3.4 EXPERIENCES WITH A PROBABILISTIC FORECASTING GAME FOR METEOROLOGY STUDENTS

Steven G. Decker *
Rutgers, The State University of New Jersey, New Brunswick, NJ

1. INTRODUCTION

Despite increasingly frequent calls for better communication of weather forecast uncertainty (e.g., Morss et al. 2008), exposure of students to probabilistic forecasting may still be the exception rather than the rule. Aside from a few probabilistic contests described later, many students engage in contests, in-class or otherwise, that are deterministic. For example, the WxChallenge, an intercollegiate weather forecasting contest with over 1000 student participants, is wholly deterministic in format (Illston et al. 2009). Additionally, students examining weather forecasts from the NWS obtain primarily deterministic information courtesy of the National Digital Forecast Database (Glahn and Ruth 2003).

In an effort to introduce students to issues related to the communication of weather uncertainty, the New Brunswick Forecasting Game (NBFG) at Rutgers University has been revised to become fully probabilistic. In addition to providing students with practice making probabilistic forecasts, the scoring system itself introduces students to concepts of forecast quality like reliability and sharpness. Furthermore, students can explore the dataset formed over the course of the year to create visual measures of forecast performance such as reliability diagrams and relative operating characteristic (ROC) curves.

This abstract outlines the new NBFG and shows how the database of past forecasts can be mined to introduce concepts like the ROC curve to students.

2. THE NEW BRUNSWICK FORECASTING GAME

2.1 OVERVIEW

The current NBFG is designed as follows. During the fall semester, students make forecasts once a week (i.e., one "forecast window") for four periods for two locations. The four periods run from either 00 to 12 UTC or 12 to 00 UTC and can be thought of as tonight, tomorrow, tomorrow night, and the day after tomorrow. In the spring, students make forecasts twice a week, but for only two periods, tonight and tomorrow. One of the two locations is always the Rutgers Gardens weather station located in New Brunswick, New Jersey, roughly 1.5 km east of the meteorology classrooms. The instructor chooses the other location to maximize forecast difficulty. This is referred to as the Alternate Location. Students are charged with predicting temperature intervals and the probability of precipitation (POP) for each period and location. For night (day) periods, the low (high) temperature is predicted. Participants must submit forecasts by 00 UTC.

All NBFG forecasts made during the four semesters between the spring of 2010 and fall of 2011 inclusive and their corresponding verifications have been stored in a database. As a result, 6688 forecasts for both precipitation and temperature are available for the analysis undertaken in subsequent sections.

2.2 PRECIPITATION SCORE

A previous incarnation of the NBFG (Croft and Milutinovic 1991) also employed probabilistic precipitation forecasts. In fact, POPs were provided for a variety of accumulated precipitation ranges. The ranked probability score (Wilks 1995, 7.4.8) was used to measure forecast performance, but with a distance weighting factor.

To limit forecasting burden on the students, the precipitation portion of the current NBFG was simplified to one POP for each location and forecast period. In addition to being simple, this approach allows for the straightforward construction of attributes and reliability diagrams and ROC curves.

POPs are provided to the nearest 10% and are assessed using the following half-Brier score:

$$E_P = 10(p - o)^2, \qquad (1)$$

where p is the POP expressed as a number between zero and one, o is the observed

* *Corresponding author address:* Steven G. Decker, Dept. of Environmental Sci., 14 College Farm Rd., New Brunswick, NJ 08901; e-mail: decker@envsci.rutgers.edu

probability of precipitation (either 0 or 1), and E_P represents the error points attributed to precipitation incurred by the forecaster. Appropriate E_P values are then summed for each forecast–observation pair contributed by that forecaster to arrive at the forecaster's overall precipitation error score. With this score, a forecast POP of 50% generates 2.5 error points regardless of the verification. A perfect deterministic forecast of 0% or 100% generates no error points, but an imperfect deterministic forecast (i.e., forecasting 0% when it does precipitate) generates 10 error points. Despite the instructor's admonitions not to be overconfident, students do construct 10-error-point forecasts on a regular basis, as the reliability diagrams will show.

2.3 TEMPERATURE SCORE

The NBFG temperature interval score is presented in the following way. Let l be the forecast lower bound for temperature, u be the forecast upper bound for temperature, and T be the observed temperature. Temperature error points are given by $E_T = E_S + E_R$, where E_S is given by $u - l$, which represents the *sharpness* of the temperature forecast, and E_R is given by

$$E_R = \begin{cases} 4(l - T) & T < l \\ 0 & l \le T \le u \\ 4(T - u) & T > u \end{cases}, \qquad (2)$$

which represents the *reliability* of the temperature forecast. Gneiting and Raftery (2007) show that this scoring rule is proper for a 50% central credible interval forecast (Murphy and Winkler 1974), and Hamill and Wilks (1995) use a similar rule.

Just as for precipitation, a perfect score of zero is achievable with a deterministic temperature forecast ($l = u$) that verifies. A wider interval increases the sharpness score, but also increases the chance that the verifying temperature will be within the interval, thus lowering the expected reliability score. It may seem that the best forecast strategy is to choose an interval such that the temperature occurs within it 50% of the time, and above or below it 25% of the time each. However, because of rounding, temperatures that are just outside the interval will not be penalized by the reliability score. For example, suppose the true high temperature was 72.4°F, and the forecast upper bound was 72°F. Then the reliability score is still zero because the high temperature will be reported as 72°F, whereas the true reliability score should be 1.6 in this case. Because of these considerations, the best results will be achieved when the verifying temperature falls within the interval at a rate that is somewhat larger than 50%; this will be confirmed empirically later.

3. PRECIPITATION FORECAST PERFORMANCE

The attributes diagram (Wilks 1995, 7.4.4) is a standard way to assess the quality of probabilistic forecasts for dichotomous events. Figure 1 presents the attributes diagram for all of the precipitation forecasts in the NBFG database. The attributes diagram presents information regarding the reliability and skill of the set of forecasts as a whole. In Fig. 1, the diagonal dashed line represents perfect reliability; the x-axis is the forecast, and the y-axis is the frequency with which the event occurs given a forecast of x. From the diagram, we can see that POPs of 70–90% are quite reliable in this dataset. The distribution of forecast POPs shows that the forecasts are relatively sharp (many 0% and 100% forecasts), but some of this sharpness comes at the expense of reliability. It actually precipitates 5% of the time when the forecast is 0%, and does not precipitate 6% of the time when the forecast is 100%. Except for the highest POPs, NBFG forecasters have a systemic underforecasting bias, as the observed relative frequencies tend to be above the perfect reliability line. This is most pronounced with the 30% POP; when that forecast is made, it precipitates 47% of the time.

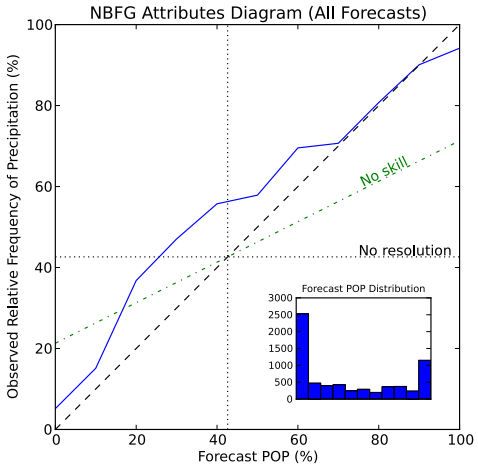

FIG. 1. Attributes diagram derived from precipitation forecasts made over four semesters (Spring 2010–Fall 2011) as part of the New Brunswick Forecasting Game.

NBFG forecasters can use this information to calibrate their forecasts. Given the data shown in Fig. 1, a first step toward that process would be to have all forecasters increase their forecasts by 10% when they are thinking of a POP between 20 and 60%. An unanswered question is whether the availability of the attributes diagram in real-time for a particular semester results in increased forecast reliability.

Figure 1 also indicates the degree of resolution and skill relative to climatology exhibited by NBFG forecasters. For the most part, forecasts display sizeable resolution, although the nearly horizontal lines between POPs of 40% and 50% and between 60% and 70% indicate somewhat less ability to resolve those forecast scenarios into separate events. All forecasts except those in the 20–40% range contribute to positive skill with respect to the NBFG climatology. Positive skill is provided when the reliability curve is below (above) the "no skill" line when the "no skill" line is below (above) the "no resolution" line (Wilks 1995).

Examining subsets of the NBFG database reveal interesting patterns in forecast reliability (Fig. 2). For example, when only forecasts for New Brunswick are considered, there is an extremely large overforecasting bias. A 50% POP leads to precipitation only 12% of the time! In part, this is likely to be an artifact of the fact that traces of precipitation at New Brunswick are not considered precipitation, and students are not able to calibrate their forecasts to this definition. Other factors that are leading to this overforecast bias are unclear [Schwartz's admonition? (Bosart 1983)], but the fact that this bias can be pointed out to students is important in and of itself.

In contrast, Fig. 2 shows that forecasts for the intentionally difficult Alternate Location are pervasively underforecast. Forecasts for the Alternate Location are considerably less sharp than those for New Brunswick as well (compare the forecast distributions in the lower right of Fig. 2), yet even with the reduced sharpness there is still very noticeable overconfidence in predicting a 0% POP for the Alternate Location.

Forecasts for Day 1 (tonight and tomorrow) mirror the overall reliability. Two interesting differences emerge when comparing Day 1 forecasts to Day 2 forecasts, however. First, there is a clear reduction in the sharpness of the Day 2 forecasts, as evidenced particularly by the greatly reduced tendency to forecast a POP of 100%. The second difference is seen in the greater amount of underforecasting that occurs for midrange (especially 30–60%) POPs in the Day 2 forecasts.

An alternative way of depicting resolution is to condition the forecasts based on what happened. Given that precipitation did not occur, how often was the forecast wrong? This is the probability of false detection (POFD), or false alarm rate. Given that precipitation did occur, how often was the forecast right? This is the probability of detection (POD), or hit rate. By applying thresholds to a set of POP forecasts, the ROC curve can be constructed (Mason and Graham 1999). Because the NBFG POPs are multiples of 10%, it is natural to use thresholds of 5%, 15%, …, 95% to construct the ROC curve, and this is what is done to create Fig. 3. When thresholds are very high, no

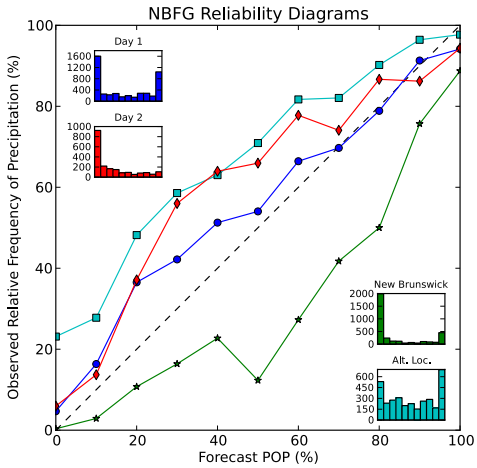

FIG. 2. Reliability diagrams for four subsets of the New Brunswick Forecasting Game database. Day 1 (blue circles), Day 2 (red diamonds), New Brunswick (green stars), and Alternate Location (green stars) forecasts are shown.

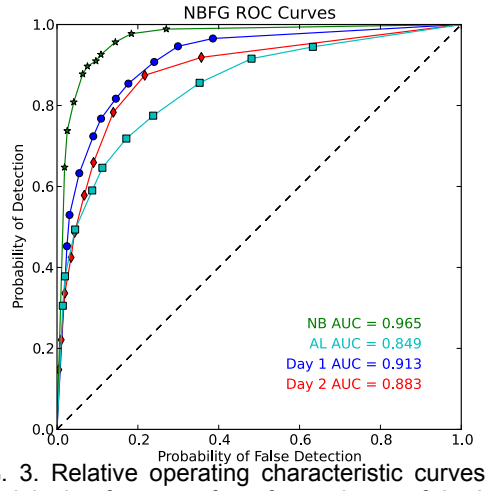

FIG. 3. Relative operating characteristic curves for precipitation forecasts from four subsets of the New Brunswick Forecasting Game database. Day 1 (blue circles), Day 2 (red diamonds), New Brunswick (NB; green stars), and Alternate Location (AL; green stars) forecasts are shown. The area under each curve (AUC) is provided in the lower right.

forecasts for precipitation are issued, so both the POD and POFD are zero. When thresholds are very low, each forecast is treated as if precipitation will occur, which necessitates that the POD and POFD be one. A skilled forecaster will be able to detect precipitation without issuing many false alarms. This is represented by points in the upper-left half of the ROC diagram. Hence, the area under a ROC curve (AUC) is a common forecast quality metric, with 1 being a perfect score, and 0.5 representing no skill.

Reassuringly for NBFG forecasters, Fig. 3 shows that the ROC curves for various subsets of the NBFG database all indicate significant skill. It is not surprising that New Brunswick forecasts are of much higher quality than their Alternate Location counterparts are, as this is by design. The reduction in skill between Day 1 and Day 2 forecasts reflects the increased uncertainty associated with a longer-range forecast. However, the difficulty of Day 2 relative to Day 1 is evidently less than the difficulty of the Alternate Location relative to New Brunswick.

4. TEMPERATURE FORECAST PERFORMANCE

Turning our attention to the temperature interval forecasts, Fig. 4 shows that the inverse relationship between sharpness and reliability posited earlier exists within the NBFG. As forecast sharpness increases (i.e., the sharpness score decreases), the reliability has a tendency to decrease (reliability scores increase). This tendency is revealed most easily by comparing Forecaster 1 (the best temperature forecaster) to Forecaster N (the worst). Forecaster 1 and other high-ranking forecasters had a roughly even mix of sharpness and reliability error points, with a small shift toward more sharpness points. Forecaster N and other poor performers had narrow intervals, leading to good sharpness scores but poor reliability scores. A few forecasters ran counter to the overall trend, particularly Forecasters C and H. Those forecasters were quite reliable, but at the cost of wide temperature intervals and hence large sharpness scores.

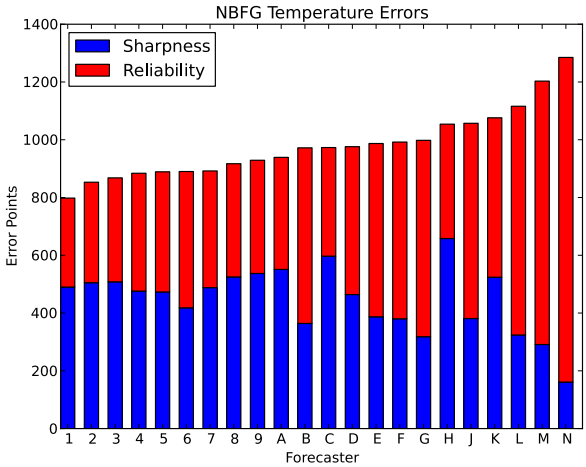

FIG. 4. Distribution of temperature error points assigned to each of the 22 forecasters participating in the New Brunswick Forecasting Game during the Fall 2011 semester. Forecasters are sorted in order of temperature error points, with Forecaster 1 achieving the lowest score, and Forecaster N recording the highest score. Blue/bottom (red/top) regions represent the portion of the temperature error attributed to lack of forecast sharpness (reliability).

An alternate way of depicting the reliability of NBFG temperature forecasts is to assess the frequency with which the observed temperature falls below, within, or above the forecast interval. Figure 5 displays such an analysis. As was suggested by Fig. 4, there is a tendency for forecasters at the bottom of the standings to perform poorly because of overconfidence. Temperatures rarely fall within their intervals, so their bars in Fig. 5 are relatively narrow. Only Forecaster H (and to some extent Forecaster C) bucks this trend.

Figure 5 also reveals that rounding indeed increases the credible forecast interval width beyond 50%. All of the top 10 temperature

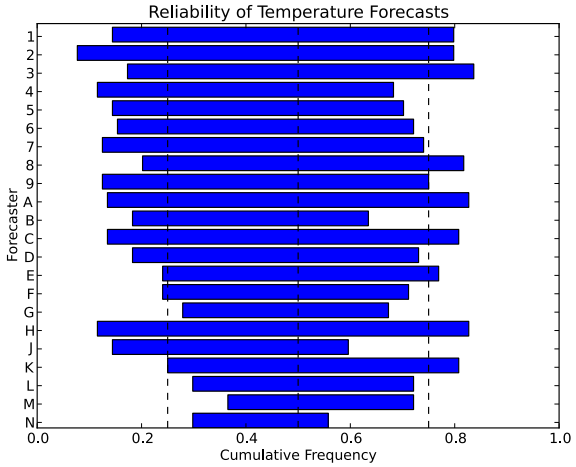

FIG. 5. Cumulative proportion of temperature forecasts that verified below the forecast range (left of blue bar), within the forecast range (blue bar), and above the forecast range (right of blue bar). Dashed lines indicate the 0.25, 0.5, and 0.75 cumulative frequencies. Forecasters are identified in the same manner as in Fig. 4.

forecasters had observed temperatures fall within their intervals more than 50% of the time. The best forecasters saw temperatures within their intervals 60–65% of the time. The degree to which the credible interval exceeds 50% depends on the variance of the forecast temperature distribution. However, the temperature variance will itself vary by location and season, which makes it difficult to determine *a priori* the credible interval a forecaster should aim for.

Finally, Figure 5 shows that most NBFG forecasters exhibited a cold bias, as it is more likely for temperatures to verify above the forecast interval than below it. Further analysis indicates that this cold bias is present in all semesters for both New Brunswick and the Alternate Location, but the source of this bias is unclear.

5. DISCUSSION

In any forecast contest, it becomes necessary to determine how the various elements of the forecast combine into an overall score. Many contests deal with this issue using skill scores (e.g., Bosart 1983, Sanders 1986, Gyakum 1986, Hamill and Wilks 1995, Newman 2003), but skill scores are not proper in general and hence may be gamed (Gneiting and Raftery 2007). The NBFG currently uses the simple approach of scaling the precipitation error score by an additional factor of six. This tends to weight precipitation and temperature errors roughly equally.

Students have been asked on end-of-semester course evaluations to respond to Likert items such as “I enjoyed the NBFG” and “I enjoyed the WxChallenge” with a response of one (five) indicating strong disagreement (agreement). Responses to these questions were collected during a period before the NBFG was revised (Fall 2007–Spring 2009) and during a period where the NBFG was fully probabilistic (Spring 2011). Table 1 provides the results. Despite the small sample size for the probabilistic NBFG, a few conclusions can be drawn. First, students appear to enjoy the probabilistic NBFG and deterministic WxChallenge equally (mean 4.46 for each). Thus, at the easily quantifiable level, there is no evidence to suggest that students prefer probabilistic to deterministic forecasting. On the other hand, there has been a notable increase in the enjoyment of the NBFG after instituting the current contest. Note that WxChallenge enjoyment increased as well; perhaps students in the spring of 2011 simply liked to forecast. However, the increase of over a point in students’ enjoyment of the NBFG is fairly significant even with the correlation between NBFG and WxChallenge ratings taken into account (p-value 0.061 for a one-sided *t*-test).

While quantitative results shed some light on students’ perceptions of these differing forecast contests, the “other comments” section of the course evaluations can provide windows into individual students’ experiences. Many students left that section blank, but the following conclusions may be drawn from those who did not. Generally, students had both positive and negative perceptions of the probabilistic NBFG. On the one hand, it seems more relevant to real-word forecasting and allows students to understand what probabilistic forecasting is more fully. On the other hand, it can be more time consuming than the WxChallenge.

6. CONCLUSIONS

The New Brunswick Forecasting Game at Rutgers University has been revised from a deterministic to probabilistic contest, and forecasts and observations from recent semesters have been stored in a database to allow for long-term verification. Participants forecast high and low temperature intervals and POPs. Course evaluations suggest that students find these changes have made the NBFG more enjoyable, but that level of enjoyment is no higher than that found for the deterministic WxChallenge. Students’ forecasts are skillful and reasonably reliable taken together, although notable biases have been found in precipitation forecasts depending on the forecast location, and temperature forecasts have a cold bias. Thus, there is room for students to improve their performance.

The results of this study will be shared with students in future semesters as a way to motivate understanding the abstract concepts behind ROC curves and attributes diagrams. Additionally, students will be able to use knowledge of the biases found by this study to improve their forecasts, at least in theory. Whether this will happen in practice remains an open question. In a few years, answering that question may become tenable, but a perfectly controlled experiment will not be possible.

Additionally, it is planned to provide figures like those shown in this study to students in real time as forecast windows are scored. The combination of attributes diagrams for the class as a whole with attributes diagrams that apply to each student individually may allow for students to calibrate their forecasts more completely.

REFERENCES

Bosart, L. F., 1983: An update on trends in skill of daily forecasts of temperature and precipitation at the State University of New York at Albany. *Bull. Amer. Meteor. Soc.,* **64,** 346–354.

Croft, P. J., and J. D. Milutinovic, 1991: The Rutgers University forecasting contest: Forecaster performance versus model guidance. *Natl. Wea. Dig.,* **16,** 2–12.

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.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378.

Gyakum, J. R., 1986: Experiments in temperature and precipitation forecasting for Illinois. *Wea. Forecasting,* **1,** 77–88.

Hamill, T. M., and D. S. Wilks, 1995: A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting,* **10,** 620–631.

Illston, B. G., J. B. Basara, M. Voss, and C. C. Weiss, 2009: An overview of the WxChallenge forecasting competition and its use as an educational tool. *Extended Abstracts, 18th Symp. on Education,* Phoenix, AZ, Amer. Meteor. Soc., 4.4.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting,* **14,** 713–725.

Morss, R. E., J. K. Lazo, B. G. Brown, H. E. Brooks, P. T. Ganderton, and B. N. Mills, 2008: Societal and economic research and applications for weather forecasts: Priorities for the North American THORPEX program. *Bull. Amer. Meteor. Soc.,* **89,** 335–346.

Murphy, A. H., and R. L. Winkler, 1974: Credible interval temperature forecasting: Some experimental results. *Mon. Wea. Rev.,* **102,** 784–794.

Newman, S. B., 2003: Encouraging student involvement through use of basic probability forecasting games. *Extended Abstracts, 12th Symp. on Education,* Long Beach, CA, Amer. Meteor. Soc., P2.5.

Sanders, F., 1986: Trends in skill of Boston forecasts made at MIT, 1966–84. *Bull. Amer. Meteor. Soc.,* **67,** 170–176.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.