Comments on "Tornado Risk Analysis: Is Dixie Alley an Extension of Tornado Alley?"

—Patrick T. Marsh
School of Meteorology,
University of Oklahoma,
Norman, Oklahoma

Cooperative Institute for Mesoscale Meteorological Studies,
Norman, Oklahoma

NOAA/National Severe Storms Laboratory,
Norman, Oklahoma

—Harold E. Brooks
NOAA/National Severe Storms Laboratory,
Norman, Oklahoma

Dixon et al. (2011, hereafter DMCA11) present a tornado risk analysis that found parts of the southeast United States are the most tornado prone in the nation, instead of "Oklahoma, the state previously thought to be the maximum for tornado activity (Schaefer et al. 1986; Brooks et al. 2003)." Because both Brooks et al. (2003, hereafter BDK03) and DMCA11 employ kernel density estimation to achieve their depictions of tornado risk, a natural question that arises is why these analyses are so different. DMCA11 attempt to explain these differences as a consequence of their focus on tornado path length instead of tornado frequency. Although there is no question that the focus on slightly different underlying datasets affects the resulting analyses, the differences between the tornado path length and tornado frequency datasets are small enough that this explanation inadequately explains the differences between BDK03 and DMCA11. This comment offers a different explanation as to why the differences in the two different studies exist, one that focuses on the differences in how the kernel density estimation was conducted.

DOI:10.1175/BAMS-D-11-00200.1

In general, kernel density estimation is a nonparametric method of estimating the underlying probability density function (PDF) of a finite dataset. This is done by choosing a weighting function, or kernel, a measure of an area subject to the weighting function (this measure is often called the kernel bandwidth or simply bandwidth) and then applying the weighting function to the finite dataset over the prescribed area. A kernel can be any function $K(u)$ that is nonnegative, real valued, integrable, and satisfies (Wilks 2006)

$$\int K(u)\, du = 1 \tag{1}$$

and

$$K(-u) = K(u), \forall u. \tag{2}$$

Several common kernels exist and include those chosen by BDK03 and DMCA11. BDK03 used a Gaussian kernel with a bandwidth of 120 km; DMCA11, however, used an Epanechnikov kernel followed by a uniform kernel, both with a bandwidth of 25 miles (40.25 km). Although different kernels will result in different probability density functions, DMCA11 state that the bandwidth, not the kernel, "is much more important for density calculations as it determines the size of the surrounding area that will influence the value at any given point." Although this is true, it is misleading. Specific bandwidth values cannot be directly compared across different kernels (Marron and Nolan 1988). Herein lie the major differences between BDK03 and DMCA11.

Before considering the impact of the difference choices of bandwidth, let us examine the impact of kernel choice in one dimension. In the case of an Epanechnikov kernel, the one-dimensional representation is

$$K(x) = \frac{3}{4h}\left(1 - \left(\frac{x}{h}\right)^2\right)\mathbf{1}\left\{\left|\frac{x}{h}\right| \le \mathbf{1}\right\}, \tag{3}$$

where h is the bandwidth and

$$\mathbf{1}\left\{\left|\frac{x}{h}\right| \le \mathbf{1}\right\}$$

is the specific indicator function generalized as

$$\mathbf{1}\{|u| \le 1\} = \begin{cases} 1; & |u| \le 1 \\ 0; & |u| > 0 \end{cases}. \quad (4)$$

The one-dimensional uniform kernel is

$$K(x) = \frac{1}{2h}\mathbf{1}\left\{\left|\frac{x}{h}\right| \le 1\right\}, \quad (5)$$

and the one-dimensional Gaussian kernel is

$$K(x) = \frac{1}{\sqrt{2\pi h^2}} e^{\frac{1}{2}\left(\frac{x}{h}\right)^2}. \quad (6)$$

Applying the Epanechnikov kernel to a single point, (0, 1), results in a PDF maximized at $x = 0$ with a value of 0.75. This PDF falls off symmetrically to 0 for $|x| < h$ and is 0 for $|x| \ge h$. Applying the uniform kernel to the resulting PDF further reduces the maximum of the PDF to slightly larger than 0.5 and broadens the PDF such that the PDF symmetrically decreases to 0 for $|x| < 2h$. Applying a Gaussian kernel to the same initial point results in a PDF that is maximized at $x = 0$ with a value of

$$\frac{1}{\sqrt{2\pi h^2}}$$

and falls off symmetrically to 0 as $|x|$ approaches infinity (Fig. 1).

The smaller amplitude of the Gaussian PDF is a consequence of the Gaussian kernel's PDF being nonzero everywhere, whereas the Epanechnikov PDF is nonzero only when $|x| \le h$ and the combined Epanechnikov–uniform kernel PDF is nonzero only when $|x| \le 2h$. Using the results of Marron and Nolan (1988), it can be shown that, when comparing Epanechnikov and Gaussian kernels, the Epanechnikov kernel must be 2.2138 times larger than the Gaussian bandwidth to achieve a similar response function. It is apparent that the resulting one-dimensional PDFs are different between kernels for the same bandwidth. However, DMCA11 used a bandwidth that was one-third the size of BDK03. This choice of a smaller bandwidth acts to amplify the differences between the Gaussian and combined Epanechnikov–uniform PDFs (Fig. 1).

As striking as the effects of the different kernels and bandwidths are in one dimension, they are magnified in two dimensions. To illustrate this, consider a tornado with pathlength zero at location (0, 0). If we apply the two-dimensional Epanechnikov kernel with the DMCA11 bandwidth, approximately 40 km, and examine the output on a 5-km grid, we find that the maximum of the PDF is 0.0117, and nonzero probabilities are found out to a radius of 40 km (Fig. 2a). Application of a uniform distribution further smoothes the PDF, resulting in a maximum of the PDF of 0.0060 and nonzero probabilities out to a radius of 80 km (Fig. 2b).

Contrast the Epanechnikov and combined Epanechnikov–Gaussian probabilities with those generated from a Gaussian distribution of identical bandwidth (Fig. 2c). The maximum of the PDF is 0.0025, but nonzero probabilities extend to infinity in both the x and y directions. The maximum Gaussian probability is 4.68 times smaller than the Epanechnikov kernel and 2.4 times smaller than the combined Epanechnikov–uniform kernel. As was the case in one dimension, when comparing the Gaussian kernel with a bandwidth of 120 km to a combined Epanechnikov–uniform kernel with a bandwidth of 40 km, the differences are magnified further. The BDK03 kernel results in a maximum PDF value of 0.0003 (Fig. 2d), whereas the DMCA kernel is 0.0060.

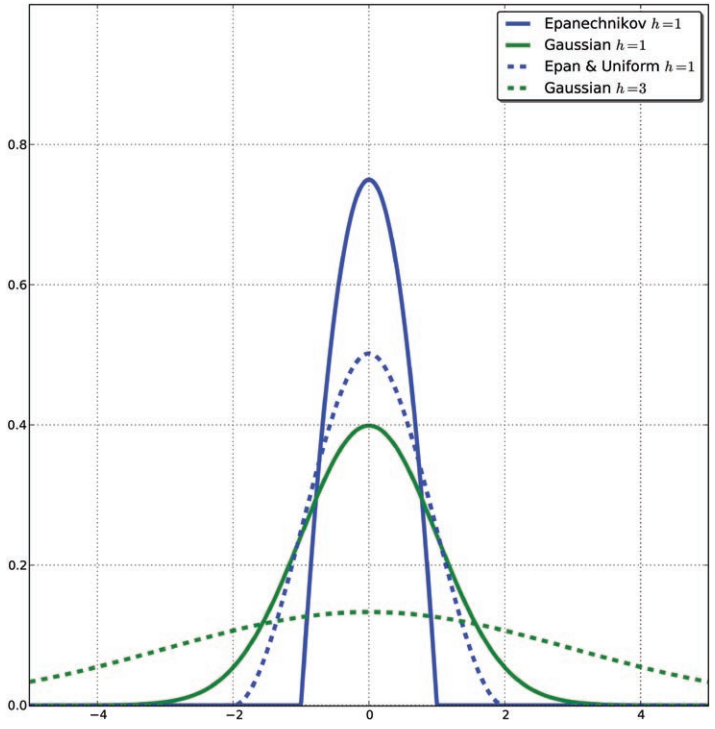

Fig. 1. Comparison of one-dimensional kernel density estimation for Epanechnikov kernel (solid blue line), combined Epanechnikov–uniform kernel (dashed blue line), and Gaussian kernel (solid green line) with bandwidth equal to 1. A Gaussian kernel with bandwidth of 3 is shown in the dashed green line.

Comparing the radii at which the different kernels produce probabilities that are 5% of the maximum PDF value offers a cursory comparison of the kernel's two-dimensional spatial extents. Using the previously described two-dimensional scenario, the Epanechnikov radius is approximately 38 km, whereas the Epanechnikov–uniform combined kernel is 64 km. The Gaussian kernels have much larger radii. The Gaussian kernel with the same bandwidth as the Epanechnikov and Epanechnikov–uniform kernels has a radius of approximately 97 km, and the Gaussian kernel used in BDK03 has a radius of approximately 293 km.

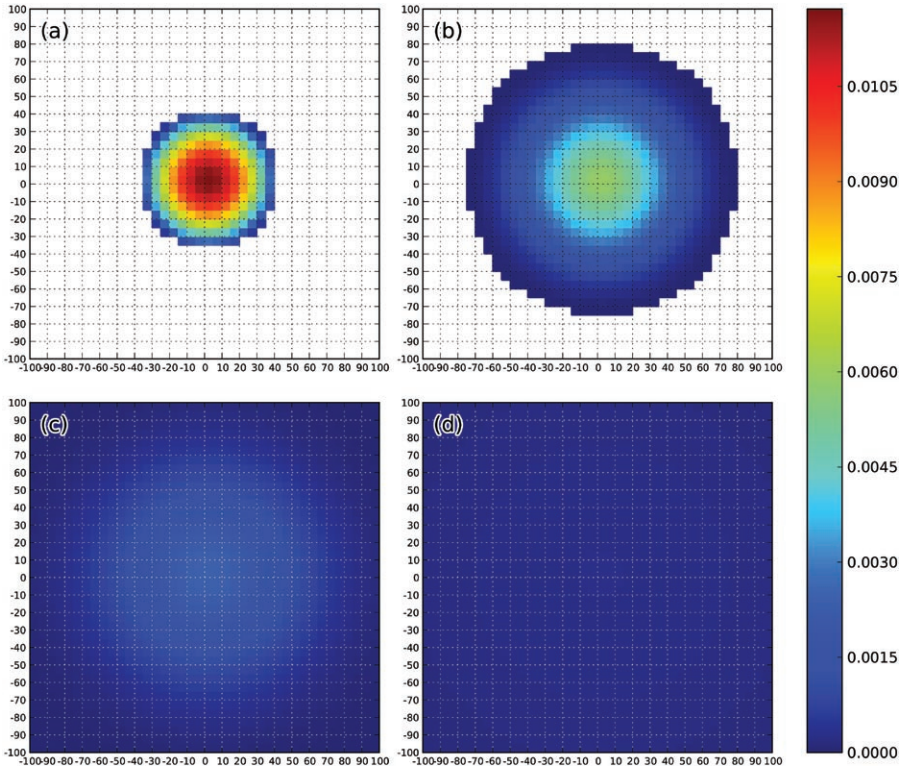

Fig. 2. Comparison of two-dimensional kernel density estimation for (a) Epanechnikov kernel, (b) combined Epanechnikov–uniform kernel, and (c) Gaussian kernel, all with bandwidth of 40 km. (d) A Gaussian kernel with bandwidth 120 km is depicted. All kernel density estimations are presented on a common grid with grid length of 5 km.

The practical implication of this for tornado risk analysis is highlighted with the following thought experiment. The area of the United States east of the Rocky Mountains (where climatologically most tornadoes occur) is approximately 6.4 million square kilometers. If we look at the area enclosed by 95% of the BDK03 smoother, an area of roughly 270,000 km^2 (4.2% of United States) is captured. However, 95% of the Epanechnikov–uniform combined smoother of DMCA11 is roughly 12,000 km^2 (0.19%). Furthermore, if we assume that each of the approximately 1,300 tornadoes per year is uniformly distributed, a point on a grid (any grid) using the BDK03 smoother will use information from approximately 55 tornadoes per year. The DMCA11 smoother uses information from approximately 2 tornadoes per year. As a result, the greater detail seen in DMCA11 arises from the choice of a much smaller kernel and a corresponding smaller effective sample size (Doswell 2007) is used to estimate the true distribution.

ACKNOWLEDGMENTS. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA17RJ1227, U.S. Department of Commerce.

REFERENCES

Brooks, H. E., C. A. Doswell, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting,* **18,** 626–640.

Dixon, P., A. Mercer, J. Choi, and J. Allen, 2011: Tornado risk analysis: Is Dixie Alley an extension of Tornado Alley? *Bull. Amer. Meteor. Soc.,* **92,** 433–441.

Doswell, C. A., 2007: Small sample size and data quality issues illustrated using tornado occurrence data. *Electron. J. Severe Storms Meteor.,* **2,** 1–16.

Marron, J., and D. Nolan, 1988: Canonical kernels for density estimation. *Stat. Probab. Lett.,* **7,** 195–199.

Schaefer, J., D. Kelly, and R. Abbey, 1986: A minimum assumption tornado-hazard probability model. *J. Appl. Meteor.,* **25,** 1934–1945.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. International Geophysics Series, Vol. 91, Elsevier, 627 pp.