Doswell, C. A. III, 2007: Small sample size and data quality issues illustrated using tornado occurrence data. *Electronic J. Severe Storms Meteor.*, **2** (5)**,** 1–16.

Electronic Journal of
SEVERE STORMS METEOROLOGY

Small Sample Size and Data Quality Issues Illustrated Using Tornado Occurrence Data

CHARLES A. DOSWELL III
Doswell Scientific Consulting
Norman, OK

(Submitted 27 January 2007; in final form 26 August 2007)

ABSTRACT

A major challenge in weather research is associated with the size of the data sample from which evidence can be presented in support of some hypothesis. This issue arises often in severe storm research, since severe storms are rare events, at least in any one place. Although large numbers of severe storm events (such as tornado occurrences) have been recorded, some attempts to reduce the impact of data quality problems within the record of tornado occurrences also can reduce the sample size to the point where it is too small to provide convincing evidence for certain types of conclusions. On the other hand, by carefully considering what sort of hypothesis to evaluate, it is possible to find strong enough signals in the data to test conclusions relatively rigorously. Examples from tornado occurrence data are used to illustrate the challenge posed by the interaction between sample size and data quality, and how it can be overcome by being careful to avoid asking more of the data than what they legitimately can provide. A discussion of what is needed to improve data quality is offered.

1. Introduction

For many research topics in meteorology, the issue of sample size is an important one. Tornado-related research, in particular, has many topics where sample sizes are insufficient to draw certain types of conclusions. There are numerous reasons for this to be an issue when studying the historical record of tornado occurrences, but at times the sheer size of the dataset can convince the unwary that sufficient data are available to draw robust conclusions, whereas that may not necessarily be the case.

The concept of sample size is evidently related to the number of observations. As discussed in Wilks (2006, p. 138), the so-called power of a hypothesis test is related to the sample size. However, the adequacy of the sample for hypothesis testing also is related to the sample *variability*. Low variability (as measured by, for example, the standard deviation of the distribution estimated from the sample) means that a relatively small sample is more likely to be sufficient than when the variability is high. But statistics is not a simple set of rules to follow in determining the meaning within a data set. Establishing the level of confidence in conclusions from the data cannot be done by rote. Careful consideration of data quality issues is also important in attempting to draw conclusions from data analysis.

Some of the notions in this paper are discussed in various other places—notably in Brooks et al. (2003 – hereafter BDK03), Doswell et al. (2005 – hereafter DBK05) and Verbout et al. (2006). The goal of this paper is to illustrate the importance of how sample size limitations, in combination with secular trends in the tornado occurrence data, can limit the ability to draw valid conclusions from tornado occurrence datasets. Commonly-used attempts to reduce the impact of these secular trends involve reducing the sample size in one way or another. Section 2

Corresponding author address: Dr. Charles A. Doswell III, Doswell Scientific Consulting, 1705 Wellesley Ct, Norman, OK 73071, E-mail: cdoswell@earthlink.net

provides an example illustrating the problem, and section 3 shows how another strong signal can be found in what is essentially the same data set. Section 4 concludes with a discussion about the implications from these two examples.

2. An illustration of small sample size problems

In order to provide a concrete example of how to recognize problems associated with small sample sizes, the database on occurrence of tornadoes that is maintained by the Storm Prediction Center (SPC) is used herein. These data are described in detail elsewhere (e.g., Schaefer and Edwards 1999) and some of the important caveats about them have also been mentioned in previous applications (e.g., BDK03). As part of a project concerning tornado outbreaks that is being done at the SPC (Doswell et al. 2006 –hereafter D06), it was decided to use recent historical tornado occurrence data, for the period 1970-2002. As discussed in BDK03 and D06, there are pronounced secular trends in the tornado occurrence data; that is, trends that are virtually certain to have a nonmeteorological origin. The farther back in time within the tornado occurrence data, the more influential these undesirable trends become on any quantitative analysis of the data. Hence, it was felt that a 33-year record of recent vintage was about as long a record as one could trust to be influenced as little as possible by these secular trends.

To find examples of tornado outbreaks, D06 made an arbitrary choice: the search began by identifying all days with seven or more reported tornadoes in the record. During the period of record, there were nearly 1400 such days, which represents roughly 10 percent of all dates within that period. This might imply to the unwary that this is a relatively large sample with which to do analysis. Note that in any given year, a "tornado day" (i.e., a day with *one* or more reported tornadoes) includes roughly half of the days in a year (Fig. 1).

Days with seven or more reported tornadoes occur considerably less often than tornado days. As Fig. 1 shows, when the threshold is raised from one to seven, to 10, 20, or 30 or more reported tornadoes in a day, the sample size decreases considerably. When considering days with 7 or more tornadoes, Fig. 1 exhibits the likely presence of a secular change in the data irrespective of the threshold chosen: a substantial increase in the frequency of days with a given number of tornadoes, after 1988. This coincides with a rapid increase in the frequency of reported F0 (on the Fujita scale for rating tornado intensity) tornadoes beginning after 1988 (Fig. 2), which likely can be associated with an increase in the emphasis within the National Weather Service on tornado warning verification statistics, and continuing growth in the number of storm chasers. Thus, even though the 33-year period was chosen to minimize the impact of known secular trends, it appears even these most recent data still contain such artifacts.

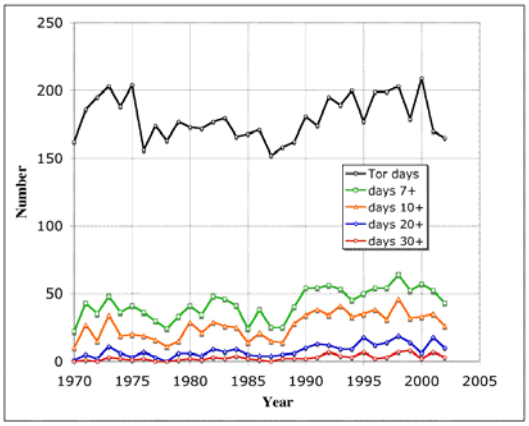

Figure 1. Number of tornado days (days with one or more reported tornadoes) per year, as well as the number of days with 7 or more, 10 or more, 20 or more, and 30 or more reported tornadoes, during the period 1970-2002.

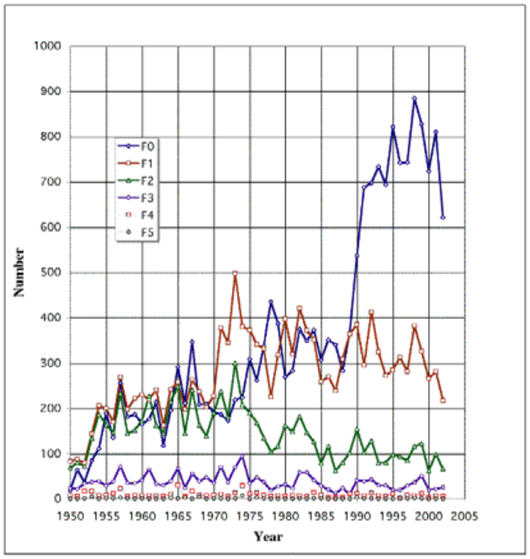

Figure 2. Time trends in the number of tornadoes, by F-scale for the period 1950-2003.

Another example of an artifact can be seen in Fig. 2: an apparent *decrease* in the frequency of tornadoes rated F2 or higher since the early 1970s. This change is likely due, at least in part, to the way tornadoes were given F-scale ratings prior to the early 1970s. There is an extensive discussion of the decrease in reports of F2 and stronger tornadoes in Verbout et al. (2006). Also, see Kelly et al. (1978) for a description of how the ratings were done for tornadoes prior the implementation of the F-Scale in the early 1970s (Fujita and Pearson 1973). Once structural engineers became involved in assessing storm damage (e.g., Minor et al. 1977), F-scale ratings began to account more carefully for the quality of home construction (see Doswell 2003) as well as the damage, per se.

Despite these known problems with the tornado occurrence data, it seems reasonable to propose that over the course of a year, there is some underlying, relatively smooth distribution of days with seven or more tornadoes. That is, if 1000 years of stable, reliable and accurate tornado occurrence observations were available, then when the frequency of such days was plotted as a function of the calendar date, the result should be fairly smooth. However, as Fig. 3 clearly demonstrates, the resulting frequency plot is far from smooth. Some dates, such as 19 April, appear to have an anomalously high frequency compared to adjacent dates on the plot. There are other dates, like 21 May, that have anomalously low frequencies. The issue becomes even more acute when the climatological frequency is low, because over the 33-year period, there are numerous dates in the fall and winter on which seven or more tornadoes did not occur at all, whereas nearby dates had several such occurrences.

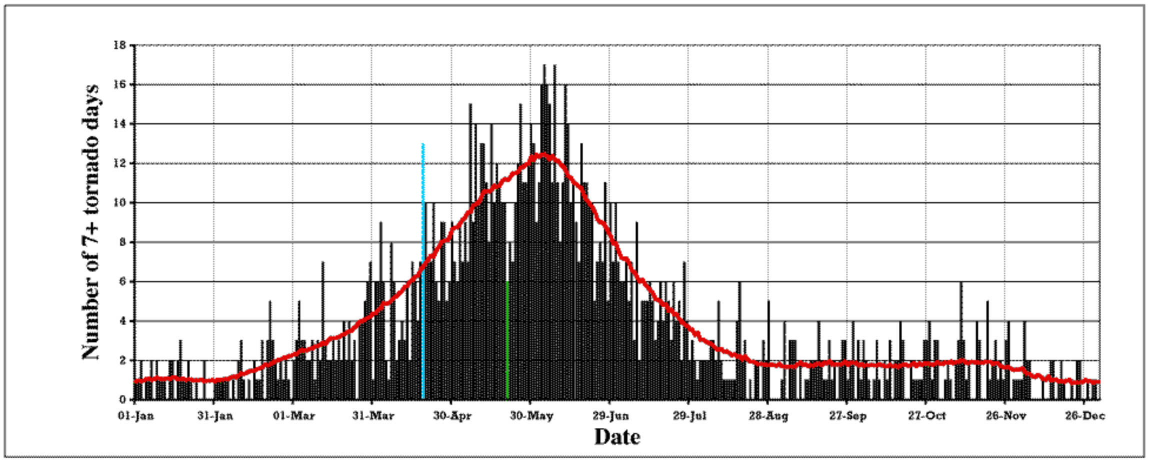

Figure 3. The number of days with 7 or more reported tornadoes during the period from 1970-2002, as a function of the calendar date (histogram). The dates of 19 April (aqua) and 21 May (green) on the histogram are highlighted. The red line is the result of passing a 61-day Gaussian smoother through the data.

It is possible to see overall trends within these data, nevertheless, and to attempt to discern the "true" day-by-day progression of the underlying frequency curve by smoothing. One type of smoothing is shown in Fig. 3—a heavy filter (a 61-day Gaussian kernel) produces a fairly smooth curve, which might represent a reasonable approximation to the unknown, underlying smooth distribution. This curve is similar to Fig. 3 in BDK03 for all tornado days as a function of the date. The fact that a heavier smoother is required to produce a smooth curve in Fig. 3 than was used for Fig. 3 in BDK03 is likely to be a direct result of the relatively small sample of tornado days with seven or more tornadoes compared to all tornado days. In fact, this can be interpreted as direct evidence of the primary issue confronting researchers seeking information from the tornado occurrence data: restricting attention to subsets of the data, which might make sense to do, creates sample size issues that either require special treatment, or that might make some research hypotheses effectively untestable with these data. For the data in Fig. 3, less heavy smoothing (not shown) results in "wiggles" on the smoothed curves that may or may not represent physically meaningful departures from a smooth curve, such as the relative minimum roughly centered on 21 May. A 33-year period of record is apparently insufficient to establish the statistical significance of such anomalies.

There is no reason to believe that one particular date is particularly favored compared to adjacent dates, even though the data seem to show this possibility. Any claim that a date like 19 April is "special" compared to 18 or 20 April would be comparable to a similar claim for the so-called "January thaw". The latter has been shown by Godfrey et al. (2002) to be virtually certain to be an illusion[1] caused by what amounts to small sample size relative to the variability in the data, albeit not so extreme as shown in Fig. 4.

The existence of high variability in the data, of course, does not imply by itself that a given sample is too small, although it *is* a clear indication that sample size might be an important issue. Suppose the mean of a random variable *X*, denoted by $\overline{X}$ and given by

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i ,$$

is of interest, where *N* denotes the sample size. The standard deviation of a sample of that size drawn from the distribution of *X* (which has a *population* mean μ and standard deviation σ), denoted as $\sigma_{\overline{X}}$, is given by

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

(Spiegel 1961, p. 144). Thus, as the variability in the data (represented by the population's standard deviation) increases, a larger sample is necessary to obtain a stable estimate of the mean. By this formula, if the variability in the distribution increases by a factor of two, the sample size must increase by a factor of four to achieve the same degree of stability in a computation of the sample mean. As *N* gets very large, of course, the variability of the sample mean becomes quite small, tending to zero as *N* increases to infinity.

In some cases, high variability is *expected* and so observations of that variability would not necessarily imply a sample size issue. For the data used to create Fig. 3, there is no plausible reason for assuming that the distribution of days with seven or more tornadoes should be anything but a relatively smooth function of the calendar date, which likely resembles that produced by the 61-day Gaussian smoother. Large variability when the expected distribution is smooth offers an important clue to the presence of potential sample size issues.

If the threshold for consideration were raised to 10, or 20, or 30 reported tornadoes on a given day, the impact of some of the secular trends (see BDK03, Verbout et al. 2006) *might* be reduced. However, the sample size issue would be even more problematic than for tornado days with seven or more reported tornadoes. Figure 1 shows that days with 10 or more reported tornadoes occur with about one-third the frequency of days with seven or more reported tornadoes. Similarly, if the F-scale criterion were raised from any tornado (F0 or stronger) to some higher threshold, such as F2 and stronger, presumably to mitigate the strong secular trends in the reporting of weak tornadoes, the resulting reduction in sample size (F2 and stronger tornadoes currently are reported at roughly one-tenth the frequency of tornadoes of any F-scale rating, as shown in Fig. 2) likely would offset this effort. Further, as already noted, there is reason to believe that a secular trend in F2 and stronger tornado occurrences is still present in the data.

Increasing the period of record *could* be attempted in order to overcome the sample size problem, but for the tornado occurrence data, the nonmeteorological artifacts in the data make this an option that could create at least as many problems as it would solve. The dilemma is that tornadoes are rare events in any one place and our knowledge of the interannual variability in tornado occurrence data is complicated by both secular trends and the specter of small sample sizes relative to the interannual variability (which can only be estimated from the data and is not known with any precision).

3. An example of a strong signal in the data

The notion of the inadequacy of the 33-year period of record for looking at relatively infrequent events like the occurrence of seven or more tornadoes on a given day should not be

[1] Note that *some* such anomalies might be statistically significant and have plausible physical explanations.

overgeneralized, however. For certain types of analysis, if the data contain a strong signal, then the sample still might be used to provide meaningful results. As an example of this, during a preliminary analysis of the data regarding the occurrence of seven or more tornadoes on a given day, it was observed that during the peak of the "tornado season" in the United States (April to June), there was a marked tendency for event days (i.e., days with seven or more reported tornadoes) to occur in “strings” of consecutive days. By contrast, outside of the tornado season, it was noticeably more likely that an event day would be isolated rather than being part of such a string. The occurrence of strings likely results from the fact that synoptic scale cyclones associated with tornadoes can take two or more days to traverse the region east of the Rocky Mountains (the most tornado-prone region within the United States). Further, more than one such system can occur in succession, so as one tornado-producing, synoptic-scale cyclone is moving through the eastern United States, another may be associated with other tornadoes on the western plains to the lee of the Rocky Mountains.

Outside the tornado season, it is common for such a synoptic-scale system to result in pushing the ingredients for tornadic storms so far apart that the next such system is unlikely to result in more than an isolated tornado or two. Within the season, however, it is relatively common for successive synoptic-scale cyclones to produce multiple tornadoes.

To reveal this apparently strong seasonal signal, each event day was categorized by the length of the string of consecutive days in which it occurred. For each date in the calendar year, the number of times every event day was associated with a string of each length was counted. On any given calendar date, the percentage of times an event day occurred as part of a string of two or more such days varied considerably (Fig. 4). Using the raw results of this count as a function of the day of the year obviously gives a noisy result. A relatively light smoother (a median filter of width 3 days, followed by a simple moving average of 15 days in length) produced a simpler distribution, although it still apparently contains some noise. Nevertheless, a strong signal seems to be emerging out of the noise.

In fact, outside of the April-June tornado season, strings of more than three consecutive event days simply did not occur in the 33-year period of record (not shown), whereas during the tornado season, strings of two or more event days were associated with more than half of all event days. Event day strings of six or more days occurred exclusively during the tornado season.

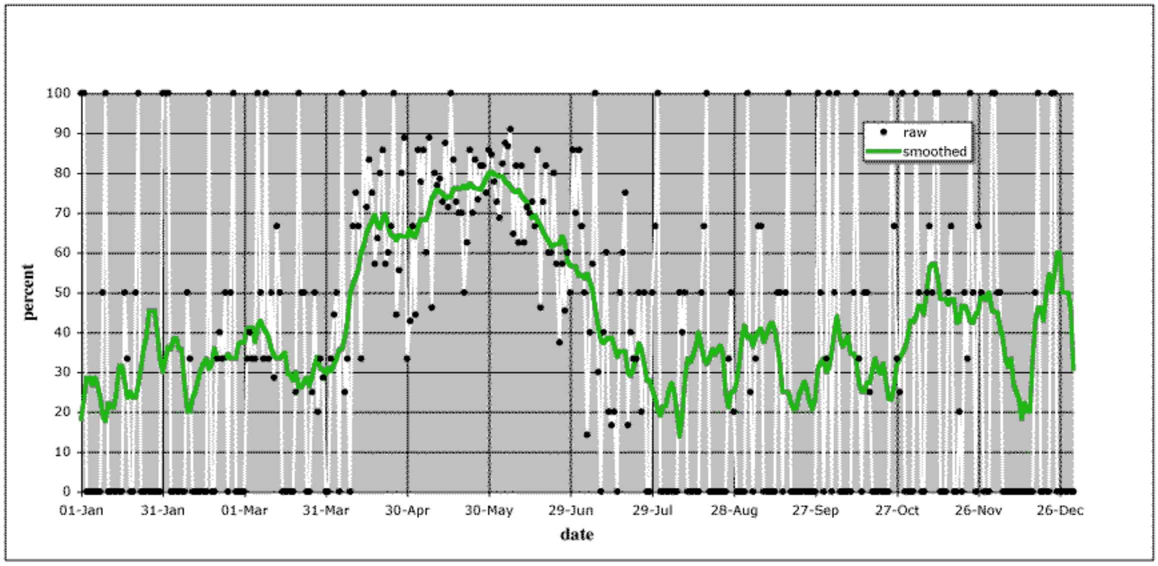

Figure 4. The percentage of event days occurring within strings of two or more consecutive days: raw (black dots and white lines) and smoothed (green line) as described in text.

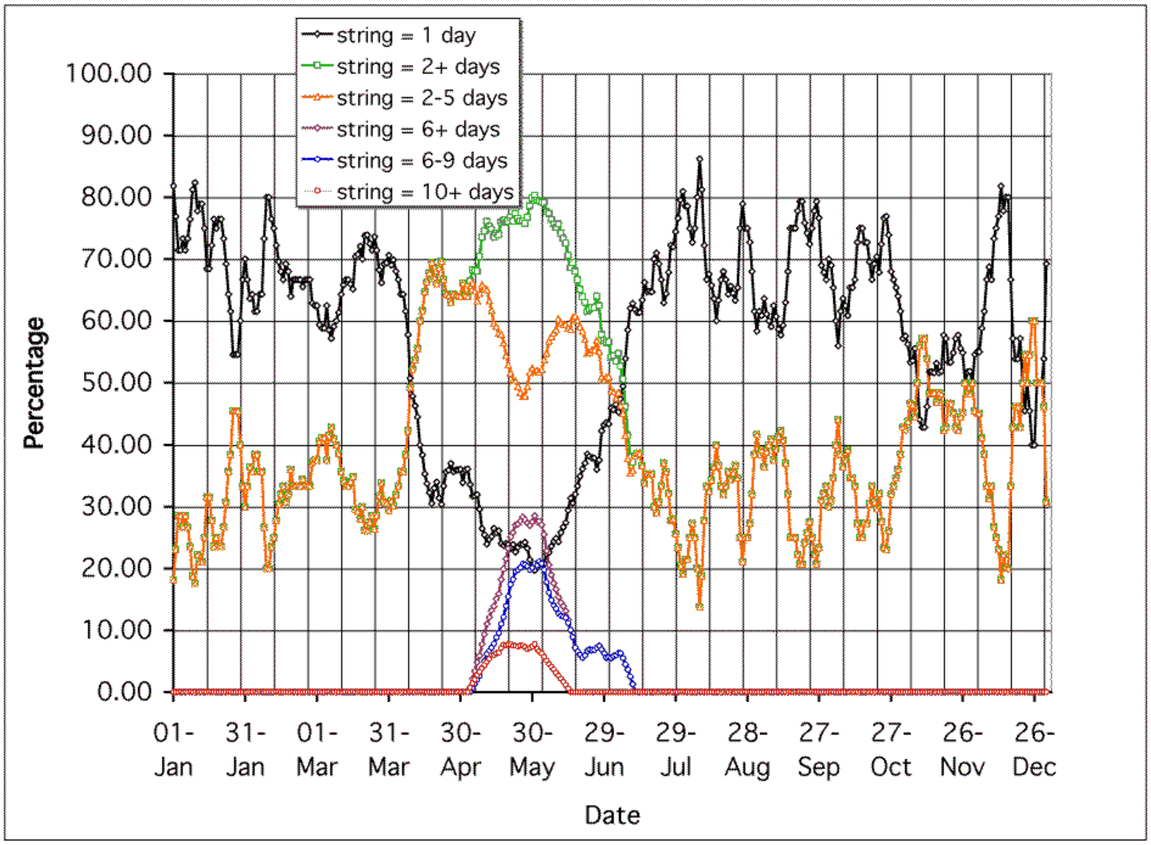

Figure 5. Filtered percentages (as in Fig. 4) of string categories of various lengths (see the key) as a function of the calendar date.

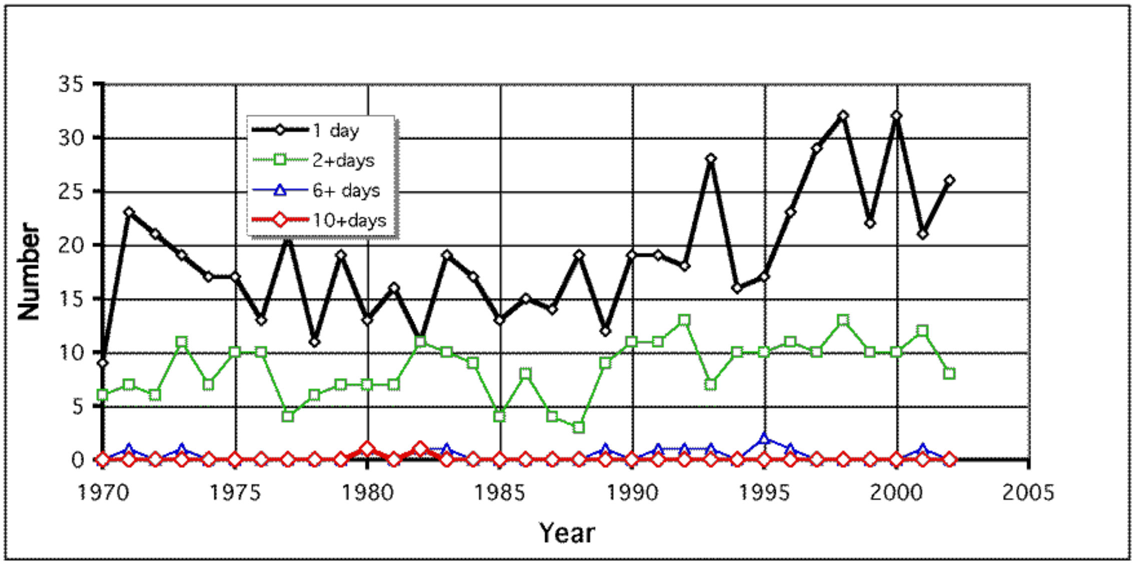

Figure 6. Number of strings of tornado event days of various lengths (see key) including "strings" of one day, by year.

To refine this emerging signal, several additional categories of string lengths were considered (Fig. 5). Any quantitative analysis or physical explanation of the reasons for this strong signal is outside of the scope of this note, but it is certainly associated with the seasonal evolution in shear- and buoyancy-related parameters (Doswell 2001). This analysis indicates that a sufficiently strong signal can emerge from fairly noisy data, but when the same data are used for other types of analysis, they could constitute too small a sample to offer reliable results.

Figure 5 also hints at the presence of a fall tornado season (note the increase in the 2+ string frequency in November), but the signal is nowhere nearly as strong as it is in the spring. The fall season is more sporadic than the spring season in terms of its interannual variability, so to be more confident in any quantitative analysis, a larger sample likely would be necessary.

The climatology of strings shows that over the 33 years, the number of strings with two or more event days hasn't changed very much (Fig. 6), despite a substantial increase in the number of tornadoes. It's also clear that the number of strings of more than five to six consecutive event days is small, so small as to suggest that the sample size for strings of more than several days is inadequate to say much. The record contains one string of 10 consecutive days with 7+ tornadoes (11 May–20 May 1982), and one with 15 consecutive days (24 May–07 June 1980). There were no strings of nine days and none between 10 and 15. Out to string lengths of about six days, the fit of the observations to an exponential decrease in the frequency as a function of string length is reasonably good (Fig. 7). Beyond that, the data suggest that a 33-year period of record is an inadequate sample. Really long strings of days with seven or more reported tornadoes (longer than 5-6 such days in a row) are too infrequent to obtain a reliable sample of them.

4. Discussion and conclusions

The temptation to draw unjustifiable conclusions from analysis of meteorological data is strong. Particularly vulnerable to this are attempts to relate tornado occurrence data to various large-scale processes with relatively long periods, such as the El Niño–Southern Oscillation (ENSO) cycle. For detecting a long-period cycle, the period of record should be long enough to contain *many* such cycles, anticipating that every cycle will not be exactly the same as every other cycle, owing to complicated interactions with other processes.

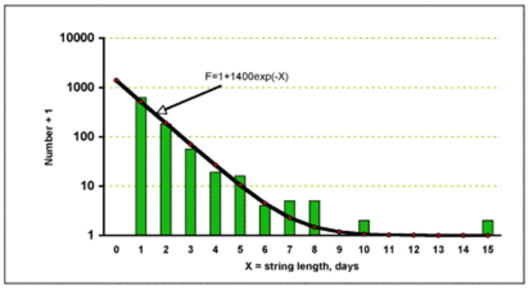

Figure 7. For the period 1970-2002, a histogram plot (in green) of the number of strings of consecutive days with 7 or more reported tornadoes. The number in each category was increased by one, to facilitate the use of a logarithmic scale. The superimposed curve represents an approximation to the distribution using the function *F*=1+1400exp(-X), where *X* represents the length of the string.

It might be tempting to consider a 33-year period of record at least marginally long enough to use for detecting cycles with periods of a few years (like the ENSO). Unfortunately, the presence of secular trends in the tornado occurrence data means that *long* periods of record contain artifacts that are difficult to deconvolve from real meteorological information. Shorter periods of record can reduce the impact of these artifacts, but are likely to create small sample size problems, unless the signal being sought is strong enough to show through the noise associated with relatively small sample sizes. The examples shown here should serve as a caveat to any researcher doing data analysis, but certainly are of considerable relevance to those using tornado occurrence data to validate hypotheses about the relationship between tornado occurrences and long-period cyclic processes affecting the weather.

The naive belief that a small sample (relative to the population variability) will be representative of what a large sample would show is unfortunately widespread. Figure 3 should make it evident that this is not a valid belief. Sample statistics (e.g., mean, variance, confidence intervals, etc.) are helpful, but do not provide conclusive evidence of the adequacy of a sample. In my own experience, an example of the dangers of a small sample comes to mind. During a study of High Plains severe weather, I considered only one year's worth of High Plains severe weather (in June and July) and found what appeared to be a very strong signal regarding the synoptic pattern (Doswell 1980) within which such events occurred. Since the chosen year included several cases, at the time I was not concerted about the sample size and the issue of my study's representativeness. Later, a more thorough study than mine was conducted by Weaver and Doesken (1991), who found that I had been lucky—their 10-year sample revealed that my one-year sample was fortuitously representative.

Unfortunately, there is no objective way to determine the sample size needed, either a priori or a posteriori. Textbooks (e.g., Wilks 2006) offer no simple formula for determining the minimum sample size, in part because the true underlying distributions are virtually never known in advance. All researchers typically have is a sample from which to make inferences about their data. To the extent that plausible assumptions about that unknown distribution can be made, based on reasoning that may include information that exists outside of the sample itself (e.g., physical arguments), we can use statistical analysis methods to test our hypotheses. But if our assumptions are violated or our reasoning is flawed, then the resulting statistical tests could be misleading or even completely invalid. Formal statistical hypothesis testing methodology can produce misleading findings about the level confidence in accepting or rejecting hypotheses. Of late, in fact, the whole formal procedure for hypothesis testing has come under considerable criticism [see Harlow et al. (1997) for some diverse essays on the topic] by statisticians.

On the basis of the first example shown here using a 33-year period of record, it should be evident that without heavy filtering, the number of days with seven or more tornadoes as a function of calendar date is highly variable from one date to the next. Any argument that a particular peak in that distribution represents a physically-based anomaly would be untenable, based on this sample. Of course, the *smoothed* version of that distribution provides some indication of what the actual day-to-day variation in the number of days with seven or more tornadoes might be. A strong seasonal preference for April-June tornadoes shows up clearly in the smoothed distributions. The smoothed distribution, however, does not show such a comparably strong signal in the fall.

The second example shows a reasonably strong seasonal signal in the number of consecutive days with seven or more reported tornadoes. During the spring tornado season, strings of 2 or more consecutive days occur frequently enough in the 33-year sample that this information could be used, for example, to plan what time of year should be set aside for researchers to conduct field observation campaigns or for storm chasers to schedule their chase vacations. In fact, the results shown in section 3 simply confirm what has been recognized by tornado researchers at least since the time of the first tornado forecasts of 1948 (see Doswell 2007): mid-April to mid-June is the best time to observe tornadoes and tornadic storms in the United States. As shown in BDK03, the majority of these tornadoes occur on the plains west of the Mississippi River, and the strong signal associated with the springtime tornado season is most evident on the plains, as well (see Fig. 8 in BDK03).

Although not demonstrated herein, BDK03's results make it similarly apparent that only the broad spatial structure of tornado occurrences[2] can be estimated with any confidence. Most of the spatial "detail" contained within the data is going to be plagued with the challenge of nonmeteorological effects on the sample, and their interaction with what is inevitably a small sample size in any given location.

Given the societal importance of severe storms and the growing concern that a changing climate might alter severe storm frequency, it seems logical to look to the tornado occurrence data to test hypotheses about possible impacts of global climate change on the temporal trend in tornado occurrence. Unfortunately, as shown herein, the existing observational data regarding tornado occurrences are plagued by secular trends that make it quite difficult to have confidence in trying to detect what might be rather subtle signals in tornado frequency as a result of climate change.

Finally, although reporting of tornado occurrences in the United States is the best in the world, it is evident that we continue to be

[2] As presented, for instance, at: http://www.nssl.noaa.gov/hazard/index.html

plagued by various problems. How might the reporting of tornado occurrences be improved? I believe that a major issue is that data about tornadoes comes primarily from the National Weather Service (NWS) staff in the offices having responsibility for the locations where tornadoes are reported.[3] Gathering such information is not a full-time responsibility for *anyone* in those NWS offices, and expertise in tornado occurrence data collection varies considerably among offices. Some individuals are highly motivated and knowledgeable, while others are less so, and a few *much* less so. Variability in this capability is inevitable, but I maintain that the range of variability in tornado rating accuracy and the average level of rating accuracy is lower than it should be, due to many factors. Among other things, changes in local and national reporting procedures repeatedly have been introduced, as discussed in DBK05, that alter the methodology essentially at the whim of administrators within the NWS. I believe that the gathering of tornado occurrence data will continue to be problematic until it can be put in the hands of full-time specialists in tornado climatology and damage assessment. In the past, state climatologists used to gather these data—something of the sort needs to be re-instituted.

The rating of tornado intensity by the F-Scale (or the recently implemented "Enhanced" F-Scale) is also problematic. Until it becomes possible to obtain accurate, detailed quantitative measurements of tornado wind velocities near the surface over the entire life cycle of every tornado, a breakthrough that is unlikely to happen any time soon, we will still be plagued with subjectivity in relating damage to wind speed and its associated uncertainty in estimating tornado intensity. This affects anything related to the distribution of tornadoes by intensity.

Unfortunately, even if a miracle occurs and a practical solution to the problems of tornado occurrence data is found and implemented in the near future, the relative rarity of tornadoes means that it would take many decades to accumulate enough accurate data even to begin to have a decent sample size. Thus, I see no near-term solution to the problem of detecting detailed spatial and temporal trends in the occurrence of tornadoes by using the observed data in its current form or in any form likely to evolve in the near future.

Acknowledgments. I appreciate the thoughtful and thorough reviews of this manuscript by Nikolai Dotzek and Barbara Mayes, which led to numerous improvements in the presentation. I also appreciate input from Harold Brooks.

REFERENCES

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

Doswell, C. A. III, 1980: Synoptic-scale environments associated with High Plains severe thunderstorms. *Bull. Amer. Meteor. Soc.*, **61**, 1388–1400.

_____, 2001: Severe convective storms – An overview. *Severe Convective Storms, Meteor. Monogr.*, No. 50, Amer. Meteor. Soc., 1–26.

_____, 2003: A Guide to F-Scale Damage Assessment. U.S. Dept. of Commerce, NOAA/NWS, 94 pp.

______, 2007: Historical overview of severe convective storms research. *Electronic J. Severe Storms Meteor.*, **2** (1), 1–18.

______, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **18**, 577–595.

______, R. Edwards, R. L. Thompson, J. A. Hart, and K. C. Crosbie, 2006: A simple and flexible method for ranking severe weather events. *Wea. Forecasting*, **21**, 939–951.

Fujita, T., and A. D. Pearson, 1973: Results of FPP classification of 1971 and 1972 tornadoes. Preprints, *8th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 142–145.

Godfrey, C. M., D. S. Wilks, and D. M. Schultz, 2002: Is the January Thaw a statistical phantom? *Bull. Amer. Meteor. Soc.*, **83**, 53–62.

Harlow, L. L., S. A. Mulaik, and J. H. Steiger, Eds., 1997: *What If There Were No Significance Tests?* Lawrence Eribaum Associates, 446 pp.

[3] A troubling aspect of this is that the severe weather occurrence data collected by NWS offices are used to verify warnings issued by those very same offices.

Kelly, D. R., J. T. Schaefer, R. P. McNulty, C. A. Doswell III and R. F. Abbey, Jr., 1978: An augmented tornado climatology. *Mon. Wea. Rev*., **106**, 1172-1183.

Minor, J. E., J. R. McDonald, and K. C. Mehta, 1977: The tornado: An engineering-oriented perspective. NOAA Tech. Memo. ERL NSSL-82, 196 pp. [NTIS PB281860/AS]

Schaefer, J. T., and R. Edwards, 1999: The SPC tornado/severe thunderstorm database. Preprints, *11th Conf. on Appl. Climat.*, Dallas, TX, Amer. Meteor. Soc., 603–606.

Spiegel, M. R., 1961: *Theory and Problems of Statistics*, Schaum Publishing, 359 pp.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954-2003. *Wea. Forecasting*, **21**, 86–93.

Weaver, J. F., and N. M. Doesken, 1991: High Plains severe weather—Ten years after. *Wea. Forecasting*, **6**, 411–414.

Wilks, D, 2006: *Statistical Methods in the Atmospheric Sciences, 2nd ed*. Academic Press, 627 pp.

REVIEWER COMMENTS

[Authors’ responses in *blue italics*.]

REVIEWER A (Nikolai Dotzek):

Initial Review:

Recommendation: Accept with Major Revisions.

Major Comments:

The bottom line of this paper is that statistical analysis should be done with care, and also with a sound knowledge of the underlying data. This is true, but neither new nor particularly exciting by itself.

Articles to EJSSM can be more tutorial than original, and that was one of my intents with this manuscript, despite my remark in the Introduction that it is not a tutorial. I’ll change this in the revised version.

However, the examples chosen for statistical analysis are interesting, as they also shed light on US tornado rating practice.

One issue with this paper is that the data problems noted are not really caused by sample size, but rather by spatial and temporal variations of data amount and quality, in other words: inhomogeneous and nonstationary sampling. Addressed by the author as “secular changes” and denoted a secondary effect, such sampling biases are in many ways the primary effect, and the mere sample size problem is only the secondary. Data sets with a consistent (even if significant) under-sampling are much superior to those with reporting efficiency varying dramatically from decade to decade or between regions.

I disagree with this as a general statement. My point is that typical efforts to mitigate the problems associated with data quality problems (“secular trends”) result in creating sample size issues. I have attempted to clarify this in the revised Introduction. I have modified the title, as well, to reflect the notion that my concern is the combined impact of nonmeteorological artifacts in the data and sample size issues; that is my primary focus.

Another issue is the way the paper is motivated, starting out as a rather academic demonstration of sampling or sample size effects, but only noting in the discussion that possible couplings of tornado occurrence to ENSO cycles or global climate change are the real motivation to address quality and quantity of tornado reports.

I’m uncertain how to respond to this. I agree that many studies of the coupling of tornado occurrence to ENSO cycles or global climate change are one motivation for this article, but such studies are not “the real” motivation. I maintain I have stated the “real motivation” in the revised text.

Lastly, after revision, the “Discussion” section can likely be omitted and in any case, a “Conclusion” section should be added.

I disagree with this, although the section has been revised.

A.1) Introduction

As briefly addressed above, there is a discrepancy between Sec. 1, Introduction and Sec. 4, Discussion concerning the motivation for writing the paper. Instead of pursuing an academic treatment of potential sample size problems in analyzing tornado reports, it would be much more compelling to focus on the a lot more relevant topics raised in the first and penultimate paragraphs of the Discussion:

- Coupling of tornado occurrence to external multiannual (or multidecadal) cycles, like ENSO;
- Trends of tornado occurrence related to climate change.

I can’t dispute that this might be more compelling for the reviewer, but it’s not the paper that I want to write. I prefer not to make these topics the sole focus for the paper.

A.2) An illustration of …

The author goes on to say that the reasons for the increase in reported F0 tornadoes from 1990 on are not clear. Is that so? The usually given explanation is that the increasing availability of digital camera and video equipment, together with a popularization of storm chasing and better information flow via e-mail and Internet strongly contributed to this trend. As the author is also a well-known storm chaser, he could provide an expert view on the role of widespread advances in technical equipment for the documentation (and reporting) of weak tornadoes.

I'll add a short discussion of this.

In [what was] the last paragraph of page 2, it is anticipated that for 1000 years of stable, reliable tornado reporting (cf. my remarks on sample size vs. sample bias above and below), a smooth annual distribution of days with 7+ tornadoes should follow, while for the 33-year period of data (roughly 3% of the 1000 years) the distribution is still quite noisy. I think this is what should be expected, not only due to the small sample size com-pared to the hypothetical 1000 years or the sampling variations over time, but also from the fact that tornado days are not completely independent events, due to the inherent persistence of weather: For a gradually advancing severe synoptic setting, today's tornadoes over the Great Plains might provide some 'foreshadowing' of the phenomena further east on the next day. In particular for the rather special criterion "7+ tornadoes per day", I would not expect a smooth distribution of the raw data on the relatively short time-span of 33 years.

For these reasons, highlighting that 19 April or 21 May have very high/low frequency of 7+ tornado reports is not justified, similar to the observation that some off-season days have zero occurrence of 7+ tornadoes in the 33-year period. Consequently, Fig. 5 can be omitted.

The particular number, 7, is arbitrary and changing it a few days either side will not affect the results in any important way. However, increasing the number would begin to impinge on the sample size issue, as I have tried to show. The reason for choosing that particular number is related to the actual historical record of tornadoes, as discussed in Doswell et al. (2006). This paper's discussion goes on in the next section to note the point that tornado days may not be completely independent, and taking that into account actually contains some useful information.

Having read the first of the two paragraphs above several times, I am still uncertain just what the reviewer is trying to say, as it seems at the end that he is saying just what I said in the paper.

Nor do I understand the logic in the second paragraph above, if it's based on the reasoning of the first paragraph. In fact, it seems that if I argue that tornado days (whatever threshold is chosen) are not completely independent – which I have – it makes the apparent anomalies of 19 April and 21 May even more evident. Calling attention to them seems important to me.

On [what was] page 3, second full paragraph of column 2, apparent climate singularities like the "January thaw" are discussed. However, some of these singularities are real. For instance, in Germany the last significant cold-air outbreak reliably occurs around mid-June. This phenomenon is even visible in 150-year average temperature series and colloquially called the "sheep's cold" as the sheep have just been shorn in early June and then have to endure the cool/cold weather. This "sheep's cold" dip is not limited to temperature records, it may also be seen in the daily count of tornado days/tornadoes in Germany, see http://tordach.org/de/gif/outbreak_T.gif (discussed in http://tordach.org/de/tornado.htm; a noisy graph like Chuck's Fig. 4, so 15-day moving central averaging has been applied). Thus, climate singularities should not be completely discarded in the clustering of 7+ tornado days.

The fact that the January thaw is now widely regarded as an artifact of the sampling size rather than a real anomaly doesn't preclude the existence of real anomalies, and I don't believe that the phrasing in the paper implies that <u>all</u> such anomalies are artifacts of sample size problems. I've revised the text to clarify this point.

In the last full paragraph on page 3, the author correctly states that high variability does not have to imply too small sample size. This could be illustrated for instance by a Gaussian signal with given variance σ^2. Here, the variability will be independent of sample size, except for unrealistically crude sampling.

I'm not certain how to respond to this, but perhaps something additional to make the point is needed. I've added some text and a reference to help explain this.

On [what was] page 4, top, ways out of the sampling dilemma for tornado 'outbreaks' are thought, and either increasing the minimum number of tornadoes per days or omitting the inhomogeneously and nonstationarily sampled F0 tornadoes are considered as options. Why not consider a different approach, and omit the biased F0 reports, and at the same time *lower* the number of remaining tornadoes per day required to define an 'outbreak'? This could help to preserve sample size and still remove much of the biased weak-tornado observations.

Perhaps I'm biased, myself, but I think any elaborate ways out of the dilemma are doomed to be of dubious value, at best.

Besides, let me utter the heretic thought that probably fixing the US definition of tornado outbreaks (as I recall it) to the magic number of 7+ tornadoes is simply inadequate today with such a high reporting efficiency of F0 events? Is an 'outbreak' of eight F0 tornadoes with near-zero damage really relevant, while a day with two F2s, one F3 and one F4, all hitting urban areas is not an 'outbreak'? Maybe a good part of the specter Chuck addresses at the end of Sec. 2 consists of the 'outbreak' definition itself.

If the reviewer will consult Doswell et al. (2006), we addressed the issue of "defining" a tornado outbreak – in particular, we did not define "outbreak" at all, nor am I doing so here. The choice of 7+ is both arbitrary and mostly irrelevant to the discussion in this paper. The revised paper attempts to make this more evident.

The end of Sec. 2 also nicely illustrates my criticism from above that the real problem here is the sampling bias (the "secular trends") and not the sheer sample size. This holds in particular for the detection of trends in the data. Imagine two countries A and B, each with 1000 tornado reports each year, and that there is some external trend in occurrence of say, 1% per decade. The question is: Can we detect this trend?

Country A has a poor, but consistent reporting system, and only 10% of all tornadoes are reported each year. 10% reported events will result in a small sample size compared to the number of actual events, but it is easy to see that the external trend is still accurately resolved.

The premise that exactly 10% of the tornadoes are reported each year is pretty dubious – in fact, I'd consider it pretty much counter-factual. If, as I believe, there's good reason to expect that percentage to vary substantially from one year to the next, depending on precisely where and when the actual tornadoes occurred in Country A, then the sample size issue remains relevant.

Country B has seen a steady increase in reporting efficiency from 10% to 90% over the last nine decades. The sample size of country B's tornado record will be tremendously larger than for country A, but it will be impossible to detect the external trend of 1% per decade, both based on all nine decades and based on two consecutive decades.

By the logic of the reviewer's hypothesized situation in Country A, if we know that reporting efficiency follows a precisely known trend, then the hypothesized 1% increase beyond that known trend should be detectable.

So, clearly the sampling process is the key, and not the sample size. A similar argument holds for regional variations of sampling efficiency, where even small regions (state or county level) with consistent sampling should not be merged with adjacent regions with less consistent sampling just to increase the sample size.

I'm sorry, but it's not at all clear to me that this thought experiment is all that useful and hence, the associated conclusion (that sample size is not relevant in detecting temporal trends) is not conclusively shown (at least to me) by this artifice.

A.3) An example of …

In the (too small!) Fig. 6,

I agree. The figure size will be increased in the revised text.

a true problem of sample size shows up out of the tornado season: The raw data points are quantized at 20%, 33%, 50%, and 66%.

I'm sorry, but this statement is demonstrably false. There are many days with percentages that differ from these values. The fact that the number of days is an integer results in certain percentage values being favored, but the results are not "quantized" as described.

I therefore suggest omitting the off-season periods. This will also enable a larger size and much better legibility of the revised picture.

Omitting the off-season periods, however, means that the point of the figure will be lost. In the season, the signal emerges pretty clearly, whereas during the off-season, the signal is overwhelmed by the inadequate sample.

The reasons for the peaks in strings of 7+ tornado days termed "out of scope of the paper" on page 5, first full paragraph, can certainly attributed at least in part to the persistence of severe weather patterns during the main season. Chuck can surely draw from his long experience as a forecaster and add some text here.

I'm certainly able to add more text, but going down this road strikes me as one that could easily turn into a "tar baby" encounter that would require a lot more text, involving a lot of meteorology that ultimately would distract the reader from the main point.

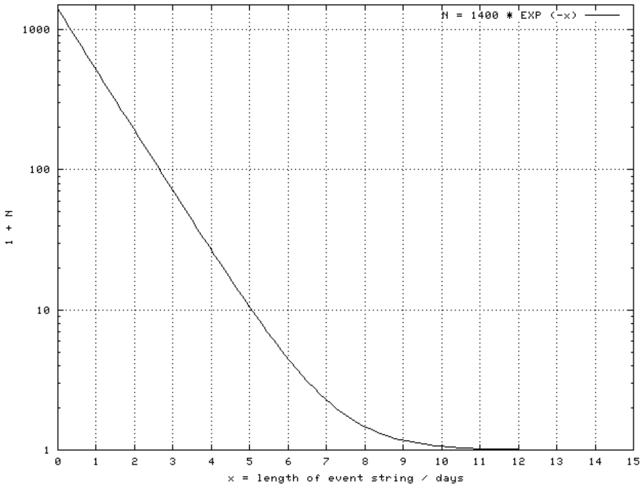

Figure 1. Exponential fit to the data of 1400 tornado days in 33 years shown in [what now is] Fig. 7 of Doswell (2007).

Discussion of Fig. 7: The (quite smooth!) distribution in Fig. 7 looks like an exponential decay of the likelihood to observe strings of length x + 1 days compared to those of length x, see my Fig. 1 [above] with a quick shot at modeling this. I think the void classes in the right tail of Fig. 7 should not be overemphasized.

I don't believe they've been overemphasized in the paper. In the revised paper, I've suggested that out to 5-6 days, the smoothness of the decay likely represents a reasonably good sample, but beyond that, the noisiness of the figure likely is virtually certain to be a sampling problem.

Surely, a larger sample should help to fill the gaps, but the extreme tails are always tough to sample and tend to remain inherently noisy. The impression of insufficient sampling comes mainly from the one 15-day string (By the way, what was so special about this 15-day period in 1980, and how reliably was it sampled? Can we learn anything from that synoptic setup?).

I haven't gone back to look at it. It's an interesting question but I consider it to be tangential to the topic of the paper.

Without this one classified event, Fig. 7 would appear much better sampled. The apparent exponential decay in Chuck's Fig. 7 and my Fig. 1 could be compared to a time series of synoptic-scale correlation of evolving weather patterns over the USA in the tornado season. I would not be surprised if this time series also displayed an exponential decay at a similar rate.

I've revised Fig. 9 to include this exponential curve. Thank you for pointing this out.

A.4) Discussion

Let me suggest omitting a separate discussion section completely for this concise paper, and instead to add a section with real conclusions – for which the present last paragraph of the Discussion could serve as a starting point.

The conclusions might also want to address points like:

- How could the sampling bias problem be overcome in the future, especially in the light of detection of possible climate change or ENSO impacts?
- What biases (= secular changes) have entered US tornado reports from 2000 on and decouple them at least in part from 20th century reports?
- Is the analysis of tornado outbreaks only plagued by sampling biases or sample size issues, or maybe also by an outbreak definition which may no longer be adequate today?
- If consistency of the sampling is the key to avoid biases, could it have been better for the USA from a statistical point of view to have kept the tornado reporting at the 1950s standards?

These are interesting points, certainly. I've attempted to say something about most of them, but I have to be careful to avoid departing too far from the paper's substantive content. The issue of the "definition" of an outbreak as 7+ tornadoes on a given day is misreading of my intentions but, in any case, it wouldn't have any impact on this paper, which is not about tornado outbreaks, per se.

[Minor comments omitted…]

Second Review:

Recommendation: Accept with Minor Revisions.

I thank Chuck for preparing this revised version. It is now in very good shape, and I have only one more technical remark... *[Minor comment omitted…]*

Pending this final change, the manuscript is ready for publication from my point of view.

One further general reply to Chuck's feedback on my review:

Outbreak definition: When I referred to 7 as the threshold number of tornadoes on a day, this is a value that you will often find cited in popular sources, such as http://en.wikipedia.org/wiki/Tornado_outbreak. So parallel to scientific discussions like by Doswell et al. (2006), it is my impression (seen from the other side of the Atlantic Ocean) that "seven or more" tornadoes has ground its way into most people's minds. That Chuck also started from the number 7 in his paper may have misled me to the conclusion that he was accepting this common "definition" here.

REVIEWER B (Barbara Mayes):

Initial Review:

Recommendation: Accept with Minor Revisions.

Major comments:

Overall, I find the paper to be an informative recapitulation of non-meteorological issues that affect the tornado database. While there is little in the way of original research in the paper, except in the way of examples, it succinctly illustrates the "known" issues in working with the tornado database, particularly for climatology applications, providing documentation that will be useful in future work using the database. I have a few suggestions to improve the presentation of the paper, as well as a couple of suggestions for additional information.

[Minor comments omitted…]

On [what was] page 4, paragraph 1, you mention that "…if the F-scale criterion were raised from any tornado (F0 or stronger) to some higher threshold, such as F2 and stronger, presumably to mitigate the strong secular trends in the reporting of weak tornadoes compared with that for the strong-to-violent tornadoes, the resulting reduction in sample size would offset this effort." This point may be better made if illustrated more specifically with the data used throughout section 2; otherwise, the assertion seems to stand out as unsubstantiated.

It's difficult for me to see this as unsubstantiated, but perhaps that's due to my familiarity with the data. All one has to do is review the distribution of tornadoes within the database when sorted by F-scale to see that significant tornadoes (F2+) comprise around 1/3 of the total, so raising the threshold to F2 eliminates around 2/3 of the sample. Is there something more specific the reviewer feels is needed?

In light of the recent attention to climate change courtesy of the IPCC release of the Summary for Policymakers in early Feb. 2007, you may want to consider moving up the discussion on tornado frequency and climate change (second to last paragraph in section 4) to the beginning of the section. This would place it closer to the ENSO discussion, thus putting the climate change and variability issues side by side.

Good suggestion.

As part of an ongoing research project (Mayes, B.E., et al, 2006: Tornado climatology and predictability by ENSO phase in the north central U.S.: A compositing study. Preprints, 19th Conf. on Climate Variability and Change, San Antonio, TX, Amer. Meteor. Soc.), we are investigating relationships between ENSO phase and tornado climatology. The methodology does include a statistical significance test, and indications are that there are legitimate statistical signals in the tornado climatology based on ENSO phase. Sample size is agreeably an issue, particularly because of the relatively small number of El Niño and La Niña years. Despite the well-known limitations with the tornado database and concerns with sample size, the study serves as an example that it is possible to draw careful conclusions based on ENSO phase using the tornado database.

The fact that the reviewer (as lead author of the cited paper) believes it is possible to draw "careful conclusions" could be considered as a biased assessment of that paper's content. If the reviewer disagrees with my assessment in this paper, I'm comfortable with her sending in comments about my paper. At this time, however, I'll stand by my conclusions.

Some researchers indicate that using moderate or strong ENSO phases in drawing conclusions helps weed out some of the interference of signal from other oscillations and weather patterns (Robert E. Livezey, personal communication), which was mentioned in your discussion as a concern. I agree that conclusions relating tornado frequency to just about any other factor should be drawn with the utmost caution – including ENSO as well as many, many other factors. In short, as you depicted with your examples, I believe it is possible to find meaningful results, as well as issues that are artifacts of the data set, using ENSO or other periodic cycles.

Regarding ENSO and other cyclic processes, I maintain that the tornado database does not lend itself to studies seeking to find signals within it that can be ascribed conclusively to ENSO. Until someone can give me a convincing argument to change my mind, I will continue to believe this. As noted above, comments on my paper will be welcome.

Second Review:

Recommendation: Accept.

The author has made many improvements in the paper, in both the presentation and in clarifying some of the substance and conclusions. I have just a few minor comments at this point.

[Minor comments omitted…]