HUMAN IMPROVEMENT TO NUMERICAL WEATHER PREDICTION AT THE HYDROMETEOROLOGICAL PREDICTION CENTER

David R. Novak, Chris Bailey, Keith Brill, Michael Eckert, Dan Petersen, Robert Rausch, and Michael Schichtel
NOAA/NWS Hydrometeorological Prediction Center, Camp Springs, MD

1. INTRODUCTION

As the skill of numerical weather prediction (NWP) and associated post-processed guidance continues to improve, recent debate asks to what degree can expert human forecasters add value to NWP (e.g., Mass 2003; Bosart 2003; Stuart et al. 2006; Novak et al. 2008; Glahn et al. 2009). The Hydrometeorological Prediction Center (HPC) has a broad mission to serve as a center of excellence in quantitative precipitation forecasting (QPF), medium range forecasting, winter weather forecasting, surface analysis, and the interpretation of operational NWP. Historically, forecasters at the HPC have had access to a large portion of the available model guidance suite, recently including multi-model ensemble information from international partners. HPC's unique national forecast mission coupled with its access to state of the art model guidance provides a rare opportunity to assess human improvement to NWP in an environment of ever-improving guidance. This work will examine human improvement to NWP at HPC using multi-year verification. Although it is recognized humans can add substantial value to NWP through interpretation and decision support, emphasis is placed on the human role in improving forecast accuracy in this work. The HPC QPF and medium range programs are examined.

2. QPF

HPC issues deterministic 6-h QPFs out to 3.5 days in 6-h increments, and a 48-h QPF spanning days 4–5. The HPC deterministic QPF is considered an area average most likely value, and is created on a 32-km grid. Forecasters assess observations of moisture, lift, and instability and make comparisons among model forecasts of these parameters. Subjective blends of model guidance are used for the day 2 and 3 forecasts, while a combination of nowcasting based on observations and short-range forecasts is used for the first 24 h of the forecast. Forecasters generally manually draw precipitation isohyets, with NWP and observations serving as a background.

**Corresponding author address:* David Novak, NOAA/NWS Hydrometeorological Prediction Center, 5200 Auth Rd., Rm. 410, Camp Springs, MD 20746. Email: david.novak@noaa.gov

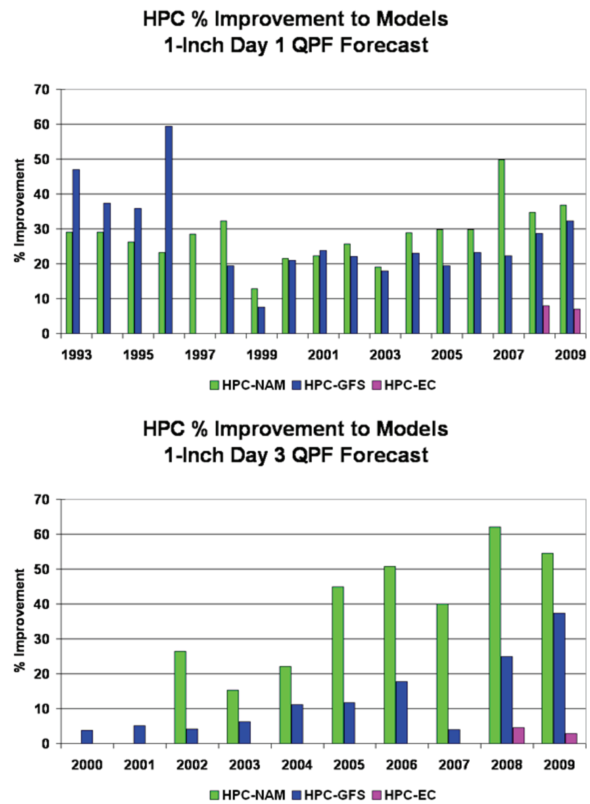

Fig. 1.HPC QPF percent improvement over the NAM (green), GFS (blue), and ECMWF (purple) for the (a) Day 1 and (b) Day 3 24-h accumulated precipitation Threat Score for the 1-in 24 h^{-1} threshold. The HPC forecast is available ~4 h after the model guidance.

Verification over the past two decades shows HPC human forecasters contribute a consistent 20–40% improvement over the NAM and GFS for the 1-in 24 h^{-1} threshold for the day 1 forecast (Fig. 1a). This value-added contribution is occurring during a period of rapid advances in NWP skill. For example, the GFS 1-in 24 h^{-1} day 1 threat score in 1993 was 0.14, while it was 0.24 in 2010. Based on the current rate of model improvement, it would take ~13 years until the GFS attains a Day 1 threat score equivalent to the current HPC threat score.

Mass (2003) and McCarthy et al. (2007) have asserted that the human is most effective for the near term forecast. However, the HPC percent improvement for the 1-in 24 h^{-1} threshold for the day 3 forecast (Fig. 1b) exhibits similar (if not larger) percent improvement as the day 1 forecast. For both the day 1 and day 3 forecast, the competitive skill of the ECWMF forecast is evident, for which the human adds small, but positive skill.

Mass (2003), Bosart (2003), Stuart et al. (2006), and McCarthy et al. (2007) have suggested that the human forecaster may be most adept at improving over NWP guidance for high-impact events. The threat score for 3-in 24 h^{-1} threshold is arbitrarily used as a proxy for a high-impact event. For this threshold for the day 1 forecast, the HPC threat score exhibits a large improvement over select NWP (Fig. 2a). This suggests the human plays an important role in improving near-term model forecasts for heavy precipitation events. However, at the day 3 forecast projection, the scores are more similar (Fig. 2b). In fact, the GFS was superior to the HPC forecast in 2001 and 2003, and the ECMWF was superior to the HPC forecast in 2009. Thus, the human improvement for extreme rainfall events may be dependent on forecast projection. In any case, the skill of model and human forecasts alike at the 3-in. threshold is rather poor when compared to the 1-in. threshold, illustrating the challenge of forecasting extreme rainfall events (Novak et al. 2011).

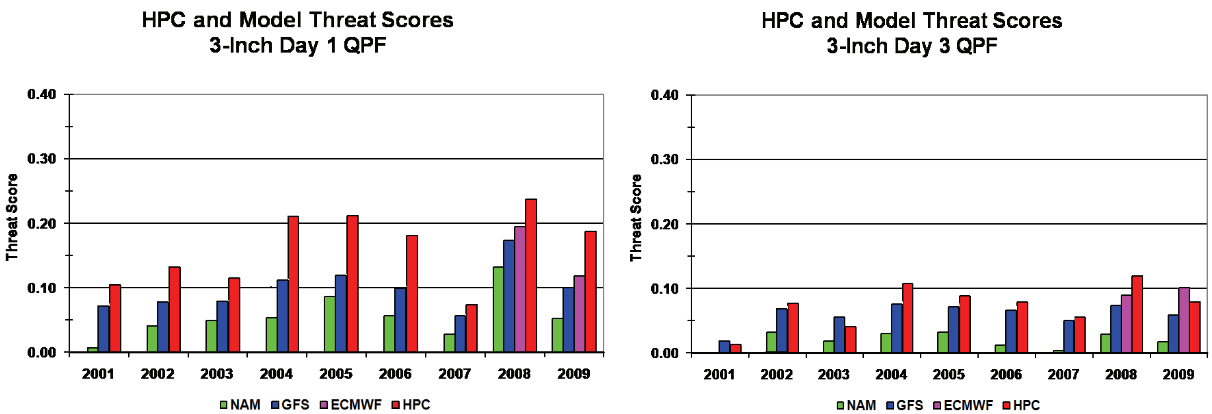

Fig. 2 Comparison of the threat score for the 3-in 24 h^{-1} threshold for a (a) day 1 and (b) day 3 forecast over the past 9 years.

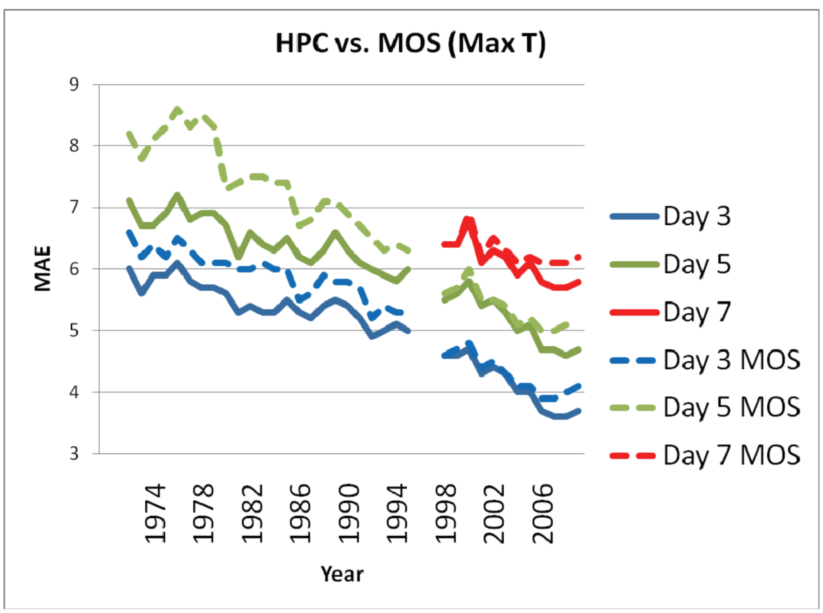

Fig.3. Time series comparison of the HPC and MOS maximum temperature forecast MAE. Data is missing between 1997 and 1998.

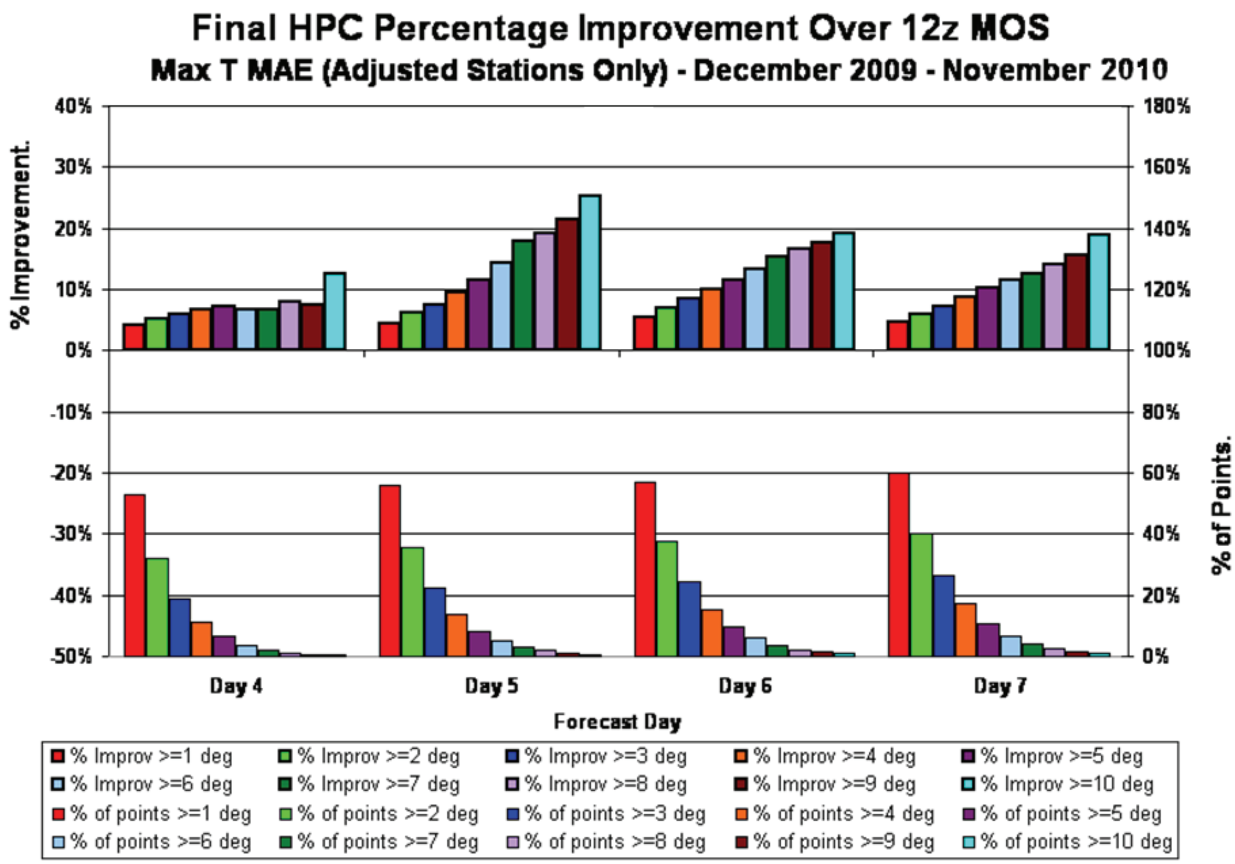

Fig. 4. (top) HPC percent improvement over the global model (GFS) MOS at stations that were adjusted from MOS over the 12-month period ending November 2010. Percent improvement for changes from ≥1–8° F are displayed for day 4 to 7 forecasts. (bottom) Corresponding percentage of points adjusted.

3. MEDIUM RANGE

HPC forecasters produce a 3–7 day forecast product suite including forecasts of sensible weather elements and associated discussion of forecast factors and confidence. Two forecasters work in tandem to complete this task and coordinate with users after assessment of observations and NWP. Forecasters apply weights to individual models and ensemble systems to derive a most likely solution.

Historical verification of maximum temperature among 93 points across the nation show the marked improvements in medium range skill. Today's 7-day maximum temperature forecast is as accurate as a 5-day forecast in the mid 1990s (Fig. 3). The human forecaster improves upon GFS MOS (Fig. 3). Using a linear trend over the past decade, it would take 5 additional years for GFS MOS to improve to the accuracy of human maximum temperature forecasts today. Before 1998 HPC forecasters were verified relative to a version of MOS termed "Kleins" (Klein and Glahn 1974). Starting in 1998, HPC forecasters were verified relative to modern MOS (Glahn et al. 2009). Differences between the Kleins and MOS are apparent, with HPC forecasters improving more against Kleins. The long term (30-year) trend shows the human is improving less over the NWP. However, within the last five years, the HPC forecasts are improving over MOS on the order of 10%. This contrary trend is intriguing, and may be the result of degraded MOS, the implementation of model blending (weighting), improved forecaster knowledge, and/or the incorporation of skillful international guidance, such as the ECMWF deterministic and ECMWF ensemble. Whatever the reason, at least for this metric the human is adding value even for the medium range forecast. This is counter to assertions that humans are most adept at improving NWP in the short range (e.g., Mass 2003; McCarthy et al. 2007).

One hypothesis explaining the improvement is that the human forecaster is adept at recognizing when MOS is in large error, and thus making large changes from MOS. Fig. 4 largely validates this hypothesis, showing that for frequent small changes the human forecaster makes small improvements (~5%). However, for rare large deviations from MOS (i.e., >8°F), forecasters are making changes in the correct direction on average, exhibiting average percent improvements >20%.

4. METHODS OF IMPROVEMENT

The value added by forecasters is a subjective process in many cases (Stuart et al. 2007). An informal survey of HPC forecasters regarding their forecast process revealed several factors, which are listed below. This feedback is in the context of providing a most likely single-valued forecast, although many factors may apply to hedging probabilitistic forecasts as well.

Pattern recognition: Recognizing environments supportive of particular events. This skill is often gained via education and experience, and is analogous to applying analogs.

Physical realism: Identifying non-physical evolutions, helping the forecaster to either reduce the weight of that particular solution or discard the solution altogether. For example, teleconnections are used by medium range forecasters.

Known biases: Incorporating **s**ubjective bias information either learned through objective verification or experience.

Past forecast system performance: Weighting of solutions based on long-term objective verification (Forecast system A is more skillful than forecast system B).

Relation to other solutions: Considering each particular solution in the context of all other available solutions. Is the forecast within the ensemble envelope? Is the forecast close to other skillful members?

Run-to-run consistency: Weighting solutions with run-to-run consistency higher than others. Solutions trending towards the ensemble mean enhance confidence in the ensemble mean solution.

Collaboration: Considering of alternative perspectives and data.

Consensus: Following the most popular opinion, and/or averaging solutions.

These key processes appear to be useful, even in an age of radical NWP improvements.

5. SUMMARY

Analysis of multi-year QPF and medium range verification of human-generated forecasts at HPC compared to automated NWP is presented. Results show that the human-generated HPC QPFs improve upon NWP, and that the percent improvement has been relatively constant over the past two decades (e.g. Fig. 1a). Medium range maximum temperature forecasts also exhibit improvement over MOS, which has been increasing over the past five years. Human forecasts of high-impact events (3-in 24 h^{-1} and changes to MOS temperatures ≥8°F) generally exhibit large improvements over NWP. These improvements are accomplished through pattern recognition, physical realism, awareness of model biases and past model performance, run-to-run consistency, collaboration, and consensus.

As computer resources advance, models will explicitly simulate more processes, and more and better observations will be used by improved data assimilation systems. These advances will lead to improved NWP guidance. Additionally, post-processing of native model guidance, including bias-correction, will improve automated forecasts of sensible weather elements. Given this environment, it may be difficult to envision the human forecaster continuing to add value.

However, as the history of NWP and the human forecaster role continues to be written, the overall evidence presented above suggests that active and engaged forecasters can continue to make incremental improvement to NWP despite radical NWP improvements. Roebber et al. (2003) cite the human ability to interpret and evaluate information as an inherit advantage over automated processes. The lead author's recent experience during the QPF component of the NOAA Hazardous Weather Testbed Spring Experiment (Barthold et al. 2011) suggests that even with the next generation of guidance (for example, a 26 member 4-km grid spacing radar assimilating storm scale ensemble), the human can recognize opportunities to improve upon the experimental guidance for QPF.

Whether there is a future point beyond which human improvements begin to asymptote to near zero is unknown. It is also recognized that the incremental improvement the human provides comes at a financial cost and that the human-generated information is generally available later than automated NWP. This cost/benefit was not considered in the above analysis. Also, statistical significance of differences between human and model forecasts was not assessed here. Additionally, only deterministic short-range QPF and medium range maximum temperature forecasts were considered. Thus, a more complete investigation of the human's role in improving upon NWP using other metrics, elements, time ranges, and formats (probabilistic) is encouraged, and may lead to new paradigms for human involvement in the forecast process.

6. ACKNOWLEDGMENTS

The views expressed are those of the authors and do not necessarily represent a NOAA/NWS position.

7. REFERENCES

Barthold, F. E., and Coauthors, 2011: The quantitative precipitation forecasting component of the 2010 NOAA hazardous weather testbed spring experiment. Preprints, *24th Conference. on Weather* *and Forecasting/ 20th Conference on NWP*, Amer. Meteor. Soc., Seattle, WA, 9A.4.

Bosart, L. F., 2003: Whither the weather analysis and forecasting process? *Wea. Forecasting*, **18**, 520–529.

Mass, C. F., 2003: IFPS and the future of the national weather service. *Wea. Forecasting*, **18**, 75–79.

Klein, W. H., H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217–1227.

McCarthy, P. J., D. Ball, and W. Purcell, 2007: Project Phoenix: Optimizing the machine–person mix in high-impact weather forecasting. Preprints, *22nd Conference on Weather Analysis and Forecasting*, Amer. Meteor. Soc. Park City, UT.

Novak, D. R., D. R. Bright, M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084.

——, and Coauthors, 2011: Quantifying extreme rainfall threats at the Hydrometeorological Prediction Center. Preprints, *24th Conference. on Weather and Forecasting/ 20th Conference on NWP*, Amer. Meteor. Soc., Seattle, WA, 14A.4.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2003: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949.

Ruth, D. P., B. Glahn, V. Dagostaro, K. Gilbert, 2009: The performance of MOS in the digital age. *Wea. Forecasting*, **24**, 504–519.

Stuart, N.A., and Coauthors, 2006: The future of humans in an increasingly automated forecast process. *Bull. Amer. Meteor. Soc.*, **87**, 1497–1501.

——, D. M. Schultz, G. Klein, 2007: Maintaining the role of humans in the forecast process: Analyzing the psyche of expert forecasters. *Bull. Amer. Meteor. Soc.*, **88**, 1893–1898.