Assessing the Skill of Updated Precipitation-Type Diagnostics for the Rapid Refresh with mPING✎

TOMER BURG[a]

National Weather Center Research Experiences for Undergraduates Program, Norman, Oklahoma, and University at Albany, State University of New York, Albany, New York

KIMBERLY L. ELMORE AND HEATHER M. GRAMS

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and National Oceanic and Atmospheric Administration/National Severe Storms Laboratory, Norman, Oklahoma

ABSTRACT

Previous work has shown that the Rapid Refresh (RAP) model severely underrepresents 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 overall performance and the performance of the individual precipitation types of the ESRL-RAP. Verification was conducted using the mobile Phenomena Identification Near the Ground (mPING) project. The overall Gerrity skill score (GSS) for the ESRL-RAP was improved relative to the NCEP-RAP at a 3-h lead time but degraded with increasing lead time; the difference is significant at $p < 0.05$. Whether this difference is practically significant for users is unknown. Some improvement was found in the bias and skill scores of ice pellets and snow in the ESRL-RAP, although the model continues to underrepresent 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 winter precipitation type (ptype) is often a significant challenge for forecasters. Considering the significant societal impacts that winter precipitation can inflict, correctly forecasting the type is crucial. As an aid, forecasters often refer to numerical weather prediction (NWP) models for guidance about the spatial distribution of different ptypes. Unfortunately, modeled precipitation types remain imperfect.

Previous studies examine modeled precipitation-type verification against surface observations. For example, Ikeda et al. (2013) verify the skill of the High Resolution Rapid Refresh model (HRRR) for predicting ptype using the Automated Surface Observing System (ASOS) network as surface observations, concluding that the forecast mixed-ptype transition zone has poorer performance scores than the rain and snow ptypes. Ikeda et al.'s (2013) results are inherently limited by usage of the ASOS network, which is spatially sparse. Additionally, of the 852 ASOS stations throughout the United States, only 15% are able to report ice pellets (PLs) only when attended, further limiting the extent of the observation dataset (Elmore et al. 2015).

More recent studies on ptype verification have incorporated observations from the mobile Phenomena Identification Near the Ground project (mPING; Elmore et al. 2014). The mPING mobile application is used to crowdsource ptype observations from the public, where users can select among various ptypes including

---

the following: none, 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 1 150 000 individual reports have been submitted to mPING as of October 2016. As observations can be submitted from any location at any time, mPING provides a network of spatially and temporally dense ptype observations. This is particularly useful in cases of highly localized variability in ptype.

Elmore et al. (2015) utilized mPING observations in 2013 to verify forecast precipitation types for the North American Mesoscale Forecast System (NAM; Janjić et al. 2005), the Global Forecast System (GFS; Moorthi et al. 2001), and the Rapid Refresh (RAP; Brown et al. 2011) models. The study similarly shows that the forecast skill for PL and freezing rain (FZRA) is substantially lower than for rain (RA) and snow (SN). In particular, the RAP model severely underrepresents PL. As a result, at that time the RAP had almost zero skill and a bias of nearly zero; thus, the RAP performed poorly for ice pellets compared to the other models analyzed in Elmore et al. (2015).

In response, the Earth System Research Laboratory (ESRL) implemented a PL diagnostic change that shifted the integrated rainwater requirement from 0.05 to $0.005 \, \text{g kg}^{-1}$ (Benjamin et al. 2016; NOAA/ESRL 2015). At the time of this writing, two versions of the updated RAP are currently active: the experimental RAP, or the ESRL-RAP, running at ESRL, and the operational RAP, or the NCEP-RAP, running at the National Centers for Environmental Prediction (NCEP).

Following is information on the changes that affected the ESRL and NCEP versions of RAP (C. Alexander 2016, personal communication):

1) NCEP RAPv2 on 25 February 2014
   IP diagnosis—if the graupel fall rate at the surface is at least $1.0 \times 10^{-6} \, \text{mm s}^{-1}$, the surface temperature is $<0°C$, the maximum rain mixing ratio in the column is $>0.05 \, \text{g kg}^{-1}$, and the graupel fall rate at the surface is greater than that for snow, then ice pellets are diagnosed. If in addition, the fall rate for graupel is greater than that for rain, ice pellets *only* are diagnosed, not freezing rain, not rain, and not snow. This diagnostic resulted in too little IP being diagnosed much of the time.
2) ESRL RAPv3 on 12 March 2014
   IP diagnosis—if the graupel fall rate at the surface is at least $1.0 \times 10^{-6} \, \text{mm s}^{-1}$, the surface temperature is $<0°C$, the maximum rain mixing ratio in the column is $>0.005 \, \text{g kg}^{-1}$ (modified 12 March 2014), and the graupel fall rate at the surface is greater than that for snow, then ice pellets are diagnosed. If, in addition,

the fall rate for graupel is greater than that for rain, ice pellets *only* are diagnosed, not freezing rain, not rain, and not snow. This change resulted in much more IP being diagnosed.
3) ESRL RAPv3 on 1 January 2015
   Switch to WRF version 3.6 with Thompson microphysics for aerosols.
4) ESRL RAPv3 during August 2015
   FZRA diagnosis—removed check on $T_{\max}$ in the column being below freezing and began using 2-m temperature instead. RA diagnosis—SN switched to RA when 2-m $T > 276.15 \, \text{K}$.
5) NCEP RAPv3 on 23 August 2016 included all ESRL RAPv3 changes

This lag between the implementation of the PL-diagnostic changes between the ESRL and NCEP versions of the RAP extends for over a year, encompassing the entire cold season of 2014/15 and so provides a unique opportunity to assess parallel versions of the RAP with and without the enhanced fix. This work seeks to determine whether the initial PL-diagnostic enhancement has improved the precipitation-type forecast skill of the NCEP-RAP since the initial work by Elmore et al. (2015) and any differences in the further refined ESRL-RAP over the NCEP-RAP by verifying both models against mPING observations.

2. Analysis

a. Data

The NCEP-RAP model output was obtained from the National Oceanic and Atmospheric Administration (NOAA) National Operational Model Archive and Distribution System (NOMADS; Rutledge et al. 2006), while the ESRL-RAP model output was obtained from the ESRL archive. The mPING observations used as ground truth come from cold season cases during 2014–2015. These observations are compared to the ptype generated from the RAP model at the grid point nearest the mPING observation.

Precipitation-type diagnosis in NWP models occurs during the postprocessing stage, utilizing raw model fields to assign ptype classes. Both versions of the RAP use a microphysics parameterization scheme based on Thompson et al. (2008). During the postprocessing stage, hydrometeor mixing ratios and fall rates for each precipitation type are assessed, along with surface temperatures, to generate a categorical yes or no value for each of the four primary or canonical precipitation types: rain, snow, ice pellets, and freezing rain. This

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

Case start time	Case end time	No. of 3-h lead-time mPING reports	Total cases (%)	Primary region(s) of impact
0900 UTC 26 Nov 2014	0300 UTC 27 Nov 2014	2976	12.6	Northeast United States
2300 UTC 1 Feb 2015	2200 UTC 2 Feb 2015	2331	9.9	Northern United States
2100 UTC 15 Feb 2015	0600 UTC 17 Feb 2015	4637	19.6	Southern United States
1200 UTC 20 Feb 2015	0000 UTC 22 Feb 2015	3015	12.8	Central, eastern United States
0600 UTC 23 Feb 2015	0000 UTC 24 Feb 2015	1347	5.7	Southern United States
1200 UTC 3 Mar 2015	2100 UTC 5 Mar 2015	9335	39.5	Central, eastern United States

procedure is discussed in further detail in Ikeda et al. (2013) and NOAA/ESRL (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 the mPING categories, such as mixed freezing rain and ice pellets; these choices are possible under the RAP's algorithm but are not options provided in mPING. To maintain consistency between the two sources, all instances of multiple precipitation types are collapsed into the four canonical types, following the approach used in Elmore et al. (2015), and a ranking is assigned in order from highest to lowest impact: FZRA, PL, SN, and RA.

b. Methods

Six cases were selected from the 2014/15 cold season for model analysis (Table 1). All observations within ±30 min of 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 grid point valid at the same hour. This procedure is performed separately for NCEP-RAP and ESRL-RAP at 3-, 6-, 9-, 12-, 15-, and 18-h forecast lead times for each case as well as a composite of all cases. Only locations for which precipitation is forecast by both versions of RAP and observed through mPING are considered for this study.

The resulting comparisons are analyzed using three different statistics: the Gerrity skill score (GSS; Gerrity 1992), the Peirce skill score (PSS; Peirce 1884), and bias. The GSS determines the skill for all four ordered precipitation types simultaneously. Like PSS, 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 antiperfect forecast, 0 is the sample climatology or constant forecast (e.g., no skill), and 1 is a perfect forecast.

The PSS and bias are applied individually to each precipitation type for each version of the RAP. 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, a bias less than 1 is an underforecast, and a bias above 1 is an overforecast of the ptype; however, a bias of 1 does not necessarily imply that the forecast was correct, only that the ptype is forecasted as often as it is observed.

A 95% confidence interval based on bootstrap resampling is provided for each statistic. Matched-pair permutation tests are used to determine how likely it is that
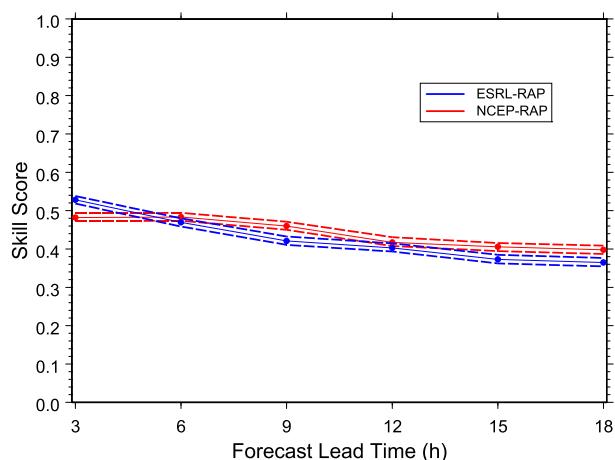

FIG. 1. GSS for the full case composite analyzed at 3-h forecast lead time intervals for the ESRL-RAP (blue) and the NCEP-RAP (red) models. Solid lines represent the mean GSS, while the encompassing dashed lines represent the 95% confidence interval based on bootstrap resampling.
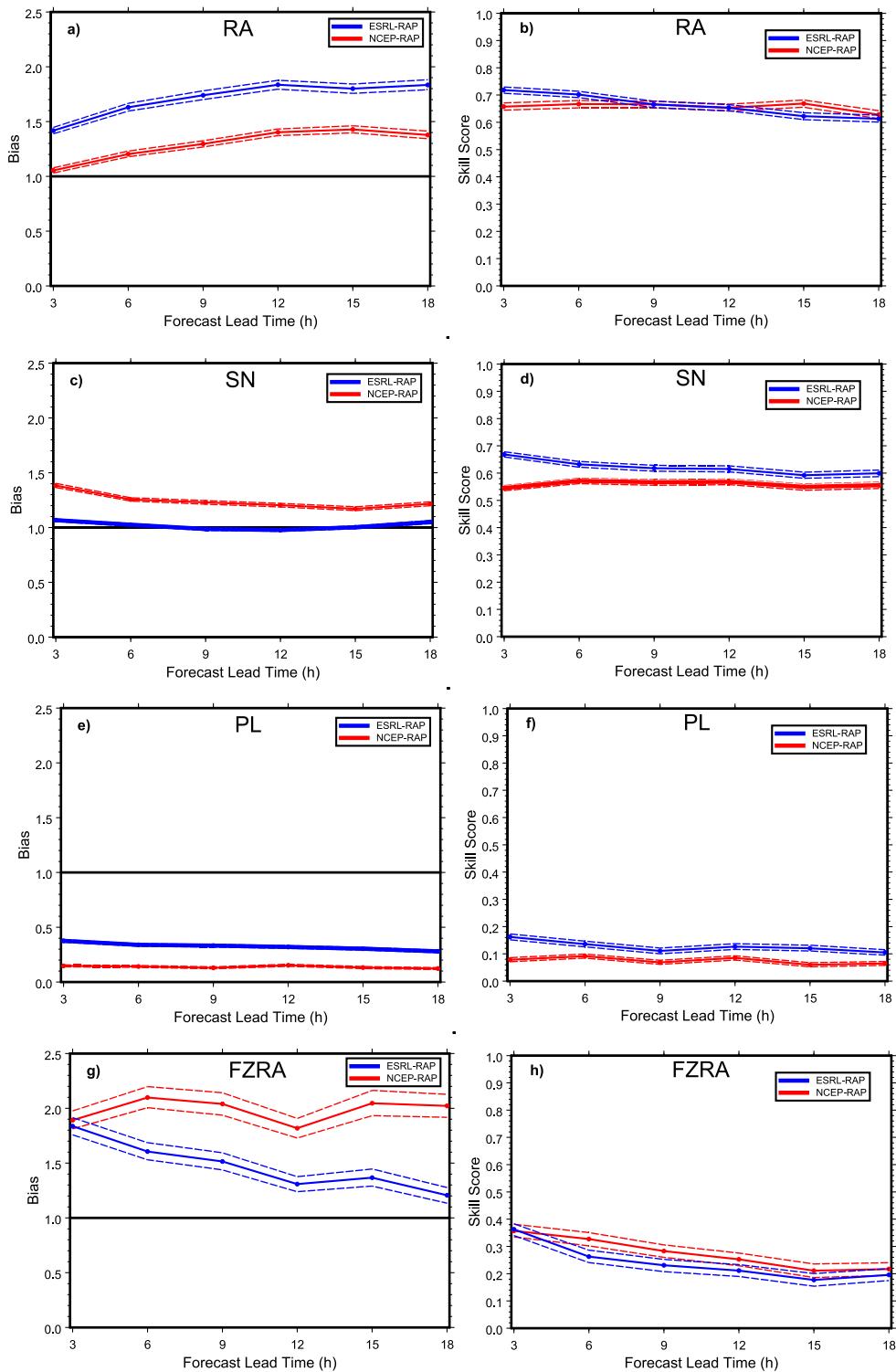
FIG. 2. For all cases, the composite of (a) bias and (b) PSS for rain. (c),(d) As in (a),(b), but for snow. (e),(f) As in (a),(b), but for ice pellets. (g),(h) As in (a),(b), but for freezing rain. The bias and PSS were calculated for the ESRL-RAP (blue) and the NCEP-RAP (red) models. Solid lines represent the mean value, while the encompassing dashed lines represent the 95% confidence interval based on bootstrap resampling.
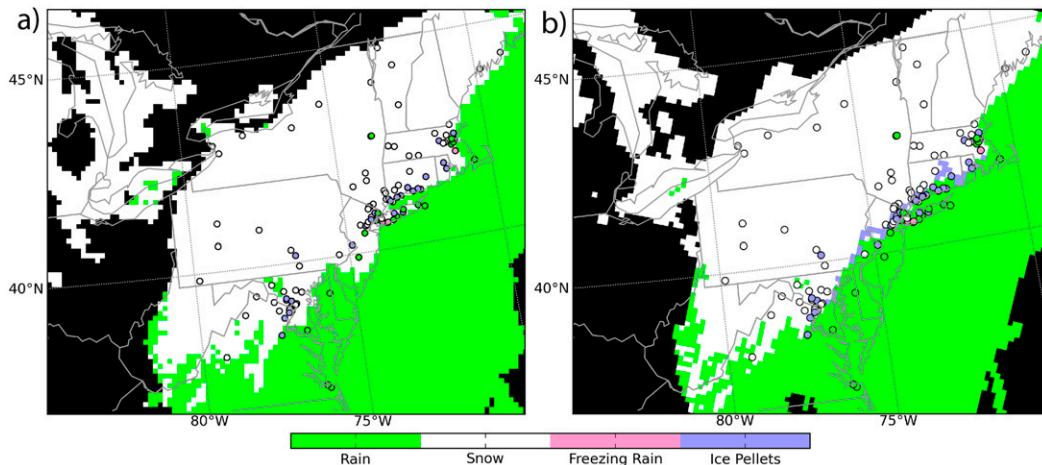
FIG. 3. (a) NCEP-RAP and (b) ESRL-RAP forecast ptypes with a 3-h lead time, valid at 1900 UTC 26 Nov 2014. Modeled ptypes are shown by filled contours, and mPING observations are represented by circles.

the difference between the means of each statistic for the ESRL-RAP and the NCEP-RAP arises by chance.

3. Results

All data presented here and used in this analysis appear in the online supplemental material.

a. Composite of all cases

The GSS, PSS, and bias are computed in 3-h intervals from 3- to 18-h forecast lead times for the composite of all cases. Figure 1 depicts the mean GSS and 95% bootstrap confidence interval for both versions of the RAP. The ESRL-RAP performs relatively well with a mean GSS of 0.525 for a 3-h forecast, whereas the NCEP-RAP has a mean GSS of 0.481. Permutation tests yield a p value < 0.002, indicating that the improved skill of the ESRL-RAP is statistically distinct from that of NCEP-RAP. The GSS of the ESRL-RAP degrades with increasing lead time, however, and by 6-h lead time the mean GSS of the ESRL-RAP is 0.461, compared against the mean GSS of the NCEP-RAP at 0.482 ($p < 0.002$). The same trend continues through the remainder of the 18-h forecast range, with the mean GSS gradually decreasing with increasing lead time with the permutation test $p < 0.01$.

The composite sample is broken down into the four primary precipitation types to help determine how the performance of any particular ptype is mirrored within the other ptypes. Figures 2a–h depict the bias and PSS for each precipitation type. Both versions of the RAP overforecast RA (i.e., bias > 1), while the ESRL-RAP has a persistently higher mean bias than the NCEP-RAP for RA. The ESRL-RAP shows a bias very near 1 for SN, while the ESRL-RAP shows a substantially lower bias

than the NCEP-RAP for FZRA, but with a continued tendency to overforecast FZRA. For PL, the category for which the earlier version performed worst, the NCEP-RAP still has a very low bias between 0.12 and 0.15 depending on lead time; however, this is much improved over the version evaluated in Elmore et al. (2015). Thus, the NCEP-RAP still underforecasts ice pellets compared with either the NAM or GFS (Elmore et al. 2015). The ESRL-RAP does better, with PL showing an underforecast bias between 0.28 and 0.38 depending on lead time. The mean bias differences between the ESRL-RAP and NCEP-RAP have p values less than 0.01 for RA, SN, PL, and FZRA, except for FZRA at the 3-h lead time, for which $0.01 < p < 0.05$.

An analysis of the PSS for the individual precipitation types shows minor differences for rain, with a higher PSS for the ESRL-RAP at 3-h lead time and lower PSS at 18-h lead time, while the PSS for freezing rain is typically lower for the ESRL-RAP than the NCEP-RAP. PSS for the ESRL-RAP is higher than for the NCEP-RAP for snow and ice pellets, where $p < 0.01$.

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 results in different outcomes for each case. To further highlight the extent of these day-to-day variabilities, two cases are analyzed in more detail below.

b. 26–27 November 2014

The 26–27 November 2014 case was dominated by a coastal low pressure system along the East Coast that produced an early season snowstorm across the mid-Atlantic and into the New England regions. This case displays the lowest skill scores out of any case analyzed in this study. The NCEP-RAP depicts very little mixed
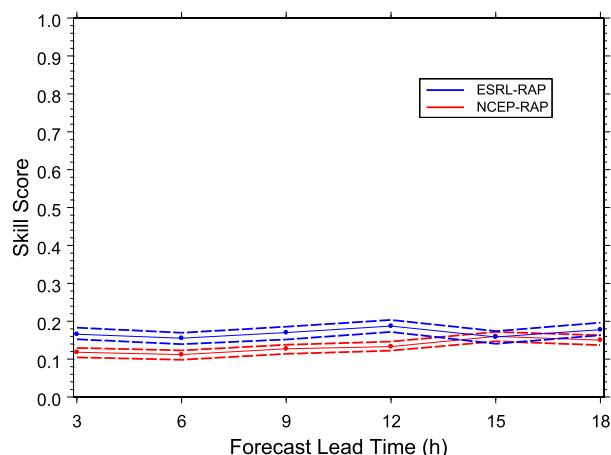
FIG. 4. As in Fig. 1, but for the 26–27 Nov 2014 event.

precipitation while mPING observations suggest otherwise. For the 3-h lead time, there are 448 spatially distributed ice pellet and 33 spatially distributed freezing rain mPING observations, while the NCEP-RAP has only 1 ice pellet and 2 freezing rain forecasts across all mPING observations.

The NCEP-RAP depicts only RA or SN (Fig. 3), despite the presence of numerous ice pellet and freezing rain mPING reports from Washington, D.C., to Boston, MA, which explains its low scores relative to the other cases. The skill of the ESRL-RAP is improved with its depiction of a narrow axis of ice pellets from Maryland into coastal New England along with the depiction of rain over southern New Jersey and Long Island, which indicates a closer match to mPING observations than the NCEP-RAP.

The mean GSS for the ESRL-RAP (Fig. 4) generally alternates between 0.15 and 0.18, which is a small but substantial ($p < 0.01$) improvement over the NCEP-RAP. The mean GSS for the NCEP-RAP alternates

between 0.10 and 0.16. Even so, this difference may be of little practical utility.

c. 3–5 March 2015

The synoptic setup for the 3–5 March 2015 case consisted of two rounds: a widespread snow and ice pellet event in the northeast United States on 3 March 2015, followed by a slow southward progression of a strong baroclinic zone extending from the southern plains into the mid-Atlantic region that produced 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 ($p < 0.05$).

Another key difference between the two versions of the RAP is the depiction of the ptype transition zones (Fig. 5). When the model output is collapsed to the four canonical precipitation types, the NCEP-RAP depicts an unrealistic transition zone from rain to snow at the 3-h lead time, particularly over Tennessee where the ptype from south to north changes from rain to snow, then to ice pellets, then to freezing rain, then back to ice pellets, and finally to 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 from rain to freezing rain to ice pellets to snow. Other cases analyzed in this study display the same characteristic. More ice pellets are being depicted within the ESRL-RAP than in the NCEP-RAP, although a bias toward overforecasting rain is apparent in the ESRL-RAP. This is particularly noticeable over western Tennessee and Arkansas, where mPING observations show sleet while the model depicts rain.

The mean GSS for the ESRL-RAP (Fig. 6) degrades from 0.497 at the 3-h forecast to 0.295 by the 18-h forecast. For the NCEP-RAP, the mean GSS peaks at 0.515 at the 6-h forecast before gradually decreasing to 0.375 by the 18-h forecast, indicating a persistent signal
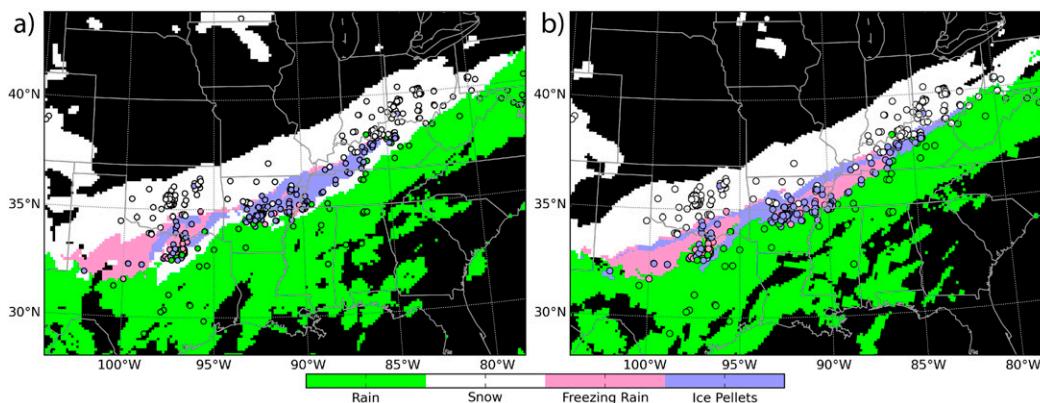

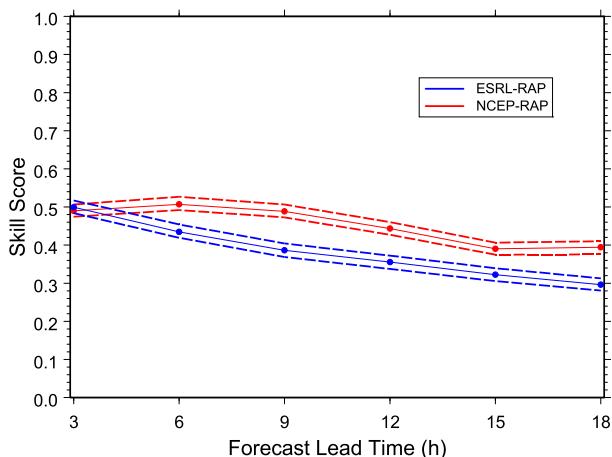FIG. 5. As in Fig. 3, but valid at 2300 UTC 4 Mar 2015.

FIG. 6. As in Fig. 1, but for the 3–5 Mar 2015 event.

that the NCEP-RAP performs better than the ESRL-RAP ($p < 0.01$), except for the 3-h lead time, where $p > 0.05$.

4. Discussion

The ESRL-RAP clearly shows an incremental improvement in forecasting PL and SN. Forecasts for FZRA and RA show either no improvement or slightly degraded skill within the ESRL-RAP. There was still case-to-case variability, although the individual cases analyzed all show improvement for PL.

Other differences between the two models may be at play, since Figs. 3 and 5 show subtle differences in the spatial extent of the precipitation between both versions of the RAP. Currently, the source(s) of these differences is not known, although indications are that other changes exist between both versions of the RAP, which prevents isolating the ice pellet diagnostic from other changes. Only six cases are analyzed in this study; thus, a complete picture of the day-to-day variability typical of precipitation events may not be captured. Finally, while most of the differences in skill scores and bias between both versions of the RAP display small p, typically $p < 0.01$, these differences probably have little practical significance because it is unlikely an observer would perceive these differences in forecast skill, though a more skillful forecast may result in some economic benefit if assessed over a long enough period. This is particularly true when the differences are very small, as is the case with the PSS for freezing rain.

5. Concluding thoughts

Both the NCEP- and ESRL-RAP models include a major ice pellet correction. Parallel runs of the ESRL-RAP,

which incorporates an ice pellet diagnosis refinement, and the NCEP-RAP, which does not, are verified with mPING observations to assess whether the refinement has improved the performance of the RAP model precipitation-type diagnosis. GSS values are computed for individual cases and the composite of all cases, along with PSS and bias for each individual precipitation type. Bootstrap confidence intervals are computed, and the statistical nature of any differences are derived from permutation tests. Results suggest that the ESRL-RAP enjoys incremental and significant improvement ($p < 0.01$) in the skill and bias for PL and SN, but either no change or decreased performance for FZRA and RA. Even with the improvements in the ice pellet diagnosis, the same bias toward underforecasting PL continues, albeit to a lesser extent. A visual analysis of the ESRL-RAP also reveals a more reasonable spatial distribution of ptypes in transition zones involving PL and FZRA. Various issues, such as the small sample size and inherent day-to-day variability, preclude higher confidence on the exact nature of the improvement in the ESRL-RAP.

REFERENCES

Benjamin, and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

Brown, J. M., and Coauthors, 2011: Improvement and testing of WRF physics options for application to Rapid Refresh and High Resolution Rapid Refresh. *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.]

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.*, **95**, 1335–1342, doi:10.1175/BAMS-D-13-00014.1.

——, 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.

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.

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.

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.

Janjić, 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. [Available online at http://www2.mmm.ucar.edu/wrf/users/workshops/WS2005/presentations/session2/9-Janjic.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.]

NOAA/ESRL, 2015: RAP and HRRR variables. NOAA/Earth Systems Research Laboratory. [Available online at http://ruc.noaa.gov/rr/RAP_var_diagnosis.html.]

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454, doi:10.1126/science.ns-4.93.453-a.

Rutledge, G. K., J. Alpert, and W. Ebisuzaki, 2006: NOMADS: A climate and weather model archive at the National Oceanic and Atmospheric Administration. *Bull. Amer. Meteor. Soc.*, **87**, 327–341, doi:10.1175/BAMS-87-3-327.

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 new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:10.1175/2008MWR2387.1.