# Probabilistic Verification of Storm Prediction Center Convective Outlooks

GREGORY R. HERMAN, ERIK R. NIELSEN, AND RUSS S. SCHUMACHER

*Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*

(Manuscript received 18 July 2017, in final form 20 November 2017)

ABSTRACT

Eight years' worth of day 1 and 4.5 years' worth of day 2–3 probabilistic convective outlooks from the Storm Prediction Center (SPC) are converted to probability grids spanning the continental United States (CONUS). These results are then evaluated using standard probabilistic forecast metrics including the Brier skill score and reliability diagrams. Forecasts are gridded in two different ways: one with a high-resolution grid and interpolation between probability contours and another on an 80-km-spaced grid without interpolation. Overall, the highest skill is found for severe wind forecasts and the lowest skill is observed for tornadoes; for significant severe criteria, the opposite discrepancy is observed, with highest forecast skill for significant tornadoes and approximately no overall forecast skill for significant severe winds. Highest climatology-relative skill is generally observed over the central and northern Great Plains and Midwest, with the lowest—and often negative—skill seen in the West, southern Texas, and the Atlantic Southeast. No discernible year-to-year trend in skill was identified; seasonally, forecasts verified the best in the spring and late autumn and worst in the summer and early autumn. Forecasts are also evaluated in CAPE-versus-shear parameter space; forecasts struggle most in very low shear but also in high-shear, low-CAPE environments. In aggregate, forecasts for all variables verified more skillfully using interpolated probability grids, suggesting utility in interpreting forecasts as a continuous field. Forecast reliability results depend substantially on the interpretation of the forecast fields, but day 1 and day 2–3 tornado outlooks consistently exhibit an underforecast bias.

## 1. Introduction

Severe weather—defined as the presence of one or more tornadoes of any intensity, convectively induced wind gusts of at least 58 mi $h^{-1}$ (93 km $h^{-1}$), or thunderstorms producing 1 in. (2.54 cm) or larger hail—poses a substantial threat to life and property over much of the United States and is collectively responsible for an annual mean of 137 fatalities and $4.69 billion in damages (NWS 2017a) over the past eight years. Outlooks and other forecasts from the Storm Prediction Center (SPC) are among the leading sources of severe weather forecast information for National Weather Service (NWS) meteorologists, broadcast meteorologists, emergency managers, and the public. The SPC routinely produces and updates numerous products from nowcasts to forecasts with 8 days of lead time, and these forecasts are publicly archived—some as far back as 2003 (SPC 2017a). However, despite the substantial viewership and reliance of end-user communities on SPC products, the specificity and concreteness of their forecast predictands, their standing as a ''gold standard'' for severe weather forecasting (e.g., Stough et al. 2010), and SPC's transparency in making available both their contemporary and historical forecasts, much remains unknown about the quality of their outlook products as a result of gaps in the published verification of SPC outlooks. This study seeks to rectify this by performing quantitative verification of these forecast products and in particular the probabilistic convective outlooks.

As alluded to above, SPC is responsible for the routine issuance of a wide variety of products, including their convective outlooks, which are the focus of this study. Convective outlooks are produced for days 1–8, when the valid period for a forecast day spans 1200–1159 UTC, but the outlook specifics vary as a function of forecast lead time. For day 1, each of the three severe weather elements—tornadoes, hail, and wind—are treated separately, while for day 2 and beyond, all three are instead treated collectively. Outlooks include both categorical and probabilistic components; the latter use neighborhood probabilities whereby contours are drawn

*Corresponding author*: Gregory R. Herman, gherman@atmos.colostate.edu

DOI: 10.1175/WAF-D-17-0104.1

to define lines of the constant probability of observing the given severe weather predictand within approximately 25 mi (40 km) of a point. For days 1–3, categorical outlooks are also provided but are simply a strict function of those neighborhood probabilities. Day 1 tornado forecasts include 2%, 5%, 10%, 15%, 30%, 45%, and 60% neighborhood probability contours; day 1 hail, day 1 wind, and day 2 and 3 aggregate severe forecasts employ a different contour set: 5%, 15%, 30%, 45%, and 60%. Furthermore, for all day 1–3 forecasts, a higher intensity level, "significant severe"—defined as a tornado with a rating of at least EF2, thunderstorms producing hail with diameters of ≥2 in. (5 cm), or convective wind gusts ≥ 75 mi $h^{-1}$ [65 kt, 33 m $s^{-1}$; Hales (1988)]—is considered, and an additional significant severe contour is drawn for 25-mi neighborhood probabilities of significant severe weather ≥ 10%. Day 4–8 probability forecasts, also made collectively for any severe weather, use only two contours, 15% and 30%, and do not directly consider the elevated significant severe criteria. Day 1 forecasts are routinely produced at 0600, 1300, 1630, 2000, and 0100 UTC, with the 0600 UTC outlook being the first day 1 outlook that covers a given valid period. Day 2 and 3 forecasts are disseminated, respectively, at 0100 and 0230 central time (CT, with the UTC time varying based on daylight saving time); day 2 receives an additional update at 1730 UTC (Hitchens and Brooks 2014; Edwards et al. 2015).

The majority of previously published severe weather verifications have focused on severe thunderstorm or tornado watch (e.g., Doswell et al. 1990; Anthony and Leftwich 1992; Doswell et al. 1993; Vescio and Thompson 2001; Schneider and Dean 2008) or warning (e.g., Polger et al. 1994; Bieringer and Ray 1996; Simmons and Sutter 2005; Barnes et al. 2007; Simmons and Sutter 2008; Brotzge et al. 2011; Anderson-Frey et al. 2016) verification. However, there has been some limited verification of convective outlooks in the literature. In particular, Hitchens and Brooks (2012) and Hitchens and Brooks (2014) verified day 1 convective outlooks, with the latter also verifying day 2 and 3 outlooks. The primary purpose of these studies was to evaluate the skill of SPC outlooks over a very long period of record—decades—to ascertain temporal trends in performance and the effects of changes in SPC forecasting philosophies on forecast skill over time. However, both of these studies only deterministically considered the verification of the *categorical* versions of the convective outlooks via contingency table statistics and did not quantitatively consider the *probabilistic* verification of the SPC probability contours. Additionally, the choice to verify categorical outlook contours rather than probabilistic ones made it impossible to perform verification broken out individually by a severe weather predictand for day 1 outlooks, since the categorical outlooks are objectively determined as a combination of the individual severe weather predictand probabilities—even if probabilities may be subjectively modified to match forecaster conceptions of categorical severity levels—and, historically, were not broken out by phenomenon at all. There has been some very limited published work exploring the verification of probabilistic SPC convective forecasts (e.g., Kay and Brooks 2000), but it is quite dated, preceding the introduction of operational convection-allowing model guidance (e.g., Kain et al. 2006), and substantial changes to both SPC outlook products and the available operational guidance in the forecast process that have been introduced in the intervening years (Edwards et al. 2015). Recently, Hitchens and Brooks (2017) have begun to investigate the verification of probabilistic SPC convective outlooks. However, while substantially advancing the literature in this area, the verification presented still employs deterministic contingency-table-based frameworks and thus, largely neglects the specific quantitative information associated with the probability contours being evaluated.

In this paper, we seek to quantify probabilistic verification properties of SPC severe weather outlooks for days 1–3, in particular forecast reliability and forecast skill. The following section describes the methods performed to conduct this verification and outlines two different verification frameworks employed in the study: a so-called traditional framework and an interpolation one. Section 3 presents verification results using the traditional analysis approach, while section 4 describes the interpolation framework results and provides a comparison between the two. The paper concludes with a synthesis of the findings and a discussion of broader applications and implications of this work.

## 2. Data and methods

Forecast data for this study come from the shapefiles in the public SPC outlook archive (SPC 2017a). SPC has changed various aspects of both their product definitions and their archive over the past 10–15 years. Importantly and consequentially, the NWS changed the definition of severe hail from a minimum hail diameter of 0.75 in. (1.9 cm) to 1 in. (2.54 cm) beginning on 5 January 2010 (Ferree 2009), considerably reducing the number of annual severe hail reports after that date. Verification in this study is performed relative to the effective severe hail criteria at the forecast issuance time. Their categorical convective outlooks were also substantially innovated in October 2014, adding "marginal" and

''enhanced'' categories to the existing classes of ''slight,'' ''moderate,'' and ''high'' (Jacks 2014). However, these changes only affected how severe weather probability contours mapped to categorical risk definitions and did not directly affect any of the probabilistic forecast contours. The public online forecast archive dates back to 23 January 2003, but forecasts are not available in shapefile format at that time. Shapefiles become available for day 1 beginning 1 January 2009 and for days 2 and 3 from around 11 April 2012. The file format is consistent for all day 1 shapefiles from the beginning of their archival, but a significant format change is incurred in the day 2 and day 3 shapefiles on approximately 13 September 2012. For these reasons, the period of record for forecast verification spans 1 January 2009–31 December 2016 (2922 total outlooks) for day 1 forecasts and 13 September 2012–31 December 2016 for day 2 and 3 forecasts (1569 and 1568 total outlooks, respectively). To maximize the period of record length, limit forecasts already affected by ongoing convection from the day of forecast issuance, and keep the issuance lead time separation—especially for days 2 and 3—as close as possible, verification in this study is based on the 1300 UTC probabilistic convective outlooks for day 1, the 0100 CT probabilistic convective outlooks for day 2, and the 0230 CT outlooks for day 3.

Archived forecast shapefiles store a list of points defining each polygon issued in the given probabilistic forecast. The verification for this study seeks to compute continental United States (CONUS)-wide verification statistics making use of the quantitative forecast probabilities in a consistent, repeatable manner for each forecast day throughout the period of record. These objectives are by far the most easily and equitably achieved using probability forecasts on a uniform grid for each forecast day. This thus requires the conversion of the contour definitions provided in the SPC shapefiles to probability grids.

Verification in this study is performed using two distinct, but complementary approaches. The first approach, the ''traditional'' verification framework, closely follows the verification performed internally at SPC (R. Edwards 2017, personal communication). Probabilities are gridded onto a CONUS-wide grid with 80-km grid spacing. Probabilities on this grid simply correspond to the value of the innermost probability contour enclosing the grid point, or are zero if no such contour exists. No interpolation is performed between probability contours in this verification scheme, and the forecast grids instead reflect a discrete number of possible forecast probabilities as determined by the allowed contour levels for the given predictand. This verification is performed to correspond with the current state of the science and for a direct comparison with internal statistics historically computed at SPC.

Recognizing however that appropriate verification is determined by the predictand definition(s) in conjunction with the forecast objectives and not on historical practice alone, a second, parallel verification approach is performed with the aim of advancing the state of the science in probabilistic severe weather forecast verification and obtaining consistency with public interpretation and the use of SPC outlooks. Ultimately, verification of neighborhood-based predictands such as those used in SPC's probabilistic convective outlooks occurs in continuous—rather than gridded or discrete—space. A severe weather observation occurs at an arbitrary physical point in space and, of course, is not constrained to occur on any grid of finite size; the circle defining 40 km centered about that point is similarly unconstrained to any particular grid. This argues against the use of grids at all, since their use only distorts the ''true'' relationship between the forecast and the observations; the distortion extent is proportional to the grid spacing, with no noise added over the true relationship in the case of infinitesimal grid spacing. Grids are used to provide a common and convenient quantitative analysis framework for comparing forecasts and observations, but in order to best represent the true relationship between these fields, it is desirable to have as small of a grid spacing as is computationally feasible, and it is certainly desirable to have a grid spacing appreciably smaller than the neighborhood radius of the predictand. Additionally, although SPC forecasters may only issue forecast probabilities at a given point corresponding to discrete levels defined by the allowable contours, some end users may reach the interpretation that within a region bounded by two contours, the verification probability is higher at a point directly adjacent to the higher-probability contour when compared with a point adjacent to the lower-probability one. Regardless of forecaster interpretation and intent, this is consistent with how the public and broader meteorological community may interpret plots of other continuous fields with a discrete number of contours (e.g., Lackmann 2011; NWS 2017b), and it is important to evaluate forecasts in a manner consistent with how forecasts are perceived by an educated end user. To this end, in addition to performing verification on an 80-km grid without probability interpolation in the so-called traditional approach, verification is also performed on a finer-resolution grid with interpolated probabilities in the interpolation approach described below.

For the interpolation approach, ArcGIS was used to process the SPC shapefiles into probability grids using a dynamic workflow divided into three different methods based upon the characteristics of the probabilities issued on a specific day. The first of these methods (hereafter

referred to as INTERP) interpolates between the SPC probability contours when more than one contour is present in the daily convective outlook and outputs the results onto the specified probability grid. The second method (hereafter referred to as CONSTANT) does no interpolation and is used when one contour level is present in the daily convective outlook since the lack of a defined probability gradient leaves the interpolation problem unconstrained. The third method (hereafter referred to as NODATA) is used when no contours are present in the daily convective outlook and outputs a constant grid of zero values over a CONUS-wide analysis domain that extends so far as the center of the neighborhood is over CONUS land. Open contours that end because of intersection with a CONUS boundary are closed using the CONUS edge as the remaining contour boundary such that all of that area is enclosed. Because of differences between the day 1, 2, and 3 outlook shapefiles, the workflow was carried about at horizontal resolutions of 0.03227° × 0.03227°, 0.05° × 0.05°, and 0.1° × 0.1°, respectively. However, for consistency, the interpolated day 2 and 3 outlooks were regridded bilinearly onto the 0.03227° × 0.03227° grid used for the day 1 forecasts. All subsequent verification for the interpolation approach is performed on this grid.

Being a grid with a fixed-degree increment in latitude and longitude, the physical area spanned varies with latitude. While the difference in physical distance of a fixed-degree increment in the latitudinal dimension varies negligibly as a function of latitude, the physical distance in the *longitudinal* dimension does vary appreciably over the domain. One degree of longitude is approximately 73 km at the northern border of the CONUS, while the same increment corresponds to near 100 km at the southern extremities. Consequently, the area in the northern CONUS is weighted slightly more—a factor of around 1.35 more in the extremes—than in the southern CONUS in the calculation of bulk verification statistics. However, with the CONUS broadly being confined to the midlatitudes, this effect has only a small quantitative effect and is not believed to appreciably impact any conclusions drawn from the analysis. Each method within the dynamic workflow is described in detail below.

The INTERP method first converts the native SPC shapefile polygons (e.g., Figs. 1a,d) to contours (e.g., Figs. 1b,e) that maintain the probability values of the original forecast. These contours are then used as input to the ArcGIS Topo-to-Raster function (Childs 2004), and the output is then extracted only over the spatial extent that was in union with the original SPC probabilities (visually depicted in the fill in Figs. 1c,f). Trial and error revealed that the output from the Topo-to-Raster function of the native probabilities tended to be lower than the initial contoured input: along an explicit intermediate contour, interpolated values tended to be approximately half of a probability bin lower than the value of that contour. This is mainly attributed to the interpolation problem becoming more unconstrained as fewer contours are present to be analyzed by the function. To correct for this, a second step was added to the INTERP process where another raster was created using the Topo-to-Raster function, but this time, using the SPC probabilities for the same day that were incremented up approximately one probability bin. With a half-bin negative bias in the interpolation procedure, the resulting values from the interpolation on the native probabilities were approximately half of a probability bin too low, and when adjusted upward by one probability bin, the resultant field was approximately half of a probability bin too high. The arithmetic mean of the two separate interpolated rasters (i.e., one from the original forecast probabilities and one from the probabilities incremented up one interval) therefore serves as an unbiased interpolated representation of the probability field. This mean thus serves as the output of the INTERP method (e.g., Figs. 1c,d). The only time this did not hold was when there was an increase in the contour interval with increasing probability; then, the bias tended to be smaller at approximately ⅓ of a probability interval, requiring less upward adjustment to produce an unbiased derived interpolated field. For example, in the case of a day 1 convective outlook that contained tornado probabilities of 2%, 5%, 10%, 15%, and 30% (e.g., Figs. 1b), the corresponding probabilities that were incremented up in the second Topo-to-Raster run are 3%, 10%, 15%, 20%, and 45%. The output of the INTERP method creates a raster that maintains a representative depiction of the SPC contours (cf. Figs. 1b,e to Figs. 1c,f), but interpolates in a manner that produces a smooth gradient between the contours and increases the maximum probabilities within the highest contour. There are still instances where the highest probability contour is slightly distorted compared to the original (cf. 30% in Figs. 1b,c and 15% in Figs. 1e,f); however, the differences in these regions are rarely larger than 1%.

The CONSTANT method, similar to the INTERP method, converts the native SPC shapefile polygons (e.g., Fig. 1g) to a constant raster (e.g., Fig. 1i) using the ArcGIS Feature-to-Raster tool. Since there is only one contour in the cases where the CONSTANT method is used (see above), no attempt is made to interpolate. As in the INTERP method, the CONSTANT method then outputs the probability raster onto the analysis domain at the resolution corresponding to the convective outlook lead time. The NODATA method simply outputs a constant grid of zero values over the analysis domain,

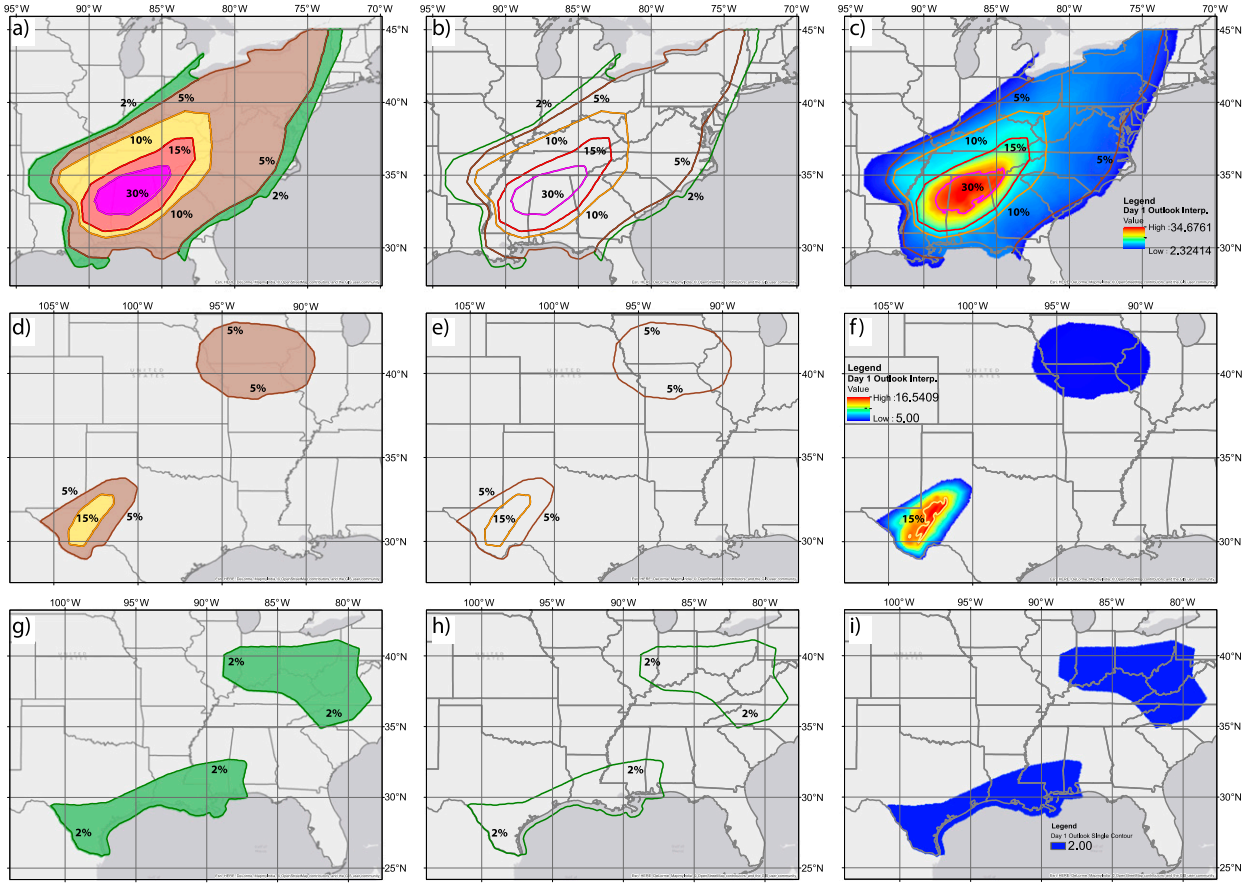


FIG. 1. Step-by-step examples of the regridding process performed on the SPC probabilities. (top) An INTERP method example for the day 1 tornado probabilities valid at 1300 UTC 27 Apr 2011. (middle) An INTERP method example for the day 1 hail probabilities valid at 1300 UTC 21 Oct 2012. (bottom) A CONSTANT method example for the day 1 tornado probabilities valid at 1300 UTC 1 May 2016. (a),(d),(g) The ArcGIS depiction of the native SPC forecast probability polygons with the colors matching the standard SPC graphic. (b),(e),(h) The contours derived from the native SPC forecast probability polygons that are used for input into the Topo-to-Raster function, where the line colors also correspond to the colors used in the standard SPC graphic. The final gridded output from the (c),(f) INTERP and (i) CONSTANT methods of the dynamic workflow, where the colored contours represent the locations of the constant probability contours in the gridded output as compared to the original input in (b), (e), and (h).

again, at the corresponding resolution for the analyzed convective outlook. The probabilities for each threat (i.e., tornado, wind, and hail) and outlook lead time (i.e., days 1, 2, and 3) are run through this dynamic workflow to create the analyzed grids that are used for the verification undertaken in this manuscript.

The traditional framework, in contrast, compares the effect the INTERP part of the dynamic workflow has on the probabilistic verification when compared with the interpolation method. In the traditional approach, all of the threat and lead time combinations are run through a simplified workflow that contains only the CONSTANT and NODATA methods and outputs to a lower-resolution grid with 80-km grid spacing.

Once all outlooks have been gridded, the same verification methods are employed in each verification framework to assess the quality of the results. These include several commonly used probabilistic verification tools. Specifically, forecast reliability, the extent to which forecasts verify at the same frequency as indicated by the forecast probability, is assessed by inspection of reliability diagrams (Murphy and Winkler 1977; Bröcker and Smith 2007; Wilks 2011). Forecast skill is quantified using Brier skill scores (BSS; Brier 1950), defined for a Brier score (BS) and using a climatological reference ($BS_{clim}$) as

$$\mathrm{BSS} = 1.0 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}} = 1.0 - \frac{\sum_c (p_c - o_c)^2}{\sum_c (p_{\mathrm{clim}_c} - o_c)^2}, \qquad (1)$$

where $p_c$ denotes the forecast probability of a case, $p_{\mathrm{clim}_c}$ denotes the climatological forecast probability for the case, and $o_c$ is a binary variable indicating whether the

forecast predictand was observed for the given case. Forecasts—defined by forecast day–latitude–longitude triplets on the prescribed analysis grid—are aggregated several different ways to ascertain various properties of SPC outlooks. In particular, they are aggregated spatially in order to determine the regional distribution of forecast skill for the different severe weather predictands, temporally both by month and year to ascertain whether there is any persistent seasonality to the forecast skill and whether forecasts are generally improving or degrading from year to year over the period of record, and meteorologically based on the conditions of the forecast point to deduce whether there is any relationship between the forecast environment and SPC severe weather forecast skill, which may also speak to the larger predictability of severe weather under different meteorological conditions.

Despite their limitations, which are discussed in more detail below (e.g., Trapp et al. 2006), verification uses storm reports as archived in SPC's Severe Weather Database (SPC 2017b), and these reports are taken to be "truth" whereby reports are taken to be certain events and nonreports are taken to be certain nonevents. All verification is performed on the CONUS-wide 0.03227° × 0.03227° grid and 80 km × 80 km grid within the interpolation and traditional frameworks, respectively. In particular, for each severe weather predictand, all points on this grid within 40 km of a severe report from the database are encoded as an observed event for the 24-h forecast day corresponding to the report; all other points in this space–time matrix are encoded as nonevents. Calculating skill scores as described above uses a climatological reference forecast; the results are calculated identically to the official SPC severe climatologies (Kay and Brooks 2000; SPC 2017c), except on a finer grid in the case of the reference within the interpolation framework. Raw verification grids are calculated as described above for 1982–2011 to match the period of the severe weather climatologies published on SPC's website at the time of this study. The 30 annual verification grids are then collectively employed to derive raw frequency grids for each day of year–latitude–longitude triplet for each of the severe weather predictands. These raw frequencies are then smoothed over time using a Gaussian filter with a 15-day standard deviation and using a "wrap" filter mode to handle treatment with respect to year beginning and year end. Finally, these steps are followed by smoothing over the two spatial dimensions with a 120-km standard deviation Gaussian filter with a "reflect" mode treatment for the domain edges; the resulting grids are said to be the climatological event probabilities (Kay and Brooks 2000).

In addition to space and time, severe weather forecasts are often viewed with respect to the prevailing environmental conditions for the forecast, and especially the CAPE-versus-shear parameter space (e.g., Schneider and Dean 2008). Verification with respect to this meteorological-regime-based parameter space can provide useful forecast insights into environments of greater and lesser forecast skill. This does however require the use of an external data source to quantify the forecast environment for each forecast. While RAP/RUC analyses are frequently employed within this sort of context (e.g., Bothwell et al. 2002; Dean et al. 2009), the transition from the RUC to RAP in May 2012 (NWS 2012) provides an undesirable potential source of inconsistency in the middle of the study period, and both received smaller changes to their data assimilation systems throughout, which may also result in changes to the analysis creation. To use a data source that is created in a consistent manner throughout the analysis period, this study employs the North American Regional Reanalysis (NARR; Mesinger et al. 2006) to determine the local meteorology at a point for a given forecast period. The NARR also has been used in analyses of severe weather environments in past studies (e.g., Gensini and Ashley 2011; Nielsen et al. 2015; Vaughan et al. 2017), and while there are some quantitative differences, errors, and regional biases—particularly in the thermodynamics (e.g., Gensini et al. 2014)—compared with other analysis and reanalysis products, the use in this study is largely to classify the general regime of the environment and not to exactly quantify the CAPE or shear at a particular point. To this extent, the NARR has been found to be qualitatively consistent with other analysis products (e.g., Vaughan et al. 2017). The NARR has 3-h temporal and 0.25° spatial resolution and includes an assortment of fields at various vertical levels from the subsurface to 100 hPa. Mean-layer convective available potential energy (MLCAPE) used for this study comes directly from the NARR and performs an averaging over the layer encompassing the lowest 180 hPa of the atmosphere. Deep-layer shear (DSHEAR), another important severe weather parameter (e.g., Doswell et al. 1993; Gallus et al. 2008; Markowski and Richardson 2010), is often expressed as the bulk wind difference between the surface (10 m) wind and 6 km above ground level. This is not available in the NARR; instead, the wind difference from the surface to 450 hPa is used for this study, and this value is rescaled based on the geopotential height at 450 hPa to approximate the value of the 0–6-km shear. All days are classified based on the maximum 3-hourly value over the 24-h 1200–1200 UTC period for each parameter.

To better ascertain the robustness of the various findings, uncertainty analysis is performed for each phase of verification. A bootstrapping procedure is

employed to generate confidence intervals for the BSS analysis. For BSSs over space and time, forecasts are resampled randomly with replacement from each analyzed grid point and time period studied—both year and month—to ascertain the uncertainty in the skill score for the given subspace. A similar method is employed for parameter space, subsampling instead for each $250\,J\,kg^{-1} \times 2.5\,m\,s^{-1}$ subregion of CAPE versus deep-layer shear parameter space. For all of this analysis, because of the small spatial scales of storms most commonly associated with severe weather and the spatiotemporally scattered nature of observed reports, points with nonoverlapping 40-km-radius neighborhoods and all forecasts on separate days are considered to be independent from one another, while forecasts on the same day with overlapping neighborhoods are, necessarily, considered to be nonindependent. Reliability uncertainty is also assessed using the methods of Agresti and Coull (1998), as described also in Wilks (2011).

## 3. Results: Traditional framework

Within the traditional verification framework, the highest BSS values for all spatial fields are generally seen in the eastern and especially central United States, with lower skill observed in the West (Fig. 2). More generally, spatially, the highest skill is often observed where the climatological event frequency is higher, as evidenced by higher climatological BSs in Fig. 2. This holds comparing across the day 1 outlooks for the individual severe phenomena as well. Severe winds (Fig. 2e), for example, have the highest BSSs over all of the CONUS at 0.093, followed by severe hail at 0.076 and tornadoes—the rarest phenomenon—coming in last of the three with a score of 0.049. The same does not hold true, however, for the ''significant'' severe phenomena. While the skill for all three phenomena (Figs. 2b,d,f) is lower at the significant criteria compared with the outlooks for the same phenomena that include events of lesser severity (Figs. 2a,c,e), significant tornado outlooks (Fig. 2b) are the most skillful in aggregate of the three significant severe outlooks with an aggregate BSS of 0.028. Significant hail (Fig. 2d) lags substantially, with an aggregate score of just 0.008, and significant wind events (Fig. 2f) have approximately no skill at all over climatology with an aggregate score of −0.001. Positive skill in significant wind events is largely confined to the Ohio River valley and middle Mississippi River valley areas, and this skill is not statistically significant. As a result of the small sample size, the negative skill in much of the West, and particularly the arid Southwest, is not found to be statistically significant either. This holds even for significant hail (Fig. 2d) events, where very negative climatology-relative skill is observed in those regions. Given the relative rarity of significant severe events, just a few events with higher (lower) predictability and large spatial coverage can drive a large degree of positive (negative) skill in a given region. The one parameter with areas of statistically significant negative skill occurs also with significant wind events, where weak but statistically significant climatology-relative skill is exhibited along a strip of the Atlantic coast from the Florida Panhandle through central New York, and secondarily in a region from the Nebraska Panhandle south through the Texas Panhandle. In the West, where larger negative BSS values are obtained, the sample size is insufficient to produce statistical significance. In no other regions is the sign of the obtained skill score, positive or negative, found to be statistically significant for significant severe forecasts. Statistically significant positive skill is seen for the regular severe convective outlooks, however. While in tornado outlooks (Fig. 2a) the statistical significance is confined to the regions such as the Tennessee River valley region and parts of the central plains, where the highest climatological event frequencies overlap with the highest skill scores, for severe hail (Fig. 2c), most of the central and southeastern CONUS excluding the immediate Gulf Coast area exhibits significantly positive skill, in addition to parts of New England. These are also, unsurprisingly, where the skill scores and event frequencies are highest for this phenomenon. With three exceptions, statistically significant positive skill is observed over all of the CONUS east of the continental divide for severe wind forecasts (Fig. 2e). The first two exceptions, in south Texas and the Florida Peninsula, have skill scores that are lower than in neighboring regions, whereas in the far northern Great Plains states, the event frequency is somewhat lower, as evidenced by smaller climatological BSs in that region. In these three areas, conditions are insufficient to garner statistical significance in the BSS. Despite different verification frameworks, these findings agree with those of Hitchens and Brooks (2017). For significant severe events, they similarly found the lowest skill among wind outlooks and the highest skill for tornadoes. Also like in this study, they found substantial skill improvement when considering outlooks pertaining to all weather exceeding the minimum severe criteria rather than the outlooks for only the significant severe events.

The same general tendencies are observed for the outlooks of all severe weather for days 2 and 3 (Figs. 2g,h) with the highest skill over the central United States and lower skill over the West. Of note, the Southeast region including the Carolinas, Georgia, and Florida has degraded skill compared to the day 1 outlooks and is largely slightly negative. Because of the considerable

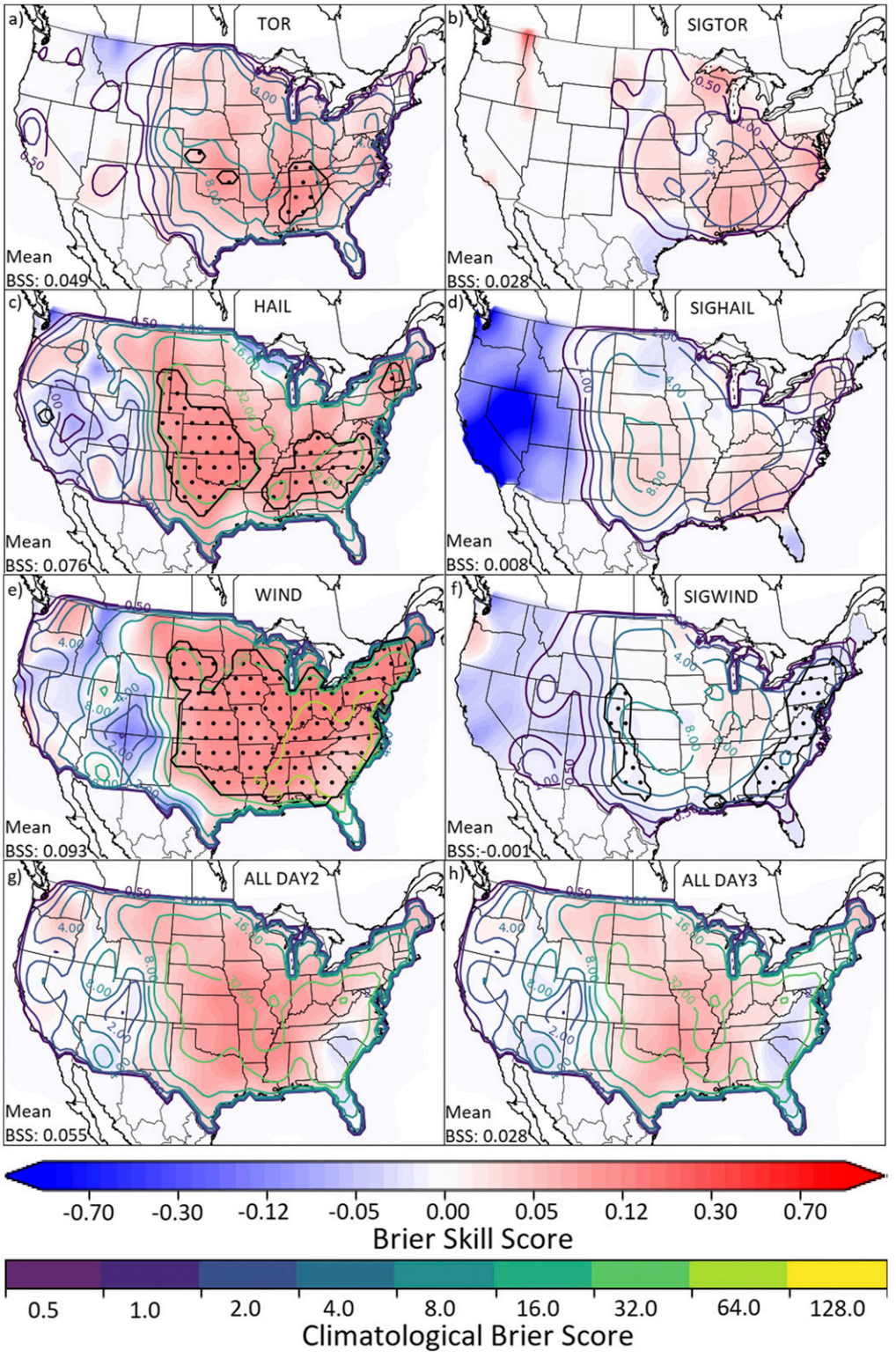


FIG. 2. BSS spatial distributions for each of the forecast sets in this study using the traditional verification framework. Results of day 1 (a) tornado probability forecasts and (b) significant tornado probabilities. Day 1 (c) hail and (d) significant hail; day 1 (e) wind and (f) significant severe wind. Any severe probabilities for days (g) 2 and (h) 3. The ''mean'' BSS (unity minus ratio of the sum of BSs at all grid points for the given variable

variability in success of the longer lead-time forecasts, none of the skill—positive or negative—was found locally to be of statistically significant sign. With domain-total skill scores of 0.055 and 0.028 for days 2 and 3, respectively, there is a clear deterioration in forecast skill with increasing forecast lead time from day 1 through day 3.

Comparing the verification results across the period of record (Fig. 3), no clear trend in skill scores is seen from the beginning of the period to the end. The verification period of this study is admittedly much shorter than that of Hitchens and Brooks (2012, 2014), when a marked improvement in forecasts was observed over the decades of SPC outlooks analyzed. In general, like with the spatial results where areas of higher event occurrence exhibited more skill, skill tends to be somewhat higher during more-active years compared with less-active ones. This is especially true for tornadoes (Fig. 3a), where the highest skill—both for all tornadoes and for just significant ones—is seen in the historically active 2011 season; this holds to a lesser extent with hail and wind as well. As a result, and given a particularly low skill year in 2009, an increase in skill is seen in the 2009–11 period, with a gradual decline in skill thereafter likely attributable to annual fluctuations in severe weather frequency, especially those associated with relatively high-predictability synoptic-scale regimes. Statistically significant positive skill is seen for all of tornadoes, hail, and wind for all years. For each phenomenon, the skill is consistently better from year to year for the regular severe criteria compared with forecasts of significant severe events. Compared with their less stringent counterparts, confidence intervals are larger for significant tornadoes and smaller for significant wind events, but with a couple of exceptions, no statistically significant skill of either sign is observed for significant severe events throughout the period of record. For day 2 and day 3 outlooks (Fig. 3d), day 2 forecasts consistently verify slightly better than day 3 forecasts from year to year despite slight fluctuations in skill overall. Confidence intervals are large compared to the day 1 forecasts for specific phenomena, largely as a result of large forecast-to-forecast variability in the successfulness of individual outlooks, in addition to an overall shorter period of record. The intervals are particularly large in 2012 because only approximately 1/3 of the year falls into the period of record, resulting in a significantly reduced sample size.

Fairly substantial amplitude seasonal cycles of forecast skill in day 1 outlooks (Figs. 4a–c) were discovered through this verification. Tornadoes, both for outlooks of any tornado and for only significant ones, exhibit two peaks: one in the spring and a particularly sharp one in the late autumn, maximizing in November. Between the two, there is a broad skill minimum throughout the summer and early autumn, consistent with prior studies (e.g., Hart and Cohen 2016). In fact, tornado outlooks suffer a skill degradation sufficiently large such that in August, tornado outlooks of both severity levels verify about equally. Confidence bounds are also much tighter in this minimum, and the seasonal differences for the EF0+ outlooks are found to be statistically significant. A somewhat similar trend is seen for hail forecasts (Fig. 4b); however the springtime maximum peaks slightly earlier, in March rather than April, and is significantly larger than either of the skill spikes in the tornado outlooks. In addition, the skill maximum in the autumn still exists but is of a much smaller magnitude and is no higher than for the winter months. While a summer local minimum of skill is observed, the primary minimum is seen in December, where appreciably negative aggregate skill is in fact observed during outlooks from that month. The monthly differences between these minima and maxima in forecast skill are found to be statistically significant. Significant hail events, in contrast, are found to have a substantially muted seasonal cycle, with a slight increase in skill during the spring found to be the primary feature. Wind events (Fig. 4c) follow a somewhat similar pattern to tornadoes, with a clear minimum in skill reaching its lowest point in August, and a maximum in skill in November, but there is no real secondary springtime peak and instead there is a gradual degradation in skill throughout the winter and spring months. This cold-season skill maximum coincides with a period whereby a higher proportion of severe weather events are from synoptically forced systems than

←

divided by the sum of the climatological BSs at all grid points for the same variable) is depicted in the bottom left of each variable's panel. A 120-km standard deviation Gaussian smoother was applied prior to plotting for (a), (c), (e), (g), and (h), while a larger 180-km smoother was applied for the significant severe variables in (b), (d), and (f) owing to the smaller sample sizes. Unfilled color contours depict the climatological BSs for the verification location; larger numbers indicate locations with more frequent events and more impactful areas toward the mean score. Note that the contour interval and color scale, shown at bottom, is nonlinear. Stippling depicts areas where the sign of the indicated skill score is statistically significant with 95% confidence using a bootstrapping procedure as described in the text. Light smoothing of the significance contours has been performed to enhance readability.

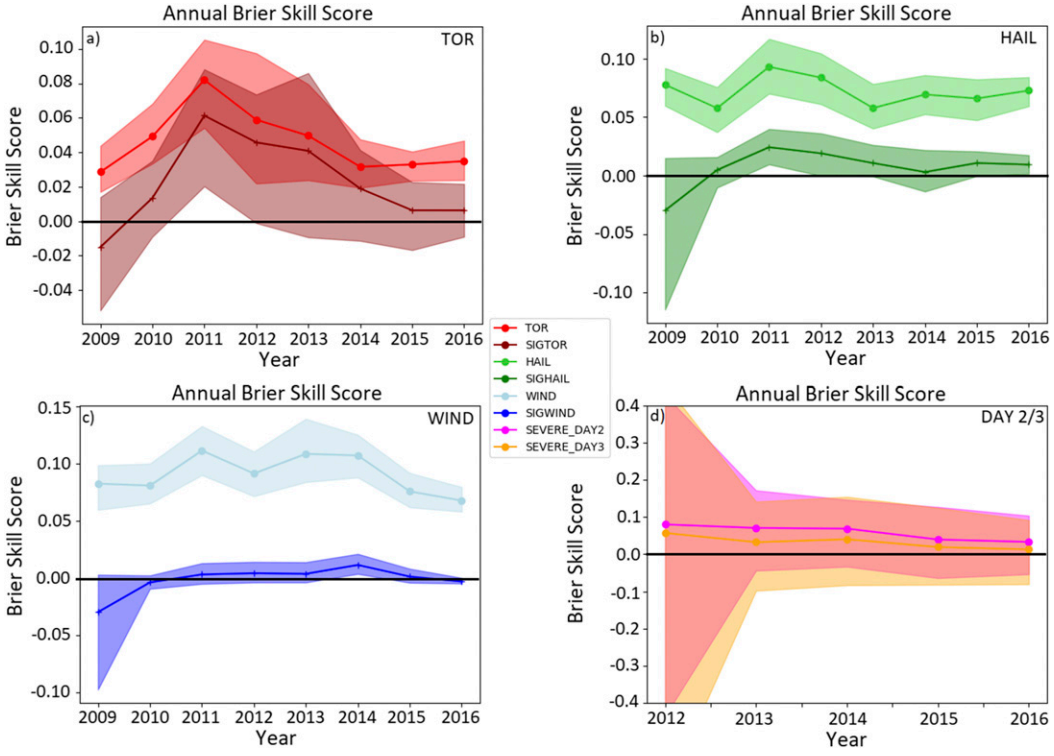


FIG. 3. BSSs for each forecast set as a function of year of forecast issuance for (a) tornadoes and significant tornadoes, (b) severe hail and significant severe hail, (c) severe wind and significant severe wind, and (d) day 2 and 3 forecasts of any severe weather using the traditional verification framework. BSs and climatological BSs have been summed over space to produce the skill scores shown in (a)–(d). Transparent shading around lines indicates 95% confidence intervals on the BSS obtained via bootstrapping as described in the text. Note that the *y* axes vary between panels.

during the warm season and, thus, likely have higher predictability, particularly at the extended range, than those with weaker or smaller-scale processes primarily responsible (e.g., Surcel et al. 2016; Nielsen and Schumacher 2016; Herman and Schumacher 2016). Significant wind events, like significant hail, feature a muted seasonal cycle with a slight maximum observed during the late autumn and early winter. Significant severe wind events also verify significantly worse than other severe wind events throughout the entire year. Finally, day 2 and 3 convective outlooks (Fig. 4d) show relatively little seasonal cycle in forecast skill, except with a slight enhancement of skill in November, as seen in many of the day 1 outlooks. Confidence intervals are generally very large, but shrink considerably in size during the warm season where the sample size is much larger. With one slight exception in October, day 2 forecasts continue to perform slightly better than day 3 forecasts from month to month; the magnitude of this difference tends to be smallest from the late summer to early autumn and largest from late winter to early spring.

Skill verification in the CAPE-versus-shear parameter space (Fig. 5) depicts positive results throughout much of the parameter space for each of the different severe phenomena forecasted in day 1 convective outlooks, as one would expect given the positive aggregate skill (Figs. 3a,c,e), with two primary regions of exception. First, low climatology-relative skill is seen in much of the parameter space with very weak deep-layer shear. For hail (Fig. 5b), this is true throughout the weak shear region of the parameter space. In contrast, for tornadoes (Fig. 5a), this is especially emphasized in the low-shear, high-CAPE region of the parameter space while for wind (Fig. 5c), the opposite is true, with the most negative scores found in the low-shear, low-CAPE environments. The second region of deflated skill is in the low-CAPE, high-shear region of the parameter space, an area frequently noted as a particularly challenging forecast region in the phase space (e.g., Evans and Doswell 2001; Davis and Parker 2014; Sherburn and Parker 2014; Sherburn et al. 2016); this degradation in skill is evident for all variables, but is particularly pronounced for tornadoes (Fig. 5a) and least apparent for wind (Fig. 5c). Despite both of these environments being quite rare (Fig. 5d), the signs of the skill score in these subregions of parameter space are found to be

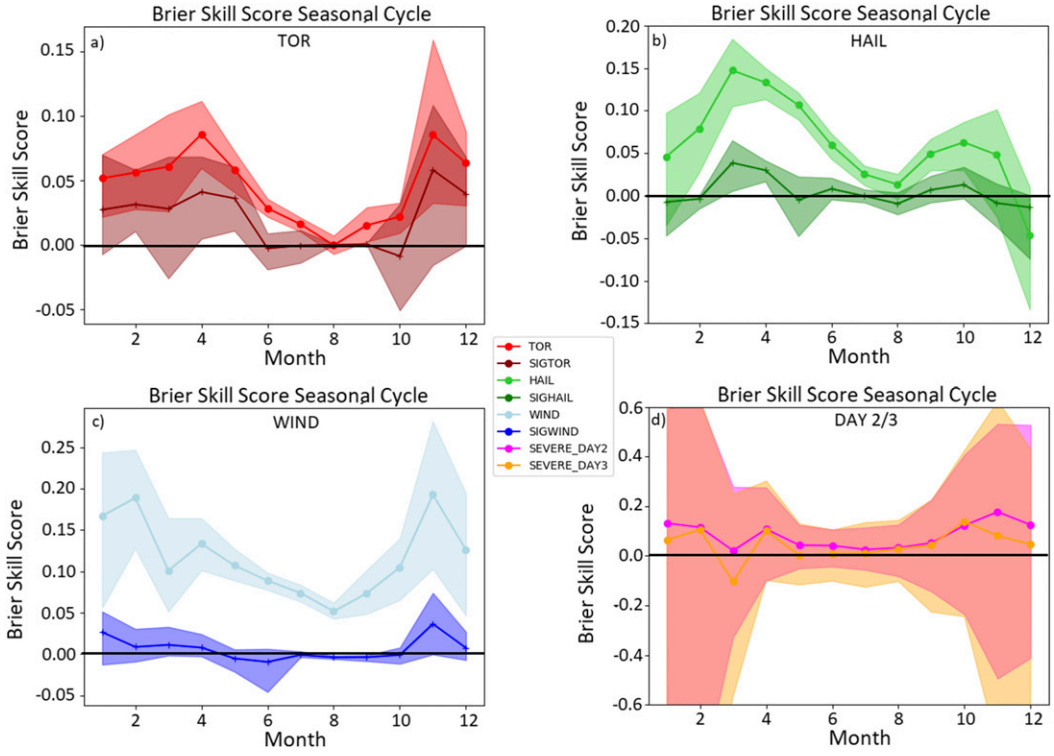


FIG. 4. As in Fig. 3, but by month of forecast issuance.

statistically significant. Much of the positive skill regions are also found to be statistically significant. For hail and wind (Figs. 5b,c), the main exception is found in portions of the high-shear, high-CAPE regions of the parameter space, where the sample size is too small (Fig. 5d) to obtain significance. For tornadoes, some of the more common lower-shear, lower-CAPE environments also fail to acquire statistical significance, in these subregions owing to deflated skill scores (Fig. 2a) compared with hail and wind. These results on outlook verification largely agree with the findings of Anderson-Frey et al. (2016) and others on tornado warning verification, who similarly found the best performance when both ingredients were highest and the worst outcomes when one ingredient or the other was lacking while the other remained relatively large. These findings are with respect to a climatological baseline, and other qualitative findings may emerge with comparison with respect to a different reference forecast.

Finally, with regard to the attributes diagrams characterizing these forecast sets (Fig. 6), one can note that the vast majority of forecasts of all variables have zero probability, with forecasts becoming increasingly rare with increasing probability. This is especially true of tornadoes, which have at least an order of magnitude fewer forecasts than other variables at probability thresholds at and above 5%. At the extreme, 60% probabilities have been issued for wind during each variable's period of record and have only been issued for approximately 1 in 100 000 forecast points. Tornado forecasts within the traditional framework appear to be quite negatively biased; observed relative frequencies are substantially higher than their corresponding forecast probabilities for probabilities at and above 5%. This is also true for day 2 and 3 convective outlooks. At forecast probabilities of 5%, the observed relative frequency is over 10% for each variable (Fig. 6b). This improves slightly at higher probabilities, but the extension of negative bias extends there as well. At the highest observed forecast probabilities during the day 2/day 3 period of record (Fig. 6a), 45% of day 3 forecasts verify with approximately that frequency, but day 2 forecasts remain statistically significantly negatively biased. Wind and hail forecasts at day 1 are reasonably well calibrated, except for wind forecasts at the 60% probability threshold, which are actually statistically significantly positively biased despite the small sample size and large associated uncertainty.

## 4. Results: Interpolation framework

The overall findings of the spatial distributions of the BSSs in the interpolation verification framework, shown in Fig. 7, are similar to those in the traditional

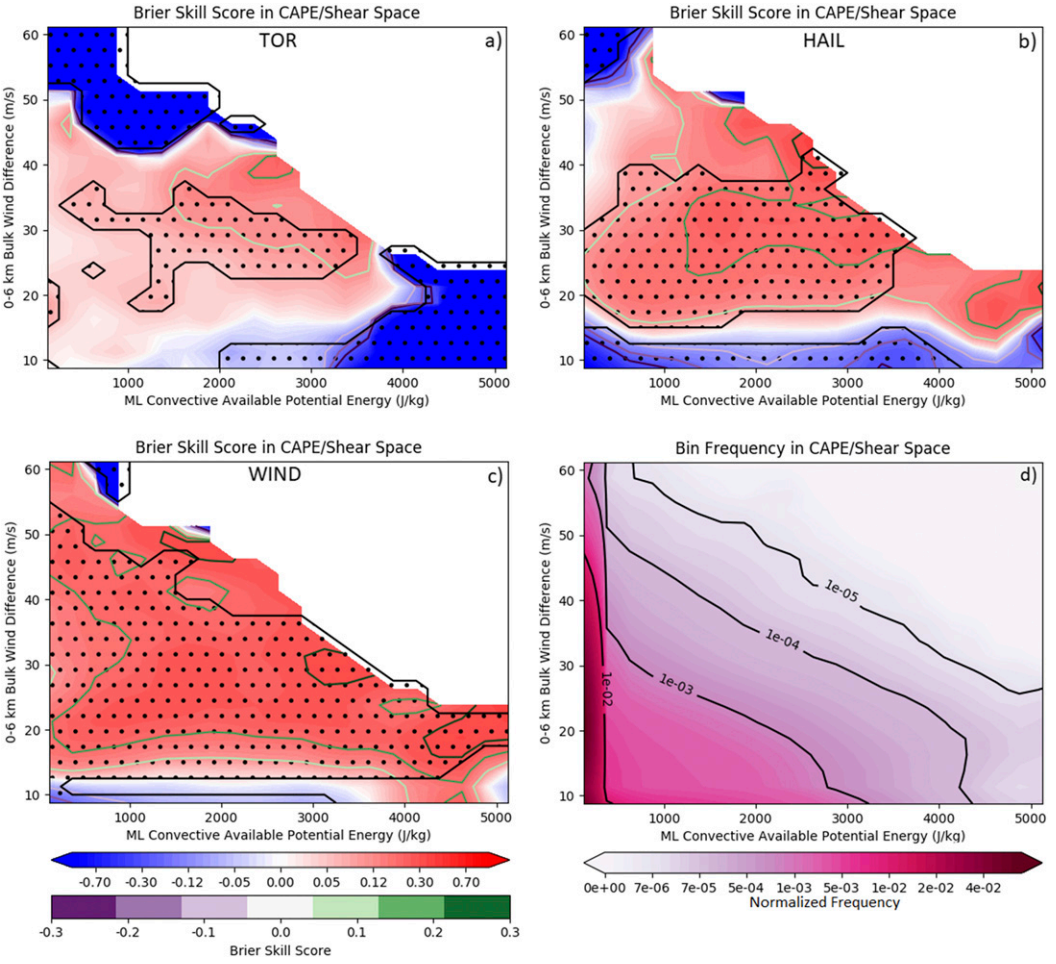


FIG. 5. BSS as a function of the prevailing MLCAPE and DSHEAR at the forecast point for day 1 (a) tornado, (b) hail, and (c) wind forecasts verified from 1 Jan 2009 to 21 Aug 2014 using the traditional verification framework. (d) The raw frequencies of points falling into each bin, separated by 250 J kg$^{-1}$ in MLCAPE space and 2.5 m s$^{-1}$ in DSHEAR space, over the verification period. Values have been lightly smoothed with a 187.5 J kg$^{-1}$, 1.875 m s$^{-1}$ Gaussian smoother for increased clarity. Stippling denotes regions of the parameter space where the sign of the indicated skill score is known with 95% confidence. Note that both the red/blue and magenta scales are nonlinear, particularly the latter one. Both the red/blue and green/purple scales depict the same BSS field, but the explicit contours in green/purple are included for quantitative clarity.

framework (Fig. 2), but there are some important and interesting differences in the details. Overall, the skill scores in aggregate are slightly higher in the interpolation framework compared to the traditional framework for multicontour forecast variables. Aggregate BSS values for tornado forecasts (Fig. 7a) are 0.059 within the interpolation context compared with 0.049 in the traditional one (Fig. 2a), 0.096 versus 0.076 for hail (cf. Figs. 7c and 2c), and 0.130 versus 0.093 for wind (cf. Figs. 7e and 2e). Similarly for the day 2 and 3 outlooks, interpolation scores are 0.066 and 0.040, respectively, compared with 0.055 and 0.028 in the traditional verification framework. All of these differences are found to be statistically significant at a 95% significance level. In contrast, for the significant severe forecasts (Figs. 7b,d,f) which use only a single 10% probability contour and where the only difference stems from the choice of the analysis grid (i.e., 80 km in traditional and 0.03227° in interpolation), the aggregate BSS differences are all 0.001 or smaller. These results strongly suggest that the act of smoothly interpolating between drawn probability isopleths has merit and results in superiorly verifying forecasts, with expected skill improvement on the order of 10%–40% in the case of severe winds.

With regard to the effects on particular regions, the overall results are fairly similar, with higher scores over the central and eastern United States and lower scores in the West, but there are some notable differences. In

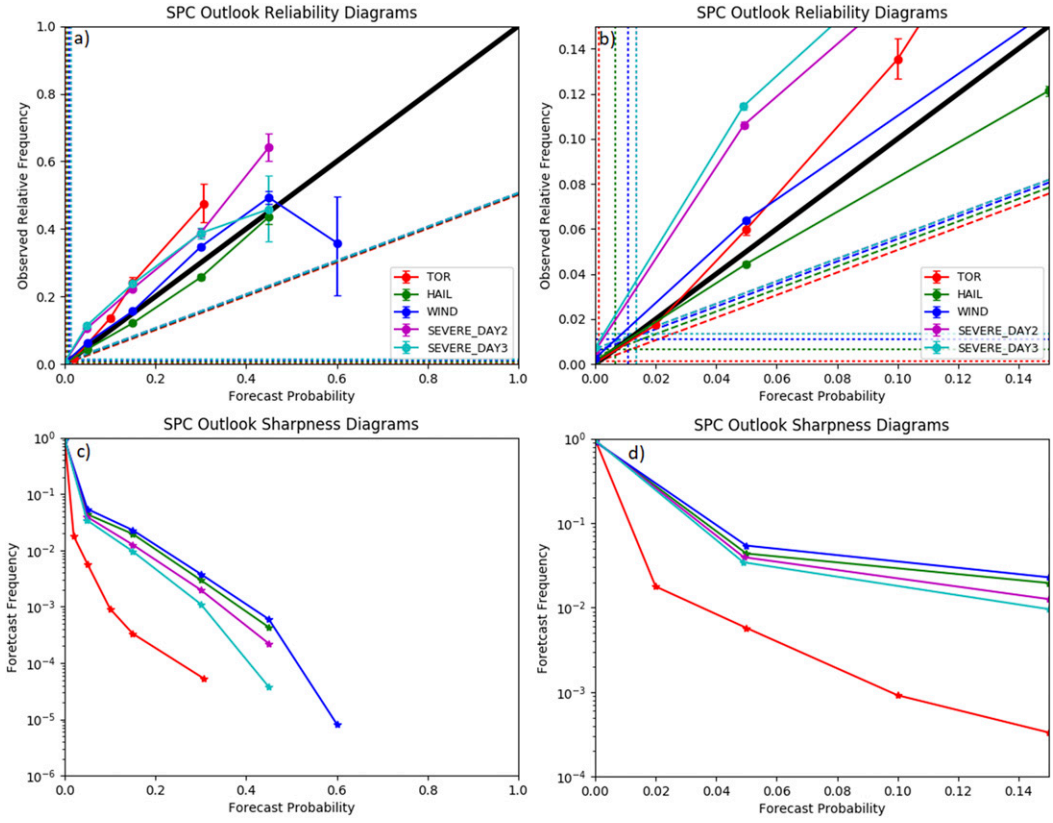


FIG. 6. Reliability and sharpness diagrams using the traditional verification framework. (a),(b) Colored lines with circular points indicate observed relative frequency as a function of forecast probability; the solid black line is the one-to-one line, indicating perfect reliability. Colors correspond to forecast sets of different parameters and lead times as indicated in the legend. (a) The entire reliability diagram, and (b) a zoomed-in image of (a), restricted to only probabilities of 0.15 or lower. Probability bins correspond to the full range of discrete probabilities that SPC can issue for the given forecast variable. Horizontal and vertical dotted lines denote the "no resolution" lines and correspond to the bulk climatological frequency of the given predictand. The tilted dashed lines depict the "no skill" line following the decomposition of the BS. Error bars correspond to 95% reliability confidence intervals using the method of Agresti and Coull (1998), where nonoverlapping neighborhoods are assumed to be independent. (c),(d) Sharpness curves, whereby lines indicate the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale shown along the *y* axis and using the same color encoding employed in (a) and (b). The *x* axes of (c) and (d) correspond to those of (a) and (b), respectively.

particular, there is a tendency for forecasts to degrade across the South and improve across the North, and this effect is especially pronounced in the severe hail and wind outlooks (Figs. 7c,e). South Texas is especially negatively impacted in the severe hail and wind outlooks, while the Atlantic Southeast including the Carolinas is especially negatively impacted for longer lead-time outlooks (Figs. 7g,h). Interpolation can make an especially large difference for hail and wind, as 5% and 15% contours are comparatively frequent, and between these contours, the probability ratios between frameworks of two to three are common. This effect appears to a lesser extent between higher-probability isopleths. If the outlooks exhibit an underforecast bias in the North and an overforecast bias in the South within the traditional framework, perhaps as a result of terrain and coastal effects inhibiting predictability over southern CONUS, the anticipated framework differences in forecast skill would be consistent with what is observed here. This effect affects statistical significance as well, with worse forecasts and less coverage of statistical significance of positive skill over the central and southern Great Plains and with more coverage of significantly skillful forecasts in the northern Great Plains and northern Rockies. For other variables, the spatial patterns of skill and statistical significance are more or less the same as the traditional approach.

The annual time series of forecast skill in the interpolation framework (Fig. 8) exhibit generally similar trends to the traditional framework verification results.

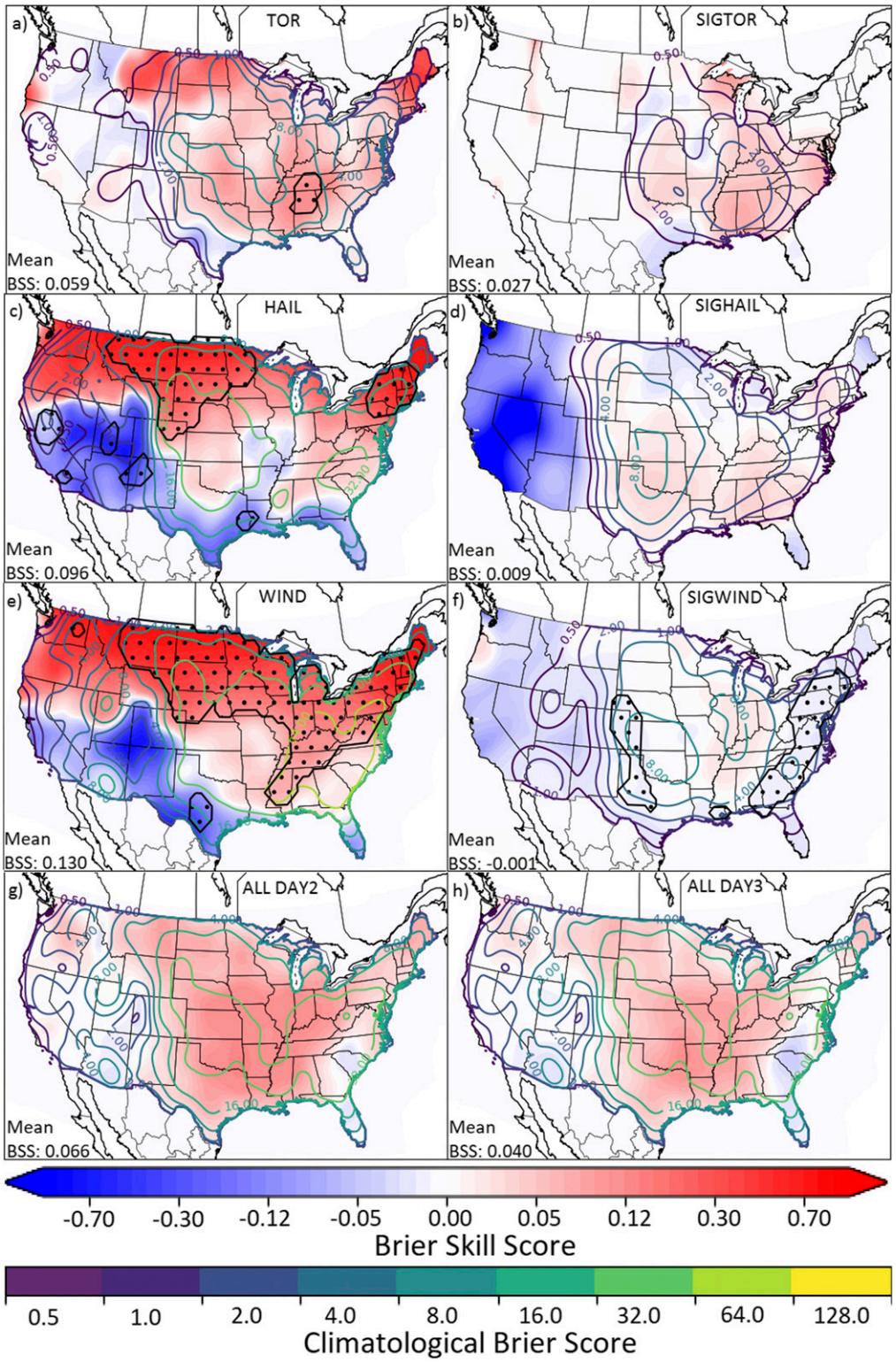


FIG. 7. As in Fig. 2, but for the interpolation verification framework.

The climatological BSs by year (Fig. 8e) further validate that the higher skill scores tend to occur during more-active years, which have correspondingly higher climatological BSs. By far the most active year of the verification period, 2011 (Fig. 8e), also featured the most skillful forecasts, consistent with the spatial findings in Fig. 7 and in accordance with what one would typically expect when forecasting very rare events (e.g., Baldwin and Kain 2006; Stephenson et al. 2008; Wilks 2011). Severe hail events were more common in 2009 owing to the lower threshold definition valid at that time, and skill was also correspondingly somewhat higher during that year compared to most other years in the period. The seasonal cycles of skill (Fig. 9), in contrast, portray both some similarities and some notable differences when compared to the traditional approach results. Tornado forecasts exhibit almost the exact same seasonal cycle of skill in both analysis frameworks (cf. Figs. 4a and 9a), with a broad spring peak, a sharp peak in late autumn, and a skill minimum in the late summer and early autumn. However, notable differences appear in the severe hail forecasts (Fig. 9b). While the large pattern is generally the same as seen in the traditional approach, the late autumn peak is more muted and, more importantly, there is a substantial performance spike in July and to a lesser extent in surrounding months that is entirely absent from the traditional results. This same spike also appears in the severe wind forecasts (Fig. 9c) and is absent from those traditional verification results as well. Like with tornadoes, day 2 and 3 forecasts (Fig. 9d) exhibit very similar skill levels for seasonal cycles within both frameworks. Unlike with years and space, there is not in general a correspondence by month between event frequency, depicted most explicitly with the climatological BSs in Fig. 9e, and forecast skill. Tornadoes have a primary peak in the spring and a coincident maximum in skill during that time, but the variables maximize in frequency in the late spring and early summer, and skill is largely quite low there except for the July skill spike in the interpolation framework. Severe weather environments often feature fewer higher-predictability, synoptic-scale forcing scenarios such as strong fronts during this period, and this may be at least partly responsible for this apparent discrepancy (e.g., Hart and Cohen 2016).

The verification results in the CAPE-versus-shear parameter space are also largely similar within the interpolation framework (Fig. 10) compared with the traditional framework (Fig. 5). The primary difference, in addition to more regions of statistical significance, is in the improvement in skill in the low-CAPE, high-shear region of the parameter space. The negative skill region is much smaller for tornado outlooks (Fig. 10a) and completely vanishes for the wind outlooks (Fig. 10c). Improvement is also seen, although the sign of the skill remains the same, across much of the moderate-to-high shear and moderate-to-high CAPE regions of the parameter space. The region of negative skill in the very-low-shear ($<10\,\mathrm{m\,s^{-1}}$) region of the parameter space remains, however, and perhaps even amplifies to an extent.

Perhaps the biggest difference between the verification regimes emerges in the analysis of forecast reliability. First and most obviously, interpolation acts to distribute the probability between different explicit probability contours and create a continuous probability field rather than a stepwise discrete one. This results in more probability bins for the attributes diagrams in the interpolation framework (Fig. 11). But perhaps more significantly, the ''redistribution'' of probability is not symmetric per se in that the total probability of the forecast is not conserved. The drawn contours define isopleths of constant forecast probability of value equal to the contour label. Within the traditional framework, all points within a given contour have an associated forecast probability in accordance with that contour label until a new interior contour is drawn. In the interpolation framework, however, the probability values are at least that large and may be larger—up to the value of the next explicit contour level, whether that contour was drawn or not. This essentially acts to strictly increase (or maintain) probabilities compared with the traditional framework and never to decrease them. This has substantial implications on the reliability of the forecast sets, which are reflected in the attributes diagrams by rotating all reliability lines clockwise. For negatively biased forecasts in the traditional framework (Fig. 6) such as tornado day 1 outlooks and the day 2 and 3 outlooks, this acts to better calibrate the probabilities by bringing them closer to the one-to-one line, even though a slight negative bias is still evident at the higher probabilities. Tornadoes, however, become slightly positively biased at lower probabilities, and as a result tornado outlooks may be characterized as underconfident. The day 1 hail and wind outlooks, in contrast, which were better calibrated in the traditional framework, are now positively biased except at the highest probabilities where a drastic increase in observed relative frequency with increasing forecast probability occurs above probabilities of approximately 0.5. For probabilities between 0.05 and 0.3, both hail and wind forecasts fall along or near the no-skill line. One does see (Fig. 11a) that within explicit probability contours, observed relative frequency does tend to increase for points closer to higher-numbered contours and in the center of closed contours where interpolated probabilities are

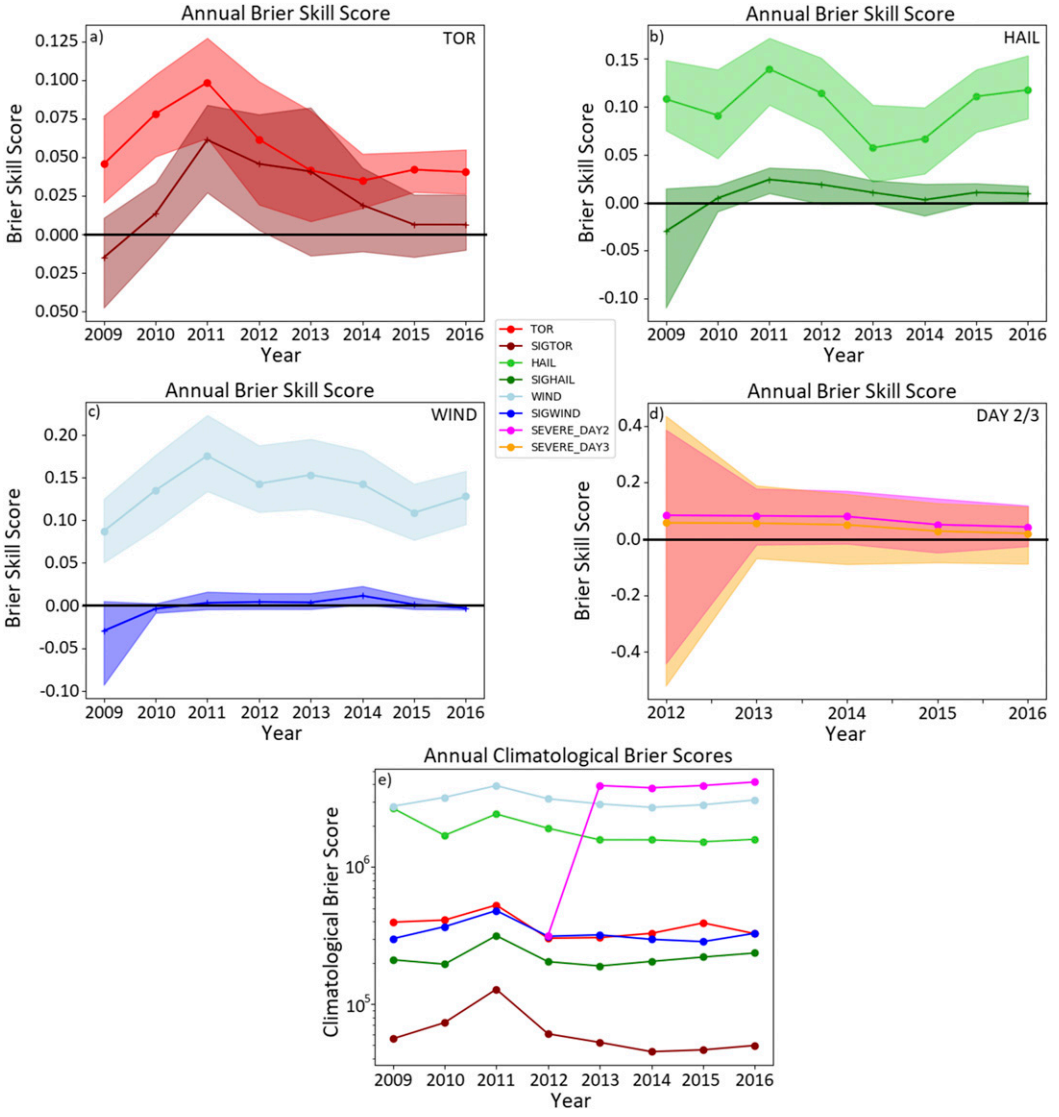


FIG. 8. As in Fig. 3, but using the interpolation verification framework. Additionally, the corresponding climatological BSs in (a)–(d) appear in (e) on a logarithmic axis using the same color coding as indicated in the legend.

higher than near contour edges, though some exceptions can be seen, particularly at lower forecast probabilities (Fig. 11b). Overall, these reliability findings in both the traditional and interpolation frameworks do contrast some with those of Hitchens and Brooks (2017), who noted positive frequency biases for essentially all forecast sets, but this discrepancy is likely attributable to differences in how that study treated the probability contours as categorical predictions as opposed to the more probabilistic treatment employed here.

Explicitly comparing the skill in the two frameworks as a function of time (Fig. 12), one sees that the interpolations scores are consistently higher than the traditional scores from year to year (Fig. 12a) with the one minor exception for the tornado forecasts in 2013. Both the magnitude of the differences within years and the uncertainty in the difference is largest for wind, with the smallest uncertainty in the difference for day 2 and 3 outlooks. Statistically significantly positive differences for all variables occur for all forecast sets in at least one year, and no significant negative difference occurs for any variable in any year. In the seasonal cycle comparison (Fig. 12b), the interpolation adds very substantial and significant skill during the summer months for the hail and wind outlooks, maximizing with BSS enhancements of over 0.1 in July. During the rest of the year, however, there is little difference, and in the case of hail, there is even a slight decrease in performance using

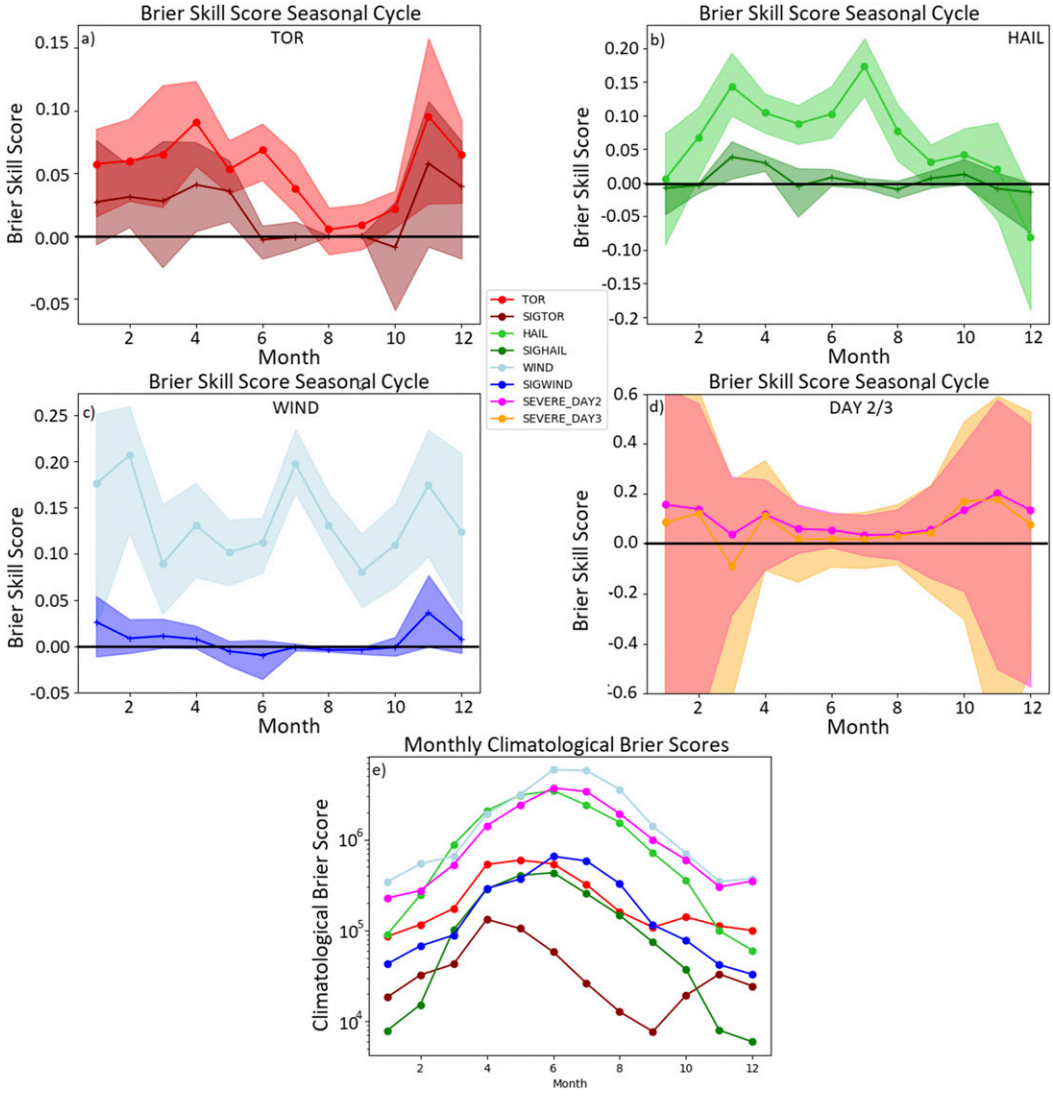


FIG. 9. As in Fig. 8, but by month of forecast issuance.

interpolated forecasts. The other variables have much less dependence on forecast month in the skill difference, but generally have the largest improvement in forecasts in the late autumn and winter months, particularly November, with tornado forecasts also having a significant peak in skill difference in June.

## 5. Discussion and conclusions

Up to eight years' worth of probabilistic SPC convective outlooks for days 1–3 were gridded onto CONUS-wide grids and evaluated using two different analysis frameworks. The first, the so-called traditional framework, uses a grid with 80-km grid spacing and does not interpolate between drawn probability contours, representing the forecast probability fields as stepwise discrete with discontinuities along the contour edges. This is performed to match the historical internal verification practice at SPC and allow direct comparison with past findings. A second approach, the so-called interpolation framework, uses a higher-resolution grid with 0.03227° spacing and instead interpolates between probability contours when two or more contour levels are depicted. Below the lowest allowable contour level for the forecast variable or when only one contour level is drawn, no interpolation is performed. The analysis period spans from January 2009 through December 2016 for day 1 forecasts and begins in September 2012 with the same end date for day 2 and 3 outlooks. The gridded outlooks were then verified using BSSs and reliability diagrams.

In general, skill verification was best when and where events were most common and when forecast lead time

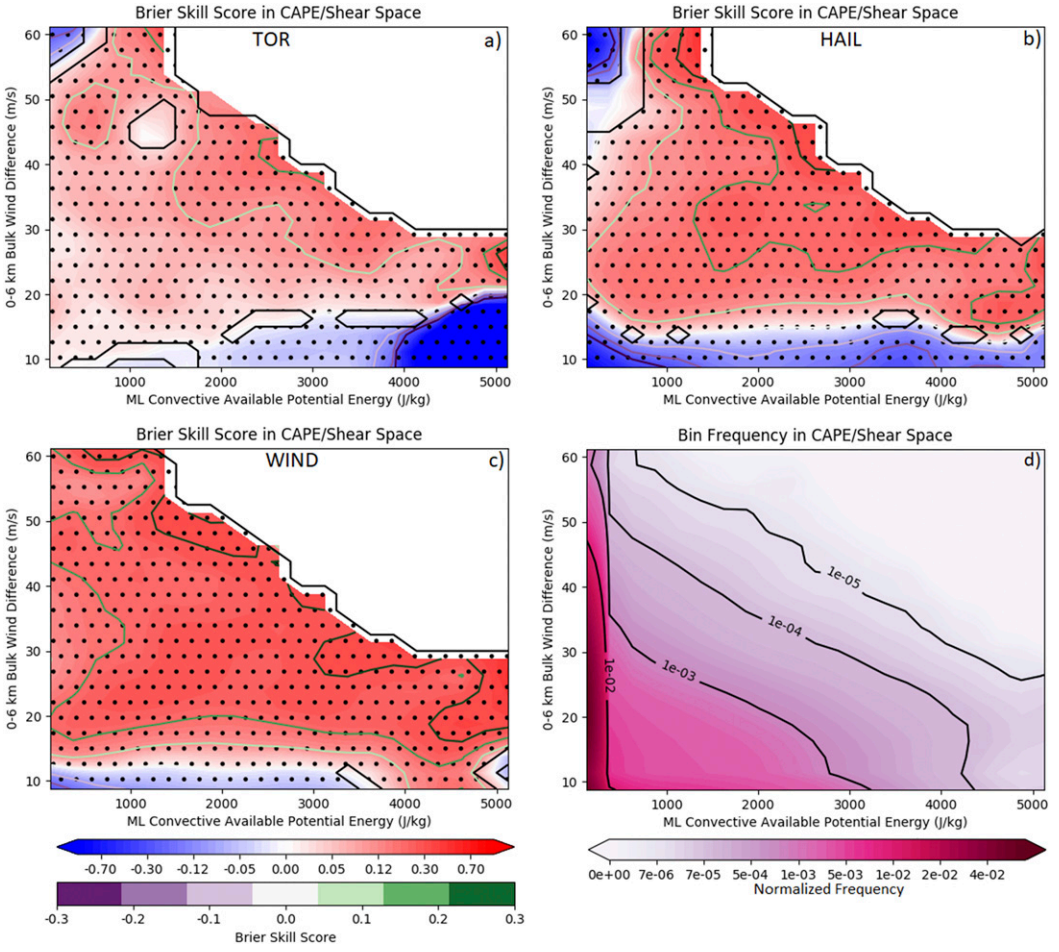


FIG. 10. As in Fig. 5, but for the interpolation verification framework.

was shortest. Among day 1 forecasts, severe winds are the most skillfully forecast by SPC in both the traditional (BSS = 0.093) and interpolation (BSS = 0.130) frameworks, followed by severe hail (BSS = 0.076 for traditional; 0.096 for interpolation) and then tornadoes (BSS = 0.049 for traditional; 0.059 for interpolation). The opposite trend, however, was observed at the significant severe threshold, with significant tornadoes (BSS = 0.027) being the best forecast, and significant severe winds (BSS = 0.00) being the worst. Forecasts were generally best in the northern and eastern parts of the country and worst in the southern and western parts of the country. Little trend was seen in the skill of SPC outlooks over the analysis period, with the most skillful forecast years coinciding with the years of highest event totals. Considerable month-to-month variability was found both between adjacent months and between variables in SPC outlook forecast skill; the highest skill was typically found in the spring and late autumn. For days 2 and 3, forecast skill was also high during winter. Forecasts were also evaluated in the CAPE-versus-shear parameter space using classifications derived from the NARR; skillful forecasts were found over the vast majority of the parameter space. Exceptionally, forecasts in the entire very-low-shear end of the parameter space were found generally not to be skillful relative to climatology for all severe weather elements. Additionally, the very-high-shear, very-low-CAPE region of the parameter space was found to be a secondary environment of forecast struggles, particularly for tornadoes and hail. In aggregate, the interpolation framework forecasts consistently yielded higher forecast skill than analogous sets in the traditional framework, suggesting that the intuitive practice of interpolating between the finite number of allowable contours yields superior forecast probabilities compared with simply using the probability associated with the nearest contour enclosing the point.

Forecasts were further analyzed to ascertain reliability; results are contrasted based on the verification framework. For traditional forecasts, day 1 tornado outlooks in addition to day 2 and 3 forecasts exhibited an underforecast bias while day 1 hail and wind outlooks

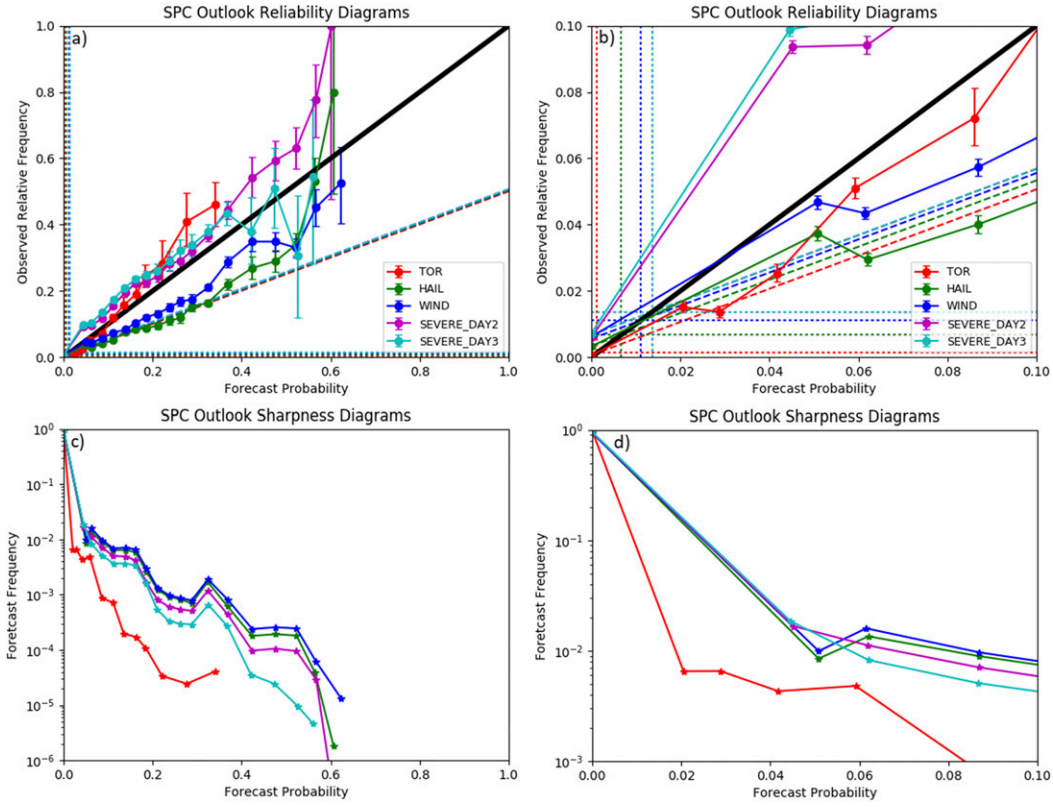


FIG. 11. As in Fig. 6, but for the interpolation framework. The zoomed-in images in (b) and (d) are to 0.1 rather than 0.15. Probability bins are delineated by 2%, 3.5%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, and 30% thresholds for day 1 tornado forecasts, and by 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 22.5%, 25%, 27.5%, 30%, 35%, 40%, 45%, 50%, 55%, and 60% for all other forecast sets.

were relatively well calibrated along the spectrum. Within the interpolation framework, in contrast, hail and wind forecasts have a moderate-to-strong overforecast bias, while day 2 and 3 forecasts have a mild underforecast bias that is alleviated compared with their traditional counterparts; tornadoes were found to have what could be considered a mild underconfidence bias, but again to a lesser extent than within the traditional framework.

There are several limitations or shortcomings of this work that should be noted. SPC outlooks, while being treated as probability grids for the purpose of this study and having grids generated internally at SPC in a similar manner, are not publicly disseminated or archived in grid format; instead, what is publicly available are equivalent to finite sets of probability contour outlines. This makes the quantitative SPC outlooks somewhat unconstrained everywhere not immediately on or directly adjacent to a drawn probability contour. Two sets of assumptions were made to convert these fixed contours to gridded probabilities, but one could certainly argue that—at least under some circumstances—the methodologies employed in this study would produce a grid from the contours that is appreciably different from what the human forecaster would have made had they produced a grid directly. Particularly consequential is the fact that no interpolation could be performed outside the lowest-probability contour because of the unconstrained nature of the problem, resulting in event probabilities being uniformly zero outside of the SPC contours. In tandem with the fact that the lowest-probability contours are generally many times larger than the climatological event frequency, this inherently inhibits SPC from gaining resolution at the lower end of the probability spectrum near the climatological frequencies. The climatological reference, in contrast, has considerable resolution in this subdomain of the probability space, and at times this can result in an uneven comparison between the SPC outlooks and the climatological reference. This effect is particularly pronounced for the significant severe forecasts, since the climatological frequencies are so low, and the forecast process effectively constrains forecasters to issue forecast probabilities of either 0 or 0.1 for any given point.

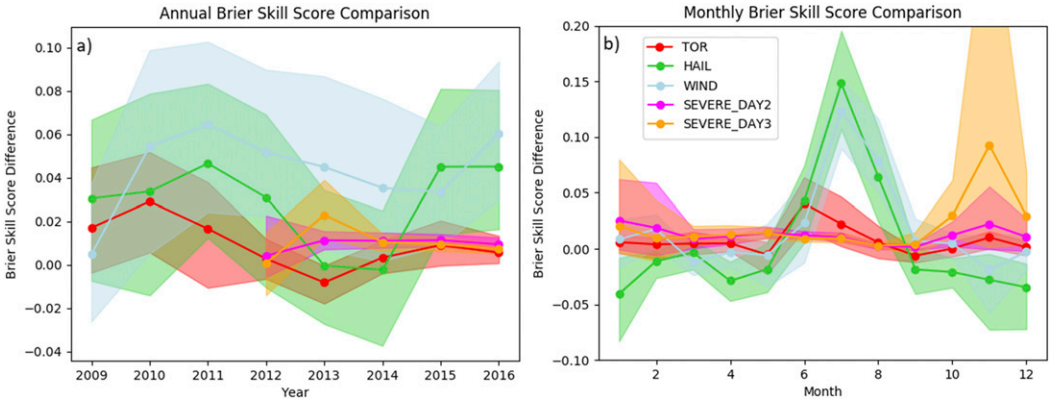


FIG. 12. Difference of verification results from the interpolation framework minus results from the traditional framework as a function of (a) forecast year and (b) forecast month for each forecast variable as indicated in the legend. Transparent shading corresponds to 95% confidence bounds on the difference obtained through bootstrapping; results are explained in greater depth in the text.

This is also seen, albeit to a lesser degree, in the verification of other fields. In particular, while forecasters draw 2% and 10% probability contours for tornado forecasts, these contours are not drawn for the remaining multicontour probabilistic forecasts, which begin at 5% and skip to 15%. Tornado forecasts therefore have some enhanced native precision; this is apparent in Fig. 12, where the probability interpolation is able to add substantially more to the forecast skill in variables other than tornado forecasts owing to their increased comparative contour granularity. These effects all work in the mean to harm the verification of SPC outlooks relative to what they would likely be if a forecaster were to adopt the operationally impractical approach of issuing continuous, subjective forecast probabilities on a point-by-point basis. More contours, particularly at the lower ends of the probability spectrum, would allow a more quantitative interpretation of the forecast probabilities by end users and would also result in more representative probability grids for verification. One way this could plausibly be addressed is by using the general thunderstorm contour from the categorical version of the convective outlooks as a 0% probability contour and interpolating between that and the lowest-probability contour. Given that the general thunderstorm contour encompasses regions where non-thunderstorm-induced severe weather is considered possible, areas outside the general thunderstorm contour can be reasonably considered to have forecast a 0% severe probability, and so this is a reasonable attempt to gain resolution at the low-probability end of the forecast spectrum. However, since this study is focused on the verification of the probabilistic convective outlooks and this contour is not included in those outlooks, applying this reinterpretation of the categorical thunderstorm line and merging it with the probabilistic convective outlooks is beyond the scope of the present study. It is, however, a worthwhile endeavor to explore in future work ways to address this important limitation within the confines of existing practices.

While several years' worth of convective outlooks have been used in this study in an attempt to obtain robust verification results, one must still recognize that severe weather is a rare phenomenon, particularly during certain times of the year and for particular regions of the CONUS. Consequently, despite this large temporal sample, the event sample, especially in certain subclasses, is still rather small and some caution should be exercised in generating conclusions from the findings. Formal uncertainty analysis and significance testing has been performed to attempt to ascertain a realistic range of true possibilities given the data sample analyzed and ascertaining which of the conclusions may be robustly made. This revealed, for example, that low skill scores in the West may be just an artifact of the sample owing to the small size and large variability, while comparably smaller magnitude scores to the east are significant owing to the higher climatological event frequency. An important and related, but distinct, point concerns the pitfalls of skill calculation for a phenomenon with varying climatological frequency (Hamill and Juras 2006). As a result, added care must be exercised when comparing skill scores across variables, regions, or times where the frequency of occurrence may vary substantially. Concerns are lessened when comparing across verification frameworks or between day 2 and day 3 outlooks when the references are identical.

Furthermore, all of this analysis uses SPC storm reports as truth. This is a sensible choice given its continuous coverage; it is generally recognized as the best

dataset for severe weather reports except for on a case-by-case basis when more thorough analysis has been conducted (e.g., Hitchens and Brooks 2012, 2014, 2017). However, this dataset has numerous limitations. Human reports, of course, require physical observations of either the phenomenon or the lasting damage it produces. Events can occur and go unreported in rural areas where few or no people are impacted (e.g., Anderson et al. 2007). Nocturnal events, and events in heavily forested areas or areas of complex terrain, also pose reporting challenges, particularly for tornadoes, because of the difficulties involved with making visual observations of an event (e.g., Anderson et al. 2007). For a multitude of reasons, including but not limited to the increasing population density (e.g., Verbout et al. 2006), increases in radar coverage (e.g., Agee and Childs 2014), and improved spotter networks and reporting practices (e.g., Trapp et al. 2006; Doswell 2007), there have also been numerous changes over time in reporting frequency and density. Fortunately, from a climatological perspective, the period of record employed by this paper is rather short, and most of these report trend considerations are not significant concerns. The unreliability and inconsistency in EF0 tornado reports (e.g., Anderson et al. 2007), the change in severe hail criteria (Ferree 2009), and particularly reporting issues associated with severe convective winds (e.g., Trapp et al. 2006; Edwards and Carbin 2016) all present additional concerns that can compromise the reliability of the database and adversely impact the validity of the verification results such as those presented herein. An additional, but related, limitation concerns the actual treatment of the reports in this study. Here, reports have been used to form binary grids of event observance, which does not account for the density of reports within a verification grid box like a practically perfect approach (e.g., Hitchens and Brooks 2014) would. However, given the high-resolution nature of the verification grid, this is not considered to be a substantial concern in the end results.

Despite these limitations, this analysis can provide utility in a variety of ways. It can help end users determine under which situations SPC outlooks exhibit more or less skill to the extent this may assist with uncertainty assessments and decision-making. As a gold standard of severe weather forecasting, these results can also help direct both operational forecasters and researchers into which particular areas could use further attention, both in the forecast process and in modeling and physical understanding. The results, and particularly the reliability findings, may invoke changes in forecasting philosophy that are of benefit to end users (Hitchens and Brooks 2012). For example, the reliability results suggest that, at least under some circumstances, SPC forecasters may benefit from increased conservatism with their day 1 hail and wind contours and more liberal usage of probability contours for their day 2 and 3 forecasts. This analysis also provides robust, quantitative benchmarks for the comparison of newly developed severe weather forecast guidance. In isolation, a skill score—other than 0 or 1—does not have much quantitative physical meaning. A positive value less than one indicates nonperfect forecasts that nevertheless have skill over the reference, of course, but the interpretation of a specific number—0.2, for example—depends both on the quality of the reference forecast and the feasibility of perfect forecasts. Having benchmarks against a robust, respected standard such as the SPC outlooks is particularly important in the severe weather forecast problem since, unlike other forecast predictands, operational models are not able to simulate or forecast most severe weather phenomena directly, further reducing the possibility of comparing new methods with existing guidance and contextualizing the results. There have been numerous forays into improving aspects of severe weather forecast guidance in recent years, some of which are already in operational use (e.g., Brimelow et al. 2006; Sobash et al. 2011), as well as other more recent work that is under development (e.g., McGovern et al. 2014; Sobash et al. 2016a,b; McGovern et al. 2017; Gagne et al. 2017); having these results will help better place those results and future similar studies within the context of existing forecasts.

Future verification work will seek to perform similar analyses for flash flooding using excessive rainfall outlooks issued by the Weather Prediction Center and explore other issues in flood and flash flood verification (e.g., Drobot and Parker 2007; Gourley et al. 2013; Schroeder et al. 2016; Herman and Schumacher 2016, among others). From there, we will also explore the overlaps and intersections of probabilistic forecasts for different weather hazards to glean additional operational insight into forecasting performance and challenges in predicting these elevated threat scenarios, such as concurrent and collocated tornado and flash flood hazards (Nielsen et al. 2015). Other work will seek to provide improved gridded probabilistic forecast guidance for these high-impact weather hazards to help yield improvements in the future verification of these operational forecasts.

*Acknowledgments.* The authors wish to thank SPC for making available their historical forecasts, which allows this sort of research to be conducted. We also wish to thank Daryl Herzmann and Matthew Taraldsen for helpful insights that helped to improve the output of the regridding process. Matthew Bunkers, Roger Edwards,

and two anonymous reviewers provided comments and questions that greatly improved the final product of this study. Funding for this research was supported by a National Science Foundation Graduate Research Fellowship Grant DGE-1321845, Amendment 3, NSF SSI Grant OAC-1450089, and NOAA Grant NA16OAR4590215. Funding was also provided from the C3LOUD-Ex Field Program, supported by the Professor Susan C. van den Heever Monfort Professorship 53633 and NSF Grant AGS-1409686.

## REFERENCES

Agee, E., and S. Childs, 2014: Adjustments in tornado counts, F-scale intensity, and path width for assessing significant tornado destruction. *J. Appl. Meteor. Climatol.*, **53**, 1494–1505, https://doi.org/10.1175/JAMC-D-13-0235.1.

Agresti, A., and B. A. Coull, 1998: Approximate is better than ''exact'' for interval estimation of binomial proportions. *Amer. Stat.*, **52**, 119–126, https://doi.org/10.1080/00031305.1998.10480550.

Anderson, C. J., C. K. Wikle, Q. Zhou, and J. A. Royle, 2007: Population influences on tornado reports in the United States. *Wea. Forecasting*, **22**, 571–579, https://doi.org/10.1175/WAF997.1.

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. 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.

Anthony, R. W., and P. W. Leftwich Jr., 1992: Trends in severe local storm watch verification at the National Severe Storms Forecast Center. *Wea. Forecasting*, **7**, 613–622, https://doi.org/10.1175/1520-0434(1992)007<0613:TISLSW>2.0.CO;2.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

Barnes, L. R., E. C. Gruntfest, M. H. Hayden, D. M. Schultz, and C. Benight, 2007: False alarms and close calls: A conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140–1147, https://doi.org/10.1175/WAF1031.1.

Bieringer, P., and P. S. Ray, 1996: A comparison of tornado warning lead times with and without NEXRAD Doppler radar. *Wea. Forecasting*, **11**, 47–52, https://doi.org/10.1175/1520-0434(1996)011<0047:ACOTWL>2.0.CO;2.

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/19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., JP3.1, https://ams.confex.com/ams/pdfpapers/47482.pdf.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brimelow, J. C., G. W. Reuter, R. Goodson, and T. W. Krauss, 2006: Spatial forecasts of maximum hail size using prognostic model soundings and HAILCAST. *Wea. Forecasting*, **21**, 206–219, https://doi.org/10.1175/WAF915.1.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Brotzge, J., S. Erickson, 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.

Childs, C., 2004: Interpolating surfaces in ArcGIS spatial analyst. ArcUser, ESRI, Redlands, CA, http://www.esri.com/news/arcuser/0704/files/interpolating.pdf.

Davis, J. M., and M. D. Parker, 2014: Radar climatology of tornadic and nontornadic vortices in high-shear, low-CAPE environments in the mid-Atlantic and southeastern United States. *Wea. Forecasting*, **29**, 828–853, https://doi.org/10.1175/WAF-D-13-00127.1.

Dean, A. R., R. S. Schneider, R. L. Thompson, J. Hart, and P. D. Bothwell, 2009: The conditional risk of severe convection estimated from archived NWS/Storm Prediction Center mesoscale objective analyses: Potential uses in support of forecast operations and verification. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 6A.5, https://ams.confex.com/ams/pdfpapers/154304.pdf.

Doswell, C. A., III, 2007: Small sample size and data quality issues illustrated using tornado occurrence data. *Electron. J. Severe Storms Meteor.*, **2** (5), http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/26/27.

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

——, 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.

Drobot, S., and D. J. Parker, 2007: Advances and challenges in flash flood warnings. *Environ. Hazards*, **7**, 173–178, https://doi.org/10.1016/j.envhaz.2007.09.001.

Edwards, R., and G. W. Carbin, 2016: Estimated convective winds: Reliability and effects on severe-storm climatology. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/28SLS/webprogram/Paper300279.html.

——, ——, and S. F. Corfidi, 2015: Overview of the Storm Prediction Center. *13th History Symp.*, Phoenix, AZ, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/95Annual/webprogram/Paper266329.html.

Evans, J. S., and C. A. Doswell III, 2001: Examination of derecho environments using proximity soundings. *Wea. Forecasting*, **16**, 329–342, https://doi.org/10.1175/1520-0434(2001)016<0329:EODEUP>2.0.CO;2.

Ferree, J., 2009: National change of the hail criteria for severe storms from 3/4 inch to 1 inch beginning January 5, 2010. National Weather Service, 8 pp., http://www.nws.noaa.gov/oneinchhail/docs/One_Inch_Hail.pdf.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gallus, W. A., Jr., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, https://doi.org/10.1175/2007WAF2006120.1.

Gensini, V. A., and W. S. Ashley, 2011: Climatology of potentially severe convective environments from the North American Regional Reanalysis. *Electron. J. Severe Storms Meteor.*, **6** (8), http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/85.

——, T. L. Mote, and H. E. Brooks, 2014: Severe-thunderstorm reanalysis environments and collocated radiosonde observations. *J. Appl. Meteor. Climatol.*, **53**, 742–751, https://doi.org/10.1175/JAMC-D-13-0263.1.

Gourley, J. J., and Coauthors, 2013: A unified flash flood database across the United States. *Bull. Amer. Meteor. Soc.*, **94**, 799–805, https://doi.org/10.1175/BAMS-D-12-00198.1.

Hales, J., Jr., 1988: Improving the watch/warning program through use of significant event data. Preprints, *15th Conf. on Severe Local Storms,* Baltimore, MD, Amer. Meteor. Soc., 165–168.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2924, https://doi.org/10.1256/qj.06.25.

Hart, J. A., and A. E. 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.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

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

——, and ——, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, https://doi.org/10.1175/WAF-D-13-00132.1.

——, and ——, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. *Wea. Forecasting*, **32**, 1321–1328, https://doi.org/10.1175/WAF-D-16-0170.1.

Jacks, E., 2014: Service change notice 14-42. Fire and Public Weather Services Branch, National Weather Service, http://www.nws.noaa.gov/os/notification/scn14-42day1-3outlooks_cca.htm.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

Kay, M. P., and H. E. Brooks, 2000: Verification of probabilistic severe storm forecasts at the SPC. Preprints, *20th Conf. on Severe Local Storms,* Orlando, FL, Amer. Meteor. Soc, 9.3.

Lackmann, G., 2011: *Midlatitude Synoptic Meteorology.* Amer. Meteor. Soc., 360 pp.

Markowski, P., and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes.* John Wiley and Sons, 424 pp.

McGovern, A., D. J. Gagne, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Mach. Learn.*, **95**, 27–50, https://doi.org/10.1007/s10994-013-5343-x.

——, K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

Murphy, A. H., and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Stat.*, **26**, 41–47, https://doi.org/10.2307/2346866.

Nielsen, E. R., and R. S. Schumacher, 2016: Using convection-allowing ensembles to understand the predictability of an extreme rainfall event. *Mon. Wea. Rev.*, **144**, 3651–3676, https://doi.org/10.1175/MWR-D-16-0083.1.

——, G. R. Herman, R. C. Tournay, J. M. Peters, and R. S. Schumacher, 2015: Double impact: When both tornadoes and flash floods threaten the same place at the same time. *Wea. Forecasting*, **30**, 1673–1693, https://doi.org/10.1175/WAF-D-15-0084.1.

NWS, 2012: Technical implementation notice 11-53. National Weather Service Headquarters, http://www.nws.noaa.gov/os/notification/tin11-53ruc_rapaae.htm.

——, 2017a: Weather fatalities 2016. Office of Climate, Weather, and Water Services, National Weather Service, http://www.nws.noaa.gov/om/hazstats.shtml.

——, 2017b: Service change notice 17-100. National Centers for Environmental Prediction, Weather Prediction Center, http://www.nws.noaa.gov/os/notification/scn17-100wpc_excessive_rainfall.htm.

Polger, P. D., B. S. Goldsmith, R. C. Przywarty, and J. R. Bocchieri, 1994: National Weather Service warning performance based on the WSR-88D. *Bull. Amer. Meteor. Soc.*, **75**, 203–214, https://doi.org/10.1175/1520-0477(1994)075<0203:NWSWPB>2.0.CO;2.

Schneider, R. S., and A. R. Dean, 2008: A comprehensive 5-year severe storm environment climatology for the continental United States. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 16A.4, https://ams.confex.com/ams/24SLS/techprogram/paper_141748.htm.

Schroeder, A. J., and Coauthors, 2016: The development of a flash flood severity index. *J. Hydrol.*, **541**, 523–532, https://doi.org/10.1016/j.jhydrol.2016.04.005.

Sherburn, K. D., and M. D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-CAPE environments. *Wea. Forecasting*, **29**, 854–877, https://doi.org/10.1175/WAF-D-13-00041.1.

——, ——, J. R. King, and G. M. Lackmann, 2016: Composite environments of severe and nonsevere high-shear, low-CAPE convective events. *Wea. Forecasting*, **31**, 1899–1927, https://doi.org/10.1175/WAF-D-16-0086.1.

Simmons, K. M., and D. Sutter, 2005: WSR-88D radar, tornado warnings, and tornado casualties. *Wea. Forecasting*, **20**, 301–310, https://doi.org/10.1175/WAF857.1.

——, and ——, 2008: Tornado warnings, lead times, and tornado casualties: An empirical investigation. *Wea. Forecasting*, **23**, 246–258, https://doi.org/10.1175/2007WAF2006027.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016a: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, https://doi.org/10.1175/WAF-D-16-0073.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

SPC, 2017a: SPC convective outlooks. Storm Prediction Center, http://www.spc.noaa.gov/cgi-bin-spc/getacrange.pl.

——, 2017b: SVRGIS (updated: 15 May 2017). Storm Prediction Center, http://www.spc.noaa.gov/gis/svrgis/.

——, 2017c: Severe weather climatology (1982–2011). Storm Prediction Center, http://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=anySvr.

Stephenson, D., B. Casati, C. Ferro, and C. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.*, **15**, 41–50, https://doi.org/10.1002/met.53.

Stough, S., E. Leitman, J. Peters, and J. Correia Jr., 2010: The role of Storm Prediction Center products in decision making leading up to severe weather events. Storm Prediction Center, 14 pp., http://www.spc.noaa.gov/publications/leitman/stough.pdf.

Surcel, M., I. Zawadzki, and M. Yau, 2016: The case-to-case variability of the predictability of precipitation by a storm-scale ensemble forecasting system. *Mon. Wea. Rev.*, **144**, 193–212, https://doi.org/10.1175/MWR-D-15-0232.1.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415, https://doi.org/10.1175/WAF925.1.

Vaughan, M. T., B. H. Tang, and L. F. Bosart, 2017: Climatology and analysis of high-impact, low predictive skill severe weather events in the northeast United States. *Wea. Forecasting*, **32**, 1903–1919, https://doi.org/10.1175/WAF-D-17-0044.1.

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, https://doi.org/10.1175/WAF910.1.

Vescio, M. D., and R. L. Thompson, 2001: Subjective tornado probability forecasts in severe weather watches. *Wea. Forecasting*, **16**, 192–195, https://doi.org/10.1175/1520-0434(2001)016<0192:FSFSTP>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.