Importance-ranking of Climate Variables for Prediction of Damaging Straight-line Winds

In the past decade there has been explosive growth in the use of machine learning (ML) to predict thunderstorm hazards such as hail, aircraft turbulence, and tornadoes. However, relatively few researchers have focused on the threat from severe non-tornadic (or "straight-line") winds, which occur much more frequently and are capable of causing great damage. We have developed machine-learning models that predict the occurrence of severe straight-line winds at lead times up to one hour. Our main objective is to have these models adopted by operational forecasters.

Three types of data are used for this project: archived radar grids from the Multi-year Reanalysis for Remotely Sensed Storms (MYRORSS); proxy soundings from the North American Regional Reanalysis (NARR); and surface wind observations from the Meteorological Assimilation Data Ingest System (MADIS), Oklahoma Mesonet, and one-minute METARs. The domain and time period are the continental United States from 2004-11 (excluding 2009). There are 306 days used for training, validation, and testing (all those with at least 100 unfiltered severe wind reports from the Storm Prediction Center).

First, storm cells are identified and tracked using w2segmotionll and w2besttrack, both of which are algorithms in the Warning Decision Support System with Integrated Information (WDSS-II). A "storm cell" is one thunderstorm at one time step; a "storm track" is the trajectory followed by a thunderstorm through time. Next, surface wind observations must be attributed to storm cells. Each wind observation is linked to the nearest storm cell, or none if there is no storm cell within 10 km. Then four types of features are calculated for each storm cell: statistics describing the shape of the bounding polygon; statistics for radar fields inside the bounding polygon; basic storm information (*e.g.*, area, speed, direction of motion); and sounding indices. Sounding indices are calculated from interpolated NARR data by the SHARPpy software. Overall, 565 features are calculated and used as ML predictors.

Due to the large number of features, our preferred ML algorithms are random forests and gradient-boosted regression trees. The dependent variable is the 90^{th}-percentile wind produced by each storm cell with a given lead time (from 15-60 minutes) and buffer distance (from 0-10 km). We have achieved good results for classification (cutoffs of 30 and 50 kt).

Several methods are used to rank feature importance in the best models, including *J*-measure ranking, sequential forward/backward selection, principal-component analysis, and permutation selection. We will present the most important features and link these, where possible, to climate change. For example, all methods rank low- to mid-level lapse rates among the most important predictors of straight-line winds. Straight-line winds are positively correlated with low- to mid-level lapse rates, which most climate models agree will decrease with global warming. This relationship alone suggest that the straight-line wind threat may decrease with global warming.

Although our main objective is operational forecasting, feature-ranking allows physical relationships to be inferred from ML models. We hope that knowledge presented here will be used to further the state of both thunderstorm and climate science.