2.3 REGIONALIZATION OF CLEAR-AIR TURBULENCE FORECASTING USING MACHINE LEARNING

Jennifer Abernethy*[1,2]
[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). At upper levels, clear-air turbulence, or CAT, is particularly hard to avoid because it is invisible to traditional remote sensing techniques. 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 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 vary and diminish with lead time.

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, they remain a sparse and qualitative data source, and 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.

* The National Center for Atmospheric Research is sponsored by the National Science Foundation
*Corresponding author address: Jennifer Abernethy, NCAR, 3450 Mitchell Lane FL2 3054
Boulder, CO 80301 , email: aberneth@ucar.edu

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 a real but modest improvement in forecasting accuracy (Kay et al. 2006). The author believes 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 use in-situ data to improve turbulence forecasting accuracy. Plentiful in-situ data enables region-specific diagnostic selection and machine learning techniques for forecasting. In this paper, I present a method using random forests for determining climatologically-relevant forecasting regions, and explain the development and initial results from my Forecasting System Testbed (FST) for prototyping regionalized turbulence forecasting systems.

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. In-situ 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, the accelerometer-based method is used by United Airlines in 101 737s and 89 757s. Southwest Airlines and Delta Airlines are scheduled to use the wind-based method when the system is deployed in their aircraft, which is currently under way.

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.

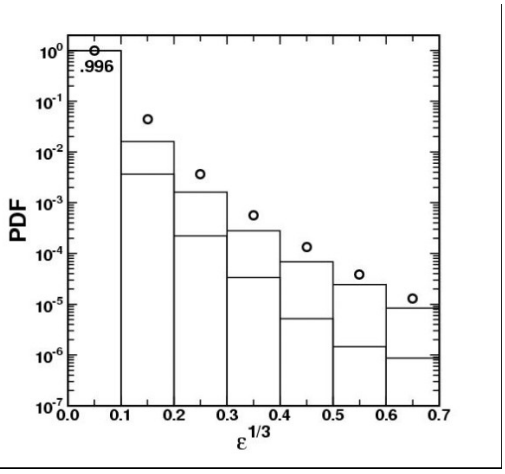

Figure 1. Taken from Sharman et al. (2006). This figure shows the probability distribution function (PDF) of three months of observed EDR values ($\varepsilon^{1/3}$) in each in-situ bin, both median (lower bar) and 95th percentile (upper bar). The open circles are estimates of the true lognormal distribution 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. Figure 2 shows the geographic distribution of available in-situ data over winter 2005-2006, which is virtually identical to distribution in winters of 2006-2007 and 2007-2008 used in this study.

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.

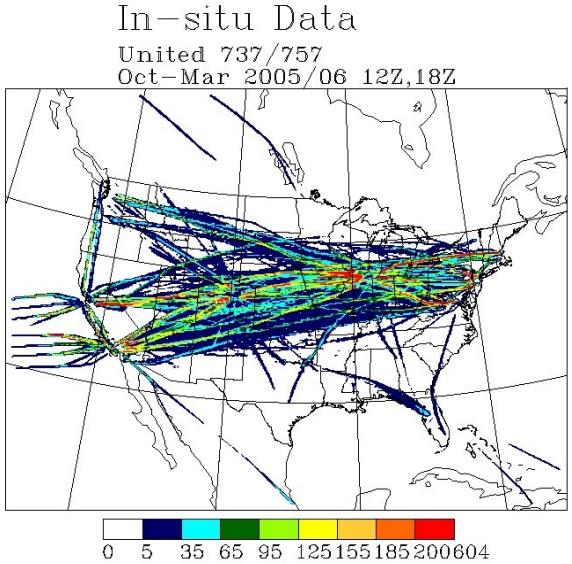

Figure 2. Geographic distribution of the in-situ data used in this study.

The effort to understand in-situ intensity values relative to PIREP intensities is ongoing.

We defined MOG turbulence as 0.35 – 4th in-situ bin - or higher. This is based on the PIREP and in-situ data comparisons (Abernethy, 2008), 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 developers 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).

While diagnostics are valuable indicators of atmospheric conditions conducive to turbulence and will continue to be integral inputs into future forecasting systems, the aim is to avoid the

hand-on human costs of determining thresholds. Machine learning techniques can aid in this goal.

4. TECHNIQUES, METRICS AND DATA

Further details, background, and work related to what is presented in this paper can be found in Abernethy (2008), but the aspects relevant to my regionalization work are summarized here.

4.1 *Machine Learning Techniques*

In this paper, I cover work using Support Vector Machines (SVMs) and random forests. SVMs are large-margin classifiers. Background on the technique of Support Vector Machines can be found in Hsu et al. (2003). For implementation of the SVM, I use the LibSVM library (Chang and Lin, 2003). I used LibSVM's probabilistic output as turbulence intensities on a scale of (0,1) in order to compare to deterministic forecasts of the current GTG algorithm and produce ROC curves.

Random forests are a collection of many decision trees trained by bootstrapping (sampling with replacement) from the training data set (Brieman, 2001). The final forest classification is a majority vote from the constituent trees' classifications. The variations between the trees enables the random forest to avoid overfitting. Random forests do not require cross-validation or a separate test set to get an accuracy estimate; estimates are made using the portion of the training data set that was not used in forming the tree, called the out-of-bag (o.o.b) data. Random forests also have implicit feature selection, which I cover in the section below. I use the PARF (Parallel Random Forest) implementation (Topic, 2004).

4.2 *Performance Metrics*

I 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. I also use the raw importance score output from the PARF random forest implementation to assess individual diagnostic performance.

Recall from Section 2 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. Both AUC and TSS closer to 1 indicate higher forecasting accuracy.

The raw importance score is a measure of the effect of a certain diagnostic on the random forest trees' final votes. Specifically, a diagnostic's values are permuted, then run through the trees in the forest for which that training example was out-of-bag, and each tree's category vote is recorded. When this has been done for all training examples, the votes are compared to the original votes (when the diagnostic's values were not permuted); the average of these differences over all the trees is the raw importance score for that diagnostic. The diagnostic with the highest raw-importance value is said to be the most important diagnostic. Likewise, sets of most important diagnostics can be made by ranking the diagnostics by raw importance score and taking the top *n* diagnostics. This variable importance can be thought of as a feature selection approach.

4.3 *Data*

This study used data from winter 2006-2007 (October – March), and 2007-2008 (November – February), 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. Only the highest intensity observation in a grid box 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.

5. CLIMATOLOGICALLY-RELEVANT REGIONALIZATION

The physics of turbulence strongly suggests that specializing forecasts to smaller regions can take advantage of spatial and temporal differences in CAT climatology. In this section, I describe how I partitioned the domain (CONUS) into regions in a climatologically-relevant way. I did this by creating many overlapping, circular regions, and used the random forest algorithm and TSS to determine which diagnostics performed well in each region. By grouping neighboring regions according to their

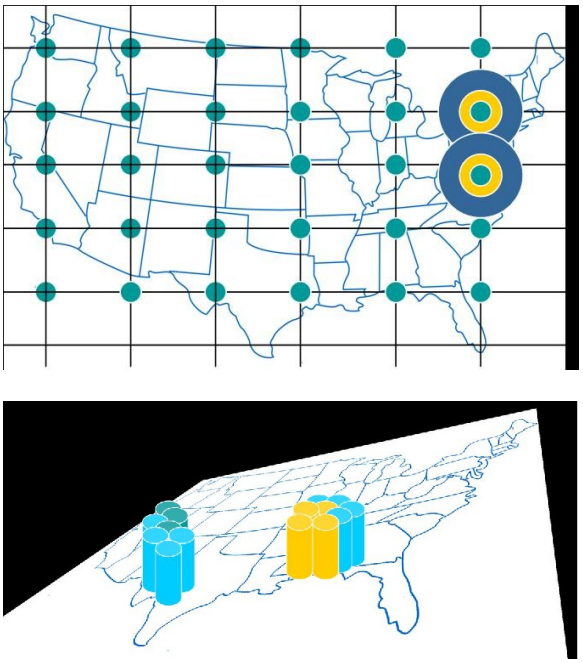

Figure 3. (top) Schematic of the point-centered region approach. (bottom) Schematic showing that regions are cylindrical and overlap. Adjacent regions can have different most important diagnostics (indicated by color).

top-performing diagnostics, I created larger areas of diagnostic similarity. The smaller regions, and the overlaps between them, allowed me to see more precisely where -- geographically -- the forecast power of different diagnostics increased or decreased between adjacent regions, defining more organic region boundaries. A regional forecasting algorithm that focuses on these individual regions can model the specific mechanisms per region, thereby increasing forecasting accuracy.

This point-centered region approach formed many more, smaller, regions with significant overlaps. I trained a random forest model on each region. Best-performing diagnostics for each region were the highest-ranked by raw importance score from the model. I determined the best-performing diagnostic for each *grid point* by tallying the best-performing diagnostics of all the regions in which the point was a member. However, sometimes all the member regions did not have the same most important diagnostic; these disagreements were rectified most simply by taking a majority “vote" of the regions' most important diagnostics As the algorithm moved point by point across the grid, I could see where the best-performing diagnostics changed. I defined the boundary lines where adjacent regions have different most-important diagnostics, which indicates the geographic location where a diagnostic's forecast accuracy starts to drop, or another diagnostic's forecast accuracy rises. Regions formed by these boundary lines presumably encompass areas that have similar mechanisms of turbulence creation.

I centered regions around every even latitude and longitude points over the CONUS domain (i.e., 100.0 W, 45.0 N), resulting in 440 regions covering the majority of the U.S. landmass between (121W,45N) and (75W,36N). A schematic of the region-creation process is shown in Figure 3. For clarity, the figure shows only a sample of the region centerpoints. The horizontal and vertical lines roughly represent latitude and longitude lines, respectively. Example region centerpoints are represented by small dots along the schematic latitude and longitude lines in the figure. I formed each region by moving the region boundary radially out from the centerpoint, until the region encompassed a sufficient amount of data to train a random forest model. The figure shows two examples of this step-wise radial boundary growth, represented as “bulls-eyes”.

I balanced the desire for small region size with the need for sufficient training data. Based on a sensitivity study Abernethy (2008) evaluating the accuracy effect of data amount, class distribution and number of trees, I found that 800 MOG observations (and therefore many more than 800 null observations), 500 trees, and several class distributions produced stable classifiers for this data. This meant I could keep regions around 400km in diameter. The training set, initially consisting of all the data located within each region's boundary, was rebalanced to be composed of 40% MOG and 60% null observations. All diagnostics were included in the training data. In the work presented here, I did not regionalize by altitude; thus, regions were essentially cylinders, as shown in Figure 3. Each region included data from 20000ft to 45000ft.

Results are shown in Figure 4. The top figure shows the regionalization using the one top diagnostic for each grid point, and the bottom figure shows the regionalization when looking at the set of two top diagnostics. The bottom figure further breaks down regions in the top figure, showing greater regional variation. In the east, the bottom figure shows the variation in mechanisms of turbulence secondary to saturated Ri.

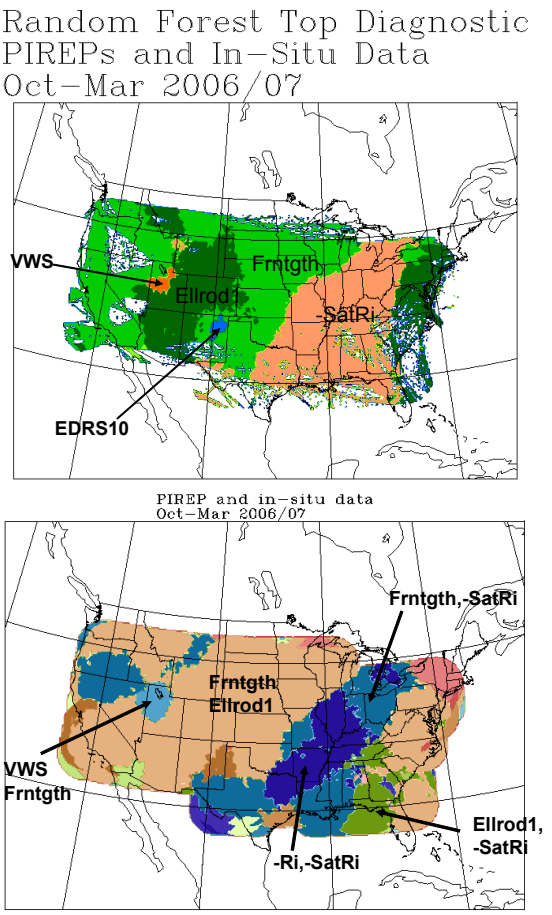

Figure 4. (top) Regions found using single top diagnostic measured by raw importance score. (bottom) Regions found using two top diagnostics measured by raw importance score. Each color represents a different diagnostic, or set of diagnostics.

6. A REGIONALIZED FORECASTING SYSTEM TESTBED: FST

Different metrics, data types, and number of diagnostics give different regionalization results (Abernethy, 2008). Since the aim of this project is to improve CAT forecasting accuracy, I needed to compare the forecasting accuracies of different regional divisions with different forecasting algorithms.To do these comparisons quickly and easily, I built a software system, called the Forecasting System TestBed (FST), that automates the steps of regional division, diagnostic set choice, algorithm training, and forecasting verification. (Diagnostic or feature set selection methods are covered in detail in Abernethy (2008)). FST takes a specific regionalization scheme such as one from Figure 4 and creates a forecasting model for each region based on region-specific diagnostics, and uses those models to produce CONUS-wide turbulence forecasts. The ability to easily test many regionalization and algorithm scenarios will enable systematic and rapid development of a final operational forecasting product.

FST divides the training, testing and validation data input by the user into regional data sets according to the geographic regions specified by the chosen regionalization scheme. The user specifies which forecasting algorithm to use: SVM, logistic regression, or random forests. Recall that to tailor a diagnostic set to a particular algorithm for forecasting, the algorithm also plays the role of the induction algorithm in the feature subset selection search. When using SVMs or logistic regression as the forecasting algorithm, FST executes a feature subset selection search on the training data within each region, and the test data set is used to assess performance of the trained model at each search step. The random forest algorithm, as before, builds decision trees using the diagnostics that produce the best split of the

data, and it outputs a list of those most important diagnostics. Thus no feature selection search was needed when the system used random forests as the forecasting algorithm; region-specific diagnostic subsets were taken from the output lists for the region model. Note that since no search is involved, a separate test set is not needed, and the training and test sets can be combined into one larger training set.

For each region, FST takes the diagnostic subset determined by the search (if using SVM or logistic regression) and uses that set to train a forecasting model with data from that region. The forecasting model can be an SVM, a logistic regression model, or a random forest, as specified by the user. In the case of a random forest model, an initial model is trained with all the possible diagnostics, and a set of the most important diagnostics is chosen from the output list. A new random forest model is then trained using only that set of diagnostics.

FST determines the accuracy of each region's forecasting model using the validation data for that region. FST calculates TSS as well as AUC for each region, and calculates these metrics CONUS-wide as well. The user is able to easily determine the skill of the regionalization scheme by looking at the regional and overall metrics output by FST.

Method	**Region**	**Diagnostic Subset Size**	**AUC**	**TSS**
RF (I,P)	1	17	0.78	0.41
RF (I,P)	2	22	0.81	0.47
RF (I,P)	3	23	0.84	0.51
RF (I,P)	4	20	0.84	0.52
RF (I,P)	5	14	0.81	0.37
		Overall:	0.833	0.50
SVM (I)	2	4	0.86	0.55
SVM (I)	3	4	0.84	0.53
		Overall:	0.848	0.54
RF (I)	2	12	0.88	0.61
RF (I)	3	13	0.89	0.61
		Overall:	0.885	0.61

Table 1. FST results for three trials: random forest forecasting algorithms using both in-situ and PIREP data (RF (I,P)), SVM forecasting algorithms using only in-situ data (SVM (I)), and random forest algorithms using only in-situ data (RF (I)).

My first trial with FST used the regionalization scheme shown in Figure 4 (top) with random forests for the forecasting algorithm for each region, using both PIREPs and in-situ data. My second and third trials with FST used same regionalization scheme, with SVMs and random forests as the forecasting algorithms. I used only in-situ data for these trials, for both winters 2006-7 and 2007-8. The AUC and TSS results are in Table 1. In all trials, I did not have enough data in some regions to train a stable model; the table shows that with both PIREPs and in-situ data, I could train models in 5 of the 8 regions, but with in-situ data alone, only two of the regions had over 100 MOG observations -- a very liberal minimum I set based on my random forest sensitivity studies (Abernethy, 2008).

The first trial produced higher forecasting accuracies than its non-regionalized counterparts (Abernethy, 2008). The first trial, using random forests with PIREP and in-situ data, had an overall AUC of 0.833 and TSS of 0.50, whereas no algorithm in had a 6-hour forecast AUC higher than 0.79; GTG 6-hour forecast AUC was 0.79, with TSS = 0.39. The second trial, using SVMs and in-situ data only, unfortunately produced an overall AUC of 0.848, which was lower than the corresponding CONUS-wide models. However, the third trial, which used random forests instead of SVMs, did improve forecasting performance; the AUC was 0.885 compared to a range of 0.82-0.87 for CONUS-wide models. It is probable that there were not enough MOG examples in the two regions for SVMs to achieve high skill as compared to the full data set used in CONUS-wide models, however, it appears that the random forest algorithm can still perform with small amounts of data.

7. CONCLUSION

Results in this paper show strong promise for the regionalization approach to turbulence forecasting. FST will continue to be used to refine the combination of geographic forecasting regions, turbulence diagnostics used, and forecasting algorithms for future versions of operational turbulence forecasting products.

8. ACKNOWLEDGMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

9. REFERENCES

Abernethy, J., 2008: A Domain analysis approach to clear-air turbulence forecasting using high-density in-situ measurements. *Ph.D. Dissertation*, Department of Computer Science, University of Colorado.

Brieman, L. 2001: Random forests. *J. of Machine Learning*, **45**, no. 1.

Chang, C. and C. Lin, 2003: LIBSVM – a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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.

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**.

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

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.*

Panofsky, H and J. Dutton, 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., 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., C. Tebaldi, G. Wiener and J. Wolff, 2006: An Integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting*, **21**, 268-287.

Takacs, A., L. Holland and M. Chapman and B. Brown and J. Mahoney and C. Fischer, 2004: Graphical Turbulence Guidance 2 (GTG2): Quality Assessment Report. *Technical Report by the FAA Aviation Weather Research Program Quality Assessment Product Development Team.*

Topic, G. 2004: PARF: Parallel Random Forest. Centre for Informatics and Computing, Institut Ruder Boskovic, Croatia. Available at http://www.irb.hr/en/cir/projects/info/parf.

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