10B.2 COMPARATIVE VERIFICATION OF EXPERIMENTAL HR QPFS WITH MODEL AND FORECASTER-PREPARED QPFS IN THE NWS

Jerome P. Charba*
Frederick G. Samplatsky
Meteorological Development Laboratory
Office of Science and Technology
National Weather Service, NOAA
Silver Spring, Maryland

1. INTRODUCTION

A high resolution, gridded model output statistics (MOS) application (HR) has been producing quantitative precipitation forecast (QPF) guidance for the contiguous United States (CONUS) in an experimental mode in the National Weather Service (NWS) since June 2008 [Charba and Samplatsky 2009; 2011a; 2011b (henceforth referenced as CS)]. The basic forecast element consists of probabilities for multiple QPF thresholds [PQPF, which includes the probability of measureable (≥ 0.01 in.) precipitation (PoP)]. Also, several supplementary QPF elements are derived from the PQPFs, which consist of the expected value [also called probability-weighted QPF (PW)], "best category" QPF (BC), and continuous QPF (CP; the derivation of each element is described in CS). The PQPF, PW, BC, and CP elements are produced for 6- and 12-h valid periods in the 12 – 156 h range from 0000 and 1200 UTC; PoP forecasts are extended to 192 h. The PQPFs are produced by geographically-regionalized linear regression equations on a 4-km grid (CS), where the geographical coverage is shown in Fig. 1. For future guidance use in the NWS National Digital Forecast Database (NDFD; Glahn and Ruth 2003), the QPF elements are interpolated to an NDFD grid.

In this study, the experimental HR QPF elements are objectively scored with various operational model and forecaster-prepared QPFs in the NWS. The model QPFs consist of those from the NWS National Centers for Environmental Prediction (NCEP) North American Mesoscale model (NAM; Rodgers et al. 2005) and Global Forecast System (GFS; Kanamitsu et al. 1991), and from the Meteorological Development Laboratory (MDL) station–based MOS (Antolik 2000) and gridded MOS programs (GMOS; Glahn et al. 2009). The human-prepared QPFs consist of gridded CONUS guidance produced by precipitation forecasting specialists at the NCEP Hydrometeorological Prediction Center (HPC) and CONUS composites of NDFD QPF sub-grids produced by forecasters at local NWS Weather Forecast Offices (referred to as NDFD).

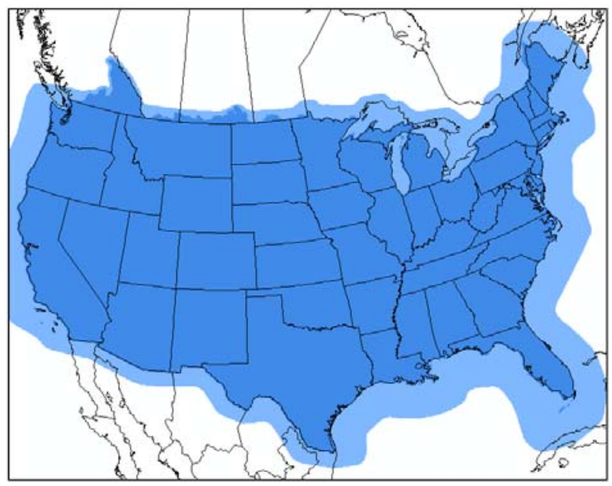

Figure 1. Coverage of the HR QPFs (light blue) and the verification domain (medium blue).

All comparative scoring here is limited to the 1200 UTC model cycle, as HRQPF scores are essentially the same for the 0000 UTC cycle. For 1200 UTC, all model QPFs are available for guidance use by human forecasters by 1800 UTC, and the QPF issuance cutoff time for HPC is 2200 UTC and that for NDFD is 0000 UTC of the following day. [It is noteworthy that QPFs are also produced operationally by NWS River Forecast Centers (RFCs) for use in their hydrologic models, but these were not included in the comparative scoring for two reasons. (1) The issuance cutoff time is several hours after the NDFD cutoff, which would give RFCs an unfair scoring advantage, as the first period QPF issuance is already 3 hours into the

* *Corresponding author address:*
Dr. Jerome P. Charba, National Weather Service
1325 East West Highway, Room 10410
Silver Spring, MD 20910-3283
email: jerome.charba@noaa.gov

6-h valid period when the forecast is issued. (2) RFC QPFs for the 1200 UTC model cycle were often missing in the verification sample archive provided by the National Precipitation Verification Unit (NPVU) of HPC possibly because of the limited schedule of RFC operations.]

The verification database is "stage IV" 6-h quantitative precipitation estimates (QPE; http://www.emc.nceo.noaa.gov/mmb/yling/pcpan/st) following supplemental quality control (QC) at MDL (CS). Most of the scoring is performed on the 4-km grid (Fig. 1), which is native for the verifying QPEs and the HR QPF elements. In the case of the alternative scoring at MOS stations, the QPE data are interpolated to irregularly-spaced NWS stations. While MOS QPFs apply to gage precipitation measurements, CS reported that gage data and QPE data interpolated to the same locations could be used interchangeably for forecasting applications. Finally, the verification periods are October 2009 - March 2010 (cool season) and April – September 2010 (warm season), except where noted otherwise.

2. GRIDDED CATEGORICAL QPF

In the gridded categorical QPF scoring, the HR model QPF is comparatively scored with the NAM, GFS, GMOS model QPF and with the HPC and NDFD forecaster-prepared QPF. Since the scoring is performed on the 4-km grid, all QPFs except that from the HR model were interpolated from a native grid, where an area-preserving method (NWS 1974) for QPFs expressed as continuous precipitation amounts was used. Then, these "continuous QPFs" and the verifying QPEs were converted to categorical (binary) form by applying precipitation threshold values. For instance, for a given threshold the associated categorical QPF and QPE assume a value of 1 when the threshold is met and 0 otherwise.

The threat score [which is the same as the critical success index (Schaefer 1990)] for six 6-h QPF sources and four selected precipitation thresholds is shown in Fig. 2. [Because of the unique properties of HR-PW and the close similarity of HR-BC with the categorical form of HR-CP (CS), only the latter QPF element was included in the scoring.] The threat scores in Fig. 2 are shown as day 1 (12 – 30 h forecast projections from 1200 UTC) and day 3 (60 – 78 h projections) CONUS averages, as the day 2 scores agreed with the day 1 - 3 trend and QPFs from the NAM, HPC, and NDFD are not available beyond day 3.

From Fig. 2, we see that HR-CP had the highest (best) threat scores across all thresholds and for both the cool season and the warm season, though the improvement over HPC (second best) and NDFD was small for the lighter precipitation thresholds during the warm season. Note also that the HR-CP threat score improvement over the three other models (NAM, GFS, and GMOS) is large over all thresholds; the improvement is also large over HPC and NDFD for the ≥ 1.00 in and ≥ 2.00 in. thresholds. Finally, the GFS threat scores are at least as high as those from the NAM and about the same as those for GMOS for upper precipitation thresholds (GMOS threat scores were usually higher for light thresholds).

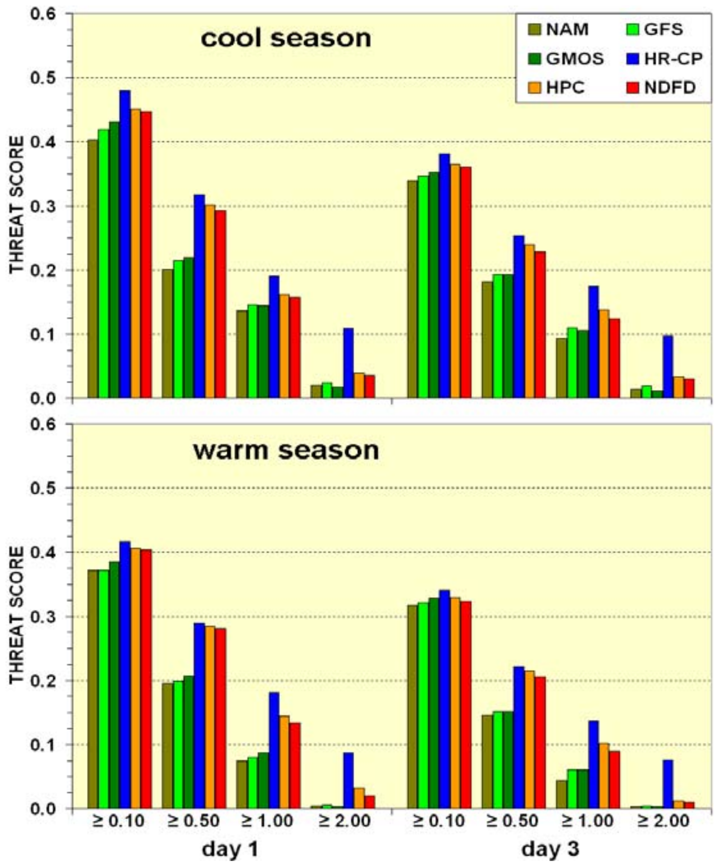

Figure 2. Threat score for gridded 6-h QPFs from six sources (see text), four precipitation thresholds (in.), and day 1 and day 3 during the cool and warm seasons.

Threat scores from all QPF sources were substantially better for the 12-h valid period, as these longer-duration forecasts are less sensitive to timing error. This can be seen by comparing the day 1 12-h QPF threat scores in Fig. 3 with corresponding 6-h QPF scores in Fig. 2. (Scores are not shown in Fig. 3 for HPC and NDFD, as human forecasters there do not issue 12-h QPFs.) Also, note from Fig. 3 that the HR-CP threat score improvement over the three other models for the ≥ 2.00 in. threshold is especially large.

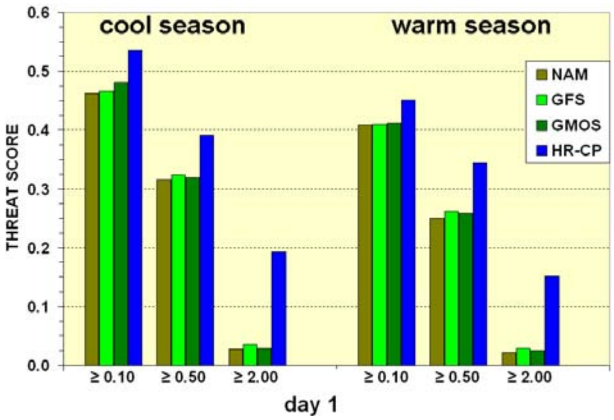

Figure 3. As in Fig. 2, except for the 12-h valid period for day 1, four models, and the ≥ 1.00 in. threshold is not included.

Bias scores (unbiased forecasts have a bias of 1.0) corresponding to the 6-h QPFs in Fig. 2 are shown in Fig. 4 (scores are shown only for day 2 as they were little different for days 1 and 3). A clear feature in Fig. 4 is that the HR-CP bias is quite good across the four precipitation thresholds (the slight overforecasting is by design), whereas all other QPF sources exhibit a striking bias drop-off from ≥ 0.10 in. to ≥ 2.00 in. The bias drop-off for GMOS is due partly to the objective analysis procedure used to obtain QPF grids from station values (Glahn et al. 2009), as evidenced by improved bias for the original station-based MOS QPFs (section 3). Finally, *the strong underforecasting of ≥ 1.00 in. and ≥ 2.00 in. for HPC and NDFD, especially during the warm season, indicates forecasters avoid predicting such heavy precipitation amounts where their accuracy is poor* (note low HPC and NDFD threat scores in Fig. 2).

3. CATEGORICAL QPF AT MOS STATIONS

Here, we consider comparative scoring between HR-CP and MOS categorical QPFs for 1647 irregularly-spaced NWS stations (within the verification domain) for which MOS QPFs are issued. The scoring procedure first required interpolating (with the area-preserving method referenced above) HR-CP and the QPE validation data to the station points. Otherwise, the scoring was identical to the grid scoring discussed in the previous section. Threat and bias scores are shown in Figs. 5 and 6, respectively, for the 6-h valid period, and both scores for the 12-h valid period are shown in Fig. 7.

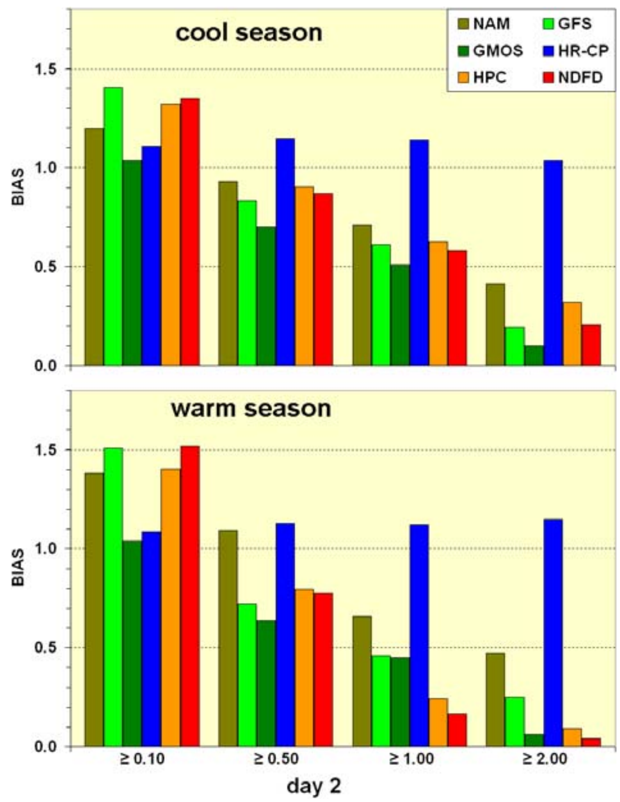

Figure 4. Six-hour QPF bias for day 2; otherwise as in Fig. 2.

From Figs. 5 and 7 (top), we find that the HR-CP threat scores are only slightly higher than MOS threat scores for both seasons and both valid periods. Recall that for the grid verification, the HR-CP threat scores were much better than the GMOS threat scores, especially for the upper precipitation thresholds (Figs. 2 and 3). Careful cross-checking of the HR-CP, GMOS, and MOS threat scores in these figures reveals that the smaller HR threat score improvement over MOS resulted as the MOS threat scores improved substantially on the GMOS threat scores, which is especially true for the higher precipitation thresholds.

The corresponding bias scores, which are shown in Fig. 6 for the 6-h period and Fig. 7 (bottom) for the 12-h period, show HR with slight over-

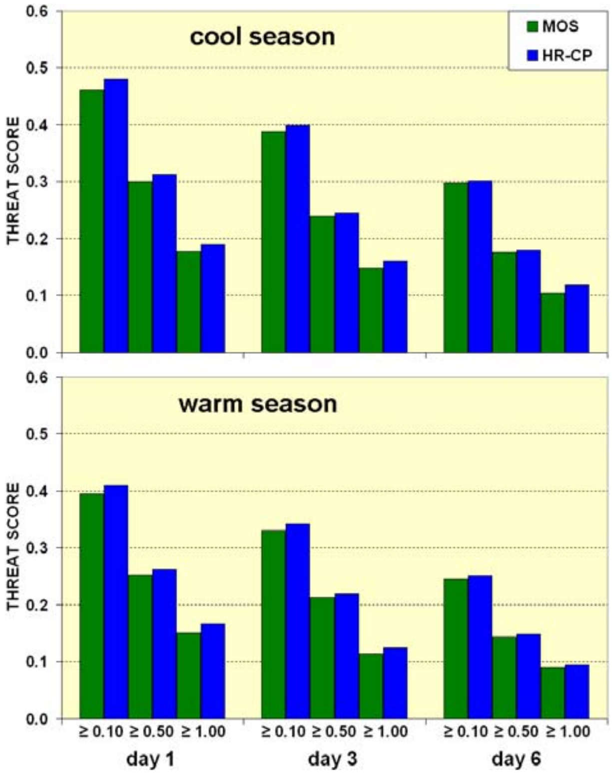

Figure 5. Threat score for categorical 6-h HR-CP and MOS QPFs at 1647 MOS stations for the indicated precipitation thresholds (in.). The color codes match those in Figs. 2 - 4.

forecasting (intentional), which is quite similar to that found for the corresponding grid verification (Fig. 4). For MOS, we again see a bias drop-off with increasing precipitation threshold, but it is less severe than for GMOS (Fig. 4).

The MOS-to-GMOS threat score and bias degradation quantifies the negative impact that GMOS gridding has on the station-based MOS QPFs. Note that this finding is not surprising, as an objective analysis procedure tuned to produce realistic values between stations (Glahn et al. 2009) cannot be expected to yield interpolated values (at the stations) that reproduce the original station values. Thus, the HR-CP grids will replace GMOS QPF grids over the CONUS when the HR model is implemented in early 2011. However, the HR QPFs will not replace the MOS QPFs contained in alphanumeric MOS guidance bulletins (http://www.weather.gov/mdl/synop/gfs.php), as these station-specific QPFs are inherently consistent with other MOS weather elements. Consequently, when the HR-CP replacement of GMOS QPF takes place, users should expect QPF discrepancies between the HR grids and the MOS station values, as *heavy precipitation* is slightly overforecast in the former

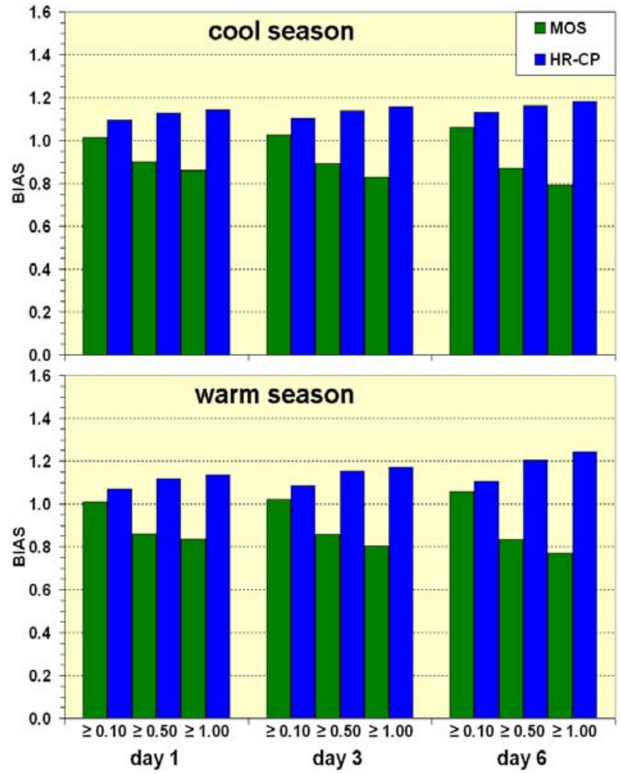

Figure 6. As in Fig. 5 except bias.

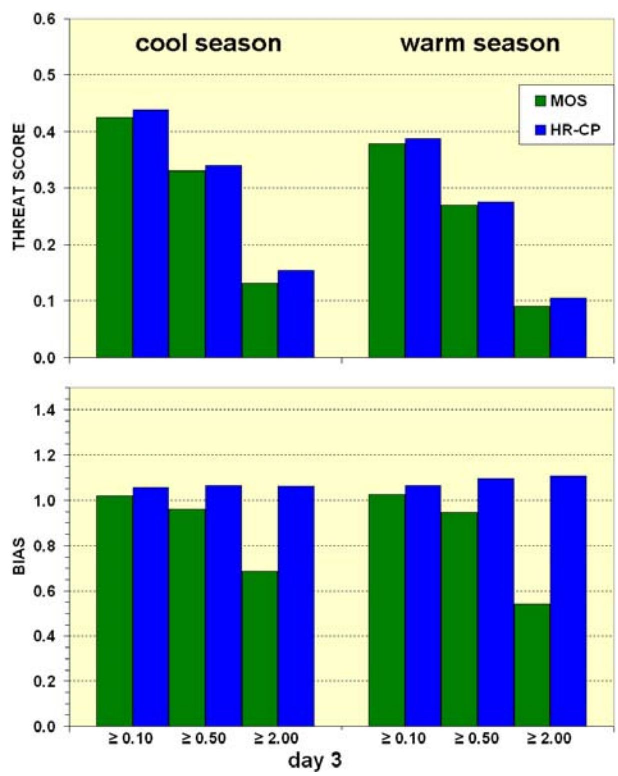

Figure 7. Threat score (top) and bias (bottom) for 12-h categorical QPFs for day 3. Otherwise as in Fig. 5.

(Figs. 4, 6, and 7) and underforecast in the latter (Figs. 6 and 7).

4. GRIDDED PQPF/POP

In this section, gridded PQPFs (including PoP forecasts) from the HR model are comparatively scored with 6- and 12-h GMOS and NDFD PoP forecasts and experimental 6-h PQPFs with a new model implemented by HPC on 1 Feb 2010 (http://www.hpc.ncep.noaa.gov/pqpf_6hr/conus_hpc_qpf_6hr.php; Brill 2010). The performance measure is the Brier score (Brier 1950) improvement on climatology [referred to as the Brier skill score (BSS)] where the incorporated climatology is taken from CS. The scoring was again performed on the 4-km grid, which necessitated interpolation (bi-quadratic interpolation was used) for the GMOS, HPC, and NDFD grids.

Fig. 8 shows HR and HPC 6-h PQPF BSSs for selected precipitation thresholds for day 1 and day 3 (the maximum range of the HPC PQPFs). Here the cool season sample is for the combined periods of 1 February 2010 (when the HPC model was implemented) to 31 March 2010 and 1 October – 30 November 2010. The figure shows HR with slightly higher skill for the lighter precipitation thresholds and HPC with slightly higher skill for heavier thresholds; the ranking switch between these models appears for the ≥ 0.25 in. or ≥ 0.50 in. thresholds for both seasons and both day 1 and day 3 [scores for day 2 (not shown) fell between those for days 1 and 3]. The skill rankings of HR and HPC are similar across the two seasons, though close inspection reveals the relative skill for HR is slightly better during the cool season and the relative skill for HPC is slightly better during the warm season.

Fig. 9 shows the reliability and sharpness (Wilks 2006) of the day 1 HR and HPC PQPFs for the ≥ 0.10 and ≥ 1.00 in. thresholds and both seasons. The reliability charts indicate the HR probabilities are slightly more reliable than those for HPC, as the (averaged) forecast probabilities for HR lie closer to the perfect reliability line than those for HPC. Also, plotted points for HR meander about the perfect reliability line, whereas HPC shows overforecasting throughout most of the probability range. On the other hand, HPC probabilities exhibit better sharpness, as the probability distribution charts show HPC with relatively more cases of 0 % and high probability values. Also, peak HPC probabilities for ≥ 1.00 in. are closer to 100 %. These contrasting reliability and sharpness properties of the HR and HPC PQPFs suggest that predictive information in the two PQPF sources may complement each other.

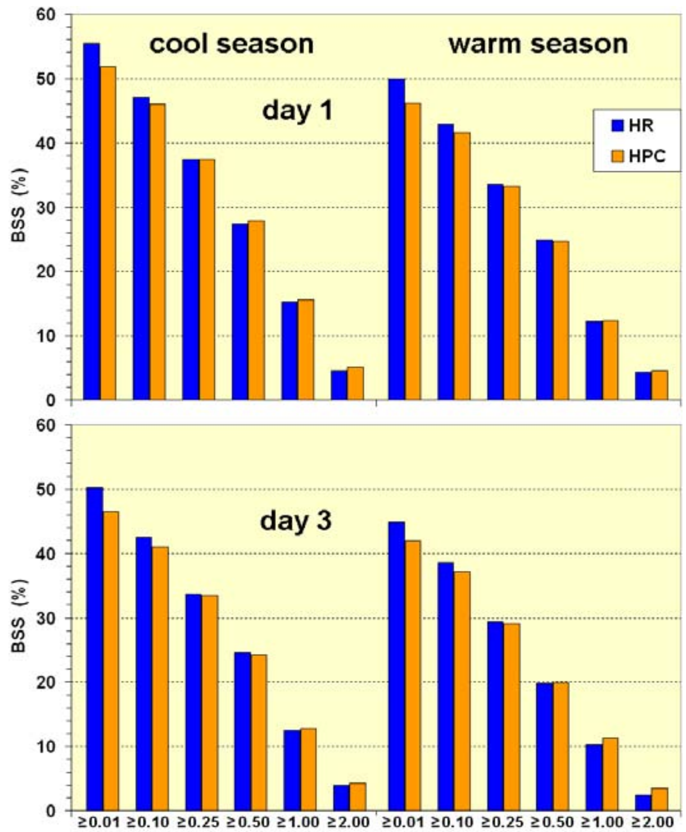

Figure 8. Brier skill score (BSS) for gridded HR and HPC 6-h PQPFs for precipitation thresholds (in.) indicated. The color codes are as in previous figures.

It is noteworthy that HPC conducts a separate 6-h PQPF comparative verification, where the HR PQPFs are included (Brill 2010). While general skill trends in the HPC verification are similar to those presented here, the HR and HPC nominal skill values are different from those in Fig. 8, especially for heavy precipitation thresholds. The score differences are due to several differences between the MDL and HPC verification procedures.

The similar overall skill of the HR and HPC 6-h PQPFs and the contrasting reliability and sharpness properties deserve comment. The competitive skill level of the HR PQPF model is attributed to full use of GFS model output together with extensive predictive use of various fine-scale climatology and topography data (CS). The comparable

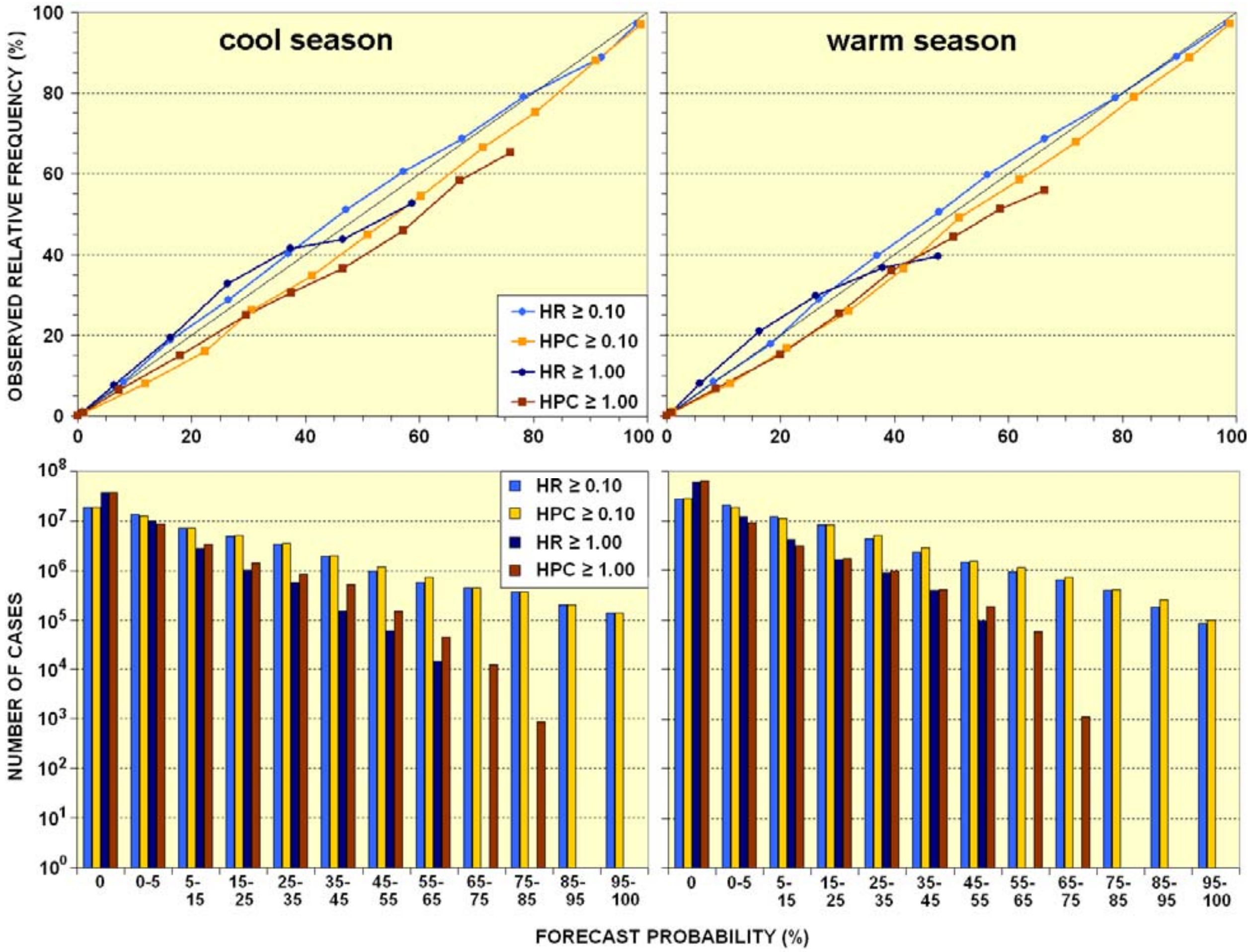

Figure 9. (top) HR and HPC 6-h PQPF day 1 reliability (perfect reliability is indicated by the straight diagonal line) for precipitation thresholds shown in the legends (in.), and (bottom) the number of cases (logarithmic scale) for the (unequal) probability intervals.

skill of the HPC PQPF model is attributed to its incorporation of multiple sources of QPFs, which includes the HPC human-prepared QPFs as well as QPFs from four numerical weather prediction models (Brill 2010). The additional finding of contrasting reliability and sharpness of the HR and HPC 6-h PQPFs suggests combining them would result in a superior multi-model PQPF product. Objective techniques for formulating such a combined 6-h PQPF product should be investigated. Also, it is noted the HR model produces gridded 12-h PQPFs, which have similar skill properties to those for the 6-h valid period. However, 12-h PQPF skill scores are not presented here as a comparable 12-h PQPF product is not available.

Turning to gridded 6- and 12-h PoP forecasts, the performance scoring compares BSSs for HR with BSSs for GMOS and NDFD out to day 8. For the 6-h valid period, where PoP BSSs are shown only for HR and GMOS (Fig. 10; HPC scores are not shown as the experimental HPC 6-h PoPs do not extend beyond day 3 and 6-h PoP is not produced in NDFD), we see that HR BSSs are substantially better over days 1 to 8. For the 12-h valid period where HR, NDFD, and GMOS PoPs are available out to day 7, HR scored clearly better than both GMOS and NDFD, though NDFD is more competitive.

5. POP AT MOS STATIONS

Since current 6- and 12-h MOS PoPs will be retained in MOS text messages with HR implementation, it is worth noting HR versus MOS PoP skill for the same MOS stations that were involved in the analogous comparative scoring of categorical QPFs (section 3). Fig. 11 shows station-based HR and MOS 6- and 12-h PoP BSSs for the cool season (the relative skill levels for these PoP sources were little different for the warm season). Here we find that the MOS BSSs are again lower than those

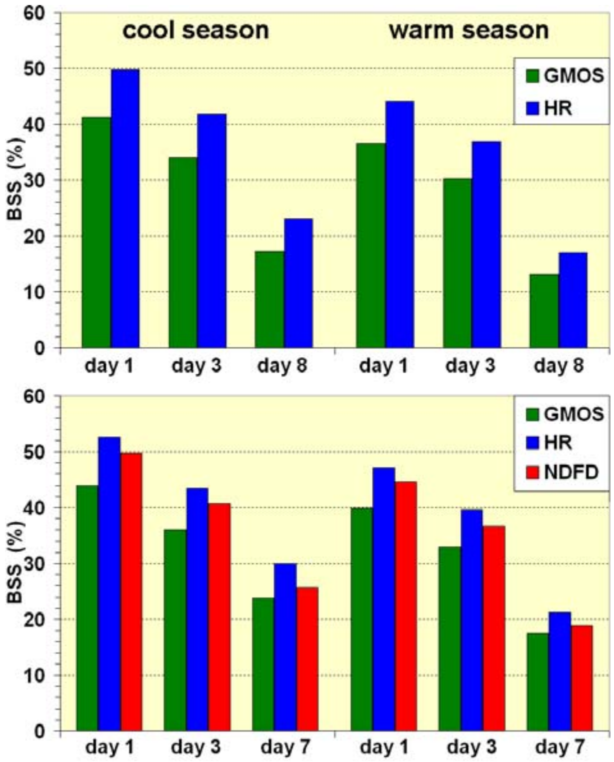

Figure 10. BSS for gridded 6-h PoPs (top) and 12-h PoPs (bottom). Otherwise as in previous figures.

for HR, but the skill differences are smaller than for the corresponding GMOS skill comparison (Fig. 10). Thus, the finding of improved forecast performance of MOS PoPs over GMOS PoPs is consistent with that noted in section 3 for the corresponding categorical QPF scoring.

6. FINDINGS AND CONCLUSIONS

Principal findings from this comparative verification study are:

(1) HR 6-h categorical QPF grids scored slightly better than human-prepared grids from HPC and NDFD and strongly better than model-produced grids (Figs. 2 - 4);

(2) HR 6- and 12-h categorical QPF grids scored clearly better GMOS QPF grids (Figs. 2 - 4); the improvement on MOS QPFs at MOS stations was smaller (Figs. 5 - 7);

(3) HR 6- and 12-h PoP forecast grids were more skillful than GMOS and NDFD grids (Fig. 10); again, the improvement on MOS PoPs at MOS stations was smaller (Fig. 11);

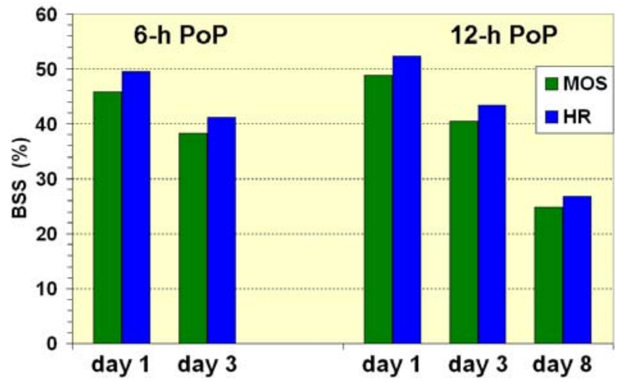

Figure 11. BSS for 6- and 12-h PoP forecasts at 1647 MOS stations during the cool season. Otherwise as in previous figures.

(4) The skill of HR 6-h PQPF grids was about the same as that for similar PQPF grids from an experimental HPC PQPF model (Fig. 8). However, PQPFs from these models have contrasting reliability and sharpness properties (Fig. 9).

Thus, we conclude that the HR model (when implemented) should provide improved gridded 6- and 12-h QPF guidance over the CONUS. Also, objectively combining the comparably-skilled 6-h PQPF grids from the HR and HPC models would probably yield a superior PQPF product. Finally, the new gridded 6- and 12-h PQPFs offer users skillful QPF uncertainty information, which is presently not available operationally.

7. ACKLOWLEDGEMENTS

Letitia Soulliard of NPVU provided archives of the stage IV precipitation analyses and the HPC and NDFD 6-h QPFs, Keith Brill of HPC provided an archive of the experimental HPC 6-h PQPFs, Valery Dagostaro of MDL furnished an archive of the NDFD 12-h PoPs, and staff of the MDL Statistical Modeling Branch led by Kathryn Gilbert furnished archives of the NCEP model, MOS, and GMOS QPF elements.

8. REFERENCES

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecast. *J. of Hydrol.*, **239**, 306-337.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Brill, K. F., 2010: Using the HPC QPF and ensemble QPF statistics to compute probabilistic QPF. NCEP Technical Document, (Available from National Weather Service/NOAA, NCEP Hydrometeorological Prediction Center, 5200 Auth Road, Suitland, Maryland 20746.)

Charba, J. P., and F. G. Samplatsky, 2009: Hi-res gridded MOS 6-h QPF guidance. Preprints, *23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., **17B.2**.

___, and ___, 2011a: Regionalization in fine grid GFS MOS 6-h quantitative precipitation forecasts. *Mon. Wea. Rev.*, **139**, 24-38.

___, and ___, 2011b: High resolution GFS-based MOS quantitative precipitation forecasts on a 4-km grid. *Mon. Wea. Rev.*, **139**, 39-68.

Glahn, B., and D. Ruth, 2003: The new digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.,* **84**, 195-201.

___, K. Gilbert, R. Cosgrove, and K. Sheets, 2009b: The gridding of MOS. *Wea. Forecasting*, **24**, 520-529.

Kanamitsu, M., J. C. Alpert, K. A. Campana, P. M. Caplan, D. G. Deaven, M. Iredell, B. Katz, H. –L. Pan, J. Sela, G. H. White, 1991: Recent changes implemented into the Global Forecast System at NMC. *Wea. Forecasting*, **6**, 425-435.

National Weather Service, 1974: Postprocessing the LFM forecasts. Technical Procedures Bulletin No. 174, National Oceanic and Atmospheric Administration, U. S. Department of Commerce, 20 pp.

Rodgers, E., Y. Lin, K. Mitchell, W.-S Wu, B. Ferrier, G. Gayno, M. Pondeca, M. Pyle, V. Wong, and M. Ek, 2005: The NCEP North American Mesoscale Modelling System: Final Eta model/analysis changes and preliminary experiments using the WRF-NMM. Preprints, *21st Conf. on Wea. Analysis and Forecasting*, Washington, D.C., Amer. Meteor. Soc., **4B.5**.

Schaefer, T. J., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 627 pp.