ASSESSING THE SKILL OF UPDATED PRECIPITATION TYPE DIAGNOSTICS FOR RAPID REFRESH WITH mPING

Tomer Burg[1,2], Kimberly L. Elmore[3,4], and Heather M. Grams[3,4]

[1]National Weather Center Research Experiences for Undergraduates Program
Norman, Oklahoma

[2]State University of New York at Albany
Albany, New York

[3]University of Oklahoma Cooperative Institute for Mesoscale Meteorological Studies
Norman, Oklahoma

[4]National Oceanic and Atmospheric Administration National Severe Storms Laboratory
Norman, Oklahoma

ABSTRACT

Previous work shows that the Rapid Refresh (RAP) model severely under-represents ice pellets in its grid, with a skill near zero and a very low bias. An ice pellet diagnostic upgrade was devised at the Earth System Research Laboratory (ESRL) to resolve this issue. Parallel runs of the experimental ESRL-RAP with the fix and the operational NCEP-RAP without the fix provide an opportunity to assess whether this upgrade has improved the performance of the ESRL-RAP, both for the models overall and for individual precipitation types, using the meteorological Phenomena Identification Near the Ground (mPING) project as verification. The overall Gerrity Skill Score (GSS) for the ESRL-RAP is improved relative to the NCEP-RAP at 3 hour lead time but degrades with increasing lead time, a difference which is statistically significant but may not have much practical significance. Some improvement was found in the bias and skill scores of ice pellets and snow in the ESRL-RAP, although the model continues to under-represent ice pellets, while rain and freezing rain were generally the same or slightly worse with the fix. The ESRL-RAP was also found to depict a more realistic spatial distribution of precipitation types in transition zones involving ice pellets and freezing rain.

1. INTRODUCTION

Forecasting precipitation types is often a significant challenge for forecasters. Considering the significant societal impacts that winter precipitation can inflict, correctly forecasting these precipitation types is crucial. As an aid, forecasters often refer to numerical weather prediction models for an approximation of the spatial distribution of potential precipitation types. Unfortunately, modeled precipitation types remain imperfect.

Previous studies examine modeled precipitation type verification against surface observations. For example, Ikeda et al. (2013) verifies the skill of the High-Resolution Rapid Refresh model (HRRR) for predicting precipitation types, using the Automated Surface Observing System (ASOS) network as surface observations, and concludes that the mixed precipitation transition zones has lower performance scores than rain or snow. This study is limited by its usage of the ASOS network which can be spatially sparse outside of metropolitan areas; additionally, as out of the 852 ASOS stations throughout the US, only 15% are able to report ice pellets, further limiting the size of the observation dataset (Elmore et al. 2015).

More recent studies on precipitation type verification have incorporated observations from the meteorological Phenomena Identification Near the Ground project (mPING, Elmore et al. 2014). mPING is a mobile application used to crowdsource precipitation type observations from the public, where users can select various precipitation types including the following: none,

[1] Tomer Burg, Center for Analysis and Prediction of Storms, The University of Oklahoma, National Weather Center, 1200 David L. Boren Blvd., Norman, OK 73072
tomerburg@yahoo.com

hail, drizzle, freezing drizzle, rain, freezing rain, ice pellets, snow, rain and snow, rain and ice pellets, and ice pellets and snow. Since its launch on 19 December 2012, over 896,000 individual reports have been submitted to mPING as of July 2015. As observations can be submitted from any location at any time, mPING provides a network of spatially and temporally dense precipitation type observations, which is especially useful during cases of highly localized variability in precipitation types.

Elmore et al. (2015) utilizes mPING observations in 2013 to verify forecast precipitation types for the North American Mesoscale Forecast System (NAM), the Global Forecast System (GFS), and the Rapid Refresh (RAP) models. The study similarly shows that forecast skill for ice pellets and freezing rain is substantially lower than for rain and snow. In particular, the RAP model severely under-represents ice pellets and accordingly has almost zero skill and a very low bias. Thus, the RAP performs poorly for ice pellets compared to the other models analyzed in the study.

In response, the Earth System Research Laboratory (ESRL) implemented an ice pellet diagnostic change that changed the integrated rain water requirement from 0.05 g/kg to 0.005 g/kg (NOAA 2015). There are two versions of the RAP currently active: the experimental RAP, or the ESRL-RAP, running through the Earth System Research Laboratory, and the operational RAP, or the NCEP-RAP, running through the National Center for Environment Predictions. The ice pellet diagnostic change was implemented in the ESRL-RAP on 12 March 2014, but as of June 2015 had yet to be incorporated into the NCEP-RAP (E. James, personal communication). This lag between the implementation of the PL-diagnostic change between the ESRL and NCEP versions of the RAP extends for over a year, encompassing the entire cold season of 2014-2015. This provides a unique opportunity to assess parallel versions of the RAP with and without the fix. This project seeks to determine whether the PL-diagnostic change has improved the precipitation type forecast skill of the ESRL-RAP over the NCEP-RAP by verifying both models against mPING observations.

2. METHODS

a. Data

The NCEP-RAP model output was obtained from the NOAA National Operational Model Archive and Distribution System (NOMADS), while the ESRL-RAP model output was obtained from the Earth Science Research Laboratory (ESRL). For the observations used as ground truth, mPING archives for the cold season of 2014-2015 were retrieved, consisting of the observation ID, time and location of observation, and the precipitation type. These observations are compared against the precipitation type output generated from the RAP model at the nearest grid point.

Precipitation type diagnosis in numerical weather prediction models occurs during the post-processing stage, utilizing raw model output to assign precipitation types. In the case of the RAP model, a microphysics parametrization scheme based on Thompson (2008) is applied to the model. In the post-processing stage, hydrometeor mixing ratios and fall rates for each precipitation type are used, along with surface temperatures, to generate a categorical yes or no value for each of the four primary precipitation types: rain, snow, ice pellets, and freezing rain. This procedure is discussed in further detail in Ikeda et al. (2013) and NOAA (2015).

As precipitation types are diagnosed independently, it is possible for the RAP to assign multiple precipitation types at the same location. Some of these overlaps do not match with mPING's categories, such as a mix of freezing rain and ice pellets which is possible under the RAP's algorithm but is not an option provided in mPING. In order to maintain consistency between the two sources, all instances of multiple precipitation types are collapsed into the four primary types, following the approach used in Elmore et al. (2015), and a ranking is assigned in order from highest to lowest impact: freezing rain, ice pellets, snow and rain.

Six cases are selected from the 2014-2015 cold season; these six cases are listed in Table 1, which contains the start and end hours of each event, as well as how many mPING observations were used from each case at 3-hour lead time, which is typical for all lead times.

Case Start	Case End	3hr Lead Time mPING Reports	Percentage of Total Cases	Primary Regions of Impact
09 UTC 11/26/2014	03 UTC 11/27/2014	2976	12.6%	Northeast US
23 UTC 2/1/2015	22 UTC 2/2/2015	2331	9.9%	Northern US
21 UTC 2/15/2015	06 UTC 2/17/2015	4637	19.6%	Southern US
12 UTC 2/20/2015	00 UTC 2/22/2015	3015	12.8%	Central, Eastern US
06 UTC 2/23/2015	00 UTC 2/24/2015	1347	5.7%	Southern US
12 UTC 3/3/2015	21 UTC 3/5/2015	9335	39.5%	Central, Eastern US

Table 1. Beginning and end times of each case, total number of locations where both versions of the RAP depict precipitation and an mPING observation exists for each case at 3-hour lead time, percentage of each case within the total composite, and the primary region(s) of impact of each event.

b. Methods

In order to verify the RAP's precipitation type forecasts, both the ESRL-RAP and NCEP-RAP are individually compared against mPING observations. All observations 30 minutes prior to or following the nearest forecast hour are centered to that hour, and each mPING observation for that centered hour is compared against the precipitation type assigned to the nearest RAP gridpoint valid at the same hour. This procedure is completed separately for the NCEP-RAP and ESRL-RAP at 3, 6, 9, 12, 15 and 18 hour forecast lead times, for each case as well as a composite of all cases. Only locations that had precipitation forecast by both versions of the RAP and observed through mPING are considered for this study.

The resulting comparisons are analyzed using three different statistics: the Gerrity Skill Score (GSS), the Peirce Skill Score (PSS), and bias. The GSS determines the skill for all four ordered precipitation types simultaneously. The GSS is an equitable score, meaning that among other factors, constant and random forecasts yield a score of zero (Gandin and Murphy 1992). Additionally, the GSS penalizes misdiagnosis of common precipitation types, such as rain, more so than misdiagnosis of rare precipitation types, such as freezing rain. The GSS ranges from -1 to 1, where -1 is an anti-perfect forecast, 0 is the sample climatology or constant forecast (e.g. no skill), and 1 is a perfect forecast (Elmore et al. 2015).

In order to analyze precipitation types individually, the PSS and bias are applied. The PSS is also an equitable score and ranges from -1 to 1, and is used to assess the skill of each individual precipitation type relative to sample climatology. The bias is the ratio of the number of forecasts of a precipitation type divided by the number of observations of the same precipitation type; a bias of 1 is an unbiased forecast, while a bias less than 1 is an underforecast and a bias above 1 is an overforecast of the precipitation type. A bias of 1, however, does not necessarily imply that the forecast is correct, as it does not account for location.

For each statistic, a 95% confidence interval is computed based on bootstrap resampling. Permutation tests are used to determine whether the difference between the means of each statistic for the ESRL-RAP and the NCEP-RAP is statistically significant. For example, a p-value less than 0.05 indicates statistical significance at the 95% confidence level, while a p-value less than 0.01 corresponds with 99% confidence level.

3. RESULTS

a. Composite of All Cases

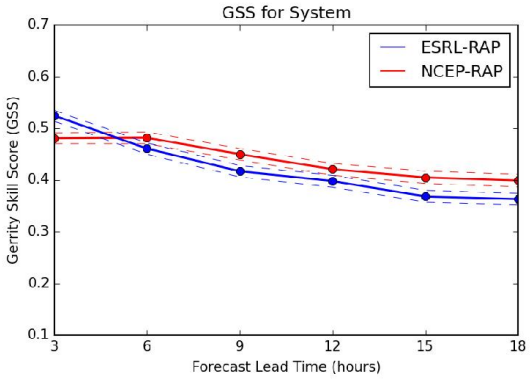

Figure 1. Gerrity Skill Score (GSS) for the full case composite analyzed at 3-hour forecast lead time intervals. Thick lines represent the mean GSS, while the encompassing dashed lines represent the 95% confidence interval based on bootstrap resampling.

For the composite of all cases, the aforementioned statistical skill scores are computed in 3-hour intervals from 3 to 18 hour forecast lead time. Figure 1 depicts the mean GSS and 95% confidence interval for both versions of the RAP. At 3 hours, the ESRL-RAP performs relatively well with a mean GSS of 0.525, whereas the NCEP-RAP has a mean GSS of 0.481. Permutation tests yield a p-value much less than 0.01, indicating that the improved skill of the ESRL-RAP is statistically significant at the 99% confidence level. The skill of the ESRL-RAP degrades with increasing lead time, however, and by 6 hour lead time the mean GSS of the ESRL-RAP is 0.461, compared against the mean GSS of the NCEP-RAP at 0.482. Permutation tests indicate that this difference is statistically significant. The same trend continues through the remainder of the 18-hour forecast range, with the mean GSS gradually decreasing with increasing lead time. Even though these differences are statistically significant, they may have little practical significance.

The composite sample is broken down into the four primary precipitation types to help determine whether any particular type is affecting the performance of the RAP. Figures 2a-2h depict the bias and PSS for each precipitation type. For rain, the ESRL-RAP has a persistently higher mean bias than the NCEP-RAP, both of which are above 1. Thus, the ESRL-RAP overforecasts rain. For snow, the ESRL-RAP shows a substantial improvement with an almost perfectly unbiased output. For freezing rain, the ESRL-RAP shows a significantly lower bias than the NCEP-RAP, but with a continued tendency to overforecast freezing rain. For ice pellets, the NCEP-RAP has a very low bias, alternating between 0.12 and 0.15, continuing to underforecast ice pellets in the grid. The ESRL-RAP does better, with the bias alternating between 0.28 and 0.38, although this is still far removed from an unbiased score. All bias differences between the ESRL-RAP and NCEP-RAP are significant at the 99% confidence level except for freezing rain at 3-hour lead time, which is significant at the 95% confidence level.

An analysis of the PSS for the individual precipitation types shows minor changes for rain, with a higher PSS for the ESRL-RAP at 3 hour lead time and lower PSS at 18 hour lead time, while the PSS for freezing rain is typically worse for the ESRL-RAP than the NCEP-RAP. The ESRL-RAP has a statistically significant higher PSS than the NCEP-RAP for snow, as well as ice pellets.

The composite of all cases reflects an improvement in ice pellet diagnosis, although typical variability in the evolution of winter storms on a daily basis result in different outcomes for each case. To further highlight the extent of these day-to-day variabilities, two cases are analyzed in more detail in the next two sections.

b. 26-27 November 2014

The 26-27 November 2014 case was dominated by a coastal low pressure system along the East Coast which produced an early season snowstorm across the Mid Atlantic into New England regions. This case produces the lowest skill scores out of any case analyzed in this study, and the NCEP-RAP depicts very little mixed precipitation, despite mPING observations suggesting otherwise; for the 3-hour lead time, there are 448 ice pellet and 33 freezing rain observations, while the NCEP-RAP has only 1 ice pellet and 2 freezing rain forecasts.

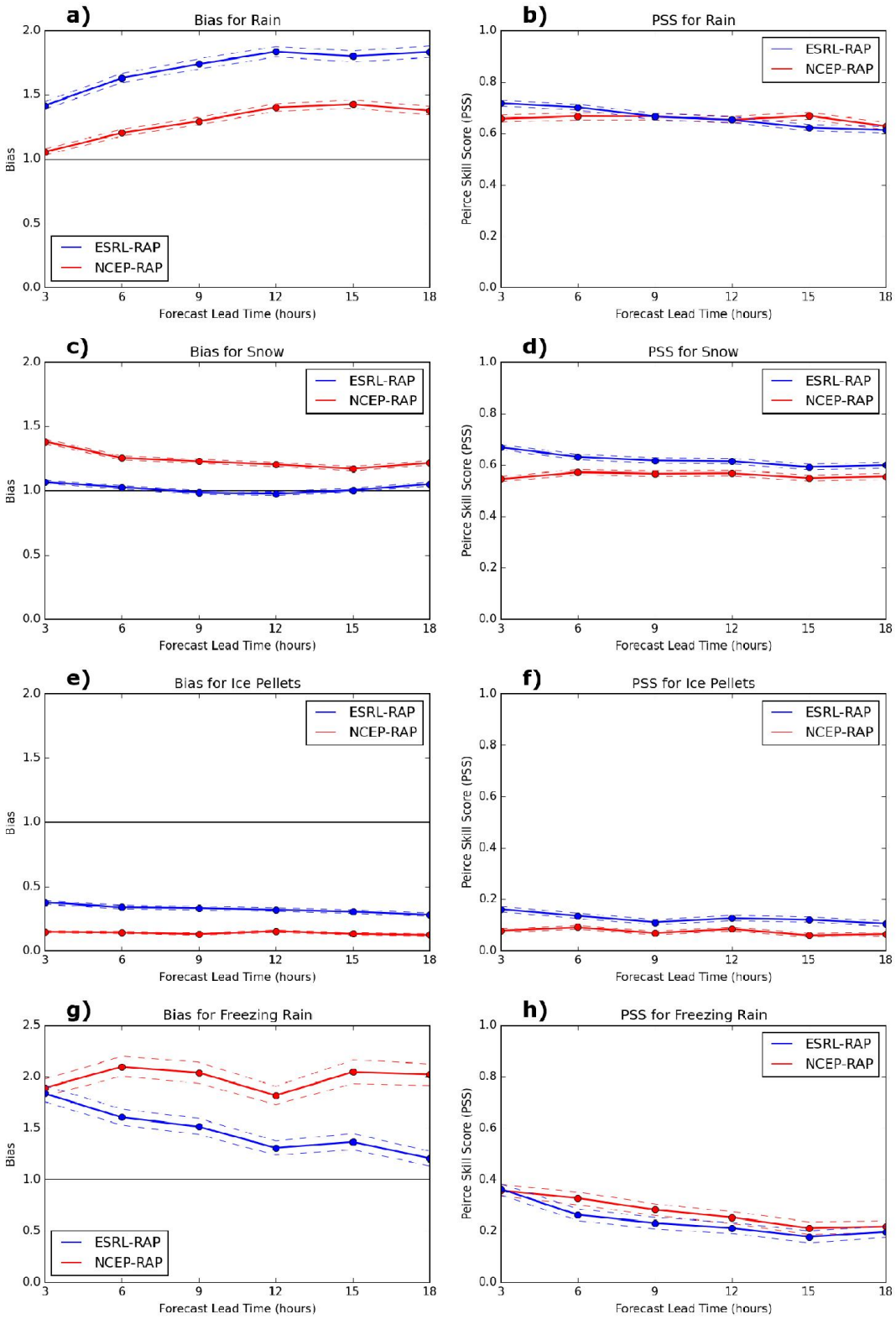

Figure 2. For the composite of all cases, a) bias and b) Peirce Skill Score (PSS) for rain, c) bias and d) PSS for snow, e) bias and f) PSS for ice pellets, and g) bias and h) PSS for freezing rain. Note the different y-axis on the freezing rain bias time series.

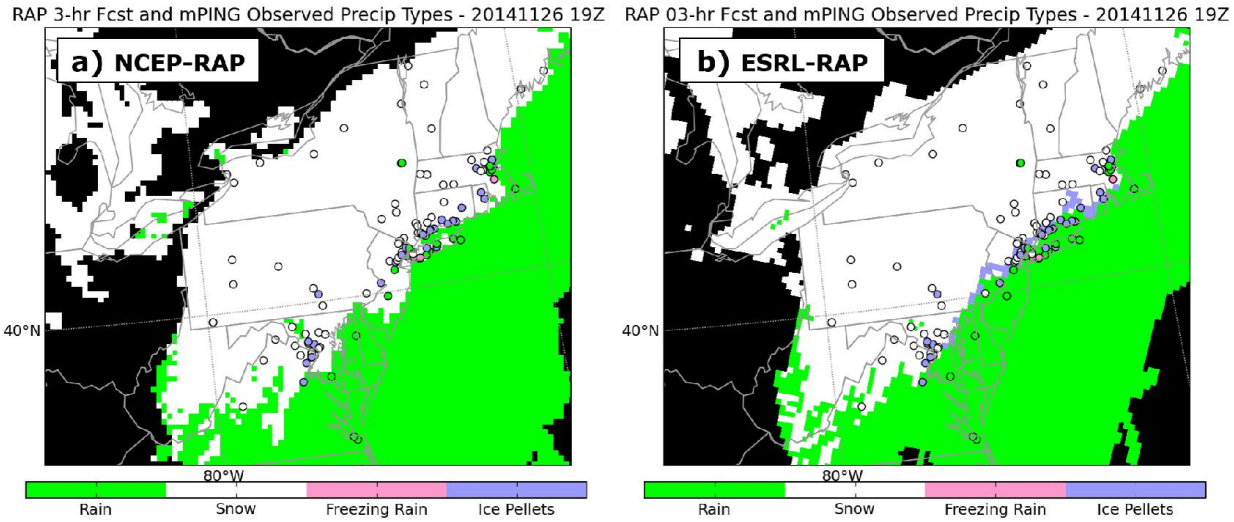

Figure 3. a) NCEP-RAP and b) ESRL-RAP forecast precipitation type with a 3-hour lead time, valid at 1900 UTC 26 November 2014. Modeled p-types are in filled contours, and mPING observations are represented in circles.

Using Figures 3a and 3b for a visual analysis, the NCEP- RAP does not depict any precipitation type other than rain or snow, despite the presence of numerous ice pellet and freezing rain mPING reports from Washington D.C. to Boston, which explains its low scores relative to the rest of the cases. The ESRL-RAP is improved with its depiction of a narrow axis of ice pellets from Maryland into coastal New England, and also outputs rain over southern New Jersey and Long Island, indicating a closer match to mPING observations than the NCEP-RAP which depicts only snow for these locations.

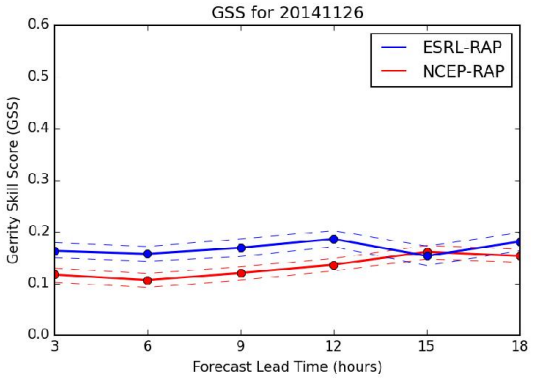

Figure 4. GSS for 26-27 November 2014 analyzed at 3-hour forecast lead time intervals. Thick lines represent the mean GSS, while the encompassing dashed lines represent the 95% confidence interval.

The mean GSS for the ESRL-RAP (Fig. 4) generally alternates between 0.15 and 0.18, which is a small but statistically significant improvement over the NCEP-RAP which alternates between 0.10 and 0.16.

c. 3-5 March 2015

The 3-5 March 2015 case synoptically consisted of two rounds: a widespread snow and ice pellet event in the Northeast US on 3 March 2015, followed by a slow southward progression of a strong baroclinic zone extending from the southern Plains into the Mid Atlantic regions consisting of a well-defined transition zone between rain, freezing rain, ice pellets and snow. This case is unusual as the GSS for the ESRL-RAP was typically lower than that of the NCEP-RAP at the 95% confidence level.

Another key difference between the two versions of the RAP becomes apparent (Fig. 5): when the model output is collapsed to the four primary precipitation types, the NCEP-RAP depicts an unrealistic transition zone from rain to snow at 3 hour lead time, particularly over Tennessee where precipitation type from south to north changes from rain to snow, then to ice pellets, then to freezing rain, then back to ice pellets and snow. One of the most noticeable changes with the ESRL-RAP is a much more realistic spatial distribution of precipitation types in the transition zone, with a south-north transition

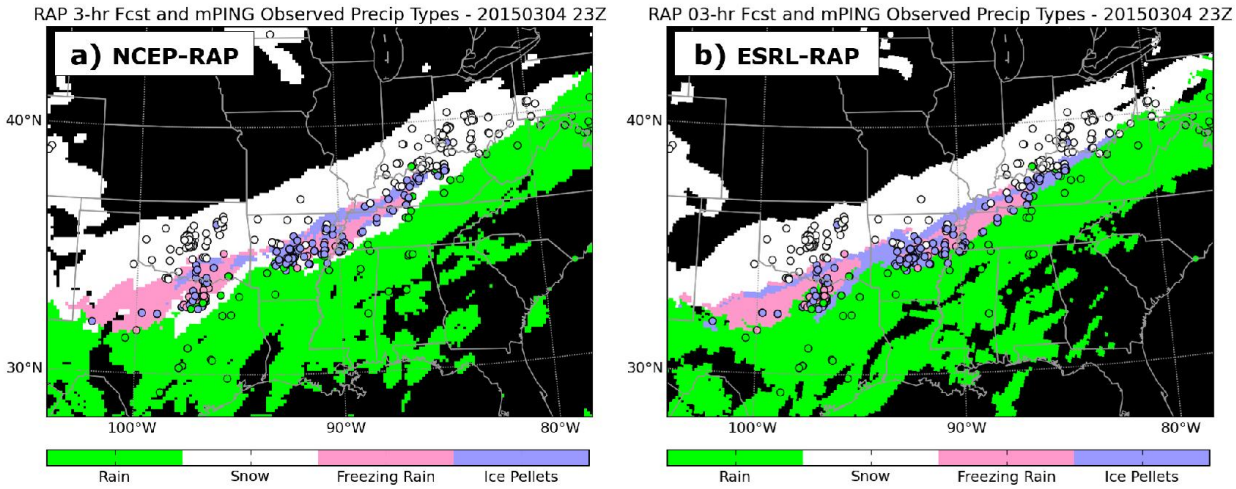

Figure 5. a) NCEP-RAP and b) ESRL-RAP forecast precipitation type with a 3-hour lead time, valid at 2300 UTC 4 March 2015. Modeled p-types are in filled contours, and mPING observations are represented in circles.

from rain to freezing rain to ice pellets to snow. Other cases analyzed in this study show the same trend. More ice pellets are being depicted within the ESRL-RAP than in the NCEP-RAP, although a bias to overforecast rain is apparent in the ESRL-RAP, particularly over western Tennessee and Arkansas where mPING observations show sleet while the model depicts rain.

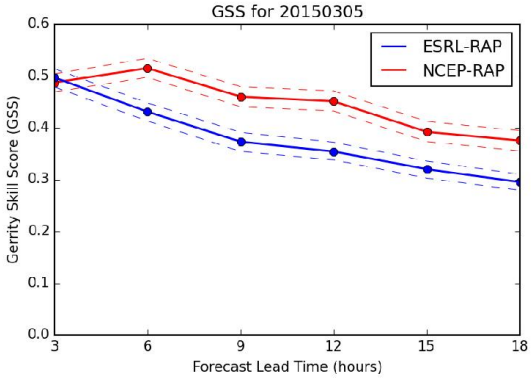

Figure 6. GSS for 3-5 March 2015 analyzed at 3-hour forecast lead time intervals. Thick lines represent the mean GSS, while the encompassing dashed lines represent the 95% confidence interval.

The mean GSS for the ESRL-RAP (Fig. 6) degrades from 0.497 at 3 hours to 0.295 by 18 hours. For the NCEP-RAP, the mean GSS peaks at 0.515 at 6 hours before gradually decreasing to 0.375 by 18 hours, indicating a persistent signal that the NCEP-RAP performs better than the ESRL-RAP. The difference between the two models is statistically significant at the 99% confidence level for all forecast lead times except 3 hours.

4. DISCUSSION

Clearly, the ESRL-RAP shows an incremental improvement in forecasting ice pellets and snow. Freezing rain and rain show either no improvement or slightly degraded skill within the ESRL-RAP. There is still case-to-case variability, although the individual cases analyzed all show improvement for ice pellets.

There may be other differences between the two models at play, because in figures 3 and 4, there are subtle differences in the spatial extent of precipitation between both versions of the RAP. Currently, the source(s) of these differences are not known, although they indicate that other changes exist between both versions of the RAP, which prevents isolating the ice pellet diagnostic from any other change. The small sample size of 6 cases is used for this analysis may not capture a complete picture of the day-to-day variability typical of precipitation events. Finally, while some of the differences in skill scores and bias between both versions of the RAP are statistically significant, they may not have much practical significance, especially if when differences are very small, as is the case with the PSS for freezing rain.

5. CONCLUSION

Parallel runs of the ESRL-RAP, which incorporates an ice pellet diagnosis upgrade and the NCEP-RAP, which does not, are verified with mPING observations to assess whether the upgrade has improved the performance of the RAP model precipitation type diagnosis. GSS values are computed for the composite and for individual cases, with PSS and bias for each individual precipitation type. Bootstrap 95% confidence intervals are computed and statistical significance of any differences are derived from permutation tests. Results suggest that the ESRL-RAP enjoys incremental and statistically significant improvement in the skill and bias for ice pellets and snow, but either no change or decreased performance for freezing rain and rain. Even with the improvements in the ice pellet diagnosis, the same bias of under-forecasting ice pellets continues but to a lesser extent. A visual analysis of the ESRL-RAP also reveals a more reasonable spatial distribution of precipitation types in transition zones involving sleet and freezing rain. Various issues, such as the small sample size and inherent day-to-day variability, preclude higher confidence on the exact nature of improvement in the ESRL-RAP.

6. ACKNOWLEDGMENTS

Appreciation is extended to Daphne LaDue for the National Weather Center Research Experience for Undergraduates, and to Eric James for providing the ESRL-RAP model data used in this study.

This work was prepared by the authors with funding provided by National Science Foundation Grant No. AGS-1062932, and 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 author(s) and do not necessarily reflect the views of the National Science Foundation, NOAA, or the U.S. Department of Commerce.

7. REFERENCES

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Framer, H. D. Reeves, and L. P. Rothfusz, 2014: mPING: Crowd-Sourcing Weather Reports for Research. *Bull. Amer. Meteor. Soc.,* doi:10.1175/BAMS-D-13-00014.1

Elmore, K. L., H. M. Grams, D. Apps, and H. D. Reeves, 2015: Verifying Forecast Precipitation Type with mPING. *Wea. Forecasting*, **30**, 656-667, doi:10.1175/WAF-D-14-00068.1, in press.

Gandin, L. S. and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370.

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

NOAA, 2015: RAP and HRRR Variables. Accessed 29 July 2015. [Available online at http://ruc.noaa.gov/rr/RAP_var_diagnosis.html.]

Thompson, G., 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.