Statistical Evaluation of Different Surface Precipitation-Type Algorithms and Its Implications for NWP Prediction and Operational Decision-Making

HEATHER DAWN REEVES,[a,b] DANIEL D. TRIPP,[a,b] MICHAEL E. BALDWIN,[a,b] AND ANDREW A. ROSENOW[a,b]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/National Severe Storms Laboratory, Norman, Oklahoma*

(Manuscript received 9 May 2023, in final form 20 September 2023, accepted 21 September 2023)

ABSTRACT: Several new precipitation-type algorithms have been developed to improve NWP predictions of surface precipitation type during winter storms. In this study, we evaluate whether it is possible to objectively declare one algorithm as superior to another through comparison of three precipitation-type algorithms when validated using different techniques. The apparent skill of the algorithms is dependent on the choice of performance metric—algorithms can have high scores for some metrics and poor scores for others. It is also possible for an algorithm to have high skill at diagnosing some precipitation types and poor skill with others. Algorithm skill is also highly dependent on the choice of verification data/methodology. Just by changing what data are considered "truth," we were able to substantially change the apparent skill of all algorithms evaluated herein. These findings suggest an objective declaration of algorithm "goodness" is not possible. Moreover, they indicate that the unambiguous declaration of superiority is difficult, if not impossible. A contributing factor to algorithm performance is uncertainty of the microphysical processes that lead to phase changes of falling hydrometeors, which are treated differently by each algorithm, thus resulting in different biases in near −0°C environments. These biases are evident even when algorithms are applied to ensemble forecasts. Hence, a multi-algorithm approach is advocated to account for this source of uncertainty. Although the apparent performance of this approach is still dependent on the choice of performance metric and precipitation type, a case-study analysis shows it has the potential to provide better decision support than the single-algorithm approach.

SIGNIFICANCE STATEMENT: Many investigators are developing new-and-improved algorithms to diagnose the surface precipitation type in winter storms. Whether these algorithms can be declared as objectively superior to existing strategies is unknown. Herein, we evaluate different methods to measure algorithm performance to assess whether it is possible to state one algorithm is superior to another. The results of this study suggest such claims are difficult, if not impossible, to make, at least not for the algorithms considered herein. Because algorithms can have certain biases, we advocate a multi-algorithm approach wherein multiple algorithms are applied to forecasts and a probabilistic prediction of precipitation type is generated. The potential value of this is demonstrated through a case-study analysis that shows promise for enhanced decision support.

KEYWORDS: Algorithms; Numerical analysis/modeling; Forecast verification/skill

1. Introduction

Correct prediction of surface precipitation type has obvious importance for winter storms. Currently, precipitation type is determined within the National Weather Service (NWS) using post-processing algorithms that are applied to NWP output. The original suite of algorithms developed for this purpose has fairly simple logic that uses properties of the vertical temperature profiles, such as the depths and temperatures of elevated warm layers or surface-based subfreezing layers to determine the phase (Ramer 1993; Baldwin et al. 1994; Bourgouin 2000). These have been in use since the early 2000s. While efficient and easy to understand, they struggle with differentiating freezing rain (FZRA) from ice pellets (PL; Bourgouin 2000; Manikin et al. 2004; Manikin 2005; Wandishin et al. 2005; Reeves et al. 2014; Reeves 2016), motivating several recent efforts to develop improved algorithms (e.g., Benjamin et al. 2016; Reeves et al. 2016; Birk et al. 2021; Harrison et al. 2022; Filipiak et al. 2023). This flush of research raises important questions, namely, what is the best method to evaluate the performance of precipitation-type algorithms and are there special considerations to an algorithm-performance assessment that could influence its use by decision-makers? The aim of this paper is to address these questions.

There are now multiple approaches used in precipitation-type algorithms. Some are simple upgrades to the first-generation algorithms used by the NWS wherein decision points that led to poor performance have been improved (Birk et al. 2021; Lu et al. 2021). Some investigators use machine learning to address this problem (e.g., Harrison et al. 2022; Filipiak et al. 2023). Recognizing that precipitation type can be influenced by microphysical controls such as riming and partial melting/refreezing, some researchers have developed algorithms that incorporate microphysical controls, such as predicted mixing ratios or that use bin microphysics to explicitly compute the phase (Benjamin et al. 2016; Reeves et al. 2016; Gascón et al. 2018; Cholette et al. 2020). Regardless of the approach, the obvious goal of all these investigators is to create an algorithm with good performance metrics that perhaps even bests the efforts of others. These two ideas are referred to herein as "goodness," which is an absolute measure of algorithm performance with respect to performance metrics, and "betterness," which is the superior performance of one algorithm over another as determined using performance metrics (Murphy 1993).

Corresponding author: Heather Reeves, heather.reeves@noaa.gov

DOI: 10.1175/WAF-D-23-0081.1

There is no clear guidance on best practices for evaluating a precipitation-type algorithm. As a result, different metrics are often used and applied to very different datasets. For example, Reeves et al. (2014) used observed sounding data to verify several algorithms and included only events where the phase was unchanged during the 40 min following sounding launch times. Others have used model soundings as input and validated against the observed precipitation type at the top of the hour (Ikeda et al. 2013; Benjamin et al. 2016). Some investigators use the Automated Surface Observation Station (ASOS; NOAA 1998) network as ground truth (e.g., Ikeda et al. 2013; Reeves et al. 2014) and others have used crowd-sourced data from the meteorological Phenomena Identification Near the Ground (mPING) project (e.g., Elmore et al. 2015; Harrison et al. 2022). Influences of the choice of input and ground truth data on algorithm verification were explored in Reeves (2016) who found that these choices can have a profound impact on the apparent performance of algorithms. An unexplored facet to precipitation-type verification is the choice of performance metrics. These also vary between investigators with the most popular being the probability of detection (POD), false alarm ratio (FAR), and one or more measures of skill such as the critical success index (CSI) or Heidke skill score (HSS). Each of these represents a unique way of defining goodness, and it is unclear what metrics are most appropriate for precipitation-type evaluation. Hence, it is unknown whether different conclusions may be drawn about the goodness and betterness of algorithms based on the choice of verification metric. How this might dampen or amplify dependencies on the choice of ground truth, as discussed in Reeves (2016), is also unknown.

The potentially different conclusions that varying metrics may yield regarding goodness and betterness raise several questions: What performance metric(s) is/are relevant for precipitation-type verification? If more than one metric is required to fully appreciate the performance of an individual algorithm, what happens when intercomparing algorithms and the different metrics suggest different "winners" or the results change depending on what phase is being evaluated? Is it possible to alter the outcomes by making honest and defendable choices in how the verification data are processed and defined? And last, is a competition to declare a winner appropriate or could the diversity of solutions provided by an array of algorithms be of greater operational utility? These questions are addressed herein through comparison of three modern precipitation-type algorithms when verified using different performance metrics and datasets.

2. Methodology

a. *Algorithms used for the evaluation*

Three modern precipitation-type algorithms are compared in this paper. These are the operational method used by the HRRR model (HP; Benjamin et al. 2016), the modified Bourgouin algorithm (MB; Bourgouin 2000; Birk et al. 2021), and the Spectral-Bin Classifier (SBC; Reeves et al. 2016). The HP output is directly accessed via the archived gridded data available through Amazon Web Services. Because the HRRR model only outputs precipitation type for forecasts and not analyses, 1-h HRRR forecasts are used. The HP precipitation type is accessed directly from the HRRR archive, while MB and SBC are computed using temperature and humidity profiles from the 1-h forecast pressure coordinate data. [Sensitivity experiments using the data in its native, terrain-following coordinates do not change the results of this study (not shown).] Hence, all three algorithms have equitable thermodynamic input.

TABLE 1. The categories of precipitation diagnosed by each precipitation-type algorithm evaluated in this paper.

Type	HP	MB	SBC
RA	✓	✓	✓
SN	✓	✓	✓
FZRA	✓	✓	✓
PL	✓	✓	✓
RASN	✓	✓	✓
RAPL	✓	✓	✓
RAPLSN	✓	✓	
PLSN	✓	✓	✓
FZRASN	✓	✓	
FZRAPL	✓	✓	✓
FZAPLSN	✓	✓	
DZ	✓		
DZSN	✓		
DZPL	✓		
DZPLSN	✓		
FZDZ	✓		
FZDZSN	✓		
FZDZPL	✓		
FZDZPLSN	✓		

The methodology used by each of the above algorithms is quite different. HP uses the HRRR-predicted mixing ratios in concert with the precipitation rate and 2-m temperature to diagnose the phase. MB uses properties of the thermal profile, such as the area between the 0°C isotherm and the elevated warm layer. SBC uses a stripped-down bin microphysics scheme to explicitly predict the liquid-water fraction (LWF) of falling hydrometeors, which it uses along with the wet-bulb temperature (T_w) to assign the phase. The manner in which all three of these algorithms were developed and initially evaluated is independent of the data being used for input/verification herein.

A minor modification to MB is required for this study. In its original form, this algorithm diagnoses probabilities of each of the four cardinal classes of rain (RA), snow (SN), FZRA, and PL. To bring it into alignment with the deterministic declarations of HP and SBC, MB is modified so that it also makes deterministic declarations. Any phase whose probability is greater than 50% is assumed to be a constituent of a deterministic category. Several modifications are also made to the SBC to address observed shortcomings as described in section 2b.

The types of categories diagnosed by each algorithm are provided in Table 1. They are all capable of diagnosing the four cardinal categories of RA, SN, FZRA, and PL as well as several mixes. All three attempt to discriminate between SN and nonclassical FZRA (i.e., FZRA that occurs in a completely subfreezing profile due to a lack of ice nucleation). The HRRR model has a layer of postprocessing that shifts some RA and FZRA diagnoses into their corresponding drizzle (DZ) category that is based on the precipitation rate. These categorical shifts are not

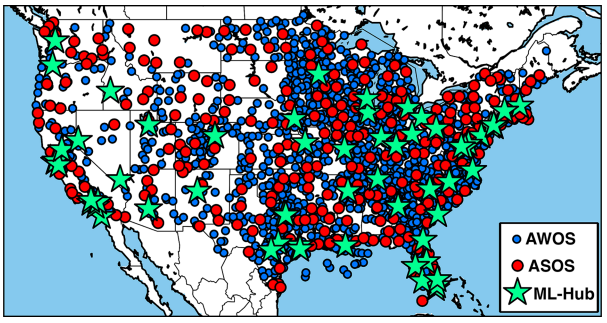

FIG. 1. The locations of all precipitation-type observing station networks used in this study. ML-Hub refers to medium- and large-hub airports that have an ASOS installation.

archived as a part of the HRRR dataset. SBC has also had some adaptations made that allow it to diagnose DZ categories (Reeves et al. 2022), but these modifications currently rely on radar observations and, therefore, are not used herein. MB does not have a capability to diagnose DZ. Since DZ is not diagnosed by all algorithms, it is not included in this study.

All three of the algorithms have certain conditions under which no phase is declared because the profile is assumed to be too dry for precipitation (or not enough precipitation occurred). In HP, this is based on a minimum hourly precipitation rate. MB uses a layer-averaged relative humidity (RH), and SBC uses a combined dewpoint temperature depression (T_{dd}) and RH as described in section 2b. Hence, the algorithms are different in this regard and have a different number of unclassified observations. HP has the highest number of these occurrences. Of particular note is SN—about 9% of these occurrences have no diagnosis. MB misses about 4% of all SN occurrences and smaller amounts of RA and RASN. SBC misses less than 1% of any one category. All statistics in this paper only include observation/forecast pairs where there is a diagnosis made by all algorithms.

b. Modifications made to the SBC

Three alterations are made to the SBC's logic to improve its discrimination of SN. First, the so-called precipitation top (P_{TOP}), which is used to declare the starting phase of hydrometeors at the top of the column, is modified to better discriminate between SN and nonclassical FZRA. In the modification, P_{TOP} is set to be the top of the bottom-most layer, where T_{dd} is less than or equal to 6°C and RH is greater than 60%. If this condition is never met, the algorithm declares P_{TOP} as the top-most layer with RH greater than 80%. If neither condition is met, no precipitation type is diagnosed. Second, the method for discriminating nonclassical FZRA from SN is modified to account for potential ice nucleation below P_{TOP} for soundings whose P_{TOP}s are greater than the assumed ice-nucleation temperature of −6°C. If the minimum T_w between the surface and 3 km above ground level is less than ice nucleation, SN is diagnosed for these kinds of profiles. Last, the integrated LWF threshold to discriminate between RASN and SN was increased from 15% to 60%. All of these changes were made in response to observed deficiencies in the SBC's performance uncovered in the process of conducting this research.

c. Verification data

The ground truth used in this paper is from the ASOS network at commercial airports and the Automated Weather Observation System (AWOS) level III/IV stations (Fig. 1). Both ASOS and AWOS-III/IV are capable of automatically detecting RA and SN. Nearly all of the ASOS sites used in this study have a FZRA sensor and/or have Contract Weather Observer (CWO) support to curate the automated reports. Categories of precipitation besides RA, SN, and FZRA (e.g., FZDZ, mixes, and PL) are only reported when a CWO augments the report (NOAA 1998). Only some AWOS sites have the capability to detect FZRA. Different combinations of ASOS and AWOS are used in the various sensitivity tests performed in section 3.

The verification period is from the 2016/17 to 2020/21 winter seasons (October–March) and includes 11 different categories of precipitation. While the ASOS archive we accessed has 5-min observations, most experiments use only the data at the top of the hour. Table 2 provides an accounting of each observation type at the top of the hour for the different sensitivity tests performed in this study. Since the High-Resolution

TABLE 2. The total number of each category of precipitation type for all experiments conducted herein. The columns labeled AP, LD, Jan17, QC, ML-Hub, and GS reference specific experiments described in section 3b.

Type	ASOS	AWOS	AP	LD	Jan17	QC	ML-Hub	GS
RA	241 833	216 069	512	170 246	9080	238 317	31 183	23 803
SN	117 708	210 906	1140	93 573	6377	113 109	9823	7846
FZRA	4651	300	838	2846	466	3703	367	133
PL	310	21	48	73	23	191	97	15
RASN	1781	—	46	—	73	1592	493	165
RAPL	454	—	49	—	28	159	144	13
RAPLSN	86	—	4	—	5	27	24	—
PLSN	555	14	71	50	18	199	157	7
FZRASN	200	—	52	—	7	186	50	14
FZRAPL	351	—	113	—	31	249	109	10
FZAPLSN	60	—	14	—	1	36	16	1
Total	376 396	427 010	2887	266 788	16 109	357 768	42 463	32 007

TABLE 3. Mathematical expressions for each of the performance metrics used in this paper.

Term	Equation
Probability of detection (POD)	$\frac{TP}{(TP + FN)}$
False alarm ratio (FAR)	$\frac{FP}{(TP + FP)}$
Success ratio (SR)	$\frac{TP}{(TP + FP)}$
Bias	$\frac{(TP + FP)}{(TP + FN)}$
Heidke skill score (HSS)	$\frac{2(TP \times TN - FN \times FP)}{(TP + FP)(FP + TN) + (TP + FN)(FN + TN)}$
False positive rate (FPR)	$\frac{FP}{(FP + TN)}$
Extremal dependence index (EDI)	$\frac{\log(FPR) - \log(POD)}{\log(FPR) + \log(POD)}$

Rapid Refresh (HRRR; Weygandt et al. 2009) data are used to initialize each algorithm, only times when archived HRRR data are available are included in Table 2. Observations that include graupel, snow grains, drizzle, freezing drizzle, or unknown precipitation type reports are discarded as these are not categories all of the algorithms diagnose. About two-thirds of the total number of ASOS observations are of RA. Refreezing precipitation types—FZRA, PL, and mixes that include these types—comprise only about 2% of the total number of ASOS observations, consistent with previous investigations (Reeves 2016; Landolt et al. 2019). By contrast, refreezing precipitation types comprise only 0.08% of all AWOS observations due to its limited sensing capabilities.

Table 2 also lists some of the various experiments performed herein. These are designed to evaluate how algorithm performance is impacted when changes are made to the verification dataset or how a "hit" is defined. The details of these experiments are discussed in section 3b.

d. *Performance metrics used herein*

Precipitation-type verification is usually performed using a contingency table wherein the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) are summed over all sites and times in the verification dataset. These four elements can be combined into a multitude of different indices, each one providing a unique definition of "goodness." An issue with verifying precipitation type is that some categories are very rare, leading to an overabundance of TNs relative to the other elements in the contingency diagram. As a consequence, performance metrics that use TNs can asymptote to their extrema if there are no measures taken to modulate its relative influence. Some metrics ignore TNs and, hence, do not have this problem. These include the POD, FAR, success ratio (SR), and bias (Table 3), which are functionally related. Hence, an evaluation of all four of these provides redundant information and two can be eliminated. Herein, we choose to eliminate FAR and SR, but this is a matter of preference. The overriding results of this study are not affected by this choice (not shown).

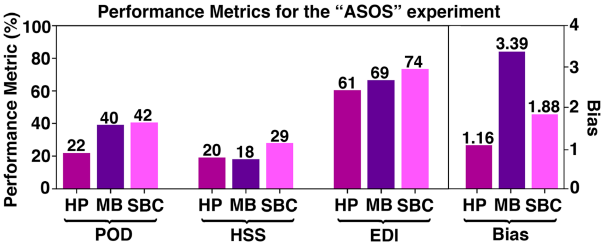

FIG. 2. The POD, HSS, EDI, and bias for each algorithm for refreezing categories of the ASOS experiment.

An additional metric that makes use of TNs is required to fully express all of the elements in the contingency table. As noted above, this can be problematic. Two metrics that regulate the impact of high TNs are considered in this study: these are HSS and the extremal dependence index (EDI; Ferro and Stephenson 2011; Table 3). In HSS, a TN appears in both the numerator and denominator, thus limiting its impacts on the overall score. EDI is a function of both the POD and the false positive rate (FPR), which is inversely proportional to TN, but its influence is tempered by taking the logarithm of both POD and FPR, which brings them into the same order of magnitude. A concern with HSS is that as an observed event becomes increasingly rare, it approaches 0, and, therefore, may not provide a reliable discrimination of either goodness or betterness. EDI, on the other hand, is independent of event rarity. But, because it is designed for rare events, it is unclear whether it can provide meaningful discrimination for more commonly observed phases like RA and SN. The reader may wonder about two other commonly used metrics: the Peirce skill score (PSS) and the critical success index (CSI). When there is a large TN relative to the other members of the contingency diagram, PSS approximates to POD. CSI uses the same three elements of the contingency diagram as POD, FAR, and bias and hence does not provide unique information about algorithm performance from POD and bias.

3. Intercomparison of the different algorithms' performance metrics

a. *Measures of algorithm goodness and betterness*

We start with a comparison of the POD, HSS, EDI, and bias for each algorithm using all ASOS refreezing categories (i.e., FZRA, PL, and mixes that include either of these forms) over the 5-yr retrospective. This evaluation assumes a hit occurs if and only if the diagnosed phase exactly agrees with the observation at the top of the hour. Hence, a diagnosis of FZRAPL is considered a miss when the observation is FZRA, even though the FZRA part of the diagnosis is correct. This experiment is referred to as "ASOS."

The results for ASOS are shown in Fig. 2 and highlight two important findings. First, the performance metrics disagree about the goodness of the algorithms. For example, HP has a very good bias, but its POD and HSS are quite low (22% and

TABLE 4. Performance metrics (expressed as percentages for POD, HSS, and EDI) broken down by algorithm and phase. The optimal values are given in boldface.

		POD	Bias	HSS	EDI
RA	HP	**97**	1.02	**89**	**97**
	MB	96	**1.01**	88	**97**
	SBC	96	**1.01**	**89**	**97**
SN	HP	**94**	**0.98**	**92**	**97**
	MB	80	0.83	82	90
	SBC	93	**0.98**	91	96
FZRA	HP	28	**0.59**	34	62
	MB	54	4.46	18	66
	SBC	**56**	1.76	**40**	**76**
PL	HP	6	1.61	5	41
	MB	8	**1.45**	**7**	46
	SBC	**21**	7.39	5	**53**

20%, respectively). Such disagreements are not surprising as algorithms can be strong in some respects and weak in others. HP's excellent bias indicates it produces refreezing precipitation at compatible rates to nature, but its low POD suggests potential flaws in its logic. MB and SBC have comparatively higher PODs, but their biases are also high, suggesting they overproduce these types of precipitation. This too suggests potential problems with logic.[1]

The second important finding from Fig. 2 is that the performance metrics disagree on betterness. The bias suggests HP is the best, while the other three metrics suggest SBC is the best, albeit by only marginal amounts. HSS and EDI also disagree on second place. HSS suggests HP is superior to MB, while the opposite is true for EDI. Hence, betterness is not able to be definitively declared.

In the preceding, all refreezing categories are combined into a single set of statistics, but the most common way algorithms are assessed is separately for each of the four cardinal categories of RA, SN, FZRA, and PL (Bourgouin 2000; Manikin et al. 2004; Manikin 2005; Wandishin et al. 2005; Reeves et al. 2014; Elmore et al. 2015; Reeves et al. 2016; Birk et al. 2021; Harrison et al. 2022; Filipiak et al. 2023). Evaluating the relative performance of algorithms this way introduces more ambiguities as goodness is strongly dependent not only on the choice of metric but also on the phase (Table 4). Consider MB—its PODs are competitive for both RA and FZRA, but it has a rather high bias for FZRA. So, it is unclear whether MB's results can be considered good. Betterness is even more problematic. Although SBC has the highest (or nearly the highest) POD for all categories, its comparatively high bias for PL makes a declaration of this algorithm as "best" a controversial one. While HP has the best bias for most categories, it too would be a controversial choice for best given its comparatively low PODs for FZRA and PL. One may wonder how the statistics in Fig. 2 and Table 4 are altered when the verification data are limited to the near −0°C range. This was tested by limiting observations to be within ±5°C. This reduces the PODs for RA by 10% for HP and MB and 11% for SBC and similarly reduces the other skill scores. Because almost all FZRA and PL observations occur within this range, these scores are not affected.

Before moving on, we stop to reflect on the relative value of HSS versus EDI as revealed in Table 4. HSS is more vulnerable to the bias than EDI. For example, MB has a comparatively low HSS for FZRA that is likely due to its high bias, whereas EDI appears to be more immune to the bias. This may lead one to favor HSS over EDI as a measure of skill. However, HSS has the problem of trending toward zero as the event frequency decreases. This is evident for the PL category, which has a comparatively small event frequency (Table 2). In spite of the fact that EDI is designed specifically for rare events, it does appear to give meaningful information about the relative performance for RA and SN. Since HSS and EDI both have different strengths and weaknesses, we continue to use them both moving forward.

One may wonder about comprehensive skill scores that provide a measure of the performance for all categories, such as the Gerrity skill score (GSS; Gerrity 1992). An advantage of the GSS is that it gives higher weighting to a less-common phenomenon, thus preventing RA and SN from monopolizing the score. When computed for all categories in Table 2, the GSSs for HP, MB, and SBC are 89%, 87%, and 87%, respectively. These are very close scores and do not convincingly indicate betterness on the part of any algorithm.

b. *Impacts of methodology on apparent goodness and betterness*

Some scores in Fig. 2 and Table 4 are close, leading one to question their sensitivity to choices made in the process of verification, such as the amount of time evaluated, the choice of how to declare a hit, and many others. Indeed, there are numerous ways the apparent goodness and sometimes betterness of algorithms can be altered by making reasonable choices about what data to include in the verification and how to define a TP (Reeves 2016). There is little consensus in the scientific community on what data should be included in a precipitation-type assessment or on how to define the elements in the contingency table. The experiments described below are conducted to evaluate how various approaches impact the verification metrics for refreezing precipitation. Each of these represent valid ways to verify a precipitation-type forecast. The number of observations in each of these experiments is included in Table 2. The performance metrics for each experiment for refreezing precipitation are provided in Table 5. All experiments are compared against the ASOS experiment, which is the same method described in section 3a and highlighted in Fig. 2.

1) IMPACTS OF CHANGING WHAT IS CONSIDERED A HIT

Two experiments are conducted that widen the aperture of what is considered a hit. In the first, "generous hit," a TP is declared if the observed phase at the top of the hour and the algorithm have at least one constituent in common. Such a

[1] We stress that low PODs may only indicate *potential* flaws in logic as model and observational error may have a nonnegligible influence on the low PODs in this assessment.

TABLE 5. The performance metrics (expressed as percentages for POD, HSS, and EDI) for refreezing categories for all experiments performed in this paper (see text for descriptions of each experiment) and all algorithms. The best score for each performance metric and experiment is indicated in boldface.

		POD	Bias	HSS	EDI
ASOS	HP	22	**1.16**	20	61
	MB	40	3.39	18	69
	SBC	**42**	1.88	**29**	**74**
Generous hit	HP	39	**24.01**	3	51
	MB	**56**	30.25	3	64
	SBC	**56**	26.84	**4**	**66**
Neighborhood approach	HP	62	**10.38**	**11**	78
	MB	**73**	17.84	7	**83**
	SBC	68	12.06	10	82
AWOS	HP	21	**3.45**	9	57
	MB	**40**	11.77	6	66
	SBC	**40**	4.31	**15**	**72**
Ambiguous profiles	HP	14	**1.35**	6	16
	MB	**36**	1.96	**19**	**41**
	SBC	30	2.42	11	29
Long duration	HP	31	**1.59**	24	69
	MB	57	5.59	17	80
	SBC	**60**	2.64	**33**	**84**
17 Jan	HP	28	**1.08**	27	64
	MB	49	2.48	28	74
	SBC	**50**	1.55	**39**	**77**
Quality control	HP	30	**1.19**	28	69
	MB	56	4.10	22	80
	SBC	**59**	1.74	**43**	**85**
ML-Hub	HP	15	**0.95**	15	53
	MB	23	2.61	13	55
	SBC	**26**	1.60	**20**	**61**
Gold standard	HP	22	**1.16**	20	61
	MB	37	3.51	16	68
	SBC	**41**	1.76	**30**	**74**

method was used in Reeves et al. (2014, 2016). This experiment has the same number of observations as ASOS (Table 2). Accordingly, generous hit has higher PODs than ASOS. However, the effectiveness of this approach is most clearly demonstrated using bias and HSS. The former shows values well over 1, indicating an extreme apparent over prediction of refreezing precipitation. The HSS scores suggest all algorithms barely perform better than random chance. EDI, because it is more immune to the bias, is only somewhat decreased, suggesting all algorithms are skillful.

A second experiment, "neighborhood approach," is conducted to evaluate how widening the aperture of a hit impacts interpretation of performance (e.g., Ebert 2009). In neighborhood approach, a TP is declared if any of the grid boxes in the 15 km surrounding each observation agree with the observation. An FP is declared if the phase in question is diagnosed in any of the surrounding grid boxes, but the phase was not observed. This experiment also has the same number of observations as ASOS (Table 2). The PODs for neighborhood approach are significantly increased relative to ASOS (Table 5). This suggests that the algorithms may suffer less from logic issues than from minor positional errors due to model uncertainty. It is interesting to note how differently the algorithms are impacted by this change. HP has a 40% increase in its POD, while SBC has only a 26% increase, suggesting HP may be more vulnerable to model error than the other methods. As with generous hit, the biases are quite high, HSSs are quite low, and EDI seems immune to excessive bias. While somewhat higher than in generous hit, the low HSSs suggest that these algorithms are barely improved relative to random chance and, as above, casts aspersions on this method of verification as an indicator of algorithm performance.

2) IMPACTS OF ENHANCING THE VERIFICATION DATASET

Two experiments are conducted that increase the number of observations for verification. In the "AWOS" experiment, both AWOS and ASOS are used for verification. This adds another 1327 sites to the list and increases the total number of observations by about 427 000 (Table 2). Because the archive we accessed for these data does not store 5-min observations, the top-of-the-hour precipitation type is assumed to be the reported type that occurs most closely to the top of the hour provided it occurs within 15 min of the top of the hour. AWOS are known to underreport refreezing precipitation (i.e., FZRA and PL), which leads to increased biases and decreased HSSs relative to ASOS (Table 5). This experiment serves to underscore an important point: A bias in the observations can exaggerate or reduce apparent biases in the algorithms making them appear better or worse than they actually are. Notice that MB's bias is significantly increased in AWOS. This algorithm already had an elevated bias for refreezing precipitation. Verifying against a dataset that has limited sensing capabilities for this form of precipitation only serves to inflate this bias, suggesting the algorithm is worse in this regard than it actually is. A similar finding was reported in Reeves (2016). While one can easily quantify the observational bias in each of these experiments relative to the ASOS observations, the true bias in the ASOS observations is not fully known. There may be biases in the ASOS dataset that impact our ability to rightly declare goodness and betterness.

3) IMPACTS OF ENVIRONMENTAL UNCERTAINTY

Two experiments are conducted that test the sensitivity of algorithm performance to environmental uncertainty. The first of these, dubbed "ambiguous profiles" (or AP in Table 2) limits the observation dataset to only those cases that are considered ambiguous. Herein, ambiguity is defined as an observation whose T_w profile has an elevated warm layer with a maximum T_w less than or equal to 1.5°C and a surface-based cold layer with a minimum T_w greater than −6°C. Melting and refreezing in such profiles are strongly dependent on microphysical controls such as the degree of riming, evaporation/sublimation, particle size distribution, and hydrometeor interactions (Reeves et al. 2016; Carlin et al. 2021; Reeves et al. 2022). Therefore, these profiles are more challenging to diagnose. Accordingly, the POD, HSS, and EDI decrease and the biases slightly increase (Table 5).

In the next experiment, "long duration" (LD in Table 1), the verification is restricted to only those times when the

observed precipitation type does not change during the 60-min following the top of the hour, similar to Reeves et al. (2014) and Reeves (2016). An additional restriction is added to include only hours where there is precipitation at the top of the hour, so algorithm performance is not injured by a dry bias in nonprecipitating soundings. Hence, the observations in this experiment are less ambiguous in terms of the temporal consistency. The POD, HSS, and EDI are improved relative to ASOS, but the biases are increased, suggesting a heightened overprediction of refreezing forms on the part of all algorithms (Table 5).

4) IMPACTS OF CHANGING THE DURATION OF THE EVALUATION PERIOD

Changing the time aperture can also impact the statistics. This is demonstrated in the "17 January" (17 Jan in Table 2) experiment, which is identical to ASOS except that only data from the month of January 2017 are used. This was a very active month for winter precipitation (Table 2). Limiting the evaluation to a smaller time range is consistent with some previous investigations (Ikeda et al. 2013; Benjamin et al. 2016). The performance metrics in 17 January are mostly improved relative to ASOS (Table 5). However, this is not the case with every individual month—some months have similar or even degraded statistics relative to ASOS, thus underscoring the fickleness of using shorter time spans to evaluate algorithm performance.

5) IMPACTS OF RESTRICTING THE VERIFICATION DATA

At some sites/times, the thermal profile in the HRRR model is incompatible with the observed phase. This could be due to either model or observational error (Reeves et al. 2014; Reeves 2016; Landolt et al. 2019). Such a problem is known to impact some precipitation-type algorithms more so than others (Reeves et al. 2014; Reeves 2016). Five incompatibilities are identified between the HRRR thermal profiles and the ASOS. These are

- observations of RA or RA mixes where the HRRR 2-m T_w is subfreezing (3063 profiles),
- observations of FZRA, FZDZ, and mixes of these where the HRRR 2-m T_w is above freezing (1103 profiles),
- observations of SN where there is an elevated warm layer (2420 profiles),
- observations of PL or mixes that include PL where there is not an elevated warm layer (841 profiles), and
- thermal profiles whose maximum RH is less than 80% (2515 profiles).

These suspicious observation/profile pairs are removed in an experiment called "quality control" (or QC in Table 2). In quality control, POD, HSS, and EDI are improved (Table 5). The changes to bias are nominal in most algorithms, indicating that the improvements to the other scores are legitimate indicators of better performance. This, of course, stands to reason—eliminating obvious erroneous pairings should improve the apparent performance of an algorithm. But there is one curious aspect of quality control, that is, the degree to which the algorithms are benefitted by this change. MB and SBC are appreciably more benefitted. The reason for this is the dominance of FZRA in the refreezing dataset—about 50% of the observations are FZRA (Table 1). As will be demonstrated below, both MB and SBC are more biased toward FZRA, whereas HP is biased away from FZRA.

Several forms of precipitation cannot be automatically detected and are only reported when a human observer augments the report. These include PL and all types of mixes. It is logical, then, to presume that the full ASOS dataset may be subject to some bias against these forms of precipitation. To account for this, an experiment called "ML-Hub" is performed. In this experiment, the observations are limited to only medium- and large-hub airports (Fig. 1) between 0600 and 2000 local time. The designation of medium and large hubs is based on the 2021 FAA categorizations. These airports/times are most likely to have a CWO on shift to curate the automatically generated observations. This experiment yields considerably lower POD, HSS, and EDI (Table 5). This is almost entirely due to mixes, which account for a larger fraction of the total number of refreezing events in ML-Hub than in ASOS (68% versus 26%; Table 2). Allowing more mixes in an algorithm increases the degrees of freedom and opportunity for inexact agreement with the observations.

6) OVERRIDING RESULTS OF METHODOLOGY IMPACTS

The overriding result from these experiments is that goodness is a function of the methodology used to measure it. Consider the range of PODs from MB. At its best, it has a POD of 73%, but at its worst, it has a POD of 23%. Taken in isolation from the other algorithms, one could very easily choose to embrace or reject this algorithm, depending on what set of statistics is presented. The same is true for the other algorithms.

Betterness is also a function of the methodology. In most experiments, HP has the best bias, while the best POD, HSS, and EDI mostly belong to MB or SBC. If one were to try and mix and match statistics by applying a different set of rules to each algorithm, the apparent betterness could change dramatically. As an example, compare the PODs for HP from neighborhood approach to the scores from SBC for ML-Hub. This comparison suggests HP is significantly better. But as Table 5 demonstrates, when these two algorithms are compared using an identical methodology and observations, SBC has the higher PODs.

Clearly, the choice of data used to verify algorithms can result in profoundly different conclusions about algorithm performance. One may wonder if a gold-standard set of observations can be created to verify algorithms that may result in more clarity on goodness and betterness. This is attempted herein by combining the rules used in the QC, long-duration, and ML-Hub experiments into a single experiment referred to as "gold standard" (GS in Table 2). This combination limits the assessment to less ambiguous environments with a higher degree of quality control and human curation. The performance metrics for this experiment are very similar to those for ASOS. The best-performing algorithm depends on the verification metric, and the differences that do exist in scores are mostly nominal, suggesting there is no obvious best. It is discouraging to note that

TABLE 6. The PODs and 95% confidence interval for select experiments. The best score for each performance metric and experiment is indicated in boldface.

Expt	Algorithm	POD
ASOS	HP	22 ± 1.0
	MB	**40 ± 1.2**
	SBC	**42 ± 1.2**
Gold standard	HP	22 ± 3.4
	MB	**37 ± 4.0**
	SBC	**41 ± 4.1**

even with a high degree of vetting, the apparent statistical performance of algorithms is similar to the unvetted ASOS experiment and underscores the very difficult problem in attempting to verify precipitation-type algorithms.

Such findings have important implications for developers and decision-makers alike. Goodness can only be understood by the receiving agent if exact details on the methodology and the potential weaknesses of that method are fully disclosed. Given the dependency of betterness on the choice of methodology, an indisputable declaration of best may not be possible.

c. *Statistical significance of results*

Before moving on, a brief discussion about the statistical significance of the results is merited. This was evaluated following the methodology of Jolliffe (2007). Namely, 1000 iterations of bootstrapping were performed to determine the 95% confidence intervals. For the sake of concision, this discussion is limited to POD and only for select experiments. The confidence intervals for ASOS indicate that the lower PODs for HP are statistically significant, while differences between MB and SBC are not (Table 6). This is the case for most of the experiments—MB and SBC do not have statistically different PODs (shown for gold standard in Table 6). What is curious to note is that the 95% confidence intervals for ASOS are quite a bit smaller than the variability of PODs obtained by changing the methodology (cf., Tables 5 and 6). This holds true even for comparatively small datasets, wherein the confidence interval is larger (e.g., gold standard; Table 6). Such a result underscores the importance of methodology—changes in methodology affect the interpretation of how an algorithm appears to perform in ways that far exceed ordinary sampling error.

4. Algorithmic biases and their impacts

a. *A case study demonstrating algorithmic bias*

The goodness of an algorithm is limited by three things: observational uncertainty, the temporal/spatial variability of precipitation type [both of which have already been investigated in previous work (Reeves 2016)], and the inherent biases within each algorithm's logic. Algorithmic biases can present decision-makers with what may appear to be a confident prediction of a certain phase when, in fact, the environment may be conducive to other outcomes. Let us consider a case study

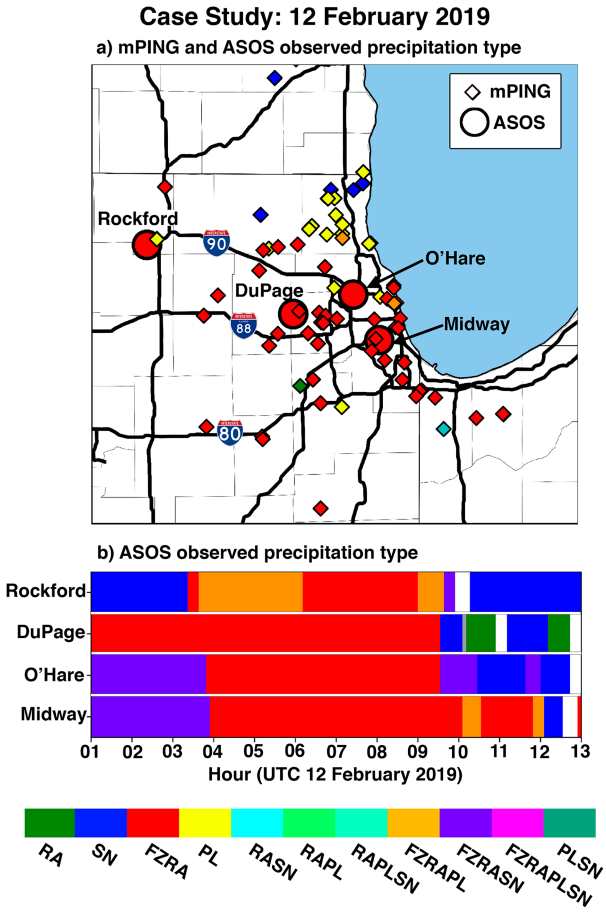

FIG. 3. The (a) mPING observations between 0400 and 0800 UTC and ASOS observations at 0600 UTC 12 Feb 2019 and (b) time sequences of precipitation type from the four ASOS stations in (a) between 0100 and 1300 UTC 12 Feb 2019. White bars in (b) indicate no precipitation type was observed.

that occurred in northern Illinois on 12 February 2019. The mPING observations between 0400 and 0800 UTC and ASOS reports at 0600 UTC on this day show FZRA between Interstates 90 and 88 (I-90 and I-88; Fig. 3a). Time sequences of the 5-min reports from four ASOS sites in this region (indicated in Fig. 3a) show that FZRA and/or FZRAPL are reported consistently between 0400 and 0930 UTC at all of these sites (Fig. 3b).

As this was not an event characterized by rapid temporal/spatial shifts in precipitation type, one might expect all three algorithms to converge on a diagnosis that includes FZRA in the region between I-90 and I-88. To evaluate whether this is true, forecasts of precipitation type are created from the 1- and 6-h HRRR forecasts valid at 0600 UTC (Fig. 4). Both lead times and all algorithms diagnose FZRA and FZRA mixes south of I-88, but between I-88 and I-90, important differences emerge. HP diagnoses PLSN, while MB and SBC diagnose FZRA and FZRA mixes. This difference may lead to different actions on the part of stakeholders. But when comparing like algorithms, differences are minor. For example,

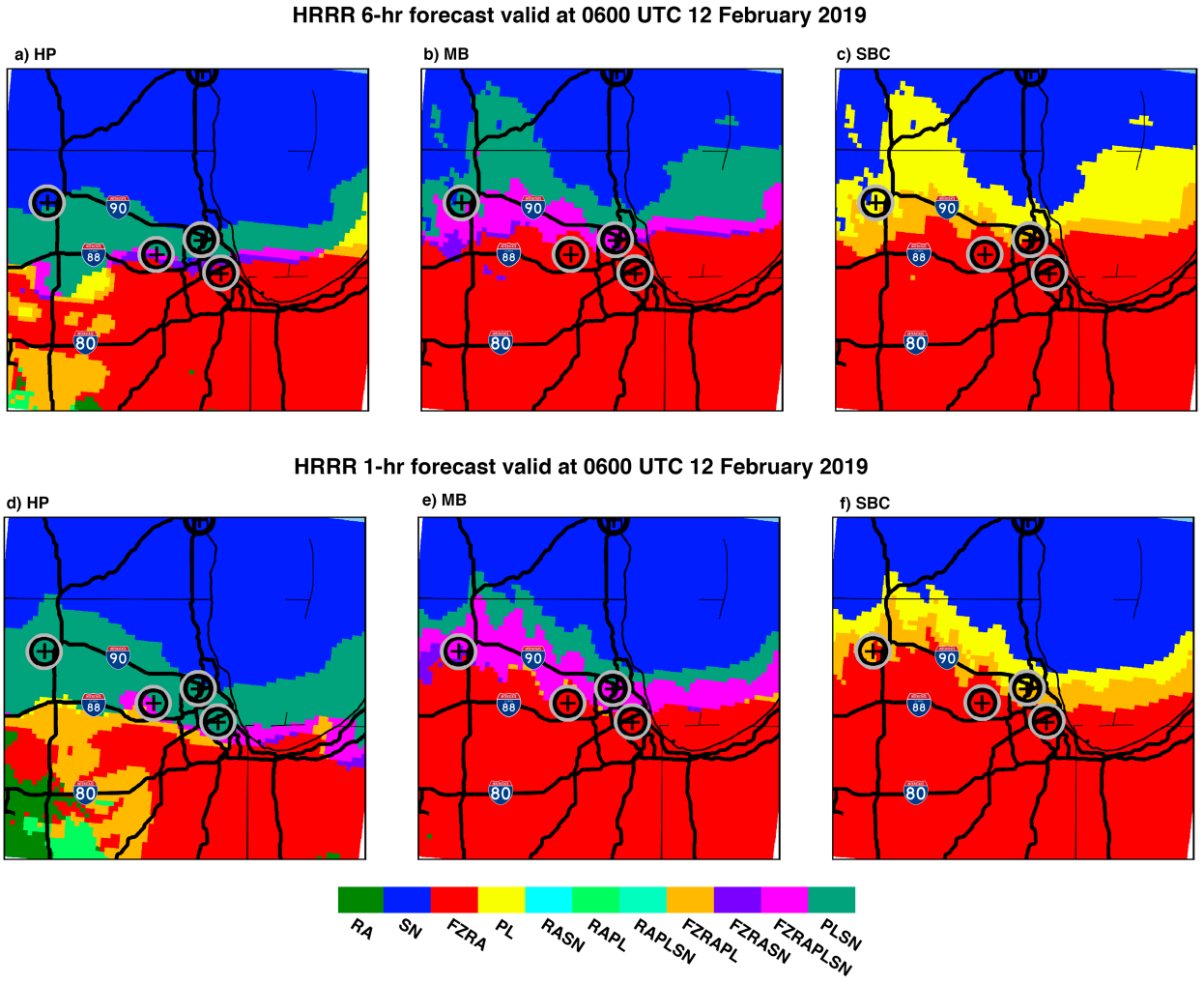

FIG. 4. The (top) 6-h and (bottom) 1-h forecasts of precipitation type by each algorithm valid at 0600 UTC 12 Feb 2019. The concentric gray and black rings with a plus sign indicate the four surface stations from Fig. 3.

both HP forecasts show this area as having mostly PLSN. Such correlation across lead times has the potential to increase forecaster confidence that the predicted outcome is more likely. In this example, the HRRR forecast may lead forecasters to believe that FZRA is unlikely.

Time trends of the predicted phase from the 0000 UTC 12 February 2019 forecast cycle give additional insight into algorithmic biases (Fig. 5). For the time range and stations shown, the majority of HP diagnoses are for SN or PLSN. Only 25% are diagnoses that include FZRA and these are not persistent in time or location, as is the case in the observations (cf. Figs. 3b and 5a). By contrast, MB more consistently produces FZRA or FZRA mixes. A bias toward FZRA is a deliberate choice on the part of the MB developers (E. Lenning 2022, personal communication), so this prediction from MB is not surprising. Last, SBC appears to be somewhat biased toward PL, at least at some times and locations.

The reason for such algorithm-to-algorithm variation is that the temperature profiles between I-90 and I-88 are ambiguous, as indicated at the four ASOS sites using the 6-h forecast valid at 0600 UTC 12 February 2019 (Fig. 6). Each profile has a modest elevated warm layer and a surface-based cold layer that is warmer than ice nucleation. Hence, refreezing is dictated by whether ice can survive the melting layer or be produced within the subfreezing layer. The subtle drying in the lowest 500 m of each sounding adds an additional complication as it suggests some of the smaller particles may be sublimated, which could impact the reported phase at the surface (Carlin et al. 2021). Based only on these soundings, no one algorithm stands out as obviously correct or incorrect.

b. *Algorithmic bias across the climatology*

The biases observed in this event are common to each algorithm as demonstrated through consideration of statistics from the ambiguous profiles experiment. HP diagnoses 33% of these profiles as SN and another 28% as PLSN (Fig. 7a). Hence, this algorithm trends toward colder diagnoses in

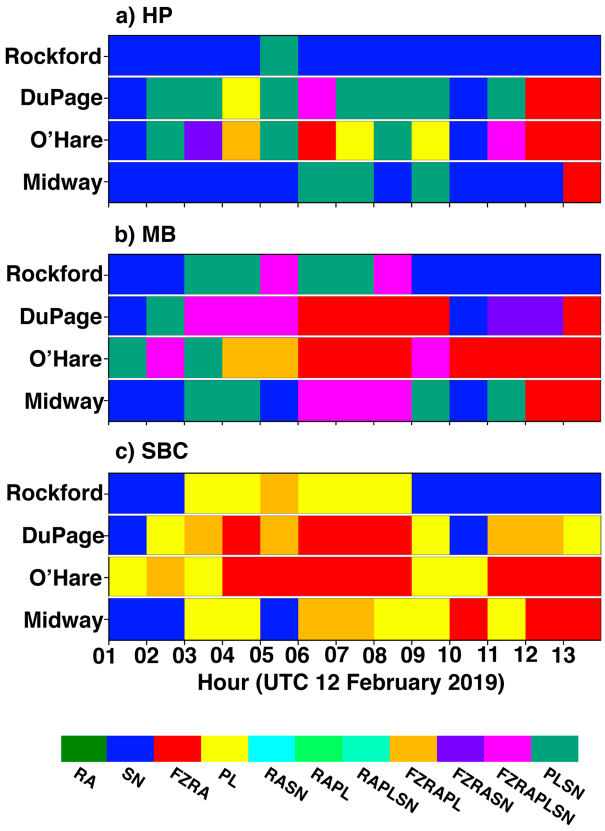

FIG. 5. The time trends of predicted phase for (a) HP, (b) MB, and (c) SBC for the forecast initialized at 0000 UTC 12 Feb 2019. The locations of the stations are indicated in Fig. 4a.

ambiguous situations. Conversely, MB is biased toward FZRA, having 66% of its diagnoses be for FZRA or FZRA mixes (Fig. 7b). SBC is biased toward PL and PL mixes (79%) with the remainder of soundings being classified as FZRA (Fig. 7c). In the case of this dataset, MB has the best representation of truth (Table 5). This is because the majority of the ambiguous profiles are observed as FZRA. So, the bias toward FZRA in this dataset is to MB's advantage. However, it is possible that observational error could be influencing these statistics and that the true precipitation phase may be better represented by one of the other algorithms.

The above biases are only evident when the thermal profile is ambiguous. To demonstrate this, the distribution of diagnoses for profiles characteristic of FZRA is computed. These profiles are defined as having an elevated warm layer whose maximum temperature exceeds 2°C and whose minimum temperature in the surface-based cold layer is warmer than −6°C. MB and SBC diagnose 95% and 100% of these profiles as FZRA or FZRA mix, respectively. HP has a lower percentage (61%). This is because HP uses temperature rather than T_w to discriminate between FZRA and RA. 28% of the soundings in this collection are diagnosed as RA or RA mix by HP because of this choice in logic. A simple change in logic would result in 89% of these profiles being diagnosed as FZRA or FZRA mix.

c. *Algorithmic biases in an ensemble forecast*

One might expect that the range of possible solutions afforded in ensemble forecasts would compensate for algorithmic bias. However, this is not the case. The 6-h HRRR ensemble (HRRR-e; Kalina et al. 2021) forecasts initialized at 0000 UTC 12 February 2019 show that the most-likely precipitation type for each algorithm is similar to the deterministic solutions (cf. Figs. 6 and 9). Herein, the most-likely precipitation type is the category that is most frequently diagnosed when the algorithms are applied to individual ensemble members. At this lead time, HP produces mostly PL between I-90 and I-88 (Fig. 8a), a slight variation on the PLSN produced by the deterministic HRRR. Ensemble forecasts for MB have mostly FZRAPLSN and FZRA in this area (Fig. 8b), and SBC produces FZRA and FZRAPL (Fig. 8c). These forecasts are nearly identical to their deterministic counterparts (Figs. 5b,c,e,f).

While the most likely precipitation type may be something other than FZRA, this does not imply a zero probability of FZRA. Time sequences of the probability of FZRA and/or FZRA mixes at each ASOS location show that at Rockford, the probabilities are low by each algorithm. This is because the HRRR-e is predicting colder-than-observed temperatures at this site in the lower levels of the atmosphere (not shown). Additionally, HP fails to give a diagnosis at Rockford, Illinois, and O'Hare Airport (Chicago, Illinois) at select times because the minimum threshold on precipitation rate was not met and no diagnosis was made for any of the members. Nevertheless, the biases noted above are apparent at other times/locations, even though there is now an ensemble of solutions. HP has very low FZRA probabilities between 0400 and 1000 UTC (Fig. 9a), MB has comparatively high probabilities (Fig. 9b) and SBC has slightly lower probabilities than MB (Fig. 9c). Hence, even though there is a diversity of solutions for the thermodynamic profiles, it is not sufficient to overcome the inherent biases in the algorithms. In the case of HP, the probability of FZRA may be too low to be considered actionable by forecasters and one could argue that this forecast of precipitation type is under dispersive. Before moving on, it is important to note that although the HP algorithm has the least-accurate forecast for this event, this is not always the case. Events can be found where the other algorithms appear to be the least accurate.

d. *Using a multi-algorithm approach*

Algorithmic biases are not an unknown phenomenon. These biases are handled within the NWS's Unified Post Processor in a "ensemble-like" fashion wherein all algorithms are applied and the most-likely solution provided to forecasters (Manikin et al. 2004; Manikin 2005). A similar approach may be merited moving forward wherein multiple algorithms are applied to each member of an ensemble, effectively expanding the ensemble membership by a factor of 3, and the probability across the entire distribution computed. This is tested with the 12 February 2019 HRRR-e forecasts. These probabilities are provided in Fig. 9d in an experiment called "multi-algorithm." The probabilities using this approach are expected—somewhat less than in MB and SBC, but more than in HP. Between about

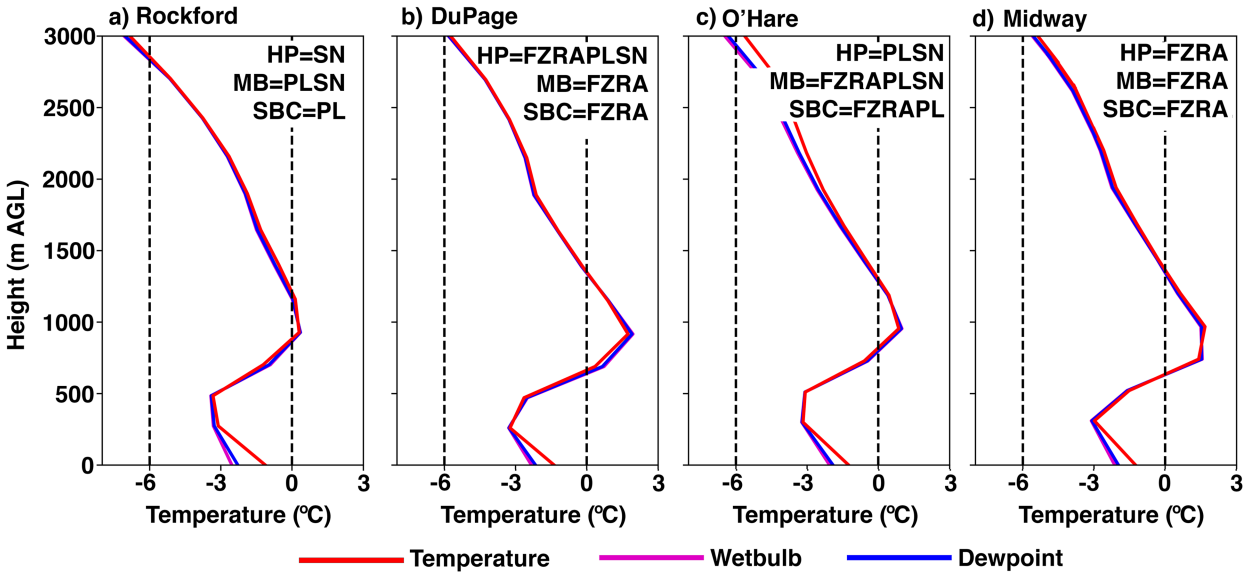

FIG. 6. Profiles of temperature, T_w, and dewpoint at (a) Rockford, (b) DuPage, IL, (c) O'Hare Airport, and (d) Midway Airport (Chicago, IL) from the 6-h forecast initialized at 0000 UTC 12 Feb 2019. The locations of the stations are indicated in Fig. 4a. The vertical dashed lines indicate 0°C and the assumed ice-nucleation temperature for SBC (−6°C).

0500 and 1000 UTC, probabilities are mostly in excess of 45%, which is a compelling fraction to motivate action.

The performance of the multi-algorithm approach is evaluated by computing the probability of precipitation type for the full ASOS dataset. In this experiment, an algorithm is assumed to produce a given category if that category occurs in isolation or as a mix. For example, suppose HP, MB, and SBC diagnose FZRASN, FZRAPL, and FZRA, respectively.

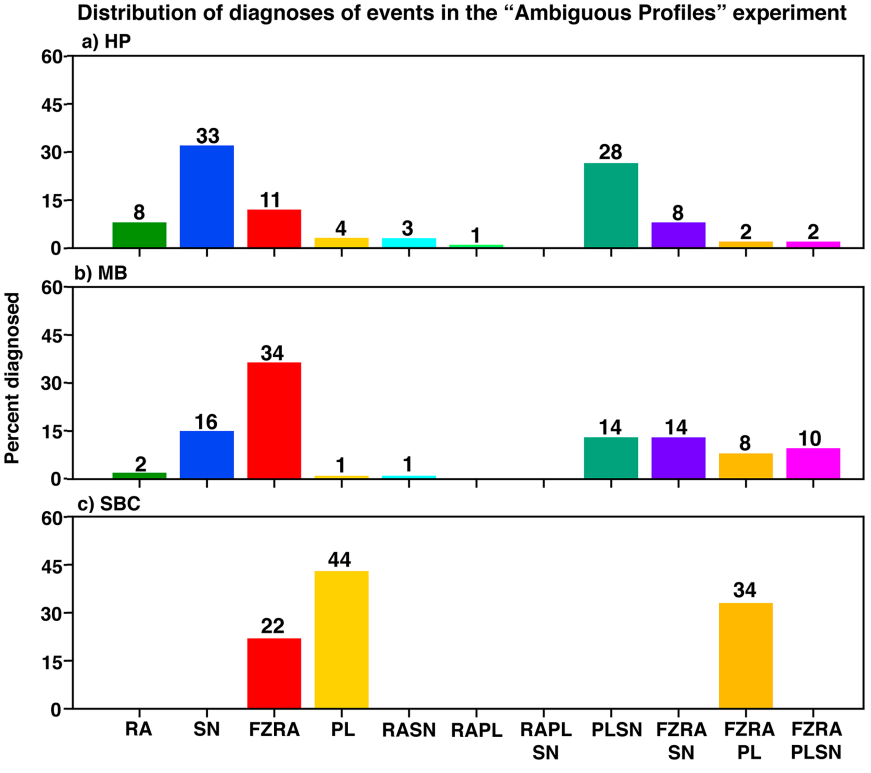

FIG. 7. Percent of each phase diagnosed for the soundings in the ambiguous profiles experiment.

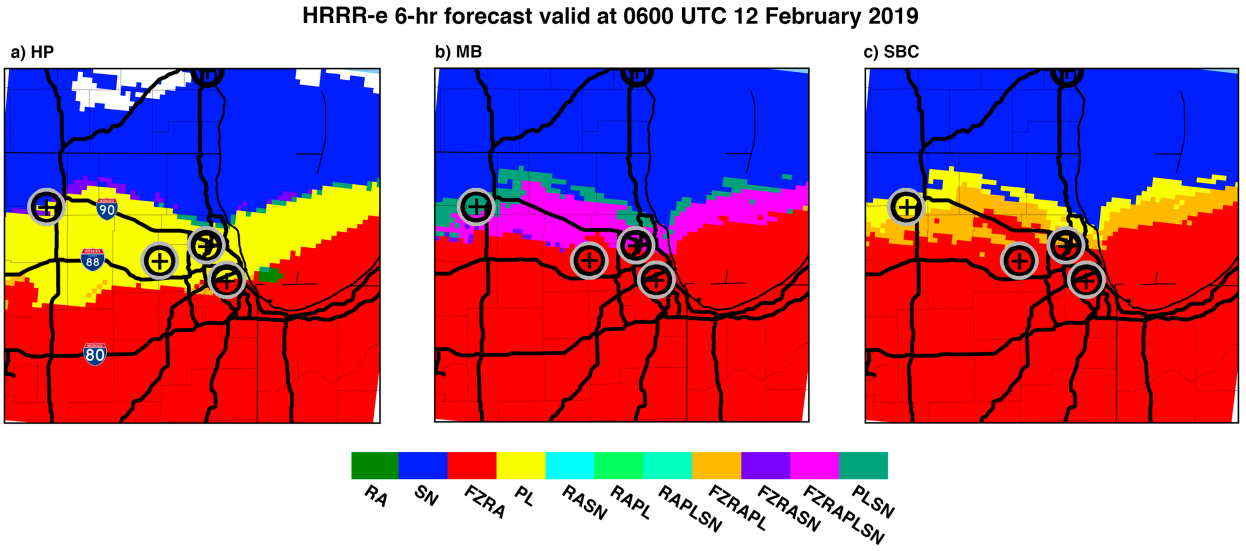

FIG. 8. The 6-h forecasts of most-likely precipitation type by each algorithm valid at 0600 UTC 12 Feb 2019 from the HRRR-e forecast system. The concentric gray and black rings with a plus sign indicate the four surface stations from Fig. 3.

Using our rules, the probability of FZRA is 100% because it is diagnosed by all three algorithms either as a constituent of a mix or in isolation. The probabilities of SN and PL are both 33.3% because they are each diagnosed as a constituent of a mix by only one algorithm apiece.

Whether or not using the algorithms in an ensemble-like fashion provides enhanced decision support depends on the phase and the method of verification. For any one category, regardless of whether using the receiver operating characteristic (ROC; Marzban 2004) curves or performance diagrams, the 66% probability has the optimal performance relative to other probability thresholds. When compared to the methodologies used in ASOS and generous hit, the relative superiority of multi-algorithm is unclear. It is obviously beneficial for RA (Figs. 10a,b). But for SN, a probability of 66% has only a slightly higher POD-FPR than for any one algorithm (1.17%; Fig. 10c). The POD + SR for SN at a probability of 66% is higher than for any one algorithm by only 0.22% (Fig. 10d). Given all of the uncertainties discussed above, the significance of such differences is highly questionable. In the case of FZRA, the POD-FPR for the generous hit method from SBC is 4.91% higher than for the 66% probability (Fig. 10e), but the POD + SR for the 66% probability is markedly higher (15.57%) than from any one algorithm (Fig. 10f). PL and refreezing precipitation are similarly ambiguous on the value of using a multi-algorithm approach when using ROC diagrams, but are clearly benefitted according to the combined performance diagram metrics (Figs. 10g–j). This may be partially due to the unfiltered influence of TNs in the ROC diagram. Last, it may be possible to see a more conclusive benefit if more algorithms are included in the analysis so that more probability thresholds can be obtained. But, we cannot definitively state there is benefit to using a multi-algorithm approach using verification metrics based on this analysis.

5. Conclusions

In this paper, the use of performance metrics to declare the "goodness" and "betterness" of precipitation-type algorithms is explored. Goodness is defined as the performance of an individual algorithm as dictated by performance metrics and betterness is defined as the superior performance of one algorithm relative to another. To assess this, three modern precipitation-type algorithms are compared using identical initial conditions for a 5-yr retrospective of precipitation-type observations from different combinations of the ASOS and AWOS networks.

There are four key findings to this work. First is that goodness and betterness are a function of the performance metric and hydrometeor phase. Because the different performance metrics define goodness in different ways, it is possible for an algorithm to have excellent scores for some metrics while having low scores for others. This was the case with HP. When applied to refreezing categories (i.e., FZRA, PL, and mixes that include either of these), its bias was quite good, while its POD was rather poor, meaning the goodness of this algorithm cannot be objectively declared. Rather, its goodness is dependent on the needs of the end user. If the need is for the algorithm to accurately predict the frequency of the phenomena and spatial/temporal errors are irrelevant, HP is clearly the best option. But if spatial/temporal accuracy is more important, this algorithm is the worst of the three considered, for this one category. The apparent betterness also depends on the phase. For example, HP had the highest POD for RA and SN, while SBC had the highest POD for FZRA and PL. This dependence on the choice of performance metric and phase makes the exercise to declare the goodness and betterness of a precipitation-type algorithm difficult and perhaps even impossible.

The second key finding of this work is that goodness and betterness are functions of the observational data used to verify the algorithms. Multiple approaches were considered herein, each

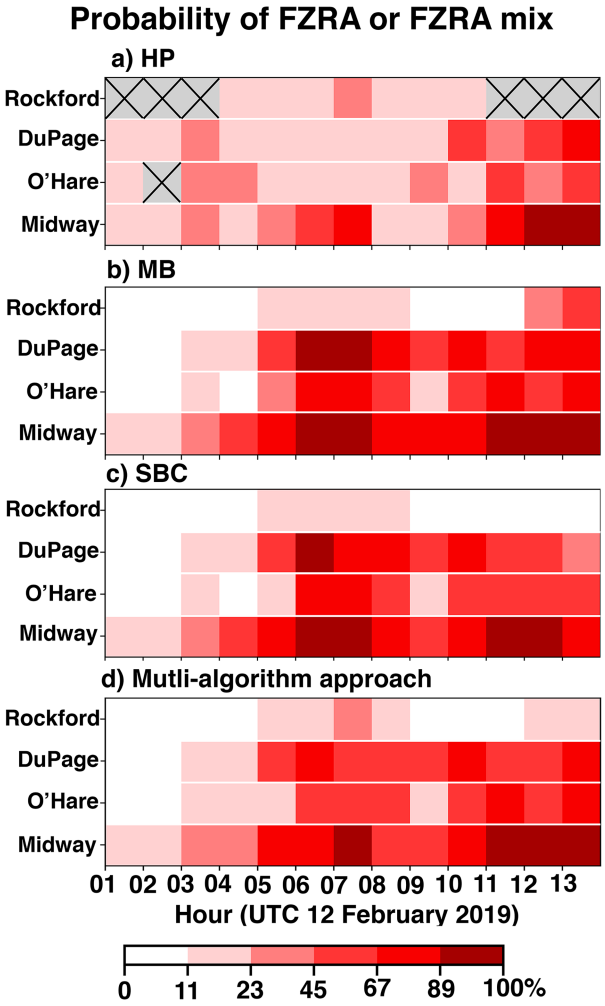

FIG. 9. The time trends of the probability of FZRA (or FZRA mix) for (a) HP, (b) MB, (c) SBC, and (d) using a multi-algorithm approach for the HRRR-e forecast initialized at 0000 UTC 12 Feb 2019. The locations of the stations are indicated in Fig. 4a. The light-gray cells with x marks in (a) represent times when the HP algorithm did not diagnose any precipitation type for any members of the ensemble.

one a plausible option. As above, goodness varied depending on the methodology. For example, MB had a POD as low as 23% and as high as 73% for refreezing categories depending on what data were used for verification. Betterness also varied. All algorithms had instances of having the most optimal of one or another metric, depending on the experiment. This further obfuscates one's ability to unambiguously declare the relative superiority of an algorithm. We stress that the different methodologies used herein are all reasonable choices one may make in the process of verification. Yet, they differently impact the interpretation of results and could influence whether or not an individual algorithm is adopted for operational use or rejected.

The third key finding of this research is that the algorithms have clear and impactful biases in ambiguous environments that are not offset even when applied to ensemble forecasts. In the case of HP, the majority of profiles in these settings are diagnosed as SN or a SN mix. MB is biased toward FZRA and FZRA mixes, while SBC is somewhat biased toward PL but tends to have a more even distribution between FZRA, PL, and FZRAPL than the other two. The impacts of these biases were demonstrated using a prolonged FZRA event in northern Illinois. In a populated corridor between I-90 and I-88, the biases of each algorithm affected the diagnosed phase. To evaluate the impacts of algorithmic biases on ensemble forecasts, the three algorithms were applied to HRRR-e forecasts of this event. The same biases were noted in the most-likely phase predicted using each algorithm. Probabilities of FZRA or FZRA mixes by each algorithm showed that HP, which tends to be biased against FZRA, were quite low, having only one member diagnosing this phase at some times and locations.

The final key finding of this research was prompted by the above. The fact that the HRRR-e forecasts did produce high probabilities of FZRA for some algorithms suggested that using all algorithms together to amplify the apparent membership of an ensemble could lead to improved decision support by providing forecasters with multiple solutions. Indeed, for the case study evaluated herein, this approach led to higher FZRA probabilities, in accordance with what was observed. However, when applied to the 5-yr climatology of events using performance metrics, the value of the multi-algorithm approach was less obvious. As above, whether a multi-algorithm approach is superior depends on the performance metric and phase.

In closing, we make the final recommendations. First, while a statistical evaluation of precipitation-type algorithms is an essential step to ensuring reasonable performance, the use of these metrics to declare the superior performance of one algorithm over another is of questionable merit. There are many contingencies that if changed in only small—but scientifically defensible—ways, could consequentially alter the conclusions of such an analysis. A potentially more valuable approach is to also consider the decision-support capabilities provided by an algorithm when evaluating their use in operations. In the case of MB, this algorithm is purposefully biased toward FZRA and FZRA mixes because it is easier for a forecaster to "tone down" an overprediction of something rather than anticipate something not predicted at all (E. Lenning 2022, personal communication). So, this algorithm's biases are not necessarily a weakness, even though statistically speaking, its bias is not as good as HP's. We recommend that future assessments of algorithms for operations include consideration of decision-support capabilities that may not be represented within a statistical performance assessment.

Second, to embrace a single-algorithm approach is to embrace its biases, even when using ensemble forecasts. These biases can have negative impacts on stakeholders. Given the many ongoing efforts to create new methods for diagnosing the precipitation phase, a more robust approach may be to use multiple algorithms. We recommend consideration of this technique as this additional axis of uncertainty may complement the forms of uncertainty already expressed in ensemble forecasts. This can provide decision-makers with a greater appreciation of the range of potential outcomes. Ultimately, this could improve a forecaster's

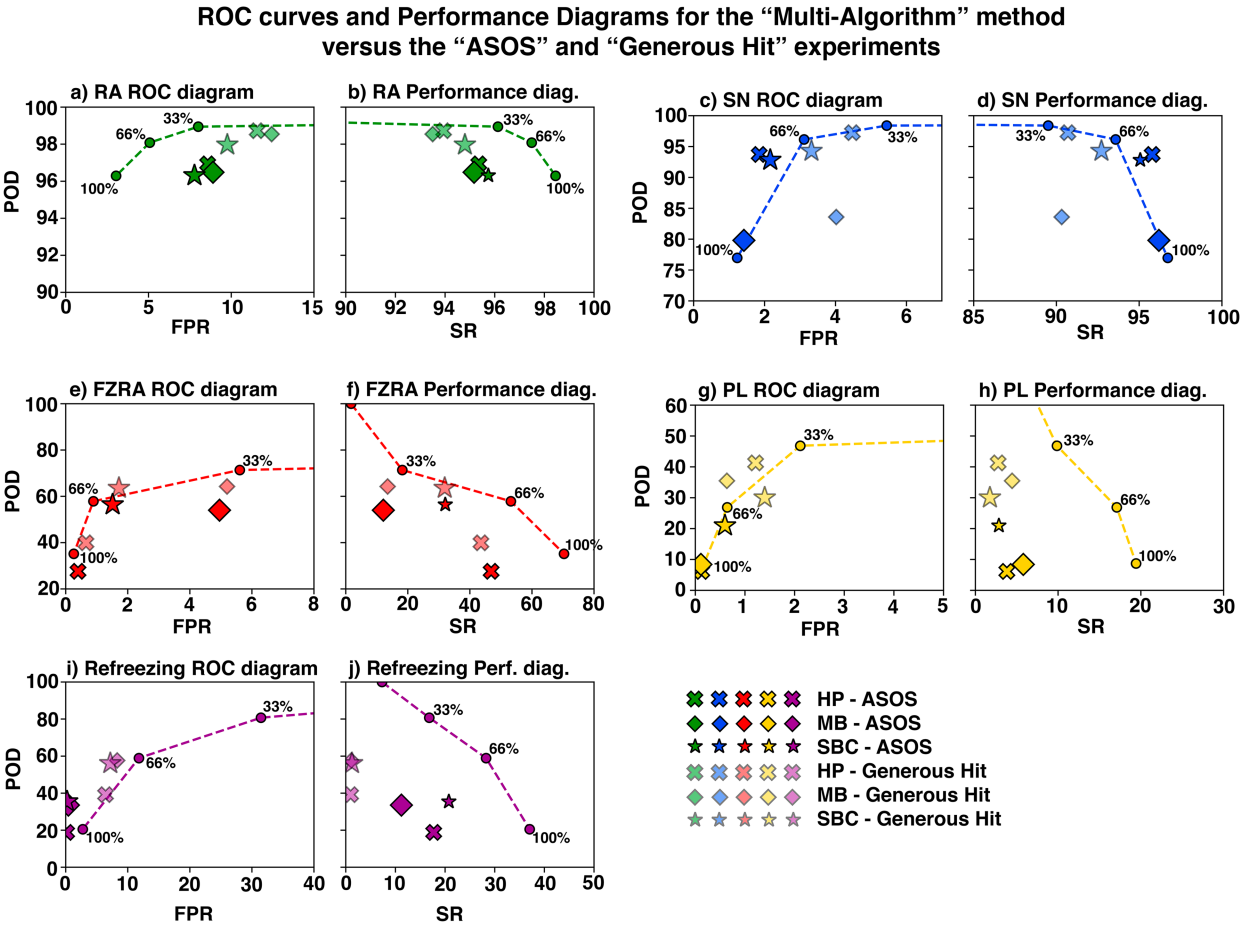

FIG. 10. Receiver operator curves (ROC) and performance diagrams for the multi-algorithm method vs the ASOS and generous hit experiments. (a),(c),(e),(g),(i) The *x* and *y* axes are the false positive rate (FPR) and probability of detection (POD) as defined in Table 3. (b),(d),(f),(h),(j) The *x* and *y* axes are the success ratio (SR) and POD also as defined in Table 3.

ability to provide more meaningful impacts-based decision support to stakeholders.

Acknowledgments. Special thanks to E. Lenning and E. James S. Cocks. This study was made possible in part due to the data made available by the governmental agencies, commercial firms, and educational institutions participating in MesoWest and through archival of HRRR forecasts by the University of Utah via Amazon web services. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA21OAR4590162, U.S. Department of Commerce.

Data availability statement. ASOS observations are archived and accessible via https://mesowest.utah.edu. The mPING observations are archived and accessible via https://mping.ou.edu. HRRR analyses are archived and accessible via https://registry.opendata.aws/noaa-hrrr-pds/. Output from each algorithm for the experiments herein is available by request from the lead author. FAA airport categorizations are available at https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/cy21_commercial_service_enplanements.

REFERENCES

Baldwin, M., R. Treadon, and S. Contorno, 1994: Precipitation type prediction using a decision tree approach with NMCs Mesoscale Eta Model. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 30–31.

Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterization. *Wea. Forecasting*, **31**, 609–619, https://doi.org/10.1175/WAF-D-15-0136.1.

Birk, K., E. Lenning, K. Donofrio, and M. T. Friedlein, 2021: A revised Bourgouin precipitation-type algorithm. *Wea. Forecasting*, **36**, 425–438, https://doi.org/10.1175/WAF-D-20-0118.1.

Bourgouin, P., 2000: A method to determine precipitation type. *Wea. Forecasting*, **15**, 583–592, https://doi.org/10.1175/1520-0434(2000)015<0583:AMTDPT>2.0.CO;2.

Carlin, J. T., H. D. Reees, and A. V. Ryzhkov, 2021: Polarimetric observations and simulations of sublimating snow: Implications for nowcasting. *J. Appl. Meteor. Climatol.*, **60**, 1035–1054, https://doi.org/10.1175/JAMC-D-21-0038.1.

Cholette, M., J. M. Theriault, J. A. Milbrandt, and H. Morrison, 2020: Impacts of predicting the liquid fraction of mixed-phase particles on the simulation of an extreme freezing rain event: The 1998 North American ice storm. *Mon. Wea. Rev.*, **148**, 3799–3823, https://doi.org/10.1175/MWR-D-20-0026.1.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, https://doi.org/10.1175/2009WAF2222251.1.

Elmore, K. L., H. M. Grams, D. Apps, and H. D. Reeves, 2015: Verifying forecast precipitation type with mPING. *Wea. Forecasting*, **30**, 656–667, https://doi.org/10.1175/WAF-D-14-00068.1.

Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699–713, https://doi.org/10.1175/WAF-D-10-05030.1.

Filipiak, B., K. L. Corbosiera, A. L. Lang, N. P. Bassill, and R. Lazear, 2023: Data fusion: A machine learning tool for forecasting winter mixed precipitation events—Updates and performance. *22nd Conf. on Artificial Intelligence for Environmental Systems*, Denver, CO, Amer. Meteor. Soc., 15A.4, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/411133.

Gascón, E., T. Hewson, and T. Haiden, 2018: Improving predictions of precipitation type at the surface: Description and verification of two new products from the ECMWF ensemble. *Wea. Forecasting*, **33**, 89–108, https://doi.org/10.1175/WAF-D-17-0114.1.

Gerrity, J. P., Jr., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709–2712, https://doi.org/10.1175/1520-0493(1992)120<2709:ANOGAM>2.0.CO;2.

Harrison, D. R., A. McGovern, C. Karstens, I. L. Jirak, and P. T. Marsh, 2022: Winter precipitation-type classification with a 1D convolutional neural network. *31st Conf. on Weather Analysis and Forecasting/27th Conf. on Numerical Weather Prediction*, Houston, TX, Amer. Meteor. Soc., J11.4, https://ams.confex.com/ams/102ANNUAL/meetingapp.cgi/Paper/394420.

Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating High-Resolution Rapid Refresh Model. *Wea. Forecasting*, **28**, 921–939, https://doi.org/10.1175/WAF-D-12-00085.1.

Jolliffe, I. T., 2007: Uncertainty and inference of verification measures. *Wea. Forecasting*, **22**, 637–650, https://doi.org/10.1175/WAF989.1.

Kalina, E. A., I. Jankov, T. Alcott, J. Olson, J. Beck, J. Berner, D. Dowell, and C. Alexander, 2021: A progress report on the development of the High-Resolution Rapid Refresh ensemble. *Wea. Forecasting*, **36**, 791–804, https://doi.org/10.1175/WAF-D-20-0098.1.

Landolt, S. D., J. S. Lave, D. Jacobson, A. Gaydos, S. DiVito, and D. Porter, 2019: The impacts of automation on present weather-type observing capabilities across the conterminous United States. *J. Appl. Meteor. Climatol.*, **58**, 2699–2715, https://doi.org/10.1175/JAMC-D-19-0170.1.

Lu, Z., H. Yongxiang, and Y. Liu, 2021: Improving the Ramer scheme for diagnosis of freezing rain in China. *Atmos. Res.*, **254**, 105520, https://doi.org/10.1016/j.atmosres.2021.105520.

Manikin, G. S., 2005: An overview of precipitation type forecasting using NAM and SREF data. *24th Conf. on Broadcast Meteorology/21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 8A.6, https://ams.confex.com/ams/pdfpapers/94838.pdf.

——, K. F. Brill, and B. Ferrier, 2004: An Eta Model precipitation type mini-ensemble for winter weather forecasting. *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 23.1, https://ams.confex.com/ams/pdfpapers/73517.pdf.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, https://doi.org/10.1175/825.1.

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

NOAA, 1998: *Automated Surface Observing System (ASOS) User's Guide.* National Oceanic and Atmospheric Administration, 74 pp., https://www.weather.gov/media/asos/aum-toc.pdf.

Ramer, J., 1993: An empirical technique for diagnosing precipitation type from model output. *Fifth Int. Conf. on Aviation Weather Systems,* Vienna, VA, Amer. Meteor. Soc., 227–230.

Reeves, H. D., 2016: The uncertainty of precipitation-type observations and its effect on the validation of forecast precipitation type. *Wea. Forecasting*, **31**, 1961–1971, https://doi.org/10.1175/WAF-D-16-0068.1.

——, K. L. Elmore, A. Ryzhkov, T. Schuur, and J. Krause, 2014: Sources of uncertainty in precipitation-type forecasting. *Wea. Forecasting*, **29**, 936–953, https://doi.org/10.1175/WAF-D-14-00007.1.

——, A. V. Ryzhkov, and J. Krause, 2016: Discrimination between winter precipitation types based on spectral-bin microphysical modeling. *J. Appl. Meteor. Climatol.*, **55**, 1747–1761, https://doi.org/10.1175/JAMC-D-16-0044.1.

——, N. Lis, G. Zhang, and A. A. Rosenow, 2022: Development and testing of an advanced hydrometeor-phase algorithm to meet emerging needs in the aviation sector. *J. Appl. Meteor. Climatol.*, **61**, 521–536, https://doi.org/10.1175/JAMC-D-21-0151.1.

Wandishin, M. S., M. E. Baldwin, S. L. Mullen, and J. V. Cortinas Jr., 2005: Short-range ensemble forecasts of precipitation type. *Wea. Forecasting*, **20**, 609–626, https://doi.org/10.1175/WAF871.1.

Weygandt, S. S., T. G. Smirnova, S. G. Benjamin, K. J. Brundage, S. R. Sahm, C. R. Alexander, and B. E. Schwartz, 2009: The High Resolution Rapid Refresh (HRRR): An hourly updated convection resolving model utilizing radar reflectivity assimilation from the RUC/RR. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 15A.6, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154317.htm.