REVIEW

Generating Probabilistic Forecasts from Convection-Allowing Ensembles Using Neighborhood Approaches: A Review and Recommendations

CRAIG S. SCHWARTZ AND RYAN A. SOBASH

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 19 October 2016, in final form 9 March 2017)

ABSTRACT

“Neighborhood approaches” have been used in two primary ways to postprocess and verify high-resolution ensemble output. While the two methods appear deceptively similar, they define events over different spatial scales and yield fields with different interpretations: the first produces probabilities interpreted as likelihood of event occurrence at the grid scale, while the second produces probabilities of event occurrence over spatial scales larger than the grid scale. Unfortunately, some studies have confused the two methods, while others did not acknowledge multiple possibilities of neighborhood approach application and simply stated, “a neighborhood approach was applied” without supporting details. Thus, this paper reviews applications of neighborhood approaches to convection-allowing ensembles in hopes of clarifying the two methods and their different event definitions. Then, using real data, it is demonstrated how the two approaches can yield statistically significantly different objective conclusions about model performance, underscoring the critical need for thorough descriptions of how neighborhood approaches are implemented and events are defined. The authors conclude by providing some recommendations for application of neighborhood approaches to convection-allowing ensembles.

1. Introduction

Over the last decade, increased computing resources have yielded a proliferation of numerical weather prediction (NWP) models with horizontal grid spacing fine enough to remove cumulus parameterization and allow organic development of convection through model dynamics. Although these “convection allowing” NWP models produce more realistic convective structures than coarser-resolution, convection-parameterizing models, traditional objective verification metrics, which usually require that forecast and observed events match at the grid scale for a forecast to be considered perfect, have not always corroborated subjective evaluations favoring convection-allowing models over convection-parameterizing models (e.g., Mass et al. 2002; Weisman et al. 2008).

Thus, in an attempt to reconcile disparities between objective metrics and subjective evaluations, several spatial verification methods have been developed that can broadly be categorized into “neighborhood,” scale-separation, object-based, and field deformation approaches (e.g., Gilleland et al. 2009, 2010). This paper focuses on neighborhood methods (e.g., Ebert 2008), which, like the other spatial verification approaches, inherently recognize limited small-scale predictability (e.g., Lorenz 1969) and acknowledge it is unrealistic to expect high-resolution models to be accurate at the grid scale. Therefore, while various flavors of neighborhood methods exist (e.g., Ebert 2008), they generally share a common trait of relaxing the traditional requirement that forecast and observed events match at the grid scale for a forecast to be considered perfect.

Neighborhood approaches were initially applied to deterministic forecasts (e.g., Theis et al. 2005; Roberts and Lean 2008), but extensions to ensemble forecasts quickly followed (e.g., Schwartz et al. 2010). Thereafter, because neighborhood approaches are easily implemented, intuitive, and effective, they have been widely adopted both as stand-alone methods to produce probabilistic guidance (e.g., Jirak et al. 2012; Clark et al. 2013; Ben Bouallégue and Theis 2014; Schwartz et al. 2015a,b; Sobash et al. 2016) and as techniques to verify deterministic and ensemble forecasts of a variety of
meteorological fields, including precipitation (e.g., Roberts and Lean 2008), updraft helicity (Sobash et al. 2011, 2016; Clark et al. 2013; Yussouf et al. 2013a), hail (Gagne et al. 2015; Snook et al. 2016), reflectivity (Stratman et al. 2013; Snook et al. 2015; Hitchcock et al. 2016), lightning (Lynn et al. 2015), and low-level vertical vorticity (e.g., Snook et al. 2012, 2015; Yussouf et al. 2013b, 2015; Wheatley et al. 2015; Zhuang et al. 2016).

This paper concerns two specific neighborhood approaches applied to convection-allowing ensembles that appear deceptively similar, but, in fact, differ both methodologically and philosophically and were the focus of Ben Bouallègue and Theis (2014, hereafter BBT14). The first method, which BBT14 called “fuzzy” probabilistic forecasts, is identical to the neighborhood ensemble probability (NEP) introduced by Schwartz et al. (2010), which applies a neighborhood approach to generate fields interpreted as gridpoint probabilities. Conversely, the second method, which BBT14 termed “upsampling,” applies a neighborhood approach to produce fields interpreted as probabilities within an area larger than the grid scale. Comparing the two methods, a primary difference is the spatial scale over which events are defined, which has historically caused confusion (e.g., Murphy et al. 1980).

Thus, although BBT14’s upsampling and fuzzy methods both leverage neighborhood approaches, the two techniques yield forecasts with different interpretations because they examine event occurrence over different spatial scales, underscoring the importance of fully describing exactly how neighborhood approaches are applied in journal articles and texts. Unfortunately, there are numerous instances in which authors have stated “a neighborhood approach was applied” without additional details, which is insufficient for readers to fully interpret the results (Yussouf et al. 2013a, b, 2015; Potvin and Wicker 2013; Duda et al. 2014, 2016; Snook et al. 2015, 2016; Luo and Chen 2015; Wheatley et al. 2015; Wang and Wang 2017). Furthermore, there appears to be some confusion; sometimes, the NEP has been incorrectly cited to describe instances where procedures more similar to BBT14’s upsampling were actually employed (e.g., Barthold et al. 2015; Hardy et al. 2016). Finally, potentially adding to further misunderstanding, BBT14’s upsampling, premised upon neighborhood maxima, is not identical to what Ebert (2008) termed upsampling, yet another method used to perform verification that employs averaging over various spatial scales (e.g., Zepeda-Arce et al. 2000; Clark et al. 2011) and is thus similar, but not identical, to BBT14’s upsampling.

Therefore, despite Ebert (2008)’s efforts to summarize various neighborhood approaches for deterministic forecasts, we believe confusion remains regarding terminology, implementation, and interpretation of neighborhood approaches, particularly for ensemble applications. Accordingly, the first part of this paper (section 2) reviews how various studies have applied neighborhood approaches to postprocess and verify convection-allowing ensemble forecasts to clarify misunderstandings. The second part of this paper (section 3) resembles BBT14, as we apply the NEP and a version of BBT14’s upsampling to ~9 months of 10-member, 3-km ensemble forecasts. However, our intent is broader than BBT14; whereas BBT14 concluded both their upscaled and fuzzy probabilities should be presented to forecasters, here, we provide more specific recommendations regarding when BBT14’s upsampling should be used in lieu of the NEP, and vice versa. Furthermore, unlike BBT14, we directly compare verification metrics computed with the two neighborhood methods to demonstrate how they can yield differing conclusions about forecast quality, highlighting the need for precise descriptions of neighborhood methodologies and event definitions.

2. Review of neighborhood approaches applied to convection-allowing ensembles

This section reviews two neighborhood techniques that have regularly been employed to postprocess and verify convection-allowing ensemble forecasts. Although the two methods share commonalities, they produce fields with very different interpretations because they define events over different spatial scales.

While we briefly consider three-dimensional neighborhoods, we primarily focus on two-dimensional spatial neighborhoods, which is sufficient to illustrate key differences between the two approaches. Furthermore, although most of the review regards generation and interpretation of probabilistic forecasts, we also describe how the probabilistic fields should be compared to observations. This verification discussion implicitly assumes that gridded observations are available, but we note that methods have been developed to perform neighborhood-based verification against point observations (e.g., Mittermaier 2014).

a. A neighborhood approach to derive grid-scale probabilities

1) METHODOLOGY

Let \( q \) denote an event threshold (e.g., \( q = 1.0 \text{ mm h}^{-1} \)) and \( f_{ij} \) forecasts at \( i = 1, \ldots, M \) grid points for \( j = 1, \ldots, N \) ensemble members. Then, the binary probability (BP) of event occurrence at the \( i \)th point for the \( j \)th member (\( BP_{ij} \)) is simply

\[
BP(q)_{ij} = \begin{cases} 
1 & \text{if } f_{ij} \geq q \\
0 & \text{if } f_{ij} < q 
\end{cases},
\]
where we have explicitly noted that $BP_{ij}$ is a function of $q$. As proposed by Theis et al. (2005), a neighborhood approach can transform $BP_{ij}$ into probabilistic forecasts of event occurrence at $i$ by first choosing a neighborhood length scale. This length scale is specified as either a number of grid boxes or physical distance, can be applied using square (e.g., Theis et al. 2005; Roberts and Lean 2008) or circular (e.g., Schwartz et al. 2009) neighborhood geometry, and defines the total number of grid boxes, $N_b$, within the neighborhood of the $i$th point. For a given neighborhood length scale, $N_b$ is constant for each of the $M$ grid points, except, perhaps, for grid points near lateral boundaries, where modifications to $N_b$ may be necessary (e.g., Nachamkin and Schmidt 2015).

Letting $S_i$ denote the unique set of $N_b$ points within the neighborhood of the $i$th point, $BP_{ij}$ can be transformed into a neighborhood probability (NP) at $i$ for the $j$th ensemble member (NP$_j$) by dividing the number of points within the neighborhood of $i$ where the event occurs by the total number of points in the neighborhood ($N_b$):

$$NP(q)_{ij} = \frac{1}{N_b} \sum_{k=1}^{N_b} BP_{ki}, \quad k \in S_i, \quad (2)$$

In other words, to produce NP$_{ij}$ from BP$_{ij}$, averaging within a neighborhood is performed. As averaging has the effect of spatial smoothing, in this case, the neighborhood length scale can be interpreted as a smoothing length scale, which we denote as $r$.

Strictly, NP$_{ij}$ is the fractional coverage of event occurrence within the neighborhood of the $i$th point for the $j$th member. Theis et al. (2005) interpreted these fractional coverages as probabilities derived from a “pseudoensemble” comprising the $N_b$ points within the neighborhood of $i$, where NP$_j$ is the fraction of the $N_b$ pseudoensemble members containing the event, and subsequent studies offered similar probabilistic interpretations of NP$_{ij}$ (e.g., Roberts and Lean 2008). Paradoxically, while neighborhoods are used to derive NP$_{ij}$, NP$_j$ itself is interpreted as a grid-scale probability (e.g., probability of precipitation exceeding a threshold at a given grid point), which was also noted by BBT14.

We concur with precedents of interpreting neighborhood-derived fractional coverages as probabilities but believe the smoothing length scale $r$ must be included in interpretation of NP$_{ij}$ to fully explain the spatial scale of the probabilities. For example, consider the hypothetical where a localized event occurs at and near the $i$th grid point but nowhere else for the $j$th member. Thus, for small $r$, NP$_{ij}$ is relatively large, and as $r$ increases, NP$_{ij}$ decreases, as progressively fewer points within the neighborhood contain the event. Clearly, $r$ governs NP$_{ij}$, yielding an interpretation of NP$_{ij}$ as a probability of event occurrence at $i$ given a smoothing length scale. Ultimately, $r$ should be chosen to focus on predictable spatial scales (e.g., Roberts and Lean 2008).

Averaging NP$_{ij}$ over all ensemble members yields the NEP, as defined by Schwartz et al. (2010) and called “fuzzy probabilities” by BBT14:

$$NEP(q)_i = \frac{1}{N} \sum_{j=1}^{N} NP_{ij}, \quad (3)$$

The NEP at the $i$th point (NEP$_i$) is the ensemble mean probability of event occurrence at $i$ given a smoothing length scale $r$. As with interpretation of NP$_{ij}$, NEP$_i$ is a gridpoint probability and function of $q$.

Alternatively, NEP can be obtained by first averaging BP$_{ij}$ over all ensemble members to obtain the gridpoint ensemble probability (EP) of event occurrence at $i$ (EP$_i$), assuming all ensemble members are equally likely:

$$EP(q)_i = \frac{1}{N} \sum_{j=1}^{N} BP_{ij}. \quad (4)$$

Then, averaging EP$_i$ over the $N_b$ points within the neighborhood of $i$ yields the NEP:

$$NEP(q)_i = \frac{1}{N_b} \sum_{k=1}^{N_b} EP_k, \quad k \in S_i. \quad (5)$$

Although the two methods to calculate NEPs are mathematically equivalent, obtaining NEPs from Eqs. (4) and (5) is computationally more efficient because it requires searching over the neighborhood just once, as opposed to $N$ times, as required by Eqs. (2) and (3). Moreover, both procedures to produce NEPs contain two distinct averaging (smoothing) steps: ensemble averaging and neighborhood averaging. The difference

\[1\] The term “neighborhood length scale” is generic and refers to a scale that defines the neighborhood about each grid point. As will be explained later, the neighborhood length scale can further be categorized as a “smoothing” or “searching” length scale, depending on what operations are performed within the neighborhood.

\[2\] When circular geometry is used, the length scale is often defined as the radius of the circle in kilometers. When square geometry is used, the length scale can be defined as either the number of grid boxes comprising the side of the square (e.g., the neighborhood is a $3 \times 3$ square centered on the $i$th point) or half- or full-width of the square in kilometers. The length scale determines the location and number ($N_b$) of grid boxes within the neighborhood of the $i$th point. Although we primarily use circular geometry herein, our arguments and results are insensitive to neighborhood geometry.
between the two procedures to compute NEPs is the order in which averaging is performed.

2) OBJECTIVE VERIFICATION OF NEPS AGAINST OBSERVATIONS

As NEPs are interpreted as grid-scale probabilities, assuming NEPs and verifying observations are on a common grid, objective verification can proceed by comparing the NEP at the $i$th point to observations at $i$, as in Schwartz et al. (2010) and BBT14. For most metrics, observed probabilities at $i$ are binary, but some metrics, like the fractions skill score (Roberts and Lean 2008), require a fractional observed field at $i$, which is obtained using Eqs. (1) and (2), where $N = 1$ and operations occur on a grid of observations. Like the NEP, observed fractions are also interpreted as grid-scale probabilities, and, overall, NEP values at $i$ can be appropriately compared to corresponding observations solely at $i$, regardless of whether observations at $i$ are binary or fractional.

3) NEP APPLICATIONS

NEPs of precipitation (e.g., Schwartz et al. 2010, 2014, 2015a,b; Johnson and Wang 2012; Duc et al. 2013; Ben Bouallègue et al. 2013; BBT14; Romine et al. 2014; Schumacher and Clark 2014; Yussouf et al. 2016) and simulated reflectivity (e.g., Kober et al. 2012; Snook et al. 2013; Scheufele et al. 2014; Hitchcock et al. 2016) from convection-allowing ensembles have regularly been verified. Interestingly, most of these studies focused on NEPs for precipitation thresholds ($q$) $\leq 12.7$ mm h$^{-1}$. For example, Kober et al. (2012) and Scheufele et al. (2014) only examined NEPs of simulated reflectivity $\geq 19$ dBZ, which corresponded to a precipitation rate of approximately 1.0 mm h$^{-1}$, and the highest precipitation threshold ($q$) for which NEPs were verified in Ben Bouallègue et al. (2013), BBT14, Schwartz et al. (2010, 2015a), and Johnson and Wang (2012) was 5.0 mm h$^{-1}$, 10 mm (6h)$^{-1}$, 10.0 mm h$^{-1}$, and 12.7 mm h$^{-1}$, respectively. Only rarely have NEPs of precipitation rates $>12.7$ mm h$^{-1}$ been explicitly examined, typically within the context of case studies of flash flooding events (e.g., Schumacher and Clark 2014; Yussouf et al. 2016).

As discussed more thoroughly in section 3, NEPs typically do not possess good reliability or resolution for rare events because of sharpness loss (sharpness refers to the tendency of probabilistic forecasts to produce values near 0 or 1; Murphy 1993). Thus, although authors seldom provide specific reasons for maximum values of $q$, this undesirable NEP property may have confined NEP applications to events that commonly occur within model climatologies. The next subsection describes methods that may be more appropriate for rarer events.

b. A neighborhood approach to derive non-grid-scale probabilities

1) METHODOLOGY

Whereas the neighborhood approach described in section 2a defines grid-scale events and yields grid-scale probabilities (i.e., NEPs), a neighborhood approach can also be applied to obtain probabilistic fields interpreted as likelihood of event occurrence over spatial scales larger than the grid length. To do so, as in section 2a, a neighborhood length scale is selected to determine which points are within the neighborhood of the $i$th point for the $j$th ensemble member. However, whereas the NEP length scale was interpreted as a smoothing scale (i.e., $r$), here, the relevant length scale defining the neighborhood about $i$ is interpreted as a searching length scale $x$, which determines the number of points in the neighborhood; $x$ is potentially distinct from $r$. Using $x$, all grid boxes within $x$ km of $i$ are searched to determine whether the event has occurred at any grid box within $x$ km of $i$ to redefine event occurrence at $i$. Alternatively and identically, event occurrence at $i$ can be determined by searching for the maximum value within the neighborhood of $i$ and comparing it to $q$. Therefore, these searching procedures can be considered “neighborhood maximum” approaches, and when either is applied to a single deterministic forecast, the resulting field is interpreted as whether the event occurs within $x$ km of $i$ and is still deterministic: either the event occurred within $x$ km of $i$ or it did not. Thus, when these searching approaches are applied in deterministic applications, the output can be considered a binary neighborhood probability (BNP).

In ensemble frameworks, the neighborhood maximum approach can be applied separately for each member, and averaging across the $N$ BNP fields yields a probabilistic product that is a “neighborhood maximum ensemble probability” (NMEP). Thus, NMEP employs ensemble averaging, like the NEP, but, unlike the NEP, production of NMEPs does not inherently contain spatial smoothing [although NMEPs can be smoothed, as discussed in section 2b(2)]. As with the NEP, NMEP is a function of $q$.

There are various flavors of NMEPs employing different grids to define probabilities. One approach explicitly subdivides the computational domain into coarser-resolution squares with horizontal grid spacing of $x$ km and searches for event occurrence among all native model grid points residing within each coarse-resolution grid box (e.g., BBT14; Sobash et al. 2016). If the event occurs at any native model grid point within a particular coarse grid box, the coarse grid box receives a
A value of 1; a value of 0 is assigned otherwise. Averaging over \( N \) ensemble members yields the NMEP and is interpreted as the likelihood of event occurrence anywhere within the coarse grid box with an effective resolution of the coarse grid length \( x \) (Fig. 1a).

Conversely, NMEPs can be computed on the native grid by searching for events within \( x \) km of each point (Figs. 1b,c). This approach yields a smoother field than the coarse grid approach and probabilities at the native grid’s resolution (cf. Figs. 1a and 1b). However, regardless of whether NMEPs are computed at each native model grid point or on a coarser-resolution grid, to produce NMEPs, binary fields from each ensemble member are averaged, whereas NEPs are derived from averaging nonbinary fields [e.g., Eqs. (3) and (5)]. Therefore, while the NEP is discretized in intervals of \( 1/(N \times N_p) \) and effectively continuous for typical \( N \) and \( r \), NMEP fields are discretized in comparatively coarse intervals of \( 1/N \) (i.e., for a 10-member ensemble, NMEPs can only be 0, 0.1, 0.2, ..., 0.9, and 1.0).

The probability of hourly accumulated precipitation \( \geq 25.4 \text{ mm h}^{-1} \) was zero everywhere over this subdomain.

Fig. 1. Various NMEP fields for probability of hourly accumulated precipitation \( \geq 25.4 \text{ mm h}^{-1} \): (a) unsmoothed NMEPs computed in 80 km \( \times \) 80 km coarse-resolution squares (see text), unsmoothed NMEPs computed at each point on the native grid with a 40-km searching length scale using (b) square and (c) circular geometry, and (d)–(f) smoothed NMEP fields corresponding to the unsmoothed fields in (a)–(c). The 40-km searching length scale (i.e., \( x = 40 \text{ km} \)) represents the radius of a circle or half-width of a square, depending on the neighborhood geometry. These probabilities were produced from the 18-h forecast of NCAR’s experimental 3-km, 10-member ensemble (Schwartz et al. 2015a) initialized at 0000 UTC 2 Jun 2016. In (d), smoothing was performed by inputting the unsmoothed NMEP field in (a) to Eq. (6) with \( \sigma = 120 \text{ km} \), and fields in (e) and (f) were smoothed by inputting the unsmoothed NMEP fields in (b) and (c) to Eq. (5) with a 40-km smoothing neighborhood length scale (i.e., \( r = 40 \text{ km} \), where \( r \) is the half-width of the square or radius of the circle). Of note, NEPs of hourly accumulated precipitation \( \geq 25.4 \text{ mm h}^{-1} \) were zero everywhere over this subdomain.

NMEP is conceptually identical to BBT14’s “upscaling,” which, in turn, is a flavor of what Ebert (2008) called “minimum coverage” after Damrath (2004), but expanded for ensemble applications. As noted by BBT14, 3 BBT14’s “upscaling” (i.e., NMEP) is not synonymous with the upscaling methods employed by Zepeda-Arce et al. (2000), Marsigli et al. (2008), Ebert (2009), or Clark et al. (2011). For example, Clark et al. (2011) averaged values within neighborhoods of each grid point in their upscaling while BBT14 took maximum values within neighborhoods in their upscaling. As both Clark et al. (2011)’s and BBT14’s methods coarsened high-resolution output, they each seem accurately described by the term “upscaling.” Therefore, given possibly different meanings and interpretations of this term, it is critical to fully describe upscaling methodologies and interpretations of the resulting fields.
the NMEP is the probability of event occurrence within x km of i, and, thus, is a probability defined over a spatial scale larger than the grid length. While there are two mathematically equivalent methods for producing the NEP (section 2a), NMEPs cannot be obtained from a field of point-based EPs [i.e., Eq. (4)]; the N individual members must be queried. If x is reduced to the native grid length, NEP and NMEP are identical and equal to point-based EPs.

2) SMOOTHING NMEP

While NMEPs are discretized based on ensemble size, a spatial smoother can transform NMEP grids into continuous probabilistic fields, analogously to producing NEPs from EPs using Eq. (5). Although any type of smoothing kernel can be employed to produce smoothed NMEP fields (NMEPsmooth), an often-used kernel with NMEP is a Gaussian filter (e.g., Brooks et al. 1998; Sobash et al. 2011, 2016; Hitchens et al. 2013; Schwartz et al. 2015b):

$$\text{NMEP}_{i,\text{smooth}} = \sum_{m=1}^{M} \frac{1}{2\pi\sigma^2} \exp \left[ -\frac{1}{2} \left( \frac{x_{i,m} - m}{\sigma} \right)^2 \right] \text{NMEP}_i,$$  \hspace{1cm} (6)

In Eq. (6), \(x_{i,m}\) is the physical distance from the ith point to the mth of M grid points, and the Gaussian standard deviation \(\sigma\) is an adjustable smoothing length scale that controls smoothing weights. Technically, Eq. (6) indicates all M grid points contribute to smoothing NMEP_i. However, points where \(x_{i,m} > 4\sigma\) contribute very little to smoothing at \(i\), and the summation in Eq. (6) can be truncated to search just the subset of the M points where \(x_{i,m} \leq 4\sigma\), increasing computational efficiency. Also, note that points where NMEP_i = 0 do not contribute to smoothed NMEP fields.

As \(\sigma\) represents a spatial scale of the filter, interpretation of Gaussian-smoothed NMEPs is potentially difficult because two possibly distinct spatial scales (x and \(\sigma\)) are combined. An alternative to the Gaussian smoother that avoids potentially conflating two spatial scales is to simply input NMEP fields into Eq. (5) with \(r = x\) and smooth over the same neighborhoods that were searched, which maintains just one length scale. Equation (5) can also be used with \(x \neq r\), which combines the NMEP smoothing length scale \(r\) with the NMEP searching length scale \(x\) to produce smoothed NMEP fields; in this case two distinct spatial scales are again used, potentially complicating interpretation. Some smoothed NMEP fields are shown in Figs. 1d–f.

A potential drawback of using Eq. (5) for smoothing is that Eq. (5) assumes all points within the neighborhood have equal weights, whereas Eq. (6), in which \(r\) is effectively replaced by \(\sigma\), assigns greater weight to points closer to \(i\). Moreover, if NMEPs are produced on a coarse-resolution grid defined by \(x\) (e.g., Fig. 1a), smoothing can be achieved with Eq. (5) only if \(r > x\), although it is cheaper to apply a smoother [either Eq. (5) or (6)] to NMEPs produced on a coarse-resolution grid (e.g., Fig. 1a) than a high-resolution grid.

However, no matter how NMEPs may be smoothed, smoothing increases the effective neighborhood length scale \(x_{\text{eff}}\). As shown in Fig. 2 for a single ensemble member (\(N = 1\)) and circular neighborhood, although the unsmoothed NMEP at the central grid box is 0 because all forecast events (shaded boxes in Fig. 2) occur outside the neighborhood of the central grid box, three points within the neighborhood of the central grid point have NMEP = 1 (Fig. 2a). Thus, thanks to these three points, when NMEP is smoothed using Eq. (5), NMEP = 0.33 at the central point (Fig. 2b). But, the three points within the neighborhood of the central grid box where unsmoothed NMEP = 1 had neighborhoods extending outside the neighborhood of the central grid box (Figs. 2c–e), illustrating how points outside the neighborhood of the central grid box contribute to NMEP at the central grid box when smoothing occurs.

Thus, to achieve NMEP_{i,\text{smooth}} \neq \text{NMEP}_i, \(x_{\text{eff}}\) must be greater than \(x\). The upper bound of \(x_{\text{eff}}\) is technically the size of the computational domain but in practice controlled by \(r\) or \(\sigma\) within the context of Eqs. (5) or (6), respectively. However, on a point-by-point basis, \(x_{\text{eff}}\) will vary considerably depending on event distribution.

Overall, because of potentially meaningful differences between NMEP_{i,\text{smooth}} and \(\text{NMEP}_i\), when generating and verifying smoothed NMEPs, the procedure and length scale used to smooth NMEPs should be clearly stated. Given the local variations of \(x_{\text{eff}}\), we suggest smoothed NMEPs simply be interpreted broadly as smoothed probabilities of event occurrence within x km of a point, where the smoothing scale is known.

3) OBJECTIVE VERIFICATION OF NMEPS AGAINST OBSERVATIONS

When objectively verifying NMEPs against observations, for consistency, application of neighborhoods to observations must follow NMEP procedures to determine observed event occurrence. Thus, as with construction of NMEPs, all points within the neighborhood of the ith observation point are searched to determine whether an observed event occurred at any point within the neighborhood (e.g., BBT14; Sobash et al. 2016), and if it has, then an observed event is deemed to have occurred at i. As a result, when verifying NMEPs, observed probabilities are necessarily binary, whereas NEP fields could potentially be verified against fractional observations.
Because observations are handled differently depending on whether NEPs or NMEPs are evaluated, observed event frequencies will differ when NEPs and NMEPs are verified. This difference should be considered when interpreting verification metrics comparing NEPs and NMEPs with observations, as discussed in section 3.

4) NMEP APPLICATIONS

Many studies have used NMEPs. For example, Clark et al. (2013), Schwartz et al. (2015b), and Sobash et al. (2016) used NMEPs when analyzing and verifying ensemble forecasts of updraft helicity. In Schwartz et al. (2015b) and Sobash et al. (2016), for comparison with observations, the native high-resolution grids were decomposed into 80 km by 80 km squares using BBT14’s approach, whereas Clark et al. (2013) computed NMEPs at each grid point. Furthermore, Schwartz et al. (2015b) and Sobash et al. (2016) used a Gaussian filter [Eq. (6)] to smooth NMEP fields, whereas Clark et al. (2013) did not. Additionally, Gaussian-smoothed NMEP fields computed at each grid point were employed by Jirak et al. (2012) and Schwartz et al. (2015a) to display probabilistic forecasts of severe weather parameters, Zhuang et al. (2016) applied NMEPs to maximum 0–5-km MSL vertical vorticity without Gaussian smoothing, and Hitchcock et al. (2016) examined Gaussian-smoothed NMEP fields of simulated reflectivity. Finally, Barthold et al. (2015) and Hardy et al. (2016) applied NMEPs to precipitation forecasts for flash flooding applications. Unlike Barthold et al. (2015), Hardy et al. (2016) performed the extra step of smoothing NMEPs with a Gaussian filter, and both Barthold et al. (2015) and Hardy et al. (2016) computed their respective NMEPs at each native grid point, unlike BBT14.
c. Vertical and temporal neighborhood extensions of NEP and NMEP

Thus far, we have only considered two-dimensional horizontal neighborhoods. However, neighborhoods can be extended both vertically and temporally. Regarding time neighborhoods, a discrete number of model output times $T$ can define a temporal neighborhood (e.g., Theis et al. 2005; Duc et al. 2013), similarly to how $r$ or $x$ selects a horizontal neighborhood length scale. As with purely horizontal neighborhoods (where $T = 1$), time neighborhoods can be incorporated within both NEP and NMEP frameworks; in the former, smoothing occurs over both space and time yielding NEPs discretized in intervals of $1/(N \times N_b \times T)$, and in the latter, searching for events occurs over all times and locations within space–time neighborhoods.

Extending purely horizontal neighborhoods to space–time neighborhoods does not alter expected behavior or interpretations of NEPs and NMEPs. For example, Duc et al. (2013) computed NEPs with several $T > 1$, and some of their results indicated that sharpness loss increased as $T$ increased, analogously to how sharpness loss increases with $r$ for horizontal neighborhoods. NMEPs incorporating space–time neighborhoods have primarily been used in association with forecasts of severe weather hazards (e.g., Jirak et al. 2012; Clark et al. 2013; Schwartz et al. 2015a,b; Sobash et al. 2016).

Neighborhood extension to the vertical dimension is also theoretically straightforward in cases where three-dimensional fields, such as vertical vorticity, are of interest. In fact, Zhuang et al. (2016) used the NMEP with three spatial dimensions to search over all model vertical levels below 5 km MSL when computing probabilistic forecasts of low-level vorticity. However, given often sharp vertical gradients, vertical smoothing is generally unlikely to be wise, and we cannot find examples when NEPs have been produced from three-dimensional spatial neighborhoods and suggest neighborhood extensions to three spatial dimensions only be performed using the NMEP.

d. Ambiguous applications of neighborhood approaches to convection-allowing ensembles

Whereas Barthold et al. (2015) and Hardy et al. (2016) incorrectly cited Schwartz et al. (2010)’s NEP to describe their methods (e.g., Potvin and Wicker 2013; Snook et al. 2015, 2016), and although these references may indeed be appropriate, given the ease with which NEP and NMEP can be confused, we suggest authors explicitly describe how they apply neighborhood approaches, rather than solely pointing to previous work. For example, to describe unsmoothed NMEP, simply stating that fields are “probabilities of exceeding $q$ within $x$ km of a grid point” may suffice.

e. Synthesis

A chief distinguisher between NEP and NMEP is smoothing versus searching. To produce NEPs, searching within neighborhoods for a specific grid box meeting certain criteria does not occur. Rather, all points within the neighborhood are treated equally, with smoothing (averaging) over both the neighborhood and ensemble (Fig. 3a). Conversely, NMEPs are derived by searching over neighborhoods for a single grid point meeting certain criteria before ensemble averaging; optional smoothing can then be applied (Fig. 3b). Although smoothing NMEP somewhat reconciles the procedures used to produce NEPs and NMEPs, smoothed NMEP fields fundamentally differ from NEPs.

The NEP has primarily been applied to forecasts of light and moderate precipitation, while NMEPs have more commonly been employed to display and verify forecasts of surrogates for explicit hazards prediction. However, decisions to use NEP rather than NMEP, or vice versa, have largely been unremarked; typically little justification is provided for using one approach over the other. Perhaps some studies of severe weather surrogates (e.g., Sobash et al. 2011) have used NMEPs, rather than NEPs, for easy comparison with Storm Prediction Center (SPC) probabilistic forecasts, which are interpreted as the likelihood of severe weather occurrence within 25 miles (40.2 km) of a point (e.g., Kay and Brooks 2000). Additionally, because NMEPs do not suffer as acutely from sharpness loss, they can more effectively evaluate events with low grid-scale observation base rates$^4$

$^4$As noted in section 2b(3), observations are handled differently depending on whether NMEPs or NEPs are objectively verified, such that base rates (observed event relative frequencies, or, sample climatologies) differ for the NEP and NMEP. Traditional base rates (Murphy and Winkler 1987) are used for the NEP, which are obtained by looking for observed events at the grid scale without considering neighborhoods. Conversely, NMEP base rates are obtained within the context of neighborhood searching. Thus, NMEP base rates are higher than NEP base rates [as shown in section 3c(2)]. In recognition of these two distinct base rates, we refer to the traditional, NEP-associated, base rate as the “grid-scale base rate.”
Overall, when deciding how to apply neighborhood approaches to convection-allowing ensembles, careful consideration should be given to the field of interest and interpretation of results. Additionally, forecasting goals and end-user requirements should be considered when determining whether to use the NEP or NMEP, especially if probabilities are displayed as a forecast product.

Moreover, because the NEP and NMEP have substantially different interpretations, it is critical that authors fully explain their methodologies and clearly state the spatial scales of events so readers can understand the resulting probabilities. Furthermore, the next section demonstrates how objective verification metrics computed from NEP and NMEP can be statistically significantly different, illustrating that without thorough descriptions of how neighborhood-based probabilities are produced and interpreted, it is impossible to fully comprehend verification results.

3. Demonstration of differences between NEP and NMEP

NEPs and NMEPs from convection-allowing ensemble forecasts were objectively verified. Our goal is not to thoroughly validate ensemble performance, but, rather, to simply demonstrate how NEP and NMEP can yield statistically significant different conclusions about model performance.

a. Forecast model

The National Center for Atmospheric Research’s (NCAR’s) experimental, real-time, 10-member ensemble forecasts with 3-km horizontal grid spacing initialized daily at 0000 UTC between 7 April and 31 December 2015 (269 total forecasts) were postprocessed to derive NEPs and NMEPs. Forecasts for each ensemble member were produced with version 3.6.1 of the WRF-ARW Model (Skamarock et al. 2008) and had common physical parameterizations (Table 1). Ensemble initial conditions were provided by downscaling analysis members of 15-km continuously cycling ensemble Kalman filter (Evensen 1994) analyses onto the 3-km computational domain (Fig. 4). Lateral boundary conditions were provided by NCEP’s Global Forecast System and perturbed for each member. For comprehensive details about the initial conditions and forecast system configurations, see Schwartz et al. (2015a).

b. Verification methods

Verification primarily focused on precipitation forecasts, but forecasts of instantaneous composite reflectivity were also examined. Precipitation forecasts were compared to NCEP stage IV (ST4) observations (Lin and Mitchell 2005); to enable this comparison, precipitation forecasts were interpolated to the ST4 grid (∼4.763-km horizontal grid spacing) using a budget interpolation method (Accadia et al. 2003). Composite reflectivity forecasts were compared to corresponding Multi-Radar Multi-Sensor (MRMS) observations (Smith et al. 2016; Zhang et al. 2016) produced on a 0.01° grid, so the MRMS data were interpolated to the 3-km model grid for verification.

NEPs and NMEPs were produced using circular geometry and NMEPs were computed at each grid point. Statistics were computed for several neighborhood length scales defined as radii about the i_th point, where...
the searching radius used for NMEP was identical to the smoothing radius used for NEP (i.e., \( r = x \) throughout section 3c).\(^5\) Unsmoothed NMEPs were verified rather than smoothed NMEPs, which avoids issues regarding modification of the neighborhood length scale when NMEPs are smoothed while illuminating the fundamental difference between NEPs and NMEPs (i.e., searching vs smoothing). All statistics were aggregated over the 269 forecasts initialized between 7 April and 31 December 2015 and verified over a region spanning the central United States (Fig. 4), hereafter referred to as the “verification region.”

Both raw and bias-corrected precipitation and reflectivity fields were verified to assess how bias impacts NEP and NMEP. Bias correction was performed using a probability matching approach (Ebert 2001) as described by Clark et al. (2010), which ensured identical forecast and observed event frequencies over the verification region for each ensemble member. However, bias correction did not impact conclusions regarding comparisons of NEP and NMEP for precipitation, so we solely present precipitation statistics based on raw, unbias-corrected fields. Conversely, as shown below, composite reflectivity forecasts were biased compared to MRMS observations, so reflectivity statistics were based on bias-corrected fields.

To determine whether differences between NEPs and NMEPs were statistically significant for a fixed neighborhood length scale, bounds of 99% confidence intervals (CIs) were determined using a bootstrap resampling technique based on pairwise differences between NEP and NMEP verification statistics (e.g., Hamill 1999; Wolff et al. 2014) with 1000 resamples. If CI bounds did not include zero, differences between the two experiments were statistically significant at the 99% level.

c. Results

The relative operating characteristic (ROC; Mason and Graham 2002), reliability statistics (Wilks 2011), and Brier skill score (BSS; Brier 1950) are common objective probabilistic verification metrics and were used to assess NEPs and NMEPs. Model climatologies were also examined to provide contexts for the objective scores.

Although NEPs and NMEPs were analyzed for many forecast lead times, results are only presented for 24-h forecasts. Examining other forecast hours did not yield additional insights, and assessing only 24-h forecasts is sufficient to demonstrate statistically significant differences between NEPs and NMEPs.

As described in section 3c(2), a fixed absolute event threshold (e.g., \( q = 1.0 \text{ mm h}^{-1} \)) corresponds to different NEP and NMEP observation and probability climatologies. However, we take the verification perspective of “fixed thresholds” and make no attempt to account for varied NEP and NMEP climatologies when comparing ROCs, BSSs, or reliability statistics to illustrate how fundamental climatological differences impact verification measures.

---

\(^5\) Because \( r = x, r \) and \( x \) can be used interchangeably in section 3c. However, for consistency with section 2, we use \( r \) when discussing NEPs and \( x \) when discussing NMEPs.
1) ENSEMBLE MEMBER CLIMATOLOGIES

As NEP and NMEP climatologies depend on ensemble members’ climatologies, percentiles of the ensemble composite reflectivity and hourly accumulated precipitation distributions (Fig. 5) were determined by compiling all 24-h forecasts of hourly accumulated precipitation and instantaneous composite reflectivity from each ensemble member over all grid points within the verification domain. Corresponding percentiles were also computed for MRMS and ST4 observations, which can be viewed as grid-scale observation base rates (Murphy and Winkler 1987).

Given a percentile \( y \) (e.g., \( y = 0.95 \), the 95th percentile), \( (1 - y) \) is the probability that precipitation or reflectivity will equal or exceed \( q_y \), the absolute value corresponding to \( y \). For example, for precipitation and \( y = 0.9975 \) (i.e., the 99.75th percentile), \( q_y = 12.5 \text{ mm h}^{-1} \) for the ensemble, indicating that 0.25% of forecast hourly accumulated precipitation values met or exceeded approximately 12.5 mm h\(^{-1}\) (Fig. 5a). Precipitation percentiles were nearly unbiased with respect to ST4 observations, while composite reflectivity forecasts were clearly biased compared to MRMS observations (Fig. 5). The top 1%, 0.1%, and 0.01% ensemble precipitation percentiles were approximately 5.0, 19.0, and 38.0 mm h\(^{-1}\), respectively, and, as shown next, properties of NEP and NMEP mean the two fields will differ most for these less-frequent events.

2) PROBABILITY DISTRIBUTIONS AND SHARPNESS

Understanding differences between NEP and NMEP requires examining how their distributions vary with event threshold (i.e., \( q \)). To perform this analysis, the number of NEP and NMEP grid points within selected probabilistic bins was calculated for a range of precipitation thresholds, considering all grid points over the verification domain and all 269 forecasts (Fig. 6). For example, for \( q = 0.25 \text{ mm h}^{-1} \) and \( r = 50 \text{ km} \), there were approximately \( 10^6 \) occurrences of NEPs between 35% and 45% over the 269 forecasts within the verification domain (green line in Fig. 6b).

Observed event frequencies corresponding to NEP and NMEP (asterisks in Fig. 6) differ, because observations were treated differently depending on whether NEPs or NMEPs were verified (section 2). Standard deviations of forecast probabilities, which measure sharpness (e.g., Mason 2004), are also shown in Fig. 6 (open circles), and Fig. 7 depicts ratios of the quantities in Fig. 6 for various pairs of NEPs and NMEPs.

As expected, for fixed \( q \), progressively higher NEPs occurred less frequently, and as \( q \) increased, NEP \( \geq 5\% \) became increasingly rare as sharpness decreased (Figs. 6b,d,f). Compared to NEPs, NMEPs were sharper and had more occurrences of probabilistic and observed events (Figs. 6 and 7a–c). When \( r \) increased, for NEP, both sharpness and frequencies of most probabilistic event occurrences decreased (Figs. 6a,b,d,f and 7d,e) while observed event frequencies were constant (by definition). Conversely, for NMEP, as \( x \) increased, frequencies of both observed and probabilistic event occurrences generally increased as sharpness improved (Figs. 6a,c,e,g and 7f,g). For neighborhood length scales \( \geq 50 \text{ km} \), when \( q \) increased to 5.0 mm h\(^{-1}\) and beyond, NEPs \( \geq 95\% \) never occurred and sharpness...
dropped precipitously (Figs. 6b,d,f), whereas NMEP values $\geq 95\%$ occurred until much larger $q$, with only a gradual sharpness decline as $q$ increased (Figs. 6c,e,g). That NEP sharpness decreased substantially at and above $q = 5.0 \text{ mm} \text{h}^{-1}$ is consistent with percentiles indicating precipitation $\geq 5.0 \text{ mm} \text{h}^{-1}$ occurred infrequently (Fig. 5a).

These collective behaviors are consistent with NEP and NMEP definitions (section 2) and illustrate their very different characters. With NMEP, as neighborhood length scales increase, the likelihood of individual ensemble members capturing events within their neighborhoods increases, yielding commensurately higher probabilities; observed event occurrence associated with NMEP also becomes more likely with increased $x$ (Figs. 7f,g). Conversely, with NEP, although increasing $r$ also means events are more likely to occur within the neighborhood, unless forecast events occur regularly, most newly included neighborhood points will not contain the event, leading to diluted probabilities and sharpness as $r$ increases (Figs. 7d,e). These fundamentally different properties and the corresponding differences regarding observation climatologies collectively engender many differences between NEPs and NMEPs, which are further manifested by the following metrics.

3) RELATIVE OPERATING CHARACTERISTIC

The ROC (Mason and Graham 2002) measures resolution, the capability to discriminate between events. To produce the ROC, probabilistic decision thresholds $p$ were chosen and a series of $2 \times 2$ contingency tables (Table 2) were populated for each $p$ and various $q$; the criteria for populating Table 2 are given in Table 3. Using elements of Table 2, the probability of detection $[\text{POD}; \text{POD} = a/(a + c)]$ and probability of false detection $[\text{POFD}; \text{POFD} = b/(b + d)]$ were computed for each probabilistic decision threshold, and the ROC was formed by plotting the POFD against the POD over the range of $p$.

Area under the ROC curve can summarize the ROC. A trapezoidal method that computes ROC areas $A$ directly from discrete (POFD, POD) pairs was used (Mason 1982), which is sensitive to specification of the probabilistic thresholds (Mason 1982; Harvey et al. 1992; Richardson 2000) and presents a challenge when comparing NEPs to unsmoothed NMEPs because the former is effectively continuous while the latter has discrete probabilities. Therefore, while unsmoothed NMEPs are relatively insensitive to how probabilistic thresholds are specified, NEP ROC areas can differ
substantially depending on the range of $p$. In particular, NEP ROC areas are sensitive to the lowest nonzero probabilistic threshold, $p_0$, interpreted as the smallest probability for which a user believes forecasting event occurrence is meaningful. To illustrate NEP ROC area sensitivity to $p_0$, $A$ was computed with three sets of probabilistic thresholds, where $p_0 = 5\%$, 1\%, and 0.5\% (Table 4).6 Because unsmoothed NMEPs were discretized in intervals of 10\%, unsmoothed NMEP ROC areas were identical across the sets (i.e., $p_0 < 10\%$ provided identical PODs and POFDs as $p_0 = 10\%$).

For precipitation, $r = x$, $p_0 = 5\%$, and constant decision thresholds, although NMEP had higher POFDs than NEP, these poorer NMEP POFDs were accompanied by higher PODs compared to NEP (Figs. 8a–c).

6 Among the three sets, the interval between adjacent decision thresholds also differed. However, the primary driver of ROC area differences was $p_0$, and omitting many intermediate probability thresholds did not significantly change ROC areas.

**TABLE 2.** Standard $2 \times 2$ contingency table for dichotomous events.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Forecast</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>Yes</td>
<td>$c$</td>
<td>$d$</td>
</tr>
<tr>
<td>No</td>
<td>$a + c$</td>
<td>$b + d$</td>
</tr>
</tbody>
</table>
As POD differences between NMEP and NEP were typically larger than POFD differences for $r = x$, NMEP had higher ROC areas than NEP (Fig. 9a). Differences between NEP and NMEP were largest for precipitation thresholds $\geq 5.0 \text{mm h}^{-1}$, after NEP rapidly lost sharpness. As the neighborhood length scale increased, NMEP ROC areas also increased for all precipitation thresholds, but for $q \geq 10.0 \text{mm h}^{-1}$, NEP ROC areas worsened for larger $r$ (Fig. 9a), due to greater sharpness loss as $r$ increased (e.g., Figs. 6a,b,d,f and 7d,e). ROC areas for reflectivity and $q$ increased, because lower $p_0$ compensated for sharpness loss at higher thresholds. Additionally, for the lowest precipitation and reflectivity thresholds, differences between NEP and unsmoothed NMEP ROC areas were smaller, and NEP ROC areas were sometimes higher than unsmoothed NMEP ROC areas (Figs. 9b,c,e,f). However, for all $p_0$ and $r = x$, NMEP ROC areas were statistically significantly highest for precipitation and reflectivity thresholds $\geq 30.0 \text{mm h}^{-1}$ and 40 dBZ, respectively. Note that if $p_0$ decreases further, NEP ROC areas will continue to increase, but it is questionable whether probabilistic decision thresholds $< 1\%$ are useful.

NEP and unsmoothed NMEP ROC areas were regularly statistically significantly different across all $p_0$ (Fig. 9), highlighting the critical need to thoroughly describe applications of neighborhood approaches. For example, consider two readers examining Figs. 8a–c and 9a without knowing how the underlying probabilistic forecasts were created, and assume that ROC areas $> 0.7$ indicate "useful" probabilistic predictions (Buzzi et al. 1999). The first person only sees NMEP statistics and concludes the ensemble is useful at most precipitation thresholds and has discriminating ability (i.e., ROC area $> 0.5$) at all thresholds. Conversely, the second person sees just NEP statistics and concludes the ensemble has either little or no discriminating ability for $q \geq 30.0 \text{mm h}^{-1}$ and is not useful (ROC areas $< 0.7$) for $q > 5.0 \text{mm h}^{-1}$.

Fortunately, these differing conclusions are reconciled once the spatial scales of events are clarified and interpretations of the probabilistic forecasts and corresponding observation climatologies are provided. For example, the likelihood of detecting events occurring within $x$ km of a point is larger than likelihood of detecting events occurring at a point, since more events occur within $x$ km of a point than at a point (Figs. 7a–c). Thus, typically higher NMEP PODs and POFDs compared to NEP were expected, which translated into generally higher NMEP ROC areas that indicated probabilistic forecasts of events occurring within a neighborhood of a point, rather than at a point, had better discriminating capability. Given the contexts of the probabilistic fields, these collective ROC results are unsurprising, complementary, and illustrate that knowing how to interpret underlying probabilistic forecasts is key to fully understanding ROC statistics.

4) RELIABILITY STATISTICS

Reliability statistics (Wilks 2011) were computed by first placing the $i$th NEP or unsmoothed NMEP value into its proper bin spanning 0%–5%, 5%–15%, 15%–25%, . . . , 85%–95%, and 95%–100%. Then, given that

<table>
<thead>
<tr>
<th>Quadrant of Table 2</th>
<th>NEP criteria</th>
<th>NMEP criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>$\text{NEP}(q) \geq p$ and $O_i \geq q$</td>
<td>$\text{NMEP}(q) \geq p$ and $O_k \geq q$ for some $k \in S_i$</td>
</tr>
<tr>
<td>b</td>
<td>$\text{NEP}(q) \geq p$ and $O_i &lt; q$</td>
<td>$\text{NMEP}(q) \geq p$ and $O_k &lt; q$ for all $k \in S_i$</td>
</tr>
<tr>
<td>c</td>
<td>$\text{NEP}(q) &lt; p$ and $O_i \geq q$</td>
<td>$\text{NMEP}(q) &lt; p$ and $O_k \geq q$ for some $k \in S_i$</td>
</tr>
<tr>
<td>d</td>
<td>$\text{NEP}(q) &lt; p$ and $O_i &lt; q$</td>
<td>$\text{NMEP}(q) &lt; p$ and $O_k &lt; q$ for all $k \in S_i$</td>
</tr>
</tbody>
</table>
the $i$th NEP or unsmoothed NMEP value was in its proper bin, observed event occurrence at $i$ was determined using the proper method for handling observations depending on whether NEP or NMEP was verified.

For $q \leq 5.0 \text{ mm h}^{-1}$, as $r$ increased, NEPs became more reliable (Figs. 10a–d). However, for $q \geq 5.0 \text{ mm h}^{-1}$, sharpness loss precluded progressively higher NEPs as $r$ increased, and for $q \geq 10.0 \text{ mm h}^{-1}$, reliability diagrams provided little information for NEPs (Figs. 10e–i). Conversely, unsmoothed NMEP provided meaningful reliability statistics for all thresholds, which indicated NMEP overforecast at most probabilities for $q \geq 20.0 \text{ mm h}^{-1}$, although increasing $x$ improved reliability (Figs. 10f–i). Statistically significant differences between NEP and NMEP also occurred for probabilities $<25\%$ for $q \leq 1.0 \text{ mm h}^{-1}$ (Figs. 10a–c). At these probabilities and $q$, NEPs suggested either perfect reliability or slight overprediction, while NMEPs suggested underprediction. Overall, shapes of NMEP curves indicated overconfidence (e.g., Hagedorn et al. 2005; Hudson et al. 2011), which was an undesirable side effect of good NMEP sharpness, and, generally, NMEP was more overconfident than NEP.

In at least some probability bins for all $q$, NEP and unsmoothed NMEP provided statistically significantly different interpretations regarding reliability. Again, if two readers examined Fig. 10 but the first and second saw solely NMEP and NEP reliability statistics, respectively, they would arrive at conflicting conclusions that are only reconcilable by understanding the underlying probabilistic fields. For example, given a 15%–25% likelihood of $q \geq 5.0 \text{ mm h}^{-1}$ within 150 km of a point (dashed orange line in Fig. 10d), observed events actually occurred within 150 km of the point approximately 30% of the time. Conversely, given a 15%–25% likelihood of $q \geq 5.0 \text{ mm h}^{-1}$ at a point (solid orange line in Fig. 10d), observed events actually occurred at the point approximately just 10% of the time. These findings are not inconsistent and again illustrate the necessity of fully understanding interpretation of probabilistic fields. Results for composite reflectivity mirrored those for precipitation (not shown).
The Brier score (BS; Brier 1950) is defined as

\[ BS = \frac{1}{M} \sum_{i=1}^{M} (P^f_i - O_i)^2, \]

where \( P^f_i \) represents any probabilistic forecast at \( i \) (e.g., NEP or NMEP) and \( O_i \) is 1 if an observed event occurred at \( i \) and 0 otherwise. BSs strongly depend on observed event frequency (Murphy 1973), and because observed climatologies differ depending on whether NEP or NMEP is verified (e.g., Figs. 6 and 7), the stand-alone BS provides little information about the relative quality of NEP and NMEP.

However, a Brier skill score comparing a BS to a reference BS can provide insights regarding differences between NEPs and NMEPs. The reference was chosen as the BS of climatological forecasts (BS\(_{\text{climo}}\), given by \( BS_{\text{climo}} = o(1 - o) \), where \( o \) is the observed event relative frequency (Murphy 1973; Fig. 6). The BSS is then

\[ BSS = 1 - \frac{BS}{BS_{\text{climo}}} = \frac{BS_{\text{RES}} - BS_{\text{REL}}}{BS_{\text{climo}}}, \]

where BS\(_{\text{RES}}\) and BS\(_{\text{REL}}\) are the respective resolution and reliability terms of the BS decomposition (e.g., Murphy 1973; Wilks 2011), and the appropriate BS\(_{\text{climo}}\) was used depending on whether NEP or unsmoothed NMEP was verified. In other words, the BSS contains a normalization directly related to base rate, which differ for NEP and NMEP; when NEPs were verified, \( o \) and associated BS\(_{\text{climo}}\) were given by grid-scale base rates.
and when unsmoothed NMEPs were verified, \( o \) was given by base rates obtained through neighborhood searching (e.g., Figs. 6c,e,g).

For precipitation thresholds \( \geq 10.0 \text{ mm h}^{-1} \), NMEP BSSs were statistically significantly higher than NEP BSSs, and BSSs increased as neighborhood lengths increased (Fig. 11a). However, for precipitation thresholds \( \geq 20.0 \text{ mm h}^{-1} \), NEP BSSs were near zero due to sharpness loss and NMEP BSSs were often strongly negative, indicating no skill compared to the corresponding climatologies. These negative NMEP BSSs are consistent with reliability statistics, which suggest NMEP had too many areas of higher probabilities for precipitation thresholds \( \geq 20.0 \text{ mm h}^{-1} \) and was overconfident (Figs. 10f–i).

Fig. 10. Reliability diagrams computed over the verification domain for hourly accumulated precipitation thresholds of (a) 0.25, (b) 0.5, (c) 1.0, (d) 5.0, (e) 10.0, (f) 20.0, (g) 30.0, (h) 40.0, and (i) 50.0 mm h\(^{-1}\), aggregated over all 24-h forecasts. Filled circles indicate those bins where differences between NEP and NMEP for a fixed neighborhood length scale were statistically significant at the 99th percentile, and the colors correspond to those in the legend (e.g., red symbols denote differences between NEP and NMEP for \( r = x = 50 \text{ km} \)). Values are not plotted for a particular bin if there were \(<1000\) grid points with forecast probabilities in that bin over the verification domain and all 269 24-h forecasts of hourly accumulated precipitation. Similarly, statistical significance was not assessed for a particular NEP–NMEP pair unless both NEP and NMEP had \( \geq 1000 \) grid points within the probability bin. The diagonal line denotes perfect reliability.
BSSs for simulated reflectivity yielded similar conclusions as those for precipitation (Fig. 11b), and as statistically significant differences occurred at all precipitation and reflectivity thresholds, BSSs again demonstrate the importance of properly describing construction and interpretation of probabilistic fields.

4. Discussion and summary

This paper reviewed two common applications of neighborhood approaches to convection-allowing ensemble forecasts in attempt to alleviate confusion regarding how the two methods are implemented and subsequently interpreted. The first neighborhood approach produces probabilities interpreted as likelihood of event occurrence at the grid scale (NEP), while the second method produces probabilities of event occurrence over spatial scales larger than the grid scale (NMEP). Numerous applications of NEP and NMEP were described and various flavors of NMEP were discussed. Furthermore, we showed how observations must be handled differently depending on whether they are used to verify NEPs or NMEPs, such that base rates associated with NEPs and NMEPs differ (Fig. 6).

Additionally, NEP and NMEP from 269 forecasts of NCAR’s experimental, real-time, 10-member, 3-km ensemble were used to demonstrate how NEP and NMEP can provide different objective conclusions about forecast performance. As NMEP has a more relaxed definition than NEP, unsurprisingly, for a fixed neighborhood length scale, NMEP yielded generally superior verification measures.

Given the differences between NEP and NMEP, which method is preferable? Although BBT14 suggested both NEP and NMEP have merits, we offer more specific recommendations for application of neighborhood approaches to convection-allowing ensembles:

1) Differences between NEP and NMEP ROC and reliability statistics were largest after the NEP began to quickly lose sharpness, which occurred for precipitation thresholds >5.0 mm h\(^{-1}\). This point of rapid sharpness loss is key; once sharpness substantially decreases, POD is diminished. Therefore, as NMEP retains sharpness better than NEP for low grid-scale base rate events, we suggest using NMEP, rather than NEP, for detection of low grid-scale base rate events, even though increased NMEP PODs will be accompanied by higher POFDs.
Calibration techniques (e.g., Johnson and Wang 2012 and references therein) could theoretically extract more value from NEPs for low grid-scale base rate events. For example, although for $r = 100\text{ km}$, over the 269 forecasts the highest 24-h NEP of hourly accumulated precipitation $\geq 50.0 \text{ mm h}^{-1}$ was just 2.56%, forecasters could recognize a 2.56% probability as climatologically extreme and adjust their forecasts accordingly; similar statistical adjustments could be performed. However, the range of NEPs for extreme events is potentially very small (e.g., 0%–2.56%), complicating calibration. Moreover, changes to model configurations could alter NEP climatologies, lessening the effectiveness of calibration methods.

2) NMEPs were overconfident for all precipitation thresholds (Fig. 10). Moreover, NMEPs and NEPs usually differed least for lighter rainfall rates (Figs. 6, 7, and 9). Therefore, for applications mostly concerned with high grid-scale base rate events, NEP may be a better choice than NMEP to limit overconfidence.

3) For fixed $x$, smoothed NMEPs will generally verify better than unsmoothed NMEPs, as, by virtue of effectively increasing the neighborhood length scale, smoothing the discrete NMEP beneficially “fills in” the probability density function (e.g., Schwartz et al. 2010). Indeed, when we smoothed NMEPs, ROC areas and BSSs increased (not shown). Thus, we suggest NMEPs be smoothed, which has the additional benefit of producing more visually pleasing fields (Fig. 1).

However, interpreting objective verification metrics applied to smoothed NMEP fields is challenging because both smoothing and searching length scales are combined. Ultimately, we suggest that smoothed NMEP fields be described using both the smoothing length scale $r$ and searching length scale $x$ such that smoothed NMEPs are interpreted as smoothed probabilities of event occurrence within $x$ km of a point, but caveats should be provided that the effective neighborhood length scale is larger due to smoothing (Fig. 2). Future work may quantify how much smoothing increases the effective neighborhood length scale and optimal smoothing distance; the amount should depend on the meteorological situation and focus on presenting model output on “skillful scales” (e.g., Roberts and Lean 2008).

4) For forecasts of severe weather using surrogates such as updraft helicity and low-level vertical vorticity, NMEP appears more suitable than NEP. As severe weather events are rare and typically occur over small spatial and temporal scales, smoothing inherent in the NEP will yield probabilities near zero that convey little meaning. Additionally, given that most convection-allowing ensembles are spread deficient (e.g., Romine et al. 2014), many users will probably find probabilistic severe weather forecasts within a distance of a point more valuable than gridpoint probabilities. Furthermore, NMEPs of 2–5-km hourly maximum updraft helicity had skill when verified against SPC storm reports as in Sobash et al. (2016), whereas NEPs had no skill (not shown).

5) Finally, forecast applications and end-user requirements should be considered. When forecasts of interest are over broader areas, like river basins, the NMEP may be most appropriate. Conversely, when point forecasts are desired, the NEP may be a better choice, even if NEPs are low.

Because neighborhood approaches are easy to implement, their use will likely continue. Moreover, every application is unique, and our recommendations may not be optimal in all situations. But, regardless of whether future studies use NEP or NMEP, our results demonstrated that verification scores based on NEP and NMEP were usually statistically significantly different, emphasizing that authors should explicitly describe their neighborhood methodologies and event definitions, rather than simply citing previous work or superficially stating “a neighborhood approach was applied.” Without providing thorough details of how neighborhood approaches are implemented, it is impossible for readers to interpret the resulting probabilistic fields and corresponding verification scores, which may translate into confusion and misunderstanding.

Acknowledgments. This work was partially supported by the Short Term Explicit Prediction (STEP) program. Glen Romine and Kate Fossell played large roles developing NCAR’s real-time convection-allowing ensemble that was used to demonstrate verification principles. Thanks to Barb Brown, Adam Clark, Elizabeth Ebert, and two anonymous reviewers for their thoughtful suggestions.

REFERENCES


