The Impact of Writing Area Forecast Discussions on Student Forecaster Performance

PATRICK S. MARKET

Department of Soil, Environmental, and Atmospheric Sciences, University of Missouri—Columbia, Columbia, Missouri

(Manuscript received 24 June 2004, in final form 9 August 2005)

ABSTRACT

A brief study is provided on the forecast performance of students who write a mock area forecast discussion (AFD) on a weekly basis. Student performance was tracked for one semester (11 weeks) during the University of Missouri—Columbia's local weather forecast game. The hypothesis posed is that student performance is no better on days when they compose an AFD. A nonparametric Mann–Whitney test cannot reject that hypothesis. However, the same test employed on precipitation forecasts (for days when precipitation actually fell) shows that there is a statistically significant difference ($p = 0.02$) between the scores of those students writing an AFD and those who do not. Similar results are found with a chi-square test. Thus, AFD writers improve their precipitation score on days when significant weather occurred. Forecaster confidence is also enhanced by AFD composition.

1. Introduction

The area forecast discussion (AFD) written by forecasters of the National Weather Service (NWS) has long been a source of critical information for the larger forecasting community. These messages allow readers to understand better not only the forecast, but also the meteorologist's thinking in the creation of the final forecast product. For some users outside of the NWS, an AFD can be a starting point for their own forecast, an additional means of guidance for their own forecasting efforts. For others, it is consulted near the end of their own forecasting process as a kind of "check" on their thinking. In any case, AFDs serve to focus a forecaster's attention on her or his problem of the day, both for the NWS author as well as the external user.

During the spring 2004 semester, student forecast performance was tracked in the local forecast game in the Atmospheric Science Program (ASP) at the University of Missouri—Columbia (UMC). The specific sample consisted of 20 students (mostly seniors) in the writing-intensive course, Daily Analysis and Forecast Interpretation. Among other assignments, a mock AFD was required of the students (usually written in groups of two) each week. This exercise is an informal writing assignment designed to foster critical thinking as discussed by Bean (1996). The resulting messages were expected to take the form of and possess similar content to a typical NWS AFD. The final product had to be sent by electronic mail to the author each day by 1700 local time.

Previous studies examine the various aspects of collegiate forecasting contests. The influence of one's proximity to her or his forecasting site is shown to give those close to the site a slight advantage over more distant forecasters (Roebber et al. 1996). However, they also conclude that the proximity advantage is also a function of forecaster experience, a factor that is explored separately by Roebber and Bosart (1996). Others use results from similar forecast contest activities to show that a blended model output statistic product can outperform the vast majority (97%) of human forecasters (Vislocky and Fritsch 1997). Yet, each of these studies focuses on the characteristics of the forecasts produced among groups of human forecasters, or how they compare to the performance of increasingly sophisticated methods of automated forecasting. No previous study has examined the influence of *how* the forecast is

Corresponding author address: Patrick S. Market, Dept. of Soil, Environmental, and Atmospheric Sciences, University of Missouri—Columbia, 331 Anheuser-Busch Natural Resources Bldg., Columbia, MO 65211.
E-mail: marketP@missouri.edu

© 2006 American Meteorological Society

Unauthenticated | Downloaded 06/18/21 03:09 PM UTC

TABLE 1. Precipitation amount categories for the UMC ASP's local weather forecasting game. For each category that a participant is off by, 5 penalty points are assigned.

Category	Precipitation range (in.)
0	None
1	Trace–0.05
2	0.06–0.24
3	0.25–0.49
4	0.50–0.99
5	≥1.00

prepared. To be sure, writing a discussion is only one facet of a NWS forecaster's daily preparation, but the AFD is a well-known and widely read product even outside the NWS community.

This note presents a brief statistical analysis of the students' performance in the local forecast game on days when they composed an AFD compared to those days when they did not. We look first at the structure of the UMC forecast game, the data generated by the game for this study, assumptions made on the data sample, and finally the statistical results.

2. Method

a. Game

The local forecast game in the UMC ASP is a required activity in the Daily Analysis and Forecast Interpretation course. A forecast is required of each contestant 4 days each week (Monday–Thursday, inclusive); they must make a prediction of maximum and minimum temperatures and a precipitation amount category for the Columbia, Missouri, airport for the next day (midnight to midnight, local standard time). Participants in the game are penalized 1 point for each Fahrenheit degree they are in error for both the maximum and minimum temperatures. Five additional points are assigned for each precipitation amount category they miss (Table 1). If a participant fails to forecast on a given day, then he or she is given the climatological maximum and minimum temperatures as well as the climatologically expected precipitation for that given day. These raw scores are what constitute the data in this study.

b. Data

Game scores were tracked for the 20 students enrolled in Daily Analysis and Forecast Interpretation for 11 weeks of the spring 2004 semester at UMC. The class was broken into two essentially equal groups, with one meeting to forecast and write their AFDs on Tuesdays, and the other group on Thursdays; only forecasts from these days were considered. Thus, there are two samples: one where students also wrote an AFD, and one where they did not. For example, on Tuesday, all of the students in the class made a forecast, but only one-half wrote an AFD as justification; on Thursdays, the groups reversed. Game forecasts from days when an AFD was composed generated "AFD day" scores. Those game forecasts from days when an AFD was not composed yielded "non-AFD" scores.

In all, there are 145 AFD-day scores and 155 non-AFD scores. Clearly these numbers do not support 20 students making 11 forecasts. In addition to absences and the other vagaries of student life, some students/contestants may have had days when a forecast was not submitted to the game, resulting in them being assigned climatology as a forecast (see above). As with days when a student was absent (and did not compose an AFD), days when a forecast was not submitted have been omitted also, as this study seeks to understand the impact of AFD composition on an *actual* forecast. An important consideration is that each group of forecasts is compared to the other (AFD-day versus non-AFD). Pairing each student's weekly AFD-day and non-AFD performance created a sample that was too small to test reliably.

Finally, a survey was administered to the student forecasters near the end of the semester when these data were gathered. Responses were requested for eight questions pertaining to AFD creation and forecast performance. Valid responses ranged from 1 (strongly disagree) to 5 (strongly agree); 3 was a neutral response.

c. Testing

The nonparametric Mann–Whitney test is used to test the hypothesis that student performance is no better (i.e., same scores) on days when they composed an AFD. The Mann–Whitney is used, in part, because of the small size of the data sample, thus making a normal distribution difficult to achieve. Additionally, while individual student schedules changed almost weekly, the Tuesday and Thursday groups were essentially left intact throughout the semester making it difficult to assume that the scores are randomly distributed. While the Mann–Whitney test also works from the assumption of a random sample, it deals with smaller samples by ranking all of the scores and then comparing the ranks from each set of scores. Details on this test may be found in any elementary statistics text (e.g., McClave and Dietrich 1991).

Unauthenticated | Downloaded 06/18/21 03:09 PM UTC

3. Results

a. Tests

The hypothesis posed with this work is that total forecaster performance (as measured by the UMC ASP forecast game) is no better on days when they composed an AFD. The median error (interquartile range) of forecast scores from AFD days is 7 (9) points, while the same measure for non-AFD scores was nearly the same 7 (8). The Mann–Whitney test of these two samples cannot reject the null hypothesis posed above. The same result is found when only the temperature or the precipitation amount portion of the aggregate score was tested.

Different results are obtained when we consider only those forecasts for days on which precipitation fell. There are 83 (96) AFD-day (non-AFD) forecast scores for such days having a median of 9 (9) total points, but an interquartile range of 8 (11). Thus, while AFD writers did not improve their median score, their typical range of error is noticeably smaller. Moreover, while no statistical difference exists between the aggregate scores of the samples or their temperature components, the Mann–Whitney test allows us to reject the null hypothesis that there is no significant difference between the precipitation forecast components of the samples ($p = 0.02$). This result suggests that the process of AFD creation benefits forecast performance on those occasions when the outcome matters most (i.e., inclement weather). These results are summarized graphically with box plots of the data (Fig. 1). The reduced error range on AFD-day forecasts is shown nicely in Fig. 1, although the graphical depiction of the difference between the precipitation forecasts on AFD and non-AFD days belies the results of the Mann–Whitney test. This is an artifact of the categorical nature of the precipitation forecast. Also, it must be noted that there was a total of 67 ties between the scores when ranked, a fact highlighted in Fig. 2. This calls into question the validity of the Mann–Whitney test, yet a chi-square test of the two samples also suggests a significant difference between the two distributions. The small size of the dataset in this initial study demands a fuller dataset from future investigations.

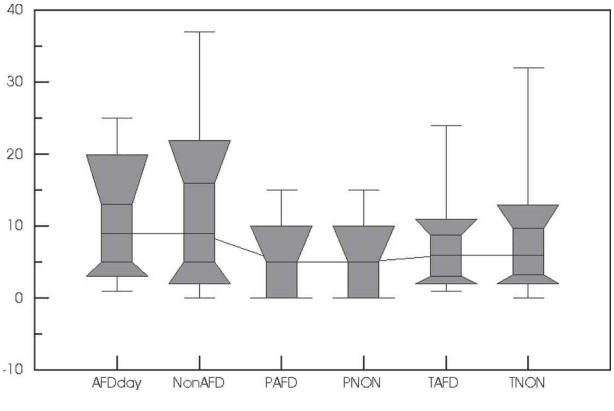

FIG. 1. A box plot of the total forecast error for days with precipitation for which a mock area forecast discussion was composed (AFDday) and not composed (NonAFD), the precipitation forecast errors on days when an AFD was composed (PAFD) and not composed (PNON), and the temperature forecast errors on days when an AFD was composed (TAFD) and not composed (TNON). From the bottom, the horizontal lines denote the minimum value; the 10th, 25th, 50th, 75th, and 90th percentiles; and the maximum value in the dataset.

A note is in order on the assumption of a random sample of student scores; several additional Mann–Whitney tests were conducted in order to assess this premise. Because the forecasting groups were largely unchanged during the course of the semester, it was assumed that 1) students in the Tuesday and Thursday groups possessed equal forecasting skill otherwise and that 2) the impact of composing an AFD on the verification score was the same on both Tuesday and Thursday. The first assumption was tested, with the null hypothesis being that there was no difference between the skill level of the two students groups. A Mann–Whitney test of the semester-end aggregate scores of students in each group was unable to reject the null hypothesis ($p = 0.90$). The second assumption was tested, with the null hypothesis being no difference between AFD writers and non-AFD writers on Tuesday (Fig. 3) as well as on Thursday (Fig. 4). Mann–Whitney tests of the difference between writers–nonwriters on Tuesday (Thursday) was unable to reject the null hypothesis, with $p = 0.62$ ($p = 0.77$). Yet, we again see the broader

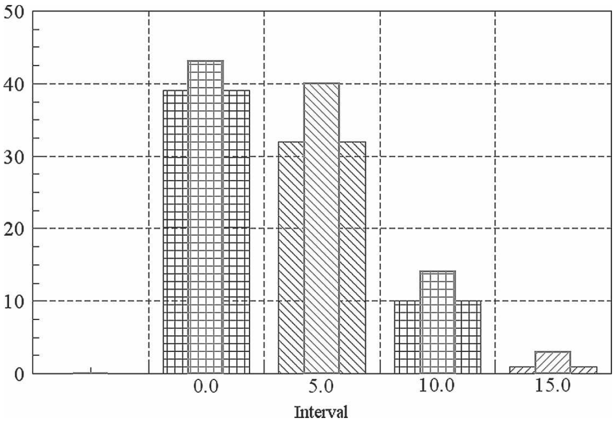

FIG. 2. Precipitation point score (x axis) vs the number of precipitation scores (y axis) earned by all forecasters on days when they wrote an AFD (wide columns) and those days when they did not (thin columns with thick borders).

Unauthenticated | Downloaded 06/18/21 03:09 PM UTC

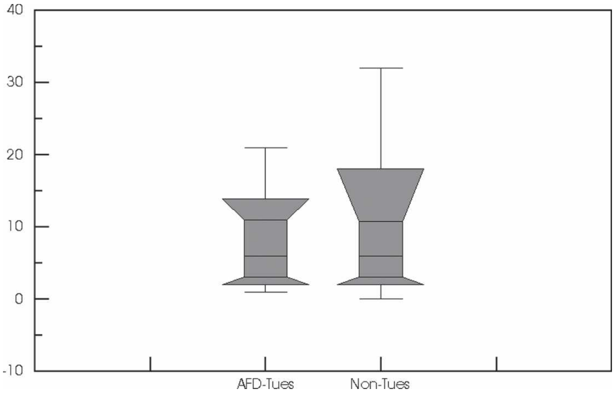

FIG. 3. A box-and-whiskers plot of the total forecast error for Tuesdays where students composed a mock area forecast discussion (AFD-Tues) and did not compose one (NonTues). From the bottom, the horizontal lines denote the minimum value; the 10th, 25th, 50th, 75th and 90th percentiles; and the maximum value in the dataset.

range of error present in both the Tuesday (Fig. 3) and Thursday (Fig. 3) non-AFD writers as opposed to the AFD writers on those days.

b. Surveys

Student response to the AFD writing activity was favorable. To the survey statement "Writing an AFD helped me to organize my thoughts about my forecast," the response was overwhelmingly favorable. All 20 respondents agreed with that statement, 12 strongly. Response was similar to the statement "I believe that writing an AFD helped me to make a better forecast." Eighteen respondents agreed, 11 strongly. When asked specifically about their perceptions of their own forecast game performance ("I expect that my forecast performance improved on days when I wrote an AFD"), confidence waned, but the majority (12) still agreed with the statement, 5 strongly so.

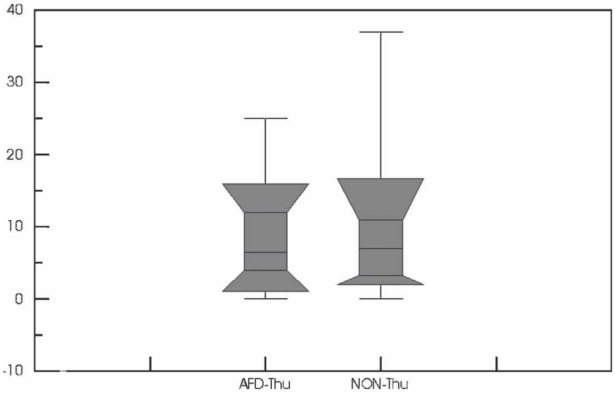

FIG. 4. A box-and-whiskers plot of the total forecast error for Thursdays where students composed a mock area forecast discussion (AFD-Thu) and did not compose one (NonThu). From the bottom, the horizontal lines denote the minimum value; the 10th, 25th, 50th, 75th and 90th percentiles; and the maximum value in the dataset.

The only real complaint regarded the time that must be invested in composing the AFD. To the statement "Writing an AFD was too time-consuming and detracted from the total forecast process," 16 disagreed, but only 4 of those strongly; the remaining 4 were neutral or strongly agreed. Indeed, the time spent in the laboratory typically doubles on days when students write their AFDs. Yet, most survey responses were favorable to the AFD as a writing assignment as well as a forecasting tool.

4. Conclusions

This analysis suggests that AFD composition is helpful in the creation of forecasts, which are more accurate during more critical, inclement weather conditions. It is noteworthy that these forecasts and AFDs were created for a midlatitude location during a meteorologically active period (February–April). Additionally, student forecasters have access to a broad range of meteorological data and forecast model output, as well as a sophisticated means of rendering and manipulating the information. Primarily, students use the General Meteorological Package (GEMPAK) Analysis and Rendering Program (GARP) for these activities.

In short, the subjects in this study had access to data and analysis tools similar to those available to NWS forecasters. In keeping with the NWS theme, these subjects were also required to create an AFD once each week. The instructor (the author) surmised that AFD creation was more than just a writing assignment to improve their composition skills, but also helped forecasters to focus their thinking, culminating in a better forecast. Although forecast scores failed to improve consistently for the group, egregious errors in forecast performance were measurably fewer on days when participants wrote an AFD in conjunction with their forecast. More importantly, precipitation amount scores do improve for AFD writers during active weather. Finally, forecaster confidence was bolstered by the idea that AFD composition demanded focused, disciplined thought and yielded a better forecast product.

Acknowledgments. The author thanks Drs. Tony Lupo, Neil Fox, Scott Rochette, John Monteverdi, and two anonymous reviewers for their thoughtful reviews of this manuscript.

Unauthenticated | Downloaded 06/18/21 03:09 PM UTC

REFERENCES

Bean, J. C., 1996: *Engaging Ideas: The Professor's Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom.* Jossey-Bass, 282 pp.

McClave, J. T., and F. H. Dietrich, 1991: *Statistics.* Dellen-Macmillan, 928 pp.

Roebber, P. J., and L. F. Bosart, 1996: The contributions of education and experience in forecast skill. *Wea. Forecasting,* **11,** 21–40.

——, ——, and G. S. Forbes, 1996: Does distance from the forecast site affect skill? *Wea. Forecasting,* **11,** 582–589.

Vislocky, R. L., and J. M. Fritsch, 1997: Performance of an advanced MOS system in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.,* **78,** 2851–2857.

Unauthenticated | Downloaded 06/18/21 03:09 PM UTC