FLEXIBLE DISCRIMINANT TECHNIQUES FOR FORECASTING CLEAR-AIR TURBULENCE

FLEXIBLE TECHNIQUES FOR TURBULENCE FORECASTING

Claudia Tebaldi[1], Doug Nychka,
Barbara G. Brown & Robert Sharman

Claudia Tebaldi, Barbara Brown and Robert Sharman are project scientists in the Research Applications Program at the National Center for Atmospheric Research (NCAR). Doug Nychka is scientist in the Climate and Global Dynamics Section of NCAR, and project leader of the Geophysical Statistics Project at NCAR. Part of this research was sponsored by NSF grant DMS-9815344 and by the Turbulence Product Development Team of the Federal Aviation Administration's (FAA's) Aviation Weather Research Program (AWRP).

[1]Claudia Tebaldi, Research Applications Program, National Center for Atmospheric Research, P.O.Box 3000, Boulder, CO 80307; `tebaldi@ucar.edu`; ph. (303) 497 2830; fax (303) 497 8401.

Summary: Forecasting aircraft (clear-air) turbulence is currently based on a system of observations by pilots combined with a mostly subjective evaluation of turbulence indices derived from numerical weather prediction models. We address the issue of improving the forecasting capability of the single indices by combining them in a non-parametric multidimensional regression model, and applying discriminant analysis to the resulting predicted values. Thus we enhance the predictive skills of the indices considered in isolation and provide a more robust algorithm. We adopt the paradigm of Flexible Discriminant Analysis (FDA), and use Multivariate Adaptive Regression Splines (MARS) and Neural Networks (NN) in the regression stage. The data for this case study covers the period March 12-15 1999, for the United States. Results of the analyses suggest that our statistical approach improves upon current practice to the point that it holds promise for operational forecasts.

Keywords: *Multivariate adaptive regression splines (MARS); neural networks (NN); flexible discriminant analysis (FDA); numerical weather prediction models; turbulence potential indices; probability of detection.*

1 Introduction

In January 1998 an airliner flying between Tokyo and the United States unexpectedly encountered severe turbulence. As a result, one woman aboard was fatally injured and several passengers were seriously harmed. This is a dramatic example of the problems caused by turbulence with regard to aviation safety. Beyond such extreme cases, the magnitude of the problem is cause for concern: US airlines experience on average 30 medical emergencies a day related to turbulence encounters; structural damage to aircraft may result; and there is a significant increase in fuel consumption during turbulent flights (Ellrod and Knapp, 1992).

One particularly troublesome type of turbulence encounters are those occurring in clear air, but predicting clear-air turbulence (CAT) is a challenge for the scientific community. Its occurrence is seldom linked to visible phenomena (viz. clouds), as the name suggests, and it is patchy in space and time. Moreover, the precise physical mechanisms that create CAT are not well understood. Operational predictions are still, for the most part, based on subjective evaluation of weather maps and recent reports from pilots who encounter turbulence. As an alternative, quantitative predictors of CAT have been proposed based on the hypothesized triggering mechanisms and computed from the output of numerical weather predictions models. Such models are dynamical in nature, relying primarily on deterministic physical models of atmospheric motions, with parameterizations of some smaller-scale phenomena. In particular, these models can not resolve instances of CAT, and so any index derived from model output is necessarily only an indirect measure of the phenomenon. This project was initiated with the belief that a comprehensive statistical approach could contribute to evaluating and improving the forecasting skill of these indices.

Tackling this as a statistical problem is complicated by the quality of the verification data. Airplanes encountering turbulence are the only source of observations. Consequently, the spatial distribution of these observations is irregular, following the routes traveled by the aircraft, with intense coverage over certain regions and poor or no 'sampling' of others (See Tables 2 and 3). In addition, the objectivity and reliability of the observations are an issue: no instrument currently is in use that can objectively measure turbulence intensity, and so the task is assigned to the pilots themselves, who are requested to log turbulence encounters and identify intensity based on their own experience or impressions. Although a small number of aircraft additionally provide automated observations of the vertical acceleration of the aircraft, this type of record is only reliable for signaling the absence of turbulence, since accelerations may also be pilot induced.

Based on pilot and instrument observations, the goal of the statistical analysis is to build a model to forecast CAT using various turbulence diagnostic indices computed from the output of a numerical weather prediction model. Our analysis features nonparametric

methods for discrimination and serves two purposes. The first is to introduce to a statistical audience a substantive problem in meteorological forecasting that benefits from an intensive statistical treatment. The second goal is to document the success of flexible discriminant analysis (FDA) for a physically based problem. Although we find the philosophy behind FDA compelling, we also note that few independent examples exist in the statistics literature indicating its value.

The paper is structured as follows. Section 2 discusses CAT from the stand point of the atmospheric science community, and outlines current practice for its prediction. Section 3 describes the observational data available, and the pre-processing needed for statistical analysis. Section 4 reports results from an initial analysis of the single indices' contribution. Section 5 gives a model-based introduction to FDA, followed in Section 6 by descriptions of two nonparametric techniques: Multiple Adaptive Regression Splines (MARS) and Neural Networks (NN). Section 7 describes the actual implementation of these techniques to the problem of CAT forecasting. Evaluation of our solutions and comparison with the single index approach and a linear (logistic) approach are performed, using standard measures of forecast quality (e.g., Brown et al., 2000). Section 8 is a discussion of open statistical questions posed by this case study and possible extensions when this work is used in an operational forecasting setting.

2 Diagnosing clear-air turbulence

2.1 The Rapid Update Cycle system

Numerical weather prediction models can predict synoptic or mesoscale conditions thought to be conducive to the initiation and maintenance of CAT. Unfortunately, CAT is a small scale meteorological phenomenon, not resolvable at the typical model grid spacing, and this fact accounts for poor statistical correlation between CAT and any quantity predicted by such models (Ellrod and Knapp, 1992). Our study aims to enhance the forecasting skill of a suite of indices routinely derived from the output of the Rapid Update Cycle (RUC) system, a numerical weather prediction model that is now in operational use by the National Oceanic and Atmospheric Administration (Benjamin et al., 1998). The RUC system produces three-dimensional analyses and short-range forecasts at regular intervals by blending observations from a network of surface stations, rawinsondes[2], aircraft and profilers with the background meteorological fields provided by its previous forecast output. In the present formulation the horizontal domain covers the continental United States and adjacent areas of Canada,

[2]Instrument packages lifted by balloon that measure meteorological quantities such as temperature, relative humidity, pressure and, being coupled to a GSP transmitter or tracked by radar, wind direction and speed

Mexico, and oceans. The vertical dimension is resolved by a number of layers extending from the surface to 18,000 meters above it, through an irregular vertical spacing based on the local thermal structure of the atmosphere. Horizontal grid spacing is 40 km. over the domain shown in Figure 1.

The RUC system produces variables to describe the state of the atmosphere, among which are relative humidity, surface temperature, dew point, sea level pressure, wind component speeds, precipitation amount, 3-hr pressure change, and gust wind speed. Conceptually such a model divides the atmosphere into a three dimensional grid of boxes. The state vector can be interpreted as the average physical quantities in each box at a particular time. The model is evolved forward to the next time step using the discretized mechanical and thermodynamical laws for fluid flow and conservation laws. The main caveat to this approach is that the spatial resolution of this model, and almost all other operational forecast models, cannot explicitly describe important features of the atmosphere. For example the influence of thunderstorms (strong convection) in providing vertical motion is accounted for in the RUC model implicitly through a technique known as *parameterization*. Clear-air turbulence also cannot be resolved from the RUC model, however the hypothesis is that the average quantities for a grid box are useful for diagnosing the local tendencies in the atmosphere, most important for weather forecasting applications. In this way they should give some information about the presence or absence of CAT.

2.2 Turbulence potential indices

From the RUC model state vector, potential indicators of turbulence (indices) are computed. A complete list of the indices that constitute the covariate set appears in the Appendix and to give the reader an example of the indices' physical base, a short derivation of Ellrod's indices is included. Depending on the scientific definition of each index, values above or below index-specific thresholds are considered indicators of CAT potential. These thresholds are calibrated by comparing index values to observed CAT episodes. This calibration has historically shown a lack of objectivity and consistency across different model outputs and sets of verification data. Indeed, to our knowledge, no rigorous approach has been undertaken so far in order to analyze, discriminate, compare and combine the efficacy of the different indices, and it is in this direction that our work is focused.

The atmospheric sciences have a well developed literature on quantifying how well a forecasting method performs (e.g., Harvey et al., 1992; Murphy, 1997; Wilks, 1995). With respect to CAT there are still unresolved issues as to what are acceptable values of the forecast verification statistics. Recent work has outlined some standards, however, that are required in order for a CAT forecast to be useful (Sharman and Cornman, 1998). Two important quantities are the probability of detection of a YES event, *pd1* from now on,

defined in our case as the proportion of observed turbulence events that were correctly forecasted, and the probability of detection of a NO event, *pd0*, defined as the proportion of observed non-turbulence events that were correctly forecasted. Due to the peculiar nature of the verification data (pilots' reports and vertical accelerometer observations), we consider *pd0* and *pd1* reliably estimated, since they are estimates of distributions conditional on the observations, but we do not look at other statistics, such as the false alarm ratio (defined as the proportion of forecasted YES events that were incorrect) since they are estimates of distributions conditional on the forecasts, and the sampling of the forecast grid by the observations is not representative enough. For a detailed justification of this choice we refer to Brown and Young (2000) and references therein.

3 The verification data

We use the model output relative to three forecast times: 15:00, 18:00 and 21:00 UTC (corresponding to 9:00, 12:00 and 3:00 EST) from March 12-15, 1999, a total of $12 = 4 \times 3$ time periods. The winter period is when CAT occurrences are more frequent, and it is also fairly unaffected by convective activity, frequent in spring and summer, that would introduce spurious sources of turbulence. We choose a limited dataset as the test bed of our statistical investigation, in order to maintain a manageable size for analysis by standard statistical packages.

The CAT observations are recorded by commercial aircraft along their regular flight path, and consist of pilots' reports (*pireps*) – signaling the location, time and intensity of turbulence episodes encountered – and vertical accelerometer data (*avars*) – measuring in real time all aircraft accelerations, whether they are due to turns, climbs, descents, or turbulence encounters. Pireps can include both positive and negative observations of turbulence. They are, however, reported sporadically and inconsistently because, in general, they are not required by regulation. Avars provide unambiguous information only when the instrument produces a null record, i.e. does not register vertical acceleration, and thus can only be used as observations for the absence of turbulence. Therefore pireps are the only source of positive turbulence observations, which are rated subjectively by the pilots on a discrete scale from 0 – No turbulence – to 8 – Extreme turbulence. Because of the degree of subjectivity involved in these grades, we prefer a coarser scale and for the purposes of this study we will focus on a binary classification:

0 for Null or Light turbulence observations, contributed by both avars and pireps; this category includes the first three grades of the original scale (0 through 2).

1 for Moderate or more severe turbulence observations, only provided through pireps; this category includes the remaining 5 grades of the original scale (3 and higher).

A three category analysis has also been done but has not been included in this study. From an operational point of view, the interest of aviation safety is well translated into the distinction between these two classes of potential turbulence encounters.

Each observation has a time stamp and is thus associated with a 3-hour time window centered at a RUC forecast time. The number of observations for each time window is on average approximately 500, with a large preponderance of avars, i.e. null turbulence, observations. To limit the spurious effect of general aviation (i.e. small aircraft) pireps, and of turbulence encountered at low altitudes, which frequently is not CAT, we will confine our attention to records taken above $20,000$ feet.

Clearly the observations are irregularly spaced and likely to underrepresent the turbulent areas, since pilots tend to avoid them when possible (e.g. when they are warned by other pilots flying ahead). A more subtle effect is the possibility that those pilots who have more experience flying over areas that are climatologically subject to frequent turbulence could be inclined to understate the turbulence phenomena and their intensity. In general, whether a turbulence encounter is judged light, moderate or severe is largely dependent on the past experience of the particular pilot reporting it. Furthermore, the severity of the turbulence experience depends on aircraft size. We hope to have limited these effects by coarsening the classification into a binary choice.

An additional source of error are delays in reporting the episode leading to inaccurate times and locations of CAT events. The 3-hour window should limit the impact of the first type of inaccuracy. The smoothness of the 3-dimensional spatial fields of index values should limit the impact of the second.

On a more positive note, the synoptic conditions of the atmosphere conducive to turbulence on average move eastward across the area of study, thus allowing for a sampling over time more comprehensive than the sparse geographic locations would suggest. Also, CAT is known to be linked at times to gravity waves breaking in the vicinity of mountainous areas, and the coverage of the westernmost part of the continental US is ensured by the large number of flight paths linking major airports of the central states to major airports on the west coast.

Despite these issues, we consider our dataset representative of the "average conditions" during this winter period, and across the geographic area of the continental US. Table 1 compares the distribution of pilot reports among turbulence categories. Similarities between the different periods to the distribution of pilot reports in our 4 days' dataset (last row), for the months of March in years previous to 1999 and for the entire March 1999, can easily be assessed. As for coverage of the air traffic routes, Tables 2 and 3 list some statistics for March 1999 and the entire two-year period 1999-2000, respectively, for the 20 national *Air Route Traffic Control Center* regions in the continental U.S., whose corresponding airport codes

are listed row-wise. The areas controlled by the 20 centers are non-overlapping polygons covering completely the continental US. One piece of information that can be gleaned from both tables is the correlation (in value of 0.6) between the numbers of flights over an area (including all commercial aircraft cruising over the area that did not take off or landed at airports in the area), in the column labeled "ovr/area" and the number of pilot reports from an altitude over 20,000 feet that were recorded in that area, in the column labeled "pireps/area". Note that both values are standardized by the area's size. The reasonable value of this correlations is an indication of the representativeness of our case study.

A key step in pre-processing the data is to match pireps and avars measurements, recorded in continuous space and time, with the indices, computed on a discrete spatial grid and at 3-hour time intervals. We match pireps and avars to grid points by first finding the eight closest grid points to the observation location, i.e. the eight RUC output locations which form a box containing the observation. For all indices but Richardson's we pair the observation to the maximum of each index's eight values since the scientific definitions associate higher turbulence potential to larger index values. For the converse reason the minimum of the eight values of Richardson number is paired to the observation.

4 Univariate analysis and linear combinations of the CAT potential indices

The traditional approach to turbulence forecasting considers each index in isolation, as an independent piece of information. Thresholds for the indices are informally determined, and regions, known as threat areas, are drawn on the maps where one or more indices exceed these thresholds.

Despite current practice, the analysis of univariate distributions of the index values is not adequate; simple data analysis indicates the necessity of a more complex solution. Figure 2 illustrates the conditional distribution of several indices for the two categories of CAT observed. Although consistent patterns in the quantiles and means of the distributions appear, the spread of the distributions is large and overlapping. Moreover, there are a number of outliers. This behavior is common to the entire suite of indicators and suggests that each index taken in isolation cannot accurately discriminate different levels of turbulence.

In order to set a standard of comparison for our solution, we fit univariate and multivariate logistic regressions to the binary classification of the observations into null and positive turbulence (YES and NO events). For the univariate models' predictions, *pd0* and *pd1* are shown not to exceed .6 at the same time (being .6 the baseline against which the forecast skill of these models is measured). For the multivariate regression the best classification (assuming equal importance of the two measures) is able to produce $pd0 = pd1 = 0.65$. We

will examine performances in more detail in Section 7.

5 Statistical models for classification

5.1 Linear Discriminants

A classical, linear solution to the statistical discrimination problem is Fisher's linear discriminant analysis (LDA). Given J categories we assume that an observation vector, say $\mathbf{x}$, associated with group j is distributed $\mathrm{MN}(\mu_j, \Sigma)$. Thus there exists a $p \times (J-1)$ matrix, B, and a $J \times (J-1)$ matrix Θ so that

$$\mathbf{x}^T B \sim \mathrm{MN}(\Theta_j, I)$$

where $\mathbf{x}$ belongs to group j and Θ_j is the j^{th} column vector for Θ. Given prior probabilities $\{p_{j,\mathrm{prior}}\}$ of group membership, the posterior probability of belonging to group j is

$$p_{j,\mathrm{post}} = D_j p_{j,\mathrm{prior}} / \sum_{k=1}^{J} D_k p_{k,\mathrm{prior}} \tag{1}$$

where

$$D_j = e^{-||\mathbf{x}^T B - \Theta_j||/2}$$

Given a training sample the goal then is to estimate B and $\{\Theta_j\}$. These quantities can be estimated by minimizing the minus log likelihood:

$$(XB - Y\Theta)'(XB - Y\Theta)/2 + C \tag{2}$$

over matrices B and Θ subject to the constraint that $\Theta^T\Theta = I$ Here Y is the $n \times J$ matrix of indicators of class membership, X is the $n \times p$ matrix of covariates for the training sample and C is an expression that does not depend on Θ or B.

In our problem, the multivariate distribution of the vector of indices' values, conditionally on having observed YES events or NO events, does not satisfy Gaussian assumptions. Moreover, the estimated within group covariances are far from constant across the two groups. These departures suggest the need to consider other approaches.

5.2 Flexible Discrimination

We adopt a "flexible discriminant analysis" (FDA) paradigm (Hastie et al., 1994). At the core of FDA is the assumption that a transformation exists, that — when applied to the original covariates' space — will map the observations to a new lower dimensional space where the assumption of normality is sustainable, and LDA may be applied. Two nonparametric techniques will be used to estimate the form of this transformation: multivariate

adaptive regression splines (Friedman, 1991), and neural networks, from now on referred to as MARS and NNs respectively.

6 Nonparametric regression estimators

Both MARS and NNs models are nonparametric regression techniques that are popular for classification and regression problems. They both accomplish flexible modeling of high dimensional data by estimating the functional form $F(\cdot)$ under the nominal Gaussian errors regression model

$$Y = F(\mathbf{x}) + \epsilon.$$

Conceptually, both of these methods have two parts: 1) a data-based determination of basis functions and 2) a linear regression on this selected basis. Of course, in terms of actual computation these two steps are carried out simultaneously.

6.1 MARS

In the MARS model,

$$F(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} [s_{km} \cdot (x_{v(k,m)} - t_{km})]_+^q. \tag{3}$$

That is, $F(.)$ takes the form of an expansion in product spline basis functions, where the number of basis functions (M) as well as the parameters associated with each one (degree q and knot locations t_{km}) are determined adaptively by the data. As mentioned above we see that $F(\mathbf{x})$ is linear in $a_0, \ldots, a_M$ once the basis functions have been specified. The procedure for determining the basis functions is an extension of the recursive partitioning approach to regression and shares its most useful properties. It is efficient in finding low local dimensionality of the function even if globally the function depends on a large number of variables.

6.2 Neural Networks

For the NNs' estimator,

$$F(\mathbf{x}) = a_0 + \sum_{l=1}^{M} a_l \Phi\left(\mu_l + \mathbf{w}_l^T \mathbf{x}\right). \tag{4}$$

The p-dimensional vector $\mathbf{x}$ of predictor variables is first transformed into M scalars by M distinct linear combinations of its components. Each of the scalars, then, becomes the argument of a sigmoid function, usually taken as

$$\Phi(x) = \frac{1}{1 + e^{-x}}$$

or a rational polynomial approximation to it. Again, note that for fixed $\{\mathbf{w}_l\}$ and μ, $F(\mathbf{x})$ is a linear function in $\mathbf{a}$. Technically, this neural network is classified as a single hidden layer feed forward network. More complicated network architectures may be used and the reader is referred to Cheng and Titterington (1994) for more background on such estimators.

6.3 Basis adaptation and dimension reduction

The specifications of the problem at hand seems to be well suited for both MARS and NN. The single indices may be good at discriminating the different levels of turbulence when they assume extreme values, but a multivariate classifier may be needed in less extreme cases. In particular we expect an adaptive, non-parametric estimator to identify more complex functional forms where interactions are needed.

The algorithms for fitting MARS models are composed of a forward fitting/backward pruning procedure: the first stage of parameter estimation is followed by the elimination of those terms that fail to provide a significant improvement in the overall goodness-of-fit. The generalized cross-validation (GCV) criterion of model selection is used in the backward stage, to control overfitting in the *training* step and to enhance the predicting skills, which is the fundamental focus of our analysis. The NN model estimates all parameters by a robust nonlinear least squares optimization, as implemented, for instance, in the FUNFITS Splus module (Nychka et al. 1996). The number of hidden units (M) is found by minimizing a GCV criterion with a cost factor of 2.

Both methods have some qualitative restrictions on the number of interactions that will be included. MARS departs from the vertical, hierarchical, tree-like structure of recursive partitioning and is able to represent functions whose dominant interactions involve only a small fraction of the total number of variables. NNs represent interactions through a sigmoidal transformation along a particular projection.

The efficacy of MARS in performing variable subset selection will also be of value in determining redundancies within the set of indices. Although NNs can perform subset selection by setting some of the coefficients (w_{lj} in the notation of (4)) to zero, this estimator is more difficult to interpret. Thus in assessing the importance of different indices, more useful information is derived from the MARS fitting, particularly with regard to possible simplifications and eliminations within the large suite of indices.

7 Prediction models for CAT

The purpose of this study is to evaluate the forecasting ability of FDA, and compare it to the traditional approaches currently in place.

Exploratory tests (not reported here) suggest the need to aggregate data from several

days to obtain a sufficiently representative number of positive turbulence records. We ideally would like to substitute the statistical estimation of a weighted combination of the indices' values for the present heuristic assessment of them, and to suggest so, we mimic an operational setting – within our limited set of test data — by the use of cross validation. One time period is omitted from the 12, the different models are fit on the basis of the remaining 11 time periods, and forecasts are made for the omitted period. Were the statistical approach to be made operational, the fit would be based on recent accumulated observations, and would be regularly updated for each issue and forecast time of the RUC model.

7.1 Forecast skill

As already indicated, we use two statistics as a measure of the prediction skills of the models: the probability of detection of positive turbulence, or YES events, (*pd1*) and the probability of detection of null turbulence, or NO events, (*pd0*). Note that 1-*pd0* is also known as the "false alarm rate", not to be confused with the false alarm *ratio* defined in Section 2. Because the probability of an event estimated by the statistical model may not be correct from a frequentist metric (the estimates may suffer lack of calibration), we are interested in evaluating the methods over a range of possible thresholds for such probabilities. The threshold is varied over the range $[0, 1]$ and areas/observations associated with probabilities above the threshold are classified as turbulent, while the areas/observations associated with probabilities below the threshold are classified as non turbulent. The net result is a relative operating curve for *pd0* and *pd1* showing the tradeoff between them as the threshold values vary (Mason, 1982; Harvey et al. 1992).

The prediction exercise for each method consists of the following steps for each time period:

1. Time period under study is omitted and the model parameters are estimated from the remaining data.
2. The probabilities of membership to the two classes (null and positive turbulence) are forecast for each observation in the omitted time period.
3. A series of threshold values uniformly spaced on $(0, 1)$ are applied to the probability of membership to the null turbulence class.
4. *pd0* and *pd1* are computed from the results of applying each threshold.

After performing these steps for each time period the probabilities of detection are combined across periods for the same value of threshold, being weighted by the number of positive/null turbulence observations in each dataset.

The final summary is a series of points (one for each threshold value) in the *pd0-pd1* plane for each method. This set of points defines the 'frontier' of performance for a method, and the relative positions of the curves associated to each method give a visual summary of performance. An ideal curve would approach the top and right sides of the *pd0/pd1* region.

The results are presented graphically in Figures 3 through 5, where the objective standards for $pd0, pd1 \geq 0.6$ are highlighted by two straight lines. The numbers appearing on the curves indicate the values of the corresponding threshold for the probability. Shown are thresholds that span $(0, 1)$ interval uniformly by steps of .1.

Figure 3 compares 7 univariate logistic regressions to the logistic fitted by the linear combination of the set of 7 indices. These 7 indices have been chosen on the basis of their individual performances, assessed by studies on larger data sets performed at NCAR. The univariate logistic classification is simply a monotonic transformation of the original index to a $[0, 1]$ range. Thus, varying the threshold across the probabilities is equivalent to varying a threshold in raw scale of the index and so we reproduce what is done in practice for a single index. The advantage of the logistic regression is that these results are in the same scale as the multivariate models. From this figure we see that the multivariate logistic regression does seem to have more discriminatory power than the univariate approach, but barely reaches the standard values of 0.6 determined as the minimum acceptable for these two statistics.

Figure 4 compares the same curve obtained from the multivariate logistic regression to the FDA+MARS model using the same 7 'best' indices and the FDA+MARS model that uses the complete suite. The gain from using the flexible technique is more important than the gain of additional indices in the nonparametric model. The FDA+MARS model performs well above the minimum standard. Last, FDA+MARS and FDA+NNs models that use the full suite of indices are compared in Figure 5, which shows that the NNs model is well above the minimum standard as well, but slightly dominated by the MARS model.

These comparisons suggest that our methods offer an important contribution in terms of robustness and reliability, compared to the single index approach. Furthermore, the *pd0-pd1* curves derived by FDA are the only ones to be in the upper right quadrant determined by the 0.6 limit on both statistics, set as an objective standard of effectiveness. One problem, however, is that the threshold probabilities seem to be distributed irregularly over the curves' span, hinting at a lack of calibration. If we notice what happens for the single index models compared to the FDA models, however, it is clear how the most severe cases of lack of calibration are found in the former and are somewhat limited in the latter. This provides another hint at the gain in consistency and robustness achieved by our approach.

7.2 Towards an operational implementation

The curves in Figure 3 through 5 only suggest a relative improvement of the nonlinear techniques over the logistic model, and assess the achievement of a standard of quality, but do not guarantee the performance of an operational implementation of such a standard. Here we demonstrate that, even if not exactly calibrated, the threshold probabilities estimated by our models are of operational utility, and we do so by showing that there is consistency of *pd0* and *pd1* for a given threshold probability, between training and test data sets (i.e. between fitted and predicted probabilities). Table 4 demonstrates this. The ten rows of Table 4 correspond to ten equally spaced intervals that cover the range $[0, 1]$. Say we fix *pd0* to be in the subinterval $[0.6, 0.7]$. There exist a range of threshold probabilities that deliver a classification of the observation in the training set satisfying the constraint on *pd0*. What is the value of *pd0* that we achieve when applying these threshold probabilities to an independent test set? Table 4 shows that, on average (we use the same cross-validation scheme described earlier), the value is still in the range $[0.6, 0.7]$, being 0.62/0.64 depending on the model used. The same can be said for any subinterval (the i^{th} row of the table contains values within the interval $((i-1)/10, i/10)$, for any i), and the same consistency is true for *pd1*. In other words, what the results in the table guarantee is the following: we can fit the parameters of our models (any of them, FDA+NN, FDA+MARS and logistic regression) on a training set of observations, choose a probability threshold on the basis of which the classification of the observations produces a desired pair of *pd0,pd1*, and apply the threshold to the predicted values of the test set, expecting to achieve the same values for *pd0,pd1*. Notice that, in order to economize space, each row of the table lists results for *pd0* and *pd1*, but it should be considered a look-up table for these two statistics separately. In fact for most of the threshold values, along a typical curve, a high value of *pd0* corresponds to a low value of *pd1* and viceversa.

7.3 Model interpretation

The results detailed above suggest that the indices considered in isolation are not very informative, and that a multidimensional approach performs better in predicting CAT. Improvement has been found in models that allow for high degrees of interactions among predictors and different subsets of predictors to be relevant in different subdomains of the multidimensional space. To substantiate these conclusions, we give a more detailed look at the variable selection procedure implicitly performed by MARS. MARS recursively chooses among the predictors set the index that — alone or in interaction terms — is more 'useful' at discriminating 'YES' and 'NO' events. This choice, based on a "least square type" criterion is indicative of the intrinsic quality of the indices. By looking at the form that this choice takes in the fitted models one can gather a measure of the relative 'discrimination quality'

of the indices in the predictors set.

In Table 5 we analyze the twelve models fitted to the data. The table indicates how many times (i.e. in how many terms) each index appears in each model, without distinguishing if it appears in isolation or in interaction terms, i.e. combined with other indices. As the table shows, some of the indices are left out consistently, or are rarely represented, thus becoming good candidates for elimination and simplification of the models. Others are obviously very important, particularly TKE3 which appears to be a cornerstone of every model.

The models do differ from each other, however, despite fundamental similarities in the choices of the predictor subset. One explanation is that the adaptive selective fit of MARS copes with the poor quality of the verification data through variation in the model structure. This also is an explanation for the dominance of our procedure over the single index approach. Through a multivariate model one borrows strength across the different predictors. Of course both flexibility and adaptivity bears with it the risk of overfitting. These results are acceptable across the entire test set and in contrast the single index's performances vary widely and have poor average performance.

To reinforce our claim that a multidimensional, nonlinear approach to the problem adds valuable information to the predictive process, we present in Table 6 the large number of two- and three-degree interaction terms. Each cell represents an interaction between two indices, and the number in the cell indicates the number of terms which use that interaction, out of all the twelve models fitted. We report results only for a subset of seven more heavily used indices.

8 Discussion

In our application we suspect that minor improvements in the classification are still possible but believe that overall we have arrived at a cogent solution for this problem. This statistical work is a significant improvement over heuristic and potentially suboptimal methods derived from subjective evaluation or linear approaches. Another measure of its success is the interest in evaluating FDA in an operational mode by the Research Applications Project (RAP) at the National Center for Atmospheric Research. This involves generating predictions every three hours over several months for the coterminous US and comparing results to pilot reports and *avar* measurements.

8.1 Extensions

We suggest some extensions of the present study in the light of its potential use as a sequential forecast methodology. Were this method to be operational, its forecast skill could be assessed as often as the weather forecast model produces forecast (every 3 hours

in its present implementation), and there would be a valuable opportunity to judiciously modify the method over time. In this perspective:

- Both space and time could be discretized differently. A narrow spatiotemporal window will possibly eliminate spurious observations unrelated to the the localized CAT phenomenon. Conversely, larger windows accommodate the errors in pilot reporting and the scarcity of turbulence observations. Although we have experimented with different windows, the width might be chosen adaptively.

- With more data one could train different models for different geographical areas and altitudes, or different synoptic conditions. This possibility raises a more strictly 'statistical' issue, one of a model selection method when prior information is available to guide such choice, if for instance we could subdivide the region of the prediction (or the synoptic conditions at the time of the forecast) into areas where large degree interactions are expected to be useful and areas where we expect simpler models to be effective. We would like to include such qualitative information in a more deliberate fashion. One benefit may be estimated probabilities with better calibration properties.

- As accelerometer data becomes available that indicates turbulence, weighting these instrument results with pilot reports would certainly improve the quality of the observations.

We believe that the quality of the available observations and the current state of the indices' development represent important limitations to the performance of the technique we propose, and we consider this as just a first attempt to address the difficult problem of CAT forecasting. Recent advances in turbulence measurement and reporting strategies (Cornman et al., 1995) promise to make available widespread quantitative observations of turbulence, with improved quality, objectivity and precision. This development should dramatically improve our methodology, whose verification and tuning have been hampered by the limited availability and subjectivity of the current set of observations. Nevertheless, despite improved measurement and verification we believe the statistical methods used in this work will remain an important tool for building forecast models.

9 Appendix

In the following, let x, y, z define a right-handed coordinate system in a plane tangent to the earth's surface and with positive x, y, z being eastward, northward, and upwards respectively. Let u be the east-west (x) wind component and v be the north-south (y) wind component, and let T be the absolute temperature. Then the indices used in this study are defined as follows:

- Vertical wind shear:
  $\mathrm{VWS} = |(\partial u/\partial z)^2 + (\partial v/\partial z)^2|^{1/2}$.

- Horizontal wind shear:
  $\mathrm{HWS} = (u/s)\partial s/\partial y - (v/s)\partial s/\partial x$,
  where $s = (u^2 + v^2)^{1/2}$ is wind speed.

- Richardson number (e.g., Kronebach 1964):
  $\mathrm{Ri} = N^2/\mathrm{VWS}^2$,
  where $N^2 = (g/\theta\ \partial\theta/\partial z)$ is stability, θ is potential temperature.

- Turbulent Kinetic Energy (Marroquin 1998):
  $\mathrm{TKE3,5} = f(\mathrm{Ri}, N^2)$.

- Colson-Panofsky index (Colson and Panofsky 1965):
  $\text{Col-Pan} = \mathrm{VWS}(1 - \mathrm{Ri}/\bar{\mathrm{R}}\mathrm{i})$,
  where the critical value $\bar{\mathrm{R}}\mathrm{i}$ is set to 0.5.

- Ellrod indices (Ellrod and Knapp 1992):

  $$\mathrm{TI1} = \mathrm{VWS} \times \mathrm{DEF}$$

  $$\mathrm{TI2} = \mathrm{VWS} \times (\mathrm{DEF} - \Delta_H),$$

where DEF $= (D_{\mathrm{ST}}^2 + D_{\mathrm{SH}}^2)^{1/2}$,
$D_{\mathrm{ST}} = \partial u/\partial x - \partial v/\partial y$ is stretching deformation,
$D_{\mathrm{SH}} = \partial v/\partial x + \partial u/\partial y$ is shearing deformation,
$\Delta_H = \partial u/\partial x + \partial v/\partial y$ is horizontal divergence.

- Endlich empirical wind index (Endlich 1964):
  $|\mathbf{v}| \times |d\psi/dz|$,
  where ψ is wind direction.

- Brown's index (Brown 1973):

  Brown1: $\Phi = (0.3\zeta_a^2 + D_{\mathrm{ST}}^2 + D_{\mathrm{SH}}^2)^{1/2}$,

  Brown2: $\Phi \mathrm{VWS}^2/24$,

  Where $\zeta_a = \zeta + f = \partial v/\partial x - \partial u/\partial y + 2\Omega \sin\varphi$, is absolute vorticity,
  f indicates the Coriolis frequency, and φ geographic latitude.

- Reap MOSS predictors (Reap 1996):

  NGM1 $=$ DEF $\times |\mathbf{v}|$

  NGM2 $=$ DEF $\times |dT/dz|$,

- Dutton's empirical index (M.J.O. Dutton 1980):
  Dutton $= 1.25\mathrm{HWS} + 0.25\mathrm{VWS} + 10.5$.

- Anomalous wind gradient (McCann, 1997):
  AWG $= \zeta_{\mathrm{curv}} + f/2$,
  $\zeta_{\mathrm{curv}} = K_S|\mathbf{v}|$,
  $K_S = -(u/s)\partial\psi/\partial x - (v/s)\partial\psi/\partial y$ is streamline curvature.

- Divergence tendency (McCann, 1997):
  Div. ten. $= \partial\Delta_H/\partial t$,
  t is time.

- Inertial-advective wind (McCann, 1997):
  ABS $= |v_i - v_c|^2$,
  where $v_i = |\mathbf{v} \cdot \bigtriangledown \mathbf{v}|/f$,
  $v_c = K_S|\mathbf{v}|^2/f$

As an example, we briefly present the definition of the Ellrod indices TI1 and TI2.

Empirical studies (Ellrod and Knapp, 1992) show that stretching deformation (D_{ST}), which is derived from the u (east-west) and v (north-south) wind components computed as

$$D_{\mathrm{ST}} = \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y},$$

relates fairly well to observed CAT, but produces excessively large threat areas. The value of D_{ST} is then combined with shearing deformation (D_{SH}), computed as

$$D_{\mathrm{SH}} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}$$

in the quantity

$$\mathrm{DEF} == (D_{\mathrm{ST}}^2 + D_{\mathrm{SH}}^2)^{1/2}$$

which is found to reduce somewhat the contour areas. Vertical wind shear (VWS), defined as

$$\mathrm{VWS} = |(\partial u/\partial z)^2 + (\partial v/\partial z)^2|^{1/2},$$

correlates significantly to CAT as well, and so the product

$$\mathrm{TI1} = \mathrm{VWS} \times \mathrm{DEF}$$

is formed. TI1 is one version of the currently used Ellrod indices. The second formulation also includes divergence, which is defined as

$$\Delta_H = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}.$$

Δ_H is typically much smaller than DEF but in some cases has been shown to contribute to CAT potential. Thus, version TI2 of the index is computed as

$$\mathrm{TI2} = \mathrm{VWS} \times (\mathrm{DEF} - \Delta_H).$$

Acknowledgements

The authors thank two anonymous reviewers for their careful reading and suggestions that have substantially improved the article since its first version.

References

Benjamin SG, Brundage KJ, and Morone LL. 1998. Implementation of the rapid update cycle. http://maps.fsl.noaa.gov/tpbruc.cgi#2.1.

Brown R. 1973. New indices to locate clear-air turbulence. *Meteorological Magazine* **102**: 347-360.

Brown BG and Young GS. 2000. Verification of icing and turbulence forecasts: why some verification statistics can't be computed using pireps.*Preprints,* 9^{th} *Conference on Aviation, Range and Aerospace Meteorology*, Orlando, FL, 11-15 September.

Brown BG, Mahoney JL, Henderson J, Kane TL, Bullock R, and Hart J. 2000. The turblence algorithm intercomparison exercise: Statistical verification results. *Preprints,* 9^{th} *Conference on Aviation, Range and Aerospace Meteorology*, Orlando, FL, 11-15 September, 466-471.

Cheng B and Titterington DM. 1994. Neural Networks: a review from a statistical perspective. *Statistical Science* **9**: 2-54.

Colson D and Panofsky HA. 1965. An index of clear air turbulence. *Quarterly Journal of the Royal Meteorological Society* **91**: 507-513.

Cornman LB, Morse CS and Cunning G. 1995. Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *Journal of Aircraft* **32**: 171-177.

Dutton MJO. 1990. Probability forecasts of clear-air turbulence based on numerical output. *Meteorological Magazine* **109**: 293-310.

Ellrod GP and Knapp DI. 1992. An objective clear-air turbulence forecasting technique: verification and operational use. *Weather and Forecasting* **7**: 150-165.

Endlich RM. 1964. The mesoscale structure of some regions of clear air turbulence. *Journal of Applied Meteorology* **3**: 261-276.

Friedman JH. 1991. Multivariate adaptive regression splines (invited paper). *Annals of Statistics* **19**: 1-141.

Harvey LO, Hammond KR, Lusk CM and Mross EF. 1992. The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review* **120**: 863-883.

Hastie T, Tibshirani R and Buja A. 1994. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* **89**: 1255-1270.

Kronebach GW. 1964. An automated procedure for forecasting clear air turbulence. *Journal of Applied Meteorology* **3**: 119-125.

Marroquin A. 1998. An advanced algorithm to diagnose atmospheric turbulence using numerical model output. 16^{th} *Conference on Weather Analysis and Forecasting*, Phoenix, AZ.

Mason J. 1982. A model for assessment of weather forecasts. *Australian Meteorological Magazine* **30**: 291-303.

McCann DW. 1997. A "novel" approach to turbulence forecasting. *Preprints*, 7^{th} *Conf. on Aviation, Range and Aerospace Meteorology*. American Meteorological Society, Long Beach, CA.

Murphy AH. 1997. Forecast verification. In *Economic Value of Weather and Climate Forecasts*, Katz RW, Murphy AH (eds.); Cambridge University Press; 19-74.

Nychka D, Bailey B, Ellner S, Haaland P and O'Connell M. 1996. FUNFITS data analysis and statistical tools for estimating functions, *North Carolina Institute of Statistics Mimeoseries* **2289**. http://www.stat.ncsu.edu/ nychka/funfits/

Reap RM. 1996. Probability forecasts of clear-air turbulence for the contiguous US.*National Weather Service Office of Meteorology Technical Procedures. Bullettin No.* **430** NOAA: 15pp.

Sharman R and Cornman L. 1998. An integrated approach to clear-air turbulence prediction. 36^{th} *Aerospace Sciences Meeting & Exhibit*, Reno, NV. Jan.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences: an Introduction*. Academic Press.

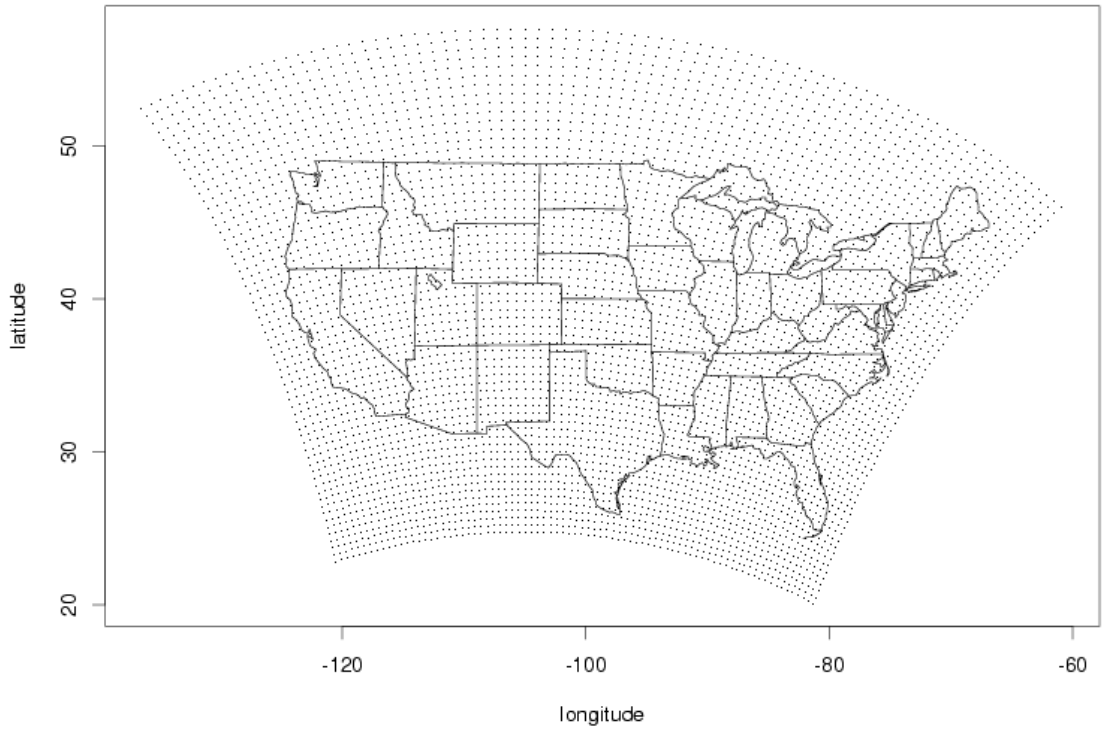

Figure 1: The RUC grid.

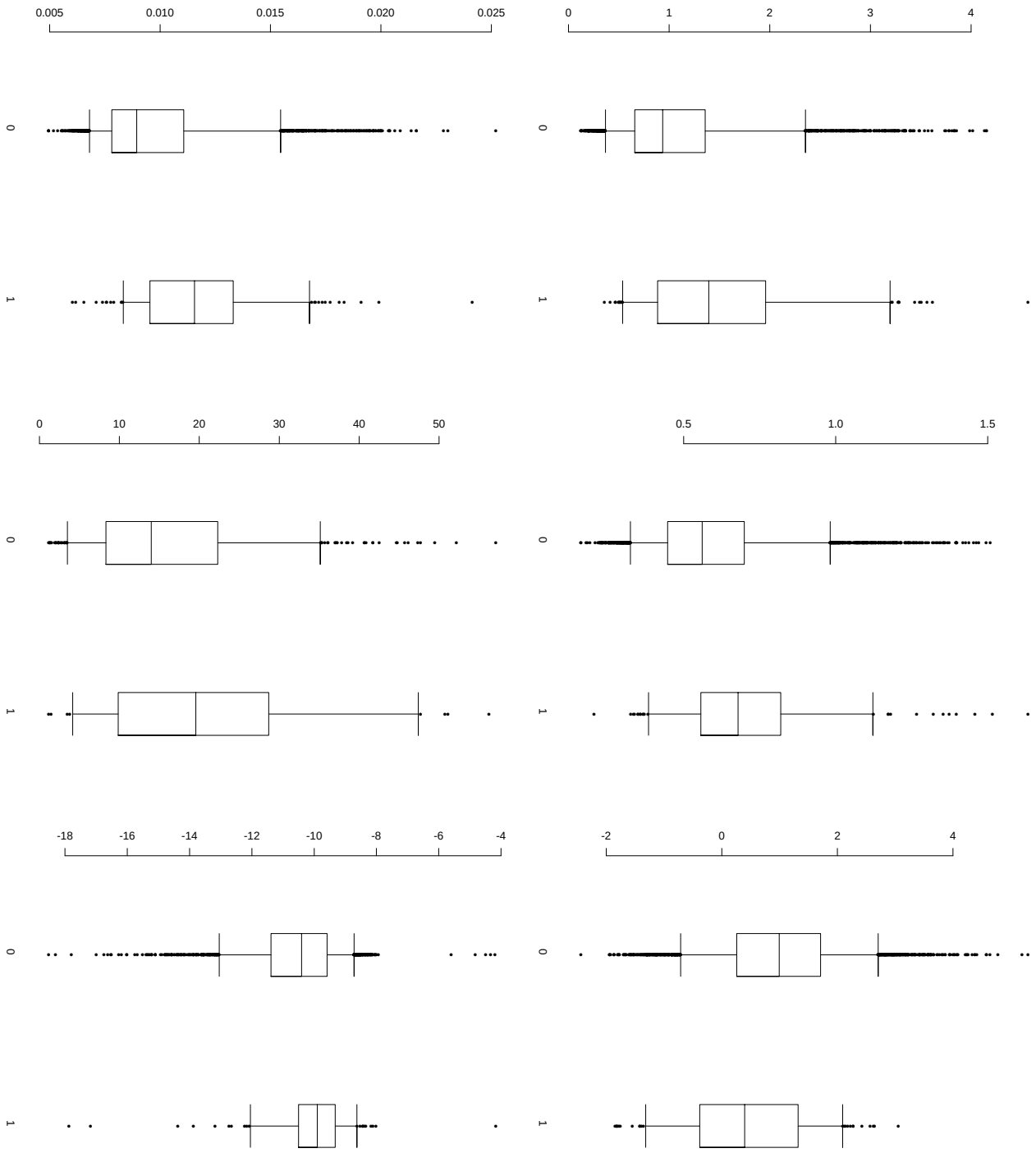

Figure 2: Distributions of several indices at the locations of the verification data set, conditional on the two different levels of turbulence observed. The boxplot at the top of each panel corresponds to values of the index at locations where no or light turbulence was reported, that at the bottom to values at locations where moderate or severe turbulence was reported. From top left to bottom right: Brown 1 (1/2 power-transformed), TKE3 (1/2 power-transformed), Col-Pan (1/2 power-transformed), Endlich (1/2 power-transformed), HWS (log-transformed), Ri (log-transformed).

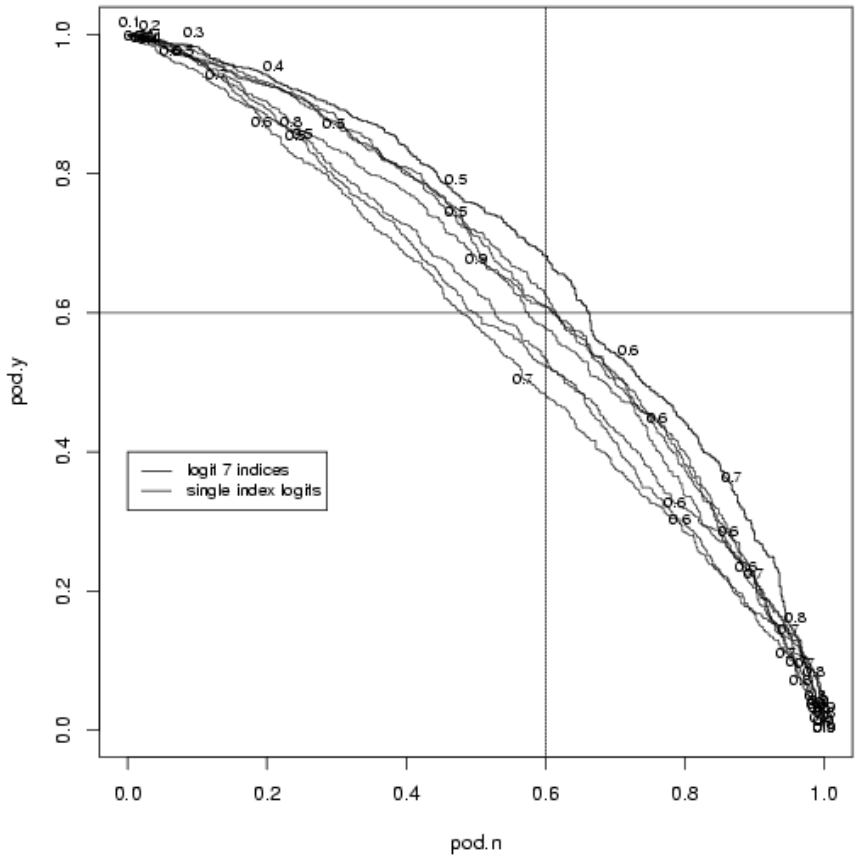

Figure 3: Operating curves for univariate and multivariate logistic regression models. On the x axis is the probability of correctly labeling an observation as "no or light turbulence" (class label 0); on the y axis is the probability of correctly labeling an observation as "moderate or severe turbulence" (class label 1). A point on the plot corresponds to a specific choice of a threshold for the posterior probabilities estimated by a model. Probabilities below threshold would translate into prediction of class 0, and conversely, probabilities above threshold would translate into prediction of class 1.

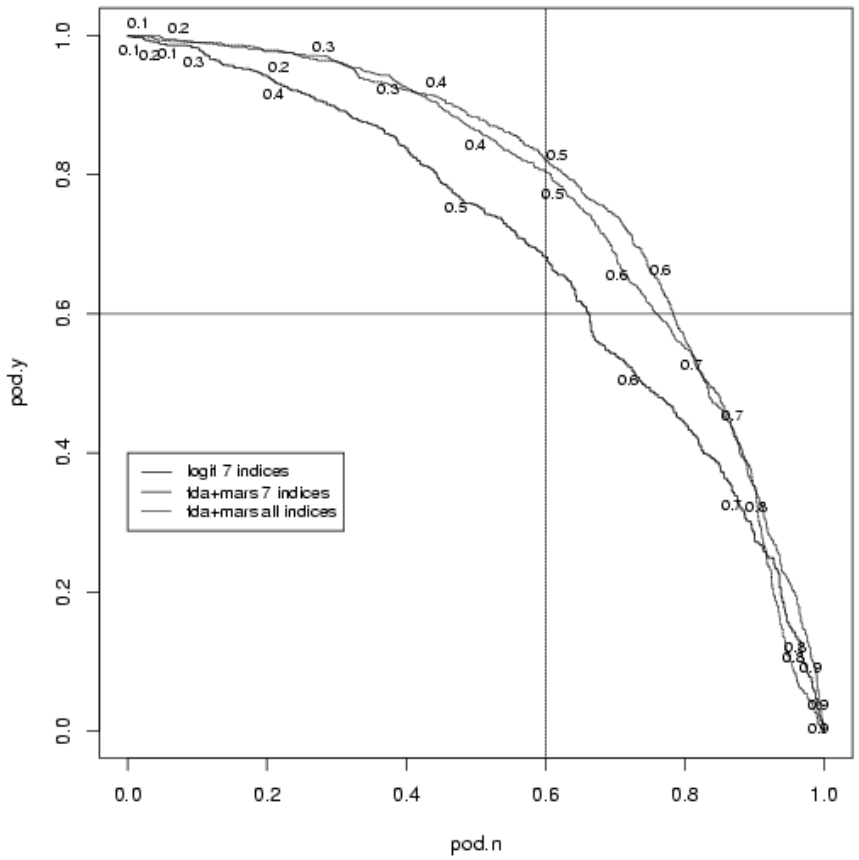

Figure 4: Operating curves for MARS models and logistic regression. Full suite of indices vs. 7 more reliable indices, chosen on the basis of univariate analyses. See caption of Figure 3 for details.

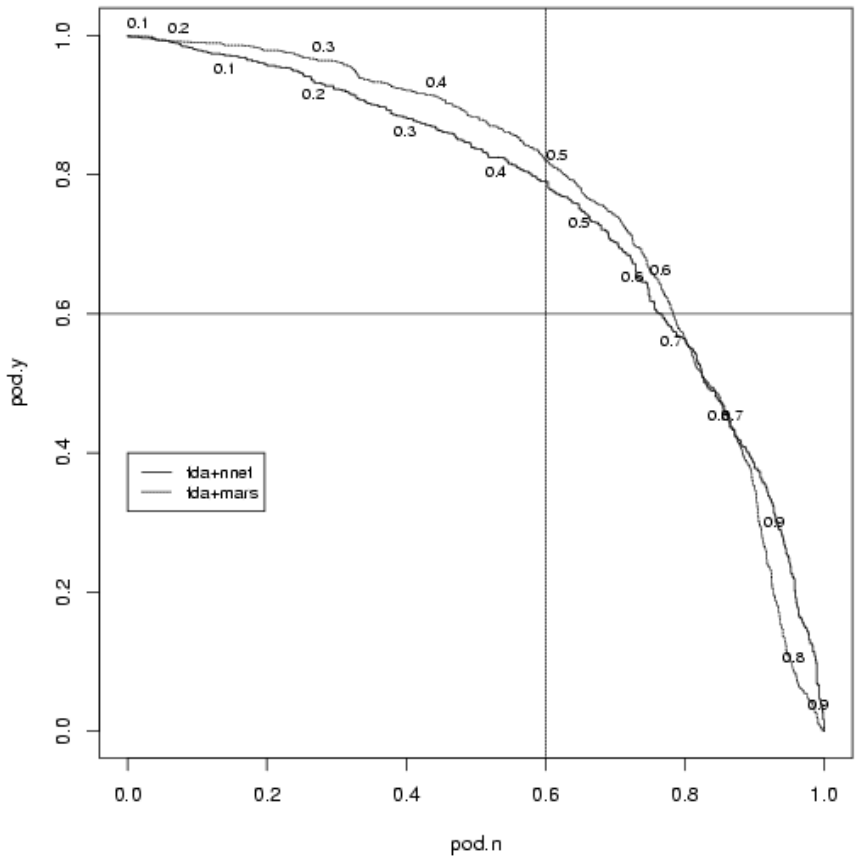

Figure 5: Operating curves for FDA+MARS and FDA+NN models. See caption of Figure 3 for details.

Table 1: Frequency distributions of pireps recorded in the month of March, for years 1993 through 1999, and of pireps in case study (last row). No turbulence encounters in categories higher than 6 was recorded, and pilots never labelled an encounter "null to light", that would correspond to category 1. Pireps represent 90% of the dataset in our case study, avars (null observations only) make up the rest.

	0	2	3	4	5	6
1993	0.42	0.24	0.08	0.21	0.02	0.02
1994	0.42	0.24	0.08	0.23	0.02	0.01
1995	0.43	0.23	0.07	0.24	0.01	0.01
1996	0.43	0.23	0.07	0.23	0.02	0.03
1997	0.39	0.22	0.04	0.32	0.01	0.02
1998	0.42	0.19	0.09	0.26	0.01	0.02
1999	0.42	0.21	0.09	0.26	0.01	0.01
4 days	0.35	0.19	0.10	0.32	0.03	0.02

Table 2: Number of flights over specific Air Route Traffic Control Center areas (listed in the first column), and pireps recorded in the same areas, for March 1999. The second column is the size of the area (km^2), the third is the total number of commercial flights that overflew the area (not landing, not taking off), the fourth is the ratio of the previous two (i.e. overflight density), the fifth is the number of pilot reports recorded over the area at an altitude of at least 20,000 feet. The sixth is the ratio of the previous column to the area (i.e. pirep density) multiplied by $1,000$. The correlation between column 4 and 6 is 0.60.

airport	area (km^2)	ovrflts	ovr/area	pireps	pireps*/area
ABQ (Albuquerque, NM)	616843	19752	0.03	283	0.46
CHI (Chicago, IL)	258921	36636	0.14	247	0.95
BOS (Boston, MA)	572151	13702	0.02	89	0.16
DC (Washington DC)	413105	41169	0.10	234	0.57
DEN (Denver, CO)	687536	48017	0.07	700	1.02
FTW (Fort Worth/Dallas, TX)	419770	27244	0.06	178	0.42
HOU (Houston, TX)	580765	11499	0.02	251	0.43
IND (Indianapolis, IN)	239752	69761	0.29	504	2.10
JAX (Jacksonville, FL)	494990	42240	0.09	106	0.21
KC (Kansas City, MO)	458859	46361	0.10	472	1.03
LAX (Los Angeles, CA)	463547	8867	0.02	162	0.35
SLC (Salt Lake City, UT)	1071881	34530	0.03	446	0.42
MIA (Miami, FL)	326840	11397	0.03	52	0.16
MEM (Memphis, TN)	368603	51556	0.14	143	0.39
MSP (Minneapolis, MN)	985467	40364	0.04	271	0.28
NY (New York, NY)	83412	29808	0.36	56	0.67
CLE (Cleveland, OH)	230969	62979	0.27	166	0.72
SEA (Seattle, WA)	673589	4175	0.01	258	0.38
ATL (Atlanta, GA)	316827	35840	0.11	132	0.42
OAK (Oakland, CA)	471378	9454	0.02	166	0.35

Table 3: Like Table 2, but for the entire years of 1999 and 2000. The correlation of column 4 and 6 is here 0.65.

airport	area	ovrflts	ovr/area	pireps	pireps*/area
ABQ	616843	778872	1.26	5266	8.54
CHI	258921	932496	3.60	5602	21.64
BOS	572151	348216	0.61	1709	2.99
DC	413105	1011000	2.45	3895	9.43
DEN	687536	1248384	1.82	13114	19.07
FTW	419770	695040	1.66	3401	8.10
HOU	580765	282528	0.49	3645	6.28
IND	239752	1562736	6.52	9497	39.61
JAX	494990	1082040	2.19	1805	3.65
KC	458859	1161216	2.53	10200	22.23
LAX	463547	255480	0.55	3458	7.46
SLC	1071881	958128	0.89	11170	10.42
MIA	326840	237048	0.72	986	3.02
MEM	368603	1323072	3.59	3401	9.23
MSP	985467	1092552	1.11	7660	7.77
NY	83412	781104	9.36	1416	16.98
CLE	230969	1480248	6.41	4626	20.03
SEA	673589	90264	0.13	6531	9.70
ATL	316827	847584	2.68	2630	8.30
OAK	471378	242760	0.52	3604	7.65

Table 4: Consistency of *pd0* and *pd1* values corresponding to same thresholds for probabilities estimated on the training set and predicted for the test set. The numbers in row i should belong to the interval $((i-1)/10, i/10)$ for the models to produce consistent estimates (notice that we are not asking that both pd0 and pd1 be in the same interval, but just that pd0 for the training set and the test set be in the same interval, and pd1 for the training set and the test set be in the same interval; in fact, given the shape of a tipical curve there is a trade off between pd0 and pd1 and low values of one correspond to high values of the other. The table is trying to economize space, but should be considered a lookup table for pd0 and pd1 independently of each other. Consistency is necessary in order to allow an operational implementation: by choosing a pair (*pd0*,*pd1*) on the curve estimated by the training set (whatever this pair's values are), and deriving the corresponding threshold for the probability of a positive event on the basis of which classifying independent observations, we can expect to see similar (*pd0*,*pd1*) for the latter.

	$FDA+MARS$		$FDA+NN$		*logistic*	
	pd0	*pd1*	*pd0*	*pd1*	*pd0*	*pd1*
$(0, 0.1)$	0.05	0.04	0.05	0.04	0.05	0.04
$(0.1, 0.2)$	0.14	0.13	0.14	0.13	0.14	0.14
$(0.2, 0.3)$	0.24	0.23	0.23	0.22	0.24	0.23
$(0.3, 0.4)$	0.34	0.33	0.33	0.32	0.34	0.33
$(0.4, 0.5)$	0.44	0.42	0.42	0.42	0.44	0.42
$(0.5, 0.6)$	0.54	0.52	0.53	0.52	0.54	0.52
$(0.6, 0.7)$	0.64	0.62	0.62	0.62	0.64	0.62
$(0.7, 0.8)$	0.73	0.72	0.73	0.72	0.74	0.73
$(0.8, 0.9)$	0.84	0.83	0.83	0.83	0.84	0.83
$(0.9, 1.0)$	0.95	0.95	0.95	0.94	0.95	0.95

Table 5: The usage of the different indices in 12 models. The number at the intersection of column i and row j indicates the number of terms in model i which include index j.

day	03.12.99			03.13.99			03.14.97			03.15.99			Total
UTC	3pm	6pm	9pm	3pm	6pm	9pm	3pm	6pm	9pm	3pm	6pm	9pm	
Brown 1	1	2	2	2	2	2	1	1	2	3	2	2	22
Brown 2	1	2	1	1	0	1	0	0	0	0	1	1	8
Col-Pan	0	8	1	1	2	0	3	0	4	6	2	4	31
Ellrod1	3	0	1	1	0	0	3	0	1	0	0	0	9
Ellrod2	0	1	0	0	1	0	0	1	0	1	1	1	6
Ri	5	1	3	3	4	5	2	4	1	4	3	1	36
TKE3	12	11	10	12	12	13	7	9	8	11	14	11	130
Dutton	0	2	0	0	0	1	0	0	0	0	0	0	3
Endlich	5	4	0	1	2	1	5	4	5	3	3	1	34
NGM 1	0	2	1	1	1	0	0	0	0	1	0	2	8
NGM 2	2	2	3	2	3	2	1	1	2	0	0	1	19
VWS	0	0	0	0	0	0	0	0	0	1	0	0	1
HWS	4	0	1	4	1	6	3	3	4	0	6	2	34
ABS	0	0	0	0	0	0	0	0	0	0	0	0	0
AWG	1	2	0	0	0	0	1	0	0	0	0	0	4
Div. Ten.	1	0	0	2	1	0	0	2	2	0	2	2	12
TKE5	0	0	1	0	0	0	0	0	0	0	0	0	1

Table 6: The interactions between the different indices in the 12 models fitted. The number under the ith column of the jth row indicates the number of terms in the 12 models which include both variable i and variable j as an interaction effect. The table is by construction symmetric. The main diagonal contains the number of terms in the 12 models that contain the correspondent index alone.

index	Brown 1	Col-Pan	Ri	TKE3	Endlich	NGM 2	HWS
Brown 1	0	9	0	22	1	0	0
Col-Pan	9	1	1	21	6	1	1
Ri	0	1	0	31	4	0	6
TKE3	22	21	31	13	14	12	19
Endlich	1	6	4	14	1	0	17
NGM 2	0	1	0	12	0	6	2
HWS	0	1	6	19	17	2	1