Verification of Quantitative Precipitation Reforecasts over the Southeastern United States

MARTIN A. BAXTER

Department of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, Michigan

GARY M. LACKMANN

Department of Marine, Earth, and Atmospheric Sciences, North Carolina State University, Raleigh, North Carolina

KELLY M. MAHONEY

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and NOAA/Earth System Research Laboratory/Physical Sciences Division, Boulder, Colorado

THOMAS E. WORKOFF

NOAA/NCEP/Weather Prediction Center, and Systems Research Group, Inc., College Park, Maryland

THOMAS M. HAMILL

NOAA/Earth System Research Laboratory/Physical Sciences Division, Boulder, Colorado

(Manuscript received 20 May 2014, in final form 12 June 2014)

ABSTRACT

NOAA's second-generation reforecasts are approximately consistent with the operational version of the 2012 NOAA Global Ensemble Forecast System (GEFS). The reforecasts allow verification to be performed across a multidecadal time period using a static model, in contrast to verifications performed using an ever-evolving operational modeling system. This contribution examines three commonly used verification metrics for reforecasts of precipitation over the southeastern United States: equitable threat score, bias, and ranked probability skill score. Analysis of the verification metrics highlights the variation in the ability of the GEFS to predict precipitation across amount, season, forecast lead time, and location. Beyond day 5.5, there is little useful skill in quantitative precipitation forecasts (QPFs) or probabilistic QPFs. For lighter precipitation thresholds [e.g., 5 and 10 mm $(24\,\text{h})^{-1}$], use of the ensemble mean adds about 10% to the forecast skill over the deterministic control. QPFs have increased in accuracy from 1985 to 2013, likely due to improvements in observations. Results of this investigation are a first step toward using the reforecast database to distinguish weather regimes that the GEFS typically predicts well from those regimes that the GEFS typically predicts poorly.

1. Introduction

Attendant with the development of advanced numerical weather prediction (NWP) systems is the need to verify the capabilities of these systems. In particular, the quantitative precipitation forecast (QPF) is important to society and challenging for NWP systems. At what forecast lead time does NWP QPF skill effectively vanish? How much additional skill is provided by the use of ensemble versus deterministic forecasts? Has QPF skill changed over time? The development of NOAA's second-generation reforecast database (Hamill et al. 2013) allows these questions to be addressed. Due to challenges in verifying QPF in areas where the precipitation climatology varies considerably (Hamill and Juras 2006), our focus is restricted to the southeastern United States (SEUS), where climatological precipitation characteristics are relatively homogeneous (Prat and Nelson 2014).

Corresponding author address: Martin A. Baxter, Dept. of Earth and Atmospheric Sciences, Central Michigan University, 314 Brooks Hall, Mount Pleasant, MI 48859.
E-mail: baxte1ma@cmich.edu

DOI: 10.1175/WAF-D-14-00055.1

The SEUS receives precipitation associated with a variety of meteorological phenomena, including tropical cyclones, baroclinic waves, mesoscale convective systems, and localized diurnal convection (Moore et al. 2014, manuscript submitted to *Mon. Wea. Rev.*, hereafter MMSCH). The juxtaposition of the Appalachian Mountains to the Atlantic Ocean and Gulf of Mexico creates an environment where QPFs are particularly challenging. While QPFs from NWP and forecasters have improved over the last 30 yr (Novak et al. 2014), challenges remain. A static model run over multiple decades allows for verifications that span time scales longer than the periods that operational models remain unchanged. Long-term verification of reforecast data provides a sufficiently large sample size to allow the development of model climatology, which can then be compared with the climatology of the atmosphere.

NWP verification can inform forecasters of the strengths, limitations, and best applications of a modeling system. Ensembles attempt to provide a measure of confidence in a particular outcome, but if the model driving the ensemble system has difficulty in predicting a phenomenon, the ensemble spread can be misleading to a forecaster and therefore has reduced utility. Thus, it is helpful to document the accuracy of a modeling system's QPFs over long periods, and to provide this information to forecasters in a manner that allows for easy incorporation into the forecast process. As the value added by human forecasters over model QPF is diminishing (Novak et al. 2014), the need to leverage any sources of available information to improve upon model QPF becomes increasingly critical.

The National Oceanic and Atmospheric Administration's (NOAA) second-generation reforecasts are approximately consistent with the operational 0000 UTC cycle of the 2012 NOAA Global Ensemble Forecast System (GEFS). The 11-member reforecasts were created at ~0.5° grid spacing out to day 8 and ~0.75° grid spacing for day-8–16 forecasts. Because of the change in grid spacing, we use only day-0–7 reforecasts. The reforecasts were verified using gridded precipitation data from NOAA/Climate Prediction Center's Daily U.S. Unified Precipitation Dataset (Chen et al. 2008). This analysis is composed of gauge observations interpolated onto a 0.25° grid. The 29 yr (January 1985–December 2013) of verification undertaken here precludes the use of a multisensor precipitation dataset. The reforecasts and precipitation data were interpolated onto a common 0.5° × 0.5° grid over the SEUS (Fig. 6, described in greater detail below, depicts the area of study) using bilinear interpolation. As the period of observed precipitation is from 1200 to 1200 UTC, and the reforecasts were initialized at 0000 UTC, the lead times used in this study are from 1.5 days (12–36 h) through 7.5 days (156–180 h). For in-depth explanation of verification metrics, the reader is directed to Wilks (2011, chapter 8) and Jolliffe and Stephenson (2012, chapters 2 and 3). To enhance readability, all time series of yearly quantities have been filtered with the 1–2–1 (or Hanning) filter (see Von Storch and Zwiers 1999). Linear regressions and their statistical significance at the 95% confidence level have been computed for all time series using unsmoothed data.

2. Verification of deterministic forecasts

The equitable threat score (ETS) evaluates a model's ability to predict a two-category event. Values range from $-\frac{1}{3}$ to 1, where 1 represents a perfect forecast and 0 or less indicates unskilled forecasts. ETS is given by

$$\frac{h - h_c}{h + \mathrm{fa} + m - h_c} \quad \text{where} \quad h_c = \frac{(h + \mathrm{fa})(h + m)}{n}. \tag{1}$$

Here, h are hits, fa are false alarms, m are misses, and h_c represents hits correct by chance (n is equal to h + fa + m + cn, where cn are correct negatives).

Generally, time series of annual ETS depict a statistically significant upward trend (Fig. 1). As the model and data assimilation system used in the reforecasts remains static over the period, the upward trend in ETS is likely due to improvements of the initial conditions used by the model through increased and higher quality observations. Alternatively, the upward trend may in part be explained by changes in the quality of the verification dataset. The year-to-year variability in ETS evidently results from variability in the model's ability to predict the phenomena that lead to precipitation in those years. Also, ETS tends to increase with fractional area coverage of the phenomenon (Hamill 1999; MMSCH). Model QPF errors arise from both the quality of the model and observations, and the model atmosphere's sensitive dependence on the initial conditions [Zhang et al. (2006), after Lorenz (1996)]. ETS decreases with increasing lead time and with increasing threshold. Beyond day 5.5, average ETS values are below 0.1 (except for the 5-mm threshold), indicating little to no skill. When the average ETS over 1985–89 is compared with 2009–13, an increase in ETS is seen for all lead times at all thresholds, though trends over the entire period are not significant for the three time series shown with dashed lines in Fig. 1. The increases in ETS over the period for day-1.5 forecasts are 0.088, 0.084, 0.073, and 0.081 for 5-, 10-, 25-, and 40-mm thresholds, respectively. For day 7.5, the increases in ETS are 0.030, 0.025, 0.015 (not significant), and 0.006 (not significant) for 5-, 10-, 25-, and 40-mm

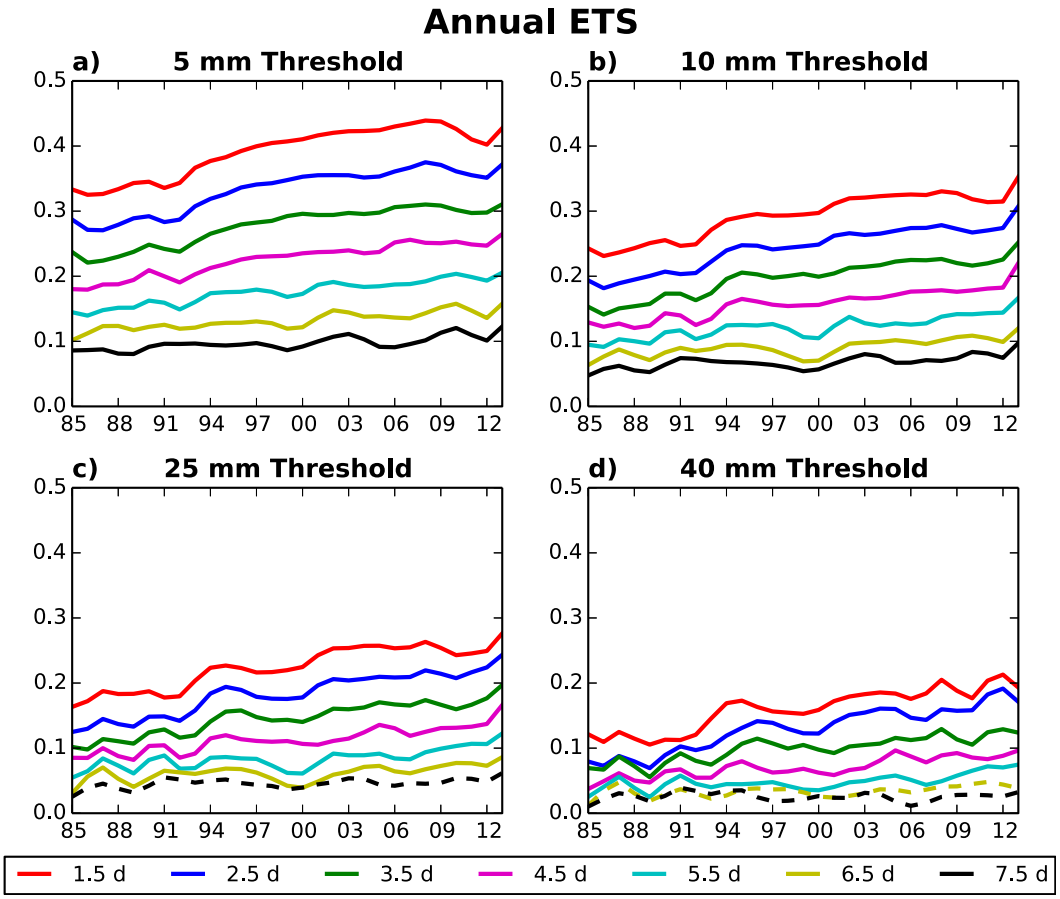

FIG. 1. Annual ETSs from 1985 to 2013 for (a) 5-, (b) 10-, (c) 25-, and (d) 40-mm thresholds. Lines are forecast lead times as indicated in the legend. Dashed lines indicate trends are not significant at the 95% confidence level. The domain used for calculation can be seen in Fig. 6.

thresholds, respectively. Thus, day-1.5 ETS values have increased by 3, 3, 5, and 14 times for 5-, 10-, 25-, and 40-mm thresholds, respectively, when compared with day-7.5 ETS values over the nearly 30-yr period. The most recent day-3.5 ETS values are approximately equivalent in accuracy to the oldest day-1.5 reforecasts, as follows for 5-, 10-, 25-, and 40-mm thresholds: 0.33 versus 0.36, 0.24 versus 0.22, 0.18 versus 0.17, and 0.11 versus 0.12. This increase in useful lead time is likely due to increases in the quantity and quality of the observations, and illustrates the considerable impact these better observations have had on QPFs over the SEUS.

Considerable seasonal variability in ETS is present for the 20-mm threshold (Fig. 2). The 20-mm threshold is chosen for further analysis, as it represents an ''intermediate'' amount of precipitation in the SEUS, and it occurs with sufficient frequency to allow meaningful analysis. Average ETS is highest in winter (0.19), followed by fall (0.16), spring (0.15), and summer (0.07), consistent with Fig. 3 of Hamill et al. (2013). This order in average ETS persists for days 1.5–5.5. At days 6.5 and 7.5, average ETSs for winter, spring, and fall are essentially the same, within 0.01. Beyond day 5.5, ETS values for all seasons are below 0.1, indicating little to no skill. As with annual ETS, seasonal ETS exhibits an increase in 5-yr-average ETS from the beginning to the end of the period, with most trends over the entire period being statistically significant. The trend in summer-day-1.5 ETS is minimal (0.01), less than that of winter (0.13), spring (0.10), or fall (0.10), suggesting that improvements in observations have not led to as much improvement in summer QPF compared with other seasons (e.g., Fritsch and Carbone 2004). Similar patterns are seen in ETS for other thresholds.

The bias score B describes the ratio of the number of yes forecasts to the number of yes observations, where h, fa, and m are as in (1):

$$B = \frac{h + \mathrm{fa}}{h + m}. \tag{2}$$

When bias exceeds one, the event is overforecast, and when bias is less than one, the event is underforecast. A

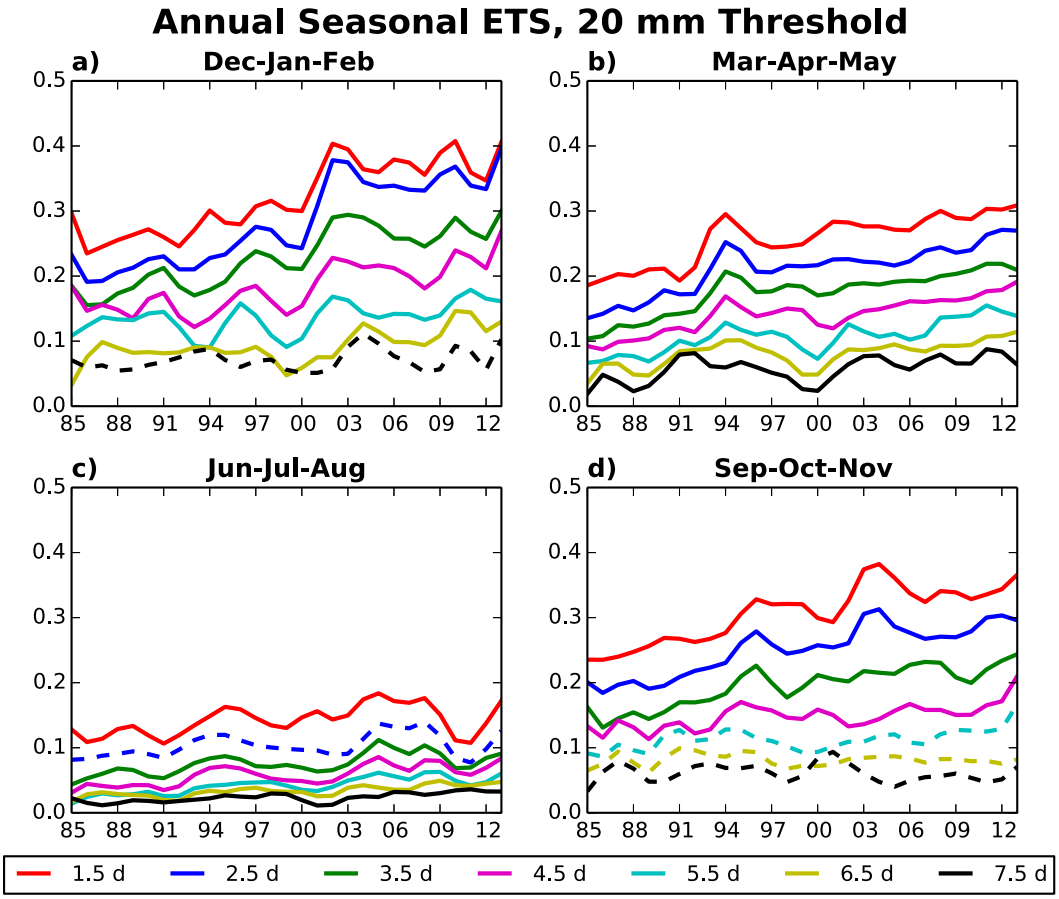

FIG. 2. Annual ETSs for the 20-mm threshold from 1985 to 2013 for (a) winter, (b) spring, (c) summer, and (d) fall. Lines are forecast lead times as indicated in the legend. Dashed lines indicate trends are not significant at the 95% confidence level.

bias of one indicates the event was forecast the same number of times as it was observed. Bias does not measure the correspondence between individual forecast–observation pairs and, thus, provides no information on model accuracy. At a 20-mm threshold, precipitation is underforecast in all seasons (Fig. 3), as indicated by averages over all years and forecast lead times: winter (0.94), spring (0.89), summer (0.62), and fall (0.72). The bias closer to unity in the winter may result from the tendency for gauges to underestimate precipitation that falls as snow (Rasmussen et al. 2012). Annual variability in the bias likely arises from variation in the number of events forecast or observed from year to year. Only four of the trends in the 28 bias time series are statistically significant.

3. Verification of ensemble forecasts

The ensemble mean versus control QPF was compared by calculating the ETS for each season over the 29-yr period for both forecasts and, then, finding the percent difference (Fig. 4). In general, the ensemble mean provides the most improvement over the control for lower thresholds and shorter lead times. As thresholds increase, the ensemble mean is more susceptible to smearing (i.e., multiple nonoverlapping positions of the precipitation maximum across ensemble members), leading to underestimation of the magnitudes found in the control. This effect is most pronounced in summer, when rain amounts are climatologically higher and precipitation is more localized. In all seasons but summer, and for all thresholds except for 40 mm, the ensemble mean has an average of 6% higher ETS values versus the control through 4.5-day lead time. In all seasons but summer, the 5- and 10-mm thresholds have an average of 10% higher ETS values in the mean for all lead times through 7.5 days.

The ranked probability score (RPS) measures the difference between the cumulative distributions of

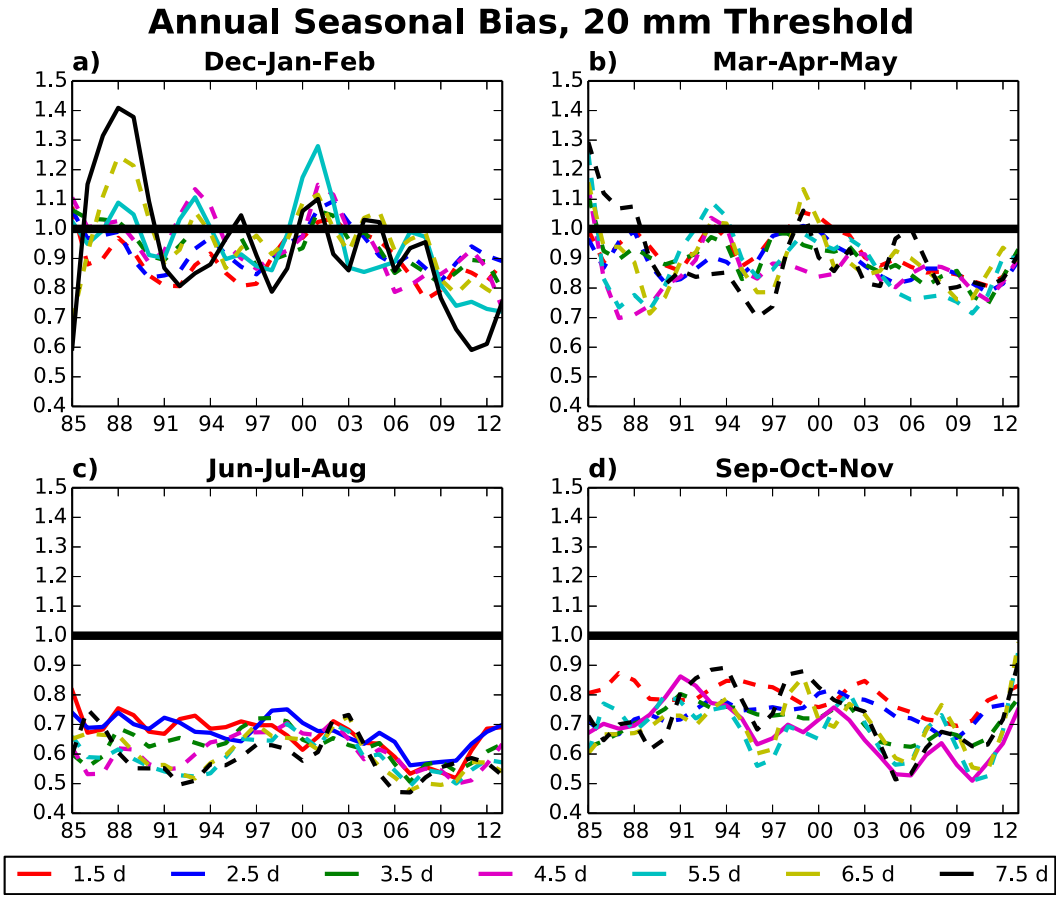

FIG. 3. Annual B for the 20-mm threshold from 1985 to 2013 for (a) winter, (b) spring, (c) summer, and (d) fall. Lines are forecast lead times as indicated in the legend. Dashed lines indicate trends are not significant at the 95% confidence level.

forecasts and observations over a set of categories, as follows:

$$\mathrm{RPS} = \sum_{k=1}^{K} (\mathrm{CDF}_{\mathrm{fc},k} - \mathrm{CDF}_{\mathrm{obs},k})^2, \qquad (3)$$

where $k = 1, \ldots, K$ indicates the number of categories, and CDF refers to the cumulative distribution of either the forecast fc or observations obs. Categories were chosen to be similar to those used for probabilistic QPF at NOAA's Weather Prediction Center.[1] A ranked probability skill score is defined as $\mathrm{RPSS} = 1 - \overline{\mathrm{RPS}}/\overline{\mathrm{RPS}_{\mathrm{CL}}}$, where $\mathrm{RPS}_{\mathrm{CL}}$ is the RPS computed using the cumulative climatological distribution to forecast the cumulative distribution of the observations. Overbars indicate that values are averaged over time and space. RPSS values less than 0 indicate that a probabilistic quantitative precipitation forecast (PQPF) from the ensemble is no better than using a climatological distribution as a probabilistic prediction. Seasonal values of $\mathrm{RPS}_{\mathrm{CL}}$ were calculated at every grid point using cumulative climatological distributions for each season. Note that comparing the RPSS values applied here with other ensemble systems with differing numbers of members will not provide an even comparison, and additional calculations would be needed to appropriately make such comparisons (Richardson 2001).

As with ETS from the control run, when RPSS are averaged over all years and forecast lead times (Fig. 5), winter has the highest skill score (0.23), followed by fall (0.19), spring (0.14), and summer (−0.05). The negative summer value indicates that RPS values are worse than those that could be achieved by using climatology. In the summer, only day-1.5 RPS values exceed those from

[1] Categories are mutually exclusive and encompass all possibilities: ≥0 and <1, ≥1 and <3, ≥3 and <5, ≥5 and <10, ≥10 and <20, ≥20 and <25, ≥25 and <40, ≥40 and <50, ≥50 and <65, ≥65 and <75, and ≥75 mm.

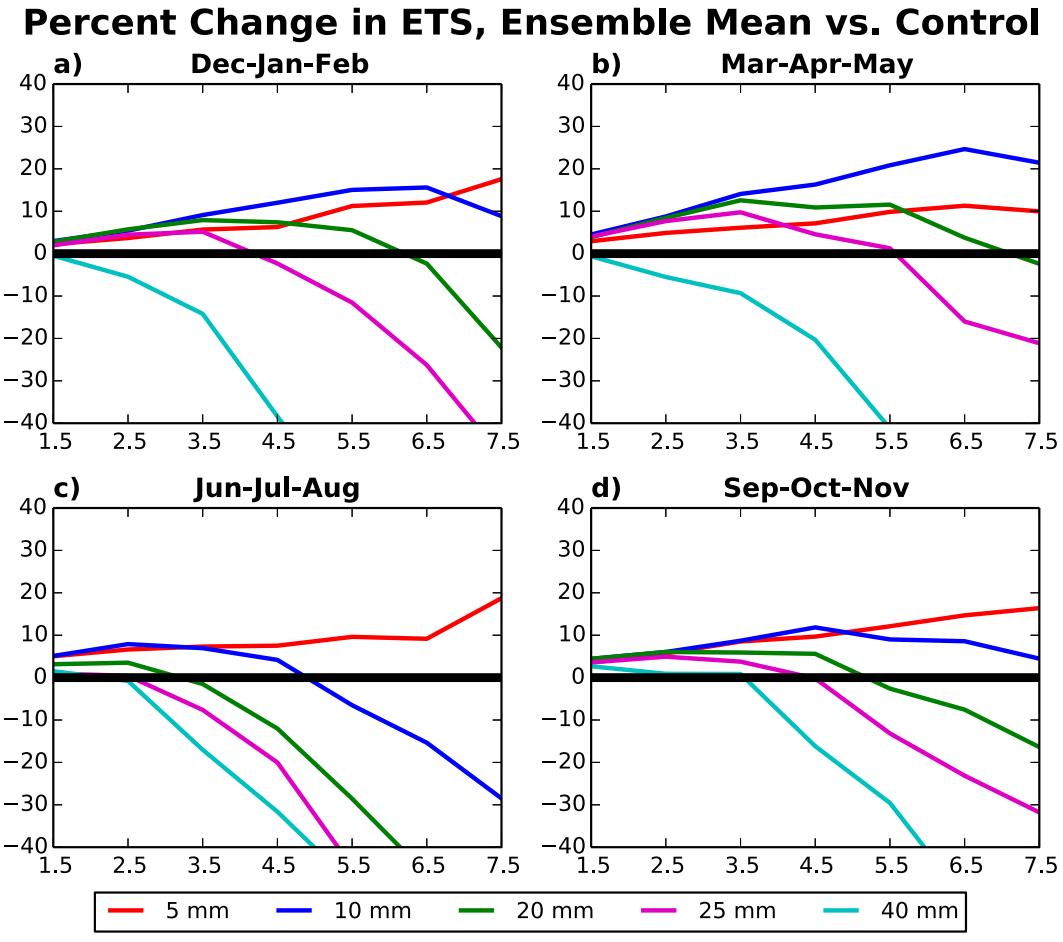

FIG. 4. Changes (%) in ETS with forecast lead time when ensemble mean is compared with control for (a) winter, (b) spring, (c) summer, and (d) fall over the 1985–2013 period. Lines are precipitation thresholds as indicated in the legend.

climatology. This result indicates that the resolution of the reforecast system is not adequate to predict the scale of the phenomena dominant in the summer months in the SEUS; intrinsic predictability in the model is reduced in the presence of convection (Zhang et al. 2003). Systematic deficiencies in the GEFS also contribute to QPF error, such as the excessive semblance of ensemble members, resulting in overconfident probabilistic forecasts (Palmer 2012). Average RPSS values beyond day 5.5 are less than 0.1, indicating little to no skill. Seasonal RPSSs experience an upward trend when 5-yr-average RPSSs from the first 5 yr are compared with those of the last 5 yr. All forecast leads in every season exhibit this upward trend, though trends for four time series are not statistically significant when linear regressions are computed over the entire period. This indicates that better observations have led to not only an increase in accuracy in deterministic QPF, but also in the ensemble's ability to provide a probabilistic QPF. The trend in day-1.5 RPSS is greatest for fall (0.27), followed by winter (0.20), spring (0.12), and summer (0.10). Interestingly, the summer exhibits a positive trend for PQPF as measured by RPSS, while no appreciable trend is observed for deterministic QPF as measured by ETS.

Spatial plots of seasonal RPSSs (Fig. 6) qualitatively match the day-1.5 curves in Fig. 5, with winter and fall having the highest RPSSs, and spring and summer the lowest RPSSs. In the summer, PQPFs are no better than climatology when measured by RPS in parts of the domain. Outside of summer, areas in the interior of the domain feature the greatest improvement over climatology. Longer lead times follow similar patterns, with the areas of greatest improvement shrinking in size toward the center of the domain.

4. Conclusions and future work

To assess QPF skill as a function of lead time, and to quantify the benefit of the ensemble mean over deterministic QPFs, verification of the NOAA second-generation

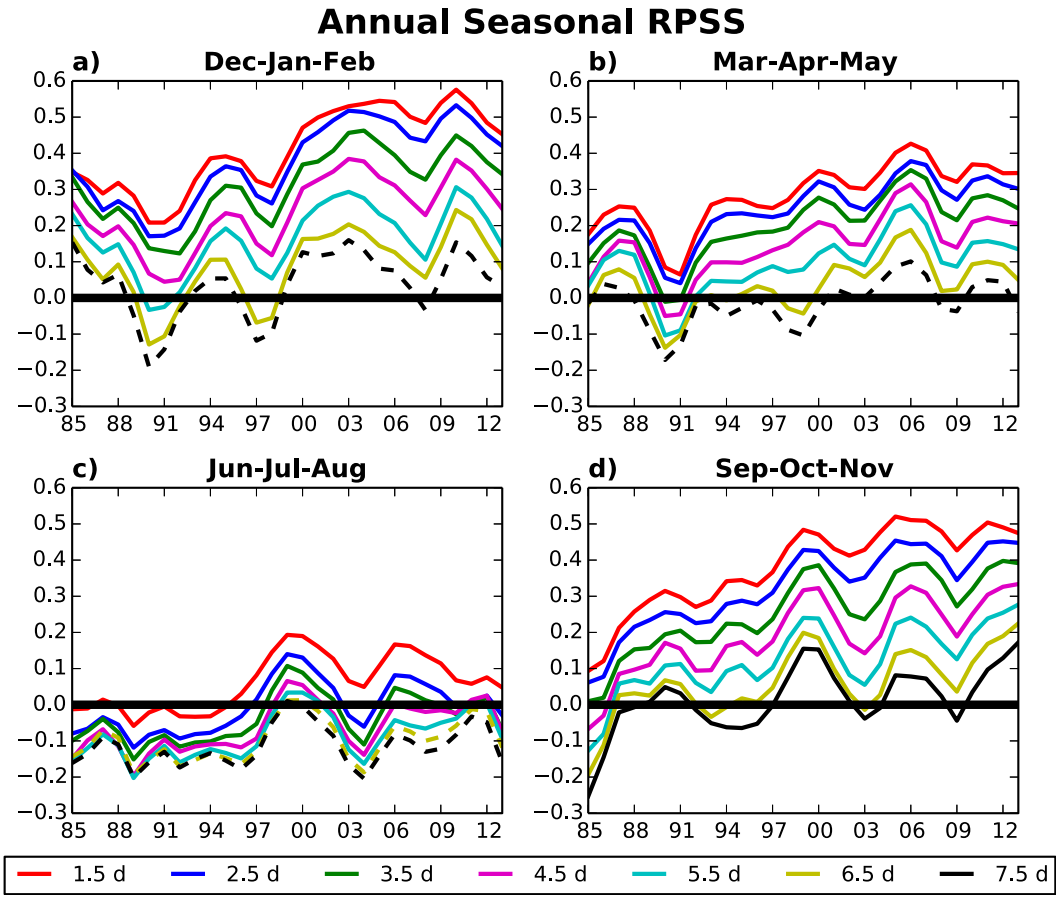

FIG. 5. Annual RPSSs from 1985 to 2013 for (a) winter, (b) spring, (c) summer, and (d) fall. Lines are forecast lead times as indicated in the legend. Dashed lines indicate trends are not significant at the 95% confidence level.

reforecast dataset has been completed for the SEUS from 1985 to 2013. This long-term verification provides forecasters with useful information on the capabilities of the current generation of NOAA's GEFS. Both deterministic and probabilistic QPFs exhibit long-term increases, likely due to the improvement of the observations. For deterministic QPFs, summer does not feature an appreciable increase, while an increase in summer occurs for PQPFs. For all seasons, day-6.5 and day-7.5 QPFs and PQPFs for the SEUS have little to no skill over random chance or climatology, respectively. The use of the ensemble mean rather than the control offers benefit on average, especially for lower thresholds and shorter lead times. Reforecasts can be used to improve forecasts of precipitation through many techniques, such as the analog-based technique of Hamill and Whitaker (2006) or logistic regression (Hamill et al. 2008). Verification of these adjusted real-time forecasts has shown them to be superior to the reforecasts themselves and the unadjusted real-time forecasts (Hamill et al. 2013).

MMSCH produced a 10-yr climatology of extreme precipitation events in the SEUS and found that events associated with stronger dynamical forcing were more accurately predicted by the reforecasts in terms of ETS, bias, and fractional area. This is corroborated by the present study of all precipitation events in the SEUS, as the summer events were poorly predicted and are more likely to be associated with weaker dynamical forcing. Future work will analyze patterns associated with the most and least accurate reforecasts of precipitation events in the SEUS. Greater understanding of how the modeled atmosphere differs from the real atmosphere will allow forecasters and researchers to identify situations where model guidance is likely to be poor. In addition, continuing analysis of the complex relationship between forecast precipitation, ensemble spread, and accuracy will help forecasters to better convert ensemble guidance into useful forecast confidence (Palmer 2012). The results of such analyses can help forecasters better allocate their time and effort in improving model guidance, and would allow researchers to better allocate

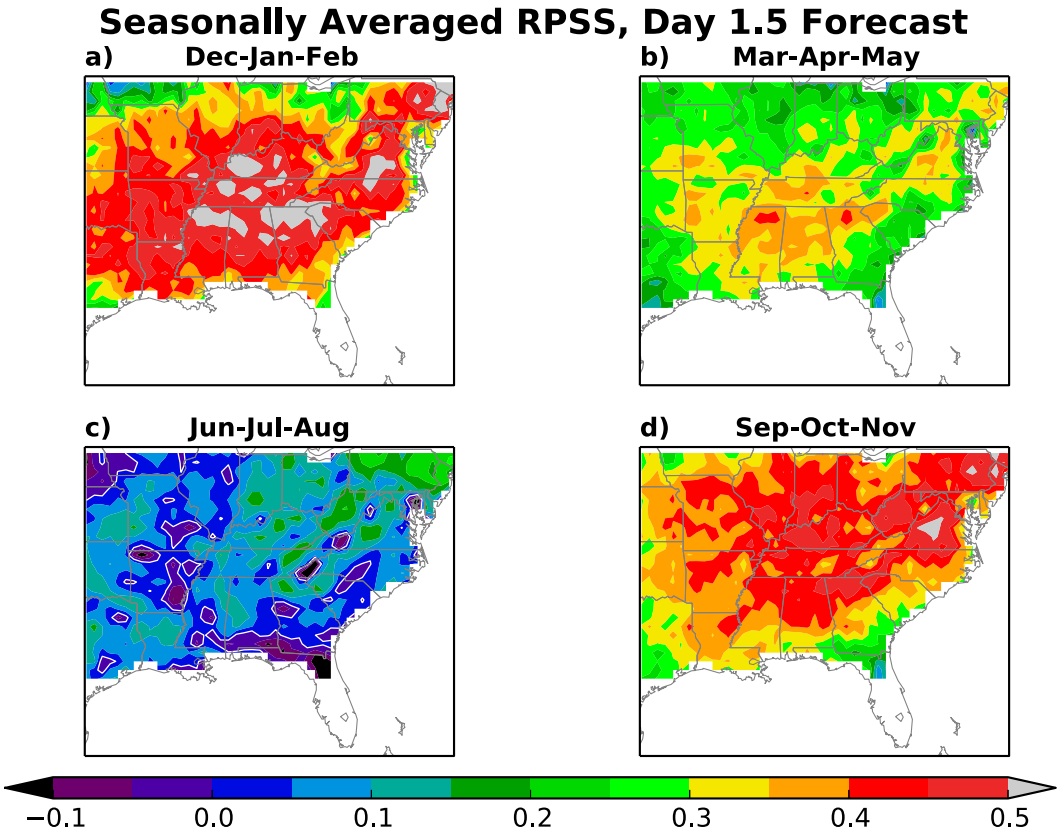

FIG. 6. Day-1.5 RPSSs from 1985 to 2013 for (a) winter, (b) spring, (c) summer, and (d) fall. Values less than zero are contoured in white.

their resources in improving observing and modeling systems.

Acknowledgments. We thank NOAA's Earth Systems Research Laboratory for support of this project, and access to their computer systems and data. The lead author's time was supported by Central Michigan University. This work was also supported by a NOAA grant to North Carolina State University. We thank NOAA/NCEP/CPC for their creation of the Unified Precipitation Dataset, and the U.S. Department of Energy for funding the creation of the Reforecast-2 dataset. We thank two anonymous reviewers for their valuable suggestions.

REFERENCES

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.,* **113,** D04110, doi:10.1029/2007JD009132.

Fritsch, M. J., 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, doi:10.1175/BAMS-85-7-955.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.,* **132,** 2905–2923, doi:10.1256/qj.06.25.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134,** 3209–3229, doi:10.1175/MWR3237.1.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.,* **136,** 2620–2632, doi:10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull. Amer. Meteor. Soc.,* **94,** 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

Jolliffe, I. T., and D. B. Stephenson, Eds, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. J. Wiley and Sons, 274 pp.

Lorenz, E. N., 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability,* Vol. I, Reading, United Kingdom, ECMWF, 1–19, doi:10.1017/cbo9780511617652.004.

Novak, D. R., C. Bailey, K. Brill, P. Burke, W. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting,* **29,** 489–504, doi:10.1175/WAF-D-13-00066.1.

Palmer, T. N., 2012: Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Quart. J. Roy. Meteor. Soc.,* **138,** 841–861, doi:10.1002/qj.1923.

Prat, O. P., and B. R. Nelson, 2014: Characteristics of annual, seasonal, and diurnal precipitation in the southeastern United States derived from long-term remotely sensed data. *Atmos. Res.,* **144,** 4–20, doi:10.1016/j.atmosres.2013.07.022.

Rasmussen, R. M., and Coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed. *Bull. Amer. Meteor. Soc.,* **93,** 811–829, doi:10.1175/BAMS-D-11-00052.1.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.,* **127,** 2473–2489, doi:10.1002/qj.49712757715.

Von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 676 pp.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.,* **60,** 1173–1185, doi:10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2.

——, A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting,* **21,** 149–166, doi:10.1175/WAF909.1.