STATISTICAL MODELS FOR PREDICTING TORNADO RATES: CASE STUDIES FROM OKLAHOMA AND THE MID SOUTH USA

JAMES B. ELSNER, TYLER FRICKER, THOMAS H. JAGGER, & VICTOR MESEV
Department of Geography, Florida State University.

ABSTRACT

The destructive impact tornadoes have on communities has sparked interest in predicting the risk of impacts on seasonal time scales. Here, the authors demonstrate how to build statistical models for predicting tornado rates. They test the models with tornado counts accumulated over a 45-year period aggregated to counties in the State of Oklahoma and to cells in a latitude/longitude grid across a large portion of south central United States. The spatial model provides a fit to the counts, which includes terms for the spatial correlation and the population effect. A space-time model not only provides a similar fit to annual counts but also includes a term for a time-varying climate factor. This work contributes to methods for forecasting severe convective storms on the seasonal time scale.

Keywords: climate, risk prediction, space-time model, statistical model, tornadoes.

1 INTRODUCTION

Seasonal climate forecasts are now a matter of routine. Predictions of how much rain and heat can be expected during the summer are issued during spring by weather agencies across the globe, by region and by country. Even single seasonal predictions of, say how many hurricanes can be expected along a coastline are available and accurate enough to warrant attention by the property insurance industry. However, what is missing from the suite of seasonal forecasting products are long-range forecasts of severe convective storm activity. Potentially useful skills (accuracy above random guess) at predicting tornado activity prior to the start of the season has been noted recently [1,2]. However, given the large gaps in our knowledge of how climate influences severe weather and the dearth of methods to forecast it on the seasonal scale, basic and applied research is needed, which focuses on statistical modeling, diagnostic understanding, and methods to predict. A major impediment to issuing seasonal convective storm forecasts is that the events of interest are too small (e.g. tornado) to be resolved within the current dynamical forecast models. Long-lead predictions of severe weather environments can be made with dynamical models but the necessary conditions do not sufficiently distinguish between days with and without tornadoes.

An alternative approach is to fit a statistical model to a historical tornado database. Climate patterns related to active and inactive seasons provide the essential information to make predictions. However, population growth and changes to procedures for rating tornadoes result in a heterogeneous database. Various methods for dealing with data artifacts have been proposed [3–5] with most assuming a uniform region of activity and estimating occurrence rates within a subset of the region likely to be most accurate. For example, tornado reports are

This paper is part of the Proceedings of the 10th International Conference on Risk Analysis (RISK 2016)
www.witconferences.com

ISSN: 2041-9031 (paper format), ISSN: 2041-904X (online), http://www.witpress.com/journals
DOI: 10.2495/SAFE-V6-N1-1-9

often aggregated using kernel smoothing [6–8]. Spatial density maps that show regions of higher and lower tornado frequency are useful for exploratory analysis and hypothesis generation. However, correctly interpreting the patterns is a problem since there is no way to control for environmental factors. Another drawback is the implicit assumption that tornadoes occur randomly. This is not the case in general as a single thunderstorm can spawn a cluster of tornadoes over a compact area [9]. In addition, tornado reports tend to be more numerous near cities as compared to rural areas but this spatial variation is decreasing with time [10]. Improvements in observing practices over time tend to result in more tornado reports, especially reports of weak tornadoes [11,12].

The purpose of this paper is to show how to build statistical models to forecast the rate of tornadoes (see also [13]). The models can be used to establish benchmarks against which future statistical and eventually dynamical models can be judged. The models are written with the open-source R language using freely-available government data including tornadoes from the U.S. Storm Prediction Center (SPC) in Oklahoma, population and administrative boundaries from the U.S. Census Bureau. The paper begins by considering the long-term risk of tornadoes. This is done at the county level of an individual state to establish a baseline level (climatology) of forecast skill and to illustrate how to utilize the uneven and incomplete tornado database. Then it considers how long-term risk gets modulated by climate factors. This is done with a space-time model applied to the data aggregated in cells of a regular grid. An example is presented that quantifies the influence of El Niño on tornado activity across the central United States.

2 LONG-TERM RATES

The long-term risk of tornado presence at the county level is examined using data from the State of Oklahoma. Oklahoma, with an area of about 180,000 km^2, is located in the south central region of the United States where tornadoes are common occurrence. County administrative boundaries are downloaded and read into R as vector polygons at a resolution of 1:5 million and subset by the area of interest using the Federal Information Processing Standard (FIPS) code (40 for Oklahoma). The 2012 population estimate is added to each of the 77 counties as part of the attribute table. Osage County in Oklahoma is the largest with an area of 5,912 km^2 and Washington County is the smallest with an area of 1,089 km^2 .

The SPC maintains the most comprehensive and up-to-date tornado database in the world. Records extend back to 1950 and include information on time of occurrence, location, magnitude, track length and width, fatalities, injuries, and property loss for tornadoes in the United States. The version of the SPC database used in this study generates “shapefiles,” with each tornado represented as a straight-line track in a Lambert conformal conic (LCC) projection centered on 96° W longitude and parallels at 33° and 45° N latitudes.

In this paper, we consider tornadoes from the database over the period 1970–2014, inclusive (45 years). The start year coincides with a period of reliable records in the database for even the weakest and least damaging tornadoes. We buffer each track using the track-width dimension to create a polygon that approximates the tornado path. The width of the actual damage path varies along the track. Prior to 1994 the track-width dimension was estimated as the average damage path width along the track. Since then it is estimated as the maximum damage path width along the track. On an average, paths are wider for tornadoes with a higher damage rating [14]. Reports having coincident time, location, length, and width represent 1.8% of all tornado reports and are removed from further analysis.

There are 2,421 tornadoes in the database whose paths intersected at least part of the state. The statewide average number of tornadoes in the counties is 36 with a standard deviation of 14. Osage County in the northeast has been hit most frequently with 78 tornadoes over the period of record (Fig. 1). Choctaw County in the southeast has been hit least frequently with only 14 tornadoes. The correlation between the number of tornadoes and the size of the county is +0.44 [(+0.27, +0.58) 90% confidence interval (CI)]. The correlation between the number of tornadoes and the number of people is slightly higher at +0.45 [(+0.28, +0.59) 90% CI]. Larger counties tend to be somewhat less populated with a correlation between area and population at −0.13.

Raw counts are not directly useful for assessing tornado rate because the counties vary in size and population. To address this we employ a spatial statistical model. The model includes population density as a fixed effect.

In addition, to account for improvements in the procedures to rank tornadoes by the amount of damage, the calendar year and an interaction term of year with population are included. Finally, to account year-to-year changes, a random effect term is added. Mathematically, the number of tornadoes in each county s (T_s) is assumed to be described by a negative binomial distribution with parameters probability p and size r [1]. If X is a random sample from this distribution, then the probability that $X = k$ is $P(k|r, p) = \binom{k+r-1}{k}(1-p)^r p^k$, for $k \in 0, \ldots, \infty$, $p \in (0, 1)$ and $r > 0$. This relates the probability of observing k successes before the r failure of a series of independent events with the probability of success equal to p. The distribution is generalized by allowing r to be any positive real number and it arises from a Poisson distribution whose rate parameter has a gamma distribution [13].

The distribution is re-formulated using the mean $\mu = r\dfrac{p}{1-p}$ and the size r. This allows a separation of the mean effect from the dispersion parameter. The mean of the negative binomial distribution, μ_s is linked to a linear combination of the predictors and random effects, ν_s through the exponential function and the area of the cell, A_s (exposure). The

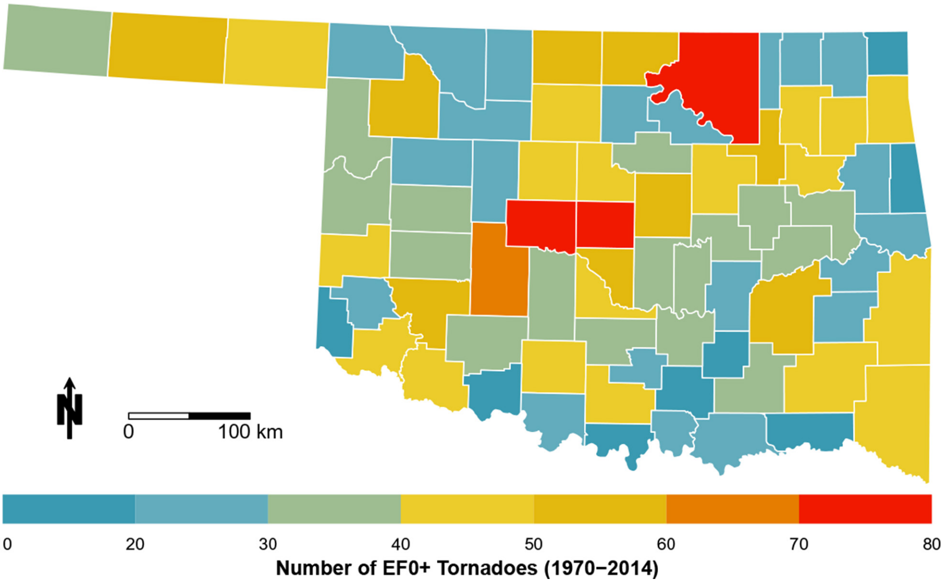

Figure 1: Number of tornadoes over the period 1970–2014.

dispersion is modeled with a scaled size parameter n where $n = r_s / A_s$ giving a dispersion of $1/p_s = 1 + \mu_s / n = 1 + \exp(v_s)/n$ that depends only on the tornado rate and n. More concisely, the model is:

$$T_s \mid \mu_s, r_s \sim \text{NegBin}(\mu_s, r_s) \tag{1}$$

$$\mu_s = A_s \exp(v_s) \tag{2}$$

$$v_s = \beta_0 + \beta_1 \,\text{lpd}_s + \beta_2 (t - t_0) + \beta_3 \,\text{lpd}_s (t - t_0) + u_s + v_t \tag{3}$$

$$r_s = A_s \, n \tag{4}$$

where NegBin(μ_s, r_s) indicates that the conditional tornado counts ($T_s \,|\mu_s$, r_s) are described by a negative binomial distribution with mean μ_s and size r_s, lpd$_s$ represents the base two logarithm of the population density during 2012 for each county, and t_0 is the base year set to 1991 (middle year of the record). The spatially correlated random effects u_s follows an intrinsic Besag formulation with a sum-to-zero constraint [15].

$$u_i \mid \left\{u_{j,j\neq i,}\tau\right\} \sim N\left(\frac{1}{m_i}\sum_{i\sim j} u_j, \frac{1}{m_i}\tau\right), \tag{5}$$

where N is the normal distribution with mean $1/m_i \cdot \sum_{i\sim j} u_j$ and variance $1/m_i \cdot 1/\tau$ where m_i is the number of neighbors of cell i and τ is the precision; $i \sim j$ indicates that cells i and j are neighbors. Neighboring cells are determined by contiguity (queen's rule). The annual uncorrelated random effect, v_t, is modeled as a sequence of normally distributed random variables, with mean zero and variance $1/\tau'$. The prior on the vector of spatial random effects is statistically independent from the vector of annual random effects. Gaussian priors with low precision are assigned to the β's. To complete the model specification, the scaled size (n) is assigned a log-gamma prior and the precision parameters (τ and τ') are assigned a log-Gaussian prior [13]. Bayes rule results in posterior distributions for the model parameters using the method of integrated nested Laplace approximation (INLA) [16,17].

The random-effects term is the spatially correlated set of county-level residuals, which quantifies tornado occurrence statewide accounting for population, exposure, and trends. Values of this term indicate where tornadoes are more likely, relative to the state average. Multiplying these county-level values by the statewide rate for 2014 gives the expected annual tornado rate per county (Fig. 2). Values range from a minimum rate of 0.72 per year in McCurtain County in the southeast corner of the state to a maximum rate of 1.19 per year in Grant County in the north central part of the State. Although the differences between the lowest and highest rates are less than a factor of two, the map features an axis of the highest rates from southwest to northeast through the Oklahoma City area. The rates are for the county as a whole and assume uniform risk within. They are normalized for exposure, and therefore, integrating the rate over the area and over the period of record yields an estimate of the total count. Uncertainty on the magnitude of these values is measured by the posterior standard deviation and range from .08 to .18 per year. Standard deviations are lower (precision higher) in counties with more neighbors (away from the state borders). A combination of high rate (1.11 per year) and low standard deviation makes Kiowa County in the southwest arguably the most vulnerable county in the state.

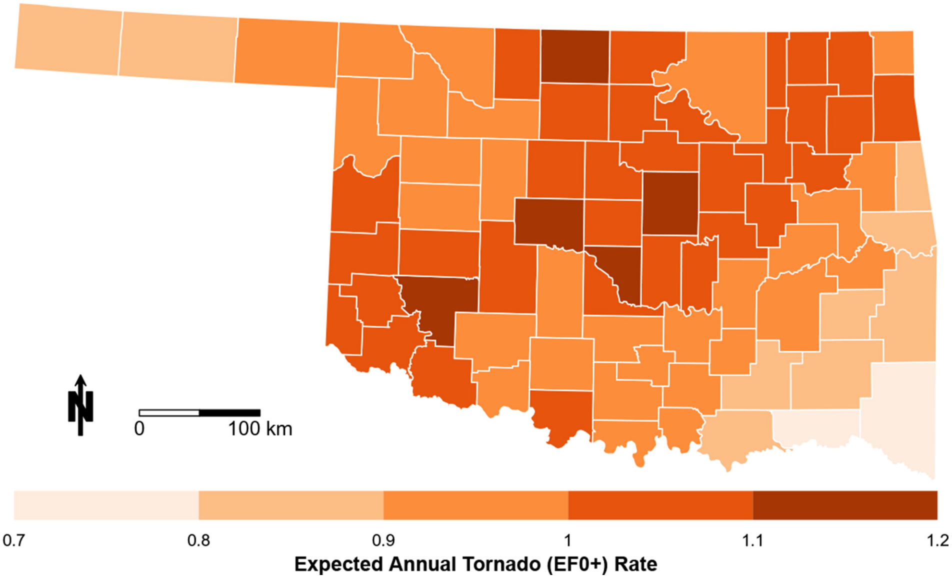

Figure 2: Annual rate of tornadoes accounting for exposure and population.

3 CONDITIONAL RATES

The county-level model is an example of how to create a baseline climatology of tornado risk using the available database. The model allows the insurance industry to set realistic rates of losses and emergency managers to effectively allocate state resources. However, depending on the climate pattern during the tornado season each year is often quite different in terms of the tornado risk. In particular, variations in sea-surface temperature and atmospheric convection in the tropical Pacific associated with the El Niño/Southern Oscillation (ENSO) modulate global weather and climate patterns including the risk of tornadoes [2, 18–21]. During the La Niña phase, a strengthened Inter-American Seas (IAS) low-level jet enhances the spread of moisture across the southeast during spring. Greater instability associated with the extra moisture is coupled with increased shear from a strengthened upper-level jet, setting the stage for severe convective storms.

Next, we demonstrate a space-time model fit to data aggregated in cells on a regular grid that quantifies how much the long-term rates should be adjusted based on a climate factor. Tornado counts are accumulated in each two-degree grid cell using all tornado paths that intersect the cell during each year (Fig. 3). The result is a space-time data set with constant-time attributes that include grid area, elevation, and variable-time attributes that include the annual number of tornadoes and population density. The spatial framework is raster [22] with a domain that extends from eastern Colorado to western Virginia and from the Mexican Gulf coast to southern Minnesota. The period of record runs from 1954 to 2014 for a total of 38,690 tornadoes representing 67% of all U.S. tornadoes and 87.5% of all high-energy tornadoes (EF4+).

The space-time model extends the spatial model above. Subscripts on parameters and variables now indicate a time component. Specifically, the tornado count in grid cell s for year t is given as:

$$T_{s,t} \mid \mu_{s,t}, r_{s,t} \sim \text{NegBin}\left(\mu_{s,t}, r_{s,t}\right)$$

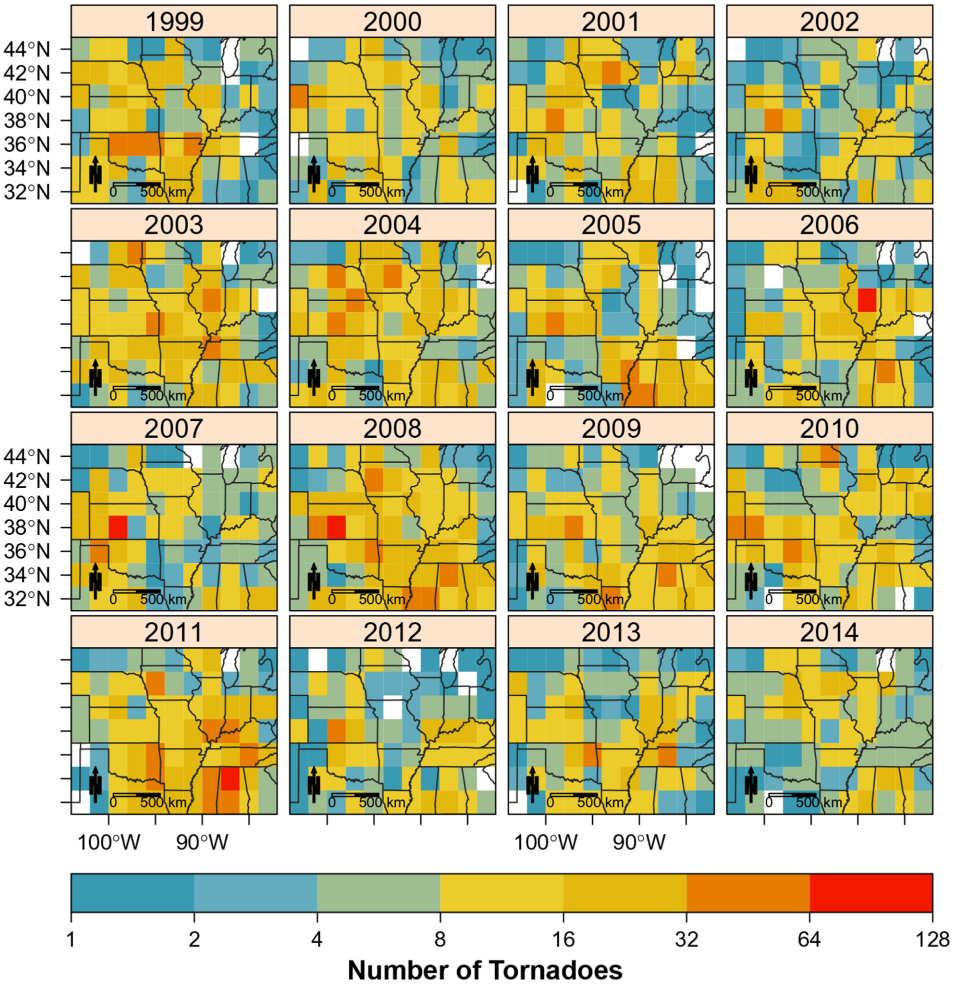

Figure 3: Annual tornado counts over the period 1999–2014 in two-degree grid cells.

$$\mu_{s,t} = A_s \exp(\nu_{s,t})$$

$$\nu_{s,t} = \beta_0 + \beta_1 \,\mathrm{rd}_s + \beta_2 (t - t_0) + \beta_{3,s}\,\mathrm{ENSO}_t + u_s + v_t$$

$$r_{s,t} = A_s\, n$$

where the conditional tornado count in each cell is described by a negative binomial distribution with mean $\mu_{s,t}$ and size $r_{s,t}$. At this scale, road density (rd) replaces population density to account for the observation bias. The effect of ENSO varies spatially through the spatial-effects term ($\beta_{3,s}$), which has an intrinsic Besag formulation (see Eqn. (5)). The variable ENSO$_t$ is the bi-variate ENSO time series averaged from March through May. The monthly series combines a standardized Southern Oscillation Index with a standardized Niño 3.4 sea-surface temperature series obtained from the Earth System Research Laboratory,

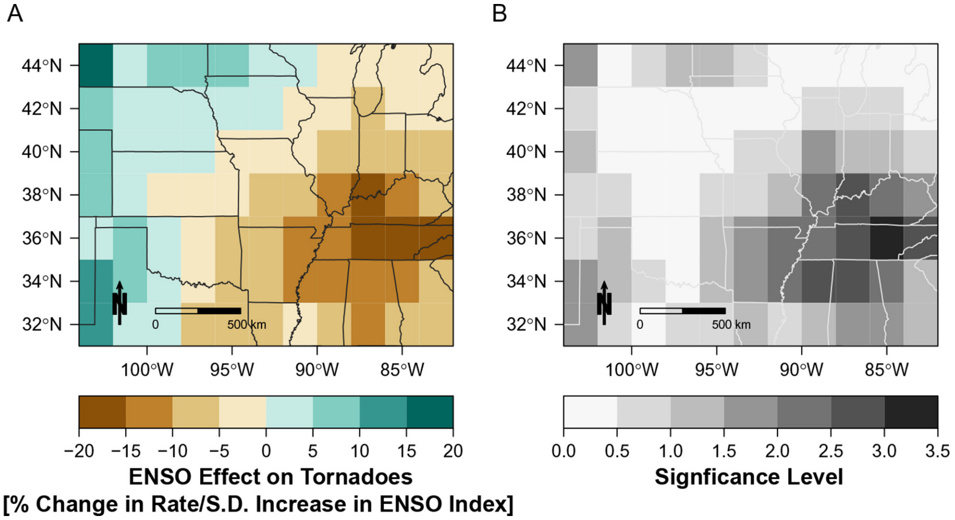

Figure 4: ENSO effect on tornadoes. (Left) Magnitude of the effect in units of percentage change in tornado rate per standard deviation (s.d.) increase in the springtime (Mar–May) value of the bi-variate ENSO index. (Right) Significance of the effect as computed by the ratio of the s.d. to the mean from the posterior distribution.

Physical Science Division. The cellarea times the number of years indicates the square-meter-years exposed to tornadoes. The exposure is normalized to have a mean of one. Again, INLA is used to obtain posterior distributions for the model parameters.

The posterior mean of the spatial-effects term (Fig. 4) answers the question: what is the geographic pattern of the ENSO effect on tornadoes controlling for data biases? The ENSO effect is most pronounced over the southeast with a reduction in the annual tornado rate exceeding 15% over a large part of Tennessee. The effect over this region exceeds two and three standard deviations (right panel). The reduction in tornado activity extends westward to northeastern Texas and southeastern Kansas and northward into eastern Wisconsin and Michigan. An enhancement in tornado activity occurs over the western High Plains from western Texas northward to western South Dakota. A physical explanation behind this relationship is tied to a reduction of the moist low-level jet from the Caribbean and Gulf of Mexico (the Intra-Americas Sea) [20] during the El Niño [23]. A prediction calling for an El Niño during spring would, according to this model, indicate a reduced risk of tornadoes across the mid south and an enhanced risk of tornadoes across the High Plains.

4 SUMMARY

This paper describes a method to produce a baseline climatology that accurately reflects where tornadoes are more and less likely to occur independent of the observation biases in the database. Further, it describes a space-time model to forecast tornado frequency from time-varying climate factors (e.g. El Niño). The research advances practices in tornado climatology through application of spatial statistical models. The available storm reports and covariate information can be classified into subsets by area of interest and aggregated by areal units (regular grids or irregular polygons – e.g. state counties). Aggregation accommodates additional human and environmental data (population, terrain, percent agriculture, etc.).

The models are fit using the method of INLA to solve the Bayesian integrals. This setup facilitates non-normally distributed counts and correlated residuals. The random-effects term quantifies where tornado activity is high (and by how much) relative to the regional average. The models make it simple to test hypotheses about the relationship between tornadoes and climate. The spatial-effects term quantifies where the climate factor has the greatest influence on tornado activity.

The models are practical and portable. Climatological rate estimates at the county level can be used by the property insurance industry to set homeowner insurance policy rates. They can be used by emergency managers to allocate resources weighted by areas of the state more prone to tornadoes. The conditional rate model can be used for planning for the next year given the predicted states of the climate factors. Models can also be fit to data separated by seasons. The computer code to estimate the long-term and conditional rates is freely available on github and can be modified with little effort to other tornado-prone states and regions.

REFERENCES

[1] Elsner, J.B. & Widen, H.M., Predicting spring tornado activity in the central great plains by March 1st. *Monthly Weather Review*, **142**, pp. 259–267, 2014.
http://dx.doi.org/10.1175/MWR-D-13-00014.1

[2] Allen, J.T., Tippett, M.K. & Sobel, A.H., Influence of the El Niño/Southern Oscillation on tornado and hail frequency in the United States. *Nature Geosciences*, **8**, pp. 278–283, 2015.
http://dx.doi.org/10.1038/ngeo2385

[3] King, P., On the absence of population bias in the tornado climatology of southwestern Ontario. *Weather and Forecasting*, **12**, pp. 939–946, 1997.
http://dx.doi.org/10.1175/1520-0434(1997)012<0939:OTAOPB>2.0.CO;2

[4] Ray, P.S., Bieringer, P., Niu, X. & Whissel, B., An improved estimate of tornado occurrence in the central plains of the United States. *Monthly Weather Review*, **131**, pp. 1026–1031, 2003.
http://dx.doi.org/10.1175/1520-0493(2003)131<1026:AIEOTO>2.0.CO;2

[5] Anderson, C.J., Wikle, C.K. & Zhou, Q., Population influences on tornado reports in the United States. *Weather and Forecasting*, **22**, pp. 571–579, 2007.
http://dx.doi.org/10.1175/WAF997.1

[6] Brooks, H.E., Doswell, C.A. & Kay, M.P., Climatological estimates of local daily tornado probability for the United States. *Weather and Forecasting*, **18**, pp. 626–640, 2003.
http://dx.doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2

[7] Dixon, P.G., Mercer, A.E., Choi, J. & Allen, J.S., Tornado risk analysis: is dixie alley an extension of tornado alley? *Bulletin of the American Meteorological Society*, **92**, pp. 433–441, 2011.
http://dx.doi.org/10.1175/2010BAMS3102.1

[8] Shafer, C.M. & Doswell, C.A., Using kernel density estimation to identify, rank, and classify severe weather outbreak events. *Electronic Journal of Severe Storms Meteorology*, **6**, pp. 1–28, 2011.

[9] Doswell, C.A. & Burgess, D.W., On some issues of United States Tornado climatology. *Monthly Weather Review*, **116**, pp. 495–501, 1988.
http://dx.doi.org/10.1175/1520-0493(1988)116<0495:OSIOUS>2.0.CO;2

[10] Elsner, J.B., Michaels, L.E., Scheitlin, K.N. & Elsner, I.J., The decreasing population bias in tornado reports. *Weather, Climate, and Society*, **5**, pp. 221–232, 2013. http://dx.doi.org/10.1175/WCAS-D-12-00040.1

[11] Doswell, C.A., Brooks, H.E. & Kay, M.P., Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Weather and Forecasting*, **20**, pp. 577–595, 2005. http://dx.doi.org/10.1175/WAF866.1

[12] Verbout, S.M., Brooks, H.E., Leslie, L.M. & Schultz, D.M., Evolution of the U.S. tornado database: 1954-2003. *Weather and Forecasting*, **21**, pp. 86–93, 2006. http://dx.doi.org/10.1175/WAF910.1

[13] Jagger, T.H., Elsner, J.B. & Widen, H.M., A statistical model for regional tornado climate studies. *PLoS ONE*, **10**(8), p. e0131876, 2015.

[14] Elsner, J.B., Jagger, T.H. & Elsner, I.J., Tornado intensity estimated from damage path dimensions. *PLoS ONE*, **9**(9), p. e107571, 2014.

[15] Besag, J., Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **24**, pp. 179–195, 1975. http://dx.doi.org/10.2307/2987782

[16] Rue, H., Martino, S. & Chopin, N., Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, pp. 319–392, 2009. http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x

[17] Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A. & Krainski, E.T., *INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion*, 2014 + 0100). R package version 0.0-1417182342.

[18] Marzban, C. & Schaefer, J.T., The correlation between US tornadoes and Pacific sea surface temperatures. *Monthly Weather Review*, **129**(4), pp. 884–895, 2001. http://dx.doi.org/10.1175/1520-0493(2001)129<0884:TCBUST>2.0.CO;2

[19] Cook, A.R. & Schaefer, J.T., The relation of El Niño-Southern Oscillation (ENSO) to winter tornado outbreaks. *Monthly Weather Review*, **136**, pp. 3121–3137, 2008. http://dx.doi.org/10.1175/2007MWR2171.1

[20] Muñoz, E. & Enfield, D., The boreal spring variability of the Intra-Americas low-level jet and its relation with precipitation and tornadoes in the eastern United States. *Climate Dynamics*, **36**(1–2), pp. 247–259, 2011.

[21] Lee, S.K., Atlas, R., Enfield, D., Wang, C. & Liu, H., Is there an optimal ENSO pattern that enhances large-scale atmospheric processes conducive to tornado outbreaks in the United States? *Journal of Climate*, **26**, pp. 1626–1642, 2013. http://dx.doi.org/10.1175/JCLI-D-12-00128.1

[22] Hijmans, R.J., *raster: Geographic Data Analysis and Modeling*, 2015. R package version 2, pp. 4–18.

[23] Krishnamurthy, L., Vecchi, G.A., Msadek, R., Wittenberg, A., Delworth, T.L. & Zeng, F., The seasonality of the great plains low-level jet and ENSO relationship. *Journal of Climate*, **28**, pp. 4525–4544, 2015. http://dx.doi.org/10.1175/JCLI-D-14-00590.1