The Uncertainty of Precipitation-Type Observations and Its Effect on the Validation of Forecast Precipitation Type

HEATHER DAWN REEVES

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 7 April 2016, in final form 22 September 2016)

ABSTRACT

Herein, an evaluation of the uncertainty of precipitation-type observations and its effect on the validation of forecast precipitation type is undertaken. The forms of uncertainty are instrument/observer bias and horizontal/temporal variability. Instrument/observer biases are assessed by comparing observations from the Automated Surface Observing Station (ASOS) and Meteorological Phenomena Identification Near the Ground (mPING) networks. Relative to the augmented ASOS, mPING observations are biased toward ice pellets (PL) and away from rain (RA). However, when mPING is used to validate precipitation-type algorithms, the probabilities of detection (PODs) for both RA and PL are decreased relative to those from the augmented ASOS. The decreased POD for RA is the result of numerous mPING reports of RA in the presence of a surface-subfreezing layer in the nearest observed sounding. Temporal and spatial variability effects are also assessed. The typical lifespan of transitional forms of precipitation is between 10 and 40 min, with many events having two or more forms of precipitation reported in a 1-h time frame. Depending on how one defines a hit for these rapidly evolving events, inherent biases in the forecasts may be dampened or masked altogether. Spatial variability also exerts a strong control on the performance of postprocessing algorithms, as both FZRA and PL often have spatial scales that are too small to be resolved, even by convection-allowing forecast models. However, the degree of variability is not strongly dependent on the distance separating any two observation pairs and, consequently, validation statistics do not change significantly as a model's grid spacing is increased, all else being equal.

1. Introduction

The correct specification of the surface precipitation type during winter is quite important, as it profoundly affects the economy and public safety (Ralph et al. 2005). Numerous assessments show that although rain (RA) and snow (SN) are usually well predicted, freezing rain (FZRA) and ice pellets (PL) are not—a result generally attributed to uncertainty effects (Bourgouin 2000; Manikin et al. 2004; Manikin 2005; Wandishin et al. 2005; Ikeda et al. 2013; Reeves et al. 2014; Elmore et al. 2015, hereafter EMAR). Model and algorithm uncertainties have been addressed by several previous investigators (Manikin et al. 2004; Manikin 2005; Wandishin et al. 2005; Reeves et al. 2014), but little attention has been given to the apparent effects observational uncertainty has on precipitation-type validation statistics. Such is the aim of this paper.

Corresponding author e-mail: Heather Dawn Reeves, heather.reeves@noaa.gov

DOI: 10.1175/WAF-D-16-0068.1

Forecasts of precipitation type are created by postprocessing algorithms that are applied to numerical model output. There is a wide range of approaches and degrees of complexity to precipitation-type algorithms. Some use bulk properties from the temperature and humidity profiles (e.g., Baldwin et al. 1994; Bourgouin 2000; Schuur et al. 2012; Elmore and Grams 2015; Chenard et al. 2015), others attempt to calculate or infer the liquid-water content of falling hydrometeors (Ramer 1993; Czys et al. 1996; Reeves et al. 2016), while still others use mixing ratios from microphysical parameterization schemes as the primary discriminant (Thériault et al. 2010; Ikeda et al. 2013). As a result, different algorithms may produce very different results, particularly when the environmental temperature is near 0°C (Manikin et al. 2004; Manikin 2005; Reeves et al. 2014). In some cases, strong biases exist. For example, the Baldwin algorithm has a well-known bias toward PL. The Ramer algorithm is known to be biased toward FZRA, and the algorithm described in Schuur et al. (2012) is strongly biased toward a FZRA–PL mix (Baldwin et al. 1994; Manikin et al. 2004; Manikin

2005; Reeves et al. 2014). All of these factors combined can result in a wide array of diagnoses from the various algorithms, even when supplied with identical input (Reeves et al. 2014).

Poor forecasts may also be due to model uncertainty. The degree to which an algorithm suffers from model uncertainty is a function of its assumptions as some algorithms use discriminants that have a higher range of uncertainty (Wandishin et al. 2005; Reeves et al. 2014). But the proximity of the environment to 0°C is also a contributor as even small errors may be sufficient to change the low-level temperature from subfreezing to above freezing or vice versa. Indeed, uncertainty effects, which have been shown to be quite detrimental for PL and FZRA (since these forms usually occur at temperatures near 0°C), render some algorithms useless for discriminating between these classes (Wandishin et al. 2005; Reeves et al. 2014).

One source of uncertainty that has hitherto received minimal attention is the observational uncertainty. Most investigators use the Automated Surface Observing Station (ASOS) network as ground truth. An advantage of the ASOS is that the instruments continuously monitor the environment and are sensitive enough to detect changes in the precipitation type before the human eye, especially at night (NOAA 1998). Some ASOS sites are manned by trained human observers who can augment an automated report if they see that it is in error. At these locations, the observer may enter a variety of mixes or change the precipitation type entirely. Presumably, the augmented observations are highly reliable since they marry the strong sensitivity of the instruments with the quality control of a trained observer.

ASOS sites that are not augmented are known to have reporting errors in certain situations. They cannot, for example, diagnose PL or freezing drizzle (FZDZ) and tend to classify these forms of precipitation as RA, SN, or mist. It is also possible for SN to be misdiagnosed as FZRA or RA under certain conditions. The automated sites also do not report mixes. [For more information on the ASOS instrumentation, the reader is referred to NOAA (1998).] There have been several studies that use the ASOS network as ground truth, but that do not discriminate between the augmented and nonaugmented observations (e.g., Manikin et al. 2004; Manikin 2005; Wandishin et al. 2005; Ikeda et al. 2013; Thompson et al. 2014; Benjamin et al. 2016; Johnson and Shepherd 2016; Scheuerer et al. 2016; Shafer and Antolik 2016). Whether using these observations interchangeably affects validation statistics for precipitation-type classifiers is unknown.

A new observation dataset has recently been collected by the National Severe Storms Laboratory called the Meteorological Phenomena Identification Near the Ground (mPING; Elmore et al. 2014). These observations are made by ordinary citizens from their computers or mobile devices and include the following classes: RA, drizzle (DZ), RASN, FZDZ, FZRA, PL, RAPL, SNPL, and SN. This dataset has been used in at least one assessment of precipitation-type algorithms (EMAR) and is used as ground truth to build a new statistically driven classifier (Elmore and Grams 2015). Preliminary comparisons between mPING and ASOS suggest the mPING observations are reasonably robust (Elmore et al. 2014, EMAR), but the exact degree to which mPING differs from ASOS has not been quantified, and it remains an open question whether using this dataset over ASOS affects validation statistics.

Other observational uncertainties beyond those of the instruments/observers exist, particularly the temporal and spatial representativeness of the observations. Several investigators have found that environments near 0°C have significant temporal and spatial variability in their precipitation types (Crawford and Stewart 1995; Bernstein 2000; Cortinas 2000; Rauber et al. 2000, 2001; Robbins and Cortinas 2002; Changnon 2003; Cortinas et al. 2004; Thériault et al. 2010; Reeves et al. 2014; EMAR). These forms of variability have not previously been quantified. Hence, it is unclear how they should be treated in the verification of classifiers and whether different approaches will alter the conclusions and recommendations of the investigator.

The aim of this study is to assess both of the above forms of observational uncertainty (instrument/observer bias and temporal/spatial variability) and their effects on the apparent performance of precipitation-type classifiers. In section 2, different observing networks are compared to assess the presence of biases and their effects on validation statistics. In section 3, the temporal and spatial variabilities are quantified and their effects assessed. Concluding thoughts are provided in section 4.

2. Constituency of precipitation types at the surface

Let us first consider the frequency of the various precipitation types during winter months (October–March) for the ASOS and mPING networks. For this exercise, we will consider separately the augmented and nonaugmented ASOS sites, the locations of which are shown in Figs. 1a and 1d. The augmented sites are at commercial airports and air force bases where human observers (either Federal Aviation Administration–trained contractors or trained air force personnel) can alter the automated observations if they see that they are in error. The 5-min observations from 2000 to 2015 are used herein. Because these reports are augmented,

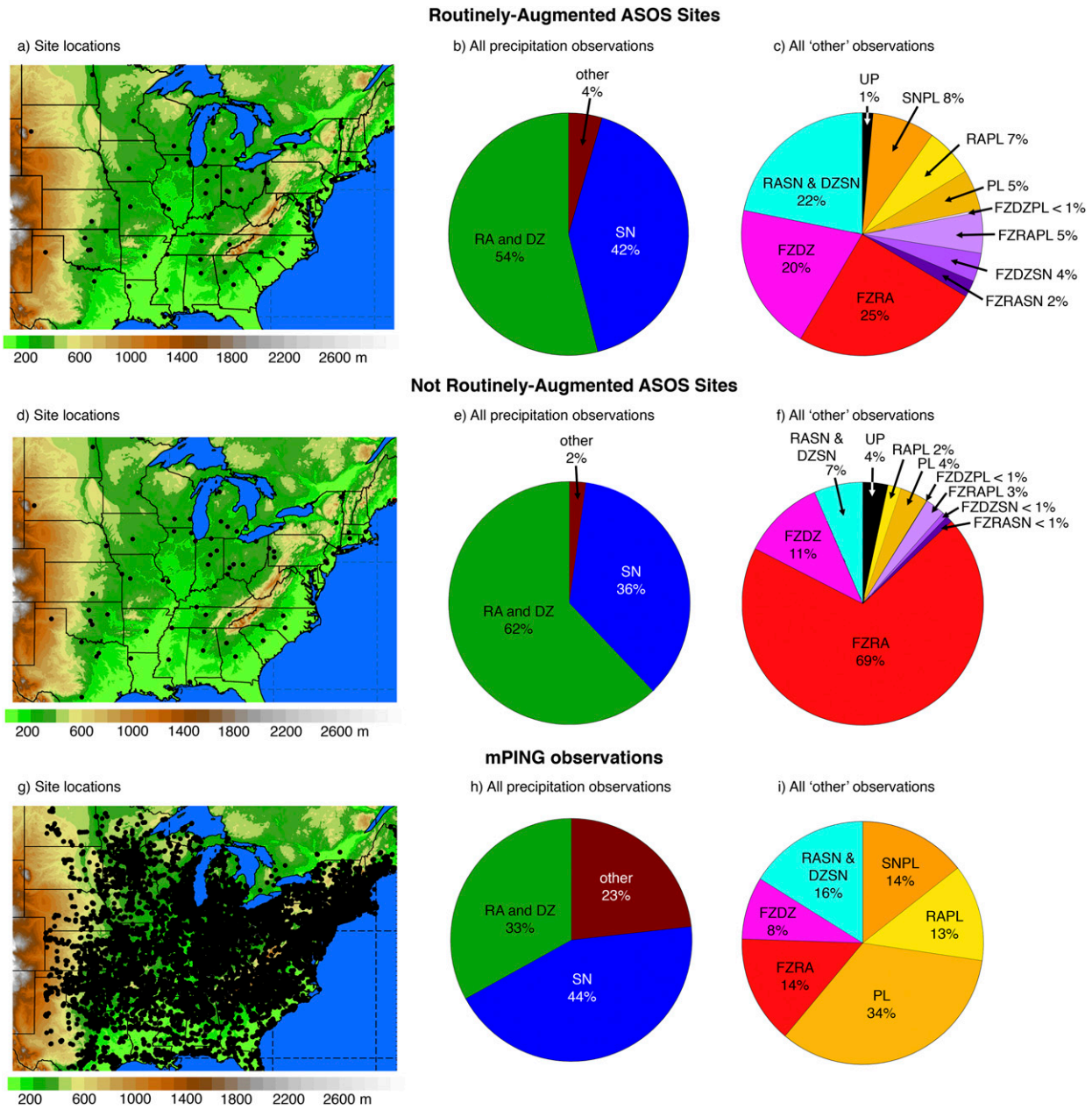

FIG. 1. (a),(d),(g) The locations of augmented, nonaugmented, and mPING observations used in this study. (b),(e),(h) The percentage (out of the total number of observations available) of each class indicated. (c),(f),(i) The percentage (out of the total number of "other" observations) of each class indicated.

they are treated as the relative "truth," against which the other reporting systems are compared. This is not to say that the augmented reports are completely free from error or may not have any biases of their own. Rather, the comparison is merely to highlight how the other networks that have lesser degrees of quality control perform.

Not all nonaugmented ASOS sites are used. Rather, only select sites that are relatively close to the augmented sites (most pairs are within 40 km of each other and no pairs are more than 60 km apart, as indicated in Fig. 1d) are included to facilitate a more direct comparison. Specifically avoided are sites within and west of the Rocky Mountains where the spatial representativity is questionable. The nonaugmented sites do not have a dedicated human observer. However, if present personnel discover an observation is in error, they can alter the report. So, there are some occasions when an augmented

report has been made at the nonaugmented sites in Fig. 1d, but the sites used herein are augmented less than 10% of the time. There is no set guidance on when to alter reports at locations that are usually nonaugmented (NOAA 1998). Likewise, there are occasions when the augmented sites do not have the augmentation flag, suggesting the observer made no alterations to the report. Only those sites with at least 95% of the observations flagged as augmented during precipitation are included in this group to ensure a high measure of quality control. The ASOS observations were carefully assessed at all sites to ensure changes in the instrumentation did not alter the frequency of the various precipitation types. While some years and sites have comparatively high frequencies of SN, FZRA, or PL that appear to be due to seasonal/latitudinal variations, there are no permanent trends or changes.

The mPING observations are available from December 2012 to March 2015. Their locations vary in time and space, as indicated in Fig. 1g. No mPING observations within and west of the Rocky Mountains are used in this study for the same reasons mentioned above. This yields 398 595 observations. (There are 1 980 987 augmented and 1 606 582 nonaugmented observations.)

In this section and all others, the classifiers used are those employed by Ramer (1993), Baldwin et al. (1994), and Bourgouin (2000). These are chosen because they can be initialized with observed sounding data and they represent a spectrum of biases according to previous research (Manikin et al. 2004; Manikin 2005; Reeves et al. 2014). There are two versions of the Baldwin algorithm, which are identical save for the discrimination between SN and PL. While both have a pronounced PL bias, the later version, referred to as Baldwin2, is the less biased and is used herein. All three algorithms diagnose SN, RA, PL, and FZRA. The Ramer scheme also has a FZRAPL class, but Reeves et al. (2014) demonstrate that this class is rarely diagnosed. In those instances where it is diagnosed, if just one of these categories agrees with the observation, that instance is considered a hit.

For the sake of concision, in this and subsequent sections, the validation statistics are limited to the probability of detection (POD). Other verification scores such as the false alarm ratio and the Heidke and Pierce skill scores were computed, but the overriding interpretation is the same as what can be gleaned through consideration of only the POD.

a. *Augmented versus nonaugmented ASOS observations*

The constituency of precipitation type for the augmented ASOS is depicted in Fig. 1b. The most commonly reported categories are RA/DZ at 54% and SN at 42%. The "other" category composes 4% of all observations. Figure 1c shows that FZRA, FZDZ, and RASN/DZSN each account for 20%–25% of the other category. Conspicuously small is the fraction of PL observations (5%). Though it is not unreasonable to presume that some instances of PL are missed, note that the fraction of all classes that include PL is 26%, which is comparable to FZRA. Therefore, it appears more likely that PL cases are not underreported, but occur more often in combination with other forms, consistent with previous research (Hanesiak and Stewart 1995; Cortinas et al. 2004).

There are two key differences between the augmented and nonaugmented sites. The first is that RA/DZ compose a higher fraction of the total number of observations in the nonaugmented observations (Fig. 1e). This comes at the expense of PL and PL mixes and some instances of SN and is a direct consequence of the way in which the ASOS detects RA. The system uses the hydrometeor terminal velocity, which is nearly identical for RA and PL. Large, wetted snowflakes can also have a terminal velocity similar to RA (NOAA 1998). The FZRA reports also appear to differ significantly between the augmented and nonaugmented data (Fig. 1f). However, in both systems FZRA composes about 1.25% of the *total* number of observations. As expected, PL, FZDZ, and RASN have nearly negligible percentages, as these forms of precipitation cannot be automatically detected with present technology (NOAA 1998).

b. *Augmented ASOS versus mPING observations*

The frequency of SN observations in the mPING dataset is consistent with the augmented ASOS reports (cf. Figs. 1b and 1h), but the fraction of RA/DZ is considerably less (33%) and the other category much higher (23%). The majority of other reports are PL (34%) or some combination that includes PL (27%; Fig. 1i). This is much higher than for the augmented ASOS. Conversely, FZRA and FZDZ are 11% and 12% *less* than in the augmented ASOS dataset. It is not a stretch to assume that the mPING users are less motivated to report RA/DZ, as these forms are rather commonplace, thus explaining the lower fraction for this class, but the apparent biases toward PL and against FZRA are worth additional thought.

Differences in the length of time over which observations are collected, the frequencies and locations of reports, and amount of user training could all contribute to the discrepancies between Figs. 1c and 1i. To better compare these two datasets, we mimic an analysis from Elmore et al. (2014; see their Fig. 8). In this exercise, the fraction of augmented ASOS observations that surrounds and agrees with the mPING observations of FZRA,

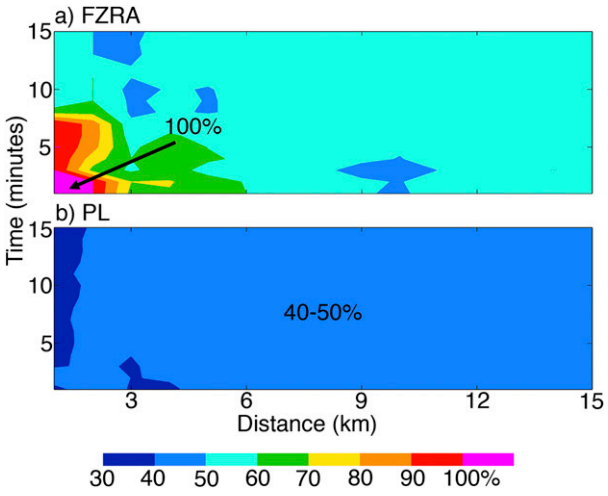

FIG. 2. The percentage of surrounding ASOS observations that agree with mPING observations of (a) FZRA and FZDZ and (b) PL.

FZDZ, and PL are computed as a function of the amount of time and distance that separates them. Since it is unclear whether citizen observers can reliably distinguish between FZRA and FZDZ, these two classes are combined. A hit is defined as any instance where the ASOS observation agrees with or is in a mix that includes the mPING observation type. Though all observations that are within 100 km and 100 min are compared, only shown are those that are within 15 km and 15 min since the fractions change by less than 2% outside of this range (Fig. 2).

The FZRA analysis shows that for most times and distances, the ASOS observations only agree with mPING about 50% of the time (Fig. 2a). As the time and distance are decreased to 2 min and 2 km, the agreement increases to 100%. This suggests that when mPING observers report FZRA, nearby augmented ASOS sites will very likely report FZRA. The same exercise can also be performed in reverse (i.e., augmented observations are compared to the surrounding mPING observations), which results in a much lower rate of agreement (ranging from 25% to 40% for pairs that are within 10 min and 10 km), indicating that mPING observations of FZRA are consistent with the ASOS, but are strongly underreported. This is not a surprising result. The ASOS system is able to detect FZRA before ice accretion is visible, which may partly account for the discrepancy, but likely of greater import is the location of the mPING observers, most of whom report from urban and suburban areas where ground temperatures may very well be above freezing even when nearby open areas have subfreezing surface temperatures.

The agreement for PL is not as good. For the span of time and distance considered, the rates of agreement range from 33% to 49% (Fig. 2b).[1] Performing this exercise in reverse gives a rate of 100% for pairs that are within 2 km and 2 min, suggesting that mPING is indeed biased toward PL relative to the augmented ASOS. There are two potential explanations for this. First, it is possible that some instances of PL are missed in the augmented reports. The fact that ASOS observers do not always augment the observations is evidenced by a small percentage of unknown precipitation (UP) reports in Fig. 1c. Another potential explanation is that some mPING users do not know what PL are. The PL option in the mPING cellphone application reads "ice pellets/sleet/graupel" or "ice pellets/sleet" depending on the reporter's operating system. The term "sleet" has a rather wide range of definitions. In Europe and some parts of the United States, it is used to refer to wet snow. In some industries, it refers to any form of precipitation that forms a coating of ice on overhead electrical wires and trees, which can happen with either SN or FZRA (Abbe 1916). These colloquial variations in the definition persist to this day (Glickman 2000). Among the augmented ASOS reports that are within 2 km and 2 min of mPING reports of PL, 65.5% are for SN or FZRA, suggesting that some mPING users probably are confusing PL with these forms of precipitation.

c. *Effects of observation biases on algorithm validation*

Let us now consider whether the PODs from the precipitation-type algorithms are affected by the above biases. To test this, observed soundings are identified that are associated with long-duration (i.e., the precipitation type does not change during the 4 h surrounding the launch time according to the augmented ASOS) SN, RA, PL, and FZRA events from December 2012 to March 2015. There are 429 such soundings. Each is run through the classifiers described above and the output compared against the observations. Only sites within 35 km and 1 h of the radiosonde launch are included. The 35-km distance is consistent with that used in Ramer (1993) and Reeves et al. (2014). The 1-h window is consistent with the typical validation period for a mesoscale model. Since several mPING observations may correspond to a single sounding, two approaches are considered. In the first, only the closest observation in time and space is used. In the second, all

[1] This range is about 25% smaller than that in Elmore et al. (2014). The discrepancy between this analysis and theirs is the length of time considered; their analysis only includes reports from December 2012 through March 2013. When this exercise is repeated using only that time frame, the analyses agree.

TABLE 1. The PODs (%) for different classification algorithms using soundings that are associated with long-duration (i.e., >4 h) SN, RA, FZRA, and PL episodes. For mPING-closest only the closest mPING in time and space are used while mPING-all uses all mPING observations within 35 km and 1 h of the sounding time and location.

	SN	RA	PL	FZRA
Augmented ASOS				
Baldwin2	89.7	87.6	64.4	16.4
Bourgouin	86.1	92.4	50.5	42.9
Ramer	86.7	87.5	28.8	47.5
Nonaugmented ASOS				
Baldwin2	91.9	83.3	60.0	18.2
Bourgouin	86.3	86.8	80.0	45.5
Ramer	88.8	82.1	20.0	54.5
mPING-closest				
Baldwin2	85.4	69.1	48.3	14.3
Bourgouin	79.2	67.9	36.7	39.1
Ramer	81.3	69.1	33.3	52.2
mPING-all				
Baldwin2	92.6	58.0	86.4	22.4
Bourgouin	86.4	56.6	73.9	31.8
Ramer	91.4	57.9	72.1	44.8

observations within 35 km and 1 h of the launch location and time are used. The PODs for all networks and algorithms are provided in Table 1.

The augmented ASOS have trends similar to what previous investigators have found (Bourgouin 2000; Manikin et al. 2004; Manikin 2005; Reeves et al. 2014). Namely, Baldwin2 has very high (low) PODs for PL (FZRA), Ramer is the opposite, and Bourgouin has similar PODs for both. All three algorithms have relatively high PODs for SN and RA. Except for PL, the PODs for the nonaugmented sites agree quite well with those for the augmented sites, suggesting that for these categories it matters little whether one distinguishes between augmented and nonaugmented observations. The PODs for PL are significantly different between the augmented and nonaugmented datasets. However, sample size prohibits a statistically meaningful interpretation as there are only five soundings that are long duration and occur coincident with a PL observation at the nearby nonaugmented site.

When validated against the closest mPING observation in time and space, the trends for RA and PL change markedly. For all classifiers, PODs for RA decrease by about 20%, with the most common errant diagnosis being for FZRA. As noted above, this is likely due to mPING observations mostly coming from urban and suburban settings. If one discounts all of the soundings with a surface-based cold layer, the PODs for RA are similar to those for the augmented ASOS reports (not shown). Given that the mPING observers have a strong PL reporting bias, the reduced PODs for this class are also expectable. Nevertheless, similar trends to the augmented ASOS emerge. For example, the POD for PL is highest for Baldwin2 and lowest for Ramer, though the spread between these PODs is not as big. The reverse is true for FZRA; Ramer has the highest PODs and Baldwin2 the lowest and the spread between them is greater than for the augmented ASOS.

When this exercise is repeated only taking into account all mPING observations within 1 h and 35 km of the radiosonde launch time and location, all algorithms appear to perform much more poorly for RA and much better for PL than what is observed for any of the other three validation approaches. The culprits are several RA and PL soundings for which there are 10 or more nearby agreeing mPING reports that unduly influence the results. The effects of this are only somewhat mitigated by reducing the distance in time and space from the radiosonde since mPING reports tend to be temporally and spatially clustered (not shown). This exercise highlights an important cautionary note regarding the use of mPING observations: when multiple observations are associated with the same profile, this can lead to misleading statistics, particularly when the multiple observations are all in agreement. In this situation, the effect is to make the algorithms appear to be much better at detecting PL and much worse at RA than they really are.

3. Temporal and spatial variability

Let us now consider the effects temporal and spatial variability have on algorithm statistics. This form of uncertainty is not the same as that investigated in section 2 (i.e., it is not a consequence of instrument or observer error), but, nevertheless, it can have profound impacts on validation approaches and statistics.

a. Temporal trends in precipitation type

In this section, the length of typical episodes of the various forms of precipitation is determined using the 5-min augmented ASOS observations. Herein, an episode is defined as a sequence of two or more consecutive reports of the same type of precipitation. According to these rules, RA/DZ and SN are comparatively long lived, each typically lasting more than 1 h and having several events lasting longer than 12 h (Fig. 3). The majority of events from the remaining classes are less than 1 h. Note that the average FZRA and FZDZ event lasts 35 and 40 min, respectively; the typical PL event only 10 min; and the typical RASN event 20 min.

Such variability has important implications for any task that uses numerical model output since the typical validation period (1 h) is longer than the average

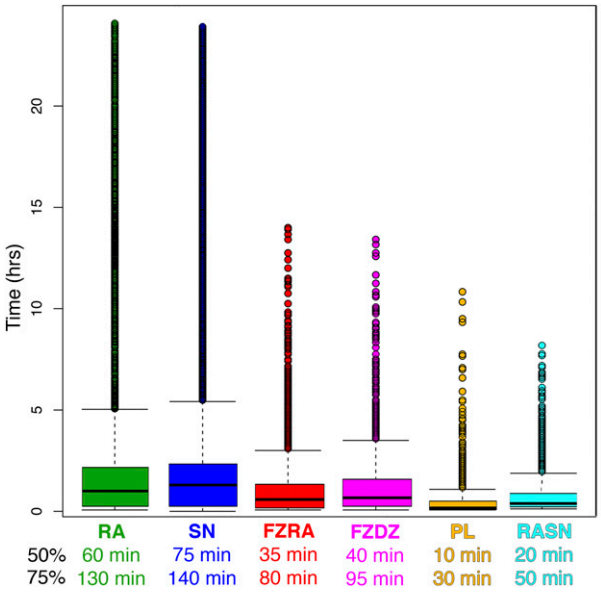

FIG. 3. Box-and-whisker plots showing the length of all episodes of RA, SN, FZRA, FZDZ, PL, and RASN, where an episode is defined as two or more consecutive reports for the same kind of precipitation. The 50th and 75th percentiles are provided beneath the graph.

duration of the transitional forms. Therefore, one is prompted to question the method by which model forecasts of precipitation type should be validated. There are three approaches one can use. The first, referred to as the top-of-the-hour (TOH) method, acknowledges that the model outputs the instantaneous fields valid at the stated time (typically at the top of the hour) and assumes that the diagnosed precipitation type should exactly agree with the observed precipitation type *at that time*. The second method, referred to as mean conditions (MC), assumes that the optimal situation occurs when the diagnosed precipitation type agrees with the most frequently observed type over the period of time for which the analysis or forecast is valid. The last method, any one counts (AOC), declares as a hit each time the diagnosed precipitation type agrees with *any* of the observed precipitation types during the validation period. For the sequence of observations in Table 2, only SN would be considered a hit for the TOH method, only FZRA for the MC method, and either SN, PL, or FZRA would be considered hits using the AOC technique.

To test the different methods described above, observed soundings associated with augmented ASOS sites where the precipitation type changes during the hour following the launch time are fed to each of the three classifiers. Only events that include FZRA or PL are used in this exercise. If the observed precipitation type according to the methods described above is a mix, that case is discarded. The POD for each method and algorithm is provided in Table 3.

TABLE 2. ASOS observations from 1100 to 1200 UTC 1 Feb 2008 at Albany, NY. The dash (-) indicates the precipitation is of light intensity.

Time (UTC)	Type
1100:31	SN
1105:31	-SNPL
1110:31	-SNPL
1115:31	PL
1120:31	PL
1125:31	-FZRAPL
1130:31	-FZRA
1135:31	-FZRA
1140:31	-FZRA
1145:31	-FZRA
1150:31	-FZRA
1155:31	-FZRA

The different methods yield different PODs in accordance with what one might expect: TOH has the lowest PODs and AOC the highest. What is surprising is the range of differences between these methods (roughly 25%–70%). Using AOC results in PODs that are far greater than what is obtained for persistent PL and FZRA events (Table 1). Even MC outperforms most of the PL and FZRA scores in Table 1. No matter which method is used, the biases in Table 1 are reduced or eliminated altogether. Such results prompt one to question whether it is fair to use all observations in an assessment of algorithm performance or only those from persistent events. Limiting an assessment to only persistent events allows one to assess whether an algorithm can adequately discriminate between these classes on those occasions when the environment is unambiguous, but ignores the majority of incidences of FZRA and PL since they are generally short lived.

b. *Spatial trends in precipitation type*

It has previously been argued that there is a high degree of horizontal variability in precipitation type when the temperature is near 0°C (Crawford and Stewart 1995; Bernstein 2000; Cortinas 2000; Rauber et al. 2000, 2001; Robbins and Cortinas 2002; Changnon 2003; Cortinas et al. 2004; Thériault et al. 2010). This has also been demonstrated using the mPING network (Reeves

TABLE 3. The PODs obtained when using the various validation methods discussed in section 3a.

	TOH	MC	AOC
Baldwin2	25.9	43.8	73.1
Bourgouin	26.5	53.0	67.4
Ramer	26.1	45.0	74.9

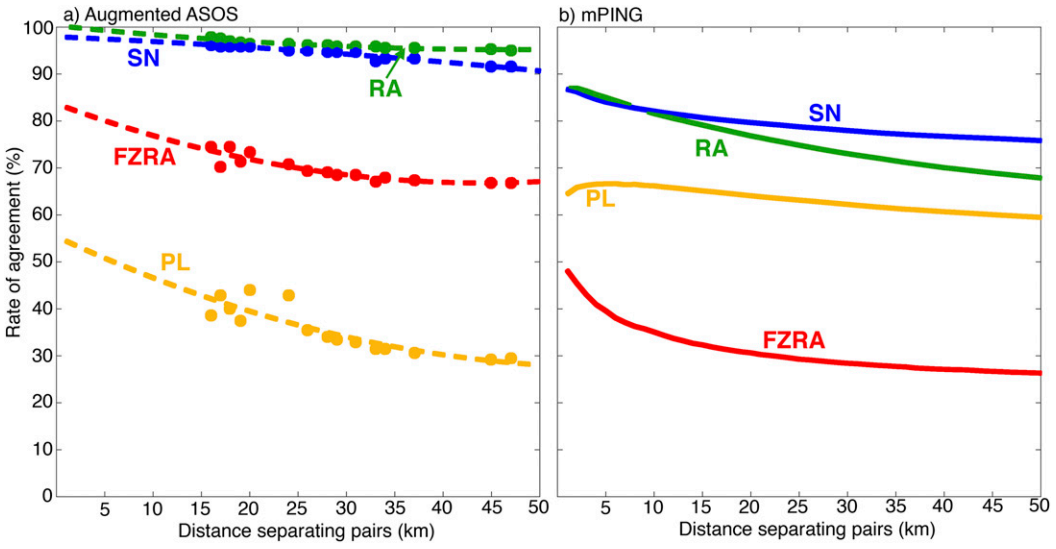

FIG. 4. The rate of agreement for (a) all augmented ASOS observations (dots) as well as a second-order least squares fit (dashed) and (b) mPING observations.

et al. 2014; EMAR). But this variability has never been quantified nor its effects on model validation examined. One might expect that SN and RA would be more horizontally uniform since either can cover rather large regions in the cold and warm sectors of midlatitude cyclones, respectively. However, FZRA and PL, since they often occur at temperatures near 0°C and are quite sensitive to small changes in temperature, precipitation rate, etc. (Thériault et al. 2010), may be more often intermixed with other forms of precipitation.

Figure 4a shows the rate of agreement between pairs of augmented ASOS observations that are taken within 10 min of each other as a function of the distance that separates them. As argued above, RA and SN have very high rates of agreement (91%–98%) for the range of distances shown. The agreement is not as good for FZRA and PL. For the pairs that are 16 km apart (the shortest distance between augmented sites used in this study), the average rate of agreement for FZRA (PL) is about 75% (39%). For comparison, a similar analysis was conducted using the mPING observations (Fig. 4b). Note that the minimum distance considered for this analysis is 1 km. The variability at subkilometer scales is not assessed since this is beyond the current capabilities of operational numerical weather prediction (NWP) models. The rates of agreement are lower for RA, SN, and FZRA and higher for PL, a consequence of the relative biases in this network. There is also a slight decrease in the rates of agreement in the mPING-PL curve for distances less than 5 km. But this decrease is not statistically significant. Overall, both networks show that the curves are comparatively flat, indicating that the distance separating pairs does not exert a strong control on the likelihood that the observations will agree. Such is counterintuitive. One might expect that two randomly selected observations that are only 5 km apart would have a much better chance of experiencing the same precipitation type than pairs that are 50 km apart. Likewise for pairs that are 50 versus 500 km apart, but this is not the case and, hence, stands as a sobering limitation on precipitation-type prediction.

The rates of agreement in Fig. 4 give some sense of the limit one can expect on the POD for precipitation-type algorithms when applied to NWP output. Because the closest ASOS pair is 16 km apart and the mPING observations have the reporting biases noted in section 2, the points in Fig. 4a are fitted by a second-order polynomial curve to extrapolate the rates of agreement for distances less than 16 km. These curves have very similar slopes to those from the mPING analysis (Fig. 4b). Assuming these curves adequately represent the actual variability, then the rates of agreement for pairs that are 8.5 km apart (the maximum distance between any observation location and the nearest grid point for a model with a 12-km spacing) are 97.1%, 98.6%, 77.8%, and 47.8% for SN, RA, FZRA, and PL, respectively. For a model with a grid spacing of 3 km, the maximum distance between a grid point and an observation is about 2 km. These percentages (97.8%, 99.9%, 82.2%, and 53.5% for SN, RA, FZRA, and PL) are slightly improved but still suggest that one cannot expect high PODs for PL.

The above results suggest that as a model's grid spacing is increased, its ability to resolve the various

TABLE 4. The PODs for the 1-h HRRR forecasts for select events using the native 3-km grid spacing and then thinning to 12, 24, and 36 km.

		SN	RA	PL	FZRA
Baldwin2	3 km	86.0	87.4	11.6	42.1
	12 km	86.4	86.1	11.6	41.2
	24 km	86.0	86.4	12.2	42.3
	36 km	86.0	84.1	12.2	42.7
Bourgouin	3 km	86.4	86.6	35.4	36.7
	12 km	84.9	86.9	32.0	36.0
	24 km	84.9	86.6	33.3	36.2
	36 km	81.1	85.1	37.4	38.9
Ramer	3 km	88.5	90.2	58.5	39.5
	12 km	88.9	92.0	57.1	38.4
	24 km	88.1	90.7	59.2	40.1
	36 km	88.2	91.7	55.8	37.8

regions of SN, RA, FZRA, and PL is not markedly changed. To gauge whether this is true, 1-h High-Resolution Rapid Refresh (HRRR; Brown et al. 2011) forecasts are obtained for those mPING events with 200 or more reports. There are 281 hourly and 80 881 mPING observations that meet these criteria. The effects of changing the grid resolution are tested by thinning the HRRR forecasts from their native grid spacing of 3 km to grid spacings of 12, 24, and 36 km. Though the mPING observations have some biases when compared to the augmented ASOS and model uncertainty is nontrivial for some forms of precipitation (e.g., Reeves et al. 2014), the interest here is in relative rather than absolute statistics.

The PODs for each algorithm and each grid spacing are provided in Table 4. Regardless of the precipitation type or algorithm used, *as the grid spacing is increased, the PODs do not significantly change*. Therefore, all else being equal, forecast accuracy for precipitation is not a function of the model grid spacing, at least for the range of grid spacings considered herein. However, there are three caveats worth noting with the above approach. First, preliminary work by Johnson and Shepherd (2016) indicates that there is a preference for warmer forms of precipitation to occur in urban areas. Indeed, the mPING observations do indicate that the greatest variability may exist in urban and suburban areas (not shown). Urban heat effects may not be well handled, even in a 3-km model. Second, the approach used here is intentionally designed to isolate the effects of grid spacing from other possible contributors, such as differences that occur as the result of cumulus parameterization or the development of scale-dependent flow patterns in one forecast versus another. Comparisons of models that are identical save for the grid spacing and use of a cumulus parameterization scheme at longer lead times may very well have different validation statistics. Third, this study neglects the validation in complex terrain, which may benefit from increased resolution. Work is under way to investigate this latter issue.

4. Conclusions

The effects of uncertainty in observations of precipitation type on the apparent performance of precipitation-type algorithms are assessed. Three datasets are considered: augmented ASOS, nonaugmented ASOS, and the Meteorological Phenomena Identification Near the Ground (mPING). Three different algorithms that represent a range of inherent biases are used in the assessment. These are Baldwin2, Bourgouin, and Ramer (Baldwin et al. 1994; Bourgouin 2000; Ramer 1993). Though some reporting errors may have occurred, only augmented ASOS sites where 95% or more of the reports associated with precipitation have an augmentation flag are included, thus allowing this dataset to be used as the ''truth'' against which the other networks are compared.

The frequencies of the different precipitation types vary by network. The augmented ASOS dataset is dominated by RA/DZ and SN. Of the remaining categories, RASN, FZDZ, FZRA, and PL/PL mixes are reported at near-equal frequencies. The nonaugmented ASOS results suffer from an inability to report mixes, freezing drizzle, and ice pellets and so tends to be biased toward RA, but otherwise agrees well with the augmented ASOS. The mPING network has a lower frequency of RA reports, which is likely due to the fact that this form of precipitation is not as novel and, hence, may not inspire as many reports. It is also biased toward PL and away from FZRA.

The effects of network biases on the probability of detection (POD) from the precipitation-type classifiers is not significant for the nonaugmented ASOS, excepting for PL, which for obvious reasons, is not well validated when using this network. However, when validated against mPING, the PODs for RA and PL have marked decreases. There are many situations where FZRA is diagnosed at a time and location where RA is reported by the mPING network. This appears to be due to the preference for mPING reports to come from urban and suburban settings, where sensible heat sources may prevent the surface temperature from dipping below 0°C. When mPING observations of RA that correlate to soundings with a surface-based subfreezing layer are removed, the PODs agree more closely with the augmented ASOS data. The lower PODs for PL are a consequence of the mPING's apparent bias toward that form of precipitation. These results are only obtained when the mPING network is thinned to include only the

closest observation in time and space to the sounding. Using all observations results in strongly reduced PODs for RA and marked increases in PODs for PL.

Since crowdsourcing is a growing trend for collecting meteorological observations, it is appropriate at this juncture to say a few words about reports made by untrained observers and the degree of reliability one should ascribe to these types of data. In the case of mPING, quality control is especially difficult. It is possible for malicious users to log in and purposely make false reports, but the bigger issue appears to be education of the users and their ability to properly distinguish between the various forms of precipitation. Even if they are able to correctly identify the precipitation type, it is unclear what criteria they are employing to decide what form of precipitation to report. For example, ASOS instruments will only report FZRA when ice accretion begins at 2 m above ground level. It is possible that mPING users are gauging the presence of FZRA by looking at power lines, treetops, or even media reports. One must also be wary of the overrepresentation of the observations; it is possible for a user at one location to make continuous reports of the same kind of precipitation over an extended period of time. With ASOS, one would customarily choose only the closest observation in time, but with mPING, since each observation is treated individually (as opposed to coming from one source), one must be careful to ensure overrepresentation does not occur.

Temporal variability is another source of uncertainty considered herein. While RA and SN are comparatively long lived, FZRA, PL and other transitional forms of precipitation typically last less than an hour. This is problematic when the model validation period is one or more hours in duration. Different validation approaches were considered to assess whether they have an effect on the apparent performance of the classifiers. Depending on how a hit is defined, PODs ranging from about 25% to 75% are obtained. These PODs differ quite dramatically from those obtained for long-duration events, making all of the algorithms appear to perform much more accurately than they really do.

Spatial variability is also considered. In comparison to RA and SN, FZRA and PL have rather high variability, meaning that they are more likely to be observed in proximity to other forms of precipitation. The spatial variability of PL, according to the augmented ASOS, is sufficiently high that even for a model with a 3-km grid spacing one cannot expect the PODs for PL to exceed about 54%. The spatial trends according to the mPING observations are somewhat different, suggesting that FZRA is the most variable and that the maximum POD one can expect for this category is about 48%. Nevertheless, the variability for all of the precipitation types is only weakly dependent on the distance. This is true for both the augmented ASOS and mPING networks and suggests increasing the horizontal grid spacing of a model will not significantly impact its ability to predict the various forms of precipitation. Simple tests with the HRRR model corroborate this assumption.

It is this author's opinion that to truly measure the accuracy of an algorithm, one should restrict one's self to unambiguous observations (i.e., cases where the temporal and spatial representativity is resolved by the model in question). However, this says nothing of an algorithm's performance for the vast majority of events in which a great deal of ambiguity does exist. One approach for mitigating this is to include mix classes in the algorithms. Another, perhaps more desirable approach, is to use ensemble prediction to provide probabilistic forecasts of precipitation type, thus providing the forecaster with greater insight into the likelihood of wintry mixes verses long-lived episodes of FZRA and PL.

Acknowledgments. 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. Special thanks to the internal reviewer. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce.

REFERENCES

Abbe, C., Jr., 1916: American definition of "sleet." *Mon. Wea. Rev.*, **44**, 281–286, doi:10.1175/1520-0493(1916)44<281:ADOS>2.0.CO;2.

Baldwin, M., R. Treadon, and S. Contorno, 1994: Precipitation type prediction using a decision tree approach with NMC's 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, doi:10.1175/WAF-D-15-0136.1.

Bernstein, B. C., 2000: Regional and local influences on freezing drizzle, freezing rain, and ice pellet events. *Wea. Forecasting*, **15**, 485–508, doi:10.1175/1520-0434(2000)015<0485:RALIOF>2.0.CO;2.

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

Brown, J. M., and Coauthors, 2011: Improvement and testing of WRF physics options for application to Rapid Refresh and High Resolution Rapid Refresh. Preprints, *14th Conf. on Mesoscale Processes/15th Conf. on Aviation, Range, and Aerospace Meteorology*, Los Angeles, CA, Amer. Meteor. Soc., 5.5. [Available online at https://ams.confex.com/ams/14Meso15ARAM/webprogram/Paper191234.html.]

Changnon, S. A., 2003: Urban modification of freezing-rain events. *J. Appl. Meteor.*, **42**, 863–870, doi:10.1175/1520-0450(2003)042<0863:UMOFE>2.0.CO;2.

Chenard, M., P. N. Schumacher, and H. D. Reeves, 2015: Determining precipitation type from maximum temperature in the lower atmosphere. Preprints, *23rd Conf. on Numerical Weather Prediction/27th Conf. on Weather Analysis and Forecasting*, Chicago, IL, Amer. Meteor. Soc., 6B2. [Available online at https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273342.html.]

Cortinas, J. V., Jr., 2000: A climatology of freezing rain in the Great Lakes region of North America. *Mon. Wea. Rev.*, **128**, 3574–3588, doi:10.1175/1520-0493(2001)129<3574:ACOFRI>2.0.CO;2.

——, B. C. Bernstein, C. C. Robbins, and J. W. Strapp, 2004: An analysis of freezing rain, freezing drizzle, and ice pellets across the United States and Canada: 1976–1990. *Wea. Forecasting*, **19**, 377–390, doi:10.1175/1520-0434(2004)019<0377:AAOFRF>2.0.CO;2.

Crawford, R. W., and R. E. Stewart, 1995: Precipitation type characteristics at the surface in winter storms. *Cold Reg. Sci. Technol.*, **23**, 215–229, doi:10.1016/0165-232X(94)00014-O.

Czys, R., R. Scott, K. C. Tang, R. W. Przybylinski, and M. E. Sabones, 1996: A physically based, nondimensional parameter for discriminating between freezing rain and ice pellets. *Wea. Forecasting*, **11**, 591–598, doi:10.1175/1520-0434(1996)011<0591:APBNPF>2.0.CO;2.

Elmore, K. L., and H. M. Grams, 2015: Using mPING data to drive a forecast precipitation type algorithm. Preprints, *13th Conf. on Artificial Intelligence*, Phoenix, AZ, Amer. Meteor. Soc., TJ1.4. [Available online at https://ams.confex.com/ams/95Annual/webprogram/Paper266757.html.]

——, Z. L. Flamig, V. Lakshmanan, B. T. Kaney, H. D. Reeves, V. Farmer, and L. P. Rothfusz, 2014: mPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342, doi:10.1175/BAMS-D-13-00014.1.

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

Glickman, T. S., Ed., 2000: *Glossary of Meteorology*. 2nd ed. Amer. Meteor. Soc., 885 pp.

Hanesiak, J. M., and R. E. Stewart, 1995: The mesoscale and microscale structure of a severe ice pellet storm. *Mon. Wea. Rev.*, **123**, 3144–3162, doi:10.1175/1520-0493(1995)123<3144:TMAMSO>2.0.CO;2.

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, doi:10.1175/WAF-D-12-00085.1.

Johnson, B., and J. M. Shepherd, 2016: Assessing urban impact on winter precipitation type using dual polarization radar. Preprints, *30th Conf. on Hydrology*, New Orleans, LA, Amer. Meteor. Soc., J20.4 [Available online at https://ams.confex.com/ams/96Annual/webprogram/Paper286394.html.]

Manikin, G. S., 2005: An overview of precipitation type forecasting using NAM and SREF data. Preprints, *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. [Available online at https://ams.confex.com/ams/WAFNWP34BC/techprogram/paper_94838.htm.]

——, K. F. Brill, and B. Ferrier, 2004: An Eta model precipitation type mini-ensemble for winter weather forecasting. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 23.1. [Available online at https://ams.confex.com/ams/84Annual/techprogram/paper_73517.htm.]

NOAA, 1998: Automated Surface Observing System user's guide. National Oceanic and Atmospheric Administration, 61 pp + appendixes. [Available online at http://www.nws.noaa.gov/asos/pdfs/aum-toc.pdf.]

Ralph, F. M., and Coauthors, 2005: Improving short-term (0–48 h) cool-season quantitative precipitation forecasting: Recommendations from a USWRP workshop. *Bull. Amer. Meteor. Soc.*, **86**, 1619–1632, doi:10.1175/BAMS-86-11-1619.

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

Rauber, R. M., L. S. Olthoff, and M. K. Ramamurthy, 2000: The relative importance of warm rain and melting processes in freezing precipitation events. *J. Appl. Meteor.*, **39**, 1185–1195, doi:10.1175/1520-0450(2000)039<1185:TRIOWR>2.0.CO;2.

——, ——, ——, and K. E. Kunkel, 2001: Further investigation of a physically based, nondimensional parameter for discriminating between locations of freezing rain and ice pellets. *Wea. Forecasting*, **16**, 185–191, doi:10.1175/1520-0434(2001)016<0185:FIOAPB>2.0.CO;2.

Reeves, H. D., K. L. Elmore, A. Ryzhkov, T. Schuur, and J. Krause, 2014: Source of uncertainty in precipitation-type forecasting. *Wea. Forecasting*, **29**, 936–953, doi: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, doi:10.1175/JAMC-D-16-0044.1.

Robbins, C. C., and J. V. Cortinas Jr., 2002: Local and synoptic environments associated with freezing rain in the contiguous United States. *Wea. Forecasting*, **17**, 47–65, doi:10.1175/1520-0434(2002)017<0047:LASEAW>2.0.CO;2.

Scheuerer, M., S. Gregory, T. Hamill, P. Shafer, and G. Wagner, 2016: Probabilistic forecasts of precipitation type based on global ensemble forecasts. Preprints, *23rd Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., P7.3 [Available online at https://ams.confex.com/ams/96Annual/webprogram/Paper283244.html.]

Schuur, T. J., H.-S. Park, A. V. Ryzhkov, and H. D. Reeves, 2012: Classification of precipitation types during transitional winter weather using the RUC model and polarimetric radar retrievals. *J. Appl. Meteor. Climatol.*, **51**, 763–779, doi:10.1175/JAMC-D-11-091.1.

Shafer, P. E., and M. S. Antolik, 2016: Development of a new NAM-based MOS precipitation type system. Preprints, *23rd Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., P1.4 [Available online at https://ams.confex.com/ams/96Annual/webprogram/Paper285539.html.]

Thériault, J. M., R. E. Stewart, and W. Henson, 2010: On the dependence of winter precipitation types and temperature, precipitation rate, and associated features. *J. Appl. Meteor. Climatol.*, **49**, 1429–1442, doi:10.1175/2010JAMC2321.1.

Thompson, E. J., S. A. Rutledge, B. Dolan, V. Chandrasekar, and B.-L. Cheong, 2014: A dual-polarized radar hydrometeor classification algorithm for winter precipitation. *J. Atmos. Oceanic Technol.*, **31**, 1457–1481, doi:10.1175/JTECH-D-13-00119.1.

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, doi:10.1175/WAF871.1.