Importance-ranking of Climate Variables for Damaging Straight-line Winds
Ryan Lagerquist\(^1\), Amy McGovern\(^2\), Travis Smith\(^3\), Michael Richman\(^2\), Valliappa Lakshmanan\(^3\)
\(^1\)Univ. of Oklahoma/CIMMS, \(^2\)Univ. of Oklahoma, \(^3\)The Climate Corporation

### Motivation
- Straight-line winds (microbursts, gust fronts, bow echoes, derechos, etc.) are one of the most damaging and least understood thunderstorm-related hazards.
- Machine learning (ML) has been used successfully in operational environments to predict thunderstorm-related hazards such as hail, tornadoes, and aircraft turbulence.
- We have developed ML models to predict the occurrence of damaging (50 kt or greater) straight-line winds at lead times of 15-60 minutes.
- We have used several methods to rank the importance of input variables to the best-performing ML models, some of which can be related to future climate scenarios.
- Our models will be incorporated into the Probabilistic Hazard Information (PHI) tool for the National Oceanic and Atmospheric Administration’s (NOAA) Spring 2016 Hazardous Weather Testbed (HWL), which allows forecasters to test new research products.

### Input Data and Processing (more)
- Four processing steps:
  1. **Storm ID and tracking.** Storms are identified from -10 °C reflectivity and tracked with two algorithms, w2segmotonil (Lakshmanan and Smith 2010) and a MATLAB adaptation of w2besttrack (Lakshmanan et al. 2015). w2besttrack processes and improves tracks from w2segmotonil (Figure 1).
  2. **Linkage of wind observations to storm cells.** Each wind observation is linked to the nearest storm track within 10 km (Figure 2).
  3. **Creation of proxy soundings.** NARR data are interpolated in space and time to each storm cell (Figure 4).
  4. **Feature calculation.** Four types of features for each storm cell:
     - a) **Sounding parameters.** Calculated from NARR soundings with SHARPy software (Halbert et al. 2015).
     - b) **Radar features.** Statistics (e.g., mean, median, skewness) for each radar variable (e.g., comp reflectivity, MESH, VI) inside storm cell (pixelated outline in Figure 3).
     - c) **Basic storm info.** Speed, direction, area, etc.
     - d) **Shape characteristics** (e.g., eccentricity, curvature, solility).

### Importance-ranking Procedure

#### J-measure Ranking
- Storm cells were classified by the 90th-percentile wind speed (U\(_{90}\)) produced at 15-60 minutes lead time (Y = 1 if U\(_{90}\) > 50 kt).
- Ranks each variable by the divergence between its probability density functions (PDFs) for positive (Y = 1) and negative events (Y = 0).
- Thus, the J-measure of a variable \(X\) is as follows.
  \[
  J(X) = \frac{1}{2} \left( \frac{P(X|Y=1) - P(X|Y=0)}{P(X|Y=1) + P(X|Y=0)} \right)
  \]

- Also, we generalized J-measure ranking to do explicit variable selection:
  1. Find the remaining variable with the highest \(J(X)\).
  2. Eliminate remaining variables for which the 95% CI J-measure does not overlap with that of \(X\) and 95th-percentile absolute Pearson correlation (also based on bootstrapping) with \(X\), \(r > 0.3\). (In other words, eliminate variables that are correlated with but less important than \(X\)).
  3. Repeat steps 1-2 until there are no variables left.

#### Sequential Forward Selection (SFS)

- SFS is a wrapper approach (independent of underlying ML model).
- It is a stepwise approach (considers effect of variable on performance of underlying ML model).
- Underlying model was logistic regression, trained to predict whether \(Y = 0 \) or \( Y = 1 \) (defined above).
- At each step \( k \), SFS adds the best remaining variable to the model, until model performance no longer improves.