Elizabeth J. Suess, Dr. Cinzia Cervato, Dr. William A. Gallus, Jr., Jon Hobbs
Department of Geological and Atmospheric Sciences Iowa State University, Ames, IA

1. Introduction

Weather has always been a popular conversation starter. Even non-scientists find weather a fascinating topic. The biggest interest related to weather is usually forecasting. Meteorologists are constantly being asked "What is the weather going to be like?"

Weather forecasting is a crucial skill for any aspiring meteorologist. Most meteorology programs have some form of forecasting contest or a class in which the basics of forecasting are taught and applied. A few studies have been done in the past to determine how forecasting skill improves over time for these human forecasters.

Bosart (1975) studied how forecasting skill changed over time for meteorology students at the State University of New York in Albany. Forecasting skill was defined as an improvement over persistence forecasts. He discovered that changes in forecasting skill appeared to be only a function of temperature change within the forecast. Other aspects of the forecast, such as changes in wind speed or precipitation, did not seem to have an overall effect on the change in forecasting skill for the students. He also noted that as students forecast, their skill reaches a plateau after which they no longer improve.

Bond and Mass (2009) did a similar study of meteorology students taking an upper-level forecasting class at the University of Washington. They studied how forecasting skill improved over time by looking at 10 years worth of forecasting data. They found that the forecasters improved for the first 25 forecasts, after which time they showed minimal improvement.

Cervato et. al. (2009) 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 tended to do in the class.

The objective of this research is to expand on those findings to see if the use of the DWF allowed students to improve their forecasting skills over time or if the plateau observed by Bosart (1975) and Bond and Mass (2009) was also found in this student population. The key difference between this study and previous studies is the focus on non-meteorology students and the more extensive set of questions included in the DWF.

2. Data and Methodology

Forecasts for 218 students were collected during the Spring 2010 semester of MTEOR 206 at Iowa State University. Students are required to make at least 25 forecasts throughout the semester as well as listen to 5 to 10 minute weather discussions at the beginning or end of the class period. MTEOR 206 is open to all majors at Iowa State University and is required for freshmen in the meteorology program.

Forecasts were made by entering values in DWF. Parameters include 12 and 18 UTC temperature, cloud cover, temperature advection, and frontal passage as well as 18 UTC wind direction and wind speed, the likelihood of precipitation over 24 hours (12-12 UTC), and the factors that could cause precipitation (Table 1). All forecasts were made for Des Moines, IA (KDSM).

The DWF was originally developed by Yarger et. al. (2000) and is designed to allow many students to enter a forecast. It also is unique in that there are questions relating to advection, cloud cover, and fronts, items for which a student cannot quickly find an expert forecast via the internet. The best use of this DWF is in large-scale courses where there are

Corresponding author address: Elizabeth J. Suess, 4733 Toronto Street Apt, 206, Ames, IA. Email: ejsuess@iastate.edu.

Question	Correct Range	Type 1/2
12,18 UTC Temperature	± 5 degrees C	Type 1
12, 18 UTC Clouds		Type 2
12, 18 UTC Advection		Type 2
18 UTC Wind speed	±5 knots	Type 1
18UTC Wind direction	± 90 degrees	Type 1
24 hr Precipitation		Type 2
Precipitation Factors		Type 2

Table 1. Details on the forecasting questions. The middle column shows the acceptable ranges counted as correct. The last column classifies the question into type 1 (traditional) or a type 2 (more complex) forecast.

upwards of 300 students. The DWF automatically grades each forecast as the data comes in, so students get the results for their forecast promptly and accurately.

Forecasts were scored out of 36 total points. Each student was given one point per question for an incorrect forecast and 3 points for getting the question correct. The forecasting period was between 13 January and 28 April 2010. Forecasts could be made on any day of the week as long as they were submitted by 6 UTC the day before the forecasting period. Two expert forecasters were chosen to provide an "expert" forecast for each day of the forecasting period that could be used as a guideline to evaluate student performance The two experts were chosen because they showed exemplary forecasting skill in upper level meteorology classes.

The data were sorted both by date and by the number of forecasts each student made to eliminate any trends due to the date on which they made their forecasts. Because forecasts were made during the spring semester, they tended to be harder to make towards the end of the semester due to the changing nature of Iowa weather in the spring. Since the weather changes dramatically, a persistence forecast does markedly worse after the beginning of meteorological spring (See Fig 1.). The persistence forecast was calculated taking the correct answer from the previous forecast and applying it as the forecast for the next day.

To try to eliminate any trends that were simply a function of tougher forecasts, each student's forecast was compared to the persistence forecast and adjusted using (1) below.

$$FS = F - P \qquad (1)$$

The persistence forecast (P) was subtracted from every student's total forecast (F) similar to Bond and Mass (2009). Because the forecast scores were tallied differently than a typical forecast (high scores indicate good forecasts as opposed to the typical lowest score being the best forecast), it was not necessary to divide the equation by the persistence forecast.

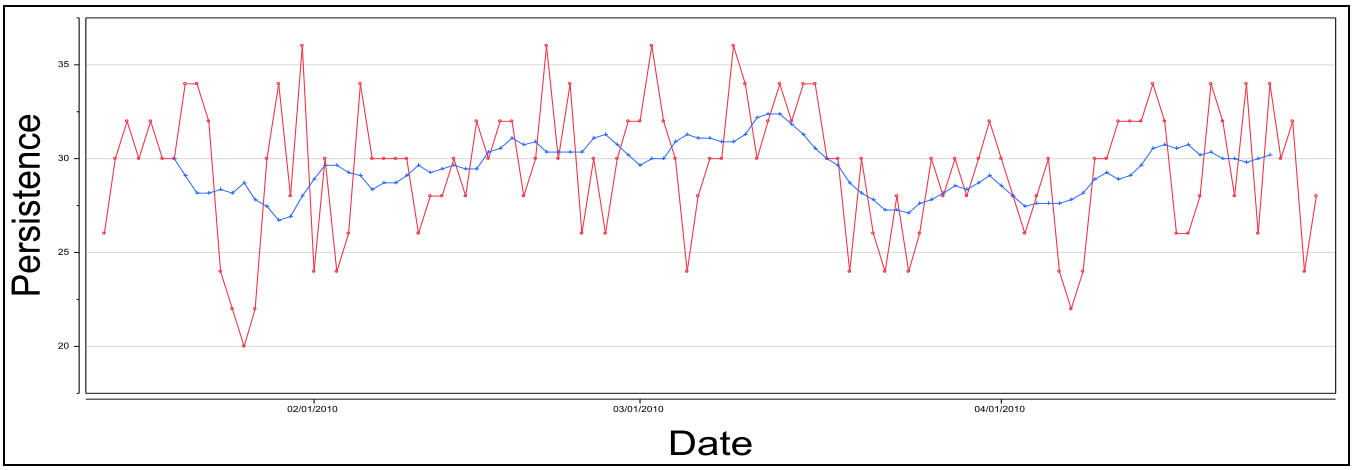


Figure 1: Scores for the persistence forecast (red line) for each day for the semester. The blue line is a running 10 day average.

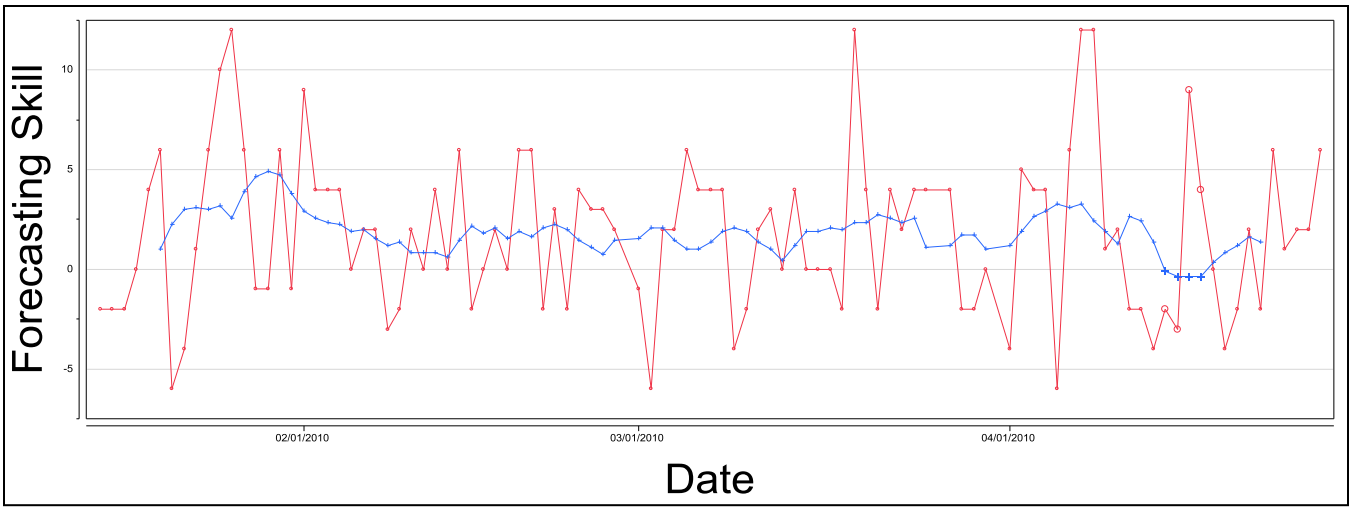


Figure 2.:Forecasting skill for expert forecasters (red) with a 10-day average (blue) overlaid to indicate trends.

To be sure that the equation adequately accounted for changes in forecast difficulty, the forecasting skill equation was applied to the expert forecasters. Minimal change was observed in the expert forecaster performance (Fig. 2), suggesting that the equation may be adequate for eliminating any biases due to difficult forecasts (Bond and Mass 2009).

The data were also divided into different types of forecasts to help understand trends in the scores. A "type 1" forecast was similar o most forecasts used traditionally in classrooms, and took just the scores for the 12 and 18 UTC temperature, 18 UTC wind speed, and wind direction and added them up for a total type 1 score out of 12 (See Table 1). The "type 2" forecast included the 12 and 18 UTC cloud cover, advection, precipitation in the past 24 hours, and the factors that could potentially cause precipitation, giving a maximum possible type 2 score of 18 (See Table 1). There was an issue with the grading of frontal passages so theses questions were ignored.

Similar to Cervato et. al. (2009), the data were also categorized by the date of the first forecast to see if students who started forecasting early did better than those students who waited until the last minute to make all of their forecasts.

These different methods of sorting the data were analyzed using JMP (See Appendix A1) to see if there were any noticeable trends in the forecasting skill. There is also a setting in JMP where a curve can be fit to the data using the data itself instead of a specific equation. This method was used to see how well the original curves fit the data and which of them best approximates the data. Specific trends that were analyzed were linear (straight line), parametric curve with a pre-specified shape (the plateau) shown in (2), and a non-parametric approach that give the data the freedom to define the trend.

$$f(x_{ij}) = y_l \exp(\beta x_{ij}) + y_h (1 - \exp(\beta x_{ij})) \quad (2)$$

3. Statistical Analysis

The data were examined as a function of date in order to see how forecasting scores changed with time. These data were fit with both a best fit line as well as a best fit quadratic curve. Then the p-values for each were evaluated to see which had the best fit.

A spline method was also used to attempt to find a best fit. The spline method allows the data to define the curve instead of fitting a pre-defined curve to the data to determine if any of the pre-defined curves were a potential good fit.

An F-distribution is a statistical model used to test the goodness of the fit of the nonlinear plateauing curve (3). In order to calculate F, the sum of squares error for the nonlinear function ($SSE_{reduced}$) was compared with the sum of squares error for the curve that was fit to the data (SSE_{full}). The difference in degrees of freedom

$$F = \frac{\frac{SSE_{reduced} - SSE_{full}}{df_{reduced} - df_{full}}}{\frac{SSE_{full}}{df_{full}}} \quad (3)$$

between the models was also needed ($df_{reduced}$-df_{full}) to compute this. Once the F-value was computed, the value was used in JMP along with the difference in degrees of freedom and the degrees of freedom of the curve acquired from the data. By using JMP to calculate the F-distribution, the p-value is easily obtained.

A final way used to evaluate the trends in the data was to normalize the number of forecasts for each student. In order to normalize the forecast, each student's forecasts were tallied into total number of forecasts (T) and each was assigned a count (C), with the earliest forecast being assigned "1." Then, (4) was used to calculate the variable, normalized counter (N). This allowed all of the forecasts to be plotted starting at 0 and ending at 1 to eliminate outliers since not every student forecasted the same number of times.

$$N = \frac{C}{T} \quad (4)$$

4. Results

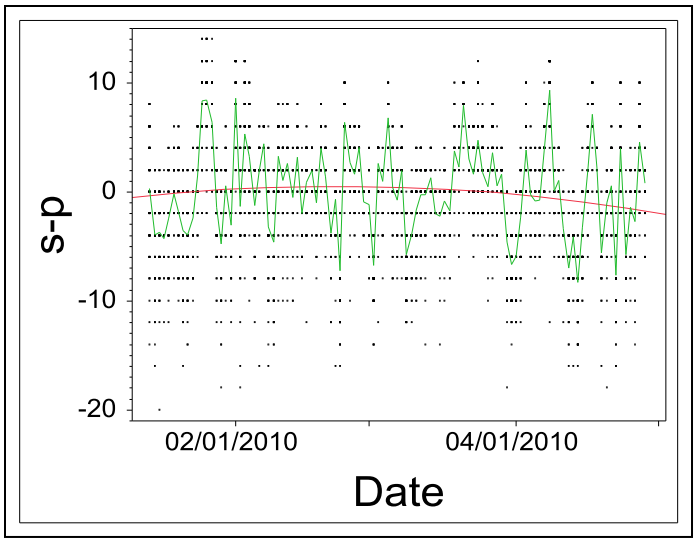


Figure 3. Student forecasting skill as a function of date in the semester. The red line shows the trend of the scores through the semester and the green line is the average forecast skill for each day.

a) Analysis by date

When each student's forecast was examined as a function of date, an initial improvement could be seen followed by a peak and then a decrease in overall forecasting skill. The t-test on this best fit curve showed that it fell within a 95% confidence interval. There are several factors that could indicate why such trends are present. It is possible that more students are forecasting towards the end of the semester as opposed to the beginning, which could cause the

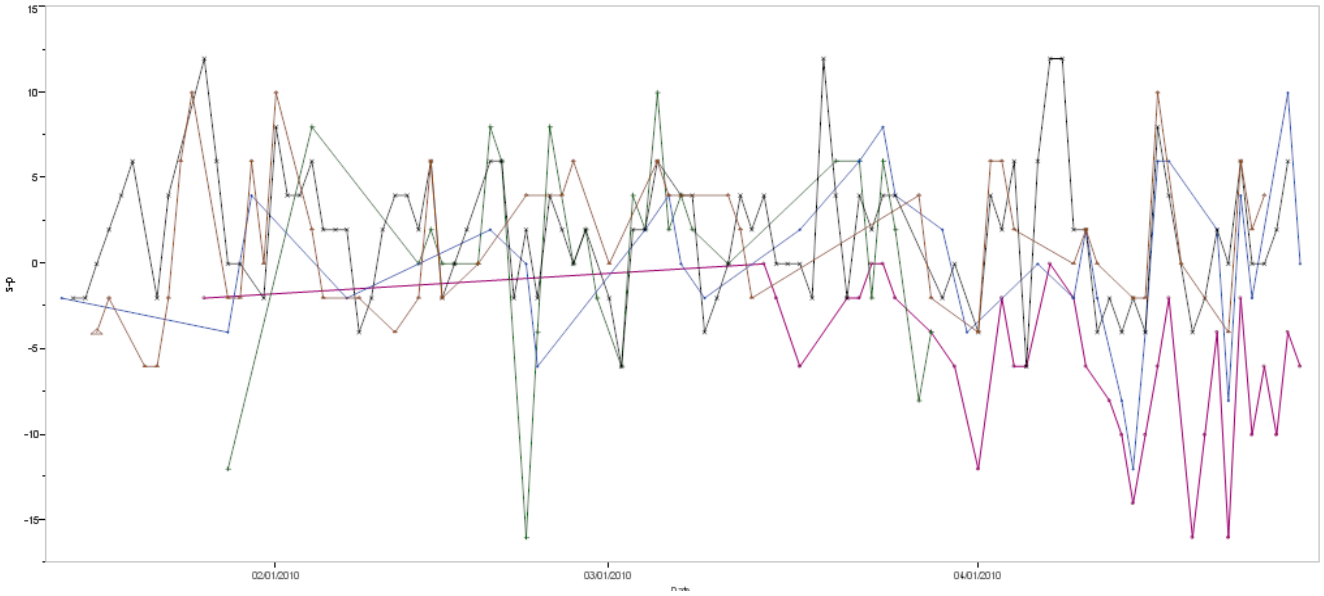


Figure 4.: A plot of the lowest student, the median student, the high student, and the two expert forecasters. The black and brown lines are the two expert forecasters, while the purple line is the low student, the green line is the highest student, and the blue line is the median student.

average to go down if more students that had difficulty with forecasting were trying then. Also, typically forecasts get harder in the spring due to more dramatic temperature changes, and increasingly convective precipitation in Iowa.

To eliminate any biases due to changing forecasting difficulty during the forecasting period, equation 1 was applied as discussed earlier. This application did not change results noticeably, with an initial upward trend followed by a decline towards the end of the semester (Fig. 3). This trend implies that perhaps the persistence adjustment does not fully account for changes in forecasting difficulty.

Student scores were also compared to those of the two expert forecasters. The student with the highest average score, the student with the median average score, the student with the lowest average score, and the two expert forecasters were plotted on one graph (Fig. 4). This plot suggests that students' forecasting scores decrease as the semester advances because forecasts indeed are more difficult later in the semester, since the lowest-scoring student made almost all forecasts at the end of the semester, while the highest-scoring student made all of them at the beginning. The median-scoring student made forecasts throughout the semester.

b) Analysis by count

Since not all students forecasted on the same day, another analysis was performed where normalization occurred, and each forecast was assigned a number, with the earliest forecast given the number "1." A nonlinear plateauing curve was then fit to the data to see if an upward trend similar to Bond and Mass (2009) was plausible. In our case, the students showed an initial upward trend through 10 or so forecasts and then leveled off (Fig. 5).

c) Normalized analysis

Due to the fact that not all of the students forecasted the same number of times, the forecasts were normalized. The normalized forecasts showed a decline in overall forecasting score (Fig. 6). Even when the data were separated into those students who forecasted 25 times or less and those who forecasted more than

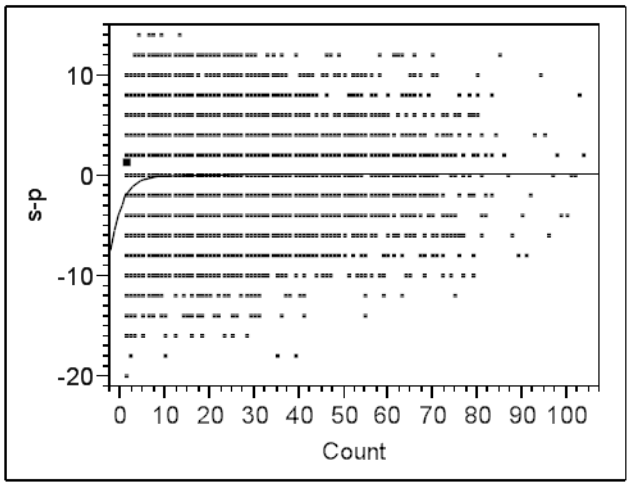


Figure 5: The nonlinear plateau curve fit to the students' forecasts plotted by a counter variable.

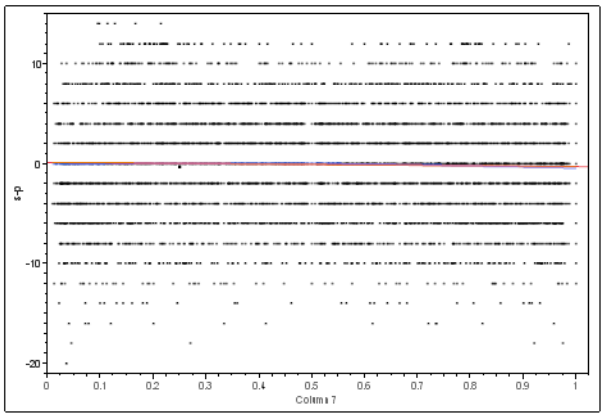
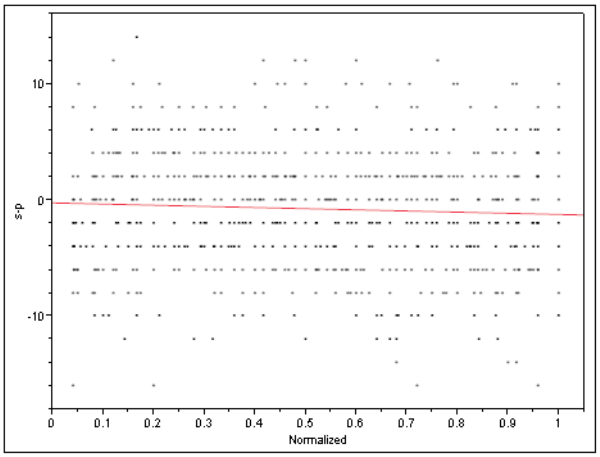


Figure 6: The normalized data plotted with a best fit line, shown in red.

25 times, a decreasing trend could be seen (Fig. 7a,b). One thing to note is that the students who forecasted more than the required 25 forecasts showed less of a decrease. This seems to imply that the students who forecast more show more of an improvement than those who don't forecast very much.

The fact that some questions used in this forecast were very different from traditional forecasting activities might also account for the downward trend later in the semester. To explore this hypothesis, the data were sorted into type 1 and type 2 forecasting scores. The type 1 scores showed that students could consistently forecast temperature and wind at a level above a persistence forecast; however, there was still a decrease in their forecasting abilities throughout the semester (Fig. 8). For type 2 questions forecasting scores had around the same



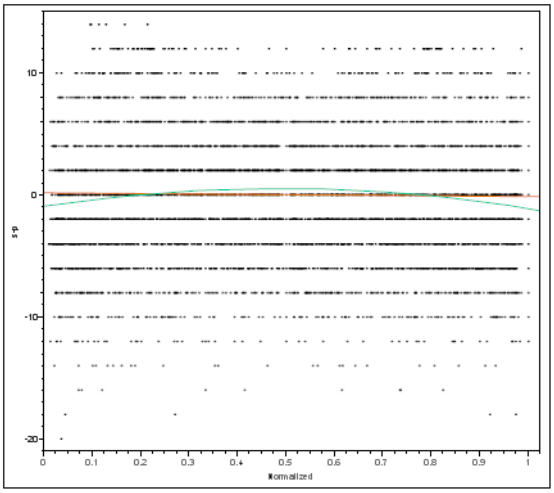


Figure 7: a) Normalized forecast skill for those who forecasted 25 times or less. The red line is the best fit line for these data. b) Normalized forecast skill for those students who forecasted greater than 25 times. Both a best fit line (red) and a polynomial curve (green) showed a statistically significant approximation of the data.

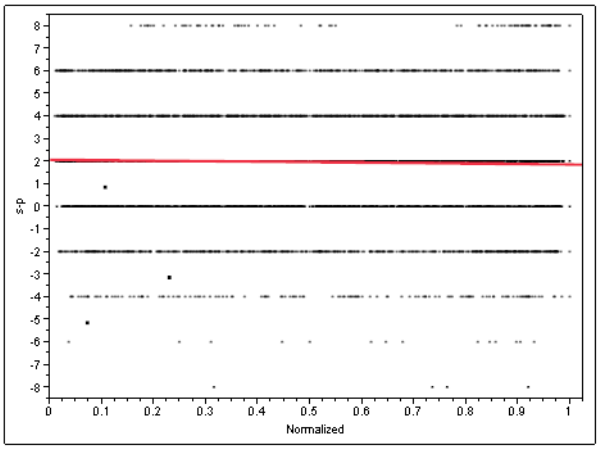


Figure 8: Normalized type 1 forecast scores with trend line plotted in red.

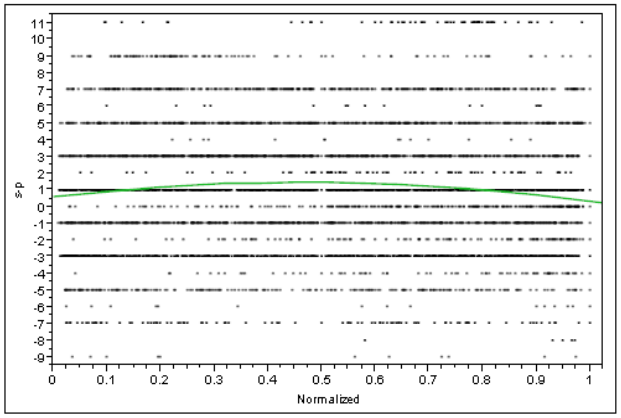


Figure 9: Normalized type 2 forecast scores with trend line plotted in green.

forecasting skill as persistence, and students again decreased in forecasting skill as they continued to forecast (Fig. 9).

5. Conclusions

An analysis of student forecasting skill in a 12 question forecasting activity undertaken by a large-lecture introductory meteorology course shows that while students may initially improve during the first few forecasts they make, they never reach a higher skill than persistence and overall show a slight downward trend. Significant improvement is restricted to the first 10-15 forecasts, a period shorter than that found in upper-level major courses (e.g. Bond and Mass 2009). The plateau appears to be right at the zero line, which indicates that an average student's forecasting skill will never be above the persistence forecast.

Because of the non-traditional nature of some of the questions in this DWF activity, the forecasts were separated into two types. 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 slightly above that of a persistence forecast. For type 2 forecasts, however, student forecasting skill remains close to zero, suggesting that these types

of questions may explain the relatively poor performance of students overall in the activity.

The results of this study imply several items. First, students may not get enough of a background in this introductory meteorology class to be able to understand the questions that are asked. The upper level students from Bond and Mass (2009) may have shown improvement over persistence due to the fact that they have a better understanding of how the atmosphere works and thus have a larger toolkit from which to draw to improve their forecasts. Another possible reason for the students never doing better than persistence is that forecasts are especially challenging in the spring in this region.

Future work should include evaluation of the activity during the fall semester to see if the trends are similar to those found in the spring for an introductory level course. Also, it would be interesting to study the trends among students within different majors of study taking the course. Is it possible that science majors would consistently do better than non-science majors? Gender may also play a role in affecting forecasting performance.

6. Acknowledgements

This work began as a senior 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.

REFERENCES

Bond, Nicholas A., Clifford F. Mass, 2009: Development of Skill by Students Enrolled in a Weather Forecasting Laboratory*. *Wea. Forecasting*, **24**, 1141–1148.

Bosart, Lance F., 1975: SUNYA Experimental Results in Forecasting Daily Temperature and Precipitation. *Mon. Wea. Rev.*, **103**, 1013 –1020.

Bosart, Lance F., 2003: Whither the Weather Analysis and Forecasting Process?. *Wea. Forecasting*, **18**, 520–529.

Cervato, C., W. Gallus, P. Boysen, and M. Larsen (2009), Today's Forecast: Higher Thinking With a Chance of Conceptual Growth, *Eos Trans. AGU*, *90*(20).

Roebber, Paul J., Lance F. Bosart, 1996: The Contributions of Education and Experience to Forecast Skill. *Wea. Forecasting*, **11**, 21–40.

Yarger, Douglas N., William A. Gallus, Michael Taber, J. Peter Boysen, Paul Castleberry, 2000: A Forecasting Activity for a Large Introductory Meteorology Course. *Bull. Amer. Meteor. Soc.*, **81**, 31–39.

7. Appendix A

1. JMP

JMP is a statistical program that allows data sets to be imported as tables and manipulated. Columns can be added to account for more data, and formulas can be applied to columns if the same statistic needs to be calculated for multiple lines. The software can also be used to plot graphs of whatever data sets are imported into the program. It also has the capability to assign best fit lines or curves to the data set. A trend line using the data itself can also be set (called a spline). There are a number of other functions that can be used in this program but this research mainly focuses on assigning best fit curves and splines.