J1.3 TOWARDS A REGIONAL CLASSIFIER APPROACH TO AVIATION TURBULENCE PREDICTION

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

1. 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 you have neither? This is the problem facing aviation turbulence forecasters, who are charged with the task of 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.

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 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

* 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

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.

The specific goal of this project is to intelligently integrate this new data source. In this paper, we use artificial intelligence techniques to produce turbulence forecasts. We combine a wrapper method for feature selection with Support Vector Machines and logistic regression to produce turbulence forecasts. We tested both algorithms on a real-time system to compare forecasting accuracy. With these baseline results and initial regional forecasting results, we can begin to design a regional CAT forecasting system which intelligently uses all available observational and atmospheric data to produce a forecast.

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.

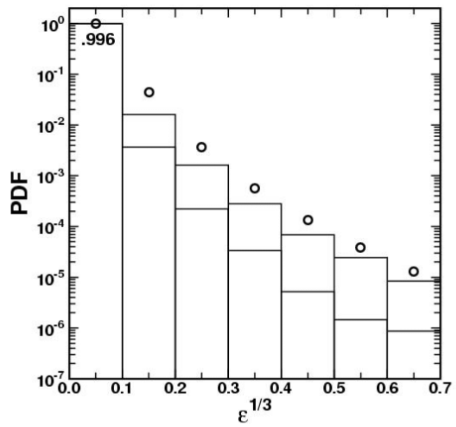

Figure 1. Taken from Sharman et al. (2006). This figure shows the probability distribution function (PDF) of three months of observed EDR values ($\varepsilon^{1/3}$) in each in-situ bin, both median (lower bar) and 95th percentile (upper bar). The open circles are estimates of the true lognormal distribution 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 2 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.

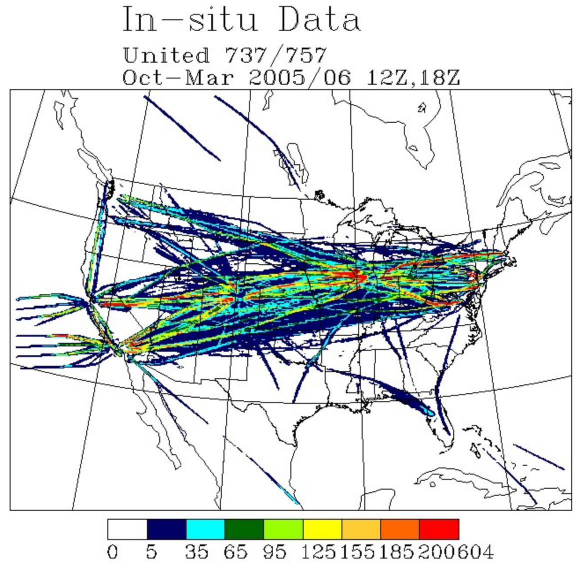

Figure 2. Geographic distribution of the in-situ data used in this study.

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 (Sharman et al. 2006).

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. Using fuzzy logic, GTG weights the diagnostics dynamically depending on their recent agreement with PIREPs, and the weighted values are combined to produce a turbulence forecast (Sharman et al. 2006).

4. METHODOLOGY

Background on the technique of Support Vector Machines can be found in Hsu et al. (2003). For implementation of the SVM, we will use the LibSVM library (Chang and Lin, 2003). Background on the technique of logistic regression can be found in Hosmer and Lemeshow (1989). Although logistic regression produces probabilities, we used its outputs as turbulence intensities on a scale of (0,1) in order to compare to deterministic forecasts of the current GTG and the SVM model.

To compare the two techniques to the current GTG algorithm, we first established baseline performance of GTG. We then implemented both SVM and logistic regression models as “global” (one forecast over the CONUS) forecasting systems to measure their global performance as compared to GTG, which uses one set of parameters (weights) to make a forecast for the entire CONUS. We looked at the performance of static (one model) versus dynamic training (training a new model each forecast time) of the algorithms, and implemented both in a realtime simulation system to measure performance and performance variability. Included are some initial regionalization results using SVMs.

4.1 *Data*

This study used data from winter 2005-2006 and 2006-2007 (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 several daytime hours at analysis time (zero-hour forecast) and the six-hour forecasts. Diagnostics were matched by location and hour on the RUC13 grid to PIREP and in-situ data from the In-Situ Reporting System. If there was more than one in-situ and or PIREP reading in a grid box during the hour, only the highest intensity reading was used. Thus, one 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 for the 2005-2006 winter is shown as an example in Figure 2. The full 2005-2006 winter contained over two million observation/diagnostic matches for each due to more planes reporting in-situ data by late 2006, the 2006-2007 winter contained over nine million observation/diagnostic matches for each of the zero-hour and six-hour forecast times.

4.2 *Feature Subset Selection Searches*

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)). 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.

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 and logistic regression (Kohavi and Sommerfield, 1995). 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 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.

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 (zero-hour forecast) observation/diagnostic matches, and the test and holdout sets consisted of either different zero-hour matches or the set of six-hour observation/diagnostic matches (divided between the two files).

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). SVMs, for instance, 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 Chang (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). We found 20% MOG and 80% nulls to be a good distribution for logistic regression training data.

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 both algorithms reflects the number of samples in each class. 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 a primary part of the scoring function in GTG (Sharman et al., 2006).

5. RESULTS

We used the winter 2006-2007 data to assess baseline GTG algorithm forecast accuracy. We found that the TSS for zero-hour and six-hour forecasts, respectively, were 0.35 and 0.31. They rose slightly for hour 18Z to 0.353 and 0.36, presumably due to more observations on which to base a forecast.

5.1 *Subset Searches*

Our first subset searches focused on zero-hour forecasts.

Our global subset searches yielded sets of diagnostics with higher TSSs than that of the current GTG. Our search using SVMs as the induction algorithm yielded a TSS of 0.501 on holdout data using the diagnostic subset CP, ET2, DTF3, DTF5, DIV, TrophTinv, TempG and ABSW.

Our subset search using logistic regression as the induction algorithm yielded a TSS of 0.467 on the set of holdout data. The diagnostics chosen in the logistic search were: B1, B2, -Ri, Frntg, LAZ, -NGM2, HS, STABinv, DefSQ, Vortsq, AG_inv, UBF, -SatRi, PVgrad, DIV, -RTW, TrophTinv, TempG, TropGovz, NCSU2, ABSW, RoL, EDRS10 and SIGW10.

While some of the diagnostics in the chosen subsets are also in the GTG combination, our study found that other diagnostics, such as STABinv and ABSW, 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 current GTG combination). These initial results support our group performance approach.

The number of diagnostics that differ between the GTG combination and those our machine learning techniques chose is larger than expected. We can attribute this at least in part to the difference in the algorithms 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.

5.2. *Realtime Simulation 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. The system uses past data, from 2006-2007 winter season. The system is running internally at NCAR/RAL for development only at this time. In addition, the system is capable of training a model for every forecast or using a pre-trained model so that we may test performance differences between dynamic and static weighting, respectively.

Using this real-time system, we ran the logistic regression algorithm over data from the month of February 2007, using the diagnostics chosen in the search detailed in 5.1. We tested zero-hour forecasts. Through preliminary trials, we found that logistic regression performed well using around seven hours of training data. Thus, in this real-time simulation, the system gathered the “previous” seven hours of data to train the model for predicting the “current” forecast. We tested both static and dynamic weighting of the logistic model. For the static weighting test, we used the set of weights (regression parameters) trained from the larger data set used in the subset search. For the dynamic weighting test, we used the same subset of diagnostics but trained a new logistic model for each forecast time. We found that on average there was almost no difference in the approaches; static weighting produced an average TSS of 0.46, and dynamic weighting produced an average TSS of 0.466.

We replicated the February 2007 real-time simulation of zero-hour forecasts using SVMs and the chosen diagnostic subset listed in 5.1. In general, the SVM models need more training data than do the logistic regression models. Through preliminary trials, we found that SVM performance increased from 0.4 to 0.46 as we increased the amount of training data from two days' to five days' worth. Performance stabilized and even decreased slightly with more than five days worth of training data. However, we have only tested the dynamic weighting option thus far. In this trial, the found that the SVM model, using five days of training data, produced an average TSS of 0.467. For further comparison, we reran the test using the original ten diagnostics used in the current GTG algorithm (see Sharman et al. 2006) and found an average TSS of 0.4. We also treated the current GTG forecast value as an additional diagnostic and added to the SVM subset (chosen by the search). This increased the average TSS to 0.499.

The next step in comparing SVM and logistic regression techniques to the current GTG algorithm is to compare the amount of forecasted turbulence and the spatial accuracy of the forecasts. Ultimately, GTG is a graphical tool for aviation. It is important to note that both SVMs and logistic regression models used thus far are only binary classifiers; the current GTG algorithm categorizes CAT into light, moderate and severe intensities. Nevertheless, Figure 3 shows a sample forecast for GTG, logistic regression and SVM, respectively. The SVM's spatial pattern is more similar to that of GTG, but both overforecast turbulence significantly.

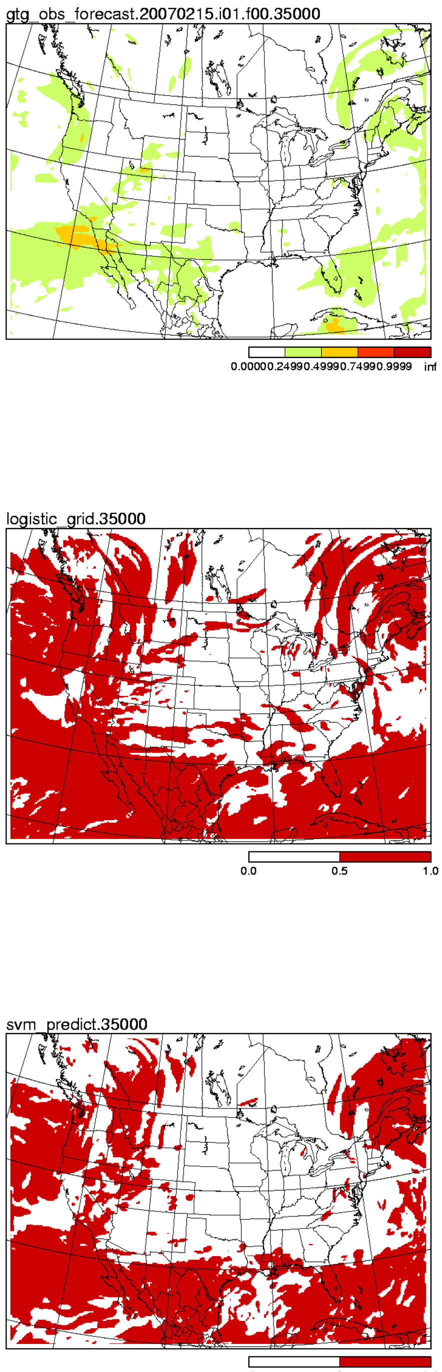

Figure 3. a)GTG forecast b)logistic regression forecast, and c) SVM zero-hour forecast for 01Z 2/15/2007 at 35000ft. Both (b) and (c) overforecast turbulence but do follow the same spatial pattern as (a).

REGION	TSS	SET OF DIAGNOSTICS
West	0.465	ETI1, STABinv,AGinv, netRI, TempG, SIGW10
East	0.562	CP, ETI1, Frntg, UBF
>=30000ft	0.447	ETI1, AB, TempG, Stone, SIGW10
<30000ft	0.607	ETI1, Ri, TempG, NCSU1, SIGW10
High west	0.441	NGM1, VWS, NCSU1, SIGW10
High east	0.516	Frntg, AGinv, DIV, RoL
Low west	0.614	ETI1, Frntg, AGinv, UBF, TempG
Low east	0.519	PVORT, TempG

Table 1. Sets of CAT diagnostics found for different regions of the CONUS by subset selection searches using TSS derived from SVMs as the heuristic function.

5.3 *Regionalization*

Thus far, we have conducted regionalization studies using SVMs only, on 2005-2006 data. We employed subset searches for each of these regions: west of 100W meridian, east of 100W, above and below 30000ft, and by both geography and altitude (e.g., east of 100W and below 30000ft: low east). 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. Results are in Table 1.

We found improvement in forecast accuracy in almost every region. In addition, the fact that different diagnostics were chosen in the different 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.

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. We have shown promise in our machine learning approaches globally, and plan to replicate our global model approach for six-hour forecasts globally and zero- and six-hour forecasts regionally. Our next steps are to develop an approach for defining 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 and logistic regression provided a general classification algorithm for this study, other algorithms such as random forests may be suitable, also. 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, 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.

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.

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

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**.

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.

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.

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*.

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.*