Evaluation of the Storm Prediction Center's Day 1 Convective Outlooks

NATHAN M. HITCHENS AND HAROLD E. BROOKS

NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 19 June 2012, in final form 8 August 2012)

ABSTRACT

The Storm Prediction Center has issued daily convective outlooks since the mid-1950s. This paper represents an initial effort to examine the quality of these forecasts. Convective outlooks are plotted on a latitude–longitude grid with 80-km grid spacing and evaluated using storm reports to calculate verification measures including the probability of detection, frequency of hits, and critical success index. Results show distinct improvements in forecast performance over the duration of the study period, some of which can be attributed to apparent changes in forecasting philosophies.

1. Introduction

The National Weather Service's Storm Prediction Center (SPC) issues forecast products that provide information about the threat of severe thunderstorms and tornadoes for the 48 contiguous states on time scales of hours up to 8 days. A wide variety of products are used to convey this information, including mesoscale discussions, severe thunderstorm and tornado watches, and convective outlooks (COs). The discussions and watches are unscheduled products and are issued as needed when conditions warrant. They cover a range of areas and time scales up to several hours, typically. COs, on the other hand, are issued at the same time every day and cover the entire country. The SPC starting issuing COs in 1955 (Corfidi 1999) and, over the years, has added COs of increasing lead time and with more frequent updates.

Evaluation of the SPC's forecasts has posed significant problems over the years. Severe thunderstorm and tornado reports are inherently ''target of opportunity'' observations. An observer and a system to collect such reports are needed in order to get information. Unfortunately, from the standpoint of evaluation, reporting practices have varied widely through time and over space in the United States (e.g., Doswell et al. 2005, 2009). Watches, particularly tornado watches, have been the focus of evaluation studies in the past (Doswell et al. 1993, Vescio and Thompson 2001), but less attention has been paid to the COs. The regularity of issuance of the COs, however, makes them an attractive target for study and makes some aspects of the interpretation of performance more amenable than for watches. This paper represents an initial effort to look at the quality (Murphy 1993) of COs over a long period of time. For now, we will look only at the product (the so-called 1200 UTC CO) that has been issued continuously since 1973 and consider the evolution of quality over time.

2. Data and methods

The SPC early morning convective outlooks currently are issued at 0600 UTC, and are valid for the 24-h period beginning at 1200 UTC and ending at 1200 UTC the following day. The early morning outlook currently is updated at 1300, 1630, 2000, and 0100 UTC, but the updates expire at the same time as the initial forecast. This study focuses on the early morning outlooks, as only those outlooks have consistently been issued with the same 24-h valid time period through the duration of the dataset (1973–2010). There are three risk levels that can be issued in a CO: slight, moderate, and high. During the first two years of the digital dataset (1973–74), only the equivalent of moderate and high risks are available, although at that time the SPC did forecast outlooks for a ''few severe storms,'' which are analogous to slight risks (S. F. Corfidi 2012, personal communication). When assigning heightened risk areas, the current convention is to enclose them within the lower risk areas; for instance, a high risk area is contained within both moderate and

Corresponding author address: Dr. Nathan M. Hitchens, National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, OK 73072.
E-mail: nathan.hitchens@noaa.gov

DOI: 10.1175/WAF-D-12-00061.1

slight risk areas. This was not always the case, as through at least the mid-1980s, a moderate risk might be issued without a surrounding slight risk, or a high-risk area might be contained within a "slight," without any "moderate."

The CO risk areas are stored as a series of coordinates that outline individual polygons, but in order to facilitate evaluation the risk areas are gridded on a 0.1° × 0.1° (~10 km) latitude–longitude grid. This is accomplished by plotting the outline of a polygon, then using a contour-encoding technique (Gourret and Paille 1987) to fill the risk area with a numeric value representing each risk level. For those instances in which a risk level was skipped by the SPC, the area is assigned the same risk as the next highest risk level.

Individual storm reports that meet the National Weather Service's criteria for "severe" are used to verify the COs (storm report data available from the SPC at http://www.spc.noaa.gov/wcm/#data). Reports are grouped into 24-h blocks beginning at 1200 UTC each day and plotted on grids with the same specifications as those used with the COs. Each cell in the report grid is considered dichotomous such that cells assigned multiple reports in a 24-h period do not have more influence than a cell with a single report. The CO and storm report data are analyzed at a variety of larger grid sizes—40, 80, 160, 320, 640, 1280, 2560, and 6100 km (a single grid cell representing the entire domain)—in order to assess the usefulness of the forecasts on a variety of scales, as well as to identify the scale of optimal performance. However, the majority of the analysis in this study uses the 80-km grid spacing since it is the intended scale of the products issued by the SPC (within 25 mi of a point).

Because both COs and reports are considered dichotomous on a grid-cell by grid-cell basis, they can be evaluated using a 2 × 2 contingency table:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} Y{:}Y & Y{:}N \\ N{:}Y & N{:}N \end{pmatrix} = \begin{pmatrix} \text{hit} & \text{false alarm} \\ \text{miss} & \text{null} \end{pmatrix}, \tag{1}$$

where a is the number of correct predictions, b is the number of incorrect predictions, and c is the number of unpredicted events. The second part of Eq. (1) denotes the status (yes or no) of forecasts and observations for each contingency (forecast:observation). A variety of verification measures are calculated from this table, but those used herein are the probability of detection (POD), frequency of hits (FOH), critical success index (CSI), and bias, which are defined as

$$\text{POD} = \frac{a}{a+c}, \tag{2}$$

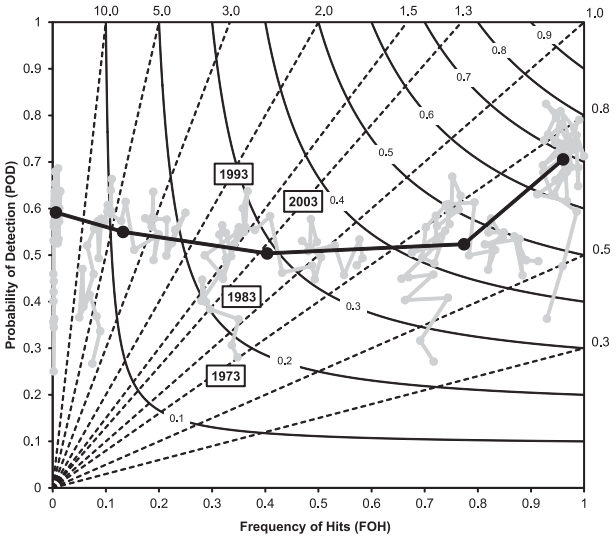

FIG. 1. Performance diagram (Roebber 2009) of convective outlook performance in terms of POD and FOH. The dashed lines represent bias (B), while the curved lines are for the CSI. The gray points indicate annual performance from 1973 to 2010, and are grouped according to the grid spacing used (from left to right: 10, 80, 320, 1280, and 6110 km). The black points correspond with the average performance over all years for each grid spacing displayed. Labeled years are provided for context and are applicable among all curves.

$$\text{FOH} = \frac{a}{a+b}, \tag{3}$$

$$\text{CSI} = \frac{a}{a+b+c}, \quad \text{and} \tag{4}$$

$$\text{bias} = \frac{a+b}{a+c}. \tag{5}$$

A complete description of these and other measures is found in Doswell et al. (1990). Additionally, a method developed by Roebber (2009) for displaying these four measures on a single graphic (performance diagram) is used to efficiently evaluate CO quality through time and at differing spatial scales.

3. Results

The performance of the SPC's day 1 CO slight risk areas are evaluated over time and at different spatial scales using the measures from a performance diagram (Fig. 1). Previous work by Brooks et al. (2011) demonstrated the utility of performance diagrams in evaluating severe storms forecasts. Comparing the results of increased grid spacing, the greatest improvement is seen in FOH values, with little change in the POD across spatial scales. This is to be expected since increases in grid spacing decrease the frequency of "false alarms" at a greater rate

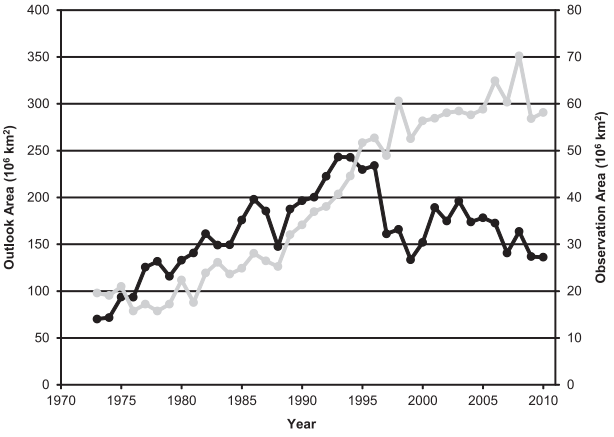

FIG. 2. Annual sums of convective outlook area (black line, left axis) and observed storm report area (gray line, right axis).

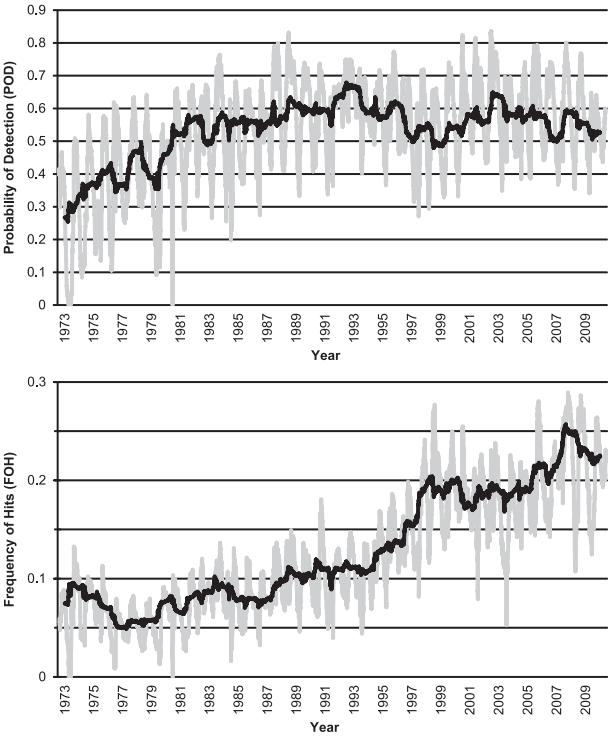

FIG. 3. (top) POD and (bottom) FOH calculated using 91- (gray line) and 365-day (black line) running means.

than the frequency of ''hits'' decreases. On the other hand, the coarsening of the grid does not significantly alter the frequency of ''misses'' with respect to the hits. Focusing on a single spatial scale (80 km, for instance) reveals that between 1973 and 1993 the POD increases more than any other measure, but changes very little for the remainder of the period of study, while the FOH improves a small amount over this time. From a forecasting perspective this suggests that during the first two decades risk areas over time were better located to enclose more severe events, accounting for the improvement in the POD, but the consistent FOH values over this time indicate that the POD improvement is also a result of an increase in the size of risk areas. It should be noted that it is unlikely that this change could be achieved through increases in severe weather reports. If more reports are found, then false alarms become hits and null events become misses, which would have little effect on POD but would increase FOH. The increase in the FOH for the remainder of the period is indicative of well-placed risk areas that decreased in size, reducing the number of false alarms. These assertions are supported by plotting the annual CO area (Fig. 2), showing an upward trend until 1993. The area of COs remains constant for 4 years, then decreases rapidly and remains relatively stable for the remainder of the study period. Accompanying data illustrating the annual area covered by storm reports provide additional support, suggesting the post-1996 decline in CO area is not related to a decline in reported severe weather.

We can see information about the seasonal and annual trends by looking at smoothed time series. A 91-day running mean mimics a seasonal period without arbitrarily defining the seasons, while a 365-day running mean shows annual performance (Fig. 3). From the beginning of the period of record through 1993, the 365-day POD increases steadily from 0.27 to 0.67, a significant improvement. After a slight decline the value of this measure stabilizes at ~0.55, although the 91-day POD peaks to at least 0.80 several times during this span. This value was only reached twice prior to the year 2000. The 365-day FOH varies between 0.05 and 0.10 from the beginning of the period until 1994, after which it steadily increases over three years, stabilizing at values between 0.17 and 0.26 from 1998 onward. It should also be noted that the range of 91-day FOH values prior to 1994 is 0.15, but the range of this measure increases to more than 0.22 after 2000. These results further support the previous observations in this study, highlighting the apparent change in forecasting philosophy during 1993–94, which resulted in improvements in the FOH. One factor contributing to this apparent philosophical change is organizational restructuring and an influx of new forecasters preceding the physical relocation of the SPC during the 1995–97 time period (Corfidi 1999).

Additional investigation into the seasonal behavior of the POD and FOH is performed by identifying the ordinal date of the maximum and minimum value of each measure using a 365-day moving window (Fig. 4). Maxima for the POD occur most frequently during the spring (March–May), while minima occur most frequently in autumn (September–November). In the late 1980s and

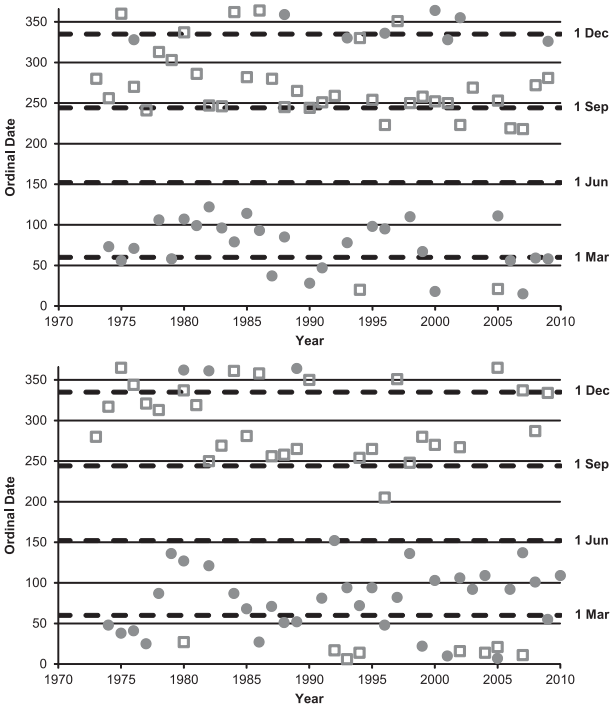

FIG. 4. Maximum (circles) and minimum (squares) values of the 91-day running mean of (top) POD and (bottom) FOH calculated using a 365-day moving window.

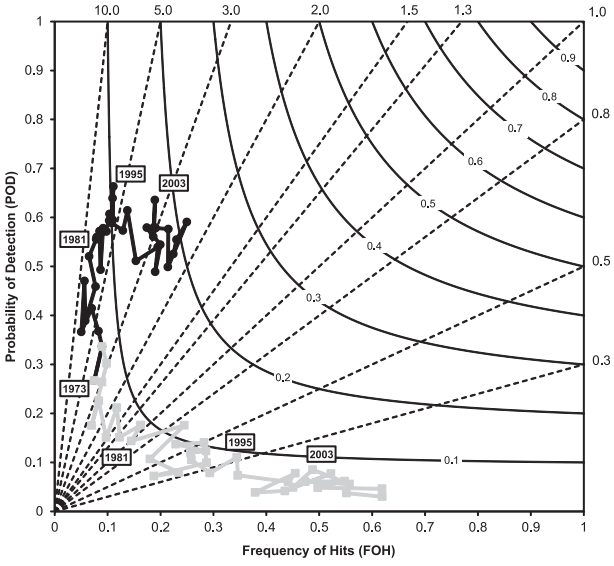

FIG. 5. Performance diagram [as in Fig. 1; see Roebber (2009)] of annual convective outlook performance for slight risk areas (black circles and line) and moderate risk areas (gray squares and line) from 1973 to 2010 at 80-km grid spacing. Labeled years are provided for context. The plot associated with the slight risk areas depicted here is also seen in Fig. 1.

early 1990s, however, the minima become more concentrated near 1 September and the occurrence of maxima in the early winter becomes more frequent, likely as a result of increased focus by forecasters on the late-autumn peak in severe storms (S. J. Weiss 2012, personal communication). The first POD maximum occurs in 1974 on day 73 (14 March), while the first minimum occurs on day 280 (7 October) of 1973. Similarly, the FOH maxima occur most frequently in the spring, and minima occur most frequently during autumn and into December. The first FOH maximum occurs in 1974 on day 48 (17 February), and the first minimum of this measure occurs on day 280 (7 October) of 1973. These occurrence patterns of maxima and minima are explained by the seasonal climatology of severe weather, with the measures maximized during the springtime when severe weather events are most frequent, and the conditions leading to these events are more synoptically evident (Johns and Doswell 1992). Conversely, the conditions leading to severe weather events in autumn can sometimes be less apparent, resulting in many of the minima occurring during these months. This may also be reflective of the annual cycle of false alarms and misses.

Although slight risk areas have been the focus of this study, the same performance measures are also calculated for moderate risk areas. Comparing the performance of both risk areas using a performance diagram (Fig. 5), it is immediately apparent that the POD values of moderate risk areas are much lower than slight risk area values, but the FOH values for moderate risk areas are much higher. For example, for the years 1981, 1995, and 2003 the POD values for slight risk areas are 0.52, 0.57, and 0.64, and for moderate risk areas the POD values are 0.15, 0.11, and 0.09. The FOH values for the same years are 0.07, 0.13, and 0.19 for slight risk areas, and 0.12, 0.34, and 0.49 for moderate risk areas. This result is not unexpected since beginning in 1975 moderate risks were almost exclusively subregions within slight risk areas, and are meant to focus on areas of heightened risk, especially high-severity events. Therefore moderate risk areas "miss" many low-severity events that decrease the POD, but suffer from fewer false alarms due to their smaller size and more targeted placement, accounting for large improvements in the FOH. For these reasons the results presented in Fig. 5 should not be interpreted as one risk type outperforming the other. Rather, each performed as expected given their intended purposes. Additionally, these data generally follow a line of constant CSI on a performance diagram, with the exception of slight risk areas after 1995.

4. Conclusions

This study sought to assess the performance of the SPC's day 1 convective outlooks by translating risk

areas and storm reports to the same grid, and used common verification measures obtained from a 2×2 contingency table to evaluate this product's performance since 1973. Specifically, POD and FOH values were examined, revealing a steady increase in the POD of slight risk areas over the first two decades, followed by an increase in the FOH. These results are indicative of an increase in the size of slight risk areas until 1993, followed by smaller but better-placed risk areas.

Ongoing research is being conducted to evaluate the performance of the entire suite of convective outlook products issued by the SPC including updates to the initial day 1 outlook, the day 2 and day 3 outlooks and their updates, as well as the probabilistic risk areas introduced publicly in 2001 that accompany categorical risk areas. Analysis of these products will explore changes in forecast performance as these products are adjusted over the course of days and within days. This analysis can be coupled with an evaluation of watch and warning quality. Eventually, the goal is to create a complete and seamless evaluation that will allow for the stratification (Murphy 1995) of forecasts by other forecasts (e.g., evaluate the warnings given that a slight risk and a tornado watch are in effect) and environmental conditions.

Acknowledgments. This research was performed while the first author held a National Research Council Research Associateship Award at the National Severe Storms Laboratory. The authors thank the Storm Prediction Center's Andy Dean for providing the convective outlook dataset and Steve Weiss for his helpful comments. The constructive comments and suggestions made by Steve Corfidi and Pieter Groenemeijer helped improve the manuscript.

APPENDIX

Contingency Table Values

TABLE A1. Annual contingency table values for the SPC's slight and moderate risk areas.

	Slight risk				Moderate risk			
Year	a	b	c	d	a	b	c	d
1973	818	10 151	2246	902 205	818	10 151	2246	902 205
1974	1003	10 208	1976	902 233	1003	10 208	1976	902 233
1975	1208	13 422	2073	898 717	991	9098	2290	903 041
1976	1021	13 588	1441	901 878	650	6632	1812	908 834
1977	986	18 634	1706	894 094	473	6312	2219	906 416
1978	1157	19 422	1303	893 538	555	6112	1905	906 848
1979	1047	17 057	1645	895 671	406	3775	2286	908 953
1980	1604	19 182	1890	895 252	739	5555	2755	908 879
1981	1430	20 560	1317	892 113	416	2978	2331	909 695
1982	2134	23 055	1596	888 635	657	3389	3073	908 301
1983	2016	21 280	2071	890 053	587	3468	3500	907 865
1984	2134	21 211	1559	893 024	649	2002	3044	912 233
1985	2159	25 327	1726	886 208	416	1894	3469	909 641
1986	2456	28 485	1930	882 549	349	1142	4037	909 892
1987	2298	26 672	1835	884 615	299	1296	3834	909 991
1988	2254	20 797	1690	893 187	392	973	3552	913 011
1989	3040	26 253	1966	884 161	703	1788	4303	908 626
1990	3168	27 537	2169	882 546	732	2469	4605	907 614
1991	3438	27 864	2339	881 779	713	1781	5064	907 862
1992	3803	30 984	2146	880 995	709	2051	5240	909 928
1993	4222	33 778	2143	875 277	670	1924	5695	907 131
1994	4097	33 848	2867	874 608	546	1316	6418	907 140
1995	4633	31 276	3445	876 066	892	1699	7186	905 643
1996	5025	31 544	3153	878 206	602	1137	7576	908 613
1997	3866	21 311	3694	886 549	456	577	7104	907 283
1998	5053	20 881	4311	885 175	730	875	8634	905 181
1999	3966	16 901	4143	890 410	314	515	7795	906 796

TABLE A1. (*Continued*)

	Slight risk				Moderate risk			
Year	a	b	c	d	a	b	c	d
2000	4744	19 015	3970	890 199	364	473	8350	908 741
2001	5148	24 421	3741	882 110	685	627	8204	905 904
2002	5091	22 225	3983	884 121	443	540	8631	905 806
2003	5805	24 849	3327	881 439	778	818	8354	905 470
2004	5102	22 059	3690	887 077	492	391	8300	908 745
2005	5317	22 546	3877	883 680	558	620	8636	905 606
2006	5791	21 171	4257	884 201	619	511	9429	904 861
2007	4703	17 275	4725	888 717	456	461	8972	905 531
2008	6389	19 180	4424	887 935	500	308	10 313	906 807
2009	4925	16 459	3954	890 082	268	165	8611	906 376
2010	4775	16 510	4310	889 825	339	276	8746	906 059

REFERENCES

Brooks, H. E., and Coauthors, 2011: Evaluation of European Storm Forecast Experiment (ESTOFEX) forecasts. *Atmos. Res.,* **100,** 538–546.

Corfidi, S. F., 1999: The birth and early years of the Storm Prediction Center. *Wea. Forecasting,* **14,** 507–525.

Doswell C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

——, S. J. Weiss, and R. H. Johns, 1993: Tornado forecasting: A review. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards, Geophys. Monogr.,* Vol. 79, Amer. Geophys. Union, 557–571.

——, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting,* **20,** 577–595.

——, ——, and N. Dotzek, 2009: On the implementation of the enhanced Fujita scale in the USA. *Atmos. Res.,* **93,** 554–563.

Gourret, J. P., and J. Paille, 1987: Irregular polygon fill using contour encoding. *Comput. Graph. Forum,* **6,** 317–325.

Johns, R. H., and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting,* **7,** 588–612.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1995: A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.,* **123,** 1582–1588.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting,* **24,** 601–608.

Vescio, M. D., and R. L. Thompson, 2001: Subjective tornado probability forecasts in severe weather watches. *Wea. Forecasting,* **16,** 192–195.