Application of Artificial Intelligence to Operational Real-Time Clear-Air Turbulence Prediction

Jennifer Abernethy[1,2], Robert Sharman[1], Elizabeth Bradley[2]

National Center for Atmospheric Research[1]
3450 Mitchell Lane, Boulder, CO 80301
University of Colorado at Boulder[2]
430 UCB UC-Boulder, Boulder, CO 80309
aberneth@ucar.edu

Abstract

Turbulence prediction is an important challenge to the aviation community because accurate forecasts are critical for the safety of the millions of people who fly every year. This paper details work in applying two AI techniques, support vector machines and logistic regression, to clear-air turbulence prediction. We show not only improved forecast accuracy over the current product performance, but also complete feasibility as part of a real-time operational turbulence forecasting system.

Introduction

The main challenges in predicting the weather are insufficient computational power and gaps in our understanding of the complex dynamics of atmospheric phenomena. There are comparatively straightforward solutions to these problems: enough teraflops, the right equations. But what happens when one has neither? This is the problem facing aviation turbulence forecasters, who are charged with predicting turbulent conditions that would affect aircraft, but who have neither the computational resources to predict it explicitly nor a complete understanding of how to derive it accurately from available meteorological data. Yet, commercial and private aviation communities expect accurate, timely turbulence forecasts.

Turbulence forecasting is an important challenge to the aviation community because while severe turbulence is rare, predicting it correctly is critical for the safety of the millions of people who fly commercial and private aircraft every year. Although fatalities are low, 65% of all weather-related commercial aircraft incidents can be attributed to turbulence encounters, and major carriers estimate that they receive hundreds of injury claims and pay out ``tens of millions" per year (Sharman et al., 2006).

Turbulence exhibits structure at all scales, all of which trade energy with one another in complicated ways, and numerical methods simply cannot keep up. Clear-air turbulence (CAT) forecasting is particularly challenging, because this phenomenon is invisible to both the eye and radar (unlike convective turbulence in/near thunderstorms, for instance). Faced with simulations that are too coarse to truly resolve the behavior that is of interest, plus sparse, subjective observations of 'light' or 'moderate' turbulence reported by pilots (PIREPs; further description of PIREPs and their limitations as a data source can be found in Schwartz (1996) and Abernethy et al. (2006)), the numerical weather prediction community reasons about large-scale quantitative atmospheric data and qualitative PIREPs in order to identify regions where aircraft-scale eddies are likely to form. The goal is to produce an automated system that detects rare events but does not over-predict them.

The current automated turbulence forecasting system, funded by the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP) and used by the National Oceanic and Atmospheric Administration's Aviation Weather Center (NOAA/AWC), integrates qualitative and quantitative data using fuzzy logic to produce a forecast. This tool, called Graphical Turbulence Guidance (GTG), was developed by the National Center for Atmospheric Research (NCAR) and NOAA's Global Systems Division (NOAA/GSD).

Recently, a new, better source of turbulence observations, termed *in-situ data*, has become available. In-situ data are sensor data from aircraft: measures of atmospheric eddy dissipation rate (Cornman et al., 2004). While the study of CAT is necessarily limited to that directly experienced by aircraft since it cannot be seen, in-situ data is so much more plentiful than PIREP observations that researchers now have enough data to explore additional AI techniques for forecasting. The specific goal of the project described in this paper is to intelligently exploit this new data source in a forecasting system using artificial intelligence techniques. We present two methods, support vector machines (SVM) and logistic regression — each combined with a wrapper

method for feature selection — for potential usage in the next version of GTG. We compare these two AI techniques by their improvement in forecasting accuracy over the current GTG. Obviously, better data should improve a forecast. Because of the complexity of the software, system and verification process, however, there are significant challenges involved. These challenges — of predicting turbulence itself, the requirements of an operational product, and how both of these affect the use of artificial intelligence techniques —are discussed throughout.

Clear-Air Turbulence Heuristics

Through the years when forecasts were done manually, forecasters developed "rules of thumb" about what atmospheric conditions typically indicated turbulence. These rules of thumb were an attempt to link the available large-scale meteorological data and the micro-scale CAT that was the subject of the forecast (Hopkins, 1977). Forecasters later quantified these rules, creating CAT *diagnostics*. A CAT diagnostic is a simple turbulence model (equation) calculated from numerical weather prediction (NWP) model data. For instance, a major cause of CAT is the Kelvin-Helmholtz instability: when gravity waves become steep and unstable, they may break into a chaotic motion (Dutton and Panofsky, 1970). This typically happens in areas of strong vertical shear (the difference in velocity between horizontal layers) and low local Richardson number (Ri, the ratio of static stability and wind shear), so many CAT diagnostics involve shears and Ri. There are many different diagnostics linking a large-scale condition to small-scale turbulence. Their predictive powers vary, depending upon the large-scale condition that each represents and how directly it is linked to turbulence. A full explanation of the forty CAT diagnostic equations can be found in Sharman et al. (2006).

Forecasters use these CAT diagnostics by mapping their values to different turbulence severity levels. As an example, low Ri indicates high turbulence. Early on, forecasters determined some unofficial thresholds to quantify the severity of turbulence that corresponded to a given diagnostic value — "Ri < 0.25 = moderate or greater turbulence," for example (Dutton & Panofsky, 1970). In this way forecasters were able to transform their qualitative knowledge to a quantitative form that could be used in automated systems. GTG developers used several years' worth of PIREPs to develop threshold values for each diagnostic, mapping them to different levels of PIREP turbulence severity. PIREPs range from 'smooth' (0) to 'extreme' (7), with 'moderate' being intensity 3. This mapping allows the diagnostics to work neatly with the qualitative PIREP observations.

Semi-Quantitative Observation Data

In-situ turbulence measurements are sensor data that are recorded by special software on commercial aircraft every minute during flight. Detailed coverage of in-situ data and the associated data-acquisition methods can be found in Cornman et al. (2004). An in-situ measurement is a measurement of the eddy dissipation rate (EDR) around an aircraft. Eddies are irregular currents of air, and the rate at which eddies break down is recognized as a good measure of atmospheric turbulence intensity (Panofsky and Dutton, 1983). Compared to PIREPs, in-situ data are more objective, more accurate, more plentiful, and more representative of the actual distribution of turbulence in the atmosphere (Dutton, 1980 and Sharman et al., 2006).

Currently, in-situ measurements of EDR are being gathered from 197 United Airlines aircraft. Several other airlines will deploy the data-gathering system in the coming year. An example of this data is shown in Figure 1.

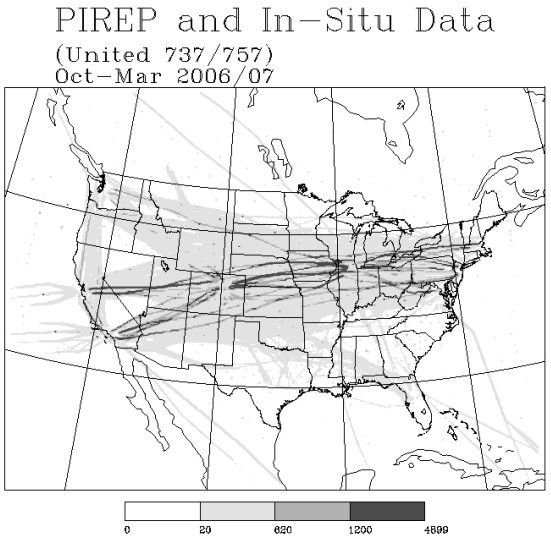

Figure 1. An example of the geographic distribution of in-situ data. PIREP data are included in this plot, though they are all but invisible under dense in-situ data along United flight paths. Color indicates frequency of observations.

Each in-situ data report is a location triple (latitude, longitude, altitude) and a median and peak (95th percentile) EDR reading from measurements taken over the corresponding minute. Each of the two EDR fields is binned and the two binned values are combined to reduce transmission costs. The binning turns otherwise continuous quantitative observation data into a set of eight discrete values that are cognate to the eight PIREP intensity levels. Currently we consider bin 4 to correspond to a 'moderate' PIREP, although study is ongoing (Abernethy et al. 2006).

Methodology

Our initial application of AI techniques to operational turbulence prediction consisted of testing Support Vector Machine (SVM) and logistic regression algorithm performance over our entire prediction domain, the continental U.S. ("CONUS"), and comparing their forecasting accuracies. For each algorithm, for both zero-hour and six-hour forecasts, we used a subset selection search to pick a subset of CAT diagnostics which together had the highest forecast accuracy. We then tested the performance of each model in a simulated operational real-time system using either a static model for each hour's forecast or dynamic training of the model using the previously-chosen subset. The following subsections summarize our data and methods.

AI Techniques

There are many choices of AI techniques for this task; here, we chose SVMs because they are good general classifiers and can give probabilistic output. We chose logistic regression for its similarity to the current GTG algorithm in its use of weights, its speed of computation, and its probabilistic output. Future product versions will need to produce probabilistic forecasts.

For brevity, we refer the reader to background on the SVM classification technique in Hsu et al. (2003). For implementation of the SVM, we used the LibSVM library (Chang and Lin, 2003). From previous studies (Abernethy, 2005), we know the radial basis function kernel, with parameters $C = 2$ and $\gamma = 8$ and probabilistic output, gives good performance for our domain. Background on the technique of logistic regression can be found in Hosmer and Lemeshow (1989). Although logistic regression produces probabilities, we used its and the probabilistic SVM's outputs as turbulence intensities on a scale of (0,1) in order to compare to deterministic (0,1) intensity forecasts of the current GTG product.

Performance Metrics

It is not trivial to assess the accuracy of a forecast because we do not know the 'truth'; we must use available observation data, however flawed or irregular (while in-situ data might have less random error than do PIREPs, the data are still spatially and temporally irregular since they exist only where/when airlines fly). We followed the verification practices of Takacs et al. (2005), which include the Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) (Marzban 2004), and True Skill Score (TSS), because these are the metrics by which our forecasting product will be measured when deployed operationally.

Recall from the subsection "Semi-Quantitative Observation Data" that both PIREPs and binned in-situ data have eight intensity levels and that we currently consider an in-situ bin 4 to be most similar in intensity to a 'moderate' pilot report. Bin 4 defines the moderate or greater (MOG) threshold, with values below bin 4 part of the class of less than moderate (LTM) observations. A ROC curve measures how well an algorithm discriminates between two classes such as MOG and LTM. To construct the curve, we vary the threshold that separates these two intensity classes over a (scaled) range of 0 to 1 and measure the discrimination accuracy at each threshold. Two numbers are used to capture this: PODY, "probability of detecting a yes" (forecast made a correct positive (MOG) prediction), and PODN, which corresponds to a correct negative (LTM) prediction. Higher PODY/PODN combinations over the range of thresholds implies greater classification accuracy, so the AUC is a useful single-number metric for forecast accuracy. The TSS considers PODY and PODN at one threshold (such as bin 4) : TSS = PODY + PODN – 1.

Data

Data used in the work described here consist of weather model and observation data – both PIREPs and in-situ data – from October through March 2006/7, shown in Figure 1. The weather model is Rapid Update Cycle (RUC) model at 13km resolution, run operationally and disseminated every hour by the National Center for Environmental Prediction. RUC model data was used to calculate forty CAT diagnostics for each RUC model grid point (see CAT Heuristics subsection) and observation data was matched by time and location to the forty diagnostics for a grid point. Only matches above 20000ft were used due to data quality issues and different mechanisms of turbulence below 20000ft.

Over 98% of the observations were of LTM turbulence. The distribution of the data used during the training process is a very important factor in the ability of a classifier to discriminate between the two classes (Japkowicz, 2000). Classifiers aim for the lowest overall error rate; one could simply classify everything as LTM and have a less than 2% error. This is well-supported in the literature (Japkowicz (2000), Weiss and Provost (2001), Wu and Chang (2005)). To work well, the training data set must have a large number of examples from each class. We found that rebalancing the training data such that 40% of the data were of MOG and 60% were LTM produced stable results: these proportions resulted in the best SVM classification rate in an earlier study of SVMs with CAT diagnostics and in-situ data (Abernethy, 2005). We did this by keeping all the MOG observations and choosing LTM observations randomly to be 60% of the set. We found 20% MOG and 80% LTM to be a good distribution for logistic regression training data.

Analysis of the data reveals that PIREPs dominate the MOG category (>92%) and in-situ data dominates the LTM category (>98%). Thus, PODY is effectively a measure of the algorithm's ability to predict PIREPs and PODN becomes a measure of in-situ prediction capability. We know using only in-situ data to train the algorithm improves performance (Abernethy et al. 2006). However, our forecasting product will be verified using PIREPs (at least in part), thus we would be foolhardy not to use the same data and metrics with which the FAA will decide its fate.

Subset Selection Search

Turbulence forecasting, in its current state, is essentially the task of classifying atmospheric indicators of turbulence: the forecast reflects the number of diagnostics which indicate turbulence in an area. While it might seem obvious to simply use the individually best-performing diagnostics for forecasting, as was done with GTG, that approach allows one to possibly miss a different set of diagnostics that might perform better, as a group, than the set of the *individually* top-ranked diagnostics (Kohavi (1995,1997), Guyon and Elisseef (2003)). Our search for the best subset of diagnostics is essentially the task of *feature subset selection* (Guyon and Elisseef, 2003). We are faced with the choice between 40 diagnostics, knowing that some may not improve our current forecasting accuracy. In addition, it is infeasible to calculate and use all 40 in a real-time operational system. The wrapper method in feature subset selection executes a state space search for a good feature subset, estimating prediction accuracy using an induction algorithm – here, we used SVMs and logistic regression (Kohavi and Sommerfield, 1995). Using the induction algorithm output, we calculated TSS as the accuracy metric. We used a simple hillclimbing search. Each state is a subset of diagnostics, and the search operator is "add a diagnostic". The search chooses the best addition to the current subset based on the classification skill (TSS) of the induction algorithm using the current subset plus an additional diagnostic. This approach to the search is called *forward selection*. Thus, we start with an empty subset and added diagnostics stepwise; our stopping condition was no further classification performance improvement. Searches were performed for SVM and logistic models for both zero and six-hour forecasts using training, testing and holdout data sets from 18Z over winter 2006/7.

Simulated Real-Time Operational System

We have created a simulated real-time forecasting system capable of using either SVMs or logistic regression to create a turbulence forecast every hour for the CONUS. The system trains a model for every forecast hour or uses a pre-trained model so that we may test performance differences between dynamic and static weighting, respectively. For both, we use the sets of diagnostics found in the searches explained above. In this paper, we present results from trials over the fifteen-day period of 2/1/2007 to 2/15/2007. Thus far we have concentrated on zero-hour forecasts in this step. We also applied our dynamically trained models to the entire RUC grid – since GTG uses dynamic weighting, also – to make a full forecast in order to assess geographic accuracy and get an idea of amount of turbulence predicted as compared to the current GTG forecasting system.

	AUC	TSS	Subset size
GTG 0hr fcsts	*0.795*	*0.390*	*10*
Log search 0hr	0.801	0.478	13
SVM search 0h	0.7825	0.471	8
GTG 6hr fcsts	*0.78*	*0.366*	*10*
Log search 6hr	0.79	0.467	6
SVM search 6h	0.78	0.4643	12
GTG 0-hr 15days	*0.799*	*0.350*	*10*
Log static 0-hr	0.823	0.466	13
SVM static 0-hr	0.796	0.459	8
Log dyn. 0-hr	0.786	0.45	13
SVM dyn. 0-hr	0.775	0.464	8

Table 1. Area under the Curve, True Skill Score, and subset size results for feature selection searches and 0-hr 15-day real-time simulation runs. GTG skills for the same data are in italics. Higher TSS and AUC indicate greater skill.

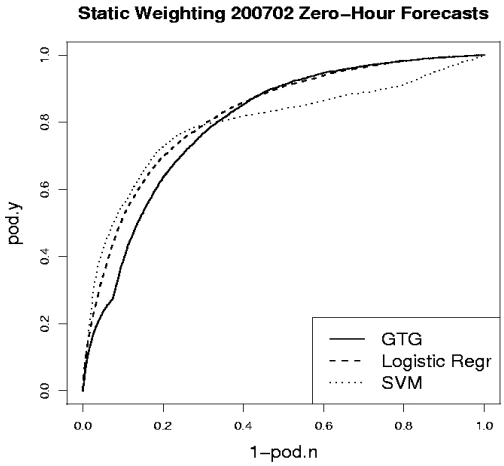

Figure 2. Receiver operating characteristic (ROC) curves comparing performance for 15-day real-time simulation of 0-hr forecasts using static weighting. The solid line is the current GTG performance for the same 15-day period. Lines closer to top left corner indicate better forecasting performance. See Table 1 for areas under the curves.

The LibSVM library did not come ready to handle such large real-world data sets. Since LibSVM uses ascii files, 13km-resolution gridded RUC data caused each forecast to take over an hour. To mitigate this, we built a NetCDF file format interface onto the library and replaced the

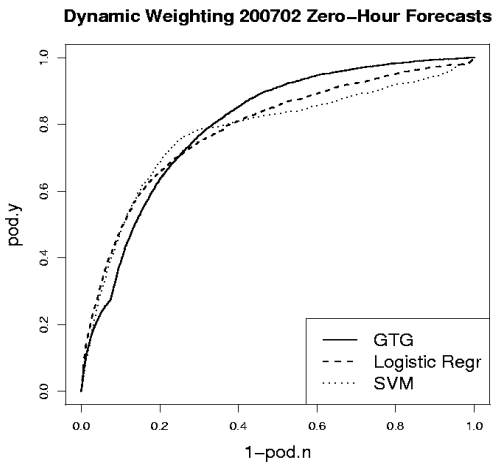

Figure 3. Like Figure 2, for dynamically-trained models.

exponential function with an approximation. Both changes cut the forecast time down to a more operationally appropriate five minutes.

Results

Results of our forward selection subset searches and real-time simulations are shown in Table 1. While we do not list the exact diagnostics chosen by each search for the sake of simplicity, we did find that there was significant — though not complete — overlap in the diagnostics chosen by each search, indicating high predictive capability for a core subset of four or five diagnostics. Logistic regression shows a small improvement in AUC over the overall performance of the current GTG algorithm for both 0 and 6-hr forecasts (about a 0.01 difference), however, the true-skill scores (TSSs) for both algorithms are significantly improved over GTG (0.09 – 0.1 improvement). This is most likely due to the fact that our search used TSS as the heuristic to choose the sets of diagnostics.

Figures 2 and 3 show the ROC curves for our static- (model trained in the search step is applied to data from each hour) and dynamic-weighting (new model is trained every hour) 15-day real-time simulations. It should be noted that GTG has been tuned using years of PIREPs, thus its PODY scores are highest (since PIREPs dominate the PODY category). Logistic regression using pre-determined (static) weights improves significantly upon the current GTG product, increasing the AUC from 0.799 to 0.823 and the TSS from 0.35 to 0.466. While the static-weighting SVM and both dynamically-weighted models had similar improvements in TSS over GTG, we saw no improvement in AUC. TSS is discrimination skill at the MOG threshold, 0.375; AUC measures classification skill at many thresholds. Thus, we have improved forecasting performance at the operational MOG threshold, although

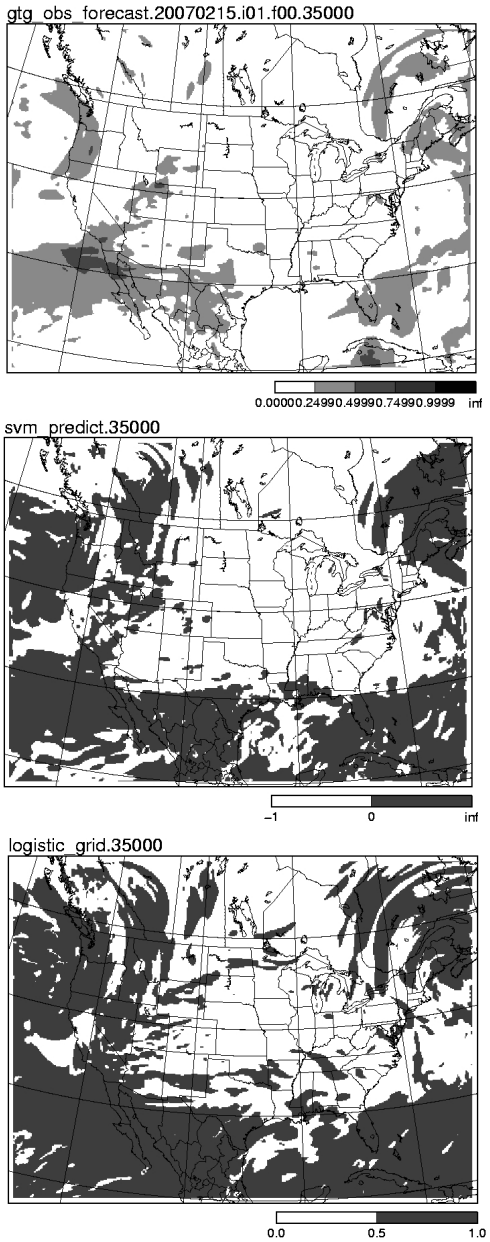

Figure 4. Comparison of results produced by GTG (top), SVM (middle) and regression (low). Note different color scale for GTG.

the ROC curves show us that there is still need for improvement in the algorithms overall.

An example of the graphical display for a zero-hour forecast (2/15/2007 at 1Z at 35000ft) of the forecasting product is shown in Figure 4. Note that GTG predicts low, moderate, severe and extreme turbulence categories, while the current SVM and regression implementations just discriminate between MOG and LTM — though their probabilistic outputs could also be used to define intensity thresholds. While both methods predict more turbulence

than does GTG, especially in areas not covered by in-situ data — indicating need for more algorithm tuning — they capture similar general patterns of turbulence.

Conclusions

Forecasting clear-air turbulence is critical to aviation safety. AI techniques can be very useful in meeting the challenges inherent in this process because they smoothly handle sparse, noisy data sets, significant levels of uncertainty, and gaps in the understanding of the underlying physical mechanisms — all of which are characteristics of the turbulence-prediction domain. This paper has detailed the first steps in applying the artificial intelligence techniques of support vector machines and logistic regression to clear-air turbulence forecasting, with promising results. While the GTG product uses fuzzy logic, past algorithmic choices were limited by the sparse PIREP observation data; now, the more objective and plentiful in-situ data vastly widens the choices for prediction algorithms. We have shown not only improvement in forecasting performance for static weighted models using new subsets of CAT diagnostics found by feature subset selection, but also feasibility of implementing these AI algorithms in a real-time operational product setting. Currently, logistic regression outperforms SVMs, and static weighting outperforms the dynamically-weighted modeling approach, although further tuning of the algorithms, training data sets, and a longer test period — all planned next steps — could make the differences more clear and further improve performance. Other future work includes continued study of these algorithms for regionally-specific forecasting and probabilistic forecasting.

References

Abernethy, J., 2005. Domain Analysis Approach to Clear-Air Turbulence Forecasting Using In-situ Data. Ph.D. diss. prop., Department of Computer Science, Univ. of Colorado. Boulder, CO.

Abernethy, J.; Bradley, E.; Sharman, R. 2006. Qualitative Reasoning About Small-Scale Turbulence in an Operational Setting. *In Proceedings of the 21st Qualitative Reasoning Workshop*. Hanover, NH.

Chang, C. and C. Lin. LIBSVM – A Library for Support Vector Machines.http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cornman, L.; Meymaris, G.; Limber, M., 2004. An Update on the FAA Aviation Weather Research Program's *In situ* Turbulence Measurement and Reporting System. Preprints, *Eleventh Conf. on Aviation, Range, and Aerospace Meteorology*, Hyannis, MA: Amer. Meteor. Soc., P4.3.

Dutton, M. J. O. 1980. Probability Forecasts of Clear-Air Turbulence Based on Numerical Output. *Meteor. Mag.* 109: 293-310.

Dutton, J., and Panofsky, H. 1970. Clear Air Turbulence: A Mystery May be Unfolding. *Science* 167: 937-944.

Guyon, I. and Elisseef, E. 2003. An Introduction to Variable and Feature Selection. *J. Machine Learning Research*, 3, 1157-1182.

Hopkins, R. H., 1977. Forecasting Techniques of Clear-Air Turbulence Including That Associated with Mountain Waves. WMO Technical Note No. 155, 31 pp.

Hosmer, D. and Lemeshow, S. eds. 1989. *Applied Logistic Regression.* John Wiley and Sons, Inc.

Hsu, C.; Chang, C.;Lin, C. 2003. A Practical Guide to Support Vector Classification. Published with LibSVM documentation: http://www.csie.ntu.edu.tw/~cjlin/libsvm**.**

Japkowicz, N. 2000. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*, Menlo Park, CA.

Kohavi, R., and Sommerfield, D. 1995. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. *First International Conference on Knowledge Discovery in Data Mining (KDD-95).* Montreal, Quebec, Canada.

Kohavi, R. and John, G. 1997. Wrappers for Feature Subset Selection. *J. Artificial Intelligence*, 97(1-2): 273-324.

Marzban, K. 2004. The ROC Curve and the Area Under It as Performance Measures. *Weather and Forecasting*, 19: 1106-1114.

Panofsky, H and J. Dutton, eds. 1983. *Atmospheric Turbulence: Models and Methods for Engineering Applications*. John Wiley & Sons.

Schwartz, B., 1996. The Quantitative Use of PIREPs in Developing Aviation Weather Guidance Products. *Weather and Forecasting*, 11: 372-384.

Sharman, R.; C. Tebaldi; G. Wiener; J. Wolff, 2006. An Integrated Approach to Mid- and Upper-Level Turbulence Forecasting. *Weather and Forecasting*, 21(3): 268-287.

Takacs, A.; L. Holland; R. Hueftle; B. Brown; A. Holmes. 2005. Using In-situ Eddy Dissipation rate (EDR) Observations for Turbulence Forecast Verification. NCAR Report to the FAA Aviation Weather Research Program.

Weiss, G. and F. Provost. 2001. The Effects of Class Distribution on Classifier Learning: An Empirical Study. Technical Report ML-TR-44, Department of Computer Science, Rutgers University.

Wu, G and E. Chang. 2005. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Transactions on Knowledge and Data Engineering,* 17(6): 786-795.