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

Ryan Lagerquist[1], Amy McGovern[2], Travis Smith[1], 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, derechoes, 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 (HWT), which allows forecasters to test new research products.

Input Data and Processing

Table 1: Input data for ML models.

Data Type	Sources	Characteristics
Radar grids	Multi-year Reanalysis for Remotely Sensed Storms (MYRORSS)	1-km and 5-minute resolution, available for 2004-11 (excl. 2009) over CONUS
Near-storm environment	North American Regional Reanalysis (NARR)	32-km and 3-hour resolution, available for 1979-pres over CONUS
Surface wind observations	Meteorological Assimilation Data Ingest System (MADIS), Oklahoma Mesonet, 1-minute METAR reports	Variable resolution, available for 2001-pres over CONUS

- Datasets overlap for 2004-11 (excl. 2009) over the CONUS.
- We used 306 days for training, validation, and testing (all days in the seven years with at least 100 severe-wind reports from the Storm Prediction Center).

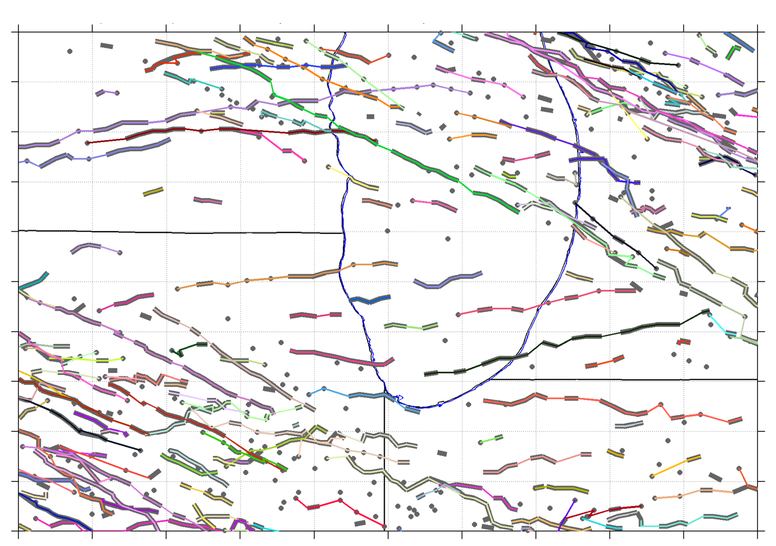

FIG. 1. Difference between w2segmotionll (thick grey lines) and w2besttrack (thin multi-coloured lines). w2besttrack results in longer storm tracks, which allows predictions to be made at longer lead times.

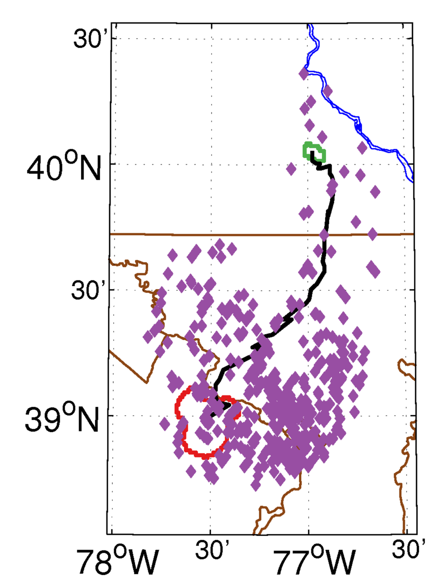

FIG. 2. Wind observations (purple) linked with a single storm track (black) from beginning (green polygon) to end (red polygon).

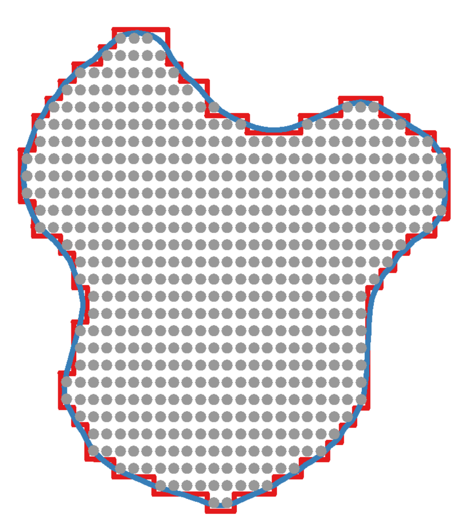

FIG. 3. Pixelated (red) and smoothed (blue) outlines of a storm cell. Pixelated outline comes from the 1-km MYRORSS grid and is used as a "cookie-cutter" to extract data from grid points (grey) inside.

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, w2segmotionll (Lakshmanan and Smith 2010) and a MATLAB adaptation of w2besttrack (Lakshmanan *et al.* 2015). w2besttrack processes and improves tracks from w2segmotionll (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 SHARPpy software (Halbert *et al.* 2015).
     b) **Radar features.** Statistics (*e.g.,* mean, median, skewness) for each radar variable (*e.g.,* comp reflectivity, MESH, VIL) inside storm cell (pixelated outline in Figure 3).
     c) **Basic storm info.** Speed, direction, area, etc.
     d) **Shape characteristics** (*e.g.,* eccentricity, curvature, solidity).

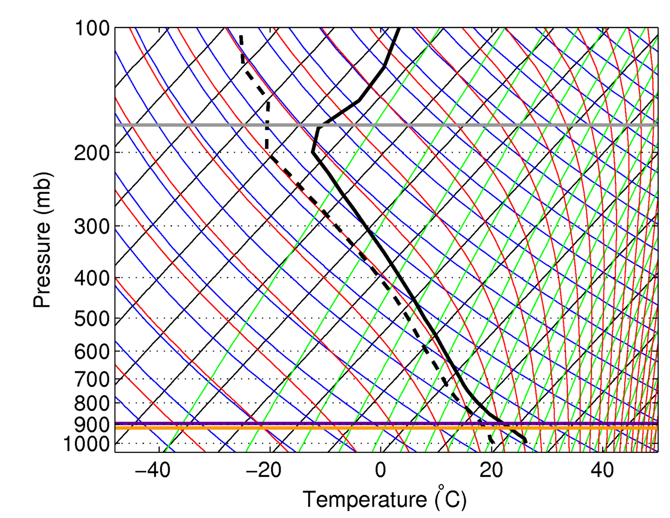

FIG. 4. Proxy sounding from NARR, interpolated to time and centroid of storm cell.

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 $U_{90} \geq 50$ 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_i is as follows.

$$J_i = \sum_{x \in X_i} [P(X_i = x \mid Y = 0) - P(X_i = x \mid Y = 1)] \log_2 \left\{ \frac{P(X_i = x \mid Y = 0)}{P(X_i = x \mid Y = 1)} \right\}$$

- Also, we generalized *J*-measure ranking to do explicit variable selection:
  1. Find the remaining variable with the highest *J*-measure (X_i*).
  2. Eliminate remaining variables for which the 95% CI *J*-measure does not overlap with that of X_i* and 5th-percentile absolute Pearson correlation (also based on bootstrapping) with X_i* is ≥ 0.3. (In other words, eliminate variables that are correlated with but less important than X_i*).
  3. Repeat steps 1-2 until there are no variables left.

Sequential Forward Selection (SFS)

- *J*-measure ranking is a filter approach (independent of underlying ML model).
- SFS is a wrapper approach (considers effect of variable on performance of underlying ML model).

- Underlying model was logistic regression, trained to predict whether Y = 0 or 1 (defined above).
- At each step *k*, SFS adds the best remaining variable to the model, until model performance no longer improves.

Results

- Table 2 shows the top 20 variables selected by both methods.

Table 2: Top 20 variables selected by importance-ranking.

Rank	*J*-measure Ranking	Sequential Forward Selection
1	700—500-mb lapse rate	3—6-km lapse rate
2	3—6-km lapse rate	Cosine of 0—1-km mean wind
3	850—500-mb lapse rate	Storm-cell age
4	5th percentile of 18-dBZ echo top	Magnitude of 0—3-km mean storm-relative wind
5	5th percentile of 0°C reflectivity	0—3-km lapse rate
6	25th percentile of 0°C reflectivity	Precipitable water
7	5th percentile of composite reflectivity	25th percentile of -20°C reflectivity
8	Mean 0°C reflectivity	CIN (convective inhibition)
9	Minimum 18-dBZ echo top	Mean -20°C reflectivity
10	Minimum composite reflectivity	5th percentile of -20°C reflectivity
11	Mean lowest-altitude reflectivity	Mean column mixing ratio
12	75th percentile of lowest-altitude reflectivity	Standard deviation of gradient of composite reflectivity
13	25th percentile of lowest-altitude reflectivity	Mean composite reflectivity
14	Time change (over 5 minutes) of minimum MESH	*v*-term in bulk Richardson number (akin to shear of v-wind)
15	Median 0°C reflectivity	Surface relative humidity
16	Minimum 0°C reflectivity	Sine of 0—3-km shear
17	Median lowest-altitude reflectivity	Standard deviation of gradient of 50-dBZ echo top
18	Max gradient of 50-dBZ echo top	Maximum gradient of composite reflectivity
19	Skewness of 50-dBZ echo top	Mean -10°C reflectivity
20	Time change (over 5 minutes) of 5th-percentile MESH	MCS (mesoscale convective system) maintenance probability

- We focus on sounding parameters, since these are the easiest to relate to climate.
- The top three variables for *J*-measure ranking, and two of the top five for SFS, are low- and mid-level lapse rates.
- Precipitable water, mixing ratio, and other moisture variables frequently appear.

- PDFs used to calculate *J*-measures (below) show that lapse rates (moisture variables) are positively (negatively) correlated with damaging straight-line winds.
- In general, climate models suggest that lapse rates (moisture) will decrease (increase) in the mid-latitudes, both of which would decrease the threat from damaging straight-line winds.
- Our method can "red flag" such relationships for further investigation by modelers (data science feeding physical science).

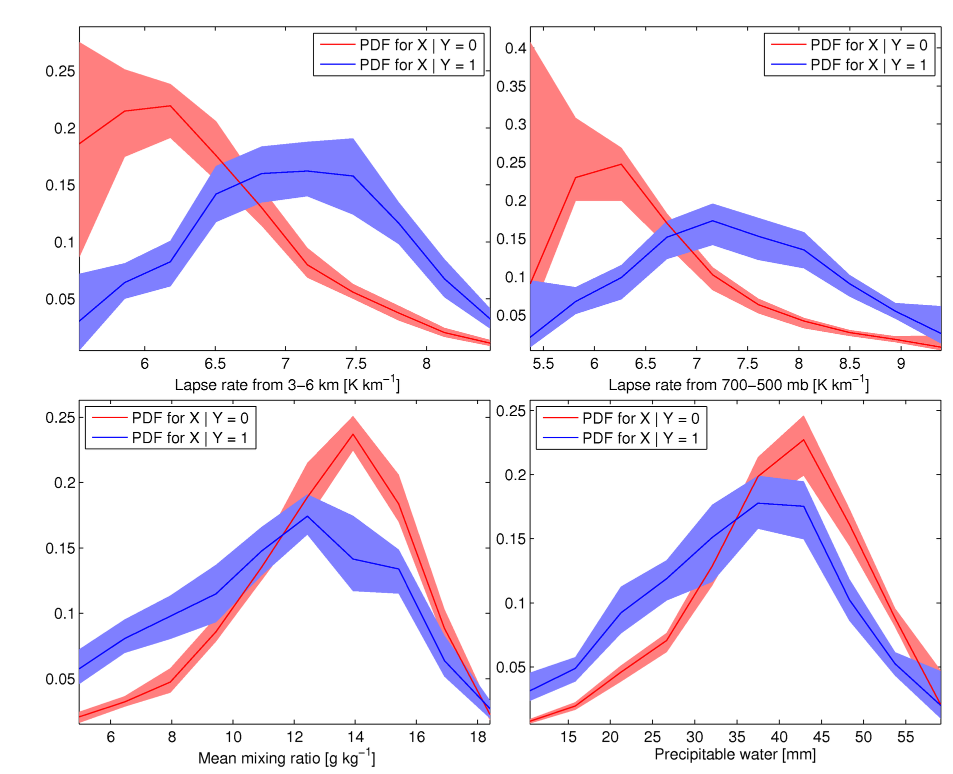

FIG. 5. Empirical probability density functions of different predictor variables under Y = 1 (90th-percentile storm wind > 50 kt) and Y = 0. The y-axis is probability.

WORKS CITED

Lakshmanan, Valliappa, and Travis Smith. "Evaluating a Storm Tracking Algorithm." *26th Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology.* 2010.

Lakshmanan, Valliappa, Benjamin Herzog, and Darrel Kingfield. "A Method for Extracting Postevent Storm Tracks." *Journal of Applied Meteorology and Climatology* 54.2 (2015): 451-462.

Halbert, K.T., W.G. Blumberg, and P.T. Marsh, 2015: "SHARPpy: Fueling the Python Cult." Preprints, *5th Symposium on Advances in Modeling and Analysis Using Python*, Phoenix AZ.