Verifying Forecast Precipitation Type with mPING*

KIMBERLY L. ELMORE AND HEATHER M. GRAMS

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

DEANNA APPS⁺

State University of New York at Oswego, Oswego, New York

HEATHER D. REEVES

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

(Manuscript received 24 June 2014, in final form 2 March 2015)

ABSTRACT

In winter weather, precipitation type is a pivotal characteristic because it determines the nature of most preparations that need to be made. Decisions about how to protect critical infrastructure, such as power lines and transportation systems, and optimize how best to get aid to people are all fundamentally precipitation-type dependent. However, current understanding of the microphysical processes that govern precipitation type and how they interplay with physics-based numerical forecast models is incomplete, degrading precipitation-type forecasts, but by how much? This work demonstrates the utility of crowd-sourced surface observations of precipitation type from the Meteorological Phenomena Identification Near the Ground (mPING) project in estimating the skill of numerical model precipitation-type forecasts and, as an extension, assessing the current model performance regarding precipitation type in areas that are otherwise without surface observations. In general, forecast precipitation type is biased high for snow and rain and biased low for freezing rain and ice pellets. For both the North American Mesoscale Forecast System and Global Forecast System models, Gilbert skill scores are between 0.4 and 0.5 and from 0.35 to 0.45 for the Rapid Refresh model, depending on lead time. Peirce skill scores for individual precipitation types are 0.7–0.8 for both rain and snow, 0.2–0.4 for freezing rain and freezing rain, and 0.25 or less for ice pellets. The Rapid Refresh model displays somewhat lower scores except for ice pellets, which are severely underforecast, compared to the other models.

1. Introduction

Winter precipitation clearly presents a threat to human health and safety, but also results in significant economic losses, with ice pellets and freezing rain posing particularly dangerous threats. Goodwin (2003)

estimates that about 7000 deaths, 600 000 injuries, and 1.4 million accidents per year can be attributed to the effects of winter precipitation on roadways. Changnon (2003a) notes that \$22.2 billion (inflation adjusted to 2014 U.S. dollars) in economic losses can be attributed to just 87 freezing rain events that occurred between 1949 and 2000. Precipitation-type prediction was a focus of a recent U.S. Weather Research Program initiative on winter weather (Ralph et al. 2005). Despite the obvious need for improvements, only limited attention has been paid to the current state of wintertime precipitation-type (ptype) forecasting in the United States. In this work, we assess the performance of precipitation-type forecasts from three operational numerical weather prediction models. As we demonstrate later, all forecast models

* Supplemental information related to this paper is available at the Journals Online website: <http://dx.doi.org/10.1175/WAF-D-14-00068.s1>.

⁺ Additional affiliation: National Weather Center Research Experiences for Undergraduates Program, Norman, Oklahoma.

Corresponding author address: Dr. Kimberly L. Elmore, NSSL, 120 David L. Boren Blvd., Norman, OK 73072.
E-mail: kim.elmore@noaa.gov

provide reliable guidance on snow and rain, but skill scores are substantially worse for freezing rain and ice pellets.

There are four types of precipitation that occur at the surface in winter storms: snow, rain, ice pellets, and freezing rain (SN, RA, PL, and FZRA). It is possible for multiple forms to occur simultaneously, such as an FZRA–PL mix. The type(s) of precipitation at ground level is determined by a number of factors including surface temperature, the vertical profiles of temperature and humidity, precipitation rate, drop size distribution, and latent heating/cooling effects (e.g., [Stewart and King 1987](#); [Kain et al. 2000](#); [Lackmann et al. 2002](#); [Theriault et al. 2010](#)). Subtle changes in any of the above can lead to a change in the surface precipitation type. Small-scale forcing, such as urban heat effects or heat transfer from water bodies, can also lead to a local alteration of the precipitation type ([Changnon 2003b](#)), sometimes resulting in a heterogeneous mix of various precipitation types.

Precipitation-type forecasts in numerical models are typically performed by one or more postprocessing algorithms. Explicit algorithms use the predicted mixing ratios of snow, graupel, and rain in concert with vertical profiles of temperature to assign the type. The North American Mesoscale Forecast System (NAM; [Janjic et al. 2005](#)) uses an explicit scheme as part of a “mini-ensemble” of precipitation-type methods. The implicit members of the mini-ensemble are the Baldwin1, Baldwin2 (known simply as “NCEP” and “NCEP Revised” at NCEP and within NCEP documentation), Bourgoiuin, and Ramer algorithms ([Ramer 1993](#); [Baldwin et al. 1994](#); [Bourgoiuin 2000](#)). The NAM explicit scheme starts with explicit mixing ratios of frozen and liquid particles. If less than 50% of all particles are frozen and the skin temperature is greater than 0°C, the method produces RA; if the skin temperature is less than 0°C, the method produces FZRA. If greater than 50% of all particles are frozen and the riming factor is high, the method produces PL; if the riming factor is low, it produces SN. The Rapid Refresh and High-Resolution Rapid Refresh (RAP and HRRR; [Brown et al. 2011](#)) models use only an explicit classifier derived from, but not identical to, that of [Thompson et al. \(2004, 2008\)](#). The performance of this algorithm within the HRRR model was assessed in [Ikeda et al. \(2013\)](#), who found it yields reliable forecasts of SN and RA, but that the prediction of transition lines between SN and RA and/or transitional forms, such as PL and FZRA, is not as good.

Implicit algorithms rely solely on predicted temperature and humidity profiles. Both the NAM and Global Forecast System (GFS; [Moorthi et al. 2001](#)) models predominantly rely on implicit classifiers. Previous

studies investigating the quality of various implicit classifiers show that some algorithms have strong biases toward one form of precipitation or perform better in some situations than in others ([Manikin et al. 2004](#); [Manikin 2005](#); [Wandishin et al. 2005](#)). Implicit classifiers also tend to yield substantially worse forecasts of PL and FZRA than for SN or RA ([Bourgoiuin 2000](#); [Reeves et al. 2014](#)). This tendency is due in part to poor assumptions made by different classifiers, but is also a consequence of model uncertainty ([Reeves et al. 2014](#)). The comparatively poor performance for PL and FZRA has prompted the National Weather Service to adopt an ensemble technique for the NAM and GFS models. In the ensemble approach, five different classifiers are used in the NAM, four in the GFS, and the dominant category is the declared precipitation type ([Manikin et al. 2004](#)). However, the statistical analyses in [Reeves et al. \(2014\)](#) suggest this method does not provide significant improvements over any one classifier’s results. Even so, the original NCEP algorithm had a distinct and intentional bias toward FZRA and PL. This caused problems for operational forecasters and the ensemble method is intended to assist in these particular cases. Such cases may be rare in the [Reeves et al. \(2014\)](#) dataset and so the intended improvement may not be evident.

The mini-ensemble used by the GFS uses the same four implicit schemes as does NAM. With the ensemble approach, it is possible to have ties between classes; for example, two methods could result in SN, two methods in FZRA, and in the case of the NAM the fifth method could be something different yet. In these circumstances, the assigned category is based on which forms have the greatest severity. Any tie with FZRA will result in an FZRA class. Any tie without FZRA, but with SN, will be classified as SN. Any tie consisting of only RA and PL will be classified as PL. Thus, from lowest to highest precedence, NCEP ptypes are RA < PL < SN < FZRA.

A key way in which the assessment and development of classifiers is hampered is in the observations of precipitation type at the surface. Previous investigators have used the Automated Surface Observing System (ASOS) network as ground truth (e.g., [Bourgoiuin 2000](#); [Manikin et al. 2004](#); [Manikin 2005](#); [Wandishin et al. 2005](#); [Ikeda et al. 2013](#); [Reeves et al. 2014](#)). This network has some deficiencies that make it flawed for ptype verification work. For example, the only condition under which PL is reported in the ASOS data is if a human observer augments the report. In 2011 there were 852 ASOS sites across the United States. Of these, there were 73 service level A (24-h dedicated contract observer with full augmentation and backup responsibilities), 55 service level B (24-h contract observers with more limited augmentation and backup responsibility), 296 service level C (variable

hours for contract observers; do not provide backup observations of PL), and 428 service level D (no augmentation or backup observation available) stations. Only A and B level stations report PL. Thus, 85% of ASOS stations do not report PL as a present weather type.

Reeves et al. (2014) note that many cases of RA are reported in the presence of a deep surface-based sub-freezing layer. To eliminate these false reports, one must manually quality control the observations using vertical profiles of temperature and humidity, but sounding data are rarely available at the same time and location as ASOS reports. Another limitation of the ASOS reports is their spatial representativeness; these observations are rather sparsely distributed. Elmore et al. (2014) and Reeves et al. (2014) demonstrate that it is possible for various forms of precipitation to coexist in a small geographical area. The ASOS data give the user little information as to the horizontal extent of a particular form of precipitation.

The Meteorological Phenomena Identification Near the Ground (mPING; Elmore et al. 2014) project was launched on 19 December 2012 and uses crowd sourcing to provide spatially and temporally dense observations of precipitation type across the United States. These observations provide a way to validate numerical model output at a large number of points in time and space, thus yielding a more robust measure of model performance in forecasting precipitation type. These observations potentially allow for a more comprehensive and thorough assessment of how any one model performs. We do this for three different forecast models (RAP, NAM, and GFS) and for lead times ranging from 3 to 24 h.

2. Verification data, models, and algorithms

The analysis for this study combines two periods—from 0000 UTC 19 December 2012 to 2300 UTC 30 April 2013 and from 0000 UTC 1 October 2013 to 2300 UTC 31 January 2014—and contains 126 576 ptype observations across approximately 928 cycles of both the NAM and GFS and 1846 cycles of the RAP. Forecast data in GRIB2 format for the NAM and GFS come from NOAA's Operational Model Archive and Distribution System (NOMADS) for 3-hourly intervals from 3 to 24 h, and forecasts from the RAP are for 3-hourly intervals from 3 to 18 h. Although the RAP is available hourly, only forecast hours also available in the NAM and GFS forecasts are evaluated. Thus, the only forecast times evaluated were ones at which all three models had a forecast available. The analysis is performed at the native grid resolution for each model (13 km for the RAP, 12 km for the NAM, and 27 km for the GFS) using the nearest grid point that contains a precipitation type, which is consistent with mPING reports since mPING

has no precipitation accumulation threshold. Thus, each mPING report is assessed at a single model grid point.

Verification data come from the mPING project (Elmore et al. 2014) wherein “citizen scientists” use a mobile app to send reports of the precipitation type they are experiencing. The data consist of GPS time (in UTC), GPS location, precipitation type, and platform operating system (OS; either iOS or Android). There are currently 11 different precipitation types for mPING observers to report: drizzle, freezing drizzle, rain, freezing rain, ice pellets–sleet, snow, mixed rain and snow, mixed rain and ice pellets, mixed ice pellets and snow, hail, and none. Hail is not considered for this study, and the “none” category, while potentially useful, is also not considered because this study only verifies ptype, not whether precipitation is occurring.

The models used in this study diagnose only SN, RA, PL, and FZRA (and some mixes in the case of the RAP model). Therefore, to allow for a more direct comparison of the model-predicted and observed precipitation types, some manipulation of the observations is necessary. While there are several ways to collapse the mPING types into four classes, two seem particularly intuitive. In the first, called here the rain-biased or simply RA-biased method, rain, drizzle, rain mixed with snow, and rain mixed with PL are cast as RA. Snow and ice pellets mixed with snow are cast as SN. Freezing rain and freezing drizzle are cast as FZRA while ice pellets remain PL. At one point, graupel was merged with PL, but graupel has since been completely removed in this analysis because there is uncertainty about how mPING users were using the graupel category. In the second method of collapsing mPING reports into four classes, called the ice-pellet-biased or simply PL-biased method, rain and drizzle are cast as rain; snow, wet snow, and rain mixed with snow are all cast as SN; freezing rain and freezing drizzle remain cast as FZRA; while ice pellets, rain mixed with ice pellets, and snow mixed with ice pellets are cast as PL. In a sense, the rain-biased method tends to group together mixes with the more common constituent while the ice-pellet-biased method tends to group together mixes with the most novel constituent.

A Voronoi, or nearest neighbor, type analysis is performed to assess which collapse method is most robust. In this exercise only, all observations within a given radius of a particular observation and within 15 min of when that observation is taken are considered and the fraction of observations that agree with the particular observation is calculated. The collapse that results in the highest consistency is chosen. Only 15-min periods with at least 20 observations are used, as these tend to occur when there are significant areas having winter precipitation that is likely to be captured in the model

forecasts and because these periods provide sufficient density for such an analysis. There are 4384 such periods (1096 total combined hours) comprising 163 124 individual observations, with an average (median) number of observations per period of 37.2 (30) and a maximum of 257.

Simple categorical agreements for all 11 categories are shown in Fig. 1a. Overall agreement is less when using all 11 of the possible categories simply because there are more ways for observations to vary between categories. Overall agreement is slightly higher for the PL-biased method. Agreement at 0 km is low because users often enter a new ptype observation if the ptype changes within a 15-min period. Clearly, limiting the number of categories to four increases the overall consistency of the observations, though there is no clear signal that one method is more consistent than the other.

Figure 1b shows a similar diagram but for only when neighboring observations of RA match, thus making the comparison for RA versus not RA. Even in this case, the PL-biased collapse is more consistent. Figure 1c is for SN and shows that SN is perhaps more spatially and temporally homogeneous than is RA. Again, the PL-biased method appears to be slightly more consistent.

Finally, Fig. 1d shows results for only the PL category and here a large difference between the two methods is apparent. The PL-biased collapse produces far more spatial and temporal consistency than does the RA-biased method. In fact, when examining the PL category in isolation, the RA-biased method performs no better than when using no collapse at all.

Vexing questions arise when the FZRA category is isolated, however (Fig. 1e; note different vertical scale), because the overall agreement is considerably worse for FZRA observations than for any of the other three classes. Two obvious explanations come to mind: 1) observers are erroneously entering RA instead of FZRA because they either do not know the outside temperature or they do not go outside to see if precipitation is actually freezing upon objects, and 2) observations are reliable and FZRA simply has a high spatial and temporal variability, a characteristic noted in Robbins and Cortinas (2002), Changnon (2003a,b), Changnon and Creech (2003), and Changnon and Karl (2003). This is a serious issue because if the observations are of poor quality, any implications for the quality of model-diagnosed ptype using these data are immediately suspect.

If explanation 1 is correct, most of the disagreeing observations should appear as RA. While such a result would not prove that most FZRA observations are erroneous, it would cast considerable doubt upon their quality. Alternatively, if most of the disagreeing observations are not

RA, then the FZRA observations do not suffer from consistent errors and so are likely of high quality.

In total, 80 628 individual mPING observations appear in the Voronoi analysis and 3114 of these are FZRA. These 3114 FZRA observations form 21 542 pairs with other mPING observations that are separated by no more than 50 km. Within these pairs 8914 (41.4%) consist only of freezing rain. Of the remaining 12 628 pairs, 17.4% (2200) of the “not FZRA” observations are RA, 23.4% (2954) are SN, and 59.2% (7474) are for PL. While the number of individual observers cannot be known, it is certainly less but probably proportional to the number of observations. That such a large proportion of observations (and so probably observers) would routinely and systematically misclassify FZRA as either PL or SN is implausible. Therefore, we conclude that mPING observations of FZRA, and by implication for the other categories, are indeed reliable.

In an effort to better assess the quality of the mPING reports, they are compared to data from the subset of 68 ASOS stations used in Reeves et al. (2014). Of the four basic precipitation types, only PL cannot be automatically generated by the ASOS equipment. But contract observers do equate to dedicated observers. In Federal Aviation Administration (FAA) Order 7900.5B “Surface Weather Observing,” four different ASOS service levels, A–D, are defined. For all but service level A, observers perform augmentation and backup as duties allow. Only service level A and B stations can generate present weather types of PL. The 68 ASOS stations used here are a mix of service levels A and B. But FAA Order 7900.5 explicitly mentions instances under which ASOS present-weather backup observations will be in error. Specifically, appendix D, section 3 states “Augmentation and backup at A, B, and C locations is provided by a combination of Federal and non-Federal personnel and existing contract weather observers through implementation of an ASOS basic weather watch. During a basic weather watch, the observer may be required to perform other duties as their observing workload permits. Because of this and other restrictions (station location, structural design, etc.) which limit the observer’s capability to continuously view and evaluate weather conditions, observers performing a basic weather watch cannot be expected to detect and report all weather changes as they occur.”

In addition, ASOS SPECI (special) augmentation may occur no more frequently than every 10 min and there may be no more than six SPECI issuances per hour. Thus, ASOS misses are expected, especially in the case of PL. There is no documentation available that shows an analysis of how often this occurs. However, this probably occurs enough to have an effect on any comparison between mPING and ASOS, even if mPING observations are

FIG. 1. (a) Overall categorical agreement. Blue shows agreement between each of the 11 possible categories. Red shows agreement between the four resulting categories when using a rain-biased collapse and green when using an ice-pellet-biased collapse. Light lines are raw values; heavy lines are smoothed values. (b)–(e) As in (a), but for only the RA, SN, PL, and FZRA categories, respectively. Note that for (e), the vertical scale is different and only two traces are shown because FZRA is unaffected by the nature of the four-class collapse.

TABLE 1. Analysis of agreement between ASOS and mPING ptype observations. This confusion matrix is ASOS oriented, meaning that ASOS results are considered the observations (columns) and mPING results the predictions of the ASOS report.

mPING obs	ASOS obs				Total
	RA	SN	PL	FZRA	
RA	27 982	2135	596	987	31 700
SN	3355	41 836	2132	512	47 835
PL	2145	3100	3438	1900	10 583
FZRA	898	586	474	2002	3960
Total	34 380	47 657	6640	5401	94 078

perfect. Overall, more mPING observations of PL than ASOS observations of PL may be reasonably expected.

However, mPING does not provide for FZRAPL (freezing rain with ice pellets) mixes while ASOS does. In such cases, mPING observers are required to choose which component to report. Based on other analyses and anecdotal evidence we suspect, but cannot prove, that mPING observers will tend to report the “novel” component of a mix if they cannot explicitly report the mix. Since mPING does not currently support FZRAPL mixes, we suspect that in such mixes mPING reports are biased toward PL.

Table 1 shows results of all available 5-min ASOS reports using mPING observations collected during the 5 min prior to the ASOS report time and within 50 km of the ASOS site. All observations are collapsed into the four major classes (RA, SN, PL, and FZRA). There is nothing meteorological driving the choice of 50 km; that radius simply collects enough mPING reports for meaningful statistics in all constituent categories. There is no minimum scale for which ptype is homogeneous and it is well known that ptype in winter transition zones can display high local variability.

Table 1 is ASOS oriented, meaning that ASOS may be considered the observation (columns) and mPING may be considered the prediction of the ASOS report. A notable characteristic is the “bias” in mPING reports, which are 0.922, 1.000, 1.590, and 0.733 for RA, SN, PL, and FZRA, respectively. Biases for PL and FZRA mean that there are more observations of PL from mPING than from ASOS yet fewer observations of FZRA. There is no significant bias in reports of RA or SN. In this four-class collapse, ASOS classes of FZRAPL and FZRAPLSN (freezing rain with ice pellets and snow) are merged into FZRA. There is already reason to suspect that ASOS will miss some instances of PL though there is no way of knowing how many. There is also reason to suspect that mPING misses instances of freezing rain within mixes because there is no way for mPING to report either FZRAPL, FZRAPLSN, or FZRASN (freezing rain with snow), all of which would

be collapsed into FZRA for mPING purposes. Again, how often this happens is unknown.

Finally, biases in PL and FZRA are nearly offsetting, meaning that moving some FZRA observations into the PL category and some PL observations into the FZRA category would improve the bias. This leads to a hypothesis that the disagreement between ASOS and mPING for these two categories is a complex interaction between the flexibility of mPING observations (not limited to every 10 min), intrinsic misses of PL by ASOS, and intrinsic misses of FZRA by mPING. Despite these issues, mPING observations are deemed to be of sufficient quality to demonstrate their use in a verification exercise.

Models and algorithms

The classifier used by the RAP is briefly described in section 1, but in more detail in Ikeda et al. (2013; see their Table 1). It uses the forecast rain, snow, and graupel mixing ratios on the lowest model level as the primary discriminator between the categories. Therefore, it is strongly modulated by the choice of the microphysical parameterization, which for this model is version 3.2 of the Thompson microphysical parameterization scheme. As noted above, this classifier allows for multiple forms of precipitation at a single grid point. The same logic used for collapsing mixed types for the mPING observations is used for these data. However, the RAP has one additional mix not available in mPING: a SN–FZRA mix. In this study, these mixes are collapsed into the FZRA category.

An mPING observation is matched to the closest model grid point location within ± 30 min of the forecast valid time. Observations are matched to model grid points only when the model produces more than a trace of precipitation. Thus, observations where the model does not produce precipitation or where only a trace is produced are excluded as are model grid points that produce precipitation for which there are no observations. Because of this, mPING observations cannot be used to assess general precipitation forecasts, either as probability of precipitation or the overall positional location accuracy of precipitation forecasts. No spatial characteristics can be verified because only the intersection between mPING reports and grid points with precipitation are used. Disjoint data, where the model produces precipitation but there are no mPING observations, or where there are mPING observations but no precipitation in the model, are not assessed.

3. Analysis and results

A total of 126 576 mPING observations are available to compare against the model forecasts. At most only

about one-third ($\sim 42\,200$) of these are expected to be used because only one out of every three is being verified, though the exact number depends upon how the observations are distributed across the hour intervals. The number of observations that go into each verified forecast hour varies somewhat by model and depends on the joint distribution between available observations, determined by population density, time of day, and precipitation distribution within the model forecast, and grid points where the model generates precipitation. For the RAP, the mean number of observations used in the verification statistics at any hour is 30 919, ranging from 33 206 at the 3 h to 29 203 at the 18-h lead times. For the NAM, the average number of observations for any hour is 21 490, ranging from 22 181 at 3 h and 20 937 at 24 h. Finally, the GFS enjoys an average of 35 914 observations for any hour, ranging from 37 170 at 3 h to 34 737 at 24 h. Anecdotally, we find that the NAM has the least area covered by precipitation while the GFS has the largest, explaining some of the differences in available verifying observations.

Here, forecast performance is discussed in terms of bias and skill scores. Bias (or bias ratio) is the number of forecasts for a specific type divided by the number of observations of that same type. Put differently, bias is the ratio of the ratio of the average forecast divided by the average observation (Wilks 2011). A bias of 1 means that the event is forecast with the same frequency it is observed. If the bias is above 1 (“high” or overforecast), the event is forecast more often than it is observed, and if the bias is less than 1 (“low” or underforecast) it is predicted less often than it is observed. The bias could be perfect but the forecast poor.

Skill scores address how well an event is forecast relative to the climatology contained within the sample data. The Peirce skill score [PSS; also known as the Hanssen–Kuipers skill score and the true skill statistic; Wilks (2011)] measures proportion correct referenced with the proportion correct that would be achieved by random forecasts that are statistically independent of the observations, such that the reference hit rate is random and unbiased. PSS is used here for dichotomous events (or categories). PSS is a ratio and ranges from -1 to 1 , where 1 is perfect, 0 is the sample climatology, and -1 is antiperfect. The PSS is an equitable score, which means that it does not reward either a “constant” or random (climatology) forecast or classification and thus does not discourage the forecasting or classification of rare events or types.

The Gerrity skill score (GSS; Gerrity 1992) is used to assess the skill of a model on all four precipitation types simultaneously (all 4×4 confusion matrices and the mPING data used in the analysis are provided in the supplemental material). The GSS requires a ranking or

ordering of the ptypes and is sensitive to the chosen ranking. The GSS rewards proper classifications of rare ptypes and penalizes improper classifications of common ptypes more than rare ptypes, gives credit for a forecast that is in a “close” class, is equitable (Gandin and Murphy 1992), and is a recommended scoring metric for such forecasts (Livezey 2004). NCEP uses a ranking (order from least to most severe) of $RA < PL < SN < FRZA$. Our ranking differs from that used by NCEP and from least to most severe is $RA < SN < PL < FZRA$; this ranking also orders the different ptypes from most common to least common. This order is chosen because it makes sense for impact severity to aviation operations and because it remains in the spirit of the Gerrity score, which is to reward rare classes that are correctly selected and penalize missing proper selection of common classes. Thus, RA within the ranking chosen here is “closer” to SN than it is to PL, which affects the Gerrity score computation in that an improper classification of RA as SN is penalized less/given more credit than is an improper classification of RA as PL. GSS ranges from 0 (sample climatology) to 1 (perfect forecast). Two additional ptype class orders for the GSS have been tested, the NCEP and a “physical” ordering ($RA < FZRA < PL < SN$), and all result in the same conclusions, though the GSS values differ. For any single ptype and thus a 2×2 confusion matrix, the Gerrity score becomes the PSS and, thus, is the skill metric used for evaluating any single ptype class.

Figure 2a shows model output for all the models at 2100 UTC 5 December 2013 during a significant winter storm that stretched from the Texas Panhandle up through the upper Ohio valley. Subjectively, both the NAM and GFS do relatively well at capturing the precipitation type. The RAP, however, completely misses all of the PL, calling for FZRA or RA instead.

Figure 2b shows that the RAP forecast ptype for the 3-h lead time tends to produce FZRA where PL is reported. By 15 h, the character of the ptype forecast changes to SN in roughly half of the area in which PL is observed. However, the RAP still does not produce PL.

Figures 3a–i show various characteristics of the different model ptype forecasts. Figure 3a assesses the overall skill of each model for all four ptypes combined, while Figs. 3b–i apply to the individual ptypes. Individual ptypes are assessed in a 2×2 confusion matrix that addresses “target ptype” versus “nontarget ptype” using bias and the Pierce skill score.

Figure 3a shows the GSS for all three models along with 95% confidence limits derived from bootstrap resampling. The difference between the NAM and GFS is not statistically significant throughout the 24-h period. The GSS for both models decreases slightly with

FIG. 2. (top) Model forecast ptype at 2100 UTC 5 Dec 2013 for (from left to right) RAP, NAM, and GFS. Forecasts precipitation type is shown by the filled color contours while mPING observations are shown by the small colored dots (the color legend is shown below). Images are for the 9-h forecast. (bottom) The 3- (left) and 9-h (right) ptype forecasts for the RAP only, both valid at 2100 UTC.

increasing lead time. While the GSS for the RAP starts out only slightly less than for the NAM and GFS, it drops quickly and significantly over the 18-h forecast period (the maximum lead time for the RAP), in stark contrast to the behavior depicted by the NAM and GFS.

Of the three models, the RAP tends to display the largest bias for any given ptype.

Compared to mPING observations, the RAP produces too much RA (Fig. 3b), too little SN (Fig. 3d), far too little PL (Fig. 3f), and, out to about a 12-h lead time, too much FZRA (Fig. 3h). Both the NAM and GFS biases are comparable and close to 1 for all but PL, when both are biased low. The most serious bias within the RAP is for PL. For all practical purposes, the scheme used in the version of the RAP available at the time of these observations simply fails to produce PL where precipitation is observed.

These biases negatively affect the RAP PSS for individual ptypes. Figures 3c, 3e, 3g, and 3i show that for the

RAP, in particular, PSS tends to decrease far more with increasing lead time than do the NAM and GFS. At 3-h lead time, the RAP FZRA bias (Fig. 3h) is high (~ 1.6) and the RAP PSS (Fig. 3i) is the best (0.43). As lead time is extended, in FZRA both the bias and PSS decrease and are no better than those for the NAM and GFS. Overall, the RAP struggles more with ptype than do the NAM or GFS.

No model does particularly well with either PL or FZRA. Is there pattern to these misses? For example, when PL or FZRA are observed, what does each model predict? Table 2 shows the “miss analysis” when FZRA is observed for each of the models. While each model behaves differently, the behavior is consistent across all forecast lead times. On average across all cases, when FZRA is observed and the RAP incorrectly diagnoses FZRA, the RAP misdiagnoses RA at a rate of 70% and SN at a rate of 30%. When the NAM incorrectly diagnoses FZRA, the NAM has misdiagnosed RA in 50% of the cases, SN in 27%, and PL in

FIG. 3. (a) GSS for the RAP, NAM, and GFS models. Forecast lead time is along the x axis and the GSS is along the y axis. The 95% confidence limits for the GSS based on bootstrap resampling are shown by the dashed lines. (b),(d),(f),(h) Biases and (c),(e),(g),(i) PSS values, respectively, for RA, SN, PL, and FZRA. Note that vertical scales for PSS are different for PL and FZRA.

23%. Finally, when the GFS incorrectly diagnoses FZRA, RA is misdiagnosed at a 44% rate, SN at a 34% rate, and PL at a 22% rate. Thus, all models misdiagnose RA when FZRA is observed, with the RAP most biased toward RA.

When PL is observed (Table 3), the behavior is again different between the models but consistent across all forecast lead times for any particular model. On average, the RAP tends to diagnose RA for 54% of the cases, SN

FIG. 3. (Continued)

for 26%, and FZRA for 20%. For the NAM when PL is observed, RA is misdiagnosed at a 44% rate, SN at a 40% rate, and FZRA is misdiagnosed at only a 16% rate. For PL forecasts from the GFS, RA is misdiagnosed at a 37% rate, SN at a 50% rate, and FZRA at a 13% rate. Thus, the GFS tends to diagnose SN when FZRA is observed while the NAM will misdiagnose the ptype as RA at a slightly higher rate than for SN. The RAP tends toward RA when PL is observed.

4. Discussion

The dearth of PL in the RAP may be partially explained by the scheme employed within the model. There is no published method for explicitly diagnosing ptype using the Thompson scheme save that in Table 1 of Ikeda et al. (2013), which describes a scheme developed by the Earth System Research Laboratory's Global Systems Division. In this method, a diagnosis of PL must meet three conditions, one of which requires

the graupel mixing ratio to exceed a threshold *and* to exceed the snowfall rate. The snowfall rate will typically exceed the graupel fall rate and may be the primary reason PL is diagnosed so rarely within the RAP (G. Thompson 2014, personal communication). The cause of the ptype diagnosis problems within the RAP resides within the postprocessing of the Thompson microphysics output, not in the microphysics scheme itself.

Based on the results for the RAP shown here, errors have been found in the ptype postprocessing in both the RAP and the High-Resolution Rapid Refresh (HRRR;

TABLE 2. Misdiagnosis analysis (%) when FZRA is observed; values are the average over all forecast lead times and represent the proportion of each class diagnosed by the model when FZRA is incorrectly diagnosed.

	RAP	NAM	GFS
RA	70		
SN	30	27	34
PL	0	23	22

TABLE 3. As in Table 1, but for when PL is observed.

	RAP	NAM	GFS
RA	54	44	37
SN	26	40	50
FZRA	20	16	13

S. Benjamin 2015, personal communication). As of 4 January 2015 these errors have been addressed and have been implemented in the HRRR at both ESRL and NCEP.

Overall, the model ptype classification schemes are weakest when considering the transitional precipitation types of PL and FZRA. Unfortunately, these are also the most troublesome when they occur. In addition, FZRA may be particularly problematic for any scheme in any model because it is the most transient ptype in both time and space. Because the output of each individual ptype algorithm is not available in the archived GRIB2 files used here, we have little insight into the nature of the misdiagnosed precipitation classes. It is likely, however, that because the two Baldwin schemes are highly correlated (Reeves et al. 2014) the diagnosed ptype is heavily influenced by these two in both the NAM and GFS. The original Baldwin scheme was designed to have a high probability of detection for both FZRA and PL. The Baldwin2 scheme is designed to reduce the resulting high bias for FZRA and PL of the original. Given that the two Baldwin schemes tend to behave similarly or “vote in a bloc,” for them to be “outvoted” requires at least two other schemes to generate not only the same ptype, but one that is also ranked with a higher impact than what the two Baldwin schemes generate.

Because there are few data available on how the different ptype schemes respond to model errors, trying to diagnose the nature of the model errors based on ptype diagnoses is not possible.

That said, while the GSS for all models degrades with increasing forecast lead time, the ptype schemes in the NAM and GFS seem to be relatively insensitive to how model errors evolve over a 24-h forecast period. The degradation of the RAP GSS is likely due to the decreasing performance in FZRA diagnosis with increasing lead time. Why FZRA suffers in particular is not clear.

5. Concluding thoughts

The mPING data are used to verify precipitation classification or ptype diagnoses for three operational models: the RAP, the NAM, and the GFS. The mPING data allow the verification of surface precipitation type with the detail and certainty exhibited here. Prior to the availability of these data, such a verification exercise was simply not possible. The ASOS cannot automatically report PL and

has known problems with FZRA and freezing drizzle. Human observers are not always available to augment ASOS observations and may not always be able to do so because of other workload requirements. The mPING data are shown to be self-consistent and to match characteristics documented in prior work, especially for FZRA, and are thus found to be suitable for this work.

Only forecasts valid at 3-h intervals out to 24 h (18 h for the RAP) are evaluated, and no distinction is made for the model initial times. The multiclass verification statistic is the GSS and the single-class verification statistic is the PSS. Diagnosed model precipitation classes display the weakest performance with the transitional precipitation types of FZRA and PL. While the NAM and GFS exhibit similar, though not identical, performance metrics, the RAP appears to have more serious problems. The RAP has no skill for PL forecasts because the explicit algorithm it uses almost never produces PL. The RAP shows better skill for FZRA for the first 6–9 h of lead time, but deteriorates in skill to that of the NAM and GFS by the 18-h lead time forecast.

The nature of the misdiagnosed classes is consistent within a given model, but different across models. Since the NAM and GFS use similar ptype diagnostic methods, differences in ptype misdiagnoses are likely due to the unique errors each model displays. The GSS deteriorates for all models with increasing lead time, but much less so for the NAM and GFS than for the RAP. This difference appears to be due mainly to the decreasing skill in the freezing rain diagnosis within the RAP.

Based on these statistics, even though some of the errors are certainly due to incorrect model forecasts of the temperature and moisture profiles, there is considerable room for improving surface ptype algorithms. Future work will also include comparisons of collocated mPING ptype reports and ASOS temperature reports to determine if there are significant problems in the wintertime boundary layer structure. Future work may also include assessment of hourly (instead of 3 hourly) ptype forecasts. Clearly, mPING data can play a major role in such improvements, from tuning existing algorithms to creating entirely new algorithms based on data-driven artificial intelligence (machine learning) techniques and work is already proceeding to that end. Even so, freezing rain may continue to pose particularly difficult challenges because of its high variability in space and time. Yet, the temporal and spatial density of the mPING data provide a means for regionally tuned ptype algorithms that adapt to the nature of model forecast errors.

Acknowledgments. Three anonymous reviewers provided excellent suggestions that improved the presentation and clarity of these results. This work was supported by the NEXRAD Product Improvement Program, by

NOAA/Office of Oceanic and Atmospheric Research. Funding was provided by NOAA/ Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of NOAA, the U.S. DOC, or the University of Oklahoma.

REFERENCES

- 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.
- 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](https://doi.org/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. *Proc. 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., 2003a: Characteristics of ice storms in the United States. *J. Appl. Meteor.*, **42**, 630–639, doi:[10.1175/1520-0450\(2003\)042<0630:COISIT>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<0630:COISIT>2.0.CO;2).
- , 2003b: Urban modification of freezing rain events. *J. Appl. Meteor.*, **42**, 863–870, doi:[10.1175/1520-0450\(2003\)042<0863:UMOF>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<0863:UMOF>2.0.CO;2).
- , and T. G. Creech, 2003: Sources of data on freezing rain and resulting damages. *J. Appl. Meteor.*, **42**, 1514–1518, doi:[10.1175/1520-0450\(2003\)042<1514:SODOFR>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<1514:SODOFR>2.0.CO;2).
- , and T. R. Karl, 2003: Temporal and spatial variation in freezing rain in the contiguous United States: 1948–2000. *J. Appl. Meteor.*, **42**, 1302–1315, doi:[10.1175/1520-0450\(2003\)042<1302:TASVOF>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<1302:TASVOF>2.0.CO;2).
- Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, and L. S. Rothfusz, 2014: mPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342, doi:[10.1175/BAMS-D-13-00014.1](https://doi.org/10.1175/BAMS-D-13-00014.1).
- Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370, doi:[10.1175/1520-0493\(1992\)120<0361:ESSFCF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2).
- Gerrity, J. P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709–2712, doi:[10.1175/1520-0493\(1992\)120<2709:ANOGAM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<2709:ANOGAM>2.0.CO;2).
- Goodwin, L. C., 2003: Weather-related crashes on U.S. highways in 2001. Mitretek Systems, Inc., Rep. prepared for U.S. Dept. of Transportation, 8 pp.
- 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](https://doi.org/10.1175/WAF-D-12-00085.1).
- Janjic, Z. I., T. L. Black, M. E. Pyle, H.-Y. Chuang, E. Rogers, and G. I. DeMego, 2005: The NCEP WRF-NMM core. Preprints, *Joint WRF/MM5 User's Workshop*, Boulder, CO, NCAR, 2.9.
- Kain, J. S., S. M. Goss, and M. E. Baldwin, 2000: The melting effect as a factor in precipitation-type forecasting. *Wea. Forecasting*, **15**, 700–714, doi:[10.1175/1520-0434\(2000\)015<0700:TMEAFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0700:TMEAFA>2.0.CO;2).
- Lackmann, G. M., K. Keeter, L. G. Lee, and M. B. Ek, 2002: Model representation of freezing and melting precipitation: Implications for winter weather forecasting. *Wea. Forecasting*, **17**, 1016–1033, doi:[10.1175/1520-0434\(2003\)017<1016:MROFAM>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)017<1016:MROFAM>2.0.CO;2).
- Livezey, R. F., 2004: The Gerrity equitable skill score as a near optimal alternative to the modified Heidke score for verification of categorical seasonal forecasts. Preprints, *Symp. on Forecasting the Weather and Climate of the Atmosphere and Ocean/15th Symp. on Global Change and Climate Variations*, Seattle, WA, Amer. Meteor. Soc., J13.9.
- 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/pdfpapers/94838.pdf>.]
- , 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/pdfpapers/73517.pdf>.]
- Moorthi, S., H.-L. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. NWS Tech. Procedures Bull. 484, 14 pp. [Available online at <http://www.nws.noaa.gov/om/tpb/484.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](https://doi.org/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.
- Reeves, H. D., K. L. Elmore, A. Ryzhkov, T. Schuur, and J. Krause, 2014: Sources of uncertainty in precipitation type forecasting. *Wea. Forecasting*, **29**, 936–953, doi:[10.1175/WAF-D-14-00007.1](https://doi.org/10.1175/WAF-D-14-00007.1).
- Robbins, C. C., and J. V. Cortinas, 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](https://doi.org/10.1175/1520-0434(2002)017<0047:LASEAW>2.0.CO;2).
- Stewart, R. E., and P. King, 1987: Freezing precipitation in winter storms. *Mon. Wea. Rev.*, **115**, 1270–1279, doi:[10.1175/1520-0493\(1987\)115<1270:FPIWS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1270:FPIWS>2.0.CO;2).
- Theriault, 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](https://doi.org/10.1175/2010JAMC2321.1).
- Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, doi:[10.1175/1520-0493\(2004\)132<0519:EFOWPU>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2).
- , P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:[10.1175/2008MWR2387.1](https://doi.org/10.1175/2008MWR2387.1).
- 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](https://doi.org/10.1175/WAF871.1).
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press, 627 pp.