BAMS Article

Compared to What?

Establishing Environmental Baselines for Tornado Warning Skill

Alexandra K. Anderson-Frey and Harold Brooks

ABSTRACT: In any discussion of forecast evaluation, it is tempting to fall back on statements reflecting unverified assumptions: "this tornado warning had lower skill because the underlying meteorology reflected a complicated or atypical scenario," or "that forecast performed worse than we would have expected given the straightforward setup." These statements of what is and is not a reasonable expectation for warning skill are particularly relevant as the meteorological community's focus has begun to emphasize non-classic storm environments (e.g., tornadoes spawned by quasi-linear convective systems). In this paper, we build a proof-of-concept methodology to quantify the effect of the near-storm environment on tornado warning skill, and we then test these methods on a 15-yr dataset composed of tens of thousands of tornado events and warnings over the contiguous United States. Our findings include that significant tornadoes rated (E)F2+ have a higher probability of detection (POD) than expected based on their near-storm environments, that nocturnal tornadoes have both worse POD and false alarm ratio (FAR) than even their marginal near-storm environments would suggest, and that tornadoes occurring during the summer months also show worse POD and FAR than their environment-based expectation. Quantifying these shifts in performance in an environmental skill score framework allows us to target the situations in which the greatest improvements may be possible, in terms of forecaster training and/or conceptual models. This work also highlights the essential question that should always be asked in the context of forecast verification: what, exactly, is the baseline standard to which we are comparing forecast performance?

KEYWORDS: Atmosphere; Tornadoes; Statistical techniques; Forecast verification/skill; Forecasting techniques; Mesoscale forecasting

https://doi.org/10.1175/BAMS-D-19-0310.1
Corresponding author: Alexandra K. Anderson-Frey, akaf@uw.edu
In final form 24 November 2020

AFFILIATIONS: Anderson-Frey—Department of Atmospheric Sciences, University of Washington, Seattle, Washington; **Brooks**—NOAA/National Severe Storms Laboratory, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

Evaluating the success of a forecast is a necessary step in the development of a rigorous and useful forecast system; robust forecast evaluation can identify the situations in which the most substantial improvements can be made (Brier and Allen 1950), and ideally also provides a roadmap for the application of those improvements. Choosing the metrics by which we evaluate forecasts, however, is a process that is far from trivial. Murphy (1993, hereafter M93) identifies three distinct types of forecast "goodness": consistency (correspondence between forecasters' best judgment and the forecasts they issue), quality (correspondence between forecasts and their matching observations), and value (incremental benefits realized by decision-makers through the use of the forecasts). In the work that follows, we develop forecasting metrics based on some distinguishing aspects of the near-storm environment, which touches on aspects of all three types of goodness. For the sake of illustration, in the following discussion we focus on the challenge of issuing successful tornado warnings, but many of the arguments that follow can be applied to any forecasting situation in which there are regional differences in the environmental parameters of the forecast area.

When tornado warnings are issued, M93's Type I goodness (*consistency*) is threatened if the prevailing near-storm environment of the forecasting region is not taken into account. Forecasting "rules of thumb" that may have been established with the U.S. Great Plains spring season tornadoes in mind may be inconsistent with, for instance, the actual best judgment of forecasters in the case of cool-season tornadoes in the southeastern United States. Even in the case where these non-regionalized rules are modified to reflect forecasters' perceptions, these judgments may be qualitative and run the risk of exaggerating differences between one region and another (e.g., Broomell et al. 2020).

M93's Type II goodness (*quality*) is also at stake. As the following discussion demonstrates, warning quality is consistently lower for certain near-storm environments (Brotzge and Erickson 2009, 2010; Brotzge et al. 2011, 2013; Anderson-Frey et al. 2019; Anderson-Frey and Brooks 2019). If these environments occur preferentially in a given region or during a particular season, evaluating all tornado warnings on the same neutral baseline gives a skewed perspective: even if forecasters in a particular situation perform better than their specific environmental baseline would indicate, their skill is instead compared to a common baseline reflecting a different near-storm environment. Comparing metrics of forecast quality using the same baseline, for example, in the Southeast as in the Great Plains, or the same baseline in the spring as in the winter, or the same baseline at night as during the day, is likely not an accurate picture of the quality of these respective forecasts unless the baseline is formulated in a way that more appropriately reflects environmental features.

Finally, M93's Type III goodness (*value*) is strongly impacted when the near-storm environment is not incorporated into forecast decision-making. National forecast metrics and evaluations that do not reflect the abovementioned discrepancies in consistency and quality have limited value: environment-specific baselines in skill more accurately represent the true range in skill from one National Weather Service (NWS) office to another and in one season versus another. As databases of forecast skill help determine which best practices and recommendations are adopted within the NWS, both formally and informally (Brooks and Correia 2018), a careful choice of baseline is essential to ensure that quality and consistency are being accurately assessed, reported, and evaluated.

Constructing environmental baselines

Any comparison of a metric of forecasting skill with a comparable baseline suggests the use of a skill score (Murphy 1996), the general format of which is as follows:

$$\text{Skillscore} = \frac{\text{Forecast} - \text{BaselineScore}}{\text{PerfectScore} - \text{BaselineScore}}.$$

In this equation, ForecastScore refers to the score for a particular forecast, PerfectScore refers to the mathematically best possible score obtainable, and BaselineScore is the score against which we are measuring skill: a skill score will hence measure whether or not a particular forecast's skill is better than, comparable to, or worse than our baseline. The selection of a relevant baseline score is clearly critical, and yet typically we default to using a climatological-mean or persistence reference (e.g., Brooks and Doswell 1996; Rasmussen and Blanchard 1998; Mason 2004; Aberson 2014). For the problem of issuing successful tornado warnings, as for many other forecasting problems, a single "one size fits all" baseline score is not reflective of the ways in which these forecasts are evaluated. Consider statements such as "we would expect tornado warning skill to be lower in this region because the near-storm environment presents a more difficult forecasting challenge": an environment-specific baseline score would allow us to validate and quantify this assertion.

We will focus in what follows on two simple measures of warning skill: the probability of detection (POD; the percentage of all tornadoes for which a warning was issued ahead of time) and the false alarm ratio (FAR; the percentage of all tornado warnings within which no tornado was reported for the duration of the warning). These two metrics provide the basis for a performance diagram (Roebber 2009) that graphically depicts both axes of forecast skill. In addition, a performance diagram depicts the critical success index [CSI (Schaefer 1990); the CSI is a function of POD and FAR] as curves.

To begin, we consider two environmental near-storm parameters that are traditionally used in diagnosis and prediction of tornadic near-storm environments (Thompson et al. 2003; Anderson-Frey et al. 2016, 2019; Coffer et al. 2019): mixed-layer convective available potential energy (MLCAPE; a measure of atmospheric instability) and 0–6-km shear (SHR6; a measure of wind shear from the surface to 6 km AGL). The combination of high MLCAPE and high SHR6 is associated with an increased potential for supercell thunderstorms (Thompson et al. 2003).

To establish baseline POD and FAR for different combinations of these near-storm environmental parameters, we make use of a 15-yr climatology (2003–17) of tornado events and tornado warnings established by the NOAA Storm Prediction Center (Smith et al. 2012; Anderson-Frey et al. 2018). Since, for the purposes of this study, POD and FAR are assumed to be binary, our initial approach is to create multidimensional bins that consist of ranges of each of several environmental parameters so that POD and FAR can be calculated for tornado events/warnings meeting particular environmental criteria (i.e., MLCAPE, SHR6, low-level storm-relative helicity, etc., within particular ranges). Unfortunately, even a dataset of 49,740 tornado warnings and 16,232 tornado events becomes sparse when splitting into that many bins; data were noisy and results were inconclusive.

As a result, we decided to approach the problem using a single parameter space as a proof of concept: MLCAPE–SHR6 (e.g., Craven and Brooks 2004; Anderson-Frey et al. 2016). In Fig. 1a, POD has been calculated for bins of SHR6 and MLCAPE, and Fig. 1b likewise demonstrates FAR for bins of SHR6 and MLCAPE; greater warning skill (i.e., higher POD and lower FAR) is generally most apparent for the combination of greater values of MLCAPE and SHR6 (i.e., toward the upper right quadrant of the plots).

We emphasize that the focus of this work on the MLCAPE–SHR6 parameter space represents a gross oversimplification of the forecast process, which in reality encompasses the use of facets such as longer-term outlook and watch products, mesoscale discussions, radar and satellite imagery, prior warnings in the region, severe weather reports, etc. These and many other factors influence a forecaster's decision-making process, but they are inherently more difficult to quantify than the near-storm environment. This parameter space, then, is not intended to fully represent the information available to forecasters issuing tornado warnings, nor is it intended to act as a complete proxy for the information available to forecasters about

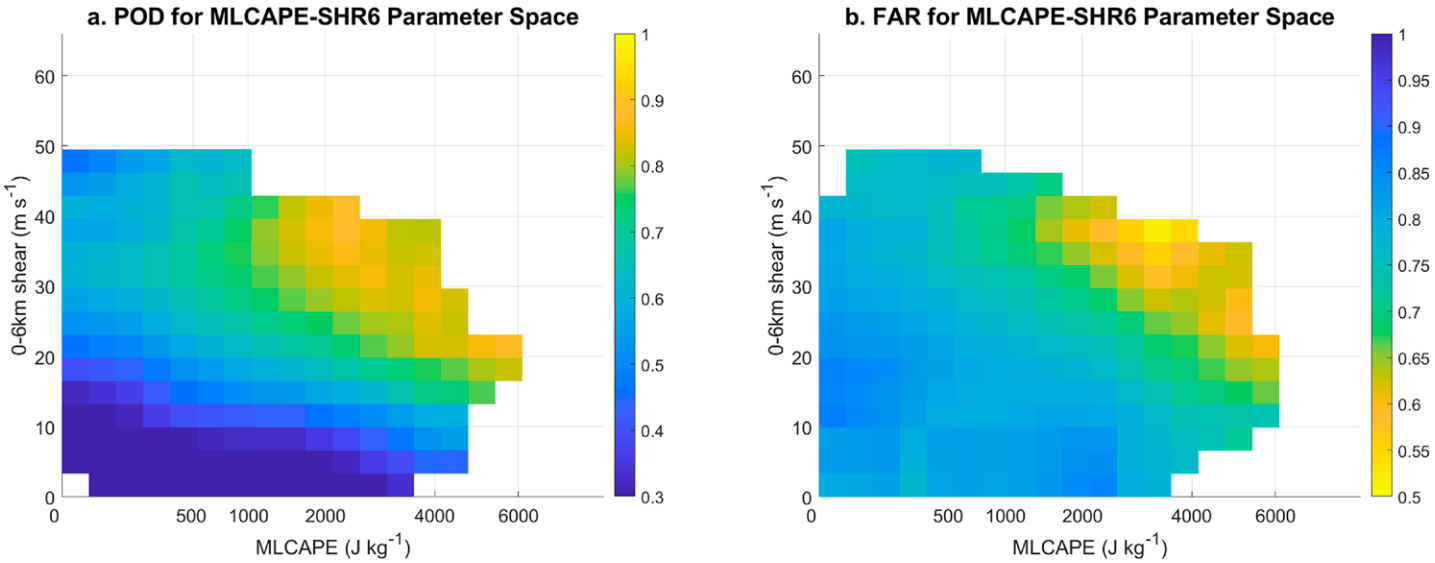

Fig. 1. Binned plots of (a) POD and (b) FAR for the MLCAPE–SHR6 parameter space. Note that the MLCAPE axis is nonlinear; MLCAPE has been converted to a maximum estimated wind speed via $w_{max} = 2 \times \text{MLCAPE}^{1/2}$. Bin sizes were determined qualitatively to provide adequate coverage of the parameter space: $w_{max} = 6.5$ m s^{-1} and SHR6 = 3.3 m s^{-1}. Smoothing is accomplished by using a mean filter over a rectangle at each location (smooth2a: www.mathworks.com/matlabcentral/fileexchange/23287-smooth2a). Note that overall trends in POD and FAR were not sensitive to the extent (or even the absence) of the smoother; smoothing is only done to facilitate discussion.

the near-storm environment. Instead, we use this simple combination of two readily available parameters to emphasize how much can be gained in gauging tornado warning skill by incorporating even this incomplete picture of the near-storm environment.

Creating environmental skill scores

In this section we calculate environmental skill scores using POD and FAR baseline values calculated from the MLCAPE–SHR6 plots in Fig. 1 as follows:

1) For each tornado event or warning that we wish to evaluate against this baseline, we obtain representative values of MLCAPE and SHR6 from the archive of SPC Mesoanalysis (Bothwell et al. 2002; Smith et al. 2012).
2) We obtain values of POD and FAR from Fig. 1 based on these values of MLCAPE and SHR6. In our skill scores, these will be the baseline skill values.
3) We calculate environmental skill scores based on POD (ESSP) and FAR (ESSF) using the following equations:

$$\text{ESSP} = \frac{\text{POD}_{\text{actual}} - \text{POD}_{\text{baseline}}}{1.0 - \text{POD}_{\text{baseline}}},$$

$$\text{ESSF} = \frac{\text{FAR}_{\text{actual}} - \text{FAR}_{\text{baseline}}}{0.0 - \text{FAR}_{\text{baseline}}} = 1 - \frac{\text{FAR}_{\text{actual}}}{\text{FAR}_{\text{baseline}}}.$$

Hence, values of ESSP and ESSF that are positive indicate that forecast skill is greater than that expected based on the near-storm MLCAPE–SHR6 environment alone, whereas values of ESSP and ESSF that are negative indicate that forecast skill is worse than that expected based on the near-storm environment alone. A performance diagram provides an easy means to expand upon these skill score results.

Season. We begin by considering how these environmental skill scores vary with season. In this context, season is defined as spring (MAM), summer (JJA), fall (SON), and winter (DJF). Figure 2 is a performance diagram reflecting the difference between the expected/baseline skill using near-storm environmental data from Fig. 1 and the actual warning performance from 2003 to 2017. Values of ESSP and ESSF for each season are compiled in Table 1. As expected based on the typical ranges of POD and FAR values, there are generally greater-magnitude discrepancies in POD than in FAR (i.e., the arrows in Fig. 2 have longer vertical components than horizontal components); keep in mind that a relatively small shift in FAR values is historically more unusual than a similar shift in POD values (Brooks and Correia 2018).

For the spring, the pink arrow pointing toward the upper right of Fig. 2 indicates that both POD and FAR were better even than the values predicted using near-storm environmental data for these tornado events. Thus, while we might on the basis of Fig. 1 expect warning skill to be quite good in the spring (since values of the relevant parameters are typically most favorable during those months; Anderson-Frey et al. 2016), Fig. 2 shows that actual warning skill is even better than expected, both in terms of detecting tornadoes and in terms of avoiding false alarms. Quantifying these differences (Table 1), ESSP = +0.064 and ESSF = +0.031, which both indicate a small improvement. To put this into perspective, these numbers reflect the difference between a POD of 0.687 (expected) and 0.707 (actual), and the difference between a FAR of 0.755 (expected) and 0.731 (actual). Note that spring is the only season with a positive ESSF, which means that false alarm ratios are lower than we would expect based on the near-storm environment.

In the summer, we see the green arrow pointing toward the lower left of Fig. 2, which shows a decrease in skill along both axes for this season compared to what we would expect based on the near-storm environment alone. Hence, during the summer months we can accurately say that while warning skill is expected to be a little lower due to the more marginal tornadic environments (Anderson-Frey et al. 2016) and the common tornado parameters have been shown to decrease in utility in the summer months (Hart and Cohen 2016), actual skill is even lower than expected. Note that summer is the only season for which ESSP is negative (−0.096), i.e., the only season in which POD is lower than that expected based on the near-storm environment. ESSF is also negative (−0.028), although the skill drop is not as dramatic as in the fall months.

The fall months, represented by the orange arrow in Fig. 2, show poorer performance in terms of FAR but better performance in terms of POD than that expected based on the near-storm environment. Thus, ESSP is positive

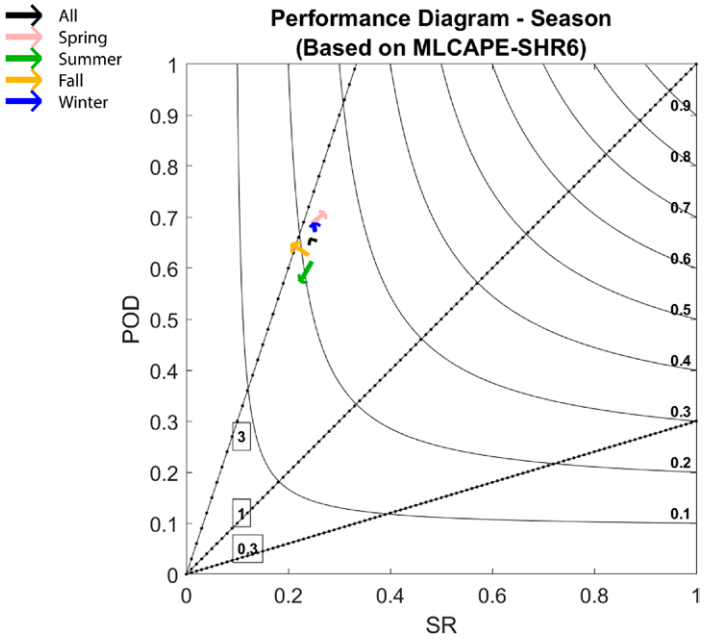

Fig. 2. Performance diagram for seasonal performance, where the baseline has been established using MLCAPE and SHR6 from Fig. 1. The abscissa is the success ratio (SR), or 1 – FAR, and the ordinate is the POD; better warning skill is hence toward the upper-right corner of the diagram, as indicated by increasing values of the curved contours of critical success index (CSI, another common measure of forecast skill). Forecast bias score is depicted with dot–dashed lines; values are printed inside boxes for reference. Each season is depicted as an arrow, where the origin of the arrow indicates the expected performance based on the near-storm environment and the tip of the arrow indicates the actual performance for that season between 2003 and 2017. Note that the skill for all events (black arrow) has not changed appreciably from expectation.

(+0.057) and ESSF is negative (−0.041). Winter (blue arrow in Fig. 2) has better-than-expected POD performance (ESSP = +0.057) and as-expected FAR performance (ESSF = −0.003).

The environmental context provided by this analysis, although limited, suggests which seasons' actual warning statistics may be better (e.g., spring) or worse (e.g., summer) than expected without the inclusion of the near-storm environment. These results are suggestive but not conclusive, given that this simple analysis has only looked at a limited subset of near-storm environmental parameters, and has completely excluded factors such as synoptic-mesoscale context and report history.

Time of day. In the discussion that follows, time of day is split into day (between local sunrise and 2 h prior to local sunset), early-evening transition (EET; between 2 h before and 2 h after local sunset[1]), and night (between 2 h after local sunset and local sunrise). The performance diagram in Fig. 3 shows how warning skill compares to the environmental baselines established in Fig. 1 for these three times of day, and Table 1 summarizes skill scores.

For daytime tornadoes and warnings (red arrow in Fig. 3), FAR is generally similar to that expected based on the environmental data (ESSF = +0.005), but the POD is slightly worse than expected (ESSP = −0.051). In contrast, for nocturnal tornadoes and warnings (blue arrow in Fig. 3), both POD and FAR are worse than expected (ESSP = −0.100 and ESSF = −0.049), even given that nocturnal tornadoes disproportionately occur in environments with more marginal values of MLCAPE (Anderson-Frey et al. 2016). The EET (green arrow in Fig. 3) has improvements in both POD and FAR (ESSP = +0.151 and ESSF = +0.014).

[1] The EET, as shown in Anderson-Frey et al. (2016), encompasses a period of maximum MLCAPE, increasing SHR6 and low-level storm-relative helicity, and decreasing lifting condensation level heights, which contribute to higher POD and lower FAR for tornadoes.

Coupled with the seasonal data in Fig. 2, these results highlight that scenarios associated with springtime tornadoes occurring within a couple hours of sunset are forecast even better than the near-storm environment alone would indicate, whereas more atypical scenarios in other seasons and times of day show improved skill only along one axis, if at all. Note, however, that this result assumes that the environment is the only control.

Table 1. Environmental skill scores according to POD (ESSP) and FAR (ESSF) by season, time of day, storm mode, and tornado intensity between 2003 and 2017 (2003–15 for storm mode data). Environmental baseline comes from MLCAPE–SHR6 data in Fig. 1. Note that storm mode and tornado intensity cannot be calculated for tornado warnings, only tornado events, and hence POD can be calculated but not FAR.

	Baseline POD	Actual POD	ESSP	Baseline FAR	Actual FAR	ESSF
All	0.654	0.656	+0.006	0.755	0.756	−0.001
Spring	0.687	0.707	+0.064	0.755	0.731	+0.031
Summer	0.613	0.576	−0.096	0.754	0.775	−0.028
Fall	0.625	0.646	+0.057	0.760	0.791	−0.041
Winter	0.671	0.686	+0.045	0.748	0.750	−0.003
Day	0.628	0.609	−0.051	0.755	0.752	+0.005
EET	0.683	0.731	+0.151	0.754	0.743	+0.014
Night	0.668	0.635	−0.100	0.753	0.790	−0.049
RMS	0.691	0.769	+0.255	—	—	—
QLCS	0.642	0.483	−0.443	—	—	—
(E)F0–1	0.640	0.625	−0.043	—	—	—
(E)F2–3	0.733	0.845	+0.420	—	—	—
(E)F4–5	0.813	0.992	+0.956	—	—	—

Storm mode. The SPC dataset also includes storm mode information between 2003 and 2015, where radar data were manually examined for each tornado event (Smith et al. 2012). Two storm modes of particular interest to tornado research are right-moving supercells (RMS) and quasi-linear convective systems (QLCS) (Thompson et al. 2012; Anderson-Frey et al. 2019). Since establishing a null dataset that could be used to identify the storm mode of a false-alarm warning is beyond the scope of this work, FAR was not calculated; hence, only ESSP values are available in Table 1, and Fig. 4 depicts POD values for the two storm modes in question.

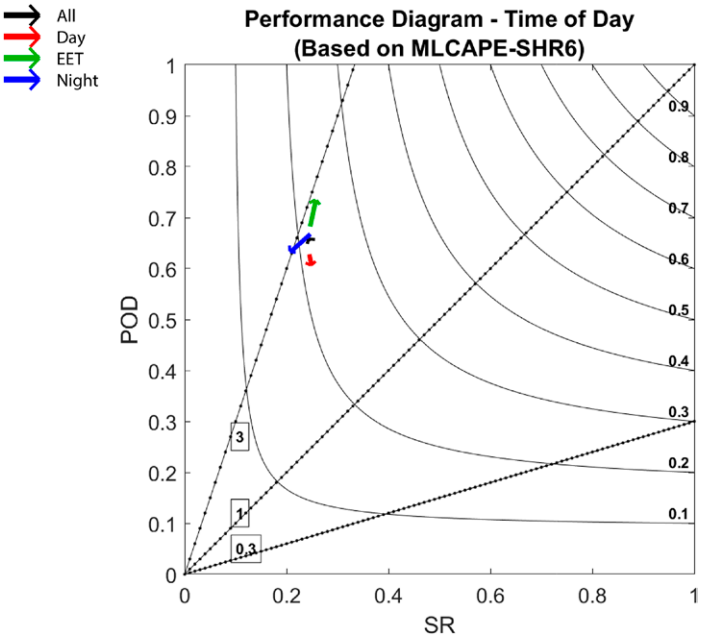

Fig. 3. As in Fig. 2, but for time of day.

In line with expectations established in the seasonal and diurnal performance diagrams, RMS tornadoes (which more typically occur in the high-MLCAPE high-SHR6 environments characterized by better warning performance; Fig. 1) have substantially higher POD values than would be expected based on the near-storm environment alone (ESSP = +0.255, corresponding to a jump in POD from 0.691 to 0.769). This result is perhaps unsurprising given all the factors outside of MLCAPE and SHR6 that can contribute to warning a tornado with a parent supercell (e.g., distinctive radar signature).

In distinct contrast, QLCS tornadoes (which tend to occur in different parts of the parameter space compared to RMS tornadoes; Thompson et al. 2012; Anderson-Frey et al. 2019) have dramatically lower POD values even than those expected based on the often-marginal near-storm environment (ESSP = −0.443, corresponding to a plunge in POD from 0.642 to 0.483). This major discrepancy also suggests that factors outside of MLCAPE and SHR6 contribute to warning a tornado in a QLCS storm, but the drop in forecast skill suggests that our understanding of these processes is limited and provides additional room for directed research and improvement.

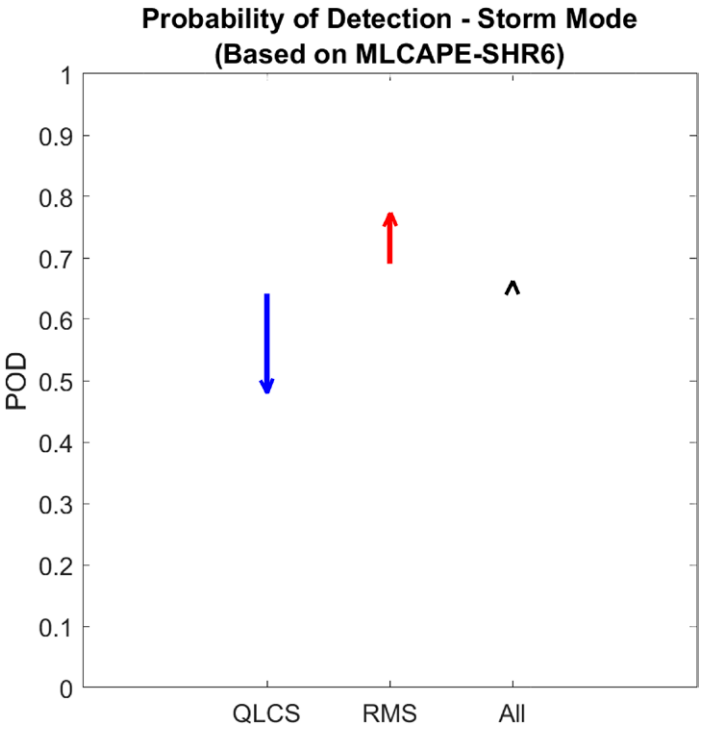

Fig. 4. As in Fig. 2, but for storm mode, where RMS is right-moving supercell and QLCS is quasi-linear convective system. Note that storm mode data are only available from 2003 to 2015 and storm mode was not determined for false alarm tornado warnings, so FAR cannot be calculated.

Intensity. Finally, we consider tornado intensity, as estimated by the F scale prior to 1 February 2007 and by the EF scale after that date. False alarm warnings cannot have an EF scale assigned to them, and so only POD can be calculated. Hence, Table 1 contains ESSP values and Fig. 5 shows POD.

We group tornado intensity into three categories: (E)F0–1, (E)F2–3, and (E)F4–5.

Even given the typically extreme environments in which (E)F2+ tornadoes occur, POD is dramatically higher than the environmental prediction for (E)F2–3 tornadoes (ESSP = +0.420, or a difference in POD from 0.733 to 0.845) and especially for (E)F4–5 tornadoes (ESSP = +0.956, or a difference in POD from 0.813 to 0.992). More marginal (E)F0–1 tornadoes are characterized by a small drop in POD compared to the environmental baseline (ESSP = −0.043).

The vast majority of tornadoes (>80%) are rated (E)F0–1, but are responsible for fewer than 5% of all tornado deaths (Anderson-Frey and Brooks 2019). Hence, high skill for warnings on (E)F2+ tornadoes is extremely important from a public safety perspective; the results of this analysis show that the most dangerous tornadoes are also those with better-than-expected warning skill.

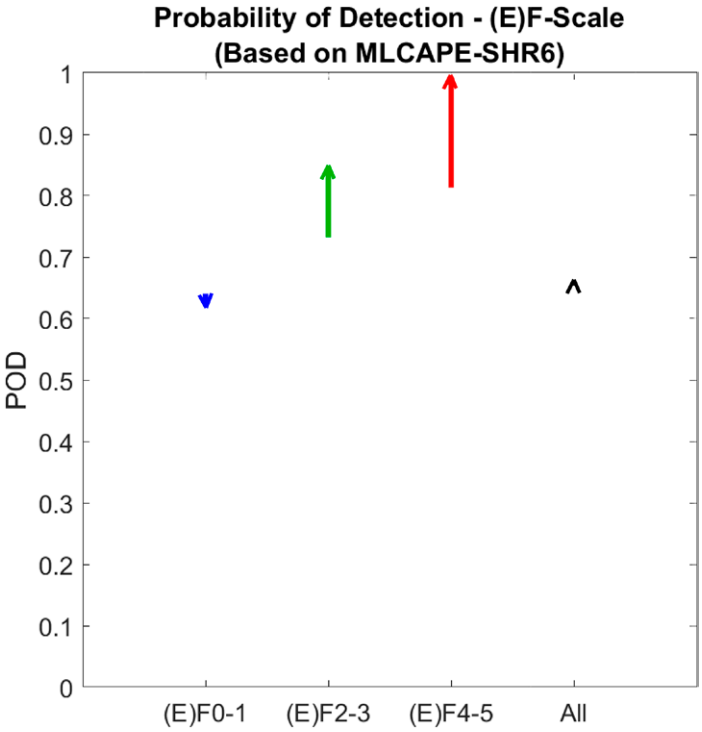

Fig. 5. As in Fig. 2, but for (E)F scale. Note that (E)F scale rating cannot be determined for false alarm tornado warnings, so FAR cannot be calculated.

Adopting an environmental framework

Metrics based on an environment-specific baseline such as ESSP and ESSF can serve as a useful post-event evaluation tool: as an example, performance data for the major tornado outbreak that occurred on 27 April 2011 (Knupp et al. 2014) are depicted in Fig. 6. Note that for the event as a whole, POD and FAR are both a substantial improvement over what we would have expected based on the near-storm environment alone (ESSP = +0.396 and ESSF = +0.111). Even separating the event into daytime, early evening transition, and nocturnal tornadoes (Fig. 6a), we still see a general pattern of great improvement in POD and FAR, with only a slight dip in POD for daytime events. The improvement in POD is common across both RMS and QLCS tornadoes (Fig. 6b) as well as across all (E)F scales (Fig. 6c). Hence, while it may have been straightforward to recognize that this event was well forecast, the environmental framework has allowed us to state with some authority that the high skill was not, for instance, entirely due to anomalously high MLCAPE and SHR6 making for a perhaps less difficult forecasting problem; other important factors contributed to the high skill in this event.

Note that, while this environmental baseline constitutes an important step toward clarifying storm warning skill, there remains a wide variety of factors beyond the environment that strongly influence metrics of skill such as POD and FAR (proximity to radar, population density, etc.). The near-storm environment remains an important piece of the tornado warning puzzle, but it is far from the only piece.

While this discussion has been entirely couched in the example of tornado warnings, the general philosophy behind careful and purposeful baseline selection for forecast evaluation applies to a broad variety of applications. When we approach forecasting from the perspective of an environmental framework—when we evaluate forecast skill and ask "compared to what?"—we gain a more nuanced understanding of some of the ways in which forecasts succeed and fail.

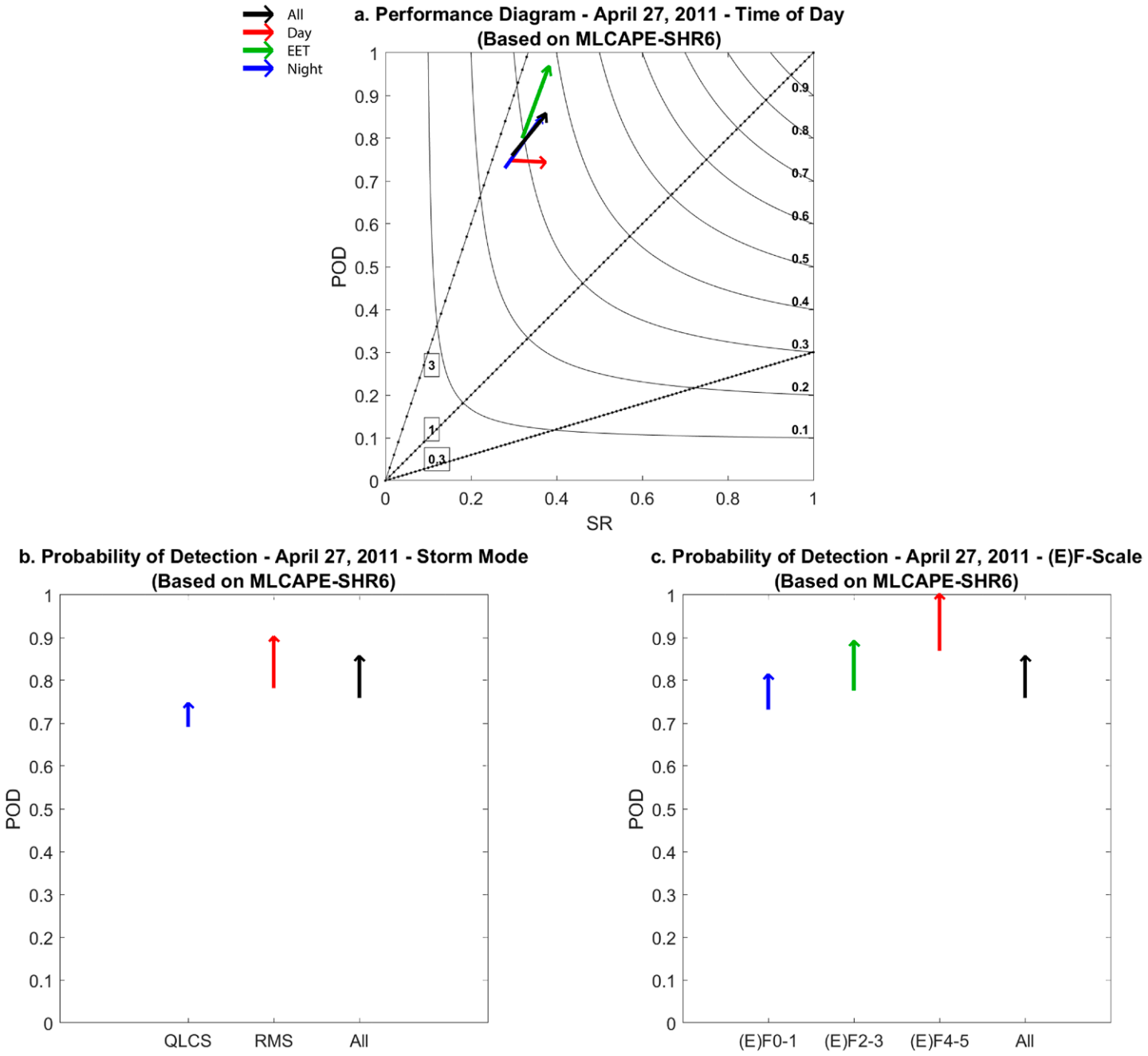

Fig. 6. Performance information for the 27 Apr 2011 tornado outbreak. (a) As in Fig. 3, but for 27 Apr 2011. (b) As in Fig. 4, but for 27 Apr 2011. (c) As in Fig. 5, but for 27 Apr 2011.

Acknowledgments. The authors thank Bryan Smith, Rich Thompson, and Andy Dean for assistance obtaining and interpreting the storm mode, warning skill, and mesoanalysis datasets. Discussions with Yvette Richardson, Burkely Gallo, and Dale Durran were extremely helpful in the formulation of these ideas. We also thank three anonymous reviewers who greatly improved and streamlined the content of this manuscript. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115, U.S. Department of Commerce.

References

Aberson, S., 2014: A climatological baseline for assessing the skill of tropical cyclone phase forecasts. *Wea. Forecasting*, **29**, 122–129, https://doi.org/10.1175/WAF-D-12-00130.1.

Anderson-Frey, A., and H. Brooks, 2019: Tornado fatalities: An environmental perspective. *Wea. Forecasting*, **34**, 1999–2015, https://doi.org/10.1175/WAF-D-19-0119.1.

——, Y. Richardson, A. Dean, R. Thompson, and B. Smith, 2016: Investigation of near-storm environments for tornado events and warnings. *Wea. Forecasting*, **31**, 1771–1790, https://doi.org/10.1175/WAF-D-16-0046.1.

——, ——, ——, ——, and ——, 2018: Near-storm environments of outbreak and isolated tornadoes. *Wea. Forecasting*, **33**, 1397–1412, https://doi.org/10.1175/WAF-D-18-0057.1.

——, ——, ——, ——, and ——, 2019: Characteristics of tornado events and warnings in the southeastern United States. *Wea. Forecasting*, **34**, 1017–1034, https://doi.org/10.1175/WAF-D-18-0211.1.

Bothwell, P., J. Hart, and R. Thompson, 2002: An integrated three-dimensional objective analysis scheme in use at the Storm Prediction Center. *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., JP3.1, https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47482.htm.

Brier, G., and R. Allen, 1950: Verification of weather forecasts. *AMS Compendium of Meteorology*, Amer. Meteor. Soc., 841–848.

Brooks, H., and C. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288–303, https://doi.org/10.1175/1520-0434(1996)011<0288:ACOMOA>2.0.CO;2.

——, and J. Correia Jr., 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

Broomell, S., G. Wong-Parodi, R. Morss, and J. Demuth, 2020: Do we know our own tornado season? A psychological investigation of perceived tornado likelihood in the southeast United States. *Wea. Climate Soc.*, **12**, 771–788, https://doi.org/10.1175/WCAS-D-20-0030.1.

Brotzge, J., and S. Erickson, 2009: NWS tornado warnings with zero or negative lead times. *Wea. Forecasting*, **24**, 140–154, https://doi.org/10.1175/2008WAF2007076.1.

——, and ——, 2010: Tornadoes without NWS warning. *Wea. Forecasting*, **25**, 159–172, https://doi.org/10.1175/2009WAF2222270.1.

——, ——, and H. Brooks, 2011: A 5-yr climatology of tornado false alarms. *Wea. Forecasting*, **26**, 534–544, https://doi.org/10.1175/WAF-D-10-05004.1.

——, S. Nelson, R. Thompson, and B. Smith, 2013: Tornado probability of detection and lead time as a function of convective mode and environmental parameters. *Wea. Forecasting*, **28**, 1261–1276, https://doi.org/10.1175/WAF-D-12-00119.1.

Coffer, B., M. Parker, R. Thompson, B. Smith, and R. Jewell, 2019: Using near-ground storm relative helicity in supercell tornado forecasting. *Wea. Forecasting*, **34**, 1417–1435, https://doi.org/10.1175/WAF-D-19-0115.1.

Craven, J., and H. Brooks, 2004: Baseline climatology of sounding derived parameters associated with deep, moist convection. *Natl. Wea. Dig.*, **28**, 13–24.

Hart, J., and A. Cohen, 2016: The challenge of forecasting significant tornadoes from June to October using convective parameters. *Wea. Forecasting*, **31**, 2075–2084, https://doi.org/10.1175/WAF-D-16-0005.1.

Knupp, K., and Coauthors, 2014: Meteorological overview of the devastating 27 April 2011 tornado outbreak. *Bull. Amer. Meteor. Soc.*, **95**, 1041–1062, https://doi.org/10.1175/BAMS-D-11-00229.1.

Mason, S., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891–1895, https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2.

Murphy, A., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

——, 1996: General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Mon. Wea. Rev.*, **124**, 2353–2369, https://doi.org/10.1175/1520-0493(1996)124<2353:GDOMBS>2.0.CO;2.

Rasmussen, E., and D. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164, https://doi.org/10.1175/1520-0434(1998)013<1148:ABCOSD>2.0.CO;2.

Roebber, P., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Schaefer, J., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Smith, B., R. Thompson, J. Grams, C. Broyles, and H. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

Thompson, R., R. Edwards, J. Hart, K. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, B. Smith, J. Grams, A. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercells and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154, https://doi.org/10.1175/WAF-D-11-00116.1.