A Comparison of Methods Used to Populate Neighborhood-Based Contingency Tables for High-Resolution Forecast Verification

CRAIG S. SCHWARTZ

National Center for Atmospheric Research,[a] *Boulder, Colorado*

(Manuscript received 28 October 2016, in final form 7 December 2016)

ABSTRACT

As high-resolution numerical weather prediction models are now commonplace, "neighborhood" verification metrics are regularly employed to evaluate forecast quality. These neighborhood approaches relax the requirement that perfect forecasts must match observations at the grid scale, contrasting traditional point-by-point verification methods. One recently proposed metric, the neighborhood equitable threat score, is calculated from 2×2 contingency tables that are populated within a neighborhood framework. However, the literature suggests three subtly different methods of populating neighborhood-based contingency tables. Thus, this work compares and contrasts these three variants and shows they yield statistically significantly different conclusions regarding forecast performance, illustrating that neighborhood-based contingency tables should be constructed carefully and transparently. Furthermore, this paper shows how two of the methods use inconsistent event definitions and suggests a "neighborhood maximum" approach be used to fill neighborhood-based contingency tables.

1. Introduction

The equitable threat score (ETS; Schaefer 1990), also called the Gilbert skill score, measures agreement between forecast and observed events (e.g., nonzero precipitation at a grid point) via a 2×2 contingency table (Table 1). Traditionally, the ith of N grid points within corresponding forecast F_i and observed O_i fields is placed into quadrant a of Table 1 and called a "hit" if both forecast and observed events occur at i; b if an event is forecast, but unobserved, at i ("false alarm"); c if an observed event occurs but is not forecast at i ("missed event"); and d if both forecast and observed nonevents occur at i ("correct negative"). Using Table 1, the ETS is defined as

$$\text{ETS} = \frac{a - a_{\text{rand}}}{a + b + c - a_{\text{rand}}}, \tag{1}$$

where

$$a_{\text{rand}} = \frac{(a + b)(a + c)}{a + b + c + d} \tag{2}$$

and interpreted as the proportion of correctly predicted observed events adjusted for hits due to random chance. Other scores are also obtained from Table 1 (e.g., Wilks 2006), including the bias B, the probability of detection (POD), and the false alarm ratio (FAR), expressed as

$$B = \frac{a + b}{a + c}, \tag{3}$$

$$\text{POD} = \frac{a}{a + c}, \tag{4}$$

and

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

PODs, biases, and ETSs of 1 are optimal, while perfect FARs are 0.

When Table 1 is populated by evaluating the agreement of forecast and observed events on a point-by-point basis, scores derived from Table 1 are considered point-by-point metrics, where perfect scores are only achievable if forecasts and observations match at each grid point. However, as numerical weather prediction models are configured with progressively higher resolution, objective verification metrics requiring grid-scale accuracy, such as the traditional ETS, have not always supported subjective impressions favoring high-resolution models over coarser-resolution

[a] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author e-mail: Craig Schwartz, schwartz@ucar.edu

DOI: 10.1175/WAF-D-16-0187.1

TABLE 1. Standard 2 × 2 contingency table for dichotomous events.

		Observed event		
		Yes	No	
Forecast event	Yes	a	b	$a + b$
	No	c	d	$c + d$
		$a + c$	$b + d$	

models (e.g., Mass et al. 2002; Done et al. 2004; Weisman et al. 2008). Thus, in an attempt to objectively corroborate these subjective assessments, various "neighborhood approaches" (e.g., Ebert 2008) have been proposed that recognize it is unrealistic to expect high-resolution models to possess grid-scale accuracy and do not require "perfect" forecasts to match observations at the grid scale.

One neighborhood-based metric was developed by Clark et al. (2010, hereafter C10), who modified definitions of hits, misses, and false alarms to account for neighborhoods around each grid point. C10 then computed a "neighborhood ETS" ($\mathrm{ETS_{neigh}}$) with neighborhood-based contingency tables and showed that while traditional point-based ETSs did not objectively indicate 4-km precipitation forecasts were better than corresponding 12-km forecasts, $\mathrm{ETS_{neigh}}$ revealed distinct 4-km advantages, matching subjective evaluations. Several studies adopted C10's methodologies to compute neighborhood-based contingency table metrics, primarily for precipitation forecasts (e.g., Schumacher et al. 2013; Dahl and Xue 2016; Ma and Bao 2016; Squitieri and Gallus 2016; Pytharoulis et al. 2016), but also for predictions of lightning (Fierro et al. 2015; Lynn et al. 2015) and drylines (Clark et al. 2015).

Furthermore, McMillen and Steenburgh (2015, hereafter MS15) employed an $\mathrm{ETS_{neigh}}$ version resembling C10's but with slightly different definitions to populate Table 1. Additionally, Table 1 can be filled using a third set of criteria based on a "neighborhood maximum" (NM) approach (e.g., Sobash et al. 2011; Ben Bouallègue and Theis 2014; Barthold et al. 2015). Although the philosophies behind C10's, MS15's, and NM criteria are similar, their subtle differences (section 2) yield varying conclusions regarding forecast quality within the context of numerical weather prediction model evaluation (section 3), demonstrating that neighborhood-based contingency tables should be carefully filled and interpreted.

2. Three neighborhood-based methods of populating contingency tables

All three neighborhood-based contingency table definitions relax the traditional requirement that forecasts and observations must match at the grid scale for a hit to occur by selecting a radius of influence (r)[1] that defines a neighborhood about the ith point and is interpreted both as the spatial scale over which errors are tolerated and the spatial scale of event occurrence. For example, using a neighborhood approach, a possible event is measurable observed precipitation *within* r kilometers of i, while point-based verification implies the spatial scale of events is the horizontal grid length (i.e., measurable observed precipitation *at* i).

However, the three variations have slightly different event definitions and consequently differ regarding how neighborhoods are searched, leading to different rules for populating Table 1. Letting q denote an event threshold and S_i the unique set of points within r kilometers of i (S_i includes i), MS15's, C10's, and NM definitions for filling Table 1 are summarized in Table 2 and now described.

a. C10's definitions

C10 employed a neighborhood approach to broaden definitions of hits, false alarms, and misses. Quoting C10, who used circular geometry with radius r to define neighborhoods:

> If an event is observed at a grid point, this grid point is a hit if the event is forecast at the grid point or at any grid point within a circular radius r of this observed event. Similarly, if an event is forecast at a grid point, the grid point is a hit if an event is observed at the grid point or at any grid point within r of this forecast event. A miss is assigned when an event is observed at a grid point and none of the grid points within r forecast the event, and false alarms are assigned when an event is forecast at a grid point and not observed within r of the forecast. Correct negatives are assigned in the same way as for the traditional ETS computation (i.e., an event is neither forecast nor observed at a single grid point).

These criteria mean event occurrence directly at the ith point determines how i is classified. For example, a hit can only occur at i if either a forecast or observed event occurs at i. Furthermore, C10's method defines forecast and observed events over inconsistent spatial scales both within individual quadrants of and across Table 1; sometimes, events are defined at the grid scale (e.g., $F_i \geq q$) but at other times over spatial scales larger than the grid length [i.e., neighborhoods (S_i) are queried].

[1] The term "radius of influence" implies circular neighborhoods about the ith point, where r is the radius of the circle, and, herein, circular geometry is used. While neighborhoods can also be implemented using square geometry, where the neighborhood is an $N \times N$ square centered on the ith point, the following results and analyses are insensitive to neighborhood geometry.

Table 2. Criteria for filling Table 1's quadrants for the ith grid point. As noted in the text, S_i denotes the unique set of grid points within the neighborhood of i, q represents a precipitation accumulation event threshold, and F_i and O_i represent the forecasts and observations at i, respectively. Although this paper applies these definitions within the context of precipitation forecasts, the definitions can be used for any dichotomous situation, and q can either be an absolute threshold (e.g., 1.0 mm h^{-1}) or a percentile threshold (e.g., the 90th percentile).

			Quadrant of Table 1		
			a (hit)	*b* (false alarm)	*c* (missed event)	*d* (correct negative)
Point-based definitions						
Forecast condition	$F_i \geq q$			$F_i \geq q$	$F_i < q$	$F_i < q$
Observation condition	$O_i \geq q$			$O_i < q$	$O_i \geq q$	$O_i < q$
C10's definitions	Criteria 1	Criteria 2				
Forecast condition	$F_i \geq q$	$F_k \geq q$ for some $k \in S_i$		$F_i \geq q$	$F_k < q$ for all $k \in S_i$	$F_i < q$
Observation condition	$O_k \geq q$ for some $k \in S_i$	$O_i \geq q$		$O_k < q$ for all $k \in S_i$	$O_i \geq q$	$O_i < q$
MS15's definitions						
Forecast condition	$F_i \geq q$			$F_i \geq q$	$F_i < q$	$F_i < q$
Observation condition	$O_k \geq q$ for some $k \in S_i$			$O_k < q$ for all $k \in S_i$	$O_k \geq q$ for some $k \in S_i$	$O_k < q$ for all $k \in S_i$
NM definitions						
Forecast condition	$F_k \geq q$ for some $k \in S_i$			$F_k \geq q$ for some $k \in S_i$	$F_k < q$ for all $k \in S_i$	$F_k < q$ for all $k \in S_i$
Observation condition	$O_k \geq q$ for some $k \in S_i$			$O_k < q$ for all $k \in S_i$	$O_k \geq q$ for some $k \in S_i$	$O_k < q$ for all $k \in S_i$

b. *MS15's definitions*

Whereas C10 employed neighborhoods for both forecasts and observations, MS15 only considered neighborhoods for observations. Quoting MS15:

> If the precipitation at the [forecast] grid point is at or above a given threshold, a hit is recorded if the [observed] precipitation at any grid point in the [neighborhood] is at or above the threshold. A false alarm is recorded if there is no [observed] precipitation at or above the threshold. If the precipitation at the [forecast] grid point is below a given threshold, a correct negative is recorded if there are no grid points in the [neighborhood] with [observed] precipitation at or above the threshold, but a miss is recorded if the [observed] precipitation at any grid point in the [neighborhood] is at or above the threshold.

Therefore, like C10, MS15 classified the ith point based on event occurrence directly at i and defined forecast and observed events over disparate spatial scales: observed events were defined over scales larger than the grid scale but forecast events were defined at the grid scale.

c. *NM definitions*

Table 1 can also be populated using an NM approach (e.g., Sobash et al. 2011; Ben Bouallègue and Theis 2014; Barthold et al. 2015), in which event occurrence at i is *redefined* based on whether the maximum value within the neighborhood of i satisfies event criteria. Alternatively, the NM method is implemented by determining whether an event occurs *anywhere* within the neighborhood of i: hits are recorded if both forecast and observed events occur *anywhere* in the neighborhood, false alarms are recorded if a forecast event occurs anywhere in the neighborhood but no observed events occur within the neighborhood, missed events are recorded when no forecast events occur within the neighborhood but an observed event occurs somewhere within the neighborhood, and correct negatives are recorded when neither observed nor forecast events occur within the neighborhood (Table 2).

Contrasting both C10's and MS15's methods, NM definitions consistently define both forecast and observed events over spatial scales larger than the grid length and always query forecast and observed neighborhoods to classify the ith point; i can be placed into any quadrant of Table 1 even if both forecast and observed events do not occur at i. Further, whereas MS15's and C10's definitions are more complicated to interpret because they define events over inconsistent spatial scales, the scale consistency of the NM definitions yields a straightforward interpretation of what the NM approach measures: correspondence of forecast and

observed events within a neighborhood about *each* grid point.

d. Synthesis

1) Comparing C10's and MS15's definitions

C10's definitions produce more hits than MS15's, as the former has two criteria for hits and the latter just one (Table 2). Additionally, C10's method yields more correct negatives than MS15's because MS15 required observed nonevents everywhere within neighborhoods, which occurs less frequently than requiring an observed nonevent at just one point, as in C10. Regarding missed events, as MS15's forecast and observation conditions are easier to satisfy than C10's, relative to C10's approach, MS15's produces more missed events.

2) Comparing NM definitions to C10's and MS15's

Unlike C10's and MS15's definitions, NM criteria are unconcerned with values directly at i and occurrence of observed and forecast events is always determined by querying neighborhoods. Accordingly, NM criteria produce the most hits and fewest correct negatives. Regarding false alarms, the three variants have identical observation criterion (Table 2) but forecast events most easily occur under the NM definition, which therefore yields the most false alarms. Finally, for missed events, relative to C10, the NM method has an identical forecast condition but a more easily satisfied observation criterion, and, thus, more misses. Using similar reasoning, NM and MS15's definitions have identical observation requirements for misses, but MS15's forecast condition is more easily satisfied, so the NM approach produces fewer missed events than MS15's method.

3) Summary

Letting a_x, b_x, c_x, and d_x denote the respective numbers of elements in Table 1's quadrants, where subscript x represents a particular variant used to populate Table 1 (C10, MS15, or NM), the relationship between the methods is

$$a_{\mathrm{MS15}} \le a_{\mathrm{C10}} \le a_{\mathrm{NM}}, \tag{6}$$

$$b_{\mathrm{MS15}} = b_{\mathrm{C10}} \le b_{\mathrm{NM}}, \tag{7}$$

$$c_{\mathrm{C10}} \le c_{\mathrm{NM}} \le c_{\mathrm{MS15}}, \quad \text{and} \tag{8}$$

$$d_{\mathrm{NM}} \le d_{\mathrm{MS15}} \le d_{\mathrm{C10}}. \tag{9}$$

As more hits and correct negatives (false alarms and misses) indicate better (poorer) forecasts, compared to C10's definitions, the NM criteria are advantaged by producing the most hits but disadvantaged otherwise. Furthermore, relative to C10's approach, MS15's penalizes forecasts by yielding fewer hits and correct negatives and more misses. Figure 1 illustrates how the three methods classify points for a hypothetical case with circular neighborhoods; differences comply with Eqs. (6)–(9).

Moreover, considering Table 2, as r increases, "for all" ("for some") conditions become harder (easier) to satisfy. Therefore, aside from NM misses and false alarms, which possess both for some and for all conditions, Table 2 provides information about how contingency table elements change as r increases (Table 3). The most important difference is with regard to missed events, which decrease under C10's definition but increase under MS15's as neighborhoods expand.

3. Application of the three methods

a. Forecast model and methodology

Forecasts from a single member of NCAR's experimental, real-time, 10-member ensemble (Schwartz et al. 2015) initialized daily at 0000 UTC between 7 April and 31 December 2015 (269 forecasts) were used to illustrate differences between the three methods of populating neighborhood-based contingency tables. As described in Schwartz et al. (2015), the forecasts had 3-km horizontal grid spacing, spanned the conterminous United States (CONUS), were produced with version 3.6.1 of the WRF-ARW model (Skamarock et al. 2008), were initialized by downscaling 15-km ensemble adjustment Kalman filter (Anderson 2001, 2003) analyses onto the 3-km computational domain, and used lateral boundary conditions from NCEP's Global Forecast System.

Hourly accumulated precipitation forecasts were verified against corresponding NCEP stage IV (ST4) observations (Lin and Mitchell 2005) over a domain spanning most of the central CONUS. For comparison with ST4 data, precipitation forecasts were interpolated to the ST4 grid (~4.763-km horizontal grid spacing) using a budget interpolation method (Accadia et al. 2003). Contingency table elements were summed over all 269 forecasts to produce aggregate statistics, and circular neighborhoods (e.g., Fig. 1a) were used.

As bias impacts contingency table scores (e.g., Baldwin and Kain 2006), precipitation forecasts were bias corrected with a probability-matching approach (Ebert 2001) described by C10. Statistical significance was assessed with a bootstrap resampling technique

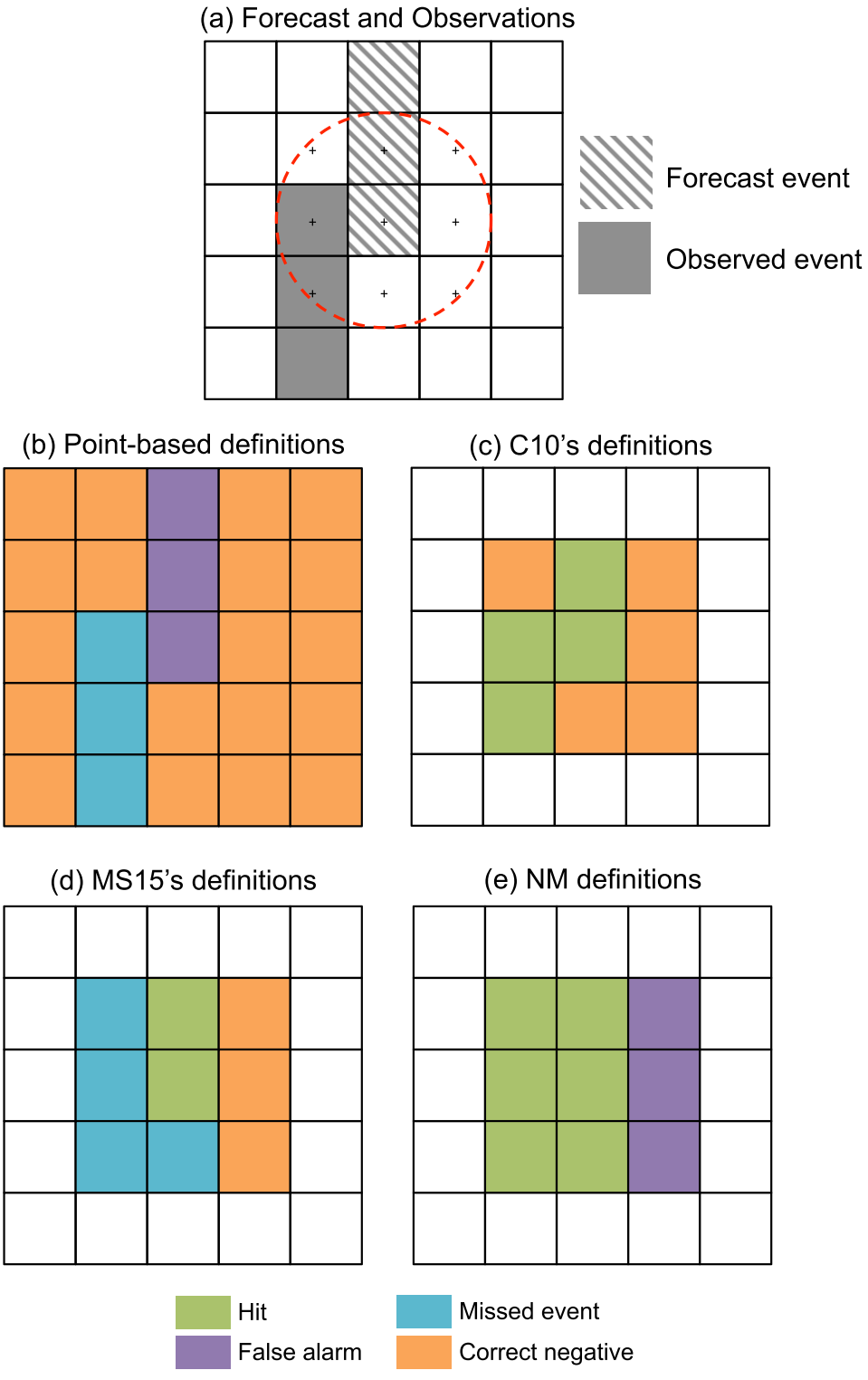

FIG. 1. Hypothetical (a) forecast and corresponding observations, where forecast (observed) events have occurred in hatched (filled) grid boxes, and contingency table (Table 1) classifications of the grid boxes based on (a) using (b) traditional point-by-point definitions and (c) C10's, (d) MS15's, and (e) NM definitions that leverage a neighborhood approach. For (c)–(e) the neighborhood is a circle with radius 1.5 times the horizontal grid spacing centered on each grid point, as illustrated in (a), where the dashed red circle denotes the neighborhood about the central grid point and grid boxes within the circular neighborhood are denoted with black plus signs (+). Note that classification of grid points in the outer rings of (c)–(e) was not possible, since the circular neighborhoods fall outside the grid.

TABLE 3. Expected variation of contingency table elements as the radius of influence (r) increases.

	Quadrant of Table 1			
	a (hit)	b (false alarm)	c (missed event)	d (correct negative)
C10's definitions	Increase	Decrease	Decrease	Constant
MS15's definitions	Increase	Decrease	Increase	Decrease
NM definitions	Increase	Unclear	Unclear	Decrease

(Hamill 1999) using 1000 resamples to compute bounds of 95% confidence intervals.

b. Results

Total hits, misses, false alarms, and correct negatives over all 269 twenty-four-h forecasts of 1-h accumulated precipitation (Figs. 2a–d and 2i–l) confirmed Eqs. (6)–(9) and Table 3, including expectations that misses and false alarms may not monotonically vary with r using NM definitions (Figs. 2b,c). Biases varied little with r under NM and C10's definitions, while MS15's approach yielded biases $\ll 1$ (Figs. 2e,m and 3a–c) that decreased with

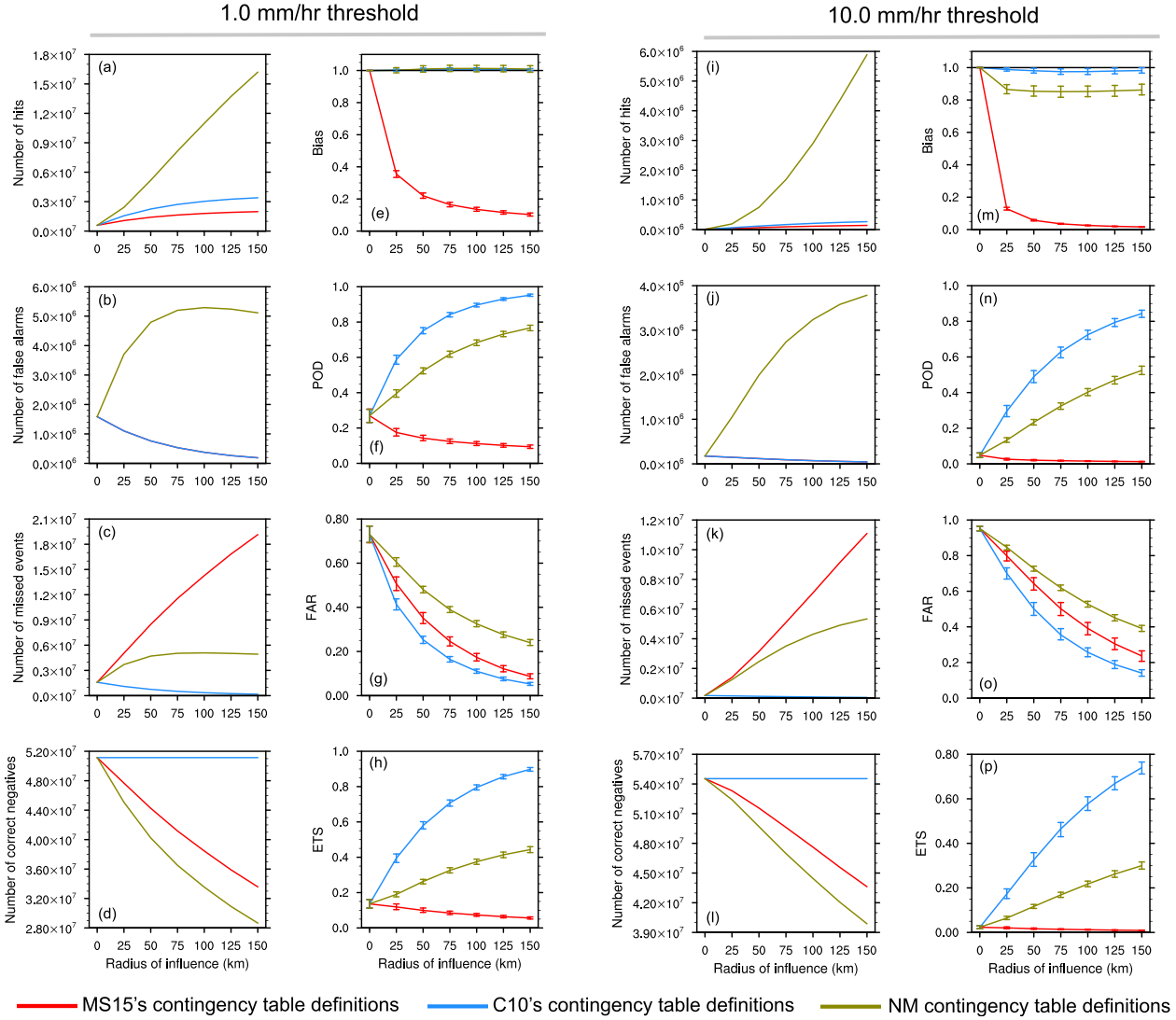

FIG. 2. Total number of (a) hits, (b) false alarms, (c) missed events, and (d) correct negatives, as well as the aggregate (e) bias, (f) POD, (g) FAR, and (h) ETS as a function of radius of influence (km) based on 24-h forecasts of 1-h accumulated precipitation over all 269 forecasts for the 1.0 mm h^{-1} event threshold. (i)–(p) As in (a)–(h), but for the 10.0 mm h^{-1} event threshold. Error bars indicate 95% confidence intervals.

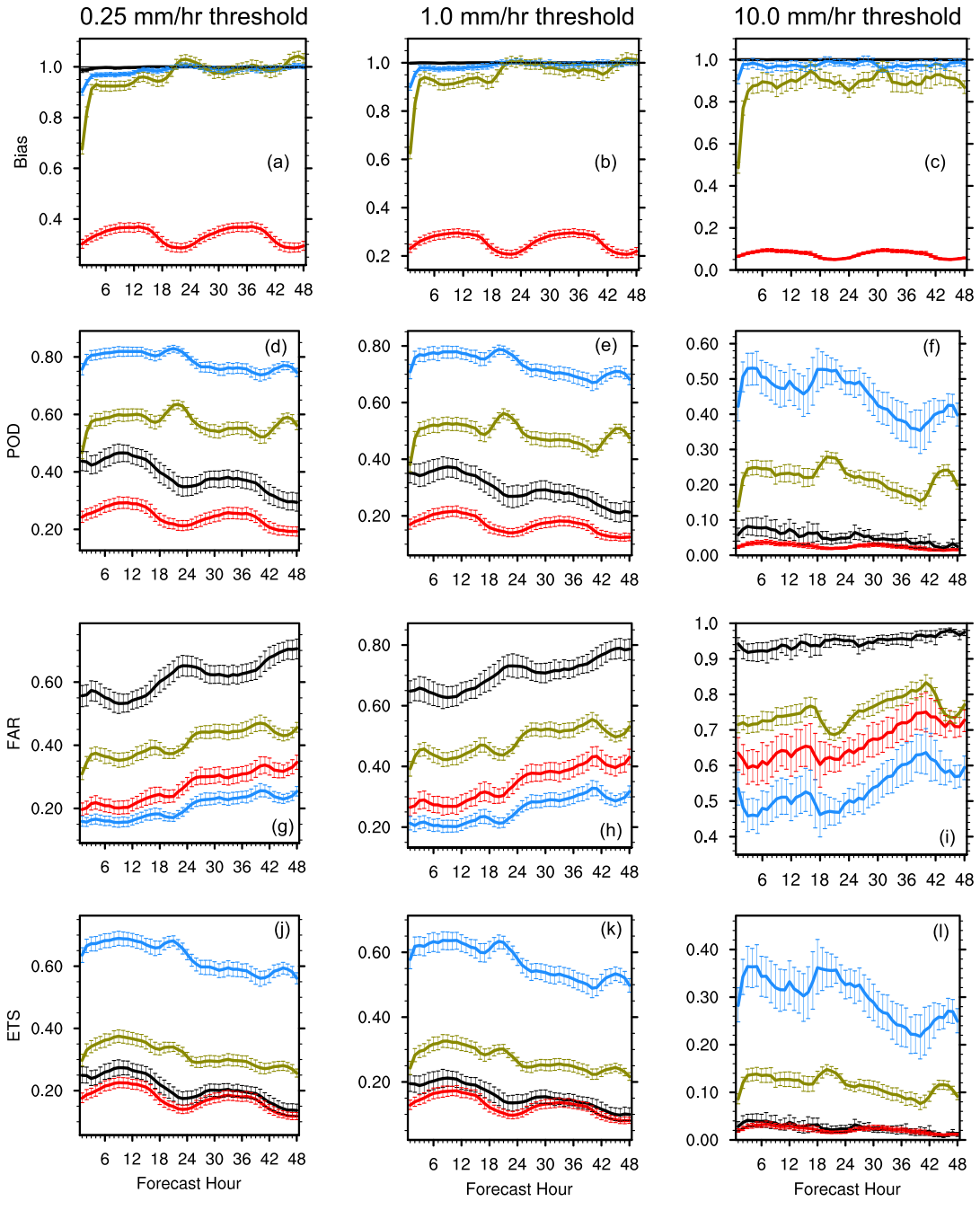

FIG. 3. (a)–(c) Bias, (d)–(f) POD, (g)–(i) FAR, and (j)–(l) ETS aggregated over all 269 forecasts as a function of forecast hour for 1-h precipitation accumulation event thresholds of (left) 0.25, (center) 1.0, and (right) 10.0 mm h^{-1} computed with various contingency table definitions. Statistics computed with NM, C10's, and MS15's criteria used r = 50 km, and error bars indicate 95% confidence intervals.

increasing r, consistent with Table 3 and Eq. (3). C10's definitions produced the highest ETSs and PODs, followed by NM and MS15's criteria (Figs. 2f,h,n,p, 3d–f, and 3j–l). Furthermore, C10's method yielded the lowest FARs, followed by MS15's and NM definitions (Figs. 2g, o and 3g–i), and FARs decreased with increasing r using all definitions. Similar results were obtained for different accumulation thresholds, r, and forecast hours not shown in Figs. 2 and 3.

However, while PODs and ETSs increased with r for NM and C10's criteria, using MS15's method, PODs and ETSs *decreased* with increasing r (Figs. 2f,h,n,p), which is undesirable and counters intuition that forecast quality should improve as neighborhoods enlarge, yet is consistent with MS15's definitions (Tables 2 and 3). Furthermore, C10 and MS15 themselves reached similar conclusions: C10 found their $\mathrm{ETS}_{\mathrm{neigh}}$ monotonically increased with r but MS15 noted their $\mathrm{ETS}_{\mathrm{neigh}}$ did not always increase as neighborhoods expanded.

Clearly, different contingency table definitions provided varying conclusions regarding forecast quality. Although application of C10's criteria indicated excellent forecast quality, C10's definitions mean fewer false alarms and misses as r increases (Figs. 2b,c,j,k), which may be inappropriate. Conversely, while MS15's definitions suggested poor forecast quality, MS15's criteria for misses overly penalizes forecasts and contributes to counterintuitive ETS and POD trends as r increases. Additionally, that biases decrease with r under MS15's definitions is undesirable.

NM definitions represent a compromise by typically producing scores between those given by MS15's and C10's methods. Additionally, NM definitions permit the possibility of increased false alarms (but not necessarily FARs) as neighborhoods increase, which, while a drawback, is nonetheless intuitive, whereas C10's method always yields fewer false alarms as r increases (Figs. 2b,j). Moreover, unlike MS15's criteria, the NM approach always indicates forecast improvement as r increases. Finally, NM event definitions have consistent spatial scales, contrasting MS15's and C10's event definitions that selectively consider neighborhoods (Table 2), and therefore, the NM approach may provide the fairest definitions to populate Table 1.

4. Summary

This paper applied three variations of populating neighborhood-based 2×2 contingency tables to deterministic 3-km forecasts and revealed statistically significant differences regarding forecast quality. Of the three flavors, C10's method appeared too lenient and MS15's too harsh, while the NM approach offered a compromise between the other two, produced expected behaviors as neighborhoods enlarged, and used consistent event definitions that required searching both forecast and observed neighborhoods to populate all contingency table quadrants. However, the primary drawback of the NM approach is potentially inflated FARs.

Overall, these findings indicate neighborhood-based contingency tables should be carefully populated to ensure forecasts are fairly evaluated and forecast verification metrics vary intuitively with r. Additionally, model intercomparisons examining neighborhood-based contingency table metrics must ensure common definitions are used across all forecast sets.

Although each method discussed here has limitations, of the three, the NM criteria are probably most appropriate, as they have consistent event definitions and yield well-behaved metrics. But, ultimately, regardless of how neighborhood-based contingency tables are filled, including with other potential definitions not discussed here, authors should explicitly state their contingency table definitions to foster clarity and interpretation of results.

Acknowledgments. Thanks to Morris Weisman and Ryan Sobash (NCAR/MMM) for internally reviewing this manuscript and Elizabeth Ebert and three anonymous reviewers for their comments. This work was partially supported by the Short Term Explicit Prediction (STEP) program.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, doi:10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, doi:10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

——, 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634–642, doi:10.1175/1520-0493(2003)131<0634:ALLSFF>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, doi:10.1175/WAF933.1.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866, doi:10.1175/BAMS-D-14-00201.1.

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, doi:10.1002/met.1435.

Clark, A. J., W. A. Gallus Jr., and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, doi:10.1175/2010WAF2222404.1.

——, A. MacKenzie, A. McGovern, V. Lakshmanan, and R. A. Brown, 2015: An automated, multiparameter dryline identification algorithm. *Wea. Forecasting*, **30**, 1781–1794, doi:10.1175/WAF-D-15-0070.1.

Dahl, N., and M. Xue, 2016: Prediction of the 14 June 2010 Oklahoma City extreme precipitation and flooding event in a multiphysics multi-initial-conditions storm-scale ensemble forecasting system. *Wea. Forecasting*, **31**, 1215–1246, doi:10.1175/WAF-D-15-0116.1.

Done, J., C. A. Davis, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, doi:10.1002/asl.72.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi:10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

——, 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:10.1002/met.25.

Fierro, A. O., A. J. Clark, E. R. Mansell, D. R. MacGorman, S. Dembek, and C. Ziegler, 2015: Impact of storm-scale lightning data assimilation on WRF-ARW precipitation forecasts during the 2013 warm season over the contiguous United States. *Mon. Wea. Rev.*, **143**, 757–777, doi:10.1175/MWR-D-14-00183.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. Preprints, *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at http://ams.confex.com/ams/pdfpapers/83847.pdf.]

Lynn, B. H., G. Kelman, and G. Ellrod, 2015: An evaluation of the efficacy of using observed lightning to improve convective lightning forecasts. *Wea. Forecasting*, **30**, 405–423, doi:10.1175/WAF-D-13-00028.1.

Ma, L.-M., and X.-W. Bao, 2016: Parametrization of planetary boundary-layer height with helicity and verification with tropical cyclone prediction. *Bound.-Layer Meteor.*, **160**, 569–593, doi:10.1007/s10546-016-0156-7.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

McMillen, J. D., and W. J. Steenburgh, 2015: Capabilities and limitations of convection-permitting WRF simulations of lake-effect systems over the Great Salt Lake. *Wea. Forecasting*, **30**, 1711–1731, doi:10.1175/WAF-D-15-0017.1.

Pytharoulis, I., S. Kotsopoulos, I. Tegoulias, S. Kartsios, D. Bampzelis, and T. Karacostas, 2016: Numerical modeling of an intense precipitation event and its associated lightning activity over northern Greece. *Atmos. Res.*, **169**, 523–538, doi:10.1016/j.atmosres.2015.06.019.

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

Schumacher, R. S., A. J. Clark, M. Xue, and F. Kong, 2013: Factors influencing the development and maintenance of nocturnal heavy-rain-producing convective systems in a storm-scale ensemble. *Mon. Wea. Rev.*, **141**, 2778–2801, doi:10.1175/MWR-D-12-00239.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., doi:10.5065/D68S4MVH.

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, doi:10.1175/WAF-D-10-05046.1.

Squitieri, B. J., and W. A. Gallus Jr., 2016: WRF forecasts of Great Plains nocturnal low-level jet-driven MCSs. Part I: Correlation between low-level jet forecast accuracy and MCS precipitation forecast skill. *Wea. Forecasting*, **31**, 1301–1323, doi:10.1175/WAF-D-15-0151.1.

Weisman, M. L., C. A. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437, doi:10.1175/2007WAF2007005.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.