



Modeling and Correcting Reporting Biases in the Storm Prediction Center Tornado Database

Corey Potvin (1,2,3), Chris Broyles (4), Patrick Skinner (1,2), and Harold Brooks (2)

(1) Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma, United States, (2) NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma, United States, (3) School of Meteorology, University of Oklahoma, Norman, Oklahoma, United States, (4) NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma, United States

Maximizing the value of tornado climatologies requires accounting for unreported or mischaracterized tornadoes, especially where people and damage indicators are sparse. Previous attempts to spatially model tornado reporting bias have used only a single variable (e.g., population density) or a combination of variables that are implicitly assumed to have mutually independent effects. The proposed methodology, on the other hand, uses multivariable polynomial regression to simultaneously account for several variables, along with potentially important variable interactions (e.g., reduced importance of population density to reporting bias for tornadoes occurring near major highways). Reporting bias estimates are computed using tornado counts valid over nearby 100K cities (population > 100,000) or NWS Weather Forecast Offices (WFOs) as truth. The regression model is then applied to these noisy (due to sampling error) estimates to produce new reporting bias estimates that eliminate most of the noise while preserving spatial relationships with the geopolitical variables. While the approach is developed using the U.S. Storm Prediction Center (SPC) tornado database, it can be valuably applied to anywhere tornado locations are recorded and GIS data on population, cities, and/or roads are available.

We model tornado reporting bias east of the Rocky Mountains during 1975-2014 with various combinations of five variables: 1) distance from nearest 100K city or WFO, 2) distance from nearest 30K city or interstate, 3) population density, 4) terrain ruggedness, and 5) road density. In cross-validation tests, variables 1-4 together account for the most variance in reporting bias. Estimates of large-scale [$O(1000\text{ km})$] reporting bias are not unduly sensitive to the number of regression variables, indicating useful information can be gained from limited geopolitical data. However, cross-validation tests and geographic maps of modeled bias suggest more complex regressions substantially improve bias estimates at smaller scales. The resulting improvements to tornado hazard models are potentially valuable to forecasters, severe storm and climate scientists, and insurance/reinsurance companies.

The regressions indicate 46 % of tornadoes in the analysis domain were reported, with reporting rate decreasing by half as distance to nearest 100K city increases to 50 km. Reporting biases are especially pronounced earlier in the record and for shorter-track tornadoes, but remain large for more recent and for longer-track tornadoes. Surprisingly, underestimation of tornado frequency increases with damage rating; for example, the actual frequency of EF/F 3-5 tornadoes appears to be nearly three times that in the record. This underscores the problem of under-rating tornadoes in rural areas.