Evaluation of the Storm Prediction Center's Convective Outlooks from Day 3 through Day 1

NATHAN M. HITCHENS

Department of Geography, Ball State University, Muncie, Indiana, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

HAROLD E. BROOKS

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

(Manuscript received 13 November 2013, in final form 31 May 2014)

ABSTRACT

The Storm Prediction Center issues four categorical convective outlooks with lead times as long as 48 h, the so-called day 3 outlook issued at 1200 UTC, and as short as 6 h, the day 1 outlook issued at 0600 UTC. Additionally, there are four outlooks issued during the 24-h target period (which begins at 1200 UTC on day 1) that serve as updates to the last outlook issued prior to the target period. These outlooks, issued daily, are evaluated over a relatively long period of record, 1999–2011, using standard verification measures to assess accuracy; practically perfect forecasts are used to assess skill. Results show a continual increase in the skill of all outlooks during the study period, and increases in the frequency at which these outlooks are skillful on an annual basis.

1. Introduction

Beginning in the mid-1950s, the National Weather Service's Storm Prediction Center (SPC) has issued convective outlooks (COs) for the 48 contiguous states on a daily basis (Corfidi 1999). Presently, four COs are issued within 48 h prior to the start of the target period, a 24-h period beginning at 1200 UTC, and four are issued after the target period has commenced. These latter four products are considered updates to the final CO issued before the start of the target period, and are valid from the time they are issued to the end of the target period.

Previous studies have evaluated the accuracy (Hitchens and Brooks 2012) and skill (Hitchens et al. 2013) of a single CO product; accuracy is described by Murphy (1993) as the average correspondence between individual pairs of forecasts and observations, and skill as the accuracy of forecasts relative to the accuracy of a forecast produced by a standard of reference. Hitchens and Brooks (2012) found increases in performance measures of COs during the nearly 40-yr study period (1973–2010), and used these results to identify periods of time during which the SPC appeared to make changes to its general forecasting philosophy, resulting in better forecasts. Following this work, Hitchens et al. (2013) described a method to assess the skill of rare-event forecasts, using COs to illustrate their approach. They found that although measures of accuracy increased during the first two decades of the study period, the forecasts showed little to no skill relative to a no-skill baseline. Over the remainder of the study period, however, there was continual, steady improvement in forecast skill.

This study seeks to continue the line of inquiry presented in the aforementioned studies by assessing the accuracy and skill of the full suite of categorical CO products issued by the SPC. We will focus only on ''slight risk'' areas, which will henceforth be referred to simply as ''outlooks,'' since these forecasts are meant to include most severe reports. Slight risk areas encompass reported events at a much higher rate than elevated (e.g., ''moderate'') risk areas, providing a larger sample size to analyze; Hitchens and Brooks (2012) showed in their Fig. 5 that since the mid-1990s slight risk areas have included at least 50% of reported events, while

Corresponding author address: Dr. Nathan M. Hitchens, Dept. of Geography, Ball State University, Muncie, IN 47306.
E-mail: nmhitchens@bsu.edu

DOI: 10.1175/WAF-D-13-00132.1

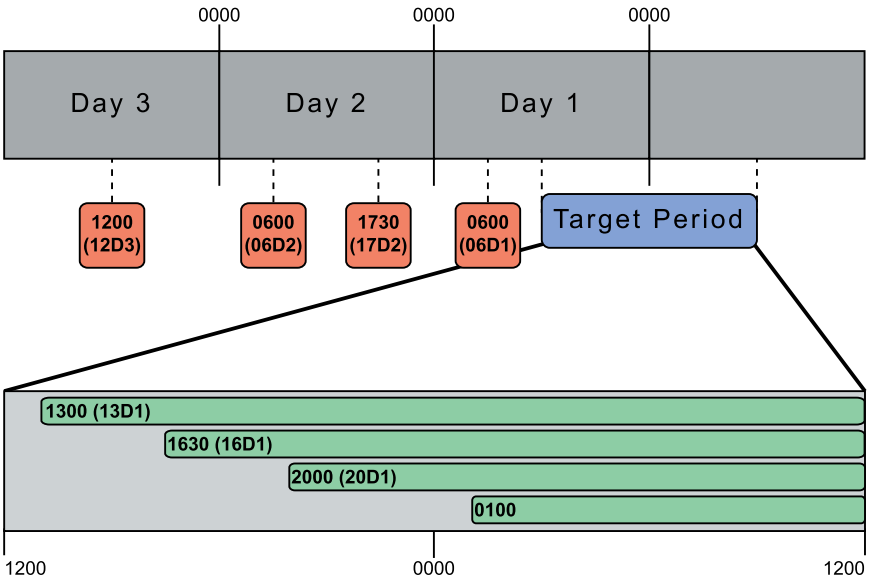

FIG. 1. The timeline (UTC) of COs issued prior to the beginning of the target period (red) and during the target period (green).

moderate risk areas have only included, at most, about 10% of these events. Of particular interest will be the evaluation of the improvement or decline in forecast accuracy and skill between consecutive forecasts. Since the forecast process differs between forecasts issued prior to the beginning of the period for which they are valid and those issued as updates during the target period, each set of forecasts will be examined separately.

2. Data and methods

The SPC currently issues four COs that are valid for the 24-h period beginning at 1200 UTC on their so-called day 1 (Fig. 1): the first is issued at 1200 UTC on day 3 (12D3; 48 h prior to the forecast's valid time), the second and third on day 2 at 0600 and 1730 UTC (06D2 and 17D2; 30 and 18.5 h prior), and the fourth at 0600 UTC on day 1 (06D1; 6 h prior).[1] Data are available from 2002 to 2011 for 12D3, 2000 to 2011 for 06D2 and 17D2, and 1973 to 2011 for 06D1.

Four additional COs are issued on day 1 as updates to the initial 0600 UTC CO (Fig. 1), with each valid from the time of issuance until 1200 UTC. The first is issued at 1300 UTC (13D1) and covers a period of 23 h, the second at 1630 UTC (16D1) with a period of 19.5 h, the third at 2000 UTC (20D1) with a period of 16 h, and the fourth at 0100 UTC with a period of 11 h. The update at 0100 UTC is usually issued after the primary severe weather threat for day 1 has passed, and since in most cases it represents a different forecast (covering the remaining convection during the overnight hours) than previous CO products, it will not be included in this study. Data are available for each CO update from 1999 to 2011.

Following Hitchens and Brooks (2012), each outlook is first gridded on a 0.1° × 0.1° (~10 km) latitude–longitude grid, and then aggregated onto a grid with an approximate spacing of 80 km × 80 km (equivalent to the area associated with SPC forecast definitions of the probability of an event occurring within 25 mi of a point). To verify these outlooks, reports of tornadoes, hail, and wind that meet the National Weather Service's criteria for severe[2] are placed on the same grid as the outlooks. Separate grids of reports are created for each target period to be evaluated: the 24-h period beginning at 1200 UTC (for 12D3, 06D2, 17D2, and 06D1) and the shortened periods associated with each day 1 update (13D1, 16D1, and 20D1). Each cell in each report grid is considered a dichotomous event; multiple reports occurring in a single grid box do not have more influence

[1] The SPC also issues COs for days 4–8, but these products only indicate areas with a probability of severe weather that exceeds 30%. Since these COs differ substantially from those of days 1–3 and have been issued for a shorter period of time, they are not included in this study.

[2] The National Weather Service currently defines a severe thunderstorm as one that produces a tornado, a wind gust of at least 50 knots (kt; 1 kt = 0.51 m s^{-1}), or hail with a diameter of at least 1.0 in. Prior to 5 Jan 2010, the min hail size criterion was 0.75 in. (National Weather Service 2014), but was increased due to research suggesting that damage to older roofs begins with hailstones that have a 1 in. diameter (Marshall et al. 2002).

TABLE 1. Contingency table (2 × 2) for forecasts and observations where quantities of interest are: POD = $a/(a + c)$, FOH = $a/(a + b)$, CSI = $a/(a + b + c)$, and $B = (a + b)/(a + c)$.

	Observed yes	Observed no	Sum
Forecast yes	a	b	$a + b$
Forecast no	c	d	$c + d$
Sum	$a + c$	$b + d$	n

than a single report. This allows for the construction of a 2 × 2 contingency table (Table 1) and subsequent calculation of standard performance measures, such as the probability of detection (POD), frequency of hits (FOH), and critical success index (CSI). For a complete description of these measures, please refer to Doswell et al. (1990).

The performance measures alone allow for the evaluation of the accuracy of these outlooks, but additional information is needed to assess aspects related to skill. Using the techniques described in Hitchens et al. (2013), a "practically" perfect (PP) forecast is developed for each target period by using nonparametric density estimation with a two-dimensional Gaussian kernel to smooth the reported events. At each grid point the PP forecast value f is given by

$$f = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{d_n}{\sigma}\right)^2\right], \quad (1)$$

where d_n is the distance from the forecast grid point to the nth location that had an event occur, N is the total number of grid points with events, and σ is a weighting function that can be interpreted as the confidence one has in the location of the forecast event. To represent the uncertainty associated with convective outlooks, the value assigned to σ is 1.5, which translates to 120 km for an 80-km grid. The PP values for each forecast range from 0 to 1, and are interpreted as the probability of a severe report occurring in that grid box. The PP forecasts are then used in conjunction with the reports to determine the maximum score of a particular performance measure (e.g., CSI) that could reasonably be attained by a forecaster, as well as the minimum score

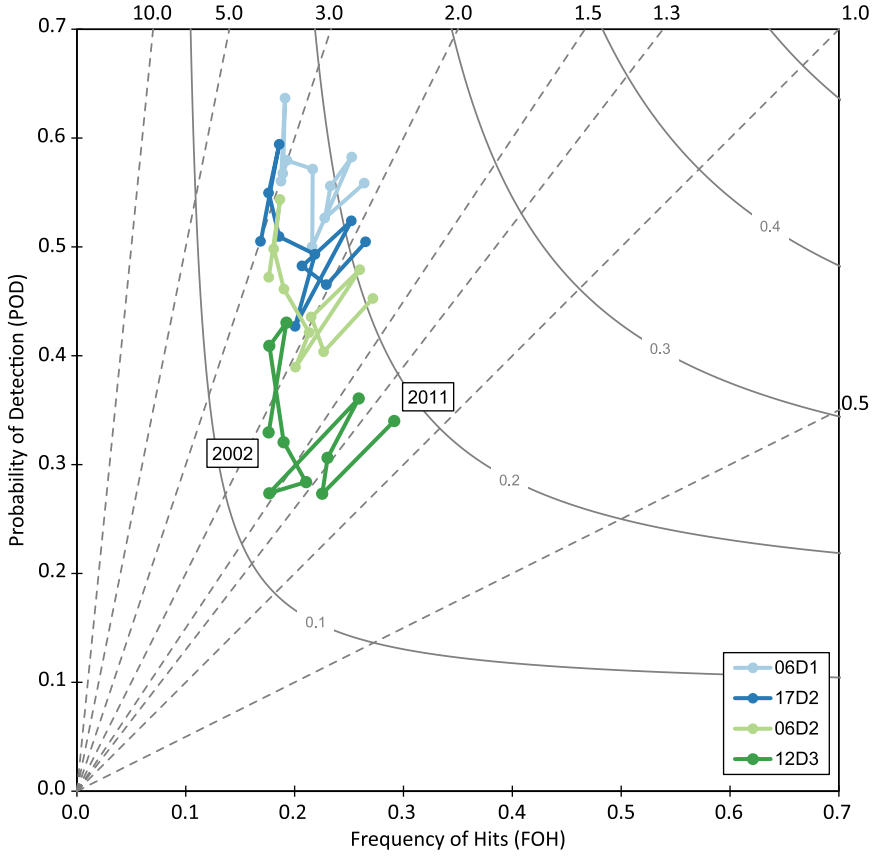

FIG. 2. Diagram (Roebber 2009) showing annual performance from 2002 to 2011 for outlook areas from day 3 through day 1 in terms of POD and FOH. The dashed lines represent B, while the curved lines are for CSI. Labeled years are provided for context.

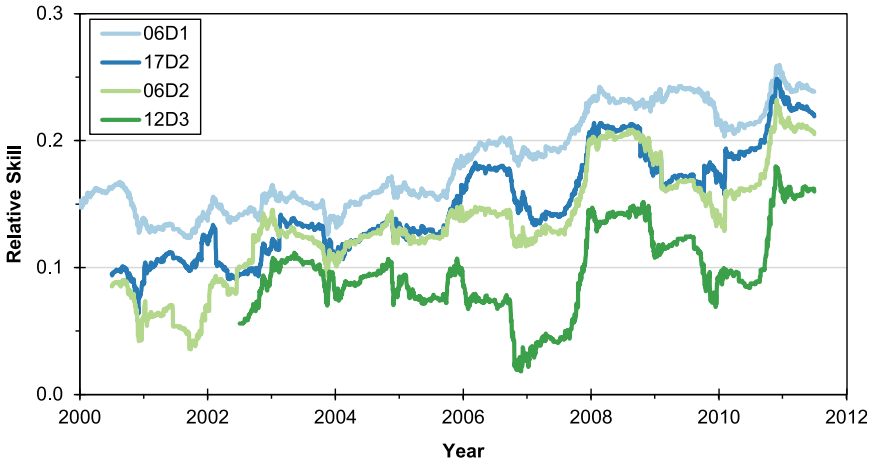

FIG. 3. Relative skill of the SPC outlook areas from day 3 through day 1 calculated as the relative position of each outlook's CSI value between the corresponding max and min CSI values using 365-day running means from the PP forecast.

(no-skill baseline). To obtain the maximum score, PP areas exceeding a threshold value, beginning at 0.01 and increased by increments of 0.01, are treated as individual, dichotomous forecasts of severe weather, and are compared to reported events, with a contingency table constructed and a CSI value calculated for each threshold. The term practically perfect is meant to describe the forecast as one that is consistent with that which a forecaster would make if given perfect knowledge of the reported events beforehand, while still adhering to the basic constraints of the forecast system (e.g., similar size and shape). All PP forecasts are placed on the same grid as the outlooks and reports.

Within the context of this study the terms "skill" and "relative skill" are both used to describe the relative position of the CSI value of an outlook between the minimum and maximum CSI values obtained from the PP forecast for the same period of time and are defined as

$$\text{skill} = \frac{\text{CSI}_{\text{outlook}} - \text{CSI}_{\text{min}}}{\text{CSI}_{\text{max}} - \text{CSI}_{\text{min}}}. \quad (2)$$

Forecasts with positive values from (2) are considered skillful, meaning a certain amount of value was added by the forecaster beyond that which could be achieved by a person with no forecasting knowledge or experience. Theoretically, the no-skill baseline would be defined as the entire domain, but in practice this minimum CSI value is calculated by linearly extrapolating the CSI values obtained using PP values of 0.02 and 0.01 (Hitchens et al. 2013).

3. Results

The analysis of these forecasts is divided into two logical parts: those forecasts issued prior to 1200 UTC on day 1, and those issued after. The 06D1 forecast

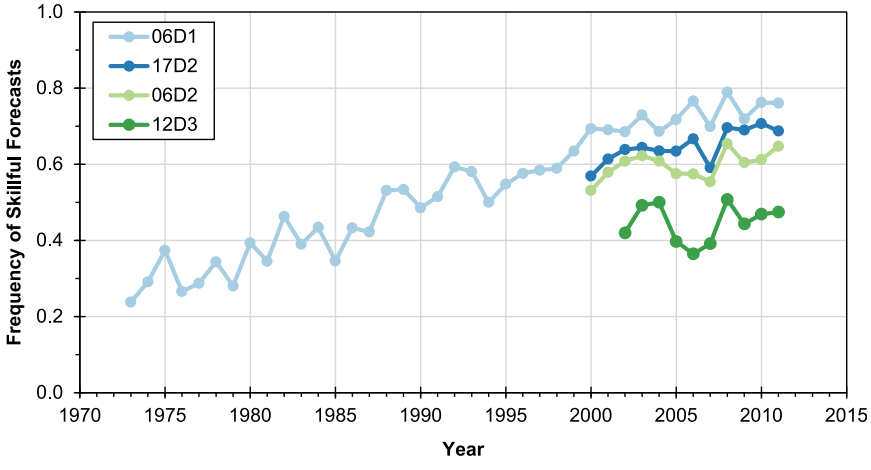

FIG. 4. Frequency of skillful daily forecasts by year for outlooks from day 3 through day 1.

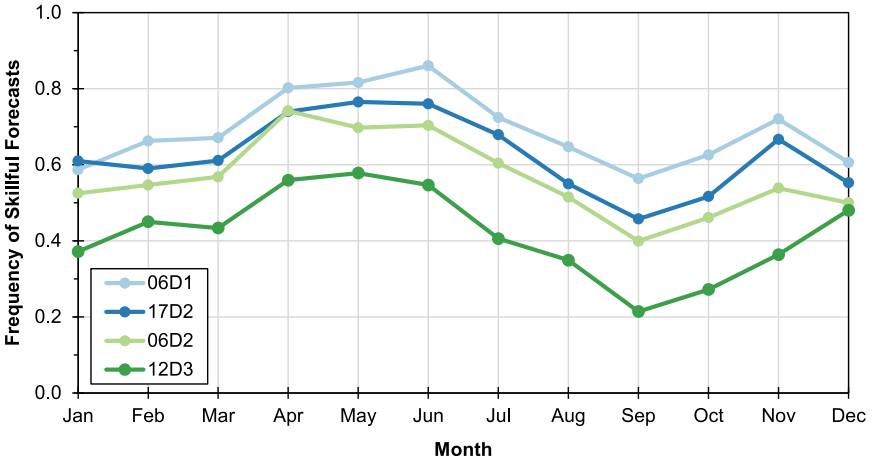

FIG. 5. Frequency of skillful daily forecasts by month for outlooks from day 3 through day 1 during 2002–11.

serves as a bridge between the two analyses because it is the last forecast issued before the beginning of the target period, and also the first forecast issued on day 1.

a. From day 3 through day 1

The annual mean accuracy[3] of the outlooks preceding the target period shows increases in values of POD with decreasing lead time, but little change in FOH and CSI (Fig. 2). For instance, in 2002 the POD values for 12D3, 06D2, 17D2, and 06D1 were 0.33, 0.47, 0.50, and 0.56, respectively, while the FOH (CSI) values for the same outlooks were 0.18 (0.13), 0.18 (0.15), 0.17 (0.14), and 0.19 (0.16). However, when examining each outlook during 2002–11 (the 10-yr period for which data were available for all four products), the POD values show no trend, with noticeable improvements in FOH and CSI from 2002 through 2011. In 2002 the POD, FOH, and CSI of the 12D3 outlooks were 0.33, 0.18, and 0.13, and in 2011 these values were 0.34, 0.29, and 0.19; POD values ranged from 0.27 to 0.43 for 12D3, 0.39 to 0.54 for 06D2, 0.43 to 0.59 for 17D2, and 0.49 to 0.64 for 06D1. From a forecasting perspective, these results indicate that outlooks were either better located and/or increased in size as lead times decreased to account for the improvements in POD. To a degree the latter appears to be true since the size of the outlooks increased with decreasing lead time more than 60% of the time, and also bias B increased with decreasing lead time. However, values of CSI also increased with decreasing lead time, suggesting that although there was an increase in the size of the outlooks, it was accompanied by improvements in location that correctly forecast a greater number of reported events. Over time the bias decreased for each outlook, indicating a decrease in size of the outlooks relative to the size of the reported events. This likely played a role in the increases in FOH and CSI over the decade, since a decrease in the rate of false alarms (1 − FOH) can result from smaller and better located outlooks.

Changes made by SPC forecasters between consecutive outlooks are influenced by both meteorological and nonmeteorological factors; with respect to the latter, there are three specific nonmeteorological factors that may affect any forecast beyond the initial 12D3 outlook. The first is the fact that a forecaster does not begin creating an outlook with a blank slate; he or she begins the process aware of the previous outlook that is being updated. The second factor is a matter of continuity between consecutive outlooks, wherein a forecaster strives to maintain a level of consistency between successive outlooks and, also, to not drastically change the message intended for the users of these forecasts. Finally, it is likely that the presence of upcoming, later outlooks provides forecasters with the opportunity to allow for future changes to these forecasts depending upon the situation.

Since 2000, each outlook has been skillful when expressed as a 365-day running mean[4] (Fig. 3), and each has continually improved over that time. The day 2 outlooks have consistently improved upon 12D3, and

[3] The annual mean value of each verification measure is calculated using a 2 × 2 contingency table that represents the totals for a calendar year.

[4] As in Hitchens et al. (2013), 365-day running means are computed by constructing a 2 × 2 table that sums all 365 forecasts centered on each day. In the case of CSI values from outlooks, the 2 × 2 table associated with each day's CSI value is used in the construction of the table for the 365-day period.

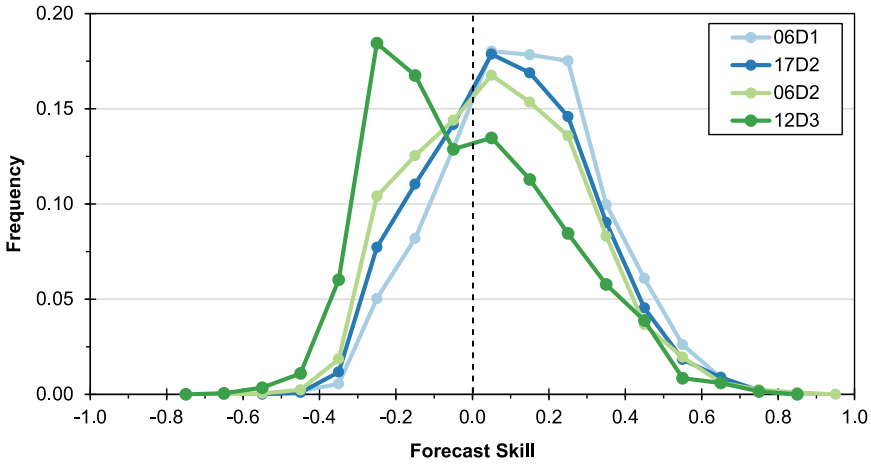

FIG. 6. Frequency of daily forecast skill binned in 10% increments for outlooks from day 3 through day 1 during 2002–11.

06D1 has consistently improved upon the day 2 outlooks, but 13% of the time 06D2 outperformed 17D2, most recently in early 2009. The similarity in skill values for the two day 2 products is likely due to the forecasting philosophy of the SPC to maintain continuity between successive forecasts discussed above, such that there are likely few substantive changes made to the 17D2 outlook compared to 06D2. The largest increases in skill between consecutive forecasts range from 0.06 to 0.10.

Expanding on the work done by Hitchens et al. (2013), the annual frequency of skillful forecasts shows increases in frequency between successive forecasts (Fig. 4), as well as a trend of improving frequencies from the early 2000s through 2011. Much like with the 365-day running means, the greatest increase in frequency occurs between 12D3 and 06D2, but in this case 17D2 is more frequently skillful on an annual basis compared to 06D2, although never by more than 0.09. Since 2000 at least 50% of the 06D2, 17D2, and 06D1 outlooks in a given year have been skillful, while only twice has the 12D3 reached or exceeded that level. However, a comparison of the frequencies in Fig. 4 shows that the values of 12D3 from 2002 to 2011 compare well to those of 06D1 during 1980–91; likewise, 06D2 and 17D2 from 2000 to 2011 compare well to 06D1 during 1991–2002. These results suggest that the frequency at which these outlooks are skillful improves by approximately one "day of lead time" every 10–12 yr.

The annual cycle of skillful forecast frequency (Fig. 5) shows each outlook peaking during the 3-month period of April–June (58% in May for 12D3, 74% in April for 06D2, 76% in May for 17D2, and 86% in June for 06D1), and each with its minimum in September (21%, 40%, 46%, and 56%). Only twice is an outlook more

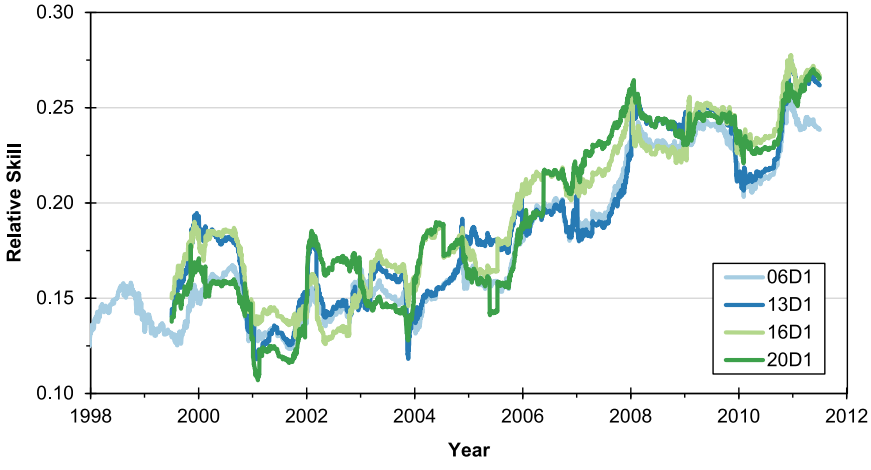

FIG. 7. Relative skill of the SPC's initial day 1 outlook (0600 UTC) and updates to this forecast calculated as the relative position of each outlook's CSI value between the corresponding max and min CSI values using 365-day running means from the PP forecast.

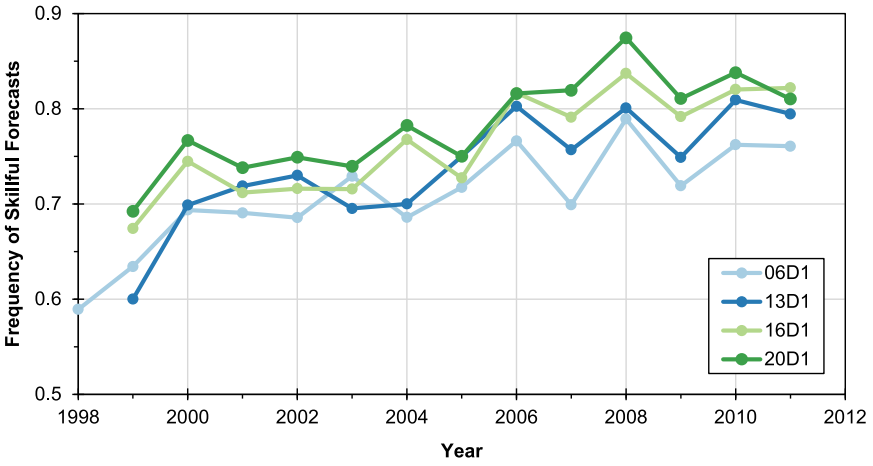

FIG. 8. Frequency of skillful daily forecasts by year for the initial day 1 outlook (0600 UTC) and the updates to this forecast.

frequently skillful than the outlook that follows it: 17D2 (61%) and 06D1 (59%) in January, and 06D2 (74.2%) and 17D2 (73.9%) in April. The annual cycle for 12D3 compares well with the annual cycle of 06D1 during 1982–91 from Hitchens et al. (2013, their Fig. 8), further supporting the idea that the frequency of skillful forecasts improves by 1 day decade^{-1}.

Examination of the frequency of relative skill values reveals an increase in the proportion of skillful forecasts with decreasing lead time (Fig. 6); 44% of 12D3 outlooks are skillful, 61% of 06D2, 66% of 17D2, and 73% of 06D1. Between 12D3 and 06D2 there are substantial improvements in the frequency of skillful forecasts with relative skill values from 0.0 and 0.4, and between 17D2 and 06D1 there are improvements in skill values from 0.1 to 0.6, but the only improvements between 06D2 and 17D2 are from 0.0 to 0.3. There is a large peak in the frequency of forecasts lacking skill in 12D3 from −0.2 to −0.3, and in 06D2 and 17D2 there is a large increase in frequency in the same range of values. This result is due to an increase in "missed" events (reported events that were not forecast), but as discussed by Hitchens et al. (2013), many of these seemingly missed events were forecast as "see text,"[5] so they were not entirely missed in the strictest sense.

b. *Day 1 updates*

A comparison of the 365-day running mean of the relative skill of 06D1 and each of the three updates to this outlook shows continual improvement of each product since the beginning of 2004 (Fig. 7), but minimal differences between individual forecasts. The largest range of skill values is 0.044 between 06D1 (0.148) and 13D1 (0.192) from the end of December 1999, and the smallest is 0.007 from the end of August 2009 between 06D1 (0.240) and 16D1 (0.247). Of these four outlooks, 06D1 had the least skill 41% of the time, while 16D1 had the most skill 50% of the time. Interestingly, 20D1 was the outlook with the second highest frequency for both the least (25%) and most (34%) skill. Being the latest of the four outlooks, the 20D1 outlook is likely adjusted more than the others based on ongoing convection and the evolving environmental conditions. In fact, relative to 16D1 the 20D1 outlook decreased in size 48% of the time, while it increased in size only 44% of the time. In comparison, 13D1 increased (decreased) in size 56% (37%) of the time, and 16D1 increased (decreased) 53% (39%) of the time.

Since 1999 the 06D1 outlook and the three updates to it have been skillful at least 60% of the time, and since 2005 these four outlooks have been skillful at least 70% of the time (Fig. 8). For 11 of the 13 years the 20D1 outlook has been the most frequently skillful, with only 16D1 exceeding the frequency of 20D1 in 2006 (by 0.1%) and 2011 (by 1.6%). Considering simply the annual frequency of improved skill between temporally consecutive outlooks,[6] the 13D1 outlook most often improved in relative skill from 06D1 (Fig. 9a). The 13D1 outlook improved on the relative skill of 06D1 more than 50% of the time for 8 of 13 years, while the other two outlooks only improved more than 50% of the time

[5] Beginning in 1999, the SPC began issuing "see text" areas within COs. These areas have no defined spatial extent, and are used when severe weather is expected but falls below the threshold of a slight risk.

[6] Also included in the frequency of "improved skill" are consecutive outlooks with positive relative skill values that stayed the same. While not strictly an improvement, we believe a forecaster should be given credit for not changing a skillful forecast.

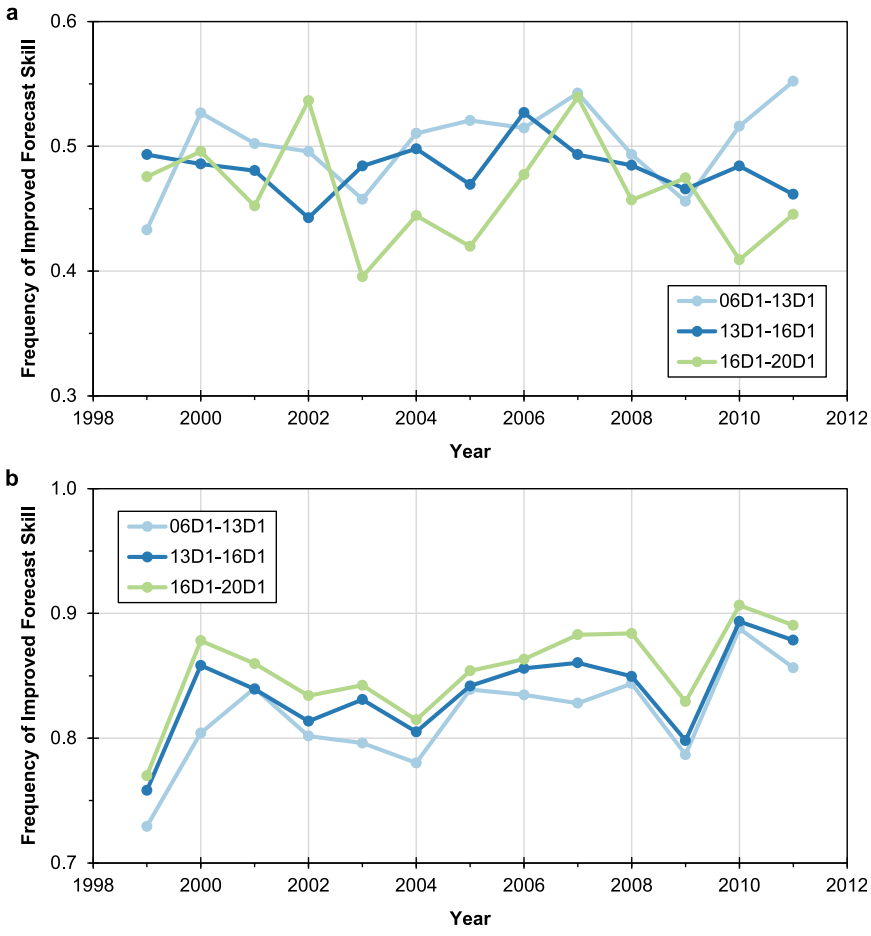

FIG. 9. Frequency of improvement in daily forecast skill between consecutive forecasts based on (a) the relative skill of each outlook and (b) the relative skill of the earlier outlook calculated with the reported events from the time period associated with the later outlook.

on the skill of the preceding outlook 3 times (16D1 in 2006, and 20D1 in 2002 and 2007). These results seem to indicate that updated outlooks only improve on the previous forecast at best about 55% of the time, but using this approach to compare two consecutive forecasts might not be appropriate. A better approach is to compare the skill of the later outlook to the skill of the previous outlook when calculated using the reported events from the time period of the later outlook (e.g., calculate the skill of 06D1 using the reported events from the time period of 13D1). For the purpose of comparison, the earlier outlook is being treated as a ''persistence'' forecast. The results using this approach are far different, with each update improving on the persistence version of the previous outlook more than 70% of the time annually (Fig. 9b). Further, each successive update improves upon the previous outlook more frequently, with the 20D1 outlook improving upon the skill of the 16D1 outlook 91% of the time in 2010. Forecasters are likely improving on the previous outlook at such a high rate because of additional information being available in the form of new data from models, observations of environmental conditions, and the initiation and evolution of convection.

4. Concluding remarks

The purpose of this study was to assess the accuracy and skill of the SPC convective outlook products beginning with day 3 and ending with the 2000 UTC day 1 outlook. This was accomplished by gridding each outlook and corresponding reported severe weather events, and constructing 2×2 contingency tables. The relative skill of each outlook was determined by using practically perfect forecasts to establish the practical maximum and minimum CSI scores for each forecast. From the analysis of the outlooks from day 3 through day 1, it was found that the relative skill levels of these forecasts have been

improving over the last decade, and there are improvements in skill between forecasts with decreasing lead time. Additionally, the frequency of skillful forecasts appears to improve by 1 day of lead time every 10–12 years for a particular outlook. The analysis of the updates to the initial day 1 outlook shows a continual improvement in the skill of each forecast over the last decade, although there is often little difference in skill between consecutive forecasts. These outlooks were frequently skillful, often showing skill more than 70% of the time in a given year. By using a persistence-style approach to calculating the skill of the updates, it was found that these outlooks improved upon the skill of the preceding outlook more than 70% of the time each year. Future work will focus on the evaluation of the SPC's probabilistic convective outlook products, with an emphasis on developing a system that incorporates the SPC's performance on past outlooks into the forecasting process.

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. The constructive comments and suggestions made by the three anonymous reviewers helped improve the manuscript.

REFERENCES

Corfidi, S. F., 1999: The birth and early years of the Storm Prediction Center. *Wea. Forecasting,* **14,** 507–525, doi:10.1175/1520-0434(1999)014<0507:TBAEYO>2.0.CO;2.

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, doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting,* **27,** 1580–1585, doi:10.1175/WAF-D-12-00061.1.

——, ——, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting,* **28,** 525–534, doi:10.1175/WAF-D-12-00113.1.

Marshall, T. P., R. F. Herzog, S. J. Morrison, and S. R. Smith, 2002: Hail damage threshold sizes for common roofing materials. Preprints, *21st Conf. on Severe Local Storms,* San Antonio, TX, Amer. Meteor. Soc., P3.2. [Available online at https://ams.confex.com/ams/pdfpapers/45858.pdf.]

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

National Weather Service, cited 2014: Why one inch hail criterion? [Available online at http://www.nws.noaa.gov/oneinchhail/.]

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