USING mPING OBSERVATIONS TO VERIFY SURFACE PRECIPITATION TYPE FORECASTS FROM NUMERICAL MODELS

Deanna Apps[1, 2] , Kimberly Elmore[3,4], and Heather Grams[3,4]

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

[2]State University of New York at Oswego, Oswego, NY

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

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

ABSTRACT

The mPING app allows the public citizen to submit reports of the weather occurring at their location from anywhere on the globe. This study uses precipitation type reports made through mPING in the continental United States to verify precipitation type forecasts of operational numerical models. The models evaluated are the North American Mesoscale (NAM) model, the Global Forecast System (GFS), and the Rapid Refresh (RAP) model. Strengths and weaknesses of each model's forecast are investigated for freezing rain, ice pellets, rain, and snow. The Heidke and Peirce skills scores are used predominantly, along with other performance measures. Overall, the models show less skill in the rare events of freezing rain and ice pellets, while overcompensating those precipitation types for rain or snow.

1. INTRODUCTION

Precipitation type during winter weather affects society and the lives of individual people. Travel at the ground and in the sky becomes difficult. Infrastructure and commerce are also greatly affected. Studies have shown that winter weather has significant impacts. A total of 87 catastrophic freezing rainstorms occurred between 1949 and 2000, resulting in losses totaling $16.3 billion (Changnon 2003). Accurate forecasts of mixed precipitation types can potentially reduce the cost of winter storms by allowing for better advanced preparation.

Precipitation type forecasting is always a challenge for meteorologists. The microphysical processes that determine precipitation type are well understood, but the forecasting of precipitation type is lacking in skill. Small variations in atmospheric parameters affect thermodynamic processes that can result in changes in the precipitation type at the surface (Lackmann 2011).

To help verify forecasted precipitation type, data was collected through the meteorological Phenomena Identification Near Ground (mPING) Project. The mPING app was introduced in December 2012 (Elmore et al. 2013) and allows the public citizen to identify precipitation type at their current time and location anywhere in the continental United States. This study uses mPING reports from two winter weather events.

The operational numerical models analyzed in this study were the Rapid Refresh (RAP) model, North American Mesoscale Model(NAM), and the Global Forecasting System(GFS). Each model's forecast ability of precipitation type during the two events was analyzed using primarily the Peirce and Heidke skill scores. Resampling techniques

[1] Deanna Apps, National Weather Center Research Experiences for Undergraduates Program, Center for Analysis and Prediction of Storms, The University of Oklahoma, National Weather Center, 120 David L. Boren Blvd, Suite 2500, Norman, OK 73072
Emaill: dapps@oswego.edu.

are used to place bounds on the accuracy of the scores and to determine the statistical significance of difference between scores from various models and across two lead times. Other performance measures are also used.

The results of this study are useful for forecasters and anyone else affected by winter weather events. The model's ability to predict different precipitation types are investigated to understand the strengths and weaknesses a model might have.

2. DATA AND METHODS

a. mPING

Observations used to verify the model's forecast were collected through mPING reports for two time periods: 8 February 2013, 1200 UTC-0000 UTC and 21-22 February 2013, 1200 UTC-0600 UTC. These events were specifically chosen due to the large number of reports and the variety of precipitation types observed. This study is focusing on the model's ability to forecast different precipitation types including freezing rain and ice pellets, so events containing only rain or snow were of lesser interest.

At the time of data collection, there were 13 different precipitation types for an mPING observer to report. These were: drizzle, freezing drizzle, rain, freezing rain, ice pellets/sleet, graupel, wet snow, snow, mixed rain and snow, mixed rain and ice pellets, mixed ice pellets and snow, hail, and none. "Test" is also a report option so users can see how the app works and to keep people from submitting false reports. Reports of "none," while available and used commonly were not investigated during this study. The goal of this project is to see how well the models forecast different precipitation types when both the model predicts precipitation and the observer reports precipitation, not if the model did or did not forecast precipitation.

In order to consolidate the data for analysis, the observations are aggregated into four categories. The four categories are: rain, snow, freezing rain, and ice pellets. If mixes were reported the report was collapsed to the precipitation type that gave better consistency or agreement among the area of precipitation. For example, a report of rain and ice pellets is collapsed to ice pellets. Furthermore, the ability of the model to forecast ice pellets has a greater impact on infrastructure and travel.

b. Numerical Models

Precipitation type forecast verification is performed on three operational numerical models, which are: the RAP, GFS, and NAM. Comparisons between model forecasts use the skill scores from both the three and six hour forecasts. In this study, the RAP model was broken down into two different variations, identified as RAP1 and RAP2. The forecasts for precipitation type are not mutually exclusive which means; two precipitation types can be forecast valid at the same time (Benjamin 2013). If the mixed precipitation type forecasts were the same as possible mPING observations they were collapsed the same as the observations were. However, there is a case when the model forecasts freezing rain and snow, implying a mix of the two. Because there is no mix of freezing rain and snow available for an mPING report, and due to the fact that the combination of the two types is rare, such reports are collapsed to both snow and freezing rain. This created two different variations of the RAP. The RAP1 is the model with those forecasts collapsed to snow. RAP2 is the model with those forecasts collapsed to freezing rain. The RAP is run hourly and uses a complex cloud scheme that uses mixing ratios, fall rates, and precipitation rate to predict precipitation type.

The NAM is run four times a day at 0000 UTC, 0600 UTC, 1200 UTC, and 1800 UTC. The NAM post processes precipitation type using five algorithms. The five algorithms are the Baldwin-Schichtel, Ramer, Bourgouin, explicit microphysics, and Revised Baldwin (UCAR 2011). The predominant precipitation type from the ensemble of algorithms is forecasted.

The GFS is also run four times a day at the same times as the NAM and uses an ensemble of four algorithms to determine precipitation type. These four algorithms are Baldwin, Revised Bladwin, Ramer, and Bourgouin (Evans and Graham 2011).

c. Algorithms

The NAM uses an explicit microphysics scheme and so uses one more algorithm than the GFS. The Baldwin-Schichtel algorithm uses a multiple step process to determine precipitation type. This process begins by finding the highest

4x4 Table	Rain	Snow	Freezing Rain	Ice Pellets
Rain	y_1, o_1	y_1, o_2	y_1, o_3	y_1, o_4
Snow	y_2, o_1	y_2, o_2	y_2, o_3	y_2, o_4
Freezing Rain	y_3, o_1	y_3, o_2	y_3, o_3	y_3, o_4
Ice Pellets	y_4, o_1	y_4, o_2	y_4, o_3	y_4, o_4

Figure 1. A 4x4 Contingency table with the four categories of precipitation type. The cells correspond to the values in the HSS and PSS equations.

saturated layer above the ground and determines the temperature of that layer. Then the melting and freezing layers a hydrometeor encounters on its way to the surface are found by calculating the area between the 0°C or -4°C isotherm and the wet-bulb temperature. This concept is applied at the surface and aloft within the layers (Evans and Graham 2011).

The revised version of the Baldwin scheme changes a condition to create a bias toward snow to avoid a bias for freezing rain and ice pellets. When ice crystals are determined from the saturated layer, then the area between the wet-bulb temperature and the 0°C wet bulb isotherm is used (Glass 2008).

The Ramer method uses relative humidity, temperature, and wet-bulb temperature to determine precipitation type. At different pressure levels where precipitation is likely to occur, the ice fraction is calculated and used to determine precipitation type in this scheme (Ramer 1993).

The Bourgouin method calculates the area enclosed for a dry, adiabatic process that follows the 0°C isotherm and the environmental temperature. The area calculated is then classified into conditions to diagnose precipitation type (Bourgouin 2000).

The NAM uses one more algorithm, the microphysics scheme. This scheme uses frozen hydrometeor fraction and skin temperature to classify precipitation type (UCAR 2011).

3. VERIFICATION

mPING observations are first used to verify the model's forecast for all precipitation types, which is done by constructing 4x4 contingency tables (Figure 1). To further test model performance, 2x2 contingency tables are formed for each precipitation type (Figure 2).

2x2 Table	Snow	Not Snow
Snow	Forecasted and Observed **a**	Forecasted, not observed **b**
Not Snow	Not forecasted, but observed **c**	Not forecasted and not observed **d**

Figure 2. A 2x2 contingency table showing what each cell letter represents. The example used here is for snow.

a. 4x4 Contingency Tables

4x4 contingency tables are constructed to evaluate the skill of each model (Figure 1). Values along the main diagonal represent a correct classification of precipitation type. Values that stray off this diagonal represent a misclassification. The Peirce and Heidke skill scores were calculated from the contingency tables. Both skill scores use the joint and marginal distributions of forecasts and observations (Wilks 2006). Equations 1 and 2 show the general form of the Heidke (HSS) and Peirce (PSS):

$$HSS = \frac{\sum_{i=1}^{I} p(y_i,o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{i=1}^{I} p(y_i)p(o_i)} \qquad \text{(Eq. 1)}$$

$$PSS = \frac{\sum_{i=1}^{I} p(y_i,o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{j=1}^{J} [p(o_j)]^2} \qquad \text{(Eq. 2)}$$

In this project *y*, represents the forecasts and *o*, represents the observations. Where $p(y_i,o_i)$, is the joint distribution and $p(y_i)p(o_i)$ is the marginal distribution. *I* is the number of possible forecasts and *J* is the number of possible outcomes.

When using a contingency table larger than the standard 2x2, only certain scores are available and the two used here are the HSS and PSS. Possible values of both the HSS and PSS skill scores span the interval [-1,1]. A score of 1 is a perfect forecast for the event. A value of 0 indicates zero skill. A value of -1 is of interest only in the case of 2x2 tables and indicates a perfect inverted forecast (the sense of the forecast is backwards). Negative skill is rarely encountered in contingency tables larger then 2x2.

b. 2x2 Contingency Tables

The 4x4 contingency tables are broken down into 2x2 contingency tables for each of the four precipitation types. This was done to provide information about which precipitation type was affecting the skill scores and to analyze where a precipitation type is problematic for the model.
An example of the 2x2 for snow would have two categories: “snow” and “not snow” (Figure 2). With the 2x2 contingency table, performance measures that demonstrate biases can be applied to the data set and skill scores can be calculated. However, the focus will stay on the HSS and PSS, and we additionally computed bias and hit rate.

$$HSS = \frac{2(ad - bc)}{(a+c)(c+d)+(a+b)(b+d)} \qquad \text{(Eq. 3)}$$

$$PSS = \frac{ad - bc}{(a+c)(b+d)} \qquad \text{(Eq. 4)}$$

$$Bias = \frac{a+b}{a+c} \qquad \text{(Eq. 5)}$$

$$HitRate = \frac{a+d}{a+b+c+d} \qquad \text{(Eq. 6)}$$

As shown in Figure 2, *a* represents a hit, *b* represents a false alarm, *c* represents a miss, and *d* represents a correct null forecast. The HSS and PSS calculations are now easier to interpret for the 2x2 table. Bias is calculated with a perfect score being a value of 1. Below one the model is under forecasting, above 1 the model is over forecasting. The hit rate is calculated by taking the correct forecasts and correct null forecasts divided by the total number of forecast occasions.

Rain and snow are the most common forms of precipitation, but rare events like freezing rain and ice pellets are still very important and can have detrimental affects on society. High hit rates may be due to many correct null forecasts. In fact in some cases, hit rates might be improved if the event were never forecast to happen (Jolliffe and Stephenson 2003). That’s why more attention will still be drawn on the skill scores because they account for the rarity of events.

4. RESULTS

a. 4x4 Analysis

Bootstrap resampling statistics (Wilks 2006) were used on the PSS and HSS from each model’s three and six hour forecasts. The box plots (Figure 3 and 4) show 95% confidence intervals for each forecast. Also, Table 1 provides the mean HSS and PSS for each model’s three and six-hour forecast. The NAM produced the best skill in both the three and six hour forecasts. As expected the three-hour forecasts for each model are more skillful than the six-hour forecasts.

Model Forecasts	Mean HSS	Mean PSS
GFS 3-hour	0.493	0.478
NAM 3-hour	0.497	0.484
RAP1 3-hour	0.469	0.449
RAP2 3-hour	0.486	0.478
GFS 6-hour	0.434	0.425
NAM 6-hour	0.444	0.455
RAP1 6-hour	0.382	0.384
RAP2 6-hour	0.370	0.397

Table 1. Each model’s three and six-hour mean HSS and PSS for all precipitation types.

Permutation tests were used to determine the statistical difference between the different scores. The threshold used for p-values to determine the “significance” in this study is 0.05. If the value is greater than, 0.05, the difference between model forecasts is not significantly different (in a statistical sense), and so one forecast cannot be declared statistically “better” or more skillful than another.

For, permutation tests on the HSS (Fig. 3), the RAP1 and RAP2 six-hour forecasts had the least skill. The GFS and NAM six-hour forecasts

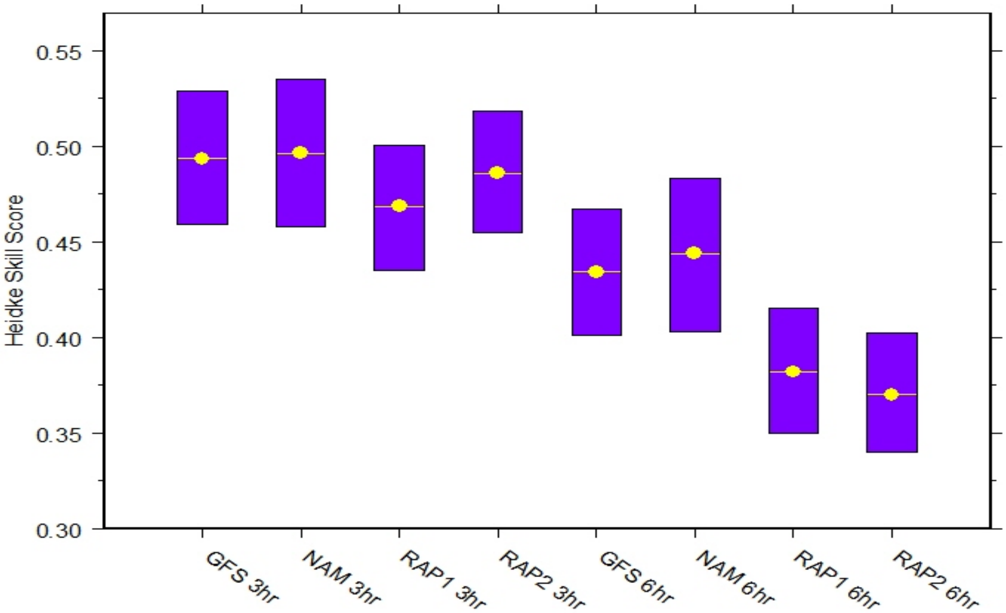

Figure 3. Distribution of the Heidke skill scores resulting from 5000-sample bootstrapping. The circle is the mean of the bias corrected bootstrapped confidence intervals. The lower end of the box is the 2.5 percentile and the upper end is the 97.5 percentile. Therefore, the whole box represents 95%.

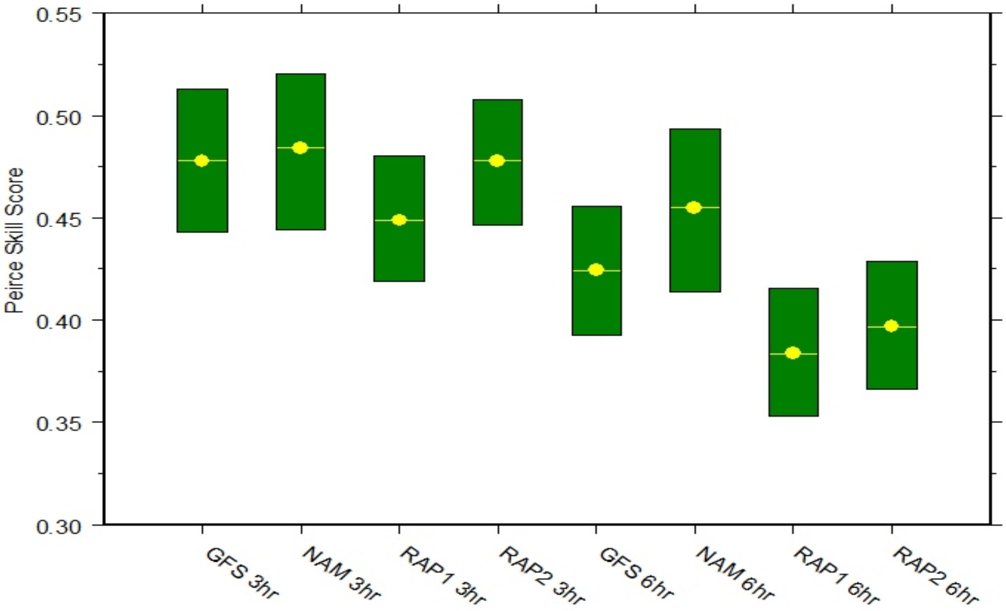

Figure 4. Distribution of the Peirce skill scores resulting from 5000-sample bootstrapping. The circle is the mean of the bias corrected bootstrapped confidence intervals. The lower end of the box is the 2.5 percentile and the upper end is the 97.5 percentile. Therefore, the whole box represents 95%.

were both significantly better than both RAP models. The RAP1 had the least skill in the three-hour forecast. The NAM had the best skill score for both the three and six-hour forecasts.

The PSS (Fig. 4) shows that the GFS six-hour forecast is not significantly different than the RAP1 and RAP2 six-hour forecast, though the p-value is close to 0.05. However, the NAM is still significantly better than the RAP1 and RAP2. A larger data set is needed to declare statistical significance. The NAM again had the best skill scores for the three and six-hour forecasts.

b. 2x2 Analysis

The three-hour (Table 2) and six-hour (Table 3) forecasts of each precipitation type didn't show conclusive evidence of one model being superior to the others. Each model had its successes, but not one model outperformed the others in all precipitation types.

For freezing rain, the six-hour forecast shows the GFS having the most skill, but for the three-hour forecast shows the RAP2 has the most skill. The RAP2 is the variation of the RAP where forecasts of snow and freezing rain were collapsed to just freezing rain. The RAP2 also has the highest bias score for all the models for freezing rain at 2.27 for the six-hour forecast. Thus, the RAP2 over-predicted freezing rain. Even the RAP1 has a bias score of over 1, however as stated previously, the hit rate is high for freezing rain; this is due to the correct null forecasts (d) that are added to the correct forecast (a).

The NAM had the best skill out of all the models for forecasting ice pellets. However, the skill scores for each model are low. All models for both the three and six hour forecast have a bias below 1. Bias scores less than 1 indicate that the event is under-forecasted. Ice pellets seem to be the greatest challenge, especially for the RAP. The skill scores and bias are both close to zero, which means that the RAP seldom forecasts ice pellets, if at all, for the two cases that were evaluated.

For rain, the result is quite different. The bias scores for all rain forecasts are above 1. The bias is especially high for the RAP1 and RAP2, scoring 2.04 on the six-hour forecast. The NAM had the best hit rates for rain, but had the best skill in the six-hour forecast. The RAP1 and RAP2 three-hour forecasts achieved a HSS of 0.664 and a PSS of 0.799 for rain.

Results are similar for snow, where the NAM six-hour forecast displays the most skill while the three-hour forecast of the RAP2 is most skillful. However, the RAP1 is the variation that is collapsed to snow. The low skill of the RAP1 achieved comes from the over-forecast of snow. Its bias is slightly higher than the RAP2, which causes the drop in skill.

It is clear in Tables 2 and 3 that models drop off in skill significantly when forecasting freezing rain and ice pellets. Figure 3 and 4 show that the skill scores for all precipitation types are in between the skill scores that are seen for rain and snow versus freezing rain and ice pellets. The overcompensation for rain and snow and the undercompensation for freezing rain and ice pellets evens out, when looking at all precipitation types.

c. Forecast Percent

In this study, we did not have a category for no precipitation forecasted, where precipitation was observed. Therefore, if the model didn't predict precipitation for an observed report, the report was not included in the analysis. This caused a different amount of forecasts and observations for each model. For the three-hour forecasts there were 1,539 mPING observations. The GFS forecasted precipitation for 94.89% of those points. The NAM forecasted precipitation for 74.94% and the RAP forecasted for 89.27% of those points.

For the six-hour forecasts there were 1,603 mPING observations. The GFS forecasted for the most again, as it forecasted precipitation for 94.67% of those points. The NAM was the lowest again, forecasting precipitation for 67.31% of the points. The RAP forecasted precipitation for 87.71% of those points.

5. CONCLUSIONS

The Heidke and Peirce skill scores were evaluated for each models forecast of precipitation type. Bootstrap statistics were used on the forecasts for all precipitation types. 95% confidence intervals of each model were shown. The NAM had the highest overall skill scores for both the three and six-hour forecasts, for all precipitation types. The highest skill score achieved by all models and forecast times was the NAM three-hour Heidke mean skill score of 0.497. This shows all though the models may have skill,

	Freezing Rain	Ice Pellets	Snow	Rain
GFS	HSS: 0.213 PSS: 0.181 Bias: 0.67 HR: 0.909	HSS: 0.286 PSS: 0.238 Bias: 0.54 HR: 0.803	**HSS: 0.580** **PSS: 0.577** Bias: 1.11 HR: 0.792	HSS: 0.626 **PSS: 0.708** Bias: 1.34 HR: 0.870
NAM	**HSS: 0.074** **PSS: 0.076** Bias: 0.92 HR: 0.874	**HSS: 0.309** **PSS: 0.246** Bias: 0.43 HR: 0.813	HSS: 0.643 PSS: 0.635 Bias: 1.11 HR: 0.826	**HSS: 0.617** PSS: 0.731 Bias: 1.45 HR: 0.882
RAP1	HSS: 0.210 PSS: 0.236 Bias: 1.26 HR: 0.881	**HSS: 0.003** **PSS: 0.003** Bias: 0.003 HR: 0.785	HSS: 0.611 PSS: 0.603 Bias: 1.19 HR: 0.809	**HSS: 0.664** **PSS: 0.799** Bias: 1.52 HR: 0.882
RAP2	**HSS: 0.223** **PSS: 0.297** Bias: 1.72 HR: 0.861	**HSS: 0.003** **PSS: 0.003** Bias: 0.003 HR: 0.785	**HSS: 0.666** **PSS: 0.660** Bias: 1.13 HR: 0.835	**HSS: 0.664** **PSS: 0.799** Bias: 1.52 HR: 0.882

Table 2. The three-hour forecast 2x2 analysis values. Heidke skill score(HSS), Peirce skill score(PSS), Bias, and hit rate (HR) shown for each precipitation type and model. The red numbers represent the worst skill score for each precipitation type. The green numbers represent the best skill scores.

	Freezing Rain	Ice Pellets	Snow	Rain
GFS	**HSS: 0.289** **PSS: 0.241** Bias: 0.636 HR: 0.913	HSS: 0.129 PSS: 0.100 Bias: 0.406 HR: 0.780	HSS: 0.569 PSS: 0.566 Bias: 1.07 HR: 0.787	HSS: 0.509 **PSS: 0.629** Bias: 1.61 HR: 0.818
NAM	HSS: 0.154 PSS: 0.175 Bias: 1.30 HR: 0.863	**HSS: 0.202** **PSS: 0.168** Bias: 0.557 HR: 0.794	**HSS: 0.582** **PSS: 0.586** Bias: 0.92 HR: 0.794	**HSS: 0.575** PSS: 0.703 Bias: 1.55 HR: 0.864
RAP1	HSS: 0.139 **PSS: 0.154** Bias: 1.25 HR: 0.853	**HSS: -0.001** **PSS: -0.001** Bias: 0.004 HR: 0.799	**HSS: 0.536** **PSS: 0.533** Bias: 1.04 HR: 0.773	**HSS: 0.500** **PSS: 0.713** Bias: 2.04 HR: 0.815
RAP2	**HSS: 0.103** PSS: 0.162 Bias: 2.27 HR: 0.783	**HSS: -0.001** **PSS: -0.001** Bias: 0.004 HR: 0.799	HSS: 0.553 PSS: 0.562 Bias: 0.883 HR: 0.776	**HSS: 0.500** **PSS: 0.713** Bias: 2.04 HR: 0.815

Table 3. The six-hour forecast 2x2 analysis values. The same set-up as in Table 2.

improvements still need to be made in forecasting precipitation type.

Analysis of each individual precipitation type was done using 2x2 contingency tables. The best skill scores were seen for rain and snow. Forecasts of freezing rain and ice pellets are where the model skill scores dropped off. The lack of skill in forecasting ice pellets and freezing rain, caused the overall drop seen in skill scores for all precipitation types.

Further analysis will need to be done with more winter weather events to have a more conclusive result. However, these initial results are important to provide information on where each model's strengths and weaknesses are. Also, model tendencies of over and under forecasting each precipitation type can help the knowledge of forecasters who are forecasting these events. Further analysis that would be beneficial is to analyze the algorithms and schemes used by the model to forecast precipitation type. This could help to further diagnose what works best in the model's forecast and where improvements can be made, especially, in diagnosing the more rare events, such as freezing rain and ice pellets.

6. ACKNOWLEDGEMENTS

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.

Much thanks to Daphne LaDue, Kevin Kloesel, and Madison Miller for their direction of the Research Experience for Undergraduates at the National Weather Center.

7. REFERENCES

Benjamin, S., cited 2013: Rapid Refresh. [Available online at: http://ruc.noaa.gov/rr/RAP_var_diagnosis.html.]

Bourgouin, P., 2000: A method to determine precipitation type. *Wea. Forecasting*, **15**, 583-592.

Changnon, S. A., 2003: Characteristics of ice storms in the United States. *J. Appl. Meteor.*, **42**, 630-639.

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, and L. P. Rothfusz, 2013: mPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.* (in review).

Evans, M., and R. Graham, cited 2011: Strengths and Weaknesses of P-Type Algorithms. [Available Online at http://www.wdtb.noaa.gov/courses/winterawoc/documents/color_PDFs/IC621.pdf.]

Glass, F. H., cited 2008: NCEP PTYPE Algorithms. [Available online at http://www.crh.noaa.gov/images/lsx/science/winter_wx/NCEP%20PTYPE%20Algorithms__Web%20Version.ppt.]

Jollife, I. T., and D. B. Stephenson, 2003: *Forecast verification a practitioner's guide in atmospheric science.* John Wiley and Sons, 240pp.

Lackmann, G., 2011: Winter Storms. *Midlatitude Synoptic Meteorology*, Amer. Meteor. Soc., 219-246.

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.

UCAR, cited 2011: Operational Models Matrix. [Available online at http://www.meted.ucar.edu/nwp/pcu2/.]

Wilks, D. S., 2006: *Statistical Methods in Atmospheric Sciences.* Academic Press, 627pp.