1.4 SUPPORT VECTOR MACHINES FOR REGIONAL CLEAR-AIR TURBULENCE PREDICTION

Jennifer Abernethy*[1,2] and Robert Sharman[2]
[1]University of Colorado, Boulder
[2]National Center for Atmospheric Research*
Boulder, Colorado

ABSTRACT

Clear-air Turbulence (CAT) is a significant safety issue for aviation at upper levels in the atmosphere. Since CAT is not observable with traditional remote sensing techniques, it is particularly difficult to avoid. The current FAA-sponsored CAT forecasting product, the Graphical Turbulence Guidance System (GTG), calculates many indicators or diagnostics of CAT potential from larger-scale numerical weather prediction model output and compares them to current turbulence observations from pilots (PIREPs). It then combines the diagnostics using a global optimization technique to provide the final CAT forecast. Theory suggests that many CAT diagnostics may vary in their predictive skill depending on the geographic region, but GTG is unable to exploit these regional dependencies due to an insufficient number of timely PIREPs. Recently, more plentiful and objective observation data have become available from the In-situ Turbulence Observation System. This system is currently installed on about 200 United Airlines' aircraft and provides data at one minute intervals. This high-resolution data now allows the development of CAT forecasts on a regional scale. For each region of the continental U.S. (determined by CAT climatology), we have used the machine learning classification technique of Support Vector Machines to determine the best subset of CAT diagnostics that together have the highest forecasting performance, regardless of the diagnostics' individual performances. To search efficiently through the state space of all feature subsets, we used a forward selection algorithm; the search is guided by a five-fold cross validation method on test sets of in-situ observation data. The results of the regional CAT forecasts determined in this manner are shown to provide better skill than the current GTG algorithm. This approach will ultimately be used to replace the current operational GTG CAT forecasting system.

1. INTRODUCTION

Pilots' ability to avoid turbulence during flight affects the safety of the millions of people who fly commercial airlines and other aircraft every year. Of all weather-related commercial aircraft incidents, 65% 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 can occur in clouds or in clear air. At upper levels, clear-air turbulence, or CAT, is particularly hard to avoid because it is invisible to traditional remote sensing techniques. One seasoned pilot noted that CAT was his “greatest worry” when flying (Salby, 2006). In order to plan flight paths to avoid turbulence, air traffic controllers, airline flight dispatchers, and flight crews must know where CAT pockets are likely to be. The dynamical scales in which CAT appears, however, are far finer than those of any current weather model. And observations of the state of the system – reports radioed in by pilots who encounter CAT – are sparse and subjective. For these reasons, no currently available CAT forecasts meet the Turbulence Joint Safety Implementation Team's (TJIST) recommended >0.8 probability of moderate-or-greater (MOG) turbulence detection and >0.85 probability of null turbulence detection.

The turbulence forecasting difficulty is due to two main factors: (1) turbulent eddies at the scales that affect aircraft (~100m) are a microscale phenomenon and NWP models cannot resolve that scale, and (2) lack of objective observational turbulence data. The prior factor has been addressed during the past 50 years, by assuming that most of the energy associated with turbulent eddies at aircraft scales cascades down from larger scales of atmospheric motion (Dutton and Panofsky (1970), Koshyk et al. (2001), Tung et al.(2003)). The turbulence forecast problem then becomes one of linking large-scale features resolvable by NWP models to the formation of aircraft-scale eddies. Numerous `rules of thumb' empirical linkages, termed turbulence *diagnostics*, were developed by the National Weather Service, airline meteorologists and academic researchers. The forecast skills of these

* The National Center for Atmospheric Research is sponsored by the National Science Foundation
*Corresponding author address: Jennifer Abernethy, University of Colorado, Department of Computer Science, 430 UCB, Boulder, CO 80309, email: aberneth@cs.colorado.edu

diagnostics depend on the forecaster (for manual forecasts) and diminish with lead time; none meet the TJIST recommendations, either alone or used together in any current implementation. The diagnostics' skills reflect in part researchers' imperfect understanding of the atmospheric processes involved.

The imperfect nature of the current diagnostics leads forecasters to depend, at least partially, on available turbulence observations. Until recently, the only available observations were pilot reports (PIREPs), and they are the second factor contributing to the difficulty of turbulence forecasting (and forecast verification). PIREPs are sparse, aircraft-dependent, subjective reports by pilots of turbulence encountered during flight. Sharman et al. (2006) shows that PIREP inaccuracy is not as large as once thought (Schwartz, 1996), however, the distribution of reports is not representative of the state of the atmosphere because most non-turbulent areas are not reported.

One major effort by the FAA's Aviation Weather Research Program (AWRP), some major airlines, and the National Center for Atmospheric Research's Research Applications Laboratory (NCAR/RAL) is the development of a better turbulence observation data source: in-situ data of eddy dissipation rate (EDR) (Cornman et al. 1995, 2004). In this system turbulence observations are recorded automatically every minute during cruise by on-board software. It addresses many of the faults of PIREPs: it is aircraft-independent, objective, less sparse, and is designed to be used quantitatively. Not only does it offer higher-resolution observations, but it also helps alleviate the inconsistent null-turbulence reporting issues with PIREPs (Takacs et al., 2005).

While the in-situ measurement and reporting system is still in its first and limited deployment, it is being incorporated already into the next release of NCAR/RAL's CAT forecasting system, the Graphical Turbulence Guidance System (GTG). However, the GTG algorithm was developed using PIREPs, and thus is designed to make the most of sparse and subjective observational data. Not surprisingly, simply adding in-situ data into the current algorithm results in only a modest improvement in forecasting accuracy (Kay et al. 2006). The authors believe that in order to fully exploit the potential of in-situ data, a new approach or forecasting algorithm is needed. This paper presents the initial work in using a machine-learning technique, support vector machines, to reevaluate the forecasting accuracy of CAT diagnostics using in-situ data. In addition, the high volume of in-situ data is used to begin looking at regional differences in diagnostics' forecasting accuracies, in an effort to further improve forecasting accuracy.

2. IN-SITU DATA

In-situ turbulence measurements are data recorded by special software on commercial aircraft during flight. This measurement and reporting system was developed at NCAR under FAA sponsorship in order to augment or replace PIREPs with a data source that has more precise location and intensity data. Insitu measurements use existing aircraft equipment and are reported using existing communications networks. Detailed coverage of in-situ data methods can be found in Cornman et al. (1995, 2004).

The in-situ-derived turbulence metric is the eddy dissipation rate (EDR), $\varepsilon^{1/3}$. EDR is recognized as an objective measure of atmospheric turbulence intensity (Panofsky and Dutton, 1983). Two methods to estimate $\varepsilon^{1/3}$ onboard aircraft were developed: the accelerometer-based method and the vertical wind-based method. Both are aircraft-independent measurements, and both result in approximately the same turbulence measurements.

Currently, only the accelerometer-based method is in use, in United Airlines 737 and 757 aircraft. Southwest Airlines and Delta Airlines are scheduled to use the wind-based method when the system is deployed in their aircraft, which is expected to happen by the end of the year.

EDR data is reported once a minute except during takeoff and landing, when data is reported more frequently depending on rate of altitude change. Each in-situ data report is a location (latitude, longitude, and altitude) and a set of statistics about various turbulence levels calculated from a number of EDR measurements taken onboard during that minute.

The set of statistics are the median eddy dissipation rate (medEDR) and the maximum

eddy dissipation rate (maxEDR). Reporting just these two fields reduces transmission costs while still providing a way to distinguish between discrete and continuous turbulence events. The medEDR is the median value of a time series. The maxEDR value is the 95% value of the time series; as a protection measure against erroneous data, peak values are not used.

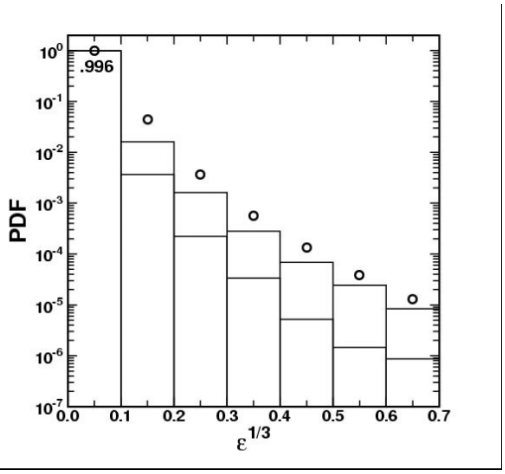

Figure 1. Taken from Sharman et al. (2006). This figure shows the probability distribution function (PDF) of three months of observed EDR values ($\varepsilon^{1/3}$) in each in-situ bin, both median (lower bar) and 95th percentile (upper bar). The open circles are estimates of the true lognormal distribution of turbulence based on the RUC20 model (Frehlich & Sharman 2004). The fact that observed EDR distribution differs from the estimated distribution may reflect the ability of commercial air carriers to avoid some turbulence during flight.

Due to transmission costs, both values are binned into 1 of 8 bins, and each possible pair of maxEDR/minEDR values for a minute is mapped to a single 8-bit character and then downloaded to the ground. The number of bins was limited by the available character sets, but a newer version of the algorithm now in development compresses the EDR data to enable more bins and thus a higher resolution of data. Currently, in-situ data is being downloaded from 89 United Airlines 757 aircraft. The software is installed on 96 757s and 101 737s. Figure 3 shows the geographic distribution of in-situ data over winter 2005-2006.

In-situ data provides a better representation of turbulence statistics in the atmosphere (Dutton (1980), Sharman et al. (2006)). Figure 1 shows that over 99% of in-situ reports are reports of null turbulence. If this distribution is representative, at any time at most 0.01% of the atmosphere at upper levels should contain MOG turbulence. In contrast, about half of PIREPs report null turbulence, 27% report light, 17% report moderate and 1% report severe; thus, pilots substantially underreport the null events. In-situ data overcomes this uncertainty by reporting data every minute during flight.

The effort to understand in-situ intensity values relative to PIREP intensities is ongoing. For instance, is a 0.45 reading moderate or severe turbulence? Comparisons to qualitative PIREPs encounter many problems such as PIREP location and time errors, and overall lack of PIREPs. A main problem is the fact that a pilot makes a report of his overall impression of the turbulent event, while in-situ data are measurements every minute; a turbulent event can span multiple minutes. How to match a series of in-situ data to one PIREP continues to be studied. Initial comparisons used the reading with the highest intensity in the event, defined as a consecutive series of 2nd-bin or higher in-situ readings (0.15 or higher), as representative of the event's severity. This value was compared to a PIREP, if there was one, from the same flight, within 40km, five minutes and 1000ft of the in-situ reading. The lack of PIREPs severely limited the number of matches – only 328 between August 2004 and November 2005 - but 2nd bin in-situ values (0.15) roughly corresponded to light/moderate PIREPs (intensity 2) and 3rd bin in-situ values (0.25) roughly corresponded to moderate PIREPs (intensity 3). There were too few matches at higher in-situ bins to draw any conclusions.

We defined MOG turbulence as 0.25 reading - 3rd in-situ bin - or higher. This is based on the PIREP and in-situ data comparisons, and that GTG considers a PIREP of intensity 3 or higher to be MOG.

3. CLEAR-AIR TURBULENCE DIAGNOSTICS

A clear-air turbulence diagnostic is a simple turbulence model (equation) derived from qualitative expert knowledge based on experience or from basic physical principles. Through the years when forecasts were done manually, forecasters developed ``rules of thumb'' about what atmospheric conditions typically indicate turbulence. These rules of

thumb were an attempt to link the large-scale meteorological data that was available and the micro-scale CAT that was the subject of the forecast (Hopkins, 1977). Forecasters later quantified these rules, creating CAT *diagnostics*. For instance, a major cause of CAT is thought to be the Kelvin-Helmholtz instability (Dutton and Panofsky, 1970). This typically happens in areas of strong vertical shear and low local Richardson number (Ri, the ratio of static stability and wind shear). Thus many qualitative CAT diagnostics concern 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. There are forty CAT diagnostics; the diagnostics cited in this paper are detailed in Appendix A.

Forecasters use these diagnostics by mapping their values to different turbulence severity levels. In this way, forecasters took their qualitative knowledge about large-scale atmospheric conditions and their relationship to small-scale turbulence, quantified it in the form of diagnostic equations, then interpreted the results using thresholds to produce a qualitative forecast. The GTG forecasting system does exactly the same thing. Its authors used several years' worth of PIREPs to develop threshold values for each diagnostic that map to different levels of PIREP turbulence severity. With the newly available in-situ data, we now have the opportunity to reevaluate these diagnostics' forecasting ability.

4. METHODOLOGY

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 (2003)).

Specifically, we were trying to determine which subset of diagnostics will give the most accurate turbulence forecast, using the new in-situ data for verification. This emphasis on *group* performance differs from the method used to pick the set of diagnostics for GTG correlation. The method used to pick the GTG set (Sharman et al. (2000, 2004)) evaluated each diagnostic's prediction accuracy in GTG (measured by TSS, see Section 4.3), and then formed a ranking of all the diagnostics. Next, diagnostics were evaluated for inter-correlation; of correlated pairs, the lowest-ranking diagnostic was replaced. Then, prediction accuracy was measured using only the highest-ranked diagnostic. Next, the second highest-ranked diagnostic was added back in, and the performance was measured; if it increased, the next diagnostic was added, until it was found that the diagnostic addition negatively affected the performance of GTG. This method resulted in the set of ten diagnostics chosen for GTG: Frntth, ETI1, TempG, SATRi, CP, EDRS10, NCSU1, DTF3, SIGW10, and UBF (see Appendix A). In this study, we instead looked for the best performing diagnostics *as a group*, regardless of their individual prediction accuracies. Results from Sharman et al. (2000) show that no single diagnostic can produce a more accurate forecast than can multiple diagnostics together, supporting this multiple-diagnostic approach.

4.1 *Support Vector Machines*

The Support Vector Machine (SVM) is a popular machine learning technique for classification. Generally, a classifier is an algorithm that predicts a data classification given (presumably) relevant data features. The SVM produces a model that predicts the class label by setting parameter values of an optimization problem based on its input data (Hsu et al., 2003)

In order to learn the relationships (parameter values) between these data features and the class label, we first train a classifier by giving it many known feature/class pairs. Each pair is known as a data instance. A data instance k consists of a set of features $x_{i,k}\ \ i = 1...n$ (in our case, the n diagnostics) and a target class label y (turbulence or no turbulence).

The SVM is trained on many data instances called a training set. The SVM prediction accuracy is estimated using a test set of data instances with known class labels which were not used during training. Using a test set instead of the training set for accuracy estimation better

reflects the SVM's ability to classify unknown data.

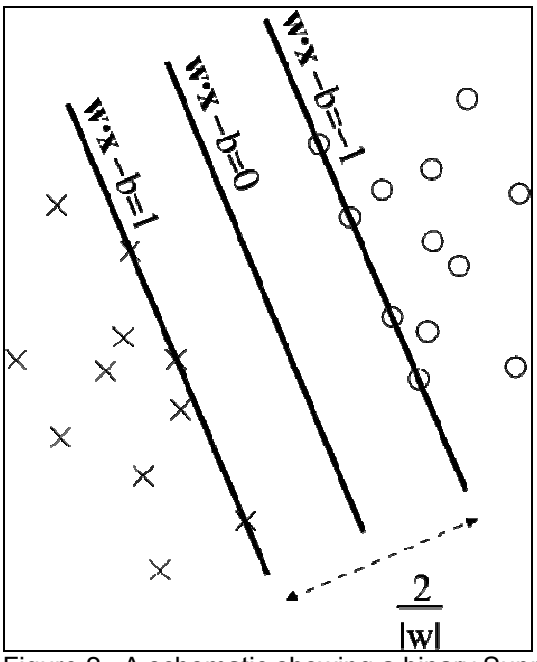

Figure 2. A schematic showing a binary Support Vector Machine classifier with a linearly separating hyperplane. The data points on the margin lines are the support vectors.

During training, each feature vector X_k is mapped into a higher dimensional space. The SVM finds a linearly separating hyperplane with the maximal margin between class means in this higher dimensional space. A schematic of this hyperplane and margins for a binary classifier is shown in Figure 2.

To classify an example, the SVM calculates the distance of that example to each class mean through a series of dot products, and classifies it in whatever class has the closest mean (Chen et al., 2003). This series of dot products is at the heart of the model and is a measure of vector similarity called a kernel function:

$$K(x_i, x_j) = \phi(x^T{}_i)\phi(x_j)$$

For implementation of the SVM, we will use the LibSVM library (Chang and Lin, 2003). LibSVM provides four basic kernels and an optional program that selects the model (i.e., does a parameter search). From previous studies (Abernethy, 2005) we know that the radial basis function kernel gives good performance for our domain:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0$$

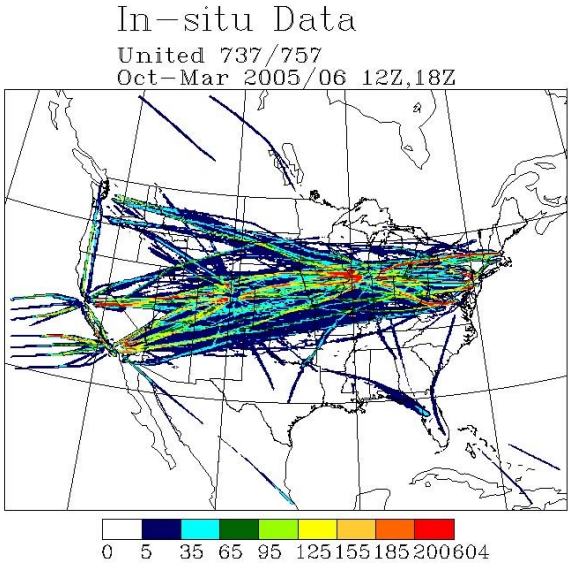

Figure 3. Geographic distribution of the in-situ data used in this study.

The radial basis function kernel only has two parameters: γ and C, a penalty parameter for the SVM error term. From the previous study, the LibSVM program chose $\gamma = 8$ and C=2, so those values are used here.

4.2 *Data*

This study used data from winter 2005-2006 (October – March), since there are more CAT events during winter (Sharman et al.,2000). The National Center for Environmental Prediction's Rapid Update Cycle model at 13km resolution (RUC13) provided the environmental data to calculate 40 CAT diagnostics at every grid point (Sharman et al., 2006). Diagnostics were calculated for hours 12Z, 15Z, 18Z and 21Z, at analysis time (zero-hour forecast) and the six-hour forecasts. Diagnostics were matched by location and hour on the RUC13 grid to in-situ data from the In-Situ Reporting System. If there was more than one in-situ reading in a grid box during the hour, only the highest intensity reading was used. Thus, one in-situ observation was matched to 40 diagnostics at a grid point. Only data at FL200 (20000ft) and higher were included, since the in-situ data was only available at these heights. The geographic distribution of the in-situ data used is shown in

Figure 3. The full winter contained 2855084 matches for the

analysis times, and 2527596 matches for set six-hour forecast times.

To find the best subset of the 40 CAT diagnostics, we executed a forward search through the space of all subsets (Kohavi and Sommerfield, 1995), using SVMs as the evaluation function. At each step, an SVM is trained on training data containing only the current subset of diagnostics and their in-situ observation matches.

Analysis-time diagnostics/in-situ matches were used to train the SVMs. The distribution of the data used during the SVM training process is a very important factor in the ability of a classifier such as SVMs to discriminate between the two classes (Japkowicz, 2000). SVMs aim for the lowest overall error rate. In our case, where in-situ data is over 99% null observations, an SVM could simply classify everything as null and have a less than 1% overall error rate. We found this to be true in preliminary tests and it is well-supported in the literature (Japkowicz (2000), Wiess and Provost (2001), Chen et al. (2004), Wu and Change (2005)). To work well, the training data set must have a large number of examples from each class. The best proportion of examples from each class to have in a training set is case-dependent. For cases such as ours, this distribution requirement means altering the distribution of the data in the training set, rather than having the training set be a representative sample from the available in-situ data. There are multiple methods for creating a new training set with acceptable proportions of MOG reports and null reports. The methods applicable to this project include altering the kernel, increasing the number of MOG reports, or decreasing the number of null reports. To increase the number of MOG reports, we could synthetically create more that look statistically similar to real MOG reports. Decreasing the number of null reports (to increase the proportion of MOG reports) means simply not including some percentage of the null reports in a training set (but including all MOG reports). Here, the latter method was chosen. Since the in-situ data set is more than 99% null turbulence (0.05, 1st bin), we rebalanced the training data such that 40% of the data were of Moderate-or-Greater (MOG) turbulence, and 60% were null (less than MOG) turbulence. We did this by keeping all the MOG observations and choosing null observations randomly to be 60% of the set. This proportion of MOG/nulls resulted in the best SVM classification rate in an earlier study of SVMs with CAT diagnostics and in-situ data (Abernethy, 2005).

4.3 *Search*

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. 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. 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 performance of an SVM 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.

At each step, sets of training data, testing data, and holdout data were generated containing only the current subset of diagnostics plus the proposed addition to that set. Training data consisted of the set of analysis-time observation/diagnostic matches, and the test and holdout sets consisted of the set of six-hour observation/diagnostic matches (divided between the two files). An SVM was trained on the training data using 5-fold cross-validation, and the resulting model was tested on the testing data, outputting overall classification accuracy.

For comparison with GTG evaluations we wanted the classification accuracies of both classes – MOG and null – to weigh equally in the estimated prediction accuracy used to choose the next node expansion. The classification accuracy given by LibSVM reflects the number of samples in each class, which was 40% MOG and 60% null. We added an extra step wherein we took the classification accuracy

of each class and factored them equally into the final assessment:

True Skill Score (TSS) = MOG classification accuracy + Null classification accuracy -1

Thus, -1 < TSS < 1.

TSS is part of the scoring function in GTG (Sharman et al., 2006). To establish a baseline, we first ran the search over data from the entire U.S (see Figure 3). We then divided the U.S into two geographic regions, one to the west of 100W meridian, and one to the east of it, and executed independent searches on both regions in order to see if diagnostics performed differently in different areas. We plan to further refine and divide regions in the near future, but for this study, we have simply isolated the mountainous terrain, and the mountain-wave turbulence, in the west region. When the hillclimbing searches terminated, a final TSS was calculated from the chosen subsets' classification performances on the holdout data set.

5. RESULTS

Our subset searches yielded sets of diagnostics with higher TSSs than that of the GTG combination, TSS = 0.453 (Sharman et al., 2006). Our baseline search, using data from the entire U.S., yielded a TSS of 0.463 using the diagnostic subset ETI1, STABinv, AGinv, netRiTW, TempG, and SIGW10 (see Appendix A). For our west region, the best set found was ETI1, ETI2, STABinv, PVORT, ABSIA, AGinv, TempG, SPEED, negNVA, SIGW10 with a TSS of 0.465. The east region search resulted in the highest TSS overall, 0.562, using the diagnostics CP, ETI1, Frntth, and UBF.

While some of the diagnostics in the chosen subsets are also in the GTG combination, our study found that other diagnostics, such as AGinv and STABinv, appear to work well as part of a group despite having a lower individual forecasting accuracy (and thus not being chosen as part of the GTG combination). These initial results support our group performance approach. In addition, the fact that different diagnostics were chosen in the east and west regions indicate that diagnostics can perform differently in different areas of the country, reflecting the geographic differences in the large-scale atmospheric processes they represent.

The number of diagnostics that differ between the GTG combination and those our SVMs chose is larger than expected. We can attribute this at least in part to the difference in the algorithms - SVM versus GTG's fuzzy logic algorithm- and the evaluation functions: True Skill Score versus area under the ROC curve (Sharman et al. 2006)), although these two are similar. Our initial assumption was that the GTG set of diagnostics, due to their high individual prediction accuracies, would also have high classification accuracies using an SVM; a forward search through the GTG set should find that all ten diagnostics produce the highest TSS. However, this was not the case. We executed a hillclimbing search using only the GTG set of and found that it terminated at a set of three diagnostics: ETI1, TempG, and SIGW10. These differences will require further investigation.

6. FUTURE WORK

Our initial study supports the idea that developing specialized forecasts for different regions of the CONUS (Continental U.S.) can improve overall turbulence forecasting accuracy. Our next steps are to develop several geographic regions that may further improve forecasting accuracy with their own sets of diagnostics, and to explore regionalizing the forecast by altitude. While SVMs provided a general classification algorithm for this study, other algorithms such as random forests may be suitable, also. We also intend to improve the search itself by using a best-first search which has shown to improve search results (Kohavi and Sommerfield, 1995). In addition, we must devise a way to merge all the regional forecasts together to make one coherent CAT forecast for the CONUS.

7. REFERENCES

Abernethy, J., 2005: Domain Analysis Approach to Clear-Air Turbulence Forecasting Using In-situ Data. Dissertation Proposal, Department of Computer Science, University of Colorado.

Bluestein, H. B., 1992: *Synoptic-Dynamic Meteorology in Midlatitudes, Vol. I.* Oxford Univ. Press, 431 pp.

Buldovskii, G. S., S. A. Bortnikov, and M. V. Rubinshtejn, 1976: Forecasting zones of intense turbulence in the upper troposphere. *Meteorologiya i Gidrologiya*, **2**, 9-18.

Chang, C. and C. Lin. LIBSVM – a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, C., A. Liaw and L. Breiman, 2004: Using Random Forests to Learn Imbalanced Data. *Technical Report 666. Statistics Department. University of University of California at Berkeley.*

Chen, P., C. Lin and B. Scholkopf, 2003: A tutorial on v-support vector machines. http://kernel-machines.org.

Colson, D., and H. A. Panofsky, 1965: An index of clear-air turbulence. *Quart. J. Roy. Meteor. Soc.*, **91**, 507-513.

Cornman, L. B., C. S. Morse, and G. Cunning, 1995: Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *J. Aircraft*, **32,** 171-177.

Cornman, L., G. Meymaris, and M. Limber, 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 H. A. Panofsky, 1970: Clear Air Turbulence: A mystery may be unfolding. *Science*, **167**, 937-944.

Ellrod, G. P., and D. L. Knapp, 1992: An objective clear-air turbulence forecasting technique: Verification and operational use. *Wea. Forecasting*, **7**, 150-165.

Endlich, R. M., 1964: The mesoscale structure of some regions of clear-air turbulence. *J. Appl. Meteor.,* **3**, 261-276.

Frehlich, R., and R. Sharman, 2004a: Estimates of turbulence from numerical weather prediction model output with applications to turbulence diagnosis and data assimilation. *Mon. Wea. Rev.*, **132**, 2308-2324.

Frehlich, R., and R. Sharman, 2004b: Estimates of upper level turbulence based on second order structure functions derived from numerical weather prediction model output. Preprints, *Eleventh Conf. on Aviation, Range and Aerospace Meteorology,* Hyannis, MA, Amer. Meteor. Soc., P4.13.

Guyon, I. and A. Elisseef, 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.

Hsu, C., C. Chang and C. Lin, 2003: A practical guide to support vector classification. Published online with Libsvm documentation at **http://www.csie.ntu.edu.tw/~cjlin/libsvm**.

Jacobi, C., A. H. Siemer, and R. Roth, 1996: On wind shear at fronts and inversions. *Meteorol. Atmos. Phys.*, **59**, 235-243.

Japkowicz, N., 2000: Learning from imbalanced data sets: a comparison of various strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*, Menlo Park, CA.

Kay, M., J. Henderson, S. Krieger, J. Mahoney, L. Holland and B. Brown, 2006: Quality assessment report: Graphical turbulence guidance (gtg) version 2.3.

Kaplan, M. L., K. M. Lux, J. D. Cetola, A. W. Huffman, J. J. Charney, A. J. Riordan, S. W. Slusser, Y.-L. Lin, and K. T. Waight, 2004: Characterizing the severe turbulence environments associated with commercial aviation accidents. A Real-Time Turbulence Model (RTTM) designed for the operational prediction of hazardous aviation turbulence environments. NASA/CR-2004-213025, 54 pp.

Knox, J. A., 1997: Possible mechanism of clear-air turbulence in strongly anticyclonic flows. *Mon. Wea. Rev.,* **125**, 1251-1259.

Knox, J. A., 2001: The breakdown of balance in low potential vorticity regions: Evidence from a clear air turbulence outbreak. Preprints, *Thirteenth. Conf. on Atmospheric and Oceanic Fluid Dynamics*, Breckenridge, CO, Amer. Meteor. Soc., 64-67.

Koch, S. E., and F. Caracena, 2002: Predicting clear-air turbulence from diagnosis of unbalance flow. Preprints, *Tenth Conf. on Aviation, Range,*

and Aerospace Meteorology, Portland, OR, Amer. Meteor. Soc., 359-363.

Kohavi, R., and D. Sommerfield, 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).*

Kohavi, R. and G. John, 1997: Wrappers for Feature Subset Selection. *J. Artificial Intelligence*, **97**, *no1-2, 273-324.*

Koshyk, J. N., and K. Hamilton, 2001: The horizontal energy spectrum and spectral budget simulated by a high-resolution troposphere-stratosphere-mesosphere GCM. *J. Atmos. Sci ,* **58***, 329-348.*

Kronebach, G. W., 1964: An automated procedure for forecasting clear-air turbulence. *J. Appl. Met.,* **3,** 119-125.

Lane, T. P., J. D. Doyle, R. Plougonven, M. A. Shapiro, and R. D. Sharman, 2004: Observations and numerical simulations of inertia-gravity waves and shearing instabilities in the vicinity of a jet stream. *J. Atmos. Sci*., **61**, 2692-2706.

McCann, D. W., 2001: Gravity waves, unbalanced flow, and aircraft clear air turbulence. *National Weather Digest*, **25**, 3-14.

Marroquin, A., 1998: An advanced algorithm to diagnose atmospheric turbulence using numerical model output. Preprints, *Sixteenth Conf. on Weather Analysis and Forecasting*, Phoenix, AZ, Amer. Meteor. Soc., 79-81.

O'Sullivan, D., and T. J. Dunkerton, 1995: Generation of inertia-gravity waves in a simulated life cycle of baroclinic instability. *J. Atmos. Sci*., **52**, 3695-3716.

Panofsky, H and J. Dutton, 1983: *Atmospheric turbulence: models and methods for engineering applications*. John Wiley & Sons.

Salby,M 2006. Personal Communication.

Schwartz, B., 1996: The quantitative use of PIREPs in developing aviation weather guidance products. *Wea. Forecasting*, **11**, 372-384.

Sharman, R., G. Wiener and B. Brown, 2000: Description and verification of the NCAR integrated turbulence forecasting algorithm. *Proceedings of the 38th Aerospace Sciences Meeting and Exhibit, Reno, NV.*

Sharman, R., J. Wolff, G. Wienter and C. Tebaldi, 2004: Technical description document for the graphical turbulence guidance product v2 (gtg2). *Technical report submitted to FAA for AWRP turbulence PDT project.*

Sharman, R., C. Tebaldi, G. Wiener and J. Wolff, 2006: An Integrated Approach to Mid- and Upper-Level Turbulence Forecasting. *Weather and Forecasting*,

Shapiro, M. A., 1978: Further evidence of the mesoscale and turbulence structure of upper level jet stream-frontal zone systems. *Mon. Wea. Rev.,***106**, 1100-1111.

Stull, R. B., 1988: *An introduction to boundary layer meteorology*. Kluwer Academic Publishers, 670 pp.

Takacs, A., L. Holland, R. Hueftle, B. Brown and A. Holmes, 2005: Using in-situ eddy dissipation rate (edr) observations for turbulence forecast verification.

Tung, K. K., and W. W. Orlando, 2003: The k^{-3} and $k^{-5/3}$ energy spectrum of atmospheric turbulence: Quasigeostrophic two-level model simulation. *J. Atmos. Sci*., **60**, 824-835.

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.*

Appendix A. GTG turbulence diagnostics

This Appendix is a partial list of the current suite of turbulence diagnostic algorithms. Note that in some cases the constituent components of a diagnostic may itself be used as a turbulence index.

1. Richardson number and its components (e.g., Endlich 1964, Kronebach 1964, Dutton and Panofsky, 1970, etc). Theory and observations have shown that at least in some situations clear-air turbulence patches are produced by Kelvin-Helmholtz (KH) instabilities. This occurs when the Richardson number (*Ri*) becomes small. Therefore, theoretically, regions of small *Ri* should be favored regions of turbulence, where

$$Ri = \frac{N^2}{S_V^2} \qquad \text{(A1.1)}$$

with $N^2 = \frac{g}{\theta}\frac{\partial \theta}{\partial z}$ or $\frac{g}{\theta_e}\frac{\partial \theta_e}{\partial z}$ (A1.2a, A1.2b)

and $S_V = \left|\frac{\partial \mathbf{v}}{\partial z}\right| = \left(\left|\frac{\partial u}{\partial z}\right|^2 + \left|\frac{\partial v}{\partial z}\right|^2\right)^{1/2}$ (A1.3)

A1.1 with A1.2b is the SatRi diagnostic. Equation 1.2a is the STAB diagnostic, and this study used the inverse of STAB (STABinv). Here θ is potential temperature, θ_e is equivalent potential temperature, g is the acceleration due to gravity, z is the vertical direction, and **v** is the horizontal wind vector with components u,v in the east-west and north-south directions respectively.

2. Turbulent kinetic energy (tke) formulations. These are based on the tke balance equation, assuming horizontal homogeneity and stationarity. The Colson-Panofsky index (Colson and Panofsky 1965) uses dimensional arguments in a stable atmosphere to estimate clear-air turbulence intensities as

$$CP = \lambda^2 S_V^2 \left(1 - \frac{Ri}{Ri_{crit}}\right), \qquad \text{(A2.1)}$$

where λ is a length scale, taken as the local value of vertical grid increment Δz, and Ri_{crit} is an empirical constant (≈ 0.5).

Marroquin (1998) DTFs (Diagnostic TKE Formulations) used k-ε closure equations (e.g. Stull 1988) and other simplifications to derive diagnostics for tke and/or ε, giving e.g. for DTF3,

$$\varepsilon = K_M \left(\frac{c_1}{c_3} S_V^2 - \frac{c_2}{c_3}\frac{N^2}{\mathrm{Pr}}\right) \qquad \text{(A2.3)}$$

where c_1=1.44, c_2=1.0, c_3=1.92 (Stull 1988, p. 219) , and K_M and Pr are taken as adjustable constants to get best agreement with observations.

3. Eddy dissipation rates estimated from second-order structure functions (Frehlich and Sharman 2004a, 2004b). The structure function of variable q is defined as

$$D_q(s) = < [q(x) - q(x+s)]^2 >$$

where <> denotes an ensemble average. The structure functions of the velocity components parallel or normal to the displacement vector **s**=(x,y,z) can be related to turbulence intensity ε (for q=u,v) or σ_w^2 (for q=w, the vertical velocity component) through

$$D_q(s) \propto C_q(s)\, \varepsilon^{2/3}\, D_{REF}(s) \qquad \text{(A3.1)}$$

$$D_w(s) \propto C_w(s)\, \sigma_w^2 \qquad \text{(A3.2)}$$

where $C_q(s)$ and $C_w(s)$ take into account NWP model specific spatial filtering effects, and D_{REF} is given by Lindborg (1999); for small s it is proportional to $s^{+2/3}$. In the text the relation (A3.1) to derive $\varepsilon^{1/3}$ is indicated as "EDR" and relation (A3.2) to derive σ_w is indicated as "SIGW."

4. Frontogenesis function. Fronts contain regions of low *Ri* and therefore may be conducive to turbulence (e.g., Jacobi et al. 1996) and can also be a source of gravity waves that may be unstable (e.g., Lane et al. 2004). The definition of the frontogenesis function is (e.g., Bluestein 1992, vol. 2, p. 253)

$$F = \frac{D}{Dt}\left|\nabla\theta\right|$$

$$\text{where } \frac{D}{Dt} \text{ is the Eulerian time derivative.}$$

This can be rewritten in two dimensions using the thermal wind relation as

$$F \propto \frac{D}{Dt}\left[\left(\frac{\partial u}{\partial\theta}\right)^2+\left(\frac{\partial v}{\partial\theta}\right)^2\right]^{\frac{1}{2}} = \left|\frac{\partial \mathbf{v}}{\partial\theta}\right|^{-1}\left[\frac{\partial u}{\partial\theta}\frac{D}{Dt}\left(\frac{\partial u}{\partial\theta}\right)+\frac{\partial v}{\partial\theta}\frac{D}{Dt}\left(\frac{\partial v}{\partial\theta}\right)\right]$$

Expanding on a constant θ surface and invoking continuity gives

$$F_\theta \propto -\frac{D}{Dt}\left[\left(\frac{\partial u}{\partial\theta}\right)^2+\left(\frac{\partial v}{\partial\theta}\right)^2\right]^{\frac{1}{2}} = \left|\frac{\partial \mathbf{v}}{\partial\theta}\right|^{-1}\left[\frac{\partial u}{\partial\theta}\frac{D}{Dt}\left(\frac{\partial u}{\partial\theta}\right)+\frac{\partial v}{\partial\theta}\frac{D}{Dt}\left(\frac{\partial v}{\partial\theta}\right)\right] \qquad \text{(A4.1)}$$

This is the form used in GTG at upper-levels. Note that its formulation is based on an isentropic coordinate system (as used at upper-levels in the RUC model).

5. Ellrod indices (Ellrod and Knapp 1992), ETI1 and ETI2. These indices are derived from simplifications of the frontogenetic function. As such it depends mainly on the magnitudes of the potential temperature horizontal gradient (proportional to S_V through the thermal wind relation) and deformation. Two variants were developed:

$$TI1 = S_V\, DEF \qquad \text{(A5.1)}$$

$$TI2 = S_V\, (DEF - \Delta_H) \qquad \text{(A5.2)}$$

where $DEF = (D_{SH}^2 + D_{ST}^2)^{1/2}$. (A5.3)

6. Potential vorticity (PVORT) (Knox 2001) or horizontal gradient of PV (Shapiro 1978)

$$|PV| \qquad \text{(A6.1)}$$

$$|\nabla PV|, \qquad \text{(A6.2)}$$

where $PV = -g\zeta_a\frac{\partial\theta}{\partial p}$.

7.Horizontal temperature gradient, TempG. This is a measure of the deformation and also vertical wind shear from the thermal wind relation, and is routinely used by airline forecasters. It was also used in Buldovskii et al. (1976).

$$\left|\nabla_H T\right| = \left\{\left(\frac{\partial T}{\partial x}\right)^2+\left(\frac{\partial T}{\partial y}\right)^2\right\}^{1/2} \qquad \text{(A7)}$$

8. Wind related indices. Besides the speed vertical shear (A1.3), the wind speed

$$s = |\mathbf{v}| \qquad \text{(A8.1)}$$

may be related to turbulence. A8.1 is the SPEED diagnostic.

9. Unbalanced flow (UBF) diagnostics (Knox 1997, Knox 2001, McCann 2001, O'Sullivan and Dunkerton 1995, Koch and Caracena 2002). There is some evidence that regions of strong imbalance may be related to turbulence aloft (e.g., Knox 1997; McCann 2001; Koch and Caracena 2002). The UBF diagnostic formulation used within GTG was developed by Koch and Caracena (2002) and McCann (2001), and derives from the residual *R* of the nonlinear balance equation

$$R = -\nabla^2\Phi + 2J(u,v) + f\zeta - \beta u \qquad \text{(A9.1)}$$

where Φ is geopotential, *J* is the Jacobian operator, and β is the Coriolis frequency gradient.

Other unbalanced flow related diagnostics developed by McCann (2001) and used in a case study by Knox (2001) include

$$\text{ABSIA} = |\text{v}_i - \text{v}_c|^2, \qquad \text{(A9.2)}$$

where $\text{v}_\text{i} = |\mathbf{v} \cdot \mathbf{v}|\ /\text{f}$ and $\text{v}_\text{c} = \text{K}_\text{s}\ |\mathbf{v}|^2\ /\text{f}$

and

$$\text{AGI} = \zeta_{curv} + \text{f}/2, \qquad \text{(A9.3)}$$

with $\zeta_{curv} = \text{K}_s|\mathbf{v}|$

and where K_S is the streamline curvature.

10. North Carolina State University Index (NCSU1) is described in Kaplan et al. (2004), and was developed from investigations of several severe turbulence encounters:

$$NCSU1 = \frac{1}{MAX(Ri, 10^{-5})} MAX\left(u\frac{\partial u}{\partial x} + v\frac{\partial v}{\partial y}, 0\right)\left|\nabla\zeta\right| \qquad \text{(A10)}$$

11. Negative vorticity advection (NVA). A rule-of-thumb forecasting approach used by the airlines is to look for regions of large NVA computed as follows (Bluestein 1992, vol. 1, p. 335):

$$NVA = MAX\left\{\left[-u\frac{\partial}{\partial x}(\zeta + f) - v\frac{\partial}{\partial y}(\zeta + f)\right], 0\right\} \qquad \text{(A11)}$$