Development of Weather Forecasting Skill for Students in the Atmospheric Sciences Program at the University of Washington

Nicholas A. Bond\textsuperscript{1} and Clifford F. Mass\textsuperscript{2}

\textsuperscript{1} University of Washington Joint Institute for the Study of the Atmosphere and Ocean
\textsuperscript{2} University of Washington Department of Atmospheric Sciences

Abstract

Daily values of forecast scores are evaluated for the students in a weather analysis and forecasting class (ATMS 452) offered by the Department of Atmospheric Sciences of the University of Washington during the spring terms of 1997 through 2007. The objective of this study is to determine the rate at which senior-level undergraduate students develop proficiency at short-term (next-day) weather forecasting. Separate analyses are carried out for different categories of forecast parameters. Time series of the average skill achieved over the course of the quarter are considered for the median student, and the highest two and lowest two student forecasters each year. An overall upward trend in student forecast skill occurs over roughly the first six weeks of the quarter, followed by minimal systematic changes in skill. The correlation coefficient between the students’ overall forecast performance and test scores in ATMS 452 is about 0.4. The results are relevant to the design of effective instructional programs for weather forecasting.
1. Introduction

Although a large body of work has examined the skill of weather forecasts, less effort has been devoted to examining the development of forecast skill, with the notable exceptions of Gedzelman (1978) and Roebber and Bosart (1996). This paper extends their work, using the results from ten years of forecasts made by students enrolled in a senior level weather analysis and forecasting class (ATMS 452) taught at the University of Washington.

The primary objective of ATMS 452 is to demonstrate how the relatively abstract information presented in classes on the dynamics and thermodynamics of the atmosphere can be applied to forecasting tangible weather. It is designed for seniors in the last term of their undergraduate studies, but has included students in different stages of their education (e.g., juniors, visiting students, and a few graduate students). The class is taught every spring, with a ten-week period of instruction from the end of March to early June. ATMS 452 is composed of a one-hour lecture three to four days a week with an emphasis on practical topics, a one-hour laboratory session four days a week that includes real-time surface map analysis and case studies illustrating specific phenomena such as mesoscale structures induced by orography, and most germane to this paper, a one-hour forecasting exercise five days a week. The lecture portion is taught by the second author (Mass); the laboratory and forecast components are taught by the first author (Bond, hereafter NAB). The focus of ATMS 452 is on understanding and forecasting day-to-day weather using the concepts and tools of modern meteorology.

The daily forecast scores for students in ATMS 452 provide a unique opportunity to analyze the development of proficiency at weather forecasting in a controlled environment. This environment resembles more of an operational setting than the forecast “game” used by Roebber and Bosart (1996), in that the forecasts are not just mandatory and graded, but are carried out under a tight time constraint. Moreover, the consistency with which the forecast component of ATMS 452 has been conducted over many years provides a substantial amount of comparable data. Our analysis of this data has had the following objectives: (1) to document the rates at which students develop skill at short-term forecasting, (2) to determine the differences in the rates at which
proficiency is gained for different kinds of forecasts, (3) to compare forecasting skill with performance on conventional written tests.

The organization of the paper is as follows. The forecasts, scoring, and methods of analysis are described in the next section. Section 3 represents the results of our analysis of the forecast scores. A summary and concluding remarks are provided in the final section.

2. Description of Forecasts and Methods

a. Weather parameters and scoring

The scores analyzed here are based on next-day forecasts made by the students and NAB on Monday through Thursday afternoons during the last hour of instruction. Three basic types of the next-day forecasts are the subjects of this study. The Type1 category is for parameters verifying at a specific time, always 1200 UTC the next morning. These parameters, itemized in Table 1, consist of categorical forecasts of ceiling, visibility and sky cover, and forecasts of wind speed and direction, temperature, and dew point. The scoring method for each parameter is also indicated in Table 1. In practice, while it is difficult for all forecasters to predict cloud cover with consistent success, the most points are generally associated with errors in forecasting dew point, followed by temperature. The Type2 category relates to precipitation over the period of 0600 to 1800 UTC the next day. There are three separate elements: a precipitation probability, a thunderstorm probability, and a categorical quantitative precipitation forecast, summarized, along with their scoring, in Table 2. Note that the probability forecasts are scored in a “proper” manner (e.g., Murphy and Epstein 1967) such that in the long run, scores are optimized when forecast probabilities match the true probabilities. After a short introductory period, the students make these two types of forecasts for the airports of four cities in different regions of the U.S.: Oklahoma City (OKC), Pittsburgh (PIT), Fargo (FAR) and Seattle (SEA). Finally, a third type of forecast is conducted in approximately the last month of class and consists of an intensive, local forecast for Seattle Tacoma Airport (SEA). This is also a next-day
forecast and consists of projections for wind and sky conditions at 6-hour intervals (0600 UTC, 1200 UTC, etc.), minimum and maximum temperatures, as well as precipitation probabilities and Quantitative Precipitation (QPF) for 0600 to 1800 UTC and 1800 to 0600 UTC. This local forecast supplants the Type1 and Type2 forecasts for SEA, with the forecasts for the other 3 cities are unchanged.

The students are subjected to an intensive forecasting experience in this class for three reasons: (1) to provide the repetition presumed to be important in gaining prowess at forecasting, (2) to yield a large number of realizations so that the students’ scores, which comprise part of their grades in ATMS 452, are statistically robust, and (3) to give the students a chance to find out how well they cope under the time pressures that they would face in an operational forecast setting.

It is important to distinguish the aspects of the forecast component of ATMS 452 from other forecast competitions that have been the subject of previous studies (e.g., Bosart 1983; Sanders 1986). First, the students’ forecasts are individual efforts. While student conversations about the weather situation in general are encouraged, comparisons of their specific forecasts during the preparation stage are discouraged. Second, the use of Model Output Statistics (MOS) guidance is prohibited. This philosophy is based on the idea that inexperienced forecasters will use MOS as a crutch, and probably should do so if forecast score is the top priority (Baars and Mass 2006). Using MOS likely delays the development of understanding of how various elements of the weather relate to the larger-scale aspects of the atmosphere.

b. Analysis procedure

The daily scores for each student, NAB, and persistence represent the data set analyzed here. A persistence forecast, i.e., what happened today will happen tomorrow, is an appropriate standard or control forecast by which a next-day forecast is evaluated. For the purpose of illustration, time series of daily scores during a typical term for three students chosen at random, NAB, and persistence for Type1 and Type2 forecasts are shown in Figs. 1a and 1b, respectively. The key trends apparent in these time series occur essentially each year. For example, the raw daily Type1 scores for virtually all students
tend to decline over the quarter, as do the scores for NAB and persistence (Fig. 1a). As mentioned earlier, the class is always held in the spring, during a period in which synoptic-scale disturbances tend to become progressively weaker, resulting in lesser day-to-day changes in temperatures and winds, and thus lower error scores. On the other hand, less systematic change is typical for Type2 (precipitation) scores, as shown in Fig. 1b. A major factor is the change in the origin of the precipitation over the quarter, with deep convection tending to play an increasingly important role in precipitation variability.

To account for the change in the weather during the period of instruction, and potentially the difficulty in forecasting, we have computed skill scores (SS) defined as

\[ SS = \frac{FP_{per} - FP_{ind}}{FP_{per}} \]  

where \( FP_{per} \) is the point total of the persistence forecast and \( FP_{ind} \) is the point total of the human forecaster. In this formulation, “1” represents perfect skill, “0” represents no skill, and it is possible to have negative skill, relative to the control forecast of persistence. Our measure of forecast skill follows the form used by Sanders 1979 and others. Sequences of daily-average SS values over many classes are formed by averaging the scores with respect to forecast day. For example, averages are computed for the scores (FPs) for the first day of forecasts over all the years, and SS values for that forecast day are then produced by using these averages in (1). This procedure is repeated for the second day scores, and so forth, through the 33 days of forecasting that occur each quarter. The alternative, computing SS values for each individual forecast and then averaging, yields results dominated by the instances for which the persistence score is small (and is ill-posed when persistence is a perfect forecast). We present results in the following section on the sequences of the average SS for the median student, and for the two most proficient and two least proficient of the student forecasters each year.
3. Forecast Skill Sequences

Time series of forecast skill in the Type1 category are shown in Fig. 2. A marked increase in median student skill occurs over the course of the first half of the quarter for the Type1 forecasts, with a flattening in this trend over the second half of the term. Since the forecast skill of NAB does not vary systematically during the term, the upward trends in student skill shows that proficiency is being gained by the students. Note also the substantial day-to-day variations in skill, for all groups of students and NAB, even though these traces reflect averages over 10 years. The consistency in the day-to-day variations in the time series illustrate that the relative degree of difficulty in forecasting on particular groups of days is similar for both novice and experienced forecasters. The top student forecasters have an immediate edge on the other students, continue to improve over the first two-thirds of the term, and by the last third of the term have skill approaching that of NAB. The skill of the less successful student forecasters increases steadily over the course of the quarter. The difference in skill between the higher and lower student groups narrows from about 0.25 to 0.15 over the duration of the quarter.

The time series for the Type2 forecasts (Fig. 3) has some interesting distinctions from those for the Type1 forecasts. First, a typical day’s median student forecast skill is actually higher for the Type2 predictions related to precipitation, than for the Type1 forecasts related to clouds, winds and temperatures. Second, while the median student skill for the Type2 forecasts declines with time over the course of instruction, NAB’s skill actually exhibits a greater rate of decline, to the point where the students’ skill closely approaches NAB’s. As mentioned above, the precipitation tends to be increasingly convective in nature during the period considered, posing an increasing challenge for all forecasters. The top students maintain their skill late into the spring as precipitation becomes less predictable, and their mean skill near the end of the term actually slightly surpasses that of NAB’s. Unlike Type1 category forecasts, the gap in skill between the poorer forecasters and the other groups is relatively constant during the quarter.

The final type of forecasts considered is the intensive local (Seattle) forecast. This forecast type is for clouds, winds, temperatures, and precipitation probability and
amount, and hence includes the elements in both the Type1 and Type2 forecast results shown above. This forecast is made only the latter portion of the class, after the students have had 5-6 weeks of forecast practice. Time series of skill as a function of forecast number are shown in Fig. 4. These time series are rather short, but suggest little in the way of obvious trends in the skills for any of the groups. For the most part, the differences in skill between NAB and the students in the three groups for the entire duration of the Seattle forecasts are similar to those differences near the end of the term for the Type1 and Type2 forecasts. This result is not surprising. The students have had forecast experience by the onset of these forecasts, and also have the benefit of familiarity with Seattle weather, based on both knowledge gained before taking ATMS 452 and the early activities of the class.

The consistency that has been maintained in the lecture portion of ATMS 452 over the years, in particular the nature of its examinations, allows assessment of the relationship between the students’ forecast scores and test scores. A scatter plot of the students’ average test scores, with the midterm and final exams weighted equally, against their overall forecast scores, combining all four types of forecasts, is shown in Fig. 5. There is a positive relationship between test and forecast scores, as might be expected, but the linear correlation coefficient is rather modest ($r \sim 0.4$). The distribution is asymmetric. A large proportion of the best forecasters also did well on tests, but for forecast scores below about 80, there is little relationship between forecast and test scores.

4. Final Remarks

The forecasts of students enrolled in ATMS 452 at the University of Washington for the years of 1997 through 2007 have been analyzed. The primary objective of this analysis was to determine the rate at which the students, primarily seniors in the last quarter of their undergraduate studies, develop skill at short-term weather forecasting. The opportunity for such an analysis is afforded by the controlled setting of ATMS 452, and the consistency over the years in its forecast segment.
The typical student requires about 6 weeks or about 25 forecasts (each involving multiple parameters at up to 4 different locations) to gain proficiency in next-day forecasts of clouds, winds and temperatures (the so-called Type1 forecasts). Beyond this point, improvements in skill are minimal on average. Based on the students’ success at forecasting for unfamiliar as well as familiar locations, it appears that this proficiency arises both from practice in the drill of forecasting and from the development of local knowledge, i.e., of the nature in the weather associated with particular locations. While the best student forecasters have comparable skill to the instructor (NAB) during the latter portion of the class, his prior experience gives him a sizable advantage early in the class. Additional support for the importance of experience is provided by the results based on the intensive forecasts for Seattle: the flat learning curves for the students in this category reflect presumably their pre-existing knowledge of Seattle weather and their prior forecast practice during the earlier portion of the class.

The sequences of daily forecast scores for the forecasts involving precipitation (Type2) reveal that typical students have almost immediate skill at a level not much lower than NAB. This result may be attributable to all forecasters relying on basically the same suite of model precipitation forecasts and the inability of humans to enhance such numerical predictions. Our results suggest that the top student forecasters are better in maintaining high skill levels through the latter portion of the term, when convective precipitation is more prevalent, compared to other groups of students and NAB.

Student forecast and test scores for ATMS 452 are only moderately correlated (r ~0.4), with some students with lesser grades proving to be excellent forecasters. The finding that some highly capable forecasters are not necessarily high academic achievers should be considered by employers seeking to fill forecasting positions.

We conclude with some musings on what it takes to become a capable forecaster based not just on the results presented above, but also our own interactions with the students. First and foremost, is the role of experience. As discussed by Roebber (2005), many and perhaps most individuals learn most effectively through the use of concrete examples, and there is nothing like a busted forecast to bring home a point. A large number of realizations are required to understand the forecasting process and to appreciate the many ways a forecast can fail. Novice forecasters tend to be overconfident...
with their probabilistic forecasts (e.g., Doswell 2004). Indeed, for the Type2 forecasts in ATMS 452, our students forecast precipitation probabilities of either 0% or 100% frequently early in the quarter, but quickly gain appreciation for the inherent uncertainties in forecasting precipitation. Based on our conversations with the students, particularly during group discussions held right after forecast preparation, it appears that the repetition also helps the students develop a consistent and complete method for examining and processing a large amount of forecast information and effectively accounting for all the elements that go into the prediction. This repetition also helps them form conceptual models of the weather tailored to specific locations. As mentioned above, for the typical student about six weeks of intensive forecasting is required to develop competency.

Further gains in proficiency in forecasting are probably more subtle and difficult to measure. The very best forecasters are distinguished by their ability to handle unusual situations. Stuart et al. (2007) suggest that this success arises through the ability to mix analytic approaches with more heuristic, intuitive methods. We expect that this necessitates the deep knowledge base that is gained with years of scrutiny of a region’s weather. Furthermore, the ability to forecast relatively infrequent occurrences demands an extended training period.

Acknowledgements

This work was supported in part by a grant from the National Science Foundation (ATM-0504028).
References


Table 1  Type 1 forecast parameters and scoring

<table>
<thead>
<tr>
<th>Parameter - Units</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceiling – Low; Medium; High</td>
<td>“0” if correct; “1” if incorrect except “2” if forecast “High” and verified as “Low”</td>
</tr>
<tr>
<td>Visibility – Low; Medium; High</td>
<td>Same as Above</td>
</tr>
<tr>
<td>Sky Cover – CLR; SCT; BKN-OVC</td>
<td>“0” if correct; “1” if incorrect</td>
</tr>
</tbody>
</table>
| Wind Direction (Degrees) | “0” Error < 40 deg.  
“1” 40 < Error < 90 deg.  
“2” Error > 90 deg. |
| Wind Speed (Knots) | “0” Error < 5 kts.  
“1” 5 < Error < 10 kts.  
“2” 10 < Error < 15 kts.  
“3” Error > 15 kts. |
| Temperature (deg. F) | “0” Error < 3 F  
“1” 3 < Error < 6 F  
“2” 6 < Error < 9 F  
“3” 7 < Error < 12 F  
“4” 8 < Error < 15 F  
“5” Error > 15 F |
| Dew Point (deg. F) | Same as Above |

Table 2  Type 2 forecast parameters and scoring

<table>
<thead>
<tr>
<th>Parameter - Units</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precipitation – Probability in %</td>
<td>[Error/10]^2</td>
</tr>
<tr>
<td>Thunderstorm – Probability in %</td>
<td>Same as Above</td>
</tr>
<tr>
<td>Categorical QPF: 0-Trace; .01-.05”; .06-.20”; .21-.50”; &gt;.50”</td>
<td>Error (# of Categories) x 10</td>
</tr>
</tbody>
</table>
Fig. 1a  Sample scores for Type1 forecasts for instructor Nick Bond (NAB) (red), persistence (orange), three randomly selected students (other colors).
Fig. 1b As in Fig. 1a, but for Type2 forecast scores.
Fig. 2 Time series of Type1 forecast skill for the median student (green), highest two students (yellow), lowest two students (blue), and NAB (red) with respect to forecast day, averaged over 1997-2007.
Fig. 3  As in Fig. 2, but for Type2 forecast skill.
Fig. 4  As in Fig. 2, but for the local (Seattle) forecasts.
Fig. 5 Overall forecast grade (ordinate) versus average (midterm and final) test grade (abscissa) in ATMS 452. The linear trend of the forecast grades on the test grades is also shown; the correlation coefficient between these two parameters is ~0.4.