3.5 A STATISTICAL METHOD FOR PROBABILISTIC TURBULANCE FORECASTS

Walter C. Kolczynski, Jr.*
Sue Ellen Haupt
Pennsylvania State University, University Park, PA

1. INTRODUCTION

Strong turbulence presents a serious hazard to commercial aircraft. Aircraft encounters with strong turbulence can cause injuries to passengers and crew and in extreme cases damage to the aircraft that can result in a crash. Therefore, quality forecasts of strong turbulence are of great importance to aviation. Unfortunately, the physical mechanisms responsible for strong turbulence are very complex, making forecasts of turbulence difficult.

This study is part of a contest to showcase artificial intelligence methods and statistical learning to predict the probability of turbulence strong enough to adversely affect aviation. A vast amount of data spanning 132 variables for 103 990 cases is provided, along with verification of whether moderate or greater (mog) turbulence was experienced. Entrants are to use this data to train their methods then make probabilistic turbulence predictions for a contest data set where the presence of turbulence is unknown to the entrants. Our study focuses on simple statistical methods for making probabilistic predictions of turbulence.

2. METHODOLOGY

To produce a probabilistic turbulence forecast, we first determine a predictive function of turbulence for each variable independently. The performance of each variable in predicting turbulence is then calculated for use as weights for an aggregate forecast combining the individual variable forecasts into a final turbulence forecast.

We assume that the probability *p* of mog turbulence is related to each observed variable by the logistic function by a generalized linear model:

$$\log\left(\frac{p(x)}{1-p(x)}\right)=\alpha+\beta x \tag{1a}$$

or, equivalently

$$p(x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{1b}$$

where *x* is the value of the variable and α and β are parameters to be determined via regression. This regression cannot be estimated using least squares, so we will use maximum likelihood to estimate the parameters α and β.

Following Casella and Berger (2001), the likelihood of this function is given by

$$L(\alpha,\beta\,|\,\mathbf{y})=\prod_{i=1}^{n}p(x_i)^{y_i}\left(1-p(x_i)\right)^{1-y_i} \tag{2}$$

If we define $F(z)\equiv e^z/(1+e^z)$ and $F_i\equiv F(x_i)$, the log-likelihood is

$$\log L(\alpha,\beta\,|\,\mathbf{y})=\sum_{i=1}^{n}\log(1-F_i)+y_i\log\left(\frac{F_i}{1-F_i}\right) \tag{3}$$

In order to determine the maximum likelihood, we must find the value where the derivative of both α and β is zero. This means we need to solve

$$\frac{\partial}{\partial\alpha}\log L(\alpha,\beta\,|\,\mathbf{y})=\sum_{i=1}^{n}(y_i-F_i)=0 \tag{3}$$

and

$$\frac{\partial}{\partial\beta}\log L(\alpha,\beta\,|\,\mathbf{y})=\sum_{i=1}^{n}x_i(y_i-F_i)=0 \tag{4}$$

simultaneously.

Equations (3) and (4) cannot be solved analytically, so we use the Newton-Raphson method to solve them numerically. Newton-Raphson is an iterative process that updates the estimated values of α and β at each iteration using the Taylor approximation. For equations (3) and (4), this is given by

$$\boldsymbol{\alpha}^{(k+1)}=\boldsymbol{\alpha}^{(k)}-H\left(\log L\left(\boldsymbol{\alpha}^{(k)}\,|\,\mathbf{y}\right)\right)\frac{\partial}{\partial\boldsymbol{\alpha}}\log L\left(\boldsymbol{\alpha}^{(k)}\,|\,\mathbf{y}\right) \tag{5}$$

where $\boldsymbol{\alpha}$ is the parameter vector $[\alpha,\beta]^T$ and *H* is the Hessian operator:

$$H=\begin{pmatrix}\dfrac{\partial^2}{\partial\alpha^2} & \dfrac{\partial^2}{\partial\alpha\,\partial\beta}\\ \dfrac{\partial^2}{\partial\alpha\,\partial\beta} & \dfrac{\partial^2}{\partial\beta^2}\end{pmatrix} \tag{6}$$

so that

$$-H\left(\log L\right)=\begin{pmatrix}\sum F_j(1-F_j) & \sum x_jF_j(1-F_j)\\ \sum x_jF_j(1-F_j) & \sum x_j^2F_j(1-F_j)\end{pmatrix} \tag{7}$$

Eq. (5) is iterated until α and β converge. In order to preserve data for testing purposes, roughly one-third of the available complete data is withheld from the regression. The portion that is used in the regression is

* *Corresponding Author Address:* Walter Kolczynski, Jr. Dept. of Meteorology, Pennsylvania State University, 503 Walker Bldg. State College, PA 16802; wck122@psu.edu.

the training data while the data withheld to evaluate the performance is the test data.

Since the final results of this method will be evaluated by the Briar Skill Score (BSS), our weighting method uses the BSS in determining the weights. First we eliminate all variables with a BSS of less than 0.15 for the training data. Then the weight assigned each remaining individual prediction is determined by the inverse square of the Briar Skill Score:

$$w_m = \frac{w'_m}{\sum w'_m} \qquad w'_m = \frac{1}{BSS^2} \tag{8}$$

where the BSS is defined by (Wilks 2006):

$$BSS = 1 - \frac{BS}{BS_{ref}} \qquad BS = \frac{1}{n}\sum_{i=1}^{n}\left(p(x_i) - o_i\right)^2 \tag{9}$$

with BS_{ref} being the Briar Score (BS) of a reference forecast of $p = \bar{o}$ and o a binary value determined by whether mog turbulence is observed (1) or not (0). Both the cutoff and the weight was determined after experimentation. The final predicted probability of mog turbulence is then the sum of the weights times the predictions:

$$p(\mathbf{x}) = \sum w_j p_j(x_j) = \sum \frac{w_j e^{\alpha_j + \beta_j x_j}}{1 + e^{\alpha_j + \beta_j x_j}} \tag{10}$$

3. RESULTS

We computed the BSS for three different, independent sets of data: the training data, the test data, and the contest data. The BSS of the training data is 0.205, indicating that this method is skillful. Skillful here is defined as having a BSS greater than zero (performing better than the reference forecast of climatological probability). The BSS for the test data is 0.201. The closeness of this score to the BSS for the training data indicates that the results are stable and that the skillful BSS for the training data is not a result from over-fitting the training data. Finally, the BSS for the contest data is 0.199, close to that for the test data. This result placed third out of four entrants, but was not statistically distinct from 2nd place.

4. CONCLUSIONS

The simple statistical method presented here shows skill in predicting moderate or greater turbulence, but there is still some room for improvement. One potential variation of this method would be to determine optimal individual variable weights using a linear regression, rather than weighting each variable prediction by the inverse of the Briar skill score squared. We could also eliminate the individual variable predictions of turbulence altogether and instead use all variables simultaneously in a generalized multi-linear model with the logistic function.

Acknowledgements. We would like to thank Steve Sullivan, Philippe Tissot and Gillian Peguero for their work in organizing this contest. We would also like to thank Caren Marzban for helpful discussion on how we may improve on out methods.

5. REFERENCES

Casella, G. and R.L. Berger, 2001: *Statistical Inference, Second Ed.* Duxbery, Pacific Grove, CA, 660 pp.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences, Second Ed.* Academic Press, Burlington, MA, 627 pp.