Weather Forecasting as a Learning Tool in a Large Service Course: Does Practice Make Perfect?

ELIZABETH J. SUESS, CINZIA CERVATO, AND WILLIAM A. GALLUS JR.
Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa

JONATHON M. HOBBS
Department of Statistics, Iowa State University, Ames, Iowa

(Manuscript received 2 October 2012, in final form 5 February 2013)

ABSTRACT

Each spring roughly 200 students, mostly nonmajors, enroll in the Introduction to Meteorology course at Iowa State University and are required to make at least 25 forecasts throughout the semester. The Dynamic Weather Forecaster (DWF) forecasting platform requires students to forecast more than just simple “numeric” forecasts and includes questions on advection, cloudiness, and precipitation factors that are not included in forecast contests often used in meteorology courses. The present study examines the evolution of forecasting skill for students enrolled in the class in spring 2010 and 2011 and compares student performance with that of an “expert forecaster.” The expert forecasters were chosen from meteorology students in an advanced forecasting course who showed exemplary forecasting skill throughout the previous semester. It is shown that these introductory students improve in forecast skill over only the first 10–15 days that they forecast, a number smaller than the 25 days found in an earlier study examining meteorology majors in an upper-level course. The skill of both groups plateaus after that time. An analysis of two types of questions in the DWF reveals that students do have skill slightly better than that of a persistence forecast when predicting parameters traditionally used in forecasting contests, but fail to outperform persistence when predicting more complex atmospheric processes like temperature advection and factors influencing precipitation such as moisture content and instability. The introduction of a contest “with prizes” halfway through the semester in 2011 was found to have at best mixed impacts on forecast skill.

1. Introduction

Because weather forecasting is a topic of broad public interest, it has been used even in introductory meteorology courses as a tool to improve student understanding of the atmosphere (e.g., Yarger et al. 2000). Since students have grown up in an environment where weather forecasts are constantly communicated to the public, it might be expected that these students would benefit from trying to forecast the weather themselves, or perhaps find the experience interesting. At the very least, such efforts convey the challenging nature of weather forecasting. Students majoring in meteorology are presumably not surprised to find forecasting a required part of at least one course they take, and many are likely exposed to some of the studies that have evaluated forecast skill (e.g., Olson et al. 1995). However, fewer studies have been done to evaluate how student forecasting skill itself develops, particularly outside the population of meteorology majors.

Bosart (1975) studied how forecasting skill changed over a several-year period for meteorology students at the University at Albany, State University of New York. Forecasting skill was defined as an improvement over persistence forecasts (i.e., a forecast that uses today’s conditions to predict tomorrow’s weather). Bosart discovered relatively little change in skill for most parameters over the years examined, although there was a small dependence of temperature forecasting skill on the standard deviation of the daily temperature from the climatological mean. He also noted that as students forecast, their skill reaches a plateau after which they no longer improve, a result also found by Sanders (1973). However, Roebber and Bosart (1996) looked at nine
semesters of temperature and precipitation forecasts made by both meteorology students and faculty at the University at Albany and found that forecast skill for these two parameters was largely dependent upon experience. They felt the relative advantage of experienced forecasters was due both to maintenance of a high level of linear consistency between the information that leads to a forecast and the forecast itself, and to recognition of events where those linear relationships would not apply. Bond and Mass (2009) used 10 years of forecasting data from meteorology students enrolled in an upper-level forecasting class at the University of Washington to study how forecasting skill improved over time. They found that the forecasters improved for the first 25 forecasts, after which time they showed minimal improvement. This number of 25 is consistent with the earlier findings of Gedzelman (1978), who proposed that forecasting skill is mostly acquired in the first 30 forecasts.

Cervato et al. (2009, 2011) took a slightly different approach and looked at how forecasting affected overall grades of students enrolled in a large-lecture introductory course, Introduction to Meteorology (MTEOR 206), at Iowa State University. The class was required to make at least 25 forecasts using the Dynamic Weather Forecaster (DWF). Cervato et al. looked to see if there was a trend in grades related to the time when students began forecasting. It was found that the earlier students started forecasting, the better they did in the class, even though the portion of the grade related to forecasting was only 25%. Even when this component of the grade was taken out of their final grade, students who started forecasting early did better overall in the class.

The objective of the present study is to expand on the findings of Cervato et al. to see how repeated use of the DWF affected the forecasting skill of this large group of primarily nonmeteorology majors over time, and if the plateau observed by Sanders (1973), Bosart (1975), and Bond and Mass (2009) also occurred with this population. The last column classifies the question as type 1 (traditional) or type 2 (more complex).

2. Data and methodology

Forecasts for 218 students enrolled in the same introductory course (MTEOR 206) at Iowa State University were collected during the spring 2010 semester and for 186 students during the spring 2011 semester. Although the present study focuses on only this 2-yr period, the large class size results in a sample size larger than that of previous studies (e.g., Bond and Mass 2009), which focused on meteorology majors over longer time periods. Students were required to make at least 25 forecasts throughout the semester; they also listened to 5–10-min weather discussions at the beginning or end of the class periods when they attended class. The MTEOR 206 course is open to all majors at Iowa State University and is required for freshmen in the meteorology program, although because some of these students take the course during fall semester, the number of meteorology majors in the course during the spring semester is generally less than 20.

Forecasts were made by entering values before local midnight in DWF for the next day’s 1200 and 1800 UTC temperature, cloud cover, temperature advection, and frontal passage as well as 1800 UTC wind direction and wind speed, the likelihood of precipitation over 24 h (1200–1200 UTC, i.e., a 12–36-h forecast), and the factors that could cause precipitation (Table 1). All forecasts were made for Des Moines, Iowa (KDSM). Additional details on DWF can be found in Cervato et al. (2011), with the entire assignment used for the course evaluated in the present study available online (http://wiki.its.iastate.edu/display/FCST/Home).

The forecasting assignment included in the DWF was originally developed by Yarger et al. (2000). This assignment is unique in that there are questions relating to advection, cloud cover, and fronts, items for which historically a student could not find quickly an expert forecast on the Internet. The best use of the DWF is in large-enrollment courses since the system automatically grades each forecast as the data come in, so students get the results for their forecast within 72 h without burdening the instructor with grading.

Forecasts are scored out of 36 total points. Each student is given one point per incorrect attempt and three
points for a correct answer to one of the 12 graded questions of the assignment. Because a problem was discovered during 2010 in the way the two frontal questions were being scored by the DWF, these two questions were eliminated from our analysis, so that the maximum point total possible for a forecast was 30. The forecasting periods for which data were collected were 13 January–28 April 2010 and 11 January–25 April 2011. Forecasts could be made on any day of the week as long as they were submitted by 0600 UTC (midnight local standard time) the day before the forecasting period. Upper-level meteorology students who had completed a one-semester junior-level course in forecasting and had shown exemplary forecasting skills were chosen to provide an “expert” forecast for each day of the forecasting period; this was used as a guideline to evaluate student performance. Two experts participated in 2010 with one participant in 2011; these small numbers follow the approach of Bond and Mass (2009), who used one expert.

One challenge in analyzing this data was the fact that individual students were free to choose on which days they would forecast and also on the total number of days they would forecast during the semester. This freedom meant that different students might be forecasting on different days, and student skill could be impacted by the type of weather occurring on each day. In addition, some students gained extensive experience forecasting by the end of February, whereas other students had not started yet, complicating our analysis of skill trends as a function of day of the semester.

To minimize problems associated with these differences, two approaches were used. First, an analysis was performed using a normalized time scale with 0 representing the first forecast made by a particular student and 1 representing the last forecast. This normalization better allows the impact of experience on forecasting skill to be examined despite students beginning to forecast at different times and choosing to forecast for different numbers of days. Second, to account for differences in skill that might be related to the complexity of weather conditions occurring on any given day, persistence forecasts were calculated by taking the correct answers from the previous forecast and using them as the forecasts for the next day; the scores of the persistence forecasts were then subtracted from the student scores, an approach similar to that used in Bond and Mass (2009). Because the forecast scores obtained with DWF were tallied differently than those in Bond and Mass, with high scores being good, unlike traditional approaches that tally error points, it was not necessary to divide the difference by the persistence forecast score as they did.

To investigate whether or not long-term trends were occurring in the level of difficulty of the forecasts as the semester progressed, persistence scores over the semesters during both 2010 and 2011 were examined (Fig. 1). No consistent long-term trends in difficulty were present over the two years, although during 2011 persistence scores were relatively high in the late March and April period.

As a check on how well the removal of the persistence score accounted for changes in forecast difficulty, forecasting skill for the expert forecasters was also examined over both semesters (Fig. 2). If it can be assumed that these experienced forecasters would not be experiencing long-term improvements in skill (Bond and Mass
the removal of the persistence score should result in relatively flat curves over time for the experts, if the persistence scores accurately reflect forecast difficulty. The daily values of the scores jump around noticeably, implying persistence forecasts likely do not fully reflect forecast difficulty. However, the smoothed curves are relatively flat, suggesting that the removal of the persistence score may be adequate to eliminate biases due to difficult forecasts (Bond and Mass 2009).

To better compare results with previous studies, the data were also divided into two groups reflecting different types of forecasts. A “type 1” forecast was chosen to use the DWF questions most similar to other forecasts used traditionally in classrooms (e.g., Bosart 1975, 1983; Bond and Mass 2009; current Weather Challenge national forecast online at http://www.wxchallenge.com), taking just the scores for the 1200 and 1800 UTC temperature, and the 1800 UTC wind speed and direction for a total possible score of 12 (Table 1). The “type 2” forecast included the other questions that mostly are designed to test student understanding of processes that affect some weather parameters. Type 2 forecasts included the 1200 and 1800 UTC cloud cover, temperature advection, precipitation in the upcoming 24 h, and the factors that could potentially cause precipitation, for a maximum possible score of 18 (Table 1).

The data were also categorized by the total number of forecasts made by the students, to see if the students who did more than the minimum of 25 experienced greater improvement or earned higher scores in general than those who failed to do the minimum amount of work. Finally, a contest “with prizes” was implemented after 1 March during 2011, and some analysis was performed comparing skill and trends in both years prior to and after 1 March.

The same statistical model was fit to each of these different subsets of the forecast skill data. To avoid requiring a specific form for the relationship between forecast skill and normalized time, the data were fit to a small set of spline basis functions. The spline approach allows for the expected forecast skill to be a smooth function of normalized time and is flexible enough to allow the data to determine where increases or decreases occur (Faraway 2006). Since individual students forecast many times, the statistical model utilizes a random coefficients framework, meaning that there is a forecast skill curve for each student. This population of individual curves can be summarized with the class average curve, and these are presented in the next section. The random coefficient splines were fit using the lme4 package in R (Bates et al. 2012; R Core Team 2012).

3. Results

a. Skill as a function of time

The best-fit curves for the students’ forecasting skill with persistence scores subtracted (Fig. 3) shows gradual improvement in both 2010 and 2011 during the first 0.2 or so of the normalized time scale, with a generally steady plateau after that time. An analysis of skill as a function of forecast number (not shown) shows this...
plateauning happening after roughly 10–15 forecasts each year. Although students were required to forecast a minimum of 25 days, most students forecast on 50–60 days since their grade was based on the best 25 forecasts. This period of only 10–15 days with improving scores is shorter than the 25-day period found in Bond and Mass (2009) for meteorology majors in an upper-level course, and implies that student skill improvement is limited by the students’ understanding of the atmosphere, a result consistent with the findings in Roebber and Bosart (1996). In the low-level course, students receive a cursory and limited understanding of how the atmosphere works, and can only improve their forecasting to a certain level. Students who have taken multiple meteorology courses may be able to improve to a higher level of forecasting skill. Of note, when the subset of roughly 20 meteorology majors is examined in both years (not shown), their skill is slightly better than that of the class average, but falls within the 95% confidence bands, indicating no statistically significant difference from the other students. This result is not surprising considering that the majors take this course during their freshman year and it is the first meteorology course to which most are exposed.

During both years the students began the semester with skill equal to persistence, and improved only slightly to have skill about 1 point better than persistence throughout much of the semester. Although this improvement is small, the analysis indicates performance above the level of persistence with greater than 95% confidence. In addition, it can be seen that this value is usually about half the skill of the expert forecasters. The differences in the expert curves between 2010 and 2011 are likely an artifact of having two experts forecasting in 2010 and only one in 2011.

Figure 3 also shows that during 2010 skill declined during approximately the last 0.1 of the normalized time scale. Such a decline was less pronounced in 2011. It is possible that the decline in 2010 reflects the fact that many students had already forecast many more times than the required 25, and were not as motivated to spend time thinking about their forecasts. The decline might also be related to end-of-semester workload increases with course projects being due and final exams approaching. The busier schedules might lead to a reduction in time spent on the forecasts as well. The 2011 data may not reflect the decline because of the implementation of a contest with rewards during the last half of that semester. This contest will be discussed later.

To gain more insight into trends in forecast skill, we also compared student scores to those of the two expert forecasters (Fig. 4) during 2010. The scores for students with the highest average score, median average score, and lowest average score over the entire semester are compared to the two expert forecasters in Fig. 4. It can be seen that the student who ended up with the highest average score made all of his/her forecasts early in the semester, whereas the student with the worst score made almost all forecasts at the end of the semester. The same trends were present in 2011 (not shown). All other factors (e.g., students’ forecasting skill, diligence, success in course, attendance) being equal, these trends might imply that forecasts became more difficult later in the semester and that students who forecasted later were more likely to receive worse scores than students who forecast early. However, it is important to note that both the student with the median score, who forecast nearly every day, did not show a drop in skill toward the end of the semester. A comparison between the scores in the four exams given throughout the course
and students’ forecasting scores (not shown) indicated that the students who did better in the exams had better forecasting skills; students with low exam scores also had lower forecasting skills. Thus, we propose that the trends shown in Fig. 4 are due to students’ skills and motivation rather than forecasting difficulty: the students who forecast early likely took more care in examining weather data before completing a forecast than students who rushed to complete the requirement before the semester ended.

To further explore the impact of student motivation on forecasting skill, the data were separated into two groups: one for students who forecast 25 times or fewer and the other for those who forecast more than 25 times (Fig. 5). In general, initial improvement can be seen during both years for both groups of students. Of note,
During 2010 the improvement was more pronounced and took place over a longer normalized time for the group that ended up forecasting the most times. During 2011, the opposite trend was present, with a more pronounced improvement in the group that forecasted 25 times or fewer. This difference in behavior suggests that each year’s student population can have some marked differences. In both years, the students who forecast more times tended to have slightly higher scores than those who forecast fewer times, although differences are not statistically significant. Also, in both years there is some evidence of a decline in performance toward the end of the semester, except for the 2011 sample of students who forecasted more than 25 times. As will be discussed later, the implementation of a contest during the last part of that semester may have helped to prevent the decline among that sample. It is of note, however, that the contest only seemed to motivate the students who were forecasting the most. This result suggests that less motivated students may not be spurred to perform better through the use of this type of reward system.

It is important to keep in mind that several of the questions used in the DWF are very different from those used in traditional forecasting activities, and this difference may explain some differences in the results between the present study and others. To explore what impact the different types of questions might have, data were sorted into type 1 and type 2 forecasting scores to examine if both semester trends and student performance are related to the question type. Type 1 scores showed that students could consistently forecast temperature and wind at a level about 2 points above a persistence forecast (Fig. 6). A small improvement in these scores also took place during roughly the first 0.2 portion of normalized time. The skill of the expert forecasters was generally slightly better than that of the students, near or just outside the 95% confidence bands, and the initial ramp upward in skill was less pronounced or missing.

For type 2 questions forecasting scores were much lower, never rising above the level of persistence (Fig. 7). This result implies that students have more difficulty understanding atmospheric processes that control advection, the influence of cloud cover on temperature, or precipitation factors. However, the initial improvement was slightly larger than for type 1 forecasts, consistent with the fact that probably all students had never previously seen forecasts for the atmospheric processes addressed in these questions, whereas they were used to hearing about forecasts for the traditional parameters covered in the type 1 questions. The difference between the skill of the expert forecaster and that of the students was slightly greater for the type 2 questions than for the type 1, but even for the experts, skill was only comparable to that of a persistence forecast. The increased improvement in the skill of the experts compared to that of the general student population for type 2 forecasts compared to type 1 is consistent with the experts having a much deeper understanding of the atmosphere and therefore better potential to forecast these more complex parameters. The expert curves behaved very differently at the start of the semester in the two years, possibly implying differences in the comfort level of the experts with forecasting the type 2 parameters (different experts were used in the two years). Because no prior studies have evaluated student skill with these types of questions, caution should be used in generalizing these results.

b. Analysis of the role of contest prizes in 2011

While the forecasting assignment was the same for the class of 2011 as it was in 2010, halfway through the
a) 2010 Type 2
b) 2011 Type 2

FIG. 7. Individual student forecast skill (points) on type 2 questions as a function of normalized time for (a) 2010 and (b) 2011. The solid lines depict the class-average spline fit, and the shaded regions indicate 95% confidence bands.

a) 2010 Contest
b) 2011 Contest Groups

FIG. 8. Individual student forecast skill (points) as a function of normalized time before and after 1 Mar (depicted by vertical line at 0.5) for (a) 2010 and (b) 2011. The solid lines depict the class-average spline fit derived separately for the two time periods in each year.

semester, a contest with prizes was started as an incentive to see if forecasting skill would show more improvement during the contest than what occurred in 2010 when no contest prizes took place. The contest prize was 5 bonus points awarded to the three students with the highest forecasting scores for the previous 2-week period. As discussed briefly earlier, average skill was relatively constant in 2010 until the last 0.1–0.15 of the normalized time scale, when it decreased (Fig. 3). This decrease was not as apparent during 2011. Also as mentioned earlier, the change in performance in 2011 compared to 2010 was restricted to those students who were most active in forecasting (i.e., forecasting over 25 times during the semester; see Fig. 5).

To get a better understanding of how the contest affected forecasting skill, the data were divided into before and after 1 March periods for both 2010 and 2011. The two time periods in each semester were normalized, with the pre–1 March period using a scale from 0 to 0.5, and the post–1 March period continuing from 0.5 to 1. Curves were fit to the data in both parts of each semester. For 2011, the data were subdivided into one group for students who had forecasted at least 10 times prior to 1 March, and those who had not forecasted at all before 1 March. (Although not shown, the curve for the full sample of all students in 2011 strongly resembled that of the group who had forecasted at least 10 times prior to 1 March.) During both years, the curves were relatively flat after the initial improvement during the period of time prior to 1 March (Fig. 8). After 1 March, however, some small differences developed between the years. During 2010, the scores after 1 March were relatively similar to those during the first part of the semester, with some decline toward the end of the course. In 2011, the scores tended to be worse during the last part of the semester, despite the contest, and were especially low.
immediately after 1 March. This result is counterintuitive and shows up both among the new forecasters who were motivated to begin forecasting due to the contest (dashed line), and those whose lack of experience results in somewhat lower skill than that of the more experienced forecasters, but also among the students who had already been forecasting consistently (solid line). A careful look at Fig. 2 shows that expert skill also dropped after 1 March in 2011. The most likely explanation for the drop in skill among all forecasters after the start of the contest is a substantial change in the general weather pattern. Figure 1 shows that the general skill of the persistence forecast changed noticeably after 1 March, with a tendency to be very high on many days but with much larger swings from day to day than was the case earlier in the semester and throughout 2010. Perhaps this behavior is indicative of a more challenging period for forecasting. At a minimum, it would be hard for students to earn positive skill scores when the persistence forecast was often very high. In addition, the dramatic day-to-day changes in the skill of persistence imply dramatic changes in the weather that would likely prove challenging for an introductory student to forecast. As a final note, at least for the students who had been forecasting consistently prior to the start of the contest, there was no longer a decline during the final days of the semester, as was the case in 2010, implying one potential positive impact from the contest.

4. Conclusions

An analysis of student forecasting skill on 10 questions of a forecasting activity undertaken by a large-lecture introductory meteorology course shows that students experience some initial improvement, before quickly plateauing for much of the semester. The small improvement is restricted to the first 10–15 forecasts, a period shorter than the 25-day period found for majors in an upper-level meteorology course by Bond and Mass (2009). Student skill initially is equal to persistence, and plateaus at an average of 1 point better than persistence, about half the skill of the expert forecasters, who were upper-level meteorology students who had performed well in a junior-level forecasting contest. It appears that the amount of improvement is likely limited by the rather cursory amount of understanding that the students can obtain in an introductory large-lecture course.

A comparison of those students who forecasted 25 times or fewer to those who forecasted more than 25 times reveals a general but statistically insignificant tendency for slightly better scores in the group that forecasts the most.

Because of the nontraditional nature of some of the questions in this DWF activity, the forecasts were separated into two types to allow better comparison to some prior studies. For type 1 questions, which are generally specific weather parameters at specific times (similar to most forecast activities used traditionally), students consistently forecast with skill above that of a persistence forecast. For type 2 forecasts, however, student forecasting skill remained worse than that of persistence, and the gap between the students and the expert forecasters was larger, suggesting that these types of questions dealing with atmospheric processes pose a special challenge to introductory students. It must be noted, however, that even the expert forecasters had skill only comparable to persistence for the type 2 forecasts, implying that the forecasts are, in general, more challenging.

The scores from 2011 suggest that turning the forecasting assignment into a contest with prizes being awarded had mixed results on student performance. On one hand, average scores during the last half of the semester in 2011 were lower than during the first half of the semester, whereas during 2010, scores were fairly uniform both before and after March 1. However, the decline in scores in 2011 is likely explained by the fact that the contest would entice many of the students who were less motivated and had not yet forecasted to begin forecasting after the contest started. These students would have lacked experience and their scores likely would lower the average. On the other hand, a drop in skill that happened at the end of the semester in 2010 did not occur during 2011. This result implies the contest may have had a small positive impact on forecasting skill. Caution must be used in comparing the two years, because some differences might arise due to different weather conditions between the two years and different student abilities. Also, the mixed results and at best modest positive impact suggest that the use of bonus points as a prize may not have been sufficient motivation for the class.

The findings from this study have several implications. First, students may not get enough of a background in introductory meteorology classes like this one to be able to understand the questions that are asked in a forecast. The upper-level students from Bond and Mass (2009) may have shown more improvement over persistence and improvement over a longer period of time due to the fact that they would have had a better understanding of how the atmosphere works and thus have a larger toolkit from which to draw to improve their forecasts. Second, as with any learning activity, motivation plays a large role into how well students may be able to forecast. Those students who choose to forecast more often tend to earn slightly higher scores than those who do not. Finally, the use of bonus points as an incentive in a contest may not be sufficient to lead to marked improvements in forecasting skill. Despite the limited improvement shown in forecasting ability, it is important to
note that use of the DWF has been shown previously to increase student learning (Cervato et al. 2009).

Numerous options exist for future work. First, an evaluation of the activity during the fall semester should be performed to see if the trends are similar to those found in the spring. Such an expansion of the study would help to ensure that results are not biased by the weather patterns common in winter and spring. Second, it would be interesting to study the trends among students within different majors of study taking the course. Would science majors consistently do better than non-science majors? As mentioned earlier, no significant differences were found in the current sample when the meteorology majors were compared to the class as a whole, but the sample size was small. Third, since forecasting often relies upon maps with graphical analysis, does gender play a role affecting forecasting performance, given that studies show that females have lower spatial skills than males (e.g., Lippa et al. 2010)? Fourth, how would the results differ if forecasts were made for the Pacific Northwest as in Bond and Mass (2009)? Finally, would a more tangible contest prize, such as a monetary gift, have a bigger impact on forecast skill?

Acknowledgments. This work began as a senior undergraduate thesis project for the lead author. Partial funding for this research was supplied by NSF Grant DUE-0618686. Thanks are also given to all of the students that took MTEOR 206 in the spring of 2010 and 2011, to Nick Krauel, the second expert forecaster in 2010, and to Josh Alland, the 2011 expert forecaster.

REFERENCES

Bates, D., M. Maechler, and B. Bolker, cited 2012: lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. [Available online at http://CRAN.R-project.org/package=lme4.]


