The Application of a Simple Method for the Verification of Weather Forecasts and Seasonal Variations in Forecast Accuracy

ANTHONY R. LUPO AND PATRICK S. MARKET

Department of Atmospheric Sciences, University of Missouri—Columbia, Columbia, Missouri

11 June 2001 and 28 January 2002

ABSTRACT

The evaluation of weather forecast accuracy has always been a difficult subject to address for many reasons. In this study, a simple semiobjective method is used to examine the accuracy of zone forecasts issued by the Weldon Spring (Saint Louis) National Weather Service (NWS) Office for mid-Missouri over a period of 416 days with the goal of demonstrating the utility of this method. Zone forecasts were chosen because these forecasts are typically what the public will receive either directly or indirectly from various media outlets. Not surprising, the evaluation method used here demonstrates that forecasts issued by the NWS and the Nested Grid Model (NGM) model output statistics (MOS) represent a considerable improvement over persistence or climatological baseline forecasts. NWS forecasts were slightly better than NGM MOS forecasts, especially when considering temperature and precipitation only. All forecasts showed distinct seasonal variability. The NWS winter-season forecasts were superior to those issued in the summer season, and this superiority was found to be a function of the precipitation forecast parameter. This technique might represent an easily understandable and concise method for providing weather forecast performance information to the general public in such a way that it would instill or reinforce public confidence in the accuracy of weather forecasts.

1. Introduction

The evaluation of forecasts issued by the National Weather Service (NWS) and received by the public directly or indirectly (via radio or television) is a difficult subject to address. Internal forecast evaluations typically may be done routinely, but information regarding weather forecast accuracy is not readily available to the general public. There are several problems that need to be addressed with regard to evaluating weather forecasts. These problems include but are not limited to the following: which parameters (e.g., temperature or precipitation) or forecasts (e.g., coded cities forecasts or the zone forecasts) to evaluate, what methods to use, and how to reduce subjectivity in scoring the forecast. These are technical issues that have been answered using various methodologies (e.g., Maglaras 1998, 1999; Thornes and Proctor 1999, hereinafter TP99; Martner and Politovitch 1999), but there remain the problems of how to present this information to the public concisely and how the public perceives forecast accuracies. TP99 discuss the issue of public perception as being critical. For example, the methodology used by the latter references above could not readily be presented to a broad segment of the general public in an easily understandable way. Also, using a ''skill'' type score generally results in percentages that would be perceived as ''low,'' even if they showed considerable skill. This information can easily be misinterpreted by the general public and, thus, may not instill much confidence in weather forecasts. In converse, the general, *but mistaken,* perception by the public is that weather forecasts are routinely ''missed'' despite the dramatic improvements made in forecasting over the last 2–3 decades (e.g., Sanders 1986; Shuman 1989; Kalnay et al. 1990; Roebber and Bosart 1996; Roebber 1998). Thus, a skeptical public may not readily accept using an evaluation method that produces very high scores. Therefore, the TP99 methodology provides an ideal compromise for evaluating forecasts in a meaningful but simple manner for public consumption.

In this investigation, we will also examine the seasonal variation of forecast accuracy over the course of more than 400 days in central Missouri. It typically might be assumed that weather forecasts would be more accurate in the summer when the weather is less variable and would show less accuracy during the winter season. TP99 evaluated forecasts in England during the late spring and found that radio broadcast forecasts were significantly better than ''persistence'' (use of preceding values) or use of climatological values (''climatology'') but that there is room for improvement. They also found that weather forecasts did improve as the season pro-

Corresponding author address: Dr. Anthony R. Lupo, Dept. of Atmospheric Sciences, 112 Gentry Hall, University of Missouri—Columbia, Columbia, MO 65211.
E-mail: lupoa@missouri.edu

TABLE 1. University of Birmingham (UB) method (TP99) and the University of Missouri (UM) method, modified from TP99, for forecast evaluation.

Element	Forecast (UB)	Forecast (UM)	Score
Max temperature	±2°C	±2°F	2
	±4°C	±4°F	1
	>±4°C	>±4°F	0
Wind speed	±1 Beaufort category	Within forecast range	1
	>±1 Beaufort category	Outside forecast range	0
Wind direction	±45°	±45°	1
	>±45°	>±45°	0
State of the sky (sky cover)	Good guide	Good guide	2
	Fair guide	Fair guide	1
	Poor guide	Poor guide	0
Weather (precipitation)	Good guide	Good guide	2
	Fair guide	Fair guide	1
	Poor guide	Poor guide	0
Total possible points			8

gressed; however, persistence improves as well such that the skill of a persistence temperature forecast peaks in the autumn in the United Kingdom (e.g., Garner 1997).

Thus, this brief investigation has two goals. The first goal is to demonstrate the utility of a simple, understandable, and effective method for forecast evaluation. This method produces forecast scores that should be easily interpreted by the general public without degrading their confidence in the weather forecasts. This method also could be useful in dispelling the notion held by a segment of the general public that weather forecasts are routinely missed. The second goal is to examine the seasonal variation of weather forecasts in central Missouri over the course of one year. In this portion of the study, we wish to demonstrate that forecasts actually result in a greater degree of improvement with respect to persistence and climatology in the cold season when the weather is generally more variable than in the summer season.

2. Data sources, methodology, and experiment design

The methodology for evaluating the Columbia, Missouri, region weather forecasts was borrowed from TP99. This methodology was applied to zone forecasts generated by the Saint Louis NWS Weather Forecast Office (WFO), Nested Grid Model (NGM) model output statistics (MOS), persistence, and climatology. Zone forecasts were chosen because these forecasts are most directly or indirectly consumed by the public via weather radio or other media outlets. The NWS zone forecasts and MOS guidance were obtained from various online sources, but the primary daily source was the Texas A&M Department of Earth and Atmospheric Sciences Web site (www.met.tamu.edu/weather/weather.shtml). The observations used in verifying the forecasts and generating persistence were obtained using hourly METAR (a French acronym that can be translated as aviation routine weather report) reports from the Columbia Regional Airport (COU) that were obtained primarily from the source cited above. The climatology records were obtained from the Missouri Climate Center.

The TP99 method for evaluating weather forecasts is a point-scoring system based on a categorization of the difference between the forecast and observations (Table 1). Their method is itself borrowed from the Met Office forecast evaluation system. TP99 modified the method to include a wind direction category. This method is preferable for evaluating forecasts for public consumption, because such forecasts typically give a range of values for a certain forecast parameter, and the evaluation of some forecast parameters can be somewhat subjective. In using the TP99 method, a perfect forecast would score 8 points. In this study, the TP99 method was adjusted to accommodate forecasts generated in the United States (Table 1). Thus, the most substantive changes are the conversion of temperature to Fahrenheit and the evaluation of wind speed in units of miles per hour. A brief description of evaluating each parameter is given below.

In evaluating the temperature, NWS forecasts for maximum and minimum temperature are assigned an ''objective'' value based on a subjective interpretation of the forecast. For example, a forecast for temperatures in the mid-50°s was interpreted to define a range between 53° and 57°F. Thus, for verification purposes we translated the NWS forecast into a single number comparable to persistence, climatology, and NGM guidance. Then, the sample forecast above would be evaluated within the scheme in Table 1 (column 3) against observations using a value of 55°F, and the ±2°F range would represent a verified forecast conforming to a typical NWS forecast and would be assigned a value of 2 points. Persistence simply uses the values observed the previous day, and climatology uses the 30-yr daily average temperature for a particular date.

Forecast wind speeds are generally given as a range

of values. For example, it is common to see a forecast of winds of 5–10 mi h^{-1}. To evaluate the forecast in a similar manner to that of TP99 and given the variable nature of wind speed, the daily observed wind speed is averaged and then ''converted'' into a wind category. For example, if the wind speed averages 7 mi h^{-1} and was between 5 and 10 mi h^{-1} all day or most of day, the speed is assigned a value of 5–10 mi h^{-1}. Climatological values and NGM MOS guidance are converted in the same manner. Thus, if the forecast is for winds of 5–10 mi h^{-1} and the winds were 10–15 mi h^{-1} during the forecast period, a score of 0 is given for that particular forecast.

Wind directions were evaluated versus the mean observed wind direction over the 12-h period. If the forecast wind direction is within 45° of the observed mean, the forecast verifies. Wind directions were the most difficult parameter to evaluate, in that rapid shifts in the wind direction (e.g., frontal passage, convective outflow) had to be accounted for and a subjective judgment then had to be made to determine whether the forecast was ''correct'' or a ''miss.'' When wind speeds were less than 10 mi h^{-1}, variable winds were reported when the wind directions were scattered over more than one cardinal quadrant in an irregular manner. These forecasts were determined to be a missed forecast for each forecast category (e.g., NWS, persistence, climatology, and the NGM MOS guidance) unless the NWS and/or MOS actually forecast variable winds (for MOS, it can be determined whether winds are light and vary in an irregular manner). Last, mesoscale events such as thunderstorms can make wind direction verification difficult. These events (individual METAR hourly reports) were subjectively filtered out of the observed data for the wind direction parameter only; thus, no time periods or days were discarded from the dataset. These events did not hamper the process of determining the prevailing synoptic wind direction. Thunderstorms occurred on 49 days (defined as from 0000 to 0000 UTC next day), of which four thunderstorm events resulted in a ''double count'' of days because of having occurred through the 0000 UTC observation time. Forecast days were broken down into two 12-h forecast periods (described below), and thunderstorms were observed during 34 overnight periods and 23 next-day periods, of which an additional double count resulted from four events that occurred through the 1200 UTC observation time. Thunderstorm events resulted in an average of three–four hourly observations being removed as due to an ''event.'' No attempt was made to count special observations or to determine the number of thunderstorms affecting COU, because this information is not germane to this study.

Sky cover was evaluated subjectively by assigning a mean value using METAR reports and manual observations. Thus, this characteristic contained the most subjectivity in comparison with the others. However, under most circumstances, it would be expected that the cloud cover at COU would be similar to that across Boone County and, thus, to manual observers at the University of Missouri. Sky cover was evaluated to be either clear (0–0.2 coverage), scattered clouds (0.2–0.5), broken clouds (0.5–0.8), or overcast (0.8–1.0). This sky-cover evaluation scheme matches the categorization given for NGM MOS. Climatology was assigned a consistent value of 50% cloudiness, which is close to the mean annual value for first-order reporting stations in Missouri. For climatology, then, the cloudiness categories were modified to include a 0.4–0.6 cloudiness category to make the climatological value fit within the evaluation without automatically losing 1 point as in the precipitation category below. Then, if a forecast was one category (two categories) different from observed, the forecast evaluation was given 1 (0) points.

Precipitation forecasts were only evaluated as to whether precipitation fell at COU. NWS forecasts were given a score of 2 if the forecast was correct to within 30%. That is, if precipitation was not forecast (forecast) and no precipitation (precipitation) fell, then the NWS forecast was given 2 points. To remove subjectivity, a forecast percentage of 0%–29% (71%–100%) was included in the ''not-forecast'' (forecast) case. Forecasts giving a value of 30%–70% were automatically given 1 point. This method also was applied to NGM MOS forecasts. Climatology was automatically assigned 1 point daily, because rain might be expected to occur on about 30% of the days in any given month (assuming the synoptic period is 3–4 days for the passage of a cyclonic system). Persistence could either receive 0 or 2, with the exception of the recording of a ''trace'' of precipitation for the 12-h period.

The experimental design was straightforward. The 12- (overnight) and 24-h (next day) forecasts issued at 1600 central daylight time [1500 central standard time] were archived each day for more than a full year (20 November 1999–8 January 2001, or 416 days). During that time, the NGM MOS forecasts for the period were also archived. The observed conditions used for assessment in most categories were the mean conditions within the 0000–1200 UTC or the 1200–0000 UTC period. The minimum temperature applied to the first period. Thus, if the daily minimum occurred more than 1 h outside this period, it was discounted and the minimum in the 12-h period was used instead. This rule accounts for wintertime minima that may occur just outside the 1200 UTC time period but filters out days on which synoptic disturbances significantly mask the normal diurnal temperature variation. A similar rule was applied to the second period for the maximum temperature. According to the scoring system above, a perfect (missed, or ''busted'') forecast would receive 8 (0) points, with the exception of climatology, which could score no better (worse) than 7 (1) points, given that it automatically lost (gained) a point from the precipitation forecast.

TABLE 2. Mean forecast scores by (a) category and (b) as a percentage. These results are based on 416 forecasts (overnight/next day). The bottom row shows the results using temperature and precipitation variables only. Climatology precipitation was assigned as 1 point for this study.

(a)	NWS	Persistence	Climatology	MOS
Min/max temperature	1.47*/1.27*	0.71/0.67	0.51/0.50	1.37/1.19
Wind speed	0.79/0.65	0.62/0.56	0.53/0.59	0.84*/0.82*
Wind direction	0.80/0.87	0.46/0.44	0.43/0.43	0.87*/0.89*
Sky cover	1.75*/1.69*	1.36/1.35	0.91/1.02	1.66/1.59
Precipitation	1.68/1.62	1.31/1.34	1.00/1.00	1.70*/1.65*
Totals	6.54*,**/6.14**	4.48/4.40	3.41/3.57	6.44**/6.14*,**
Temperature + precipitation	3.15*/2.89*	2.02/2.01	1.51/1.50	3.06/2.85
(b)	NWS	Persistence	Climatology	MOS
Min/max temperature	73.5*/63.5*	35.5/33.5	25.5/25.0	68.5/59.5
Wind speed	79.0/65.0	62.0/56.0	53.0/59.0	84.0*/82.0*
Wind direction	80.0/87.0	46.0/44.0	43.0/43.0	87.0*/89.0*
Sky cover	87.5*/85.0*	68.0/67.5	45.5/51.0	83.0/79.5
Precipitation	84.0/81.0	65.5/67.0	50.0/50.0	85.0*/82.5*
Totals	81.3*/76.4	56.0/55.0	42.0/44.0	80.4/77.1*
Temperature + precipitation	78.9*/72.0*	50.5/50.3	37.8/37.5	76.5/71.3

* Performance is better between NWS and MOS.
** Forecast score is better than persistence and climatology at the 95% confidence interval.

3. Results and discussion

a. Cumulative results

The results of this forecast evaluation exercise for the 416-day period are shown in Table 2. Table 2a shows the average daily score for each forecast parameter and the average daily total forecast score (TF), and Table 2b shows the same numbers as a percent (TF%) of the perfect score for each parameter and the total forecast score (8 points):

$$\text{TF} = \text{temperature} + \text{wind speed} + \text{wind direction} + \text{sky cover} + \text{precipitation}, \quad \text{and} \tag{1a}$$

$$\text{TF\%} = (\text{TF}/8) \times 100\%. \tag{1b}$$

These numbers are straightforward to interpret and would be very useful for public consumption. These scores could be placed side by side with climatology and/or persistence to give the general public a sense of the skill of NWS forecasts in a direct and readily understandable way as is done in Tables 2 (see also Table 4, described later). However, one problem with the TP99 methodology presented here is that, without these side-by-side comparisons, the scores in Table 2 reveal little about the actual skill or inferred value of forecasts issued for a particular area. This issue would be problematic, particularly in locations where a forecast of persistence and/or climatology might perform well.

TABLE 3. Skill scores (NWS vs all others, %), where 0 is no skill and 100 is perfect. These results are based on 416 forecasts (overnight/next day). The bottom row shows the results using temperature and precipitation variables only. Climatology precipitation was assigned as 1 point for this study.

	Persistence	Climatology	MOS
Min/max temperature	58.9/44.9	64.5/51.3	15.9/9.5
Wind speed	44.7/20.5	55.3/14.6	−31.0/−94.4
Wind direction	63.0/76.3	64.9/76.7	−54.0/−21.0
Sky cover	60.9/52.3	77.1/68.4	26.5/24.4
Precipitation	53.6/42.4	68.0/62.0	−6.7/−8.6
Totals	57.9/48.1	67.5/57.6	4.5/0.0
Temperature + precipitation			9.6/3.6

Table 3 shows comparable skill scores, with the NWS forecast being compared with each other forecast issued (persistence, climatology, and NGM MOS). Forecast skill is defined as in TP99:

$$S = [(F - B)/(P - B)] \times 100.0\%, \tag{2}$$

where S is the skill score expressed as a percentage and F represents the forecast score (e.g., NWS forecast for a particular category or its total score) being compared with B, a "baseline" forecast score (e.g., climatology, persistence, or NGM MOS). In (2), P represents a perfect forecast score (8 points for the total score; 2 or 1 point for each category in Table 1). Thus, in using this method, it is possible to get skill scores of less than 0 if the baseline forecast is better than the forecast being compared. Also, the skill score represents the percentage (improvement) that a particular forecast has achieved over the baseline (100% represents a perfect score for the forecast). Thus, unlike the TP99 method presented here, S represents a real measure of forecast skill.

The skill scores are, in general, lower than those in Table 2, and there are negative numbers. These numbers might be more difficult for the general public to interpret and to understand easily. The overall forecast performances (Table 2) showed that both the NWS and NGM MOS forecasts were better for the overnight period than for the next day, which might be expected. Also both the NWS and NGM MOS forecasts represented a sub-

TABLE 4. The total number of perfect forecasts for the overnight/next-day period for the 416-day period of study. An 8 is a perfect forecast score, and a 0 is a forecast bust. The exception is climatology, which can only score as much as 7 or as few as 1 because climatology precipitation was assigned a score of 1.

	Perfect forecasts	Forecast busts
NWS	107/69	0/0
NGM	87/80	0/0
Persistence	21/20	7/4
Climatology	5/5	31/23

stantial improvement over either persistence or climatology, and the former forecasts were better across each forecast parameter than were the latter forecasts. Statistical testing of the mean total forecast scores (e.g., Neter et al. 1988, chapter 11) demonstrated that the NWS and NGM MOS scores were better than climatology and persistence forecasts, a result that was significant at the 95% confidence interval. The significance is similar to that found by TP99, who used different statistical testing methods to test the significance of their results. It should be noted here that all of the results were tested (and those that are significant are indicated in Tables 2 and 5), but not all of the results proved to be statistically significant. The lack of statistical significance, however, does not automatically preclude these results from representing meaningful or useful information (e.g., Nicholls 2001).

A more detailed comparison between the total scores for the NWS and NGM MOS forecasts revealed that the NWS forecasts were slightly better for the overnight period, whereas the NGM MOS forecasts were slightly better for the next-day forecast period. However, when comparisons were made using only temperature and precipitation as the two variables, the NWS was better for both forecast periods. The NWS temperature forecasts were an improvement on the model guidance by about 0.1 points or roughly 4–5 percentage points. The precipitation scores were closer, but the NGM MOS guidance possessed a very slight edge in this category. In the sky-cover category, the NWS scores were better; however, the NGM MOS scores were, in some cases, much better than the NWS scores in the wind speed and direction categories. The wind categories are perhaps the most difficult to forecast, and the observations that serve as the verification may be greatly affected by small- or local-scale effects that are site dependent.

An examination of the number of perfect (8 points) and busted (0 points) forecasts (Table 4) shows that the NWS issued a perfect forecast more often than did persistence, climatology, or NGM MOS. For the overnight forecast period, NWS forecasts were perfect 26% of the time; they were perfect 21% overall. The NGM MOS forecasts were perfect 19% of the time for the next-day forecast period and were perfect 20% overall. In terms of busted forecasts, climatology performed the worst, as 6.5% of the total number of forecasts were busts (which for climatology was a 1, given that this forecast gets 1 automatic point for precipitation). The NWS and the NGM MOS issued no forecasts evaluated to be a busted forecast over the 416-day period, and, in fact, the NWS and NGM forecasts scored 1 point during only one period each during the same 416-day period.

b. Seasonal results and variations

A seasonal breakdown of the forecast evaluations is given in Table 5, which presents the average total forecast score and the standard deviation for each season. Each season is defined in the conventional sense in which December–February, March–May, June–August, and September–November represented the winter, spring, summer, and autumn seasons, respectively.

A seasonal breakdown of the NWS forecast performances (Table 5) demonstrates that these forecasts were best during the autumn and winter seasons and that each season was comparable to the other in terms of the variability in forecast scores. The NWS spring and summer forecasts also were comparable, but these scores were lower than the autumn and winter forecast scores. The forecast scores were also generally more variable in the spring and summer season, as the higher standard deviations in those seasons would suggest. These results are different than those of the TP99 study, which showed

TABLE 5. Forecast performance by season for each forecast method (mean score/std dev). Overnight (12-h) and next-day (24-h) forecasts are given.

	Autumn	Winter	Spring	Summer
NWS morning	6.70**/0.64	6.66*/0.66	6.35*/0.71	6.39**/0.79
NWS afternoon	6.47*/0.67	6.11*/0.65	6.04*/0.58	5.92**/0.68
NGM morning	6.67**/0.94	6.24*/0.90	6.36*/0.73	6.57**/0.88
NGM afternoon	6.24**/0.93	6.13*/1.08	6.27*/1.12	5.92**/0.97
Persistence morning	4.96**/1.27	4.02/0.99	4.03/1.03	5.05**/1.21
Persistence afternoon	4.66/1.14	3.98/0.79	4.29/1.08	4.79/1.26
Climatology morning	3.32/0.79	3.34/0.60	3.32/0.70	3.56/0.82
Climatology afternoon	3.42/0.85	3.31/0.69	3.74/0.59	3.92/1.04

* Forecast is better than both climatology and persistence at the 95% confidence interval.
** Forecast is better than climatology at the 95% confidence interval.

increasing scores for British radio forecasts as spring progressed into summer. An explanation for these seasonal differences can be accounted for by examining the precipitation parameter. Zone forecasts typically may issue a forecast stating that there is a 30%, 40%, or 50% probability of precipitation during situations that might favor convective activity, because in the spring and especially the summer seasons, precipitation is predominantly convective in nature. By the scoring system devised here, an automatic score of 1 is assigned for the score of the precipitation forecasts above. During the summer season there was a noticeable degradation of average precipitation scores for NWS forecasts (in winter, 1.83 for morning and 1.77 for afternoon vs in summer, 1.48 for morning and 1.33 for afternoon). None of the other NWS forecast parameters (not shown) showed any distinct seasonal variations.

Seasonal variations in the NGM MOS forecast scores were less apparent than for the NWS forecast scores. For the NGM MOS, summer and autumn 12-h forecasts were superior to winter and spring forecasts, and autumn and winter forecasts showed more variation. The 12–24-h NGM MOS forecasts were best during the transition seasons and worst during the summer season but were more variable in the winter and spring seasons. Persistence morning and afternoon forecasts were better during the summer and autumn seasons, but also the scores were more variable. Some insight into these seasonal variations might be gained by examining the variations in each NGM MOS forecast parameter. The NGM MOS winter-temperature-forecast mean scores were 0.33 points lower than the summer-forecast mean scores, which were the best when compared with the other seasons. However, NGM MOS summer-forecast point totals were the lowest because of that season scoring worse than the other seasons in the mean precipitation and mean wind direction categories by a combined margin of about 0.3 points.

Persistence might be expected to perform the best during the summer season (Table 5) for the total score because the day-to-day variations in the weather tend to be smaller during this season, with the exception of precipitation (see above discussion of NWS scores). TP99 demonstrated that persistence did perform better in the latter part of their study, which was during the late spring. Climatology might also be expected to perform better during the warm season when the day-to-day excursions from normal might also be expected to be smaller. Climatology did perform best during the summer period here, but the winter climatology forecast scores were less variable. The monthly temperature anomalies were larger for the winter seasons than for the summer season covered during the 416-day period here [mean absolute departure for winter months was 7.6°F (4.2°C) vs 2.2°F (1.2°C) for summer months]. Although the monthly temperature anomalies may not be a perfect indicator of day-to-day forecast performance, these may provide a partial explanation nonetheless. For example, a very warm winter month (with respect to normal) might suggest a poor performance by climatology for both temperatures and wind direction.

TABLE 6. As in Table 4, but stratified by season.

Perfect forecasts	Autumn	Winter	Spring	Summer
NWS	16/14	28/16	27/15	21/10
NGM	28/20	17/19	21/27	21/14
Persistence	5/6	2/1	2/4	9/9
Climatology	4/4	2/0	0/0	2/4
Busted forecasts	**Autumn**	**Winter**	**Spring**	**Summer**
NWS	0/0	0/0	0/0	0/0
NGM	0/0	0/0	0/0	0/0
Persistence	2/0	2/2	2/1	1/0
Climatology	8/5	5/4	10/7	3/4

An examination of the number of perfect and busted forecasts generally supports the discussion of the persistence and climatological forecast scores described above. Table 6 shows the number for perfect and busted forecasts normalized to 100 days to account for the unequal number of days that contributed to each season in the study. Persistence and climatology forecasts had more perfect days and fewer busted days in the summer season. NGM MOS perfect forecast scores do not have an apparent relationship to forecast performance or variations, and more of the NWS perfect forecasts were recorded in the winter and spring seasons.

4. Summary and conclusions

A simple study of the accuracy of zone forecasts issued by the Weldon Spring (Saint Louis) NWS WFO for central Missouri was done for a 416-day period from 20 November 1999 to 8 January 2001. Zone forecasts were chosen because these forecasts are most directly or indirectly consumed by the general public via radio or other media outlets. The overnight (0000–1200 UTC) and next-day (1200–2400 UTC) forecasts were examined to compare the NWS forecasts with climatology and persistence. Also, these forecasts were evaluated to examine seasonal variations in forecast accuracy. The method used to evaluate the forecasts was a simple point-scoring scheme used by TP99, who evaluated British radio forecasts over a spring season. The parameters included were temperature, wind speed and direction, sky cover, and precipitation. This scoring system was modified to make the scoring system compatible with forecasts issued in the United States.

These results show that the total scores for NWS and NGM MOS forecasts routinely scored better than climatology and persistence forecasts and that these results were significant at the 95% confidence interval when applying a simple statistical test, a result consistent with those of TP99. Forecasts for the overnight period were better than those issued for the next afternoon for both

the NWS and NGM MOS. NWS and NGM MOS performances were comparable, with NWS scoring better for the overnight period and NGM MOS scoring better for the next-day forecast period. The better scores for NGM MOS forecasts were primarily a result of better scoring for the wind speed and direction parameters. When temperature and precipitation were the only parameters used, the NWS forecasts were better for both periods. The NWS issued more forecasts scored to be perfect than did NGM MOS; neither issued a forecast scored to be busted. Climatology forecasts were most often scored as busts, but only 6.5% of the time.

An examination of the seasonal variations in the forecasts revealed that NWS forecasts were better and that the scores showed less variation during the autumn and winter seasons than in spring and summer. This could primarily be attributed to the scoring of the precipitation parameter, which was lower during the primary convective season. Convective precipitation is generally more difficult to forecast, and, in many forecast situations, the NWS forecast was evaluated to score an automatic 1. NGM MOS forecasts showed no apparent systematic seasonal variation in the total score, and seasonal variations in individual forecast parameters (e.g., winds, sky cover, or temperature) confirm this assertion, with some parameters scoring better in some seasons than in others. Climatology and persistence forecasts were best during the summer season, but still, in general, did not score nearly as well as the NWS or NGM MOS forecasts, which were a significant improvement over these baselines. The seasonal variations in the number of perfect and busted forecasts generally mirrored the variations in total score.

One goal of this work was to present a simple forecast evaluation method that would be easy to understand and to assess in an efficient manner for use by the general public. Also, this forecast evaluation procedure would provide any public or private (e.g., television or radio stations) weather forecasting entity with a methodology for compiling meaningful forecast evaluations and publishing the results. This is especially true because these forecast performance numbers would be published for direct comparison alongside baseline forecast performances derived from persistence, climatology, or some other baseline standard chosen. The numbers used here were simple forecast scores and percentages. The method presented here is easier to understand and interpret than skill scores, which typically yield relatively low scores. Without an understanding by the public of what these skill scores represent, the lower scores could be misleading. The percentages given using the method presented here would provide a standard evaluation that would also be helpful in dispelling the generally held, but mistaken, notion that weather forecasts are routinely missed or are inaccurate.

Acknowledgments. The authors thank Ron Przybylinski, Science Operations Officer, and Eric Lenning at the Saint Louis National Weather Service Office for providing missing or incomplete data when such data were requested. We also thank the three anonymous reviewers for their time and effort in providing very helpful comments regarding this manuscript.

REFERENCES

Garner, D. A., 1997: The seasonal variation of the accuracy of weather forecasts using climatology and persistence at Birmingham. M.S. dissertation, School of Geography, University of Birmingham, 93 pp.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428.

Maglaras, G. J., 1998: Verification trends at the Albany Forecast Office continue to show improvement on MOS guidance. *Natl. Wea. Dig.,* **22** (2), 9–14.

——, 1999: Temperature and precipitation forecast verification trends at the Albany Forecast Office: Forecasters continue to show improvement on MOS guidance—Part II. *Natl. Wea. Dig.,* **23** (2), 3–12.

Martner, B. E., and M. K. Politovich, 1999: Five-day temperature forecasts from Denver television stations and newspapers. *Natl. Wea. Dig.,* **23** (2), 9–20.

Neter, J., W. Wasserman, and G. A. Whitmore, 1988: *Applied Statistics*. 3d ed. Allyn and Bacon Press.

Nicholls, N., 2001: The insignificance of significance testing. *Bull. Amer. Meteor. Soc.,* **82,** 981–986.

Roebber, P. J., 1998: The regime dependence of degree day forecast technique, skill, and value. *Wea. Forecasting,* **13,** 783–794.

——, and L. F. Bosart, 1996: The contributions of education and experience to forecast skill. *Wea. Forecasting,* **11,** 21–40.

Sanders, F., 1986: Trends in skill of Boston forecasts made at MIT, 1966–1984. *Bull. Amer. Meteor. Soc.,* **67,** 170–176.

Shuman, F. G., 1989: History of numerical weather prediction at the National Meteorological Center. *Wea. Forecasting,* **4,** 286–296.

Thornes, J. E., and E. A. J. Proctor, 1999: Persisting with persistence: The verification of Radio 4 weather forecasts. *Weather,* **54,** 311–320.