The DOI for this manuscript is doi: 10.1175/WAF-D-13-00066.1

Precipitation and Temperature Forecast Performance at the Weather Prediction Center

David R. Novak, Christopher Bailey, Keith Brill, Patrick Burke, Wallace Hogsett, Robert Rausch, and Michael Schichtel

[@]NOAA/NWS/NCEP, Weather Prediction Center,
College Park, Maryland

Submitted to *Weather and Forecasting*

June 10, 2013

Revised September 21, 2013
and
October 9, 2013

Corresponding author address: David R. Novak, NOAA/NWS Weather Prediction Center, 5830 University Research Court, Rm 4633, College Park, MD 20740.
E-mail: David.Novak@noaa.gov

ABSTRACT

The role of the human forecaster in improving upon the accuracy of numerical weather prediction is explored using multi-year verification of human-generated short-range precipitation forecasts and medium-range maximum temperature forecasts from the Weather Prediction Center (WPC). Results show that human-generated forecasts improve over raw deterministic model guidance. Over the past two decades, WPC human forecasters achieved a 20–40% improvement over the NAM and GFS models for the 1-in (25.4-mm) 24 h^{-1} threshold for the day 1 precipitation forecast, with a smaller, but statistically significant, 5–15% improvement over the deterministic ECMWF model. Medium-range maximum temperature forecasts also exhibit statistically significant improvement over GFS Model Output Statistics (MOS), and the improvement has been increasing over the past five years. The quality added by humans for forecasts of high-impact events varies by element and forecast projection, with generally large improvements when the forecaster makes changes ≥8°F (4.4°C) to MOS temperatures. Human improvement over guidance for extreme rainfall events [3-in (76.2-mm) 24 h^{-1}] is largest in the short-range forecast.

However, human-generated forecasts failed to outperform the most skillful downscaled, bias-corrected ensemble guidance for precipitation and maximum temperature available near the same time as the human-modified forecasts. Thus, as additional downscaled and bias-corrected sensible weather element guidance becomes operationally available, and with the support of near-real time verification, forecaster training, and tools to guide forecaster interventions, a key test is whether forecasters can learn to make statistically significant improvements over the most skillful of this guidance. Such a test can inform to what degree, and just how quickly the role of the forecaster changes.

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 human forecasters add quality[1] to NWP (e.g., Mass 2003; Bosart 2003; Roebber et al. 2004; Reynolds 2003; Doswell 2004; Stuart et al. 2006; Homar et al. 2006; Novak et al. 2008; Ruth et al. 2009). The National Centers for Environmental Prediction (NCEP) Weather Prediction Center (WPC[2]) 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 WPC have had access to a large portion of the available model guidance suite, recently including multi-model ensemble information from international partners. The WPC's unique national forecast mission coupled with its access to state-of-the-art model guidance provides a rare opportunity to assess the quality added by humans to ever-improving NWP guidance.

This work will examine multi-year historical and contemporary verification for short-range deterministic precipitation forecasts and medium-range maximum temperature forecasts generated at the WPC. Although humans can add substantial value to NWP through retaining forecast continuity (run-to-run consistency), assuring element consistency (e.g., wind shifts with fronts), and helping users make informed decisions (e.g., Roebber et al. 2010a), this work focuses on the human role in improving forecast accuracy. In this respect, the current work examines only one component of the forecaster's role, and is limited to analysis of just two weather elements.

---

[1] Although "value-added" is often used colloquially, this work abides by the terms for forecast "goodness" defined in Table 1 of Murphy (1993), where "value" refers to the benefit realized by decision makers through the use of the forecasts and "quality" refers to the correspondence between forecasts and the matching observations.

[2] The Center's name was changed from the Hydrometeorological Prediction Center to the Weather Prediction Center on March 5, 2013.

The current work builds upon and extends previous analyses of WPC skill by Olson et al. (1995), Reynolds (2003), and Sukovich et al. (2013), and points to future verification approaches in the continuing history of NWP and the human forecaster. Section 2 presents analysis of QPF while section 3 explores human improvement to medium range maximum temperature forecasts. A discussion of the limitations of the work and implications of the verification for the future role of the forecaster is presented in section 4.

2. QPF

a. Production and verification method

The WPC forecasters create deterministic QPFs at 6-h intervals through the day 3 forecast projection, and 48-h QPFs for days 4–5 and 6–7. The focus here is on the day 1–3 forecasts. The WPC deterministic QPF during the study period was defined as the most-likely, areal-averaged amount mapped to a 32-km horizontal resolution grid. An example 24-h accumulated QPF is shown in Fig. 1a. The forecast process for QPF involves forecaster assessment of observations of moisture, lift, and instability and comparisons among deterministic and ensemble forecasts of these parameters. Objectively post-processed model-based QPFs are also available to WPC forecasters. Emphasis shifts from nowcasting based on observations in the first 6–12 h of the forecast, to an increasing use of NWP as lead time increases. For example, subjective blends of model guidance are used almost exclusively beyond 36-h. Forecasters manually draw precipitation isohyets, which operational software converts to a grid. In areas of complex topography, forecasters use monthly Parameter-elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 1994; Daly et al. 2008) output as a background.

The 24-h accumulated QPF was verified using a human quality-controlled (QC'ed) analysis valid at 12 UTC. The analyst can choose a first-guess field from either the multisensor Stage IV quantitative precipitation estimate mosaic analyses (Lin and Mitchell 2005) or Climate Prediction Center (CPC) daily precipitation analysis (Higgins et al. 1996). The analyst QCs the analysis based on gauge data and a review of radar data, and can adjust isohyets if necessary. The QC'ed precipitation analysis is mapped onto a 32-km grid, matching the forecast grid. Retrospective tests show that the relative skill difference between the WPC and NWP datasets shown in this paper are not sensitive to the precipitation analysis used (e.g., the WPC QC'ed analysis or Stage IV).

Conventional 2x2 contingency tables of dichotomous outcomes (e.g., Brill 2009) for precipitation exceeding several thresholds are created by comparing each QPF to the corresponding verifying analysis. The 2x2 contingency tables are used to calculate threat score and frequency bias for the day 1 and day 3 forecast periods. The WPC forecast period naming convention is shown in Fig. 2. Focus is placed on the threat score for the day 1 QPF valid at 12 UTC at the 1-in (25.4-mm) 24 h^{-1} threshold. This threat score is reported to Congress as part of the Government Performance and Results Act of 1993 (GPRA). Historically, the goal of the WPC QPF was to improve upon the model guidance available during the interval of forecast preparation (Reynolds 2003). Therefore the performance of the WPC QPF is compared against model forecasts that are somewhat older (i.e., time lagged) than the WPC issuance time. The latency of the WPC forecasts for the most frequently used model guidance are shown in Table 1.

The historical verification analysis was constrained to data available during the last ~50 years, which were largely deterministic forecasts. Bias-corrected forecasts were also generally not available for verification purposes during the historical timeframe. Bias correction can dramatically improve raw QPF guidance (e.g., Yussouf and Sensrud 2006; Brown and Seo 2010), and ensemble approaches can quantify predictability and reduce error. Thus, the contemporary verification compares the WPC QPF to one created by an ensemble algorithm with bias correction, issued near the time as the human-modified forecast. This product, the pseudo bias corrected ensemble QPF (ENSBC), is based on the premise that the larger the uncertainty, the smoother the forecast should be, whereas the smaller the uncertainty the more detailed the forecast should be. During the study period the ENSBC was composed of a high-resolution ensemble part comprising output from the deterministic NCEP North American Mesoscale model (NAM; Janjic 2003), Global Forecast System (GFS; Caplan et al. 1997), and European Center for Medium-Range Weather Forecasts (ECMWF; Magnusson and Kallen 2013), and a full ensemble part composed of the high-resolution ensemble plus the Canadian GEM (Belair et al. 2009), UK Met Office model (UKMET), and all members of the NCEP Short Range Ensemble Forecast (SREF; Du et al. 2006). The product is objectively downscaled from 32 km to 10 km (5 km over the west) using PRISM. A detailed description of the ENSBC algorithm is provided in Appendix A.

Additionally, the historical analysis is limited to verification metrics with a long record to facilitate historical context [threat score, frequency bias (referred to as bias hereafter), and mean absolute error (MAE)]. Metrics such as the threat score have inherent limitations, including a double penalty for false alarms (Baldwin et al. 2002) and bias sensitivity (Brill 2009; Brill and Mesinger 2009). To address this issue, a bias-removed threat score is calculated using the procedure based on probability matching (Ebert 2001) described by Clark et al. (2009). The procedure uses probability matching to reassign the distribution of a forecast field with that of the observed field, so that the modified forecast field has the same spatial patterns as the original forecast, but has values adjusted so the distribution of their amplitudes exactly matches those of the analysis. The end result is the removal of all bias. Because the NAM precipitation skill lags so severely relative to WPC and other international guidance, and to simplify interpretation, the bias-removed threat score calculation was not conducted for the NAM. This reduced skill is likely due to the use of 6-h old boundary conditions from the GFS, an earlier data cutoff, as well as a less advanced data assimilation system (G. DiMego and E. Rogers, personal communication).

Finally, it is important to quantify the statistical significance of comparisons. To accomplish this task the forecast verification system (fvs) software was used (described in Appendix B). Assessment of statistical significance in fvs is accomplished using random resampling following the method of Hamill (1999).

Thus, contemporary verification addressing these four issues (ensemble approaches, bias-corrected guidance, bias sensitivity, and statistical significance) was conducted during the latest years available (2011–12).

b. Results

Verification of the WPC QPF over the last 50 years (Fig. 3) is a testament to the advancement of precipitation forecasts. Threat scores of the 1-in (25.4-mm) 24 h^{-1} threshold for day 1 forecasts doubled during the period, while day 2 and 3 forecasts also continued to improve (Fig. 3). Improvement has accelerated after 1995. This improvement is directly tied to the quality of the NWP guidance. In fact, during the 1993–2012 period, the correlation of yearly values of the day 1 threat score for the 1-in (25.4-mm) 24 h^{-1} threshold between WPC and the NAM and WPC and the GFS was 0.91 and 0.88, respectively.

Although NWP serves as skillful guidance, verification over the past two decades shows WPC human forecasters achieved a 20–40% improvement over the deterministic NAM and GFS for the threat score of the 1-in (25.4-mm) 24 h^{-1} threshold for the day 1 forecast (Fig. 4a). This improvement was occurring during a period of advances in NWP skill. For example, the GFS 1-in (25.4-mm) 24 h^{-1} day 1 threat score in 1993 was 0.14, whereas in 2012 it was 0.25. Based on the long term rate of model improvement, it would take ~13 years until the GFS attains a day 1 threat score equivalent to the current WPC threat score. This rate is nearly identical to the 14 years reported by Reynolds (2003) for the 2001 verification year.

The ECMWF precipitation forecast information became available to WPC forecasters in the mid 2000s, and the first full year of formal verification was established in 2008. Verification of the 1-in (25.4-mm) 24 h^{-1} day 1 forecast over the 2008–2012 period shows that the WPC forecast exhibits smaller 5–15% improvements over the very skillful deterministic ECMWF model (Fig. 4a). However, WPC improvement over the ECMWF model has nearly doubled over the past 5 years.

A complete picture of precipitation verification must include bias information. In recent years the NAM, GFS, and ECWMF guidance have exhibited a low bias at the 1-in (25.4-mm) 24 h^{-1} threshold, while the WPC has sustained a more favorable bias near 1.0 (Fig. 4a). Contemporary verification using the bias-removed threat score shows that WPC has maintained a statistically significant advantage over the ECMWF and GFS during 2011 and 2012 (Fig. 4b). However, the ensemble-based post-processed QPF from the ENSBC was very competitive. In fact, the ENSBC and WPC forecasts were statistically similar for the 1-in (25.4-mm) 24 h^{-1} threshold during 2012 (Fig. 4b).

Mass (2003) and McCarthy et al. (2007) have asserted that the human is most effective for the near-term forecast. However, the WPC percent improvement over the GFS at the 1-in (25.4-mm) 24 h^{-1} threshold for the day 3 forecast is similar to the percent improvement for this threshold for the day 1 forecast (c.f. Figs. 4b and d). All guidance, including WPC, has a slight low bias at the day 3 forecast (Figs. 4c,d). For both the day 1 and day 3 forecasts, the competitive skill of the ECWMF forecast is evident, for which the human adds small, but statistically significant positive skill. However, once again, the WPC is statistically similar to the ENSBC at the 1-in (25.4-mm) 24 h^{-1} threshold for the day 3 period during 2012 (Fig. 4d). Thus, at least for precipitation at this threshold, the quality added by the forecaster does not appear dependent on forecast projection.

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 the 3-in (76.2 mm) 24 h^{-1} threshold is arbitrarily used here as a proxy for a high-impact event. The skill of both model and human forecasts at the 3-in (76.2 mm) 24 h^{-1} threshold is rather poor when compared to the 1-in threshold, illustrating the challenge of forecasting extreme rainfall events (Fritsch and Carbone 2004; Sukovich et al. 2013). However, the day 1 WPC threat score exhibits a large improvement over select NWP (Fig. 5a), with a slight dry bias. Contemporary verification accounting for bias shows WPC significantly improved over the GFS in 2012 and ECMWF in both 2011 and 2012 at this threshold. However, once again, WPC was similar in skill to the ENSBC product (Fig. 5b).

Skill comparisons for the 3-in (76.2 mm) 24 h^{-1} threshold at the day 3 lead time reveals generally less forecaster improvement, with similar model and WPC threat scores (Fig. 5c). In fact, the GFS was superior to the WPC forecast in 2001 and 2003, and the ECMWF was superior to the WPC forecast in 2009. All guidance, except the GFS, is severely under-biased. The authors speculate that the GFS had frequent grid-point storms (e.g., Giorgi 1991) during the verification period, which may have improved its bias, but degraded its threat score. Contemporary verification shows the WPC bias-removed threat score is not statistically significantly different than the corresponding threat scores from any of the competitive guidance (Fig. 5d).

All of the above results suggest humans can make statistically significant improvement over competitive deterministic model guidance for precipitation. The magnitude of quality-added by the forecaster is generally not dependent on forecast projection for the 1-in (25.4-mm) 24 h^{-1} threshold; however, human improvement for extreme rainfall events does appear dependent on forecast projection, favoring larger human improvements over deterministic model guidance in the short-range forecast. However, a downscaled, bias-corrected ensemble forecast available near the same time as the human-modified forecast exhibits similar skill – even for extreme precipitation events.

3. Maximum temperature

a. Production and verification method

WPC forecasters produce a 3–7 day forecast suite including gridded predictions of sensible weather elements to support the National Digital Forecast Database (NDFD; Glahn and Ruth 2003) (Fig. 1b), graphical depictions of the surface fronts and pressures (Fig. 1c), and associated discussion of forecast factors and confidence. Two forecasters work in tandem to complete this task and coordinate with users after assessment of NWP. Since 2004, forecasters have used a graphical interface to apply weights to individual models and ensemble systems to derive a most-likely sensible weather solution. The result of the forecaster's chosen blend can be manually edited.

Before model data is weighted by the forecaster, the data is bias-corrected and downscaled to a 5-km horizontal resolution. Bias correction of gridded model data is accomplished using the NCEP decaying averaging bias-correction method of Cui et al. (2012), applied as:

$$B_{new} = (1 - w)B_{past} + wB_{current} , \tag{1}$$

where, $B_{current}$ is the latest calculated forecast error given by the difference between the forecast and verifying analysis, B_{past} is the past accumulated bias, and B_{new} is the updated accumulated bias. The NCEP 5-km resolution Real Time Mesoscale Analysis (RTMA; De Pondeca et al. 2011) was used as the verifying analysis. The weight factor, w, controls how much influence to give the most recent bias behavior of weather systems. A w equal to 2% was used operationally. Once initialized, the bias estimate can be updated by considering just the current forecast error ($B_{current}$) and the stored average bias (B_{past}). The new bias-corrected forecast is generated by subtracting B_{new} from the current forecasts at each lead time and each grid point.

Downscaling of coarse model data to a 5-km resolution grid is accomplished using a decaying averaged downscaling increment (Cui et al. 2013). The downscaling increments are created at each 6-h time step by differencing the coarse 1° resolution GFS analysis (GDAS) and 5-km resolution RTMA according to:

$$D_{new} = (1 - w)D_{past} + wD_{current} , \tag{2}$$

where, $D_{current}$ is the latest calculated downscaling increment given by the difference between GDAS and RTMA, D_{past} is the past accumulated downscale increment, and D_{new} is the updated downscale increment. The weight factor, *w*, controls how much influence to give the most recent difference. A *w* equal to 10% was used operationally. The 6-hour grids are then downscaled using the mean downscaling increment for each 6-hour period. For maximum and minimum temperature, at each grid point, the downscaled 6-hour grids are compared to each other to find the highest (lowest) values for maximum (minimum) temperature over the 12–06 UTC period (00–18 UTC period) to get a final maximum (minimum) temperature forecast grid. The verifying maximum (minimum) temperature is taken as the highest (lowest) hourly value from the RTMA at each grid point.

The resulting maximum and minimum temperatures are extracted from the 5-km grid to 448 points for the forecaster to edit where necessary. An objective analysis is performed on the incremental changes made by the forecaster at the 448 points to create a difference grid. The forecaster-edited difference grids are added to the forecaster-weighted output grids to get a final adjusted 5-km forecast grid. Complete details of the methodology for all elements are documented at: http://www.wpc.ncep.noaa.gov/5km_grids/medr_5km_methodology_newparms.pdf

Both point and gridded verification are conducted. Points are verified by the respective observed station information, while the RTMA is used to verify gridded fields. The fvs (described in Appendix B) is used to calculate both point-based and gridded verification of sensible weather elements, including determination of statistical significance.

b. Results

Historical verification of maximum temperature at 93 points across the nation shows the marked improvements in medium range temperature forecast skill over time. Today's 7-day maximum temperature forecast is as accurate as a 3-day forecast in the late 1980s (Fig. 6). Comparison of the 00 UTC GFS MOS forecast to the 20 UTC "final" daily issuance of the WPC forecast shows the human forecaster improves upon GFS MOS (Fig. 6). Before 1998 WPC forecasters were verified relative to a version of MOS termed "Kleins" (Klein and Glahn 1974). Starting in 1998, WPC forecasters were verified relative to modern MOS (Glahn et al. 2009), and MOS was used as the starting point for their forecasts. Differences between the Kleins and MOS are apparent, with WPC forecasters improving more against Kleins (Fig. 6). The long term (30-year) trend shows the human is improving less over the NWP. However, within the last seven years, the WPC forecasts are improving over MOS on the order of 5% (Fig. 6). This improvement may be related to a change in forecast methodology in 2004, whereby forecasters use a graphical interface to apply weights to individual models and ensemble systems to derive a most likely sensible weather solution. Further, ECMWF guidance became available reliably to forecasters by 2008.

It is necessary to account for the 13-h latency between the WPC final forecast issuance (19 UTC) and 00 UTC GFS MOS (Table 2). WPC issues a preliminary forecast that substantially reduces this latency. Comparison of the preliminary WPC forecast issuance to the 00 and 12 UTC MOS is examined. This analysis also uses the full expanded set of 448 points over the contiguous United States (CONUS). The results are summarized as an aggregate of monthly scores averaged during the 2007–2012 period (60 months) for maximum temperature. WPC accomplishes a 7–9% improvement over 00 UTC MOS with an 8-h latency, and a 4–5% improvement over 12 UTC MOS with a human forecast issued 4 h *prior* to MOS (Fig. 7). Both results are statistically significant at the 90% level for all days. Using a linear trend over the past decade, it would take ~5 additional years for the 12 UTC GFS MOS to improve to the accuracy of earlier-issued human maximum temperature forecasts.

One hypothesis for the improvement over MOS is that the human forecaster is adept at recognizing when MOS is in large error, and thus makes large changes from MOS. Figure 8 shows that for frequent small changes the human forecaster makes small improvements over 12 UTC MOS (~5%). However, for infrequent large deviations from 12 UTC MOS [i.e., >8°F (4.4°C)], forecasters usually make changes in the correct direction, exhibiting average percent improvements near 15%.

Gridded verification allows examination of how the human gridded forecasts compare to downscaled, bias-corrected international model guidance and gridded MOS (GMOS; Glahn et al. 2009). The WPC final forecasts are statistically significantly better than all raw downscaled international model guidance and GMOS (Fig. 9a). However, bias-correction substantially improves the maximum temperature model guidance – so much so that the bias-corrected ECMWF ensemble mean is statistically significantly superior to the WPC gridded forecast for days 5–7 (Fig. 9b).

Given that surface pressure patterns influence temperature and precipitation patterns, further verification of the WPC mean sea-level pressure (PMSL) forecasts for days 3–7 was conducted for 2012. Verification of anomaly correlation of the deterministic ECMWF and GFS, and their respective ensemble system means is shown in Fig. 10. WPC has a higher anomaly correlation score than all guidance at all time ranges; however, WPC is only statistically significantly superior to all these gridded datasets at the day 6 forecast projection. The deterministic ECMWF, which is available near the time of the final WPC forecast issuance, exhibits similar skill to WPC at days 3 and 4. The 00 UTC ECMWF ensemble mean at day 7 is also similar to WPC skill.

4. Discussion and summary

Analysis of multi-year verification of short-range precipitation forecasts and medium-range maximum temperature forecasts from the Weather Prediction Center (WPC) are compared to automated NWP guidance. Results show that human-generated forecasts improve over raw deterministic model guidance when verified using both traditional methods as well as contemporary methods. However, perhaps the more compelling result is that on the basis of a statistical analysis of two recent years, human-generated forecasts failed to outperform the most skillful downscaled, bias-corrected ensemble guidance for precipitation and maximum temperature available near the same time as the human-modified forecasts.

Specifically, historical verification results show that the human-generated WPC QPFs improve upon deterministic raw model guidance, and that the percent improvement has been relatively constant over the past two decades (e.g. Fig. 4a). Medium range maximum temperature forecasts also exhibit improvement over MOS. The improvement has been increasing during the 2005–12 period. The quality added by humans for forecasts of high-impact events varies by element and forecast projection, with generally large improvements when the forecaster makes changes ≥8°F (4.4°C) to MOS temperatures in the medium range forecast. Human improvement for extreme rainfall events [3-in (76.2-mm) 24 h^{-1}] is dependent on forecast projection, favoring larger human improvements in the short-range forecast. Contemporary verification confirms that the human forecaster makes small, but statistically significant improvement over competitive deterministic model guidance for precipitation and maximum temperature.

However, human-generated forecasts failed to outperform the most skillful downscaled, bias-corrected ensemble guidance for precipitation and maximum temperature available near the same time as the human-modified forecasts. Such downscaled, bias-corrected ensemble guidance represents the most skillful operational benchmark. Thus, it is premature to claim superiority by the human forecaster until such forecasts are statistically significantly better than the most skillful guidance. In fact, these results raise the question of whether human-generated forecast superiority has ended.

Indeed, 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, more sophisticated post-processing of raw model guidance, including bias-correction and downscaling, will improve automated forecasts of sensible weather elements. Roebber et al. (2004) cite the human ability to interpret and evaluate information as an inherent advantage over algorithmic automated processes. However, artificial intelligence algorithms continue to strive to simulate such human decisions – for example, developing methods to automate selective consensus of ensemble members (e.g., Etherton 2007), or applying artificial neural network and evolutionary programming approaches that "learn" through time (e.g., Bakhshaii and Stull 2009; Roebber 2010b). Given this future environment, it is difficult to envision the human forecaster adding quality in terms of forecast accuracy.

On the other hand, there is a distinction between long-term statistical verification (the primary the focus of this paper) and critical deviations from skillful guidance in local regions and cases. Contemporary post-processing approaches are best at correcting repeatable, systematic errors, but struggle when the forecast sample size is small for unusual weather scenarios. The forecaster's decision to deviate from skillful automated guidance in these unusual weather scenarios often comes with substantial societal consequences, such as whether a snowstorm will affect a city (Bosart 2003), or whether a killing freeze will occur. Thus, it is especially critical that the forecaster make the very best decision in these scenarios. Figure 8 shows that when forecasters make large changes from MOS, the deviations are generally in the correct direction, providing evidence of skill in recognizing opportunities to deviate from MOS temperatures. Obviously, more evidence of this skill for other variables, benchmarked against more skillful datasets, and filtered to examine only the most critical weather scenarios is needed to more conclusively demonstrate the forecaster's skill at these deviations.

Bosart (2003) contends that as more and more automation occurs, forecasters' skill at recognizing critical opportunities to deviate from guidance may atrophy. Thus, a key component of assuring the forecaster continues to add quality to NWP is keeping the forecaster engaged in the forecast process. Indeed, the WPC forecasters appear to have learned how to improve over the ECWMF precipitation forecasts over the past 5 years (Figs. 4a,c), perhaps learning when to deviate from the skillful guidance. From the authors' experience a key to this improvement is greater emphasis on using the most skillful datasets as the forecaster's starting point, and encouraging changes only when confidence is high. Further, improvement can be gained with greater availability of near-real time verification, using the most skillful guidance as the benchmark. Finally, investment in training forecasters in the strengths and weaknesses of the most skillful guidance, and providing tools to guide forecaster modifications may lead to further forecaster improvements. An example of such a tool is ensemble sensitivity analysis, which can indicate the source of upstream uncertainties for a given forecast parameter. As demonstrated by Zheng et al. 2013, in theory, this tool allows forecasters to identify and monitor the sensitive areas using available observations (satellite, aircraft or other types) in real time to assess the likelihood of future scenarios.

Emphasis on the most skillful downscaled, bias-corrected guidance with supporting near-real-time verification, forecaster training, and tools to guide forecaster interventions has only recently been established at WPC, but has already resulted in forecasters making high-order forecast decisions. These high-order decisions include the removal of outlier forecast guidance that degrades the consensus forecast (e.g. a spurious tropical cyclone), adjusting for regime-dependent biases that are not corrected (or that are introduced) in the post-processing, and perhaps most importantly, deciding when to substantially deviate from the skillful guidance. Thus, as additional downscaled and bias-corrected sensible weather element guidance becomes operationally available, and with the support of near-real time verification, forecaster training, and tools to guide forecaster interventions, a key test is whether forecasters can learn to make statistically significant improvements over the most skillful of this guidance. Such a test can inform to what degree, and just how quickly the role of the forecaster changes.

Given that only one component of the forecaster's role (accuracy) was considered and only deterministic short-range QPF and medium range maximum temperature forecasts were assessed, the above results must not be over generalized. Downscaling and bias-correcting of a full suite of sensible weather elements is not an operational reality yet, as challenges remain with elements such as wind, sky cover, ceiling, and visibility to name a few. Additionally, the contemporary verification was limited to two years. Further the financial cost/benefit of human involvement in the forecast process was not considered in the above analysis. Finally, a critical question facing the forecasting community is if and how a forecaster may add quality to ensemble guidance of many variables (e.g., Roebber et al. 2004, Novak et al. 2008). 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.

Acknowledgements. This work benefited from insightful discussions with Lance Bosart, Brian Colle, Brad Colman, Ed Danaher, Larry Dunn, Jim Hoke, Cliff Mass, Paul Roebber, David Schultz, Neil Stuart, and Jim Steenburgh, as well as stimulating email exchanges on the University of Albany “map” listserve. Mark Klein assisted with gathering necessary data. Jim Hoke and Mike Eckert assisted with a previous version. Two anonymous reviewers provided constructive comments leading to improvements in the presentation of this work. The views expressed are those of the authors and do not necessarily represent a NOAA/NWS position.

APPENDIX A

Description of Pseudo Bias Corrected Ensemble QPF

The pseudo bias corrected ensemble QPF (ENSBC) is a series of 6-h accumulations posted at 6-h intervals. Each 6-h QPF is computed in three phases:

1. Calculate the weighted ensemble mean (WEM).
2. Perform the pseudo bias correction (PBC).
3. Apply downscaling based on data obtained from the PRISM precipitation climatology.

The first phase assumes that the larger the uncertainty, the smoother the forecast should be, whereas the smaller the uncertainty the more detailed the forecast should be. Two ensemble means are computed. The high resolution ensemble mean is the mean of an ensemble made up of relatively high-resolution deterministic single model runs (NAM, GFS, ECMWF). The full ensemble mean is the mean of a high-resolution ensemble consisting of the same deterministic runs along with the GEM and UKMET, and a standard ensemble system (e.g., NCEP Short-Range Ensemble Forecast or NCEP Global Ensemble Forecast System). The maximum QPF from the high-resolution ensemble is added as an additional member. The members of the high resolution ensemble are equally weighted in the warm season, but not in the cold season (October through April), and the weights are adjusted periodically with reference to verification. The members of the full ensemble are equally weighted. The spread of the full ensemble is obtained to compute a normalized spread, $\hat{\sigma}$, which is the full ensemble spread divided by the full ensemble mean, with a small amount added to prevent division by zero. A weight value, *w*, is computed at each grid point:

$$w = \frac{\hat{\sigma}}{\hat{\sigma}_{\max}}, \qquad \text{(A1)}$$

where $\hat{\sigma}_{\max}$ is the domain maximum of the normalized spread. Then the WEM is computed at each grid point:

$$WEM = w\,\mu + (1-w)\,\mu_{hr}\,, \qquad \text{(A2)}$$

where μ is the full ensemble mean and μ_{hr} is the high resolution ensemble mean. Thus, where the forecast uncertainty as measured by the normalized spread is relatively large the WEM is weighted toward the full ensemble mean; whereas, at points with lower normalized spread and less uncertainty, the WEM is weighted toward the high resolution ensemble mean.

In the next phase, the WEM is passed to the PBC, which has nine tuning parameters, is perpetually evolving, and undergoes fairly regular (about every six weeks or so) adjustments based on verification. Here, the PBC is described in general terms.

For WEM 6-h precipitation amounts less than about 6—9 mm the PBC algorithm uses the 10th percentile QPF from the full ensemble to reduce frequency bias (areal coverage). A weighting function, ω, is applied to modify the WEM according to

$$WEM = \omega\, WEM + (1-\omega)\, QPF_{10}\,, \qquad \text{(A3)}$$

where QPF_{10} is the 10th percentile QPF from the multi-model ensemble. The weighting function linearly increases to one as WEM values increase from 0 up to 6—9 mm, with higher limits for longer forecast projections.

For WEM precipitation amounts greater than ~10 mm, the WEM is compared to the high resolution ensemble mean, which is assumed to have better bias characteristics than the WEM based on the findings of Ebert (2001). The algorithm iterates over an arbitrary list of increasing precipitation thresholds, computes the bias of the volume of QPF exceeding the threshold for the WEM relative to the high resolution ensemble mean over the entire domain, and then applies a correction factor to bring this volumetric bias to unity for QPF exceeding the threshold. The correction factor is constrained to range between .5 and 2.0. As the threshold value increases, the high resolution ensemble mean is nudged toward the 90th percentile amount from the full ensemble. This is intended to augment bias for higher thresholds, at which ensemble means tend to be under-biased. The successive bias corrections alter the amount of precipitation but not its placement.

The final phase is a downscaling based on PRISM and accomplished using correction factors that vary monthly. Although more sophisticated downscaling techniques exist (Voisin et al. 2010), they are too complex and computationally demanding for the development and computing resources available to the WPC. This simple terrain correction is based on 5-km PRISM data over the western third of the CONUS and 10-km resolution data elsewhere. The method has some similarity to the terrain correction scaling used in Mountain Mapper (Henkel and Peterson 1996). The PRISM data are first remapped to the 32-km WPC QPF grid, preserving area averages. These values are then placed back on the high-resolution PRISM grid via bilinear interpolation. Then the ratios of the original PRISM data to the back-interpolated data are computed. Finally, the ratios are moved to the 32-km resolution by assigning the nearest-neighboring value from the high-resolution grid. A monthly varying lower bound ranging from .3 in the cold season to .9 in the warm season is imposed on the ratios. The downscaling coefficients are replaced with values smoothed using a 9-point smoother at points where the values are less than 1. Multiplication of the pseudo bias corrected QPF by the downscaling factor completes the ENSBC processing.

As various model data become available, the ENSBC is executed ten times per day to provide guidance for WPC forecast operations. However, a special configuration of ENSBC execution is performed to create a competitive, realistic benchmark for the WPC QPF suite of Day 1--3 forecasts. This configuration releases products in the same order as the WPC manual forecasts for two “final” cycles per day: 00 and 12 UTC. The execution schedule permits creation of products using the same models available to WPC forecasters, but without the human time handicap; therefore, the automated product suite is about an hour earlier than the WPC official delivery deadline for Day 1, almost two hours earlier for Day 2, and nearly four hours earlier for the Day 3 forecasts. It should be noted that WPC forecasters often send products well in advance of the deadlines, especially for Day 3. All comparisons to ENSBC in the main text are against this benchmark.

APPENDIX B

A Description of the WPC-EMC Forecast Verification System (fvs)

The fvs performs three functions:

1. Retrieves and combines data records read from one or more Verification Statistics DataBase (VSDB) text files under the control of user defined search conditions.
2. Computes performance metrics from the combined data.
3. Displays the performance metrics and optional statistical significance box-whisker elements graphically or in a text formatted output.

The VSDB records in the text files are created by comparisons of forecast objects to observed objects. This comparison is typically, but not necessarily, a forecast grid to analysis grid, a forecast grid to observation points, or point forecasts to point observations. The software systems used to generate VSDB record files are quite varied and not part of fvs. A single VSDB record usually contains summary statistics for comparisons at multiple analysis or observation points over an area or spatial volume. The summary statistics are either means or fractions. For example, for verification of standardized anomalies, the following means along with the data count are written in the VSDB record: means of forecast and observed anomalies, means of squares of forecast and observed anomalies, and the mean of the product of forecast and observed anomalies. With the data count, these means can be converted to partial sums that are combined in step 1 outlined above. Another example applies to verification of dichotomous events such as QPF exceeding a specific threshold for which a 2 x 2 contingency table is required. In this case, each VSDB record contains fractions of forecasts exceeding the threshold, observations exceeding the threshold, and both exceeding the threshold (hits). Again, multiplication by the data count turns these fractions into values that can be added in combining the data according to user specified search conditions.

In addition to the data values, each VSDB record contains information identifying the forecast source, forecast hour, valid time, verification area or volume, verifying analysis, parameter, and the statistic type. The statistic type is important because it determines what set of performance metrics can be computed once the VSDB records have been retrieved and combined. The user-defined search conditions are important because they inform the fvs as to the independent variable associated with the display of the performance metrics. Any of the identifier fields or combinations of them may be specified as the independent variable, so, the fvs will search for and combine VSDB records as a function of different values (string or numeric) for selected identifier information. The fvs will also perform consistency checks (event equalization) under user direction to assure equal comparisons of multiple forecast sources. If consistency checking is in force, the fvs saves the uncombined data from the search of VSDB records in a binary file. The uncombined data are used in random resampling following the method of Hamill (1999) if the user requests displays of box-whisker objects to depict statistical significance of differences of any performance metric for paired comparisons of different forecast sources.

Once step 1 is finished, the resulting data may be used to compute a variety of performance metrics, depending on the statistic type. The fvs performs step 2 and step 3 seamlessly, first computing the requested metric, then generating the display. If box-whisker objects are requested, the resampling is done separately at each point along the abscissa of the graphical depiction during the display process. Numerous user-specified parameters are provided to allow the user to control labels, text fonts, bar, line or marker characteristics, and colors for the objects appearing in the graphical display.

REFERENCES

Bakhshaii, A., and R. Stull, 2009: Deterministic ensemble forecasts using gene-expression programming. *Wea. Forecasting*, **24**, 1431–1451

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2002: Development of an “events-oriented” approach to forecast verification. Preprints, *19th Conf. on Weather Analysis and Forecasting and 15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 210-213.

Bélair, S., M. Roch, A.-M. Leduc, P. A. Vaillancourt, S. Laroche, and J. Mailhot, 2009: Medium-range quantitative precipitation forecasts from Canada’s new 33-km deterministic global operational system. *Wea. Forecasting*, **24**, 690–708.

Brill, K. F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting*, **24**, 307–318.

——, and F. Mesinger, 2009: Applying a General Analytic Method for Assessing Bias Sensitivity to Bias-Adjusted Threat and Equitable Threat Scores. *Wea. Forecasting*, **24**, 1748–1754.

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

Brown, J. D. and D-J Seo, 2010: A nonparametric post-processor for bias correcting ensemble forecasts of hydrometeorological and hydrologic variables. *J. of Hydrometeorology*, **11(3)**, 642–665.

Caplan, P., J. Derber, W. Gemmill, S.-Y. Hong, H.-L. Pan, and D. Parrish, 1997: Changes to the 1995 NCEP operational medium-range forecast model. *Wea. Forecasting*, **12**, 581–594.

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24,** 1121–1140.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410.

——, Zhu, Y. and Toth, Z., 2013: Development of statistical post-processor for NAEFS. Submitted to *Weather and Forecasting*.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158.

——, M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. A. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28,** 2031–2064.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612.

Doswell, C. A. III, 2004: Weather forecasting by humans—heuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126.

Du, J., J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) System: Inclusion of WRF Members. Preprints, *WMO Expert Team Meeting on Ensemble Prediction System*, Exeter, United Kingdom, World Meteorological Organization. [Available online at http://wwwt.emc.ncep.noaa.gov/mmb/SREF/reference.html].

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

Etherton, B. J., 2007: Preemptive forecasts using an ensemble kalman filter. *Mon. Wea. Rev.*, **135**, 3484–3495.

Fritsch, J. M., and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–965.

Giorgi, F., 1991: Sensitivity of simulated summertime precipitation over the western United States to different physics parameterizations. *Mon. Wea. Rev.*, **119**, 2870–2888.

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

——, K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The gridding of MOS. *Wea. Forecasting*, **24**, 520–529.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Henkel, A., and C. Peterson, 1996: Can deterministic quantitative precipitation forecasts in mountainous regions be specified in a rapid, climatologically consistent manner with Mountain Mapper functioning as the tool for mechanical specification, quality control, and verification? *Extended Abstracts, Fifth National Heavy Precipitation Workshop,* State College, PA, NWS/NOAA, 31 pp. [Available from Office of Climate, Water, and Weather Services, W/OS, 1325 East-West Hwy., Silver Spring, MD 20910.]

Higgins, R. W., J. E. Janowiak, and Y. -P. Yao, 1996: A gridded hourly precipitation data base for the United States (1963-1993). *NCEP/Climate Prediction Center Atlas 1*, U.S. Department of Commerce, NOAA/NWS, 47 pp.

Homar, V., D. J. Stensrud, J. J. Levit, and D. R. Bright, 2006: Value of human-generated perturbations in short-range ensemble forecasts of severe weather. *Wea. Forecasting*, **21**, 347–363.

Janjic, 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271–285.

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

Lin, Y. and K. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. Preprints. *19th Conf. on Hydrology, San Diego, CA, Amer. Meteor. Soc.*, Vol. 1, 2–5.

Magnusson, L., and E. Kallen, 2013: Factors influencing skill improvements in the ECMWF forecast system. *Mon. Wea. Rev.*, **141**, 3142–3153.

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

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.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

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

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting*, **10**, 498–511.

Reynolds, D., 2003: Value-added quantitative precipitation forecasts: how valuable is the forecaster? *Bull. Amer. Meteor. Soc.*, **84**, 876–878.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949.

——, M. Westendorf, and G. R. Meadows, 2010a: Innovative weather: a new strategy for student, university, and community relationships. *Bull. Amer. Meteor. Soc.*, **91**, 877–888.

——, 2010b: Seeking consensus: A new approach. *Mon. Wea. Rev.*, **138**, 4402–4415.

Ruth, D. P., B. Glahn, V. Dagostaro, and 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.

Sukovich, E., F. M. Ralph, D. R. Novak, F. E. Barthold, and D. Reynolds, 2013: Extreme quantitative precipitation forecast performance benchmarks over 11 years. *Wea. Forecasting*, in press.

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25,** 1603–1627.

Yussouf, N., and D. J. Stensrud, 2006: Prediction of near-surface variables at independent locations from a bias-corrected ensemble forecasting system. *Mon. Wea. Rev.*, **134**, 3415–3424.

Zheng, M., E. K. M. Chang, and B. A. Colle, 2013: Ensemble sensitivity tools for assessing extratropical cyclone intensity and track predictability. *Wea. Forecasting*, in press.

TABLES

Table 1. Timing of the availability of day 1 QPF guidance from the WPC, GFS, NAM, and ECMWF. The elapsed time between when guidance is available and when the WPC forecast is available (WPC latency) is shown in the right column.

Guidance Source	**Time Available**	**WPC Latency**
Overnight WPC	10 UTC	
00 UTC GFS	05 UTC	5 h
00 UTC NAM	03 UTC	7 h
00 UTC ECMWF	07 UTC	3 h
Overnight ENSBC	09 UTC	1 h

Table 2. Timing of the availability of medium range forecast guidance from the WPC and GFS. The elapsed time between when guidance is available and when the WPC forecast is available (WPC latency) is shown in the right column.

Guidance Source	**Time Available**	**WPC Latency** **Final (prelim)**
WPC Final (prelim)	19 UTC (14 UTC)	
00 UTC GFS MOS	06 UTC	13 h (8 h)
00 UTC ECMWF	08 UTC	11 h (6 h)
00 UTC ECMWF Ensemble	10 UTC	9 h (4 h)
12 UTC GFS MOS	18 UTC	1 h (-4 h)

FIGURE CAPTIONS

Fig. 1. Examples of WPC forecasts of (a) QPF, (b) medium range maximum temperature, and (c) medium range pressures and fronts. Examples are from different days.

Fig. 2. Timeline showing the WPC forecast period naming convention for the overnight issuance, including the forecast projection (h), time (UTC), and day 1, day 2, and day 3 designations.

Fig. 3. Time series of annual WPC threat scores for the 1-in (25.4-mm) 24-h^{-1} threshold for the day 1 (red), day 2 (green), and day 3 (blue) forecasts from 1960–2012. Percent areal coverage of the 1-in (25.4-mm) 24-h^{-1} threshold over the contiguous United States over the year is shown by the thin black line. Linear threat score trends are shown in respective colors. The linear trends are divided into two periods to account for increasing improvement after 1995. (Data updated yearly at: http://www.WPC.ncep.noaa.gov/images/WPCvrf/WPC10yr.gif)

Fig. 4. (a,c) WPC QPF percent improvement (bars) over the NAM (green), GFS (blue), and ECMWF (purple) for the (a) day 1 and (c) day 3 24-h accumulated precipitation threat score for the 1-in (25.4-mm) 24 h^{-1} threshold during the 2001–2012 period. The frequency bias of each data set is shown as diamonds. (b,d) As in (a, c) except calculated using bias-removed threat score and including the ENSBC product. Statistically significant differences from WPC at the 90% level are marked by the black asterisk.

Fig. 5. (a,c) Comparison of the threat score (bars) and frequency bias (diamonds) for the 3-in (76.2-mm) 24 h^{-1} threshold for (a) day 1 and (c) day 3 forecasts during the 2001–2012 period.

(b,d) As in (a,c) except using bias-removed threat score (bars) and including the ENSBC product. Statistically significant differences in threat score from WPC at the 90% level are marked by the black asterisk.

Fig. 6. Time series comparison of the WPC (solid) and 00 UTC GFS MOS (dashed) maximum temperature forecast Mean Absolute Error (MAE) (°F) at 98 major stations. Data are missing between 1996 and 1997.

Fig. 7. (a) Comparison of 2007–2012 time-averaged maximum temperature Mean Absolute Error for WPC and 00 UTC GFS MOS and WPC and 12 UTC GFS MOS for the day 3, 5, and 7 forecast projections. (b) WPC percent improvement over the 00 and 12 UTC GFS MOS.

Fig. 8. (top) WPC final forecast percent improvement over the 12 UTC GFS MOS at stations that were adjusted from MOS during 2012. Percent improvement (left axis) for changes from ≥1–10° F are displayed for day 4 to 7 forecasts. (bottom) Corresponding percentage of points adjusted out of a maximum of 448 points (right axis).

Fig. 9. Comparison of 5-km gridded maximum temperature Mean Absolute Error from WPC and (a) raw and (b) downscaled and bias-corrected 00 UTC ECMWF, ECMWF ensemble, GFS, and GEFS over the CONUS during 2012. The RTMA is used as the verifying analysis. Due to missing data, a homogeneous sample of 321 days is used in (a) and 313 days in (b). Statistically significant differences from WPC at the 90% level are shown as asterisks.

Fig. 10. Comparison of PMSL forecast anomaly correlation for the WPC final forecast and various international model guidance. The 90% confidence interval relative to the WPC forecast is shown as a box.

FIGURES

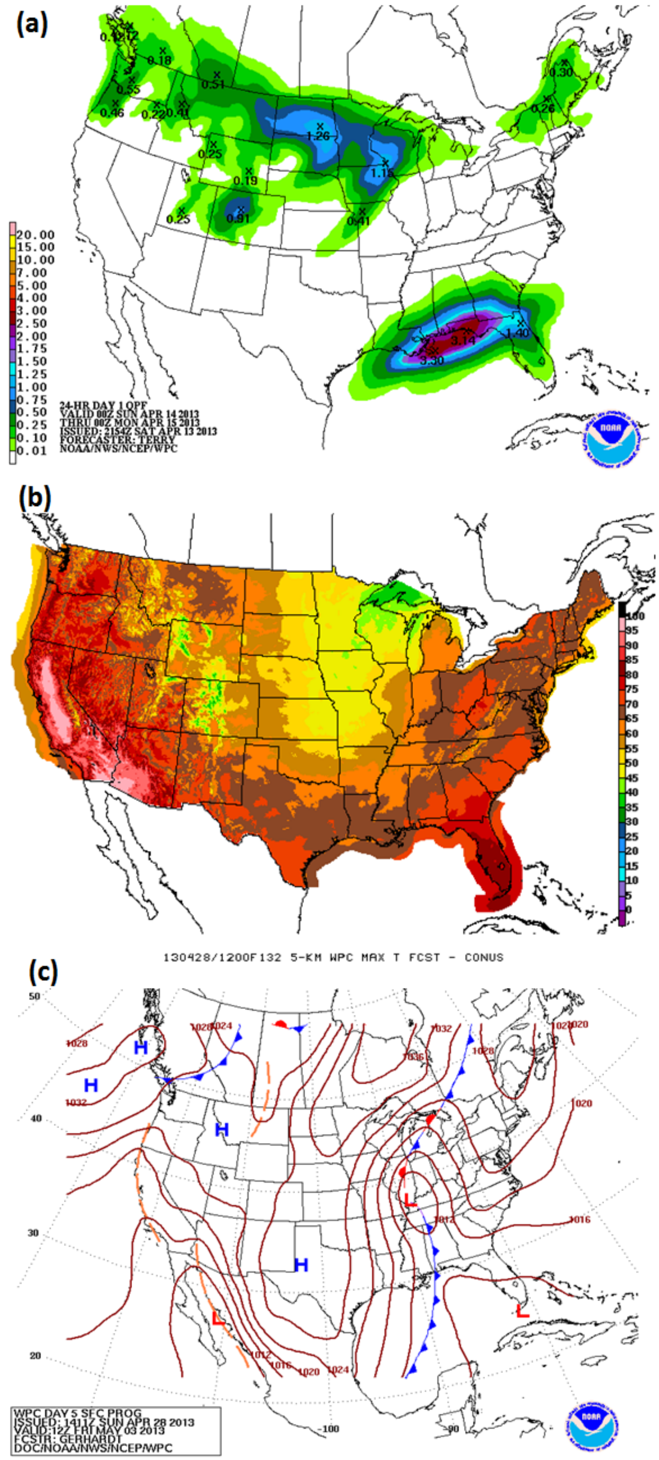

Fig. 1. Examples of WPC forecasts of (a) QPF, (b) medium range maximum temperature, and (c) medium range pressures and fronts. Examples are from different days.

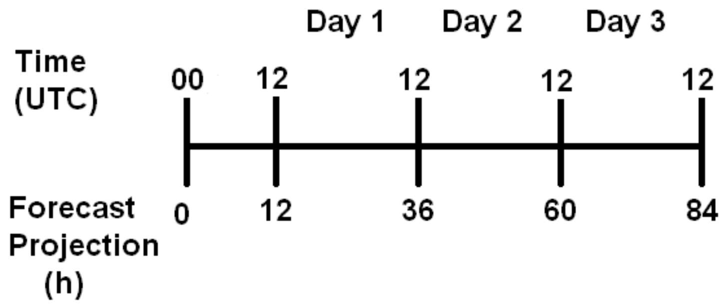

Fig. 2. Timeline showing the WPC forecast period naming convention for the overnight issuance, including the forecast projection (h), time (UTC), and day 1, day 2, and day 3 designations.

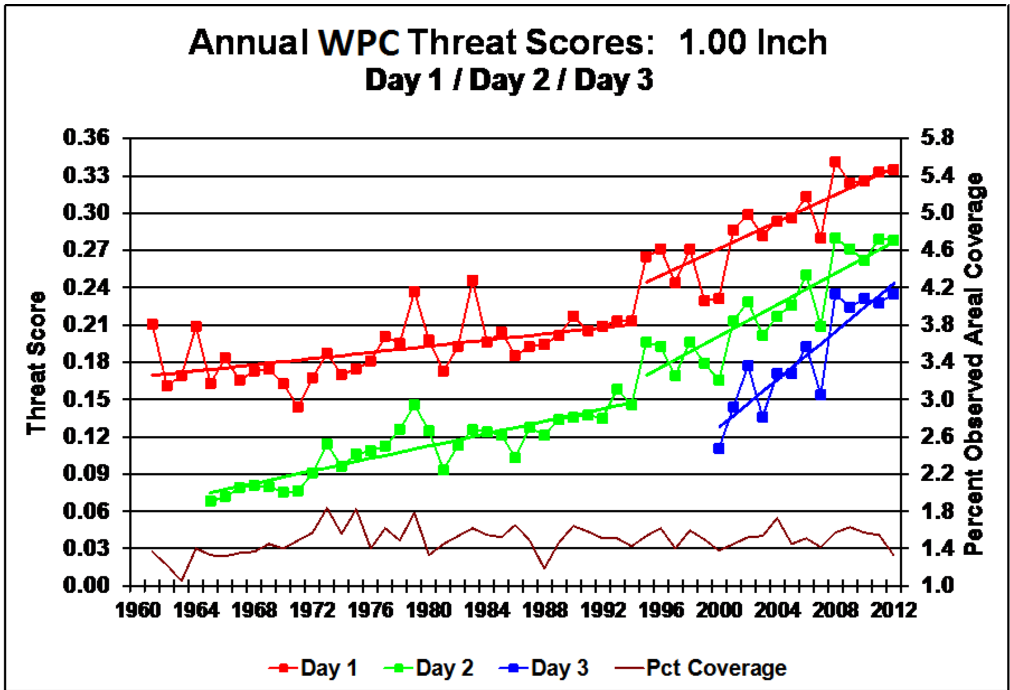

Fig. 3. Time series of annual WPC threat scores for the 1-in (25.4-mm) 24-h^{-1} threshold for the day 1 (red), day 2 (green), and day 3 (blue) forecasts from 1960–2012. Percent areal coverage of the 1-in (25.4-mm) 24-h^{-1} threshold over the contiguous United States over the year is shown by the thin black line. Linear threat score trends are shown in respective colors. The linear trends are divided into two periods to account for increasing improvement after 1995. (Data updated yearly at: http://www.WPC.ncep.noaa.gov/images/WPCvrf/WPC10yr.gif)

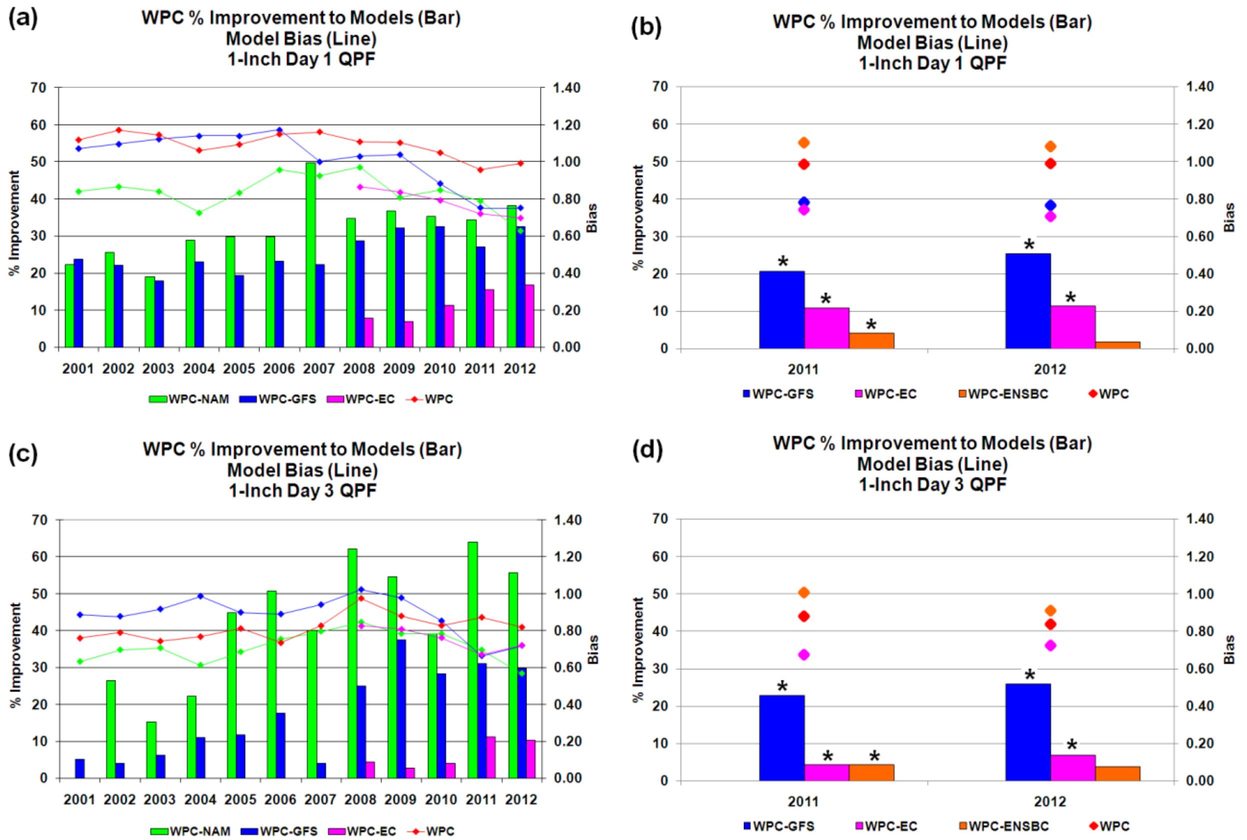

Fig. 4. (a,c) WPC QPF percent improvement (bars) over the NAM (green), GFS (blue), and ECMWF (purple) for the (a) day 1 and (c) day 3 24-h accumulated precipitation threat score for the 1-in (25.4-mm) 24 h^{-1} threshold during the 2001–2012 period. The frequency bias of each data set is shown as diamonds. (b,d) As in (a, c) except calculated using bias-removed threat score and including the ENSBC product. Statistically significant differences from WPC at the 90% level are marked by the black asterisk.

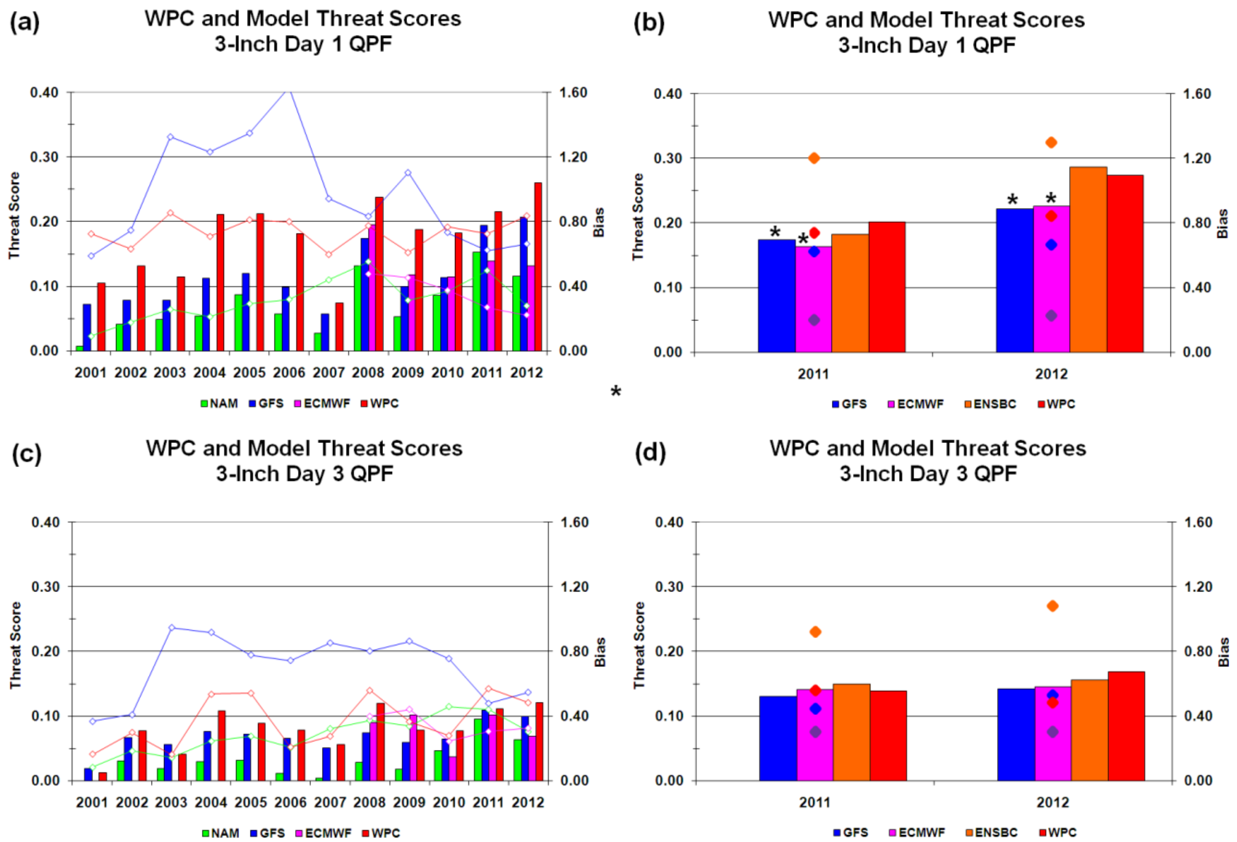

Fig. 5. (a,c) Comparison of the threat score (bars) and frequency bias (diamonds) for the 3-in (76.2-mm) 24 h^{-1} threshold for (a) day 1 and (c) day 3 forecasts during the 2001–2012 period. (b,d) As in (a,c) except using bias-removed threat score (bars) and including the ENSBC product. Statistically significant differences in threat score from WPC at the 90% level are marked by the black asterisk.

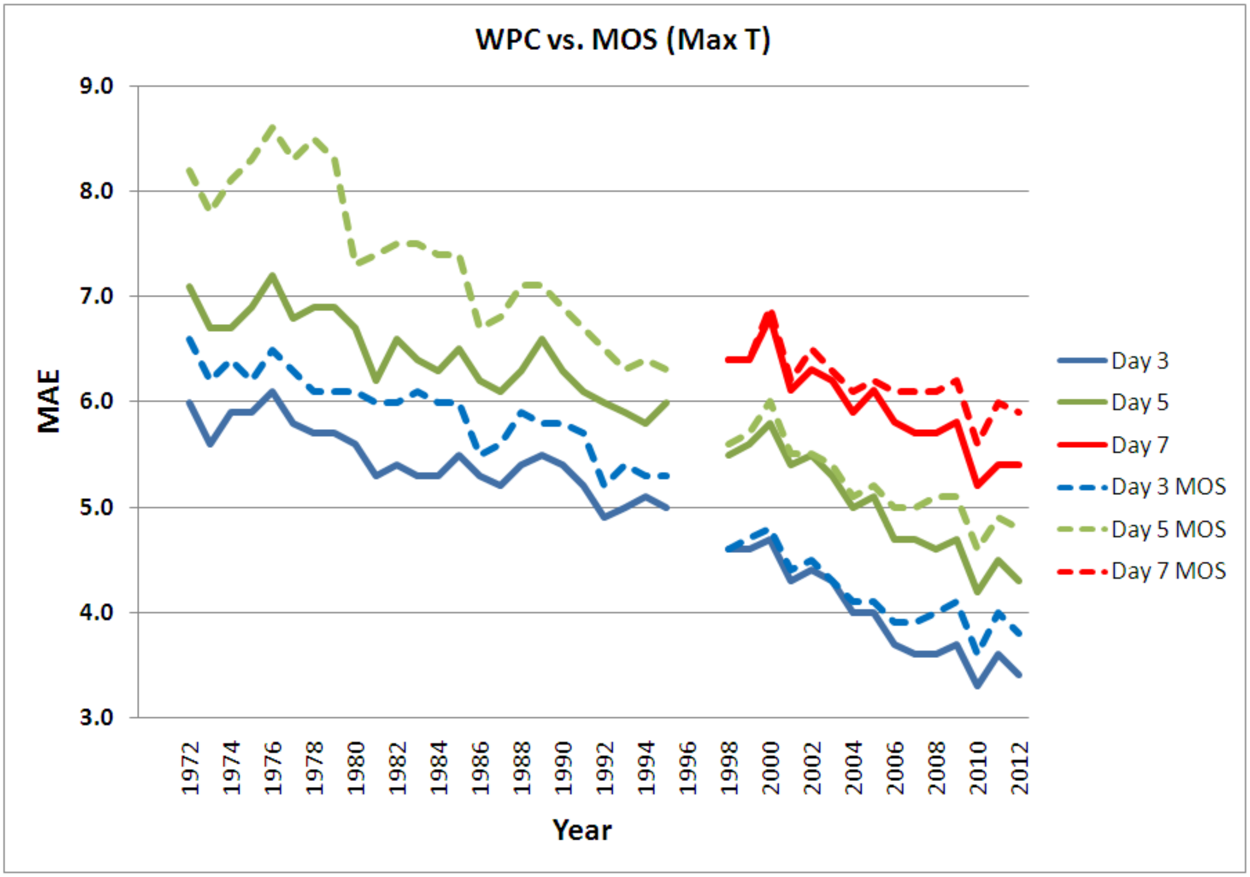

Fig. 6. Time series comparison of the WPC (solid) and 00 UTC GFS MOS (dashed) maximum temperature forecast Mean Absolute Error (MAE) (°F) at 93 major stations. Data are missing between 1996 and 1997.

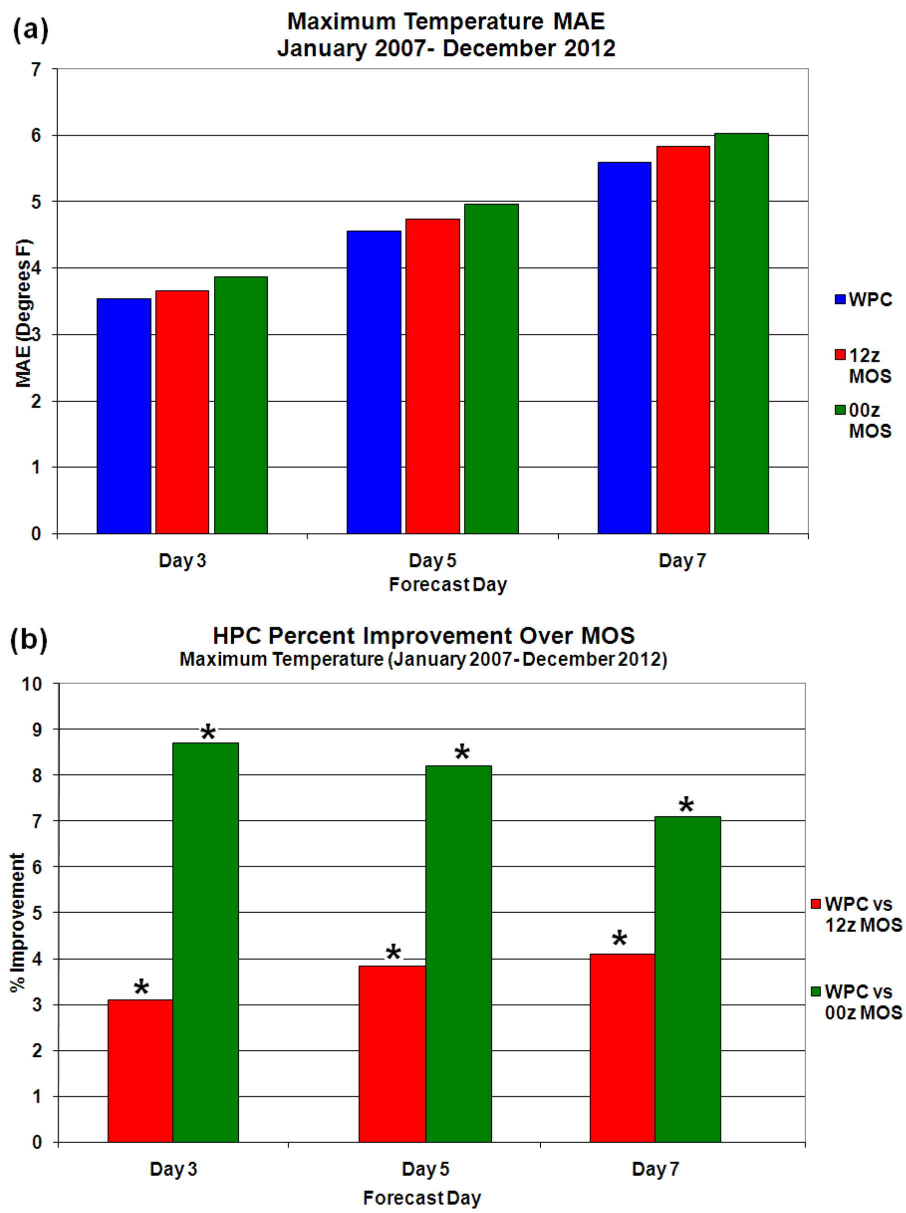

Fig. 7. (a) Comparison of 2007–2012 time-averaged maximum temperature Mean Absolute Error for WPC and 00 UTC GFS MOS and WPC and 12 UTC GFS MOS for the day 3, 5, and 7 forecast projections at 448 points. (b) WPC percent improvement over the 00 and 12 UTC GFS MOS. Statistically significant differences in percent improvement from WPC at the 90% level are marked by the black asterisk.

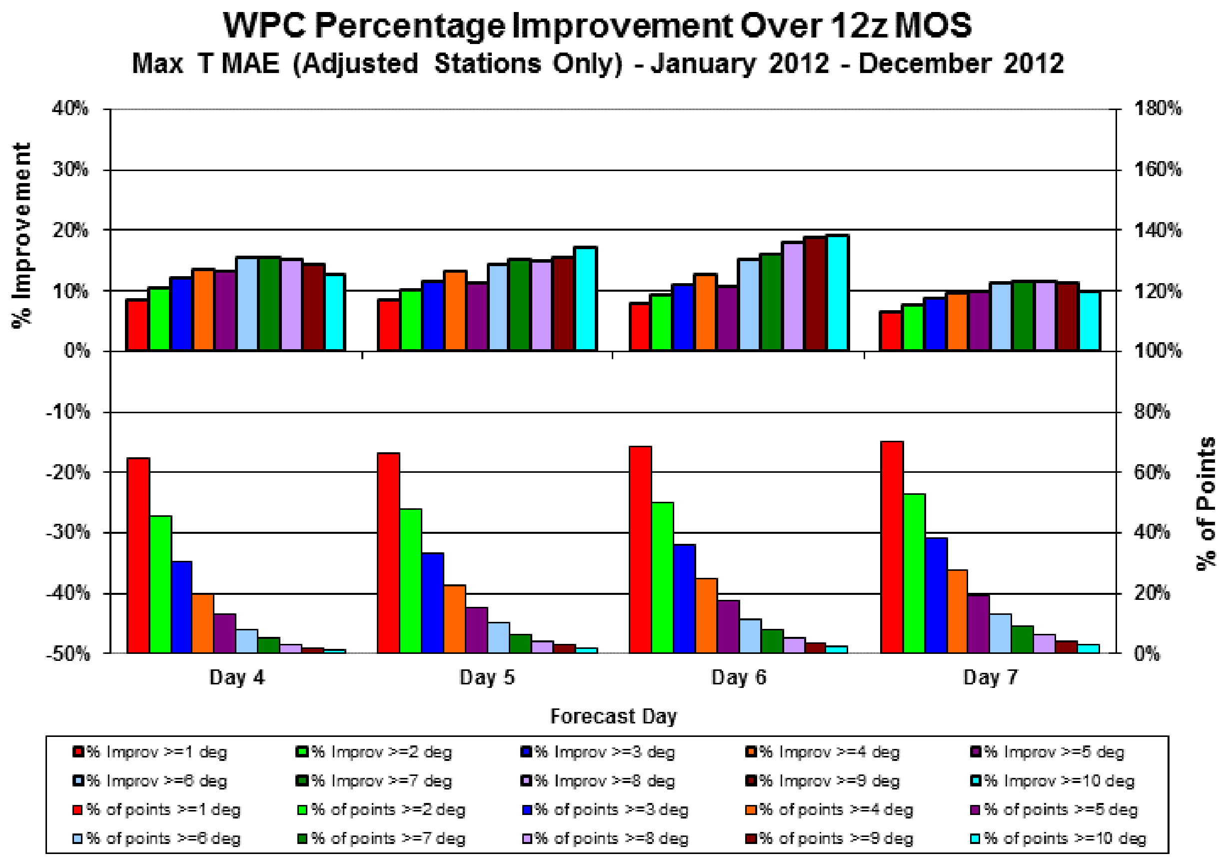

Fig. 8. (top) WPC final forecast percent improvement over the 12 UTC GFS MOS at stations that were adjusted from GFS MOS during 2012. Percent improvement (left axis) for changes from ≥1–10° F are displayed for day 4 to 7 forecasts. (bottom) Corresponding percentage of points adjusted out of a maximum of 448 points (right axis).

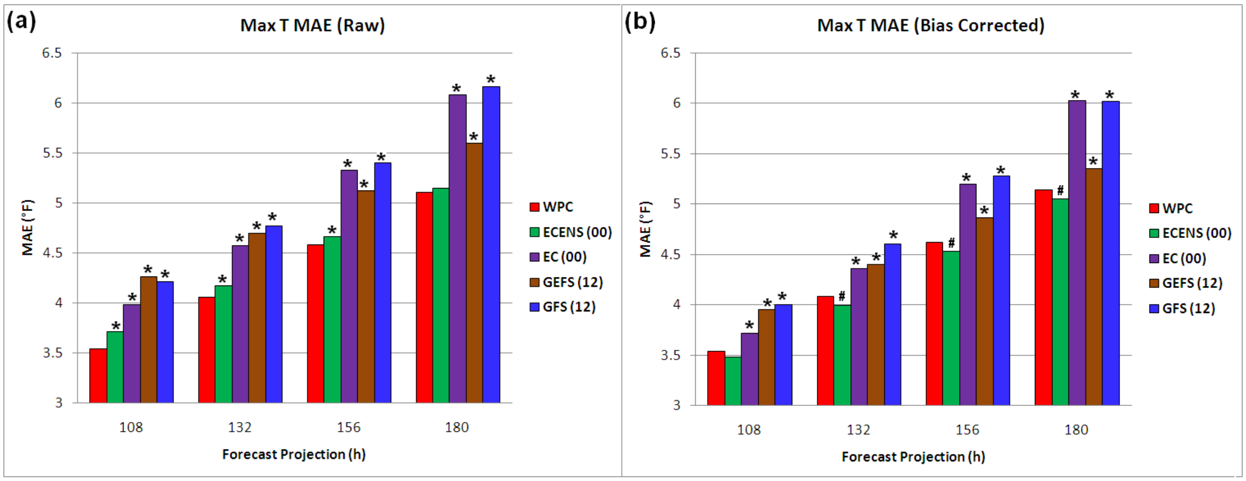

Fig. 9. Comparison of 5-km gridded maximum temperature Mean Absolute Error from WPC (red bar) and (a) raw and (b) downscaled and bias-corrected 00 UTC ECMWF, ECMWF ensemble, and 12 UTC GFS, and GEFS over the CONUS during 2012. The RTMA is used as the verifying analysis. Due to missing data, a homogeneous sample of 321 days is used in (a) and 313 days in (b). Statistically significant positive (negative) differences from WPC at the 90% level are shown as asterisks (number sign).

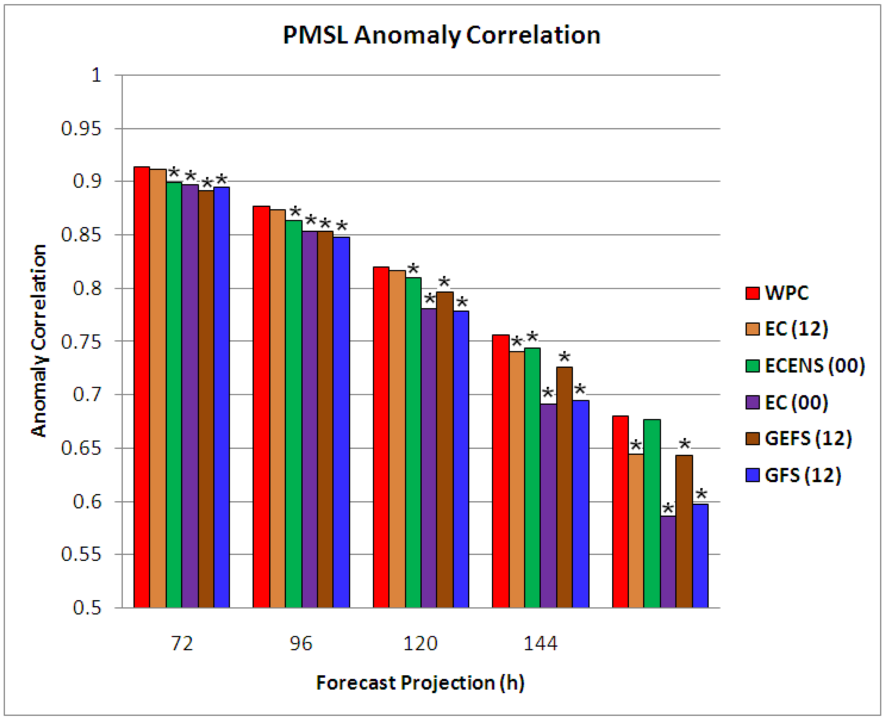

Fig. 10. Comparison of PMSL forecast anomaly correlation for the WPC final forecast and various international model guidance for 2012. Statistically significant differences from WPC at the 90% level are shown as asterisks.