likely category is shown as the category with the highest probability, along with the probabilities of the remaining blockage categories. In all of the FCAs, the most likely category is also the forecasted category, though this is not always the case. For FCA 1, the confidence in blockage category Low is 57%, for FCA 2 the confidence in blockage category None is 45%, for FCA 3 the confidence in blockage category Medium is 46%, and for FCA 4 the confidence in blockage category Medium is 48%. In all FCAs except for FCA 1, the most likely category is followed closely by a second category that is almost as likely, indicating the uncertainty in the blockage forecast, and a possible alternate scenario.

Figure 6c shows the observed weather and blockage category. The forecasted blockage verified in three of the four FCAs, and the southern routes were indeed disrupted by weather. In FCA 4 located over IL, the convective line verified further north than was forecasted and had higher Echo Tops (not shown). As a result, High impact was observed, while only Medium impact was forecasted. Observe that for FCA 4, the forecast confidence model showed High was the second most likely category, suggesting





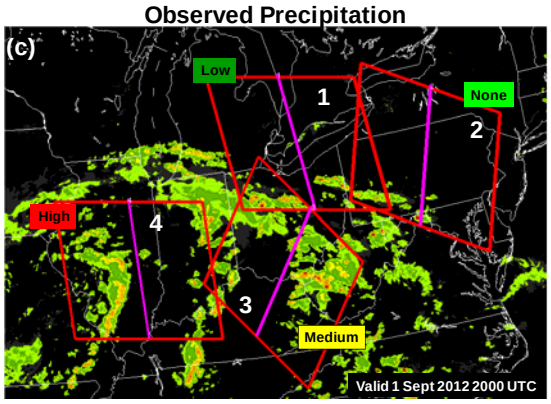


Figure 6: (a) Six hour CoSPA precipitation forecast valid 20 UTC 1 September 2012 with the forecasts blockage categories in four (numbered) sample FCA regions. (b) Histograms of the probability of occurrence for each blockage category within the FCAs. Note that in the FCA over IL, the most likely blockage category is medium, yet the second most likely category is high blockage. (c) Observed precipitation and blockage. The forecasted blockage was correct in 3 of the four cases. In the FCA 4, the second most likely forecasted category (high) occurred.

that the potential of higher impacts needed to be considered in a decision, even though this was not depicted in the single deterministic forecast.

5. SUMMARY

We have developed a capability to assess the confidence of a deterministic 8 hour aviation forecast at the time of forecast issuance. The capability provides the confidence in weather impacts in Flow Constrained Areas (FCAs), and is designed for strategic traffic flow

management in the NAS. The algorithm uses features in the current forecast, such as intensity and scale of forecasted VIL, blockage features computed from a Route blockage model, features extracted from a time-lagged ensemble of VIL forecasts, auxiliary features such as time of day and geographic location, and a historical data base to output the forecast confidence in four impact categories: No Impact, Low Impact, Medium Impact, or High Impact. The confidence in the forecasted category is provided, along with the probability of observing the other impact categories so that the users can assess forecast uncertainty and alternate impact scenarios. An example of forecast confidence involving an actual deterministic forecast was presented.

An initial quantitative assessment of the capability was performed. The probabilities generated by the confidence algorithm showed reasonable reliability for predicting the occurrence of each impact category, especially for the operationally-important None and High impact categories. The algorithm is less reliable with the Low and Medium impact categories. Future work will involve developing display concepts that allow for easy interpretation of forecast confidence for traffic flow managers, and preparing the capability for transfer to the NextGen Weather Processor.

6. REFERENCES

DeLaura, R., Robinson, M., Pawlak, M. L., Evans, J. E., “Modeling Convective Weather Avoidance in Enroute Airspace,” 13th Conference on Aviation, Range, and Aerospace Meteorology (ARAM), New Orleans, LA, Amer. Meteor. Soc., 2008

DeLaura, R. A., Lin, Y-H., Jordan, R. K., Venuti, J. C., Evans, J. E., “Evaluation of Consolidated Storm Prediction for Aviation (CoSPA) 0–8 Hour Convective Weather Forecast Using the Airspace Flow Program Blockage-based Capacity Forecast ("The Matrix")”, Project Report ATC-385, MIT Lincoln Laboratory, Lexington, MA, 2011

Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." ICML. Vol. 96. 1996.

Gibbons, William, et al. “Understanding Convective Weather Forecast Uncertainty Needs of ATM.” General Dynamics, Technical Report DOT/FAA/DTFAWA-10-00029, Task Order #6 (2013).

Iskenderian, H., C. Reiche, W. Dupree, M. Wolfson, T. Langlois, D. Morse, X. Tao, K. Haas, L. Bickmeier, P. Lamey, J. Pelagatti, and D. Moradi, J. Pinto, J. Williams, D. Ahijevych, and M. Steiner, S. Weygandt, C. Alexander, S. Benjamin, J. Mecikalski, W. Feltz, and K. Bedka. Update on CoSPA Storm Forecasts. 15th Conference on Aviation, Range, and Aerospace Meteorology, Los Angeles, CA, 1-4 August 2011. (available for download at www.ll.mit.edu/mission/aviation/publications/publication-files/ms papers/Iskenderian_2011_ARAM_WW-23958.pdf)

Pinto, J., W. Dupree, S. Weygandt, M. Wolfson, S. Benjamin, and M. Steiner, 2009: Advances in CoSPA. *14th Conference on Aviation, Range and Aerospace Meteorology, American Meteorological Society,* Atlanta, GA. (available for download at http://ams.confex.com/ams/90annual/techprogram/paper_163811.htm)

Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." Advances in large margin classifiers 10.3 (1999): 61-74.

Seiffert, Chris, et al. "RUSBoost: A hybrid approach to alleviating class imbalance." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 40.1 (2010): 185-197.

Wolfson, M. M. and D. Clark, 2006: Advanced Aviation Weather Forecasts, *Lincoln Laboratory Journal*, Vol. 16, Number 1. 31-58.

(available for download at http://www.ll.mit.edu/publications/journal/pdf/vol16_no1/16_1_3Wolfson.pdf)

Wolfson, M.M., W. J. Dupree, R. Rasmussen, M. Steiner, S. Benjamin, and S. Weygandt, 2008: Consolidated Storm Prediction for Aviation (CoSPA), *AMS 13th Conference on Aviation, Range, and Aerospace Meteorology*, New Orleans, LA. (available for download at ams.confex.com/ams/pdfpapers/132981.pdf)

Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259.